Lecture Notes
on

Information Theory

Univ.-Prof. Dr. rer. nat. Rudolf Mathar
RWTH Aachen University
Institute for Theoretical Information Technology
Kopernikusstr. 16, 52074 Aachen, Germany

Ti Chair for Theoretical
Information Technology

RWTH AACHEN
UNIVERSITY

Contents

1 Introduction

According to Merriam-webster.com *”Information is any entity or form that provides the answer to a question of some kind or resolves uncertainty. It is thus related to data and knowledge, as data represents values attributed to parameters, and knowledge signifies understanding of real things or abstract concepts.”*.

However, modern *Information Theory* is not a theory which deals with the above on general grounds. Instead, information theory is a mathematical theory to model and analyze how information is transferred. Its starting point is an article by Claude E. Shannon, ”A Mathematical Theory of Communication”, Bell System Technical Journal, 1948.

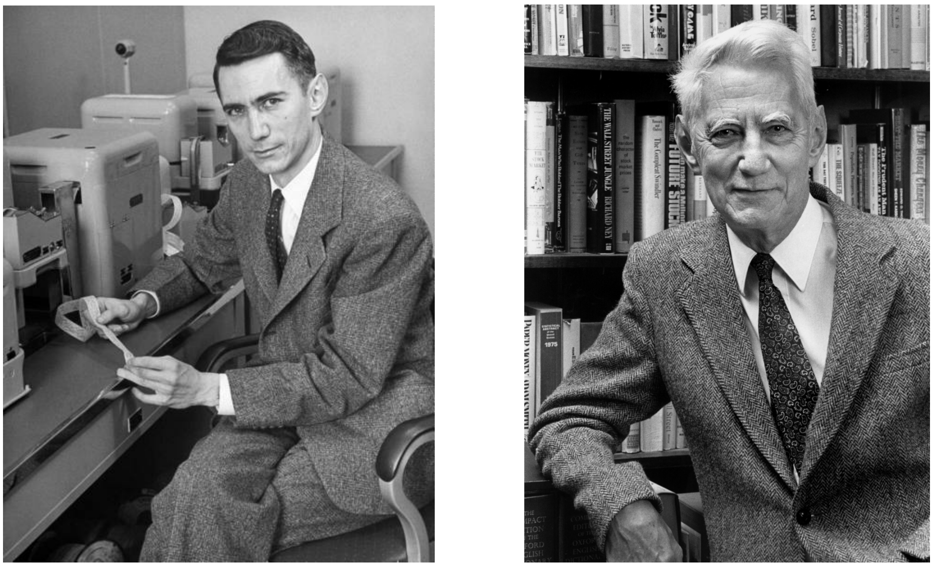


Figure 1.1: Claude Elwood Shannon (1916 – 2001)

Quoting from the introduction of this article provides insight into the main focus of information theory: ”*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning. . . . These semantic aspects of communications are irrelevant to the engineering problem. . . . The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.*”

Later, in 1964 a book by Claude E. Shannon and Warren Weaver with a slightly modified title ”The Mathematical Theory of Communication”appeared at University of Illinois Press, emphasizing the

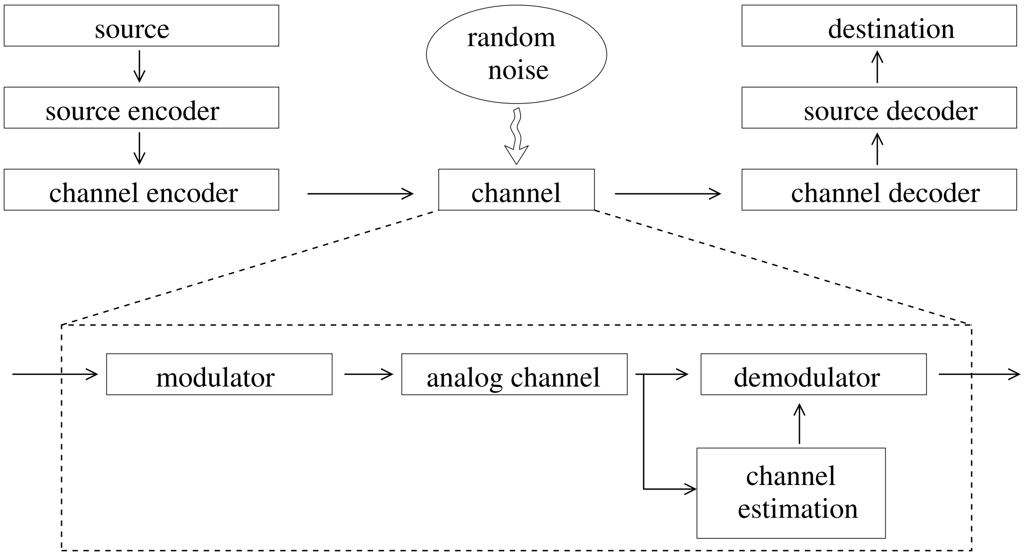


Figure 1.2: The general model of a communciation system.

generality of this work.

Information Theory provides methods and analytical tools to design such systems. The basic components of a communication are shown in Fig. 1.2. Although the starting point of information theory was in electrical engineering and communications, the theory turned out to be useful for modeling phenomena in a variety of fields, particularly in physics, mathematics, statistics, computer science and economics. It cannot be regarded only as a subset of communication theory, but is much more in general. In recent years, its concepts were applied and even further developed in biological information processing, machine learning and data science.

This lecture focuses on the latter. We first provide the basic concepts of information theory and prove some main theorems in communications, which refer to source coding, channel coding and the concept of channel capacity. This will be followed by the relation between rate distortion theory and autoencoders. Biological information processing will be modeled and analyzed by the concept of mutual information. This will finally lead to artificial neural networks and contributions of information theory to understanding how such networks learn in the training phase.

2 Fundamentals of Information Theory

2.1 Preliminary Definitions

In this chapter we provide basic concepts of information theory like entropy, mutual information and the Kulback-Leibler divergence. We also prove fundamental properties and some important inequalities between these quantities. Only discrete random variables (r.v.) will be considered, denoted by capitals X, Y and Z and having only finite sets of possible values to attain, the so called support. Only the distribution of the r.v. will be relevant for what follows. Discrete distribution can be characterized by stochastic vectors, denoted by

$$\boldsymbol{p} = (p_1, \dots, p_m), p_i \geq 0, \sum_i p_i = 1.$$

For intuitively motivating a measure of uncertainty consider the following two random experiments with four outcomes and corresponding probabilities

$$\begin{aligned}\boldsymbol{p} &= (0.7, 0.1, 0.1, 0.1)\\ \boldsymbol{q} &= (0.25, 0.25, 0.25, 0.25)\end{aligned}$$

Certainly the result of the second experiment will be more uncertain than of the first one. On the other hand, having observed the outcome of the second experiment provides more information about the situation. In this sense, we treat information and uncertainty as equivalently describing the same phenomenon.

Now, an appropriate measure of uncertainty was introduced by Shannon in his 1948 paper. He did this axiomatically, essentially requiring three properties of such a measure and then necessarily deriving *entropy* as introduced below.

We start by requesting that the *information content* of some event E shall only depend on its probability $p = P(E)$. Furthermore the information content is measured by some function $h : [0, 1] \to \mathbb{R}$ satisfying the following axioms.

(i) h is continuous on $[0, 1]$

(ii) $h(p \cdot q) = h(p) + h(q)$

(iii) there is some constant $c > 1$ such that $h(1/c) = 1$

The first axiom (i) requires that a small change in p results in a small change of the measure of its information content. Number (ii) says that for two independent events E_1 and E_2 with probabilities p and q respectively the intersection of both, i.e., the event that both occur at the same time, has information content $h(p) + h(q)$. The information content shall hence be additive for independent events. Finally, by (iii) a certain normalization is fixed.

Now, if (i),(ii), and (iii) are satisfied by some information measure h then necessarily

$$h(p) = -\log_c(p), p \in [0,1].$$

A convenient description is achieved by introducing a discrete random variables X with finite support $\mathcal{X} = \{x_1, \dots, x_m\}$ and distribution $P(X = x_i) = p_i, i = 1, \dots, m,\ p_i \geq 0,\ \sum_i p_i = 1$. *Entropy* is then defined as the average information content of the events $\{X = x_i\}$.

Definition 2.1. *Let $c > 1$ be fixed.*

$$H(X) = -\sum_i p_i \log_c p_i = -\sum_i P(X = x_i) \log P(X = x_i)$$

is called **entropy** *of X or of $\mathbf{p} = (p_1, \dots, p_m)$.*

Remark 2.2.

a) $H(X)$ depends only on the distribution of X, not on the specific support.

b) If $p_i = 0$ for some i, we set $p_i \log p_i = 0$. This follows easily from continuity.

c) The base of the logarithm will be omitted in the following. After the base has been chosen, it is considered to be fixed and constant throughout.

d) Let $p(x)$ denote the probability mass function (pmf)*, also called* discrete density *of X, i.e.,*

$$p : \mathcal{X} \to [0,1] : x_i \mapsto p(x_i) = p_i.$$

Then $H(X)$ may be written as

$$H(X) = E\Big(\log \frac{1}{p(X)}\Big),$$

the expectation of the r.v. $\log \frac{1}{p(X)}$.

Example 2.3. *a) Let $X \sim Bin(1, p)$, i.e., $P(X = 0) = 1 - p$ and $P(X = 1) = p$, $p \in [0,1]$. Then*

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

b) Let $X \sim U(\{1, \dots, m\})$, i.e., $P(X = i) = \frac{1}{m}$ for all $i = 1, \dots, m$. Then

$$H(X) = -\sum_{i=1}^{m} \frac{1}{m} \log \frac{1}{m} = -\log \frac{1}{m} = \log m$$

Particularly if $m = 26$, the size of the Latin alphabet, then $H(X) = \log_2 26 = 4.7004$

c) *Consider the frequency of characters in written English regarded as precise estimates of corresponding probabilities.*

character	*A*	*B*	*C*	$\cdots$	*Y*	*Z*
probability	*0.08167*	*0.01492*	*0.02782*	$\cdots$	*0.01974*	*0.00074*

It holds that

$$H(X) = -0.08167 \log_2 0.08167 - \cdots - 0.00074 \log_2 0.00074 = 4.219 < 4.7004.$$

The last inequality holds in general as will be clarified later.

The extension of the above definition to two-dimensional or even higher dimensional random vectors and conditional distributions is obvious.

Let (X, Y) be a discrete random vector with support $\mathcal{X} \times \mathcal{Y} = \{x_1, \ldots, x_m\} \times \{y_1, \ldots, y_d\}$ and distribution $P(X = x_i, Y = y_j) = p_{ij}, p_{ij} \geq 0, \sum_{ij} p_{ij} = 1$.

Definition 2.4. *a)*

$$H(X,Y) = -\sum_{i,j} P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j) = -\sum_{i,j} p_{ij} \log p_{ij}$$

is called joint entropy of (X, Y).

b)

$$\begin{aligned} H(X \mid Y) &= -\sum_j P(Y = y_j) \sum_i P(X = x_i \mid Y = y_j) \log P(X = x_i \mid Y = y_j) \\ &= -\sum_{i,j} P(X = x_i, Y = y_j) \log P(X = x_i \mid Y = y_j) \log P(X = x_i \mid Y = y_j) \end{aligned}$$

is called conditional entropy or equivocation of X *given* Y

Theorem 2.5. *(Chain rule) chain rule*

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(Y \mid X)$$

Proof. Denote by $p(x_i)$, $p(x_i, y_j)$ and $p(y_j|x_i)$ corresponding probability mass functions. It holds that

$$\begin{aligned} H(X,Y) &= -\sum_{ij} p(x_i, y_j) \big[\log p(x_i, y_j) - \log p(x_i) + \log p(x_i) \big] \\ &= -\sum_{i,j} p(x_i, y_j) \log p(y_j|x_i) - \sum_i \underbrace{\sum_j \log p(x_i, y_j)}_{=p(x_i)} \log p(x_i) \big] \\ &= H(Y \mid X) + H(X) \end{aligned}$$

The second equality is shown analogously by interchanging the roles of X and Y. □

Theorem 2.6. *(Jensen' s inequality) If f is a convex function and X is a random variable, then*

$$Ef(X) \geq f(EX). \tag{$*$}$$

$(*)$ *holds for any random variable (discrete,abs-continuous,others) as long as the expectation is defined. For discrete random variable with distribution* $(p_1 ....p_m)$, $(*)$ *read as*

$$\sum_{i=1}^{m} p_i f(x_i) \geq f\left(\sum_{i=1}^{m} p_i x_i\right) \forall x_1.....x_m \in dom(f)$$

Proof. We prove this for discrete distributions by induction on the number of mass points. The proof of conditions for equality when f is strictly convex is left to the reader. For a two-mass-point distribution, the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

which follows directly from the definition of convex functions. Suppose that the theorem is true for distributions with $k-1$ mass points. Then writing $p'_1 = p_i \,/\, (1-p_k)$ for $i = 1, 2....k-1$, we have

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

where the first inequality follows from the induction hypothesis and the second follows from the definition of convexity. The proof can be extended to continuous distributions by continuity arguments. □

Prior to showing relations between the entropy concepts we consider some important inequalities

Theorem 2.7. *(log-sum inequality) Let $a_i, b_i \geq 0$, where $i = 1.....m$. Then*

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \sum_i a_i \log \frac{\sum_j a_j}{\sum_j b_j}$$

with equality if and only if $\frac{a_i}{b_i}$= *constant.*

We use the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ *if* $a > 0$, $0 \log \frac{0}{0} = 0$ *(by continuity)*

Proof. The function $f(t) = t \log t \geq 0$ is strictly convex, since $f''(t) = \frac{1}{t} > 0$, for $t > 0$. Assume without loss of generality that $a_i, b_i > 0$.

By convexity of f :

$$\sum_{i=1}^{m} \alpha_i f(t_i) \geq f\left(\sum_{i=1}^{m} \alpha_i t_i\right), \alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i = 1.$$

Setting $\alpha_i = \frac{b_i}{\sum_j b_j}, t_i = \frac{a_i}{b_i}$; it follows

$$\begin{aligned} & \sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \sum_i \frac{b_i}{\sum b_j} \frac{a_i}{b_i} \log\left(\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i}\right) \\ \Leftrightarrow \quad & \frac{1}{\sum_j b_j} \sum_i a_i \log \frac{a_i}{b_i} \geq \frac{1}{\sum_j b_j} \sum_i a_i \log\left(\sum_i \frac{a_i}{\sum_j b_j}\right) \\ \Leftrightarrow \quad & \sum_i a_i \log \frac{a_i}{b_i} \geq \sum_i a_i \log \frac{\sum_j a_j}{\sum_j b_j} \end{aligned}$$

□

Corollary 2.8. *Let* $\boldsymbol{p} = (p_1, ..., p_m)$, $\boldsymbol{q} = (q_i..., q_m)$ *be stochastic vectors, i.e* $p_i, q_i \geq 0$, $\sum_i p_i = \sum_i q_i = 1$. *Then*

$$-\sum_{i=1}^{m} p_i \log p_i \leq -\sum_{i=1}^{m} p_i \log q_i,$$

equality holds if and only if $p = q$.

Proof. In theorem 2.7, set $a_i = p_i$, $b_i = q_i$ and note that $\sum_i p_i = \sum_i q_i = 1$. □

Theorem 2.9. *Let* X, Y, Z *be discrete random variable as above*

a) $0 \overset{(i)}{\leq} H(X) \overset{(ii)}{\leq} \log m$ *Equality in* $(i) \iff X$ *has a singleton distribution. i.e* $\exists x_i : P(X = x_i) = 1$ *Equality in* $(ii) \iff X$ *has a uniformly distribution. i.e* $P(X = x_i) = \frac{1}{m} \forall i = 1....m.$

b) $0 \overset{(i)}{\leq} H(X|Y) \overset{(ii)}{\leq} H(X)$ *Equality in* $(i) \iff P(X = x_i|Y = y_j) = 1 \forall (i,j)$ *with* $P(X = x_i, Y = y_j) > 0$ *, i.e* X *is totally dependent on* Y *Equality in* $(ii) \iff X, Y$ *are stochastically independent*

c) $H(X) \overset{(i)}{\leq} H(X,Y) \overset{(ii)}{\leq} H(X) + H(Y)$ *Equality in* $(i) \iff Y$ *is totally dependent on* X *Equality in* $(ii) \iff X, Y$ *are stochastically independent*

d) $H(X|Y,Z) \leq \mathit{min}\{H(X|Y), H(X|Z)\}$

Proof. a) $(i) 0 \leq H(X)$, By definition. (ii)

$$\begin{aligned} H(X) &= -\sum_{i=1}^{m} p_i \log p_i = \sum_{i=1}^{m} p_i \log \frac{1}{p_i} \\ &\leq \log \big(\sum_{i=1}^{m} p_i \frac{1}{p_i} \big) \quad \text{(Jensen Inequality)} \\ &= \log \big(\sum_{i=1}^{m} 1 \big) = \log m \end{aligned}$$

b) Equality (i) holds iff $H(X|Y) = 0$, that is if X is a deterministic function of Y.

Similarly, (ii) holds iff $H(X) = H(X|Y)$. Since

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

we have that this holds only if $I(X;Y) = 0$, that is X and Y are statistically independent.

c) i) By the Chain rule, theorem 2.5 $H(X,Y) = H(X) + \underbrace{H(Y|X)}_{\geq 0} \geq H(X)$ with "equality" from b) (ii).

ii) From b) $0 \leq H(X) - H(X|Y)$. Using Chain rule theorem 2.5, we can write $H(X) - H(X|Y) = H(X) - [H(X,Y) - H(Y)]$. Hence, we get $H(X) + H(Y) \geq H(X,Y)$ with "equality" from b) (i)

d)

$$H(X|Y,Z) = H(X|Z) - \underbrace{I(X;Y|Z)}_{\geq 0} \leq H(X|Z) \, .$$

The same can be said for $H(X|Y,Z) \leq H(X|Y)$.

□

Definition 2.10. *Let X, Y, Z be discrete random variables*

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

is called **mutual information** *of X and Y.*

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z)$$

is called **conditional mutual information** *of X and Y given Z.*

Interpretation: $I(X;Y)$ is the reduction in uncertainty about X when Y is given or the amount of information about X provided by Y.

Relation between entropy and mutual information (See fig.2.1):

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

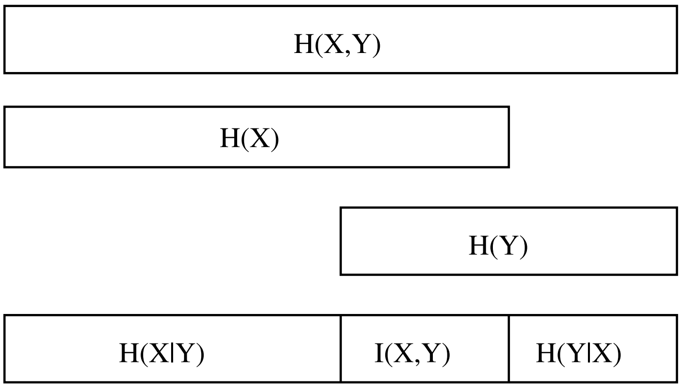


Figure 2.1: Relation between entropy and mutual information.

Note: By theorem 2.9 b), we know $I(X;Y) \geq 0$.

By definition it holds that

$$
\begin{aligned}
I(X;Y) &= -\sum_i p(x_i)\log p(x_i) + \sum_{i,j} p(x_i,y_j)\log p(x_i|y_j)\\
&= -\sum_{i,j} p(x_i,y_j)\log p(x_i) + \sum_{i,j} p(x_i,y_j)\log p(x_i|y_j)\\
&= \sum_{i,j} p(x_i,y_j)\log \frac{p(x_i|y_j)}{p(x_i)}\\
&= -\sum_{i,j} p(x_i,y_j)\log \frac{p(x_i,y_j)}{p(x_i)p(y_j)}
\end{aligned}
$$

which shows symmetry in X and Y.

Example 2.11. *(Binary symmetric channel,BSC)*

Let the symbol error with probability be ε, $0 \leq \varepsilon \leq 1$. Then,

$$P(Y=0|X=0) = P(Y=1|X=1) = 1-\varepsilon$$

$$P(Y=0|X=1) = P(Y=1|X=0) = \varepsilon.$$

Assume that input symbols are uniformly distributed $P(X=0) = P(X=1) = \frac{1}{2}$. Then for the joint distributions: $P(X=0,Y=0) = P(Y=0|X=0)P(X=0) = (1-\varepsilon) - \frac{1}{2}$, that is

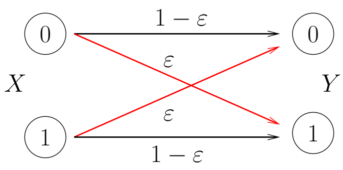


Figure 2.2: Binary symmetric channel .

X \ *Y*	*0*	*1*	
0	$\frac{1}{2}(1-\varepsilon)$	$\frac{(\varepsilon)}{2}$	$\frac{1}{2}$
1	$\frac{(\varepsilon)}{2}$	$\frac{1}{2}(1-\varepsilon)$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	

Further,

$$\begin{aligned}
P(X=0|Y=0) &= \frac{P(X=0,Y=0)}{P(Y=0)} = 1-\varepsilon \\
P(X=1|Y=1) &= 1-\varepsilon \\
P(X=0|Y=1) &= P(X=1|Y=0) = \varepsilon
\end{aligned}$$

For $\log = \log_2$

$$\begin{aligned}
H(X) = H(Y) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1bit \\
H(X,Y) &= 1-(1-\varepsilon)\log(1-\varepsilon) - \varepsilon\log\varepsilon \\
H(X,Y) &= -(1-\varepsilon)\log(1-\varepsilon) - \varepsilon\log\varepsilon \\
0 \le I(X,Y) &= 1+(1-\varepsilon)\log(1-\varepsilon) + \varepsilon\log\varepsilon \le 1
\end{aligned}$$

Definition 2.12. *(Kullbach-Leibler divergence, KL divergence)*
Let $\boldsymbol{p} = (p_i, ...p_n)$*,* $\boldsymbol{q} = (q_i, ..., q_n)$ *be stochastic vectors. Then*

$$D(\boldsymbol{p}||\boldsymbol{q}) = \sum_{i=1}^{n} p_i \log\frac{p_i}{q_i}$$

is called **KL Divergence** *between* **p** *and* **q** *(or relative entropy).*

$D(\boldsymbol{p}||\boldsymbol{q})$ measures the divergence (distance, dissimilarity) between distributions $\boldsymbol{p}$ and $\boldsymbol{q}$. However it is not a metric ,neither symmetric nor satisfies the triangle inequality.It measures how difficult it is for $\boldsymbol{p}$ to pretend it to be $\boldsymbol{q}$.

Theorem 2.13. *(Relative entropy)*

a) $D(\boldsymbol{p}||\boldsymbol{q}) \ge 0$ *with "equality" iff* $\boldsymbol{p} = \boldsymbol{q}$

b) $D(\boldsymbol{p}\|\boldsymbol{q})$ *is a convex in the pair* $(\boldsymbol{p},\boldsymbol{q})$

c) $I(X;Y) = D((p(x_i,y_j))_{i,j}\|(p(x_i)p(y_j)))_{i,j}) \geq 0$

Proof. a) Immediate by definition and Corollary 2.8

b) Use the log-sum inequality 2.7 Let p, r and q, s be stochastic vectors. For all $i = 1......n$ it holds

$$(\lambda p_i + (1-\lambda)r_i)\log\frac{\lambda p_i + (1-\lambda)r_i}{\lambda q_i + (1-\lambda)s_i} \leq \lambda p_i \log\frac{\lambda p_i}{\lambda q_i} + (1-\lambda)r_i \log\frac{(1-\lambda)r_i}{(1-\lambda)s_i}$$

summing over $i = 1.....n$ it follows $\forall\lambda \in [0,1]$,

$$D(\lambda p + (1-\lambda)r\|\lambda q + (1-\lambda)s) \leq \lambda D(\boldsymbol{p}\|\boldsymbol{q}) + (1-\lambda)D(\boldsymbol{r}\|\boldsymbol{s}).$$

c) By definition.

□

Note: $D(\boldsymbol{p}\|\boldsymbol{q}) \neq D(\boldsymbol{q}\|\boldsymbol{p})$**.**

Lemma 2.14. *For any distribution* $\mathbf{p},\mathbf{q}$ *with support* $\mathcal{X} = \{x_1,...,x_m\}$ *and stochastic matrix* $\mathbf{W} = (p(y_j \mid x_i))_{i,j} \in \mathbb{R}^{m\times d}$

$$D(\mathbf{p}\|\mathbf{q}) \geq D(\mathbf{pW}\|\mathbf{qW})$$

Proof. Let $\underline{w}_1,....,\underline{w}_d$ be the columns of $\boldsymbol{W}$, that is $\mathbf{W} = (\underline{w}_1,....,\underline{w}_d)$. Using the log-sum inequality 2.7

$$\sum_i a_i \log\frac{a_i}{b_i} \geq \sum_i a_i \log\frac{\sum_j a_j}{\sum_j b_j}$$

$$\begin{aligned}
D(p\|q) &= \sum_{i=1}^{m} p(x_i)\log\frac{p(x_i)}{q(x_i)} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{d} \underbrace{p(x_i)p(y_j \mid x_i)}_{a_i} \log\frac{\overbrace{p(x_i)p(y_j \mid x_i)}^{a_i}}{\underbrace{q(x_i)p(y_j \mid x_i)}_{b_i}} \\
&\geq \sum_{j=1}^{d} p\underline{w}_j \log\frac{p\underline{w}_j}{q\underline{w}_j}. \\
&= D(\mathbf{pW}\|\mathbf{qW})
\end{aligned}$$

□

Theorem 2.15. $H(\mathbf{p})$ *is a concave function of* $\mathbf{p} = (p_1,.....,p_m)$.

Proof. Let $\mathbf{u} = (\frac{1}{m},.....\frac{1}{m})$ be the uniform distribution. $D(\mathbf{p}\|\mathbf{u}) = \sum_{i=1}^{m} p_i \log \frac{p_i}{\frac{1}{m}} = \log m - H(\mathbf{p})$. Hence by theorem 2.13 b), $H(\mathbf{p}) = \log m - D(\mathbf{p}\|\mathbf{u})$

$$D(\lambda\mathbf{p} + (1-\lambda)\mathbf{q}\|\lambda\mathbf{u} + (1-\lambda)\mathbf{u}) \leq \lambda D(\mathbf{p}\|\mathbf{u}) + (1-\lambda)D(\mathbf{q}\|\mathbf{u}),$$

i.e., convexity in $\mathbf{p}$ thus ,$H(\mathbf{p})$ is a concave function of $\mathbf{p}$. □

2.2 Inequalities

Definition 2.16. *Random variables X, Y, Z are said to form a Markov chain in that order (denoted by X $\rightarrow$ Y$\rightarrow$ Z) if the joint probability mass function (discrete density) satisfies*

$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y)$$

For $X \rightarrow Y \rightarrow Z$, the conditional distribution of Z depends only on Y and is conditionally independent of X.

Lemma 2.17. *a) If $X \rightarrow Y \rightarrow Z$ then $p(x, z \mid y) = p(x \mid y)p(z \mid y)$.*

b) If $X \rightarrow Y \rightarrow Z$ then $Z \rightarrow Y \rightarrow X$.

c) If $Z = f(y)$, then $X \rightarrow Y \rightarrow Z$.

Proof. a)

$$\begin{aligned} p(x, z \mid y) &= \frac{p(x, z, y)}{p(y)} = \frac{p(x)p(y \mid x)p(z \mid y)}{p(y)} \\ &= \frac{p(x, y)p(z \mid y)}{p(y)} = p(x \mid y)p(z \mid y) \end{aligned}$$

b) If $X \rightarrow Y \rightarrow Z$ then $Z \rightarrow Y \rightarrow X$. $p(x, y, z) = p(x)p(y \mid x)p(z \mid y) = p(x, y)\frac{p(z,y)}{p(y)}\frac{p(z)}{p(z)}$= $p(z)p(y \mid z)p(x \mid y)$, i.e $Z \rightarrow Y \rightarrow X$

c) If $Z = f(y)$ then,

$$p(x, y, z) = \begin{cases} p(x, y), & \text{if } z = f(y). \\ 0, & \text{otherwise} \end{cases}$$

hence $p(x, y, z) = p(x, y)\mathbf{1}(z = f(y)) = p(x)p(y \mid x)p(z \mid y)$,

$$p(z \mid y) = \begin{cases} 1, & z = f(y). \\ 0, & \text{otherwise} \end{cases}$$

□

Theorem 2.18. *(Data-processing inequality) If $X \rightarrow Y \rightarrow Z$,then $I(X; Z) \leq \min\{I(X; Y), I(Y; Z)\}$, "No processing of Y can increase the information that Y contains about X".*

Proof. By chain rule

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

Since X and Z are conditionally independent given Y, we have $I(X;Y|Z) = 0$. Since $I(X;Y|Z) \geq 0$, we have

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

Equality holds iff $I(X;Y|Z) = 0$ i.e $X \to Z \to Y$ $I(X;Z) \leq I(Y;Z)$ is shown analogously. □

Theorem 2.19. *(Fano inequality) Assume X,Y are random variables with the same support $\mathcal{X} = \{x_1, ..., x_m\}$. Lets define $P_e = P(X \neq Y)$, the "error probability".*

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

This implies that $P_e \geq \frac{H(X|Y) - \log 2}{\log(m-1)}$.

Proof. We Know,

1) $H(X|Y) = \sum_{x \neq y} p(x,y) \log \frac{1}{p(x|y)} + \sum_x p(x,x) \log \frac{1}{p(x|x)}$
2) $P_e \log(m-1) = \sum_{x \neq y} \log(m-1) p(x,y)$
3) $H(P_e) = -P_e \log P_e - (1-P_e)\log(1-P_e)$
4) $\ln(t) \leq t - 1, t \geq 0$

Using this we obtain

$$\begin{aligned} H(X|Y) &= P_e \log(m-1) - H(P_e) \\ &= \sum_{x \neq y} p(x,y) \log \frac{P_e}{p(x|y)(m-1)} + \sum_x p(x,x) \log \frac{1-P_e}{p(x|x)} \\ &\leq (\log e) \left[\sum_{x \neq y} p(x,y) \left(\frac{P_e}{(m-1)p(x|y)} - 1 \right) + \sum_x p(x,x) \left(\frac{1-P_e}{p(x,x)} - 1 \right) \right] \\ &= (\log e) \left[\frac{P_e}{(m-1)} \sum_{x \neq y} p(y) - \sum_{x \neq y} p(x,y) + (1-P_e) \sum_x p(x) - \sum_x p(x,x) \right] \\ &= \log e \left[P_e - P_e + (1-P_e) - (1-P_e) \right] = 0 \end{aligned}$$

□

Lemma 2.20. *If X and Y are i.i.d random variables with entropy $H(X)$. Then*

$$P(X = Y) \geq 2^{-H(X)}$$

Proof. let $p(x)$ denote the p.m.f of of X. Use Jensen inequality $f(t) = 2^t$ is a convex function. Hence with $Y = \log p(x)$ we obtain $E(2^Y) \geq 2^{E(Y)}$, that is

$$\begin{aligned}
2^{-H(X)} &= 2^{E(\log p(X))} \leq E(2^{\log p(X)}) \\
&= \sum_x p(x) 2^{\log p(x)} \\
&= \sum_x p^2(x) \\
&= P(X = Y)
\end{aligned}$$

□

2.3 Information Measures for Random sequences

Consider sequences of random variables X_1,X_2.... denoted as $X = \{X_n\}_{n \epsilon N}$. A naive approach to define the entropy of X is

$$H(X) = \lim_{n \to \infty} H(X_1, ..., X_n).$$

In most cases this limit will be infinite. Instead consider the *entropy rate*.

Definition 2.21. *Let $X = \{X_n\}_{n \epsilon N}$ be a sequence of discrete random variables.*

$$H_\infty(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, ..., X_n)$$

is called the entropy rate of X, provided the limit exists. $H_\infty(X)$ may be interpreted as average uncertainty per symbol

Example 2.22. *a) Let $X = \{X_n\}_{n \epsilon N}$ be i.i.d random variable with $H(X_i) < \infty$, Then*

$$H_\infty(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, ..., X_n) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = H(X_i)$$

b) Let $X = \{X_n\}_{n \epsilon N} = \{(X_n, Y_n)\}_{n \epsilon N}$ be i.i.d sequence with $I(X_k; Y_k) < \infty$. Then

$$\begin{aligned}
I_\infty(X_k, Y_k) &= \lim_{n \to \infty} \frac{1}{n} I(X_1, ..., X_n; Y_1, ..., Y_n) \\
&= \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} I(X_k, Y_k) \\
&= I(X_1, Y_1)
\end{aligned}$$

Going further than i.i.d sequences, let us introduce in the following definition.

Definition 2.23. *A sequence of random variable X =$\{ X_n\}_{n\epsilon N}$ is called (strongly) stationary if*

$$p(X_{i_1}, ..., X_{i_k}) = p(X_{i_1+t}, ..., X_{i_k+t})$$

for all 1 $\leq i_1 <$ .....$< i_k$, t ϵ N.

- *The joint distribution of any finite selection of random variable from $\{X_n\}$ is invariant w.r.t shifts.*
- *An equivalent condition for discrete random variable with support X is as follows.*

$$P(X_1 = s_1, ..., X_n = s_n) = P(X_{1+t} = s_1, ..., X_{n+t} = s_n)$$

for all s_1,....,$s_n \in X, n \in N, t \in N$. For stationary sequences all marginal distributions $P(X_i)$ are the same.

Theorem 2.24. *Let $X = \{X_n\}_{n\epsilon N}$ be a stationary sequence. Then*

a) $H(X_n|X_1, \ldots, X_{n-1})$ *is monotonically decreasing*

b) $H(X_n|X_1, \ldots, X_{n-1}) \leq \frac{1}{n}H(X_1, ..., X_n)$

c) $\frac{1}{n}H(X_n|X_1, \ldots, X_n)$ *is monotonically decreasing*

d) $\lim_{n\to\infty} H(X_n|X_1, \ldots, X_{n-1}) = \lim_{n\to\infty} \frac{1}{n}H(X_1, ..., X_n) = H_\infty(X)$

Definition 2.25.

a) $X = \{X_n\}_{n\epsilon N_0}$ *is called a Markov chain (MC) with state shape* $X = \{X_1, \ldots, X_n\}$ *if*

$$P(X_n = s_n|X_{n-1} = s_{n-1}, ..., X_0 = s_0) = P(X_n = s_n|X_{n-1} = s_{n-1})\forall s_1 \epsilon N$$

b) It is called homogenous, if the transition probabilities $P(X_n = s_n|X_{n-1} = s_{n-1})$ *are independent of n.*

c) $p(0) = (p_1(0), ..., p_m(0)) \sim X_0$ *is called initial distribution.*

d) $\Pi = (p_{ij})_{1\leq i,j\leq m} = (P(X_n = j|X_{n-1} = i))_{1\leq i,j\leq n}$ *is called transition matrix.*

e) $\boldsymbol{p} = (p_1, ...p_m)$ *is called stationary if* $\boldsymbol{p}\Pi = \boldsymbol{p}$.

Lemma 2.26. *Let X= $\{ X_n\}_{n\epsilon N_0}$ be a stationary homogenous MC. Then*

$$H_\infty(X) = -\sum_{i,j} p_i(0)p_{ij} \log p_{ij}$$

Proof. By Theorem 2.24

$$\begin{aligned} H_\infty &= \lim_{n\to\infty} H(X_n|X_{n-1}, ..., X_0) \\ &= \lim_{n\to\infty} H(X_1|X_0) \\ &= -\sum_i p_i(0) \sum_j p_{ij} \log p_{ij} \\ &= -\sum_{ij} p_i(0) p_{ij} \log p_{ij}. \end{aligned}$$

□

Remark: A homogenous MC is stationary if $\boldsymbol{p}(0)\Pi = \boldsymbol{p}(0)$, i.e, if the initial distribution is a so called stationary distribution.

Example 2.27. *(2- state homogenous MC)*

Two states: $X = \{0, 1\}$ *Transition probabilities* Π = $\begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$, $0 \leq \alpha, \beta \leq 1$

Compute a stationary distribution $\boldsymbol{p} = (p_1, p_2)$,*solve* $\boldsymbol{p}\Pi = \boldsymbol{p}$

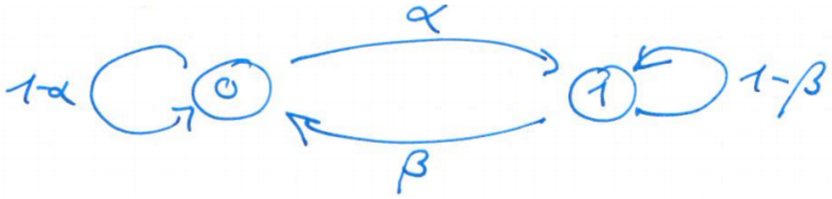


Figure 2.3: Transition graph

Solution: $\boldsymbol{p}^*$ = $(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$

Choose $\boldsymbol{p}(0) = \boldsymbol{p}^*$. *Then* $X = \{X_n\}_{n\epsilon N_0}$ *is a stationary MC with*

$H(X_n) = H(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$. *However*

$$H_\infty(X) = H(X_1|X_0) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta)$$

Example 2.28. *(Random walk as a weighted graph) Consider an indirected weighted graph*

Nodes $\{1, ..., m\}$. *Edges with weight* $w_{ij}, i < j = 1, ..., m, w_{ji} = w_{ij}$, *no edge means* $w_{i,j} = 0$.

Random walk on the graph $X = \{X_n\}_{n\epsilon N_0}$ *is a MC with support* $X = \{1, ..., m\}$ *and*

$$P(X_{n+1} = j|X_n = 1) = \frac{w_{ij}}{\sum_{k=1}^m} w_{ik} = p_{ij}, 1 \leq i, j \leq m.$$

Stationary distribution: (we guess it and then prove that it is actually the same)

$$p_i{}^* = \frac{\sum_j w_{ij}}{\sum_{ij} w_{ij}} = \frac{w_i}{w},$$
$$(p_i{}^*, ..., p_m{}^*) = \boldsymbol{p}^*$$

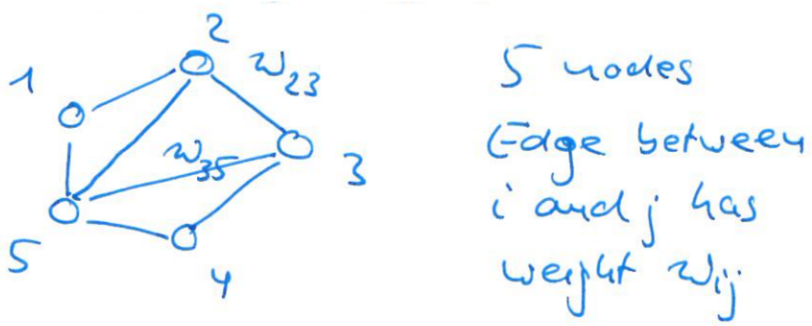


Figure 2.4: Random walk as a weighted graph

Assume that the random walk starts at time 0 *with the stationary distribution* $p_i(0) = p_i^*, i = 1, ..., m$. *Then* $X = \{X_n\}_{n\epsilon N_0}$ *is a stationary sequence (MC) and*

$$\begin{aligned} H_\infty(X) &= H(X_1|X_0) \\ &= -\sum_i p_i^* \sum_j p_{ij} \log p_{ij} \\ &= -\sum_i \frac{w_i}{w} \sum_j \frac{w_{ij}}{w} \log \frac{w_{ij}}{w_i} \\ &= -\sum_{ij} \frac{w_{ij}}{w} \log \frac{w_{ij}}{w_i} \\ &= -\sum_{ij} \frac{w_{ij}}{w} \log \frac{w_{ij}}{w} + \sum_{ij} \frac{w_{ij}}{w} \log \frac{w_i}{w} \\ &= H\left(\left(\frac{w_{ij}}{w}\right)_{i,j}\right) - H\left(\left(\frac{w_i}{w}\right)_i\right) \end{aligned}$$

If all edges have equal weight, then

$$p_i^* = \frac{E_i}{2E},$$

where E_i *=no of edges emanating from node* i *E=total no of edges*

In this case

$$H_\infty(X) = \log(2E) - H\left(\frac{E_1}{2E}, ..., \frac{E_m}{2E}\right)$$

H_∞*(X) depends only on the entropy of the stationary distribution and the total no of edges.*

2.4 Asymptobic Equipartition Property(AEP)

Information theory, the AEP is the analog of the law of large numbers (LNN).

LLN: Let $\{X_c\}$ be i.i.d r.v.s $X_i \sim X$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E(X) \text{ almost everywhere (and in prob.) as } n \to \infty\,.$$

AEP : X_i discrete i.i.d with joint pmf $p^{(n)}(X_i, ..., X_n)$ then

$$\frac{1}{n}\log\frac{1}{p^{(n)}(X_1, .., X_n)} \text{ is “close to” } H(X) \text{ as } n \to \infty\,.$$

Thus

$$p^{(n)}(X_1, ..., X_n) \text{ is “close to” } 2^{-nH(X)} \text{ as } n \to \infty\,.$$

“close to” must be made precise.

Consequence: Existence of the typical set with sample entropy close to true entropy and the nontypical set, which contain the other sequences.

Definition 2.29. *A sequence of random variable X_n is said to converge to a random variable X*

(i) in probability if $\forall\epsilon > 0,\ P(\mid X_n - X \mid > \epsilon) \to 0$ *as* $n \to \infty$

(ii) in mean square if $E(X_n - X)^2) \to 0$ *as* $n \to \infty$

(iii) with prob 1 *(or almost everywhere) if* $P(\lim_{n\to\infty} X_n = X) = 1$

Theorem 2.30. *Let $\{X_n\}$ be i.i.d discrete random variable $X_i \sim X$ with support X. $(X_1, ..., X_n)$ with joint pmf $p^n(X_1, ..., X_n)$. Then $-\frac{1}{n}\log p^{(n)}(X_1, ..., X_n) \to_{(n\to\infty)} H(X)$ in probability.*

Proof. $Y_i = \log p(X_i)$ are also i.i.d. By the weak law of Large numbers

$$-\frac{1}{n}\log p^{(n)}(X_i, ..., X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \to -E\log p(X) = H(X)$$

with convergence in probability. □

Definition 2.31.

$$A_\epsilon^{(n)} = \{(x_1, .., x_n)\} \in X^n \mid 2^{-n(H(X)+\epsilon)} \leq p^{(n)}(x_1, .., x_n) \leq 2^{-n(H(X)-\epsilon)}\}$$

is called the typical set w.r.t ϵ and p.

Consider X_i i.i.d $\sim p(x)$ (p.m.f), Then

$$-\frac{1}{n}\log p^{(n)}(X_1, ..., X_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) = E[-\log p(X_1)] = H(X)$$

(a.e, hence in probability)

$$A_\epsilon^{(n)} = \{(x_1, ..., x_n) \in X^n \mid 2^{-n(H(X)+\epsilon} \leq p^{(n)}(x_1, ..., x_n) \geq 2^{-n(H(X)-\epsilon}\}$$

typical set

Theorem 2.32.

a) If $(x_i, ..., x_n) \in A_\epsilon^{(n)}$ *then*

$$H(X) - \epsilon \leq -\frac{1}{n} \log p^{(n)}(x_1, .., x_n) \leq H(X) + \epsilon$$

b) $P(A_\epsilon^{(n)} > 1 - \epsilon$ *for* n *sufficiently large.*

c) $\mid A_\epsilon^{(n)} \mid \leq 2^{(n)(H(X)+\epsilon)}, (\mid . \mid cordinality)$

d) $\mid A_\epsilon^{(n)} \mid \geq (1-\epsilon) 2^{(n)(H(X)-\epsilon)}$ *for* n *sufficiently large*

Proof. a) obvious

b) obvious

c)

$$1 = \sum_{x \in X^n} p^n(X) \geq \sum_{x \in A_\epsilon^{(n)}} p^{(n)}(X) \geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)}\text{çç} \quad = 2^{-n(H(X)+\epsilon)} \mid A_\epsilon^{(n)} \mid$$

d) For sufficiently large n, $P(A_\epsilon^{(n)}) > 1-\epsilon$, hence $1-\epsilon < P(A_\epsilon^{(n)}) \geq \sum_{x \in A^(n)_\epsilon} 2^{-n(H(X)-\epsilon} = 2^{-n(H(X)-\epsilon} \mid A_\epsilon^{(n)} \mid$

□

For given $\epsilon > 0$ and sufficiently large n. X^n decomposes into a set $T = A_\epsilon^{(n)}$ (typical set) such that

- $P((X_1, ..., X_n) \in T^c) \geq \epsilon$
- For all $x = (x_1, ..., x_n) \in T$:

$$\mid -\frac{1}{n} \log p^{(n)}(x_1, ..., x_n) - H(X) \mid \leq \epsilon$$

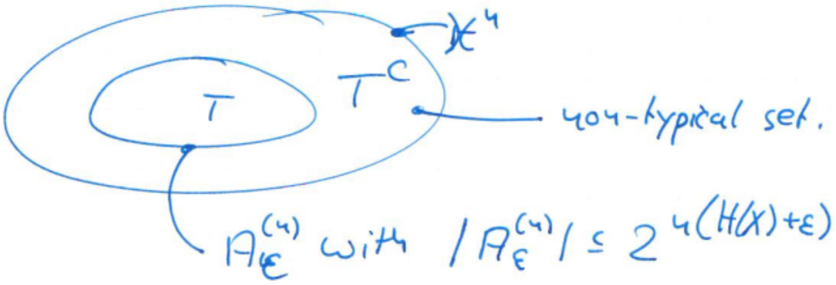
the normalized log-prob of all sequences in T is nearly equal and close to $H(X)$.

Graphically:

The AEP and Data Compression

Let $X_1, .., X_n$ i.i.d with support $\mathcal{X}, X^{(n)} = (X_1, ..., X_n)$. The aim is to find a short description/encoding of all values $x^{(n)} = (x_1, ..., x_n) \in \mathcal{X}^n$. The key idea is index coding, allocate each of the $\mid \mathcal{X}^n \mid$ values an index

- Holds $\mid A_\epsilon^{(n)} \mid \leq 2^{(n)(H(X)+\epsilon)}$ (Th 2.4.4 c). Indexing of all $x^{(n)} \in A_\epsilon^{(n)}$ requires at most $n(H(X)+\epsilon)+1$ (1 bit extra since $n(H(X)+\epsilon)$ may not be an integer)


$\mathcal{X}^n$
T
T^c
non-typical set.
$A_\epsilon^{(n)}$ with $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$

- $\mid \mathcal{X}^n \mid$ requires $n \log \mid \mathcal{X} \mid +1$ bits as indices

Prefix each code word for $x^{(n)} \in A_\epsilon^{(n)}$ by 0 and each code word for $x^{(n)} \notin A_\epsilon^{(n)}$ by 1. Let $l(x^{(n)})$ denotes the length of the code word for $x^{(n)}$. Then

$$\begin{aligned}
E[l(X^{(n)}] &= \sum_{x^{(n)} \in \mathcal{X}^n} p(x^{(n)} l(x(n) \\
&= \sum_{x^{(n)} \in A_\epsilon^{(n)}} p(x^{(n)}) l(x^{(n)} + \sum_{x^{(n)} \notin A_\epsilon^{(n)}} p(x^{(n)}) l(x^{(n)} \\
&\le \sum_{x^{(n)} \in A_\epsilon^{(n)}} p(x^{(n)})(n(H(X)+\epsilon)+2) + \sum_{x^{(n)} \notin A_\epsilon^{(n)}} p(x^{(n)})(n \log \mid X \mid +2) \\
&= P(X^{(n)} \in A_\epsilon^{(n)})(n(H(X)+\epsilon)+2) + P(X^{(n)} \notin A_\epsilon^{(n)})(n \log \mid X \mid +2) \\
&\le n(H(X)+\epsilon) + \epsilon n \log \mid X \mid +2 \\
&\le n(H(X)+\epsilon+\epsilon n \log \mid X \mid + \frac{2}{n} \\
&= n(H(X)+\epsilon'
\end{aligned}$$

for any $\epsilon' > 0$ with n sufficient large it follows:

Theorem 2.33. *$\{X_n\}$ i.i.d For any $\epsilon > 0$ there exists $n \in N$ and a binary code that maps each $X^{(n)}$ one-to-one onto a binary string satisfying*

$$E(\frac{1}{n} l(X^{(n)})) \le H(X) + \epsilon.$$

Hence, for sufficiently large n there exists a code for $X^{(n)}$ such that the expected average codeword length is arbitrary close to $H(X)$.

2.5 Differential Entropy

Theorem 2.34. *By now: Entropy for discrete random variable with finite support. Extension: Discrete random variable but countably many support points , $\mathcal{X} = \{x_1, x_2, ...\}$ distr $\boldsymbol{p} = (p_1, p_2, ...)$*

$$H(X) = -\sum_{i=1}^{\infty} p_i \log p_i$$

Note : The sum may be infinite or may not even exist. Important: Extension of entropy to random variable X with a density f.

Definition 2.35. *Let X be absolute continuous with density $f(x)$, then*

$$h(X) = -\int_{-\infty}^{\infty} f(x)\log f(x)dx$$

is called differential entropy of X.

Remarks:

a) The integral in Def 2.35 may be infinite or may not even exist (Exercises).

b) As a general implicit assumption in defining $h(x)$ we include: " provided the integral exist ".

c) $h(X) = E[-\log f(X)]$.

Example 2.36.

a) $X \sim U(0, a), f(x) = \frac{1}{a}\mathbf{1}(0 < x \leq a)$

$$\begin{aligned} h(X) &= -\int_{0}^{a} \frac{1}{a}\log\frac{1}{a}dx \\ &= \log a, \quad a > 0. \end{aligned}$$

b) $X \sim N(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, X \in \mathbb{R}^n$

$$h(X) = \frac{1}{2}\ln(2\pi e\sigma^2)$$

Definition 2.37. *a) $X = (X_1, .., X_n)$ a random vector with joint density $f(x_1, .., x_n)$. $h(X_1, .., X_n) = -\int .... \int f(x_1, .., x_n)\log f(x_1, ..., x_n)dx_1.....dx_n$ is called joint differential entropy of X.*

b) (X, Y) a random vector with joint density $f(x, y)$ and conditional density.

$$f(x \mid y) = \frac{f(x, y)}{f(y)}, iff(y) > 0,$$

and 0 otherwise. Then,

$$h(X \mid Y) = -\int\int f(x, y)\log f(x \mid y)dxdy$$

is called conditional differential entropy of X given Y.

Definition 2.38. *The mutual information between two random variable X and Y with joint density $f(x, y)$ is defined as*

$$I(X; Y) = h(X) - h(X \mid Y) = h(Y) - h(Y \mid X)$$

Interpretation: Amount of information about X from Y and vice versa.

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

also showing interchangeability of X and Y.

Definition 2.39. *The relative entropy or Kulback-Leibler divergence between two densities f and g is defined as*

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

From (*) it follows that.

$$I(X;Y) = D(f(x,y) \| f(x).f(y)). \qquad (**)$$

Theorem 2.40. $D(f\|g) \geq 0$ *with equality iff* $f = g$ *(almost everywhere)*

Proof. Let $S = \{x \mid f(x) > 0\}$ be the support of f. Then

$$\begin{aligned}
-D(f\|g) &= \int_s f \log \frac{g}{f} \\
&\leq \log \int_s f \frac{g}{f} ((la2.1.6)(f \text{ is concave} : Ef(X) \leq f(EX))) \\
&= \log \int_s g \leq \log 1 = 0
\end{aligned}$$

Equality holds iff $f = g$ a.e. □

Corollary 2.41.

a) *$I(X;Y) \geq 0$ with equality iff X and Y are independent.*

b) *$h(X \mid Y) \leq h(X)$ with equality iff X, Y are independent.*

c) *$-\int f \log f \leq -\int f \log g$.*

Proof. a) follows from $(**)$.

b) $I(X;Y) = h(X) - h(X \mid Y) \geq 0$ by a).

c) By definition of $D(f\|g)$.

□

Theorem 2.42. *(Chain rule for different entropy)*

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

Proof. From the definition it follows that

$$h(X, Y) = h(X) + h(Y \mid X).$$

This implies

$$h(X_1, .., X_i) = h(X_1, .., X_{i-1}) + h(X_i \mid X_1, ..., X_{i-1}).$$

The assertion follows by induction. □

Corollary 2.43.

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

with equality iff $X_1, ..., X_n$ are stoch independent.

Theorem 2.44. *Let $X \in \mathbb{R}^n$ with density $f(x), A \in \mathbb{R}^{n \times n}$ of full rank, $b \in \mathbb{R}^n$. Then*

$$h(AX + b) = h(X) + \log \mid A \mid$$

Proof. If $X \sim f(x)$, then $Y = AX + B \sim \frac{1}{|A|} f(A^{-1}(y - b)), x, y \in \mathbb{R}^n$

$$\begin{aligned} &- \int \frac{1}{\mid A \mid} f(A^{-1}(y - b)) \log(\frac{1}{\mid A \mid} f(A^{-1} 9y - b)) dy \\ &= - \log \frac{1}{\mid A \mid} - \int \frac{1}{\mid A \mid} f(A^{-1} y) \log(f(A^{-1} y) dy \\ &= - \log \frac{1}{\mid A \mid} - \int f(x) \log f(x) dx \\ &= \log \mid A \mid + h(X) \end{aligned}$$

□

Theorem 2.45. *Let $X \in \mathbb{R}^n$ absolute continuous with density $f(x)$ and* $\mathrm{Cov}(X) = C$*, with C positive definite. Then*

$$h(x) \leq \frac{1}{2} ln((2\pi e)^n \mid C \mid),$$

i.e $N_n(\mu, C)$ has largest entropy amongst all random variables with positive definite covariance matrix C.

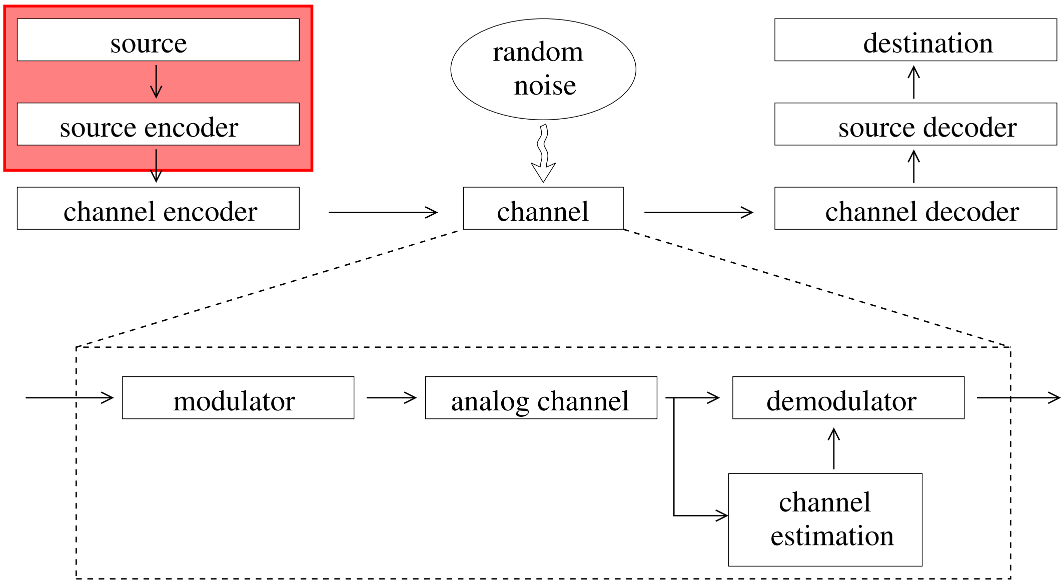
Proof. W.l.o.g assume $EX = 0$ (see thm 2.44).
Let $Q(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |C|^{\frac{1}{2}}} exp\{-\frac{1}{2} X^T C^{-1} x\}$ be the density of $N_n(0, C)$. Let $X \sim f(x), EX =$

$0, \mathrm{Cov}(X) = E(XX^T) = \int XX^T f(x)dx$

$$\begin{aligned}
h(X) &= -\int f(x) f(x) dx \\
&\leq -\int f(x) \ln q(x) dx \quad \text{Cor 2.41} \\
&= -\int f(x) \ln \left[\frac{1}{(2\pi)^{\frac{n}{2}} \mid C \mid^{\frac{1}{2}}} \exp\{-\frac{1}{2} x^T C^{-1} x\} \right] dx \\
&= -\ln \left[\frac{1}{(2\pi)^{\frac{n}{2}} \mid C \mid^{\frac{1}{2}}} \right] + \frac{1}{2} \int x^T C^{-1} x f(x) dx \\
&= \ln (2\pi)^{\frac{n}{2}} \mid C \mid^{\frac{1}{2}} + \frac{1}{2} \int \mathrm{tr}(C^{-1} x x^T) f(x) dx \\
&= \ln (2\pi)^{\frac{n}{2}} \mid C \mid^{\frac{1}{2}} + \frac{1}{2} \mathrm{tr}(C^{-1}) \int x x^T f(x) dx \\
&= \ln (2\pi)^{\frac{n}{2}} \mid C \mid^{\frac{1}{2}} + \frac{n}{2} \\
&= \ln((2\pi e)^{\frac{n}{2}} \mid C \mid^{\frac{1}{2}})
\end{aligned}$$

□

3 Source Coding



Communication Channel from an information theoretic point of view

3.1 Variable Length Encoding

Given some *source alphabet* $\mathcal{X} = \{x_1, \ldots, x_m\}$ and *code alphabet* $\mathcal{Y} = \{y_1, \ldots, y_d\}$. The aim is to find a code word formed over $\mathcal{Y}$ for each character $x_1, \ldots, x_m$. In other words, each character $x_i \in \mathcal{X}$ uniquely mapped onto a "word" over $\mathcal{Y}$.

Definition 3.1. *An injective mapping*

$$g : \mathcal{X} \to \bigcup_{\ell=0}^{\infty} \mathcal{Y}^{\ell} : x_i \mapsto g(x_i) = (w_{i1}, \ldots, w_{in_i})$$

is called encoding. $g(x_i) = (w_{i1}, \ldots, w_{in_i})$ *is called* code word *of character* x_i*,* n_i *is called* length *of code word* i*.*

Example 3.2.

	g_1	g_2	g_3	g_4
a	*1*	*1*	*0*	*0*
b	*0*	*10*	*10*	*01*
c	*1*	*100*	*110*	*10*
d	*00*	*1000*	*111*	*11*
	no encoding	*encoding,* *words are separable*	*encoding,* *shorter,* *words separable*	*encoding,* *even shorter,* *not separable*

Hence, separability of concatenated words over $\mathcal{Y}$ is important.

Definition 3.3. *An encoding g is called* uniquely decodable (u.d.) *or* uniquely decipherable*, if the mapping*

$$G : \bigcup_{\ell=0}^{\infty} \mathcal{X}^{\ell} \to \bigcup_{\ell=0}^{\infty} \mathcal{Y}^{\ell} : \big(a_1, \dots, a_k\big) \mapsto \big(g(a_1), \dots, g(a_k)\big)$$

is injective.

Example 3.4. *Use the previous encoding g_3*

	g_3
a	*0*
b	*10*
c	*110*
d	*111*

1 1 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0
1 1 1|1 0 0 0 1 1 0 1 1 1 0 0 0 1 0
1 1 1|1 0 |0 0 1 1 0 1 1 1 0 0 0 1 0
1 1 1|1 0 |0|0 |1 1 0|1 1 1|0| 0|0|1 0
d b a a c d a a a b

(g_3 is a so called prefix code)

3.2 Prefix Codes

Definition 3.5. *A code is called* prefix code*, if no complete code word is prefix of some other code word, i.e., no code word evolves from continuing some other.*

Formally:

$\boldsymbol{a} \in \mathcal{Y}^k$ *is called prefix of* $\boldsymbol{b} \in \mathcal{Y}^l$, $k \le l$, *if there is some* $c \in \mathcal{Y}^{l-k}$ *such that* $\boldsymbol{b} = (\boldsymbol{a}, \boldsymbol{c})$.

Theorem 3.6. *Prefix codes are uniquely decodable.*

Properties of prefix codes:

- Prefix codes are easy to construct based on the code word lengths.
- Decoding of prefix codes is fast and requires no memory storage.

Next aim: characterize uniquely decodable codes by their code word lengths.

3.3 Kraft-McMillan Theorem

Theorem 3.7. *Kraft-McMillan Theorem*

a) *[McMillan (1959)]: All uniquely decodable codes with code word lengths $n_1, \dots, n_m$ satisfy*

$$\sum_{j=1}^{m} d^{-n_j} \le 1$$

b) *[Kraft (1949)]: Conversely, if $n_1, \dots, n_m \in \mathbb{N}$ are such that $\sum_{j=1}^{m} d^{-n_j} \le 1$, then there exists a u.d. code (even a prefix code) with code word lengths $n_1, \dots, n_m$.*

Proof.

(a) g u.d. code with codeword lengths $n_1, ..., n_m$. Let $r = \max\{n_i\}$ maximum codeword length, $\beta_e = |\{i|n_i = l\}|$ be the number of codewords of length $l \in \mathbb{N}, l \le r$ and it holds, $k \in \mathbb{N}$

$$(\sum_{j=1}^{m} d^{-nj})^k = (\sum_{l=1}^{r} \beta_e d^{-e})^k = \sum_{l=k}^{k.r} \gamma_e{}^{d^{-e}}$$

with

$$\gamma_e = \sum_{\substack{i \le i_1,..,i_k \le r \\ i_1+..+i_k = l}} \beta_{i_1}, ..., \beta_{i_k}, l = k, ..., k.r$$

γ_e is the number of source words of length of length k which have codeword length l and d^e be the number of all codewords of length l, Since g is u.d, each code word has at most one source word. Hence

$$\gamma_e \le d^e$$

$$(\sum_{j=1}^{m} d^{-nj})^k \le \sum_{i=k}^{k.r} d^e d^{-e} = kr - k + 1 \le kr \;\forall k \;\in \mathbb{N}.$$

Further

$$\sum_{j=1}^{m} d^{-nj} \le (kr)^{\frac{1}{k}} \to 1(k \to \infty),$$

so that $\sum_{j=1}^{m} d^{-nj} \le 1$.

Example 3.8.

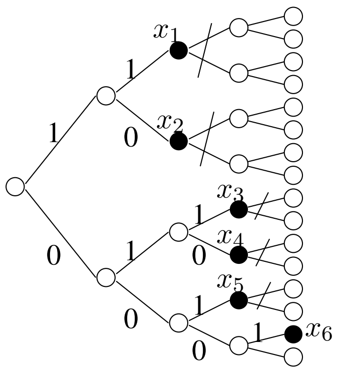
	g_3	g_4
a	*0*	*0*
b	*10*	*01*
c	*110*	*10*
d	*111*	*11*
	u.d.	*not u.d.*

For g_3*:* $2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1.$

For g_4*:* $2^{-1} + 2^{-2} + 2^{-2} + 2^{-2} = 5/4 > 1.$
g_4 *is not u.d., there is no u.d. code with code word lengths* $1, 2, 2, 2.$

(b) Proving with an example. Assume $n_1 = n_2 = 2,\ n_3 = n_4 = n_5 = 3,\ n_6 = 4$. Then $\sum i = 1^6 = 15/16 < 1$.

Construct a prefix code by a binary code tree as follows.



The corresponding code is given as

x_i	x_1	x_2	x_3	x_4	x_5	x_6
$g(x_i)$	11	10	011	010	001	0001

□

3.4 Average Code Word Length

Given a code $g(x_1), \ldots, g(x_m)$ with code word lengths $n_1, \ldots, n_m$.

Question: What is a reasonable measure of the "length of a code"?

Definition 3.9. *The expected code word length is defined as*

$$\bar{n} = \bar{n}(g) = \sum_{j=1}^{m} n_j p_j = \sum_{j=1}^{m} n_j P(X = x_j)$$

Example 3.10.

	p_i	g_2	g_3
a	*1/2*	*1*	*0*
b	*1/4*	*10*	*10*
c	*1/8*	*100*	*110*
d	*1/8*	*1000*	*111*
$\bar{n}(g)$		*15/8*	*14/8*
$H(X)$	*14/8*		

3.5 Noiseless Coding Theorem

Theorem 3.11. *Noiseless Coding Theorem, Shannon (1949) Let random variable X describe a source with distribution $P(X = x_i) = p_i$, $i = 1, \dots, m$. Let the code alphabet $\mathcal{Y} = \{y_1, \dots, y_d\}$ have size d.*

a) Each u.d. code g with code word lengths $n_1, \dots, n_m$ satisfies

$$\bar{n}(g) \geq H(X)/\log d.$$

b) Conversely, there is a prefix code, hence a u.d. code g with

$$\bar{n}(g) \leq H(X)/\log d + 1.$$

Proof. a) For any u.d. code it holds by McMillan's Theorem that

$$\begin{aligned}
\frac{H(X)}{\log d} - \bar{n}(g) &= \frac{1}{\log d}\sum_{j=1}^{m} p_j \log\frac{1}{p_j} - \sum_{j=1}^{m} p_j n_j \\
&= \frac{1}{\log d}\sum_{j=1}^{m} p_j \log\frac{1}{p_j} + \sum_{j=1}^{m} p_j \frac{\log d^{-n_j}}{\log d} \\
&= \frac{1}{\log d}\sum_{j=1}^{m} p_j \log\frac{d^{-n_j}}{p_j} \\
&= \frac{\log e}{\log d}\sum_{j=1}^{m} p_j \ln\frac{d^{-n_j}}{p_j} \\
&\leq \frac{\log e}{\log d}\sum_{j=1}^{m} p_j \Big(\frac{d^{-n_j}}{p_j} - 1\Big) \quad \text{(since } \ln x \leq x-1,\ x \geq 0\text{)} \\
&\leq \frac{\log e}{\log d}\sum_{j=1}^{m} \Big(d^{-n_j} - p_j\Big) \leq 0.
\end{aligned}$$

b) Shannon-Fano Coding

W.l.o.g. assume that $p_j > 0$ for all j.

Choose integers n_j such that $d^{-n_j} \le p_j < d^{-n_j+1}$ for all j. Then

$$\sum_{j=1}^{m} d^{-n_j} \le \sum_{j=1}^{m} p_j \le 1$$

such that by Kraft's Theorem a u.d. code g exists. Furthermore,

$$\log p_j < (-n_j + 1) \log d$$

holds by construction. Hence

$$\sum_{j=1}^{m} p_j \log p_j < (\log d) \sum_{j=1}^{m} p_j(-n_j + 1),$$

equivalently,

$$H(X) > (\log d) \big(\bar{n}(g) - 1\big).$$

□

3.6 Compact Codes

Is there always a u.d. code g with

$$\bar{n}(g) = H(X)/\log d?$$

No! Check the previous proof. Equality holds if and only if $p_j = 2^{-n_j}$ for all $j = 1, \dots, m$.

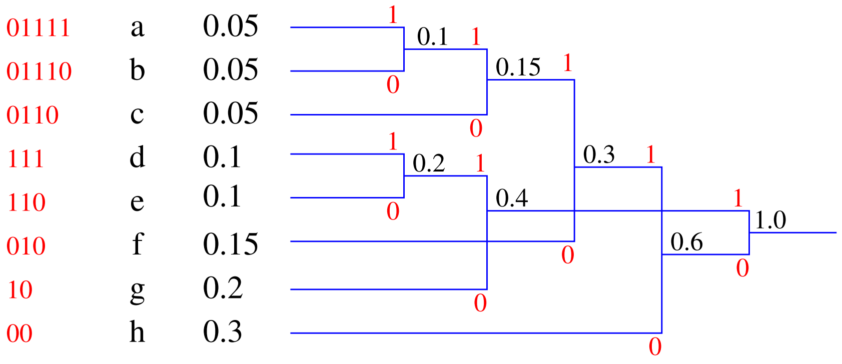
Example 3.12. *Consider binary codes, i.e., $d = 2$. $\mathcal{X} = \{a, b\}$, $p_1 = 0.6,\ p_2 = 0.4$. The shortest possible code is $g(a) = (0),\ g(b) = (1)$.*

$$\begin{aligned} H(X) &= -0.6 \log_2 0.6 - 0.4 \log_2 0.4 = 0.97095 \\ \bar{n}(g) &= 1. \end{aligned}$$

Definition 3.13. *Any code of shortest possible average code word length is called* compact.

How to construct compact codes?

3.7 Huffman Coding



A compact code g^* is given by:

Character:	a	b	c	d	e	f	g	h
Code word:	01111	01110	0110	111	110	010	10	00

It holds (log to the base 2):

$$\bar{n}(g^*) = 5 \cdot 0.05 + \cdots + 2 \cdot 0.3 = \mathbf{2.75}$$
$$H(X) = -0.05 \cdot \log_2 0.05 - \cdots - 0.3 \cdot \log_2 0.3 = \mathbf{2.7087}$$

Huffman are optimal i.e, have shortest average codeword length. We consider the case $d = 2$.

Lemma 3.14. *Let $\mathcal{X} = \{x_1, ..., x_m\}$ with probabilities $p_1 \geq, ..., \geq p_m > 0$. There exists an optimal binary prefix code g with codeword lengths $n_1 \leq, ..., n_m$ such that*

(i) $n_1 \leq ..... \leq n_m$*,*

(ii) $n_{m-1} = n_m$*,*

(iii) $g(X_{m-1})$ *and* $g(X_m)$ *differ only in the last position.*

Proof. Let g be an optimum prefix code with $n_1, , ..., n_m$.

(i) If $p_i > p_j$ then necessarily $n_i \leq n_j, 1 \leq i < j < m$. Otherwise exchange $g(x_i)$ and $g(x_j)$ to obtain code g' with

$$\begin{aligned} \bar{n}(g') - \bar{n}(g) &= p_i n_j + p_j n_i - p_i n_i - p_j n_j \\ &= (p_i - p_j)(n_j - n_i) < 0 \end{aligned}$$

contradictory optimality of g.

(ii) There is an optimal prefix code g with $n_i \leq, .., \leq m$.if $n_{m-1} < n_m$ delete $n_m - n_{m-1}$ positions of $g(x_m)$ to obtain a better code.

(iii) If $l_1 \leq ..... \leq l_{m-1} = l_m$ for an optimal prefix code g and $g(x_{m-1})$ and $g(x_m)$ differ by more than the last position, delete the last position in both to obtain a better code.

□

Lemma 3.15. *Let $\mathcal{X} = \{x_i, ..., x_m\}$ with prob $p_1 \geq .... \geq p_m > 0$. $\mathcal{X}_1 = \{x'_1, .., X'_{m-1}\}$ with prob $p'_i = p_i, i = 1..., m-2$, and $p'_{m-1} = p_{m-1} + p_m$. Let g' be an optimal prefix code for $\mathcal{X}'$ with codewords $g'(x'_i), i = 1, ..., m-1$. Then*

$$g(x_1) = \begin{cases} g'(x'_i), & i = 1, ..., m-2. \\ (g'(x'_{m-1}, 0), & i = m-1 \\ (g'(x'_{m-1}, 1), & i = m \end{cases}$$

is an optimal prefix code for X

Proof. Denote codeword lengths n_i, n'_i for g, g'respectively.

$$\begin{aligned} \bar{n}(g) &= \sum_{j=1}^{m-2} p_j n'_j + (p_m + p_{m-1})(n'_{m-1} + 1) \\ &= \sum_{j=1}^{m-2} p'_j n'_j + p'_{m-1}(n'_{m-1} + 1) \\ &= \sum_{j=1}^{m-1} p'_j n'_j + p_{m-1} + p_m = \bar{n}(g') + p_{m-1} + p_m \end{aligned}$$

Assume g is not optimal for $\mathcal{X}$ .There exists an opt prefix code h with properties (i)-(iii) of 3.14 and $\bar{n}(h) < \bar{n}(g)$.

Set

$$h'(x_j)' = \begin{cases} h(x_j), & j = 1, ..., m-2. \\ \lfloor h(x_{m-1} \rfloor, & \text{deleting the last position of } h(x_{m-1}), j = m \end{cases}$$

Then $\bar{n}(h') + p_{m-1} + p_m = \bar{n}(h) < \bar{n}(g) = \bar{n}(g') + p_{m-1} + p_m$. Hence $\bar{n}(h') < \bar{n}(g')$ contradicting optimality of g'. □

3.8 Block Codes for Stationary Sources

Encode blocks/words of length N by words over the code alphabet $\mathcal{Y}$. Assume that blocks are generated by a stationary source, a stationary sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$.

Notation for a block code:

$$g^{(N)} : \mathcal{X}^N \to \bigcup_{\ell=0}^{\infty} \mathcal{Y}^\ell$$

Block codes are “normal” variable length codes over the extended alphabet $\mathcal{X}^N$. A fair measure of the “length” of a block code is the average code word length per character

$$\bar{n}\big(g^{(N)}\big)/N.$$

Theorem 3.16. *Noiseless Coding Theorem for Block Codes*
Let $\boldsymbol{X} = \{X_n\}_{n\in}$ be a stationary source. Let the code alphabet $\mathcal{Y} = \{y_1, \dots, y_d\}$ have size d.

a) Each u.d. block code $g^{(N)}$ satisfies

$$\frac{\bar{n}(g^{(N)})}{N} \geq \frac{H(X_1, \dots, X_N)}{N \ \log d}.$$

b) Conversely, there is a prefix block code, hence a u.d. block code $g^{(N)}$ with

$$\frac{\bar{n}(g^{(N)})}{N} \leq \frac{H(X_1, \dots, X_N)}{N \ \log d} + \frac{1}{N}.$$

Hence, in the limit as $N \to \infty$:

There is a sequence of u.d. block codes $g^{(N)}$ such that

$$\lim_{N\to\infty} \frac{\bar{n}(g^{(N)})}{N} = \frac{H_\infty(\boldsymbol{X})}{\log d}.$$

3.8.1 Huffman Block Coding

In principle, Huffman encoding can be applied to block codes. However, problems include

- The size of the Huffman table is m^N, thus growing exponentially with the block length.
- The code table needs to be transmitted to the receiver.
- The source statistics are assumed to be stationary. No adaptivity to changing probabilities.
- Encoding and decoding only per block. Delays occur at the beginning and end. Padding may be necessary.

3.9 Arithmetic Coding

Assume that:

- Message $(x_{i_1}, \dots, x_{i_N}), x_{i_j} \in \mathcal{X}, j = 1, \dots, N$ is generated by some source $\{X_n\}_{n\in\mathbb{N}}$.
- All (conditional) probabilities

$$P(X_n = x_{i_n} \mid X_1 = x_{i_1}, \dots, X_{n-1} = x_{i_{n-1}}) = p(i_n \mid i_1, \dots, i_{n-1}),$$

$x_{i_1}, \ldots, x_{i_n} \in \mathcal{X}$, $n = 1, \ldots, N$, are known to the encoder and decoder, or can be estimated.

Then,

$$P(X_1 = x_{i_1}, \ldots, X_n = x_{i_n}) = p(i_1, \ldots, i_n)$$

can be easily computed as

$$p(i_1, \ldots, i_n) = p(i_n \mid i_1, \ldots, i_{n-1}) \cdot p(i_1, \ldots, i_{n-1}).$$

Iteratively construct intervals

Initialization, $n = 1$*:* $\big(c(1) = 0,\ c(m+1) = 1\big)$

$$I(j) = \big[c(j), c(j+1)\big), \quad c(j) = \sum_{i=1}^{j-1} p(i),\ j = 1, \ldots, m$$

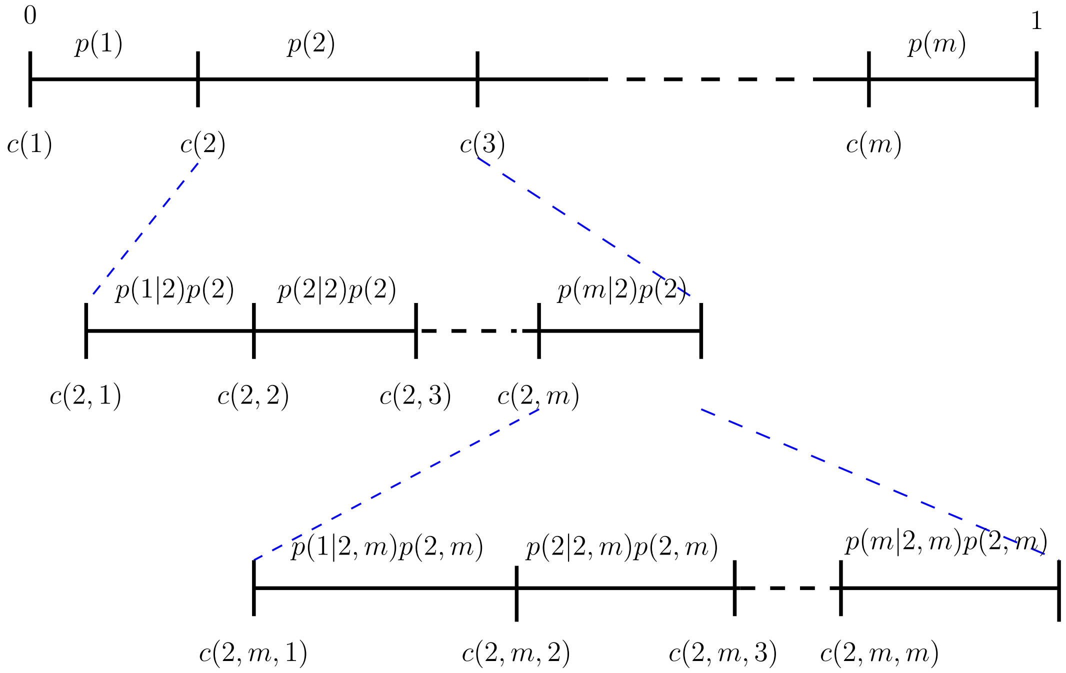
(cumulative probabilities)

Recursion over $n = 2, \ldots, N$*:*

$$\begin{aligned} &I(i_1, \ldots, i_n) \\ &= \Big[c(i_1, \ldots, i_{n-1}) + \sum_{i=1}^{i_n - 1} p(i_n \mid i_1, \ldots, i_{n-1}) \cdot p(i_1, \ldots, i_{n-1})\big) \\ &\qquad c(i_1, \ldots, i_{n-1}) + \sum_{i=1}^{i_n} p(i_n \mid i_1, \ldots, i_{n-1}) \cdot p(i_1, \ldots, i_{n-1})\Big) \end{aligned}$$

Program code available from Togneri, deSilva, p. 151, 152.

Example 3.17.



Encode message $(x_{i_1},\dots,x_{i_N})$ by the binary representation of some binary number in the interval $I(i_1,\dots,i_n)$.

A scheme which usually works quite well is as follows.

Let $l = l(i_1,\dots,i_n)$ and $r = r(i_1,\dots,i_n)$ denote the left and right bound of the corresponding interval. Carry out the binary expansion of l and r until until they differ. Since $l < r$, at the first place they differ there will be a 0 in the expansion of l and a 1 in the expansion of r. The number $0.\ a_1a_2\dots a_{t-1}1$ falls within the interval and requires the least number of bits.

$$(a_1a_2\dots a_{t-1}1) \text{ is the encoding of } (x_{i_1},\dots,x_{i_N}).$$

The probability of occurrence of message $(x_{i_1},\dots,x_{i_N})$ is equal to the length of the representing interval. Approximately

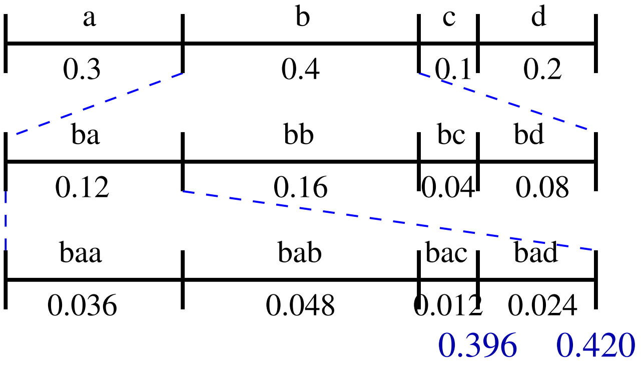
$$-\log_2 p(i_1,\dots,i_n)$$

bits are needed to represent the interval, which is close to optimal.

Example 3.18. *Assume a memoryless source with 4 characters and probabilities*

x_i	*a*	*b*	*c*	*d*
$P(X_n = x_i)$	*0.3*	*0.4*	*0.1*	*0.2*

Encode the word (bad)*:*

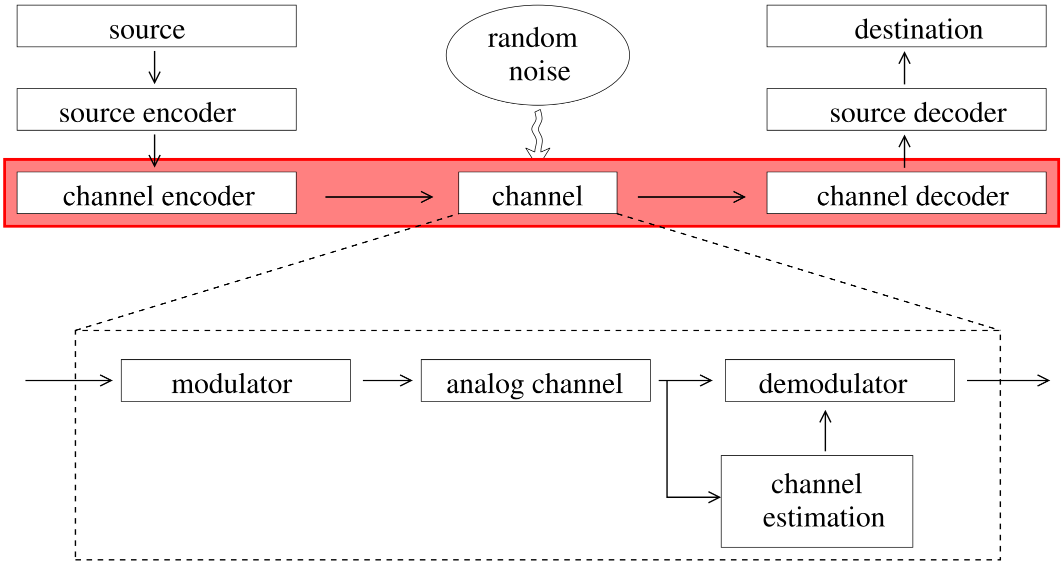


$(bad) = [0.396, 0.42)$

$0.396 = 0.01100\ldots \quad 0.420 = 0.01101\ldots$

$(bad) = (01101)$

4 Information Channels



Communication Channel from an information theoretic point of view

4.1 Discrete Channel Model

Discrete information channels are described by

- A pair of random variables (X, Y) with support $\mathcal{X} \times \mathcal{Y}$, where X is the input r.v., $\mathcal{X} = \{x_1, \ldots, x_m\}$ the input alphabet and Y is the output r.v., $\mathcal{Y} = \{y_1, \ldots, y_d\}$ the output alphabet.
- The channel matrix

  $$\boldsymbol{W} = \left(w_{ij}\right)_{i=1,\ldots,m,\ j=1,\ldots,d}$$

  with

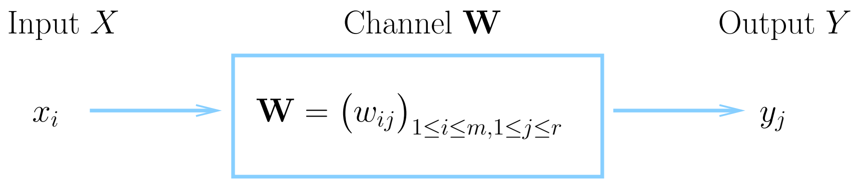
  $$w_{ij} = P(Y = y_j \mid X = x_i),\ i = 1, \ldots, m,\ j = 1, \ldots, d$$

- Input distribution

  $$P(X = x_i) = p_i,\ i = 1, \ldots, m,$$

  $\boldsymbol{p} = (p_1, \ldots, p_m)$.

Discrete Channel Model :


Input X
Channel $\mathbf{W}$
Output Y
x_i
$\mathbf{W} = \left(w_{ij}\right)_{1\le i\le m, 1\le j\le r}$
y_j

where $\boldsymbol{W}$ composed of rows $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ as $\boldsymbol{W} = \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_m \end{pmatrix}$.

Lemma 4.1. *Let X and Y be the input r.v. and the output r.v. of a discrete channel with channel matrix $\boldsymbol{W}$, respectively. Lets denote the input distribution as $P(X = x_i) = p_i$, $i = 1, \ldots, m$, with $\boldsymbol{p} = (p_1, \ldots, p_m)$. Then*

(a) $H(Y) = H(\boldsymbol{pW})$.

(b) $H(Y \mid X = x_i) = H(\boldsymbol{w}_i)$.

(c) $H(Y \mid X) = \sum_{i=1}^{m} p_i H(\boldsymbol{w}_i)$.

Proof. (a) $H(Y)$, determine the distribution of Y:

$$\begin{aligned} P(Y = y_j) &= \sum_{i=1}^{m} P(Y = y_j \mid X = x_i) P(X = x_i) \\ &= \sum_{i=1}^{m} p_i w_{ij} = (pW)_j, j = 1, .., d. \end{aligned}$$

(b) $H(Y \mid X = x_i) = H(w_i)$ by definition.

(c) $H(Y \mid X) = \sum_{i=1}^{m} p_i H(Y \mid X = x_i) = \sum_{i=1}^{m} p_i H(w_i)$.

□

4.2 Channel Capacity

The mutual information between X and Y is

$$\begin{aligned}
I(X;Y) &= H(Y) \quad - H(Y \mid X) \\
&= H(\boldsymbol{p}\boldsymbol{W}) - \sum_{i=1}^{m} p_i H(\boldsymbol{w}_i) \\
&= H\left[\sum_{i=1}^{m} p_i w_{ij}\right]_{j=1,\dots,d} + \sum_{i=1}^{m} p_i \sum_{j=1}^{d} w_{ij} \log w_{ij} \\
&= -\sum_{j=1}^{d} \left(\sum_{i=1}^{m} p_i w_{ij}\right) \log \left(\sum_{i=1}^{m} p_i w_{ij}\right) + \sum_{i,j} p_i w_{ij} \log w_{ij} \\
&= -\sum_{i,j} p_i w_{ij} \log \left(\sum_{l=1}^{m} p_l w_{lj}\right) + \sum_{i,j} p_i w_{ij} \log w_{ij} \\
&= \sum_{i} p_i \left[\sum_{j} w_{ij} \log \frac{w_{ij}}{\sum_i p_i w_{ij}}\right] \\
&= \sum_{i=1}^{m} p_i D\big(\boldsymbol{w}_i \parallel \boldsymbol{p}\boldsymbol{W}\big) = I(\boldsymbol{p}; \boldsymbol{W}),
\end{aligned}$$

where D denoting the Kulback-Leibler divergence.

The aim is to use the input distribution that maximizes mutual information $I(X;Y)$ for a given channel $\boldsymbol{W}$.

Definition 4.2.

$$C = \max_{(p_1,\dots,p_m)} I(X;Y) = \max_{\boldsymbol{p}} I(\boldsymbol{p}, \boldsymbol{W})$$

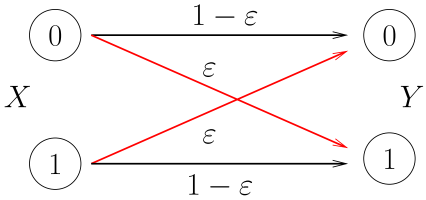
is called channel capacity.

Determining capacity is in general a complicated optimization problem.

4.3 Binary Channels

4.3.1 Binary Symmetric Channel (BSC)

Example 4.3. *BSC*



Input distribution $\boldsymbol{p} = (p_0, p_1)$

Channel matrix

$$\boldsymbol{W} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

In this case $I(X,Y)$ is

$$\begin{aligned} I(X;Y) = I(\boldsymbol{p};\boldsymbol{W}) = \quad & H(\boldsymbol{pW}) - \sum_{i=1}^{m} p_i H(\boldsymbol{w}_i) \\ &= H(p_0(1-\epsilon)+p_1\epsilon, \epsilon p_0 + (1-\epsilon)p_1) - p_0.H(1-\epsilon,\epsilon) - p_1 H(\epsilon, 1-\epsilon) \\ &= H_2(p_0(1-\epsilon)+p_1\epsilon) - H_2(\epsilon) \end{aligned}$$

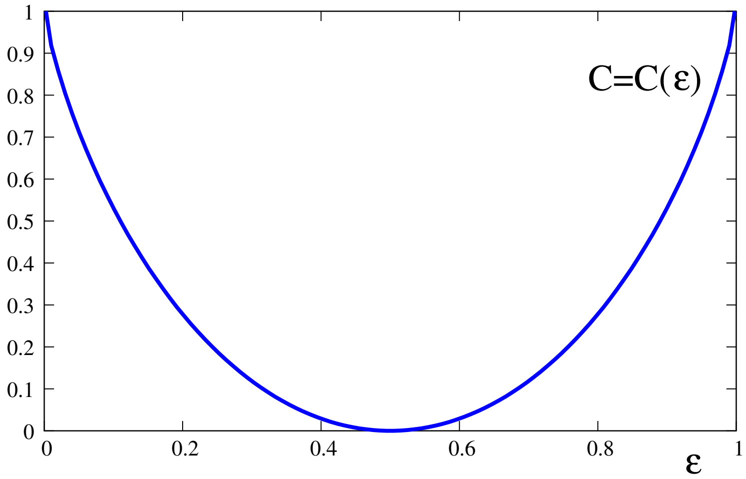
and is to be maximised over $(p_0,p_1), p_0, p_1 \geq 0, p_0+p_1 = 1$, that is

$$\begin{aligned} H_2(q) &= -q\log q - (1-q)\log(1-q), 0 \leq q \leq 1 \\ &\leq \log 2. \end{aligned}$$

This is a achieved if $p_0 = p_1 = \frac{1}{2}$. Capacity achieving distribution is $p^* = (\frac{1}{2},\frac{1}{2})$ with capacity

$$\begin{aligned} C = \max I(X;Y) &= \log 2 + (1-\epsilon)\log(1-\epsilon) + \epsilon\log(\epsilon) \\ &= 1 + (1-\epsilon)\log_2(1=\epsilon) + \epsilon\log_2\epsilon. \end{aligned}$$

Capacity of the BSC as a function of ϵ:



Remark 4.4. *To compute channel capacity for a given channel with channel matrix $\boldsymbol{W}$, we need to solve*

$$C = \max_{\boldsymbol{p}} I(\boldsymbol{p};\boldsymbol{W}) = \max_{\boldsymbol{p}} \sum_{i=1}^{m} p_i D\big(\boldsymbol{w}_i \parallel \boldsymbol{pW}\big).$$

Theorem 4.5. *The capacity of the channel $\boldsymbol{W}$ is attained at $\boldsymbol{p}^* = (p_1^*,\ldots,p_m^*)$ if and only if*

$$D(\boldsymbol{w}_i \parallel \boldsymbol{p}^*\boldsymbol{W}) = \zeta \textit{ for all } i = 1,\ldots,m.$$

for all $i = 1,\ldots,m$ with $p_i > 0$.

Moreover,

$$C = I(\boldsymbol{p}^*;\boldsymbol{W}) = \zeta.$$

Proof. Mutual information $I(\boldsymbol{p};\boldsymbol{W})$ is a concave function of $\boldsymbol{p}$. Hence the KKT conditions (cf., e.g., Boyd and Vandenberge 2004) are necessary and sufficient for optimality of some input distribution $\boldsymbol{p}$. Using the above representation some elementary algebra shows that

$$\frac{\partial}{\partial p_i} I(\boldsymbol{p};\boldsymbol{W}) = D(\boldsymbol{w}_i \| \boldsymbol{p}\boldsymbol{W}) - 1.$$

Then,

$$\begin{aligned}\frac{\partial}{\partial_{p_k}} H(\boldsymbol{p}\boldsymbol{W}) &= \frac{\partial}{\partial_{p_k}}[-\sum_j (\sum_i p_i w_{ij}) \log(\sum_i p_i w_{ij})] \\ &= -\sum_j [w_{kj} \log(\sum_i p_i w_{ij}) + \sum_i p_i w_{ij} \frac{w_{kj}}{\sum_i p_i w_{ij}}] \\ &= -\sum_j [w_{kj} \log(\sum_i p_i w_{ij}) + w_{kj}],\end{aligned}$$

thus

$$\begin{aligned}\frac{\partial}{\partial_{p_k}} I(\boldsymbol{p},\boldsymbol{W}) &= \frac{\partial}{\partial_{p_k}} H(\boldsymbol{p}\boldsymbol{W}) - \frac{\partial}{\partial_{p_k}}(\sum_i p_i H(\boldsymbol{w}_i)) \\ &= -\sum_j w_{kj} \log(\sum_i p_i w_{ij}) + \sum_j w_{kj} \log w_{kj} - 1 \\ &= \sum_j w_{kj} \log \frac{w_{kj}}{\sum_i p_i w_{ij}} - 1 \\ &= D(\boldsymbol{w}_k \| \boldsymbol{p}\boldsymbol{W}) - 1.\end{aligned}$$

The full set of KKT conditions now reads as

$$\begin{aligned}\textstyle\sum_{j=1}^m p_j &= 1 \\ p_i &\geq 0,\ i = 1,\dots,m \\ \lambda_i &\geq 0,\ i = 1,\dots,m \\ \lambda_i p_i &= 0,\ i = 1,\dots,m \\ D(\boldsymbol{w}_i \| \boldsymbol{p}\boldsymbol{W}) + \lambda_i + \nu &= 0,\ i = 1,\dots,m\end{aligned}$$

which shows the assertion. □

Theorem 4.6. *(G. Alirezaei, 2018)*

Given a channel with square channel matrix $\boldsymbol{W} = \big(w_{ij}\big)_{i,j=1,\dots,m}$. *Denote self information by* $\rho(q) = -q \log q,\ q \geq 0$. *Assume that* $\boldsymbol{W}$ *is invertible with inverse*

$$\boldsymbol{T} = \big(t_{ij}\big)_{i,j=1,\dots,m}.$$

Then, measured in nats, the capacity is

$$C = \ln\Big(\sum_k exp\{-\sum_{i,j} t_{ki}\, \rho(w_{ij})\}\Big)$$

and the capacity achieving distribution is given by

$$p_\ell^* = e^{-C}\sum_k t_{ks}\,exp\{-\sum_{i,j} t_{ki}\,\rho(w_{ij})\} = \frac{\sum_k t_{ks}\,exp\{-\sum_{i,j} t_{ki}\,\rho(w_{ij})\}}{\sum_k exp\{-\sum_{i,j} t_{ki}\,\rho(w_{ij})\}}.$$

Proof. $\boldsymbol{p}$ is capacity achieving iff $D(\boldsymbol{w}_i \| \boldsymbol{pW}) = \zeta\ \forall_i : p_i > 0$. Let $p(q) = -q\log q, q \geq 0$ and $\boldsymbol{T}$ inverse of $\boldsymbol{W}, \boldsymbol{T} = \boldsymbol{W}^{-1}$ so that $\boldsymbol{T}\boldsymbol{1}_m = \boldsymbol{TW}\boldsymbol{1}_m = \boldsymbol{I}\boldsymbol{1}_m = \boldsymbol{1}_m$.

Then H holds:

$$\begin{aligned}\zeta &= D(\boldsymbol{w}_i \| \boldsymbol{pW})\\ &= \sum_j w_{ij}\ln\frac{w_{ij}}{\sum_{l=1}^m p_l w_{lj}}\\ &= -\sum_j [w_{ij}\ln(\sum_l p_l w_{lj}) + p(w_{ij})], \quad i = 1,..,m.\end{aligned}$$

Hence $\forall k = 1, ..., m$ by summation over i

$$\begin{aligned}\underbrace{\zeta(\sum_i t_{ki})}_{1} &= -\sum_i t_{ki}\sum_j [w_{ij}\ln(\sum_l p_l w_{lj}) + p(w_{ij})]\\ &= -\sum_j \underbrace{\sum_i t_{ki}w_{ij}}_{\delta_{kj}}\ln(\sum_l p_l w_{lj}) - \sum_{i,j} t_{ki}p(w_{ij})\\ &= -\ln(\sum_l p_l w_{lk}) - \sum_{i,j} t_{ki}p(w_{ij}).\end{aligned}$$

Resolve for $\boldsymbol{p} = (p_1, ..., p_m)$: $\sum_l p_l w_{lk} = exp(-\zeta - \sum_{i,j} t_{ki}p(w_{ij})\forall k = 1,..,m.(**)$
Summation over k:

$$\begin{aligned}1 &= \sum_k exp(-\zeta - \sum_{i,j} t_{ki}p(w_{ij}))\\ &= \sum_k e^{-\zeta}e^{-\sum_{i,j} t_{ki}p(w_{ij})}\end{aligned}$$

It follows

$$\zeta = \ln(\sum_k e^{-\zeta}e^{-\sum_{i,j} t_{ki}p(w_{ij})}) = C \ \ \text{(capacity)}$$

To determine p_l multiply (**) by t_{ks} and sum over k.

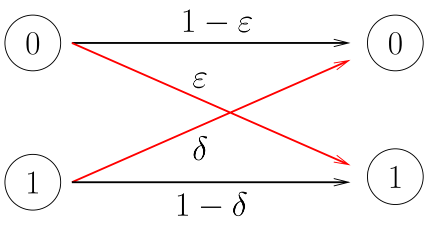
$$\begin{aligned}\sum_k t_{ks}\sum_l p_l w_{lk} &= \sum_k t_{ks}e^{-C}e^{-\sum_{i,j} t_{ki}p(w_{ij})}\\ \sum_l p_l \underbrace{\sum_k w_{lk}t_{ks}}_{\delta_{ks}} &= \sum_k t_{ks}e^{-C}e^{-\sum_{i,j} t_{ki}p(w_{ij})}\end{aligned}$$

$$P_s = e^{-C}\sum_k t_{ks}e^{-\sum_{i,j} t_{ki}p(w_{ij})}, s = 1,..,m \ \ \text{(Capacity-achieving distribution)}$$

□

4.3.2 Binary Asymmetric Channel (BAC)

Example 4.7.



$$\boldsymbol{W} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \delta & 1-\delta \end{pmatrix}$$

The capacity-achieving distribution is

$$p_0^* = \frac{1}{1+b}, \quad p_1^* = \frac{b}{1+b},$$

with

$$b = \frac{a\epsilon - (1-\epsilon)}{\delta - a(1-\delta)} \ \text{and} \ a = \exp\left(\frac{h(\delta)-h(\epsilon)}{1-\epsilon-\delta}\right),$$

and $h(\epsilon) = H(\epsilon, 1-\epsilon)$*, the entropy of* $(\epsilon, 1-\epsilon)$*.*

Note that $\epsilon = \delta$ *yields the previous result for the BSC.*

Derivation of capacity for the BAC:

By 4.5 the capacity-achieving input distribution $\boldsymbol{p} = (p_0, p_1)$ *satisfies*

$$D(\boldsymbol{w}_1 \| \boldsymbol{pW}) = D(\boldsymbol{w}_2 \| \boldsymbol{pW}).$$

This is an equation in the variables p_0, p_1 *which jointly with the condition* $p_0 + p_1 = 1$ *has the solution*

$$p_0^* = \frac{1}{1+b}, \quad p_1^* = \frac{b}{1+b}, \tag{4.1}$$

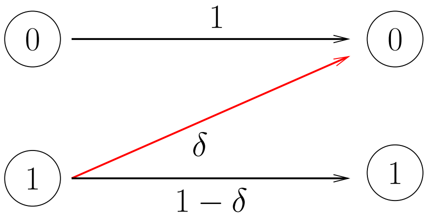
with

$$b = \frac{a\epsilon - (1-\epsilon)}{\delta - a(1-\delta)} \ \text{and} \ a = \exp\left(\frac{h(\delta)-h(\epsilon)}{1-\epsilon-\delta}\right),$$

and $h(\epsilon) = H(\epsilon, 1-\epsilon)$*, the entropy of* $(\epsilon, 1-\epsilon)$*.*

4.3.3 Binary Z-Channel (BZC)

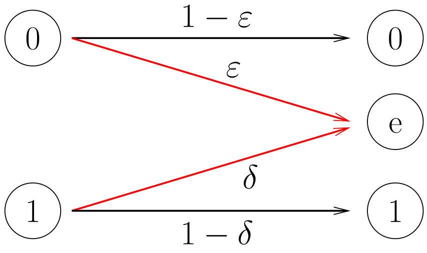
Example 4.8. *The so called Z-channel is a special case of the BAC with* $\epsilon = 0$*.*



The capacity-achieving distribution is obtained from the BAC by setting $\epsilon = 0$.

4.3.4 Binary Asymmetric Erasure Channel (BAEC)

Example 4.9.



$$\boldsymbol{W} = \begin{pmatrix} 1-\varepsilon & \varepsilon & 0 \\ 0 & \delta & 1-\delta \end{pmatrix}$$

The capacity-achieving distribution is determined by finding the solution x^* of

$$\epsilon \log \epsilon - \delta \log \delta = (1-\delta)\log(\delta + \epsilon x) - (1-\epsilon)\log(\epsilon + \delta/x)$$

and setting

$$\frac{p_0^*}{p_1^*} = x^*, \quad p_0^* + p_1^* = 1.$$

Derivation of capacity for the BAEC:

By 4.5 the capacity-achieving distribution $\mathbf{p}^* = (p_0^*, p_1^*)$, $p_0^* + p_1^* = 1$ is given by the solution of

$$\begin{aligned} &(1-\epsilon)\log\frac{1-\epsilon}{p_0(1-\epsilon)} + \epsilon \log \frac{\epsilon}{p_0\epsilon + p_1\delta} \\ &= \delta \log \frac{\delta}{p_0\epsilon + p_1\delta} + (1-\delta)\log\frac{1-\delta}{p_0(1-\delta)}, \end{aligned} \tag{4.2}$$

Substituting $x = \frac{p_0}{p_1}$, equation (4.2) reads equivalently as

$$\epsilon \log \epsilon - \delta \log \delta = (1-\delta)\log(\delta + \epsilon x) - (1-\epsilon)\log(\epsilon + \delta/x)$$

By differentiating w.r.t. x it is easy to see that the right hand side is monotonically increasing such that exactly one solution $\boldsymbol{p}^* = (p_1^*, p_2^*)$ exists, which can be numerically computed.

4.4 Channel Coding

Consider transmission of blocks of length N.

Denote:

$$\boldsymbol{X}_N = (X_1, \dots, X_N) \text{ input random vector of length } N$$
$$\mathbf{Y}_N = (Y_1, \dots, Y_N) \text{ output random vector of length } N$$

where $X_1, \dots, X_N \in \mathcal{X}, Y_1, \dots, Y_N \in \mathcal{Y}$.

Only a subset of all possible blocks of length N is used as input, the channel code.

Definition 4.10. *A set of M codewords of length N, denoted by*

$$\mathcal{C}_N = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_M\} \subseteq \mathcal{X}^N$$

is called (N, M)-code.

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

is called the code rate. *It represents the average number of bits per code word.*

Transmission is characterized by

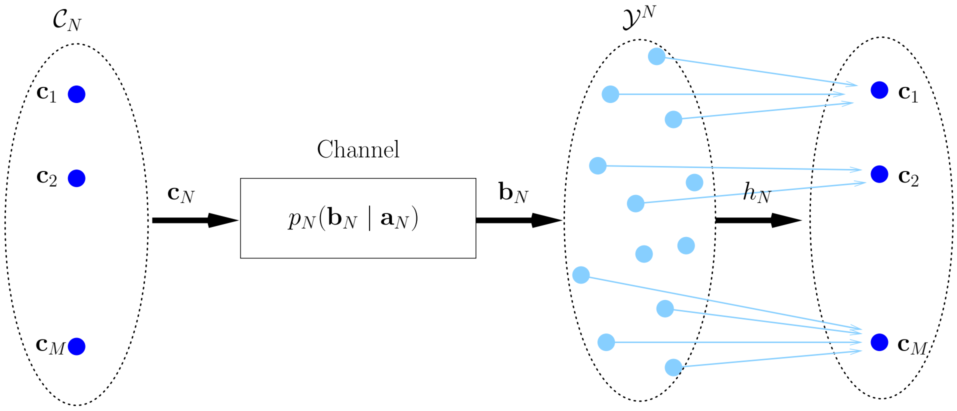
- the channel code $\mathcal{C}_N = \{\boldsymbol{c}_1, \dots, \boldsymbol{c}_M\}$
- transmission probabilities

$$p_N(\boldsymbol{b}_N \mid \boldsymbol{a}_N) = P\big(\mathbf{Y}_N = \boldsymbol{b}_N \mid \boldsymbol{X}_n = \boldsymbol{a}_N\big)$$

- the decoding rule

$$h_N : \mathcal{Y}^N \to \mathcal{C}_N : \boldsymbol{b}_N \mapsto h_N(\boldsymbol{b}_N)$$

It can be represented graphically as:



4.5 Decoding Rules

Definition 4.11. *A decoding rule* $h_N : \mathcal{Y}^N \to \mathcal{C}_n$ *is called* minumum error rule (ME) *or* ideal observer *if*

$$\boldsymbol{c}_j = h_N(\boldsymbol{b}) \Rightarrow P(\boldsymbol{X}_N = \boldsymbol{c}_j \mid \boldsymbol{Y}_N = \boldsymbol{b}) \geq P(\boldsymbol{X}_N = \boldsymbol{c}_i \mid \boldsymbol{Y}_N = \boldsymbol{b})$$

for all $i = 1, \ldots, M$. *Equivalently,*

$$\begin{aligned} \boldsymbol{c}_j = h_N(\boldsymbol{b}) \Rightarrow P(\boldsymbol{Y}_N = \boldsymbol{b} \mid \boldsymbol{X}_N = \boldsymbol{c}_j) P(\boldsymbol{X}_N = c_j) \\ \geq P(\boldsymbol{Y}_N = \boldsymbol{b} \mid \boldsymbol{X}_N = \boldsymbol{c}_i) P(\boldsymbol{X}_N = c_i) \end{aligned}$$

for all $i = 1, \ldots, M$.

With ME-decoding, $\boldsymbol{b}$ *is decoded as the codeword* $\mathbf{c}_j$ *which has greatest conditional probability of having been sent given* $\boldsymbol{b}$ *is received. Hence,*

$$h_N(\boldsymbol{b}) \in \arg \max_{i=1,\ldots,M} P\big(\boldsymbol{X}_N = \boldsymbol{c}_i \mid \boldsymbol{Y}_N = \boldsymbol{b}\big).$$

. The ME decoding rules depend on the input distribution.

Definition 4.12. *A decoding rule* $h_N : \mathcal{Y}^N \to \mathcal{C}_n$ *is called* maximum likelihood rule (ML) *if*

$$\boldsymbol{c}_j = h_N(\boldsymbol{b}) \Rightarrow P(\boldsymbol{Y}_N = \boldsymbol{b} \mid \boldsymbol{X}_N = \boldsymbol{c}_j) \geq P(\boldsymbol{Y}_N = \boldsymbol{b} \mid \boldsymbol{X}_N = \boldsymbol{c}_i)$$

for all $i = 1, \ldots, M$.

With ML-decoding, $\boldsymbol{b}$ *is decoded as the codeword* $\mathbf{c}_j$ *which has greatest conditional probability of* $\boldsymbol{b}$ *being received given that* $\mathbf{c}_j$ *was sent. Hence,*

$$h_N(\boldsymbol{b}) \in \arg \max_{i=1,\ldots,M} P\big(\boldsymbol{Y}_N = \boldsymbol{b} \mid \boldsymbol{Y}_N = \boldsymbol{c}_i\big).$$

4.6 Error Probabilities

For a given Code $\mathcal{C}_N = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\}$,

- $$e_j(\mathcal{C}_N) = P\big(h_N(\boldsymbol{Y}_N) \neq \boldsymbol{c}_j \mid \boldsymbol{X}_N = \boldsymbol{c}_j\big)$$

  is the probability for a decoding error of code word $\boldsymbol{c}_j$.

- $$e(\mathcal{C}_N) = \sum_{j=1}^{M} e_j(\mathcal{C}_N)\, P(\boldsymbol{X}_N = \boldsymbol{c}_j)$$

  is the error probability of code $\mathcal{C}_N$.

- $$\hat{e}(\mathcal{C}_N) = \max_{j=1,\dots,M} e_j(\mathcal{C}_N)$$

  is the maximum error probability.

4.7 Discrete Memoryless Channel

Definition 4.13. *A discrete channel is called* memoryless (DMC) *if*

$$P\big(\boldsymbol{Y}_N = \boldsymbol{b}_N \mid \boldsymbol{X}_n = \boldsymbol{a}_N\big) = \prod_{i=1}^{N} P\big(Y_1 = b_i \mid X_1 = a_i\big)$$

for all $N \in$, $\boldsymbol{a}_N = (a_1, \dots, a_N) \in \mathcal{X}^N$, $\boldsymbol{b}_N = (b_1, \dots, b_N) \in \mathcal{Y}^N$.

Remark 4.14. *From the above definition it follows that the channel*

- *is memoryless and nonanticipating*
- *transition probablities of symbols are the same at each position*
- *transition probabilities of blocks only depend on the channel matrix*

Definition 4.15. *Suppose a source produces R bits per second (rate R). Hence ,NR bits in N seconds. Let the total no of messages in N seconds is 2^{NR}(assigned as integer) and M codewords available for encoding all messages.*

$$M = 2^{NR} \Longleftrightarrow R = \frac{\log M}{N}$$

(No of bits per channel use)

Lemma 4.16. *$(\boldsymbol{X}_N, \boldsymbol{Y}_N)$ is a DMC iff $\forall l = 1 \dots N$*

$$P(Y_l = b_l \mid X_1 = a_1, \dots, X_N = a_N, Y_1 = b_1, \dots Y_{l-1} = b_{l-1}) = P(Y_1 = b_l \mid X_1 = a_l)$$

Proof. " $\Longleftarrow$ "
$P(\boldsymbol{Y}_N = \boldsymbol{b}_N \mid \boldsymbol{X}_N = \boldsymbol{a}_N)$

$$= P(\boldsymbol{Y}_N = b_N \mid \boldsymbol{X}_N = \boldsymbol{a}_N, \boldsymbol{Y}_{N-1} = \boldsymbol{b}_{N-1}).\frac{P(\boldsymbol{Y}_{N-1} = \boldsymbol{b}_{N-1}, \boldsymbol{X}_N = \boldsymbol{a}_N)}{P(\boldsymbol{X}_N = \boldsymbol{a}_N)}$$
$$= P(Y_1 = b_N \mid X_1 = a_N)P(\boldsymbol{Y}_{N-1} = \boldsymbol{b}_{N-1}, \boldsymbol{X}_N = \boldsymbol{a}_N)$$
$$= P(Y_1 = b_N \mid X_1 = a_N)P(\boldsymbol{Y}_{N-1} = \boldsymbol{b}_{N-1} \mid X_1 = a_{N-1})P(\boldsymbol{Y}_{N-2} = \boldsymbol{b}_{N-2} \mid \boldsymbol{X}_N = \boldsymbol{a}_N)$$
$$= \dots$$
$$= \Pi_{i=1}^{N} P(Y_1 = b_i \mid X_1 = a_i)$$

$\Longrightarrow$

$$\begin{aligned} P(Y_l = b_l \mid \boldsymbol{X}_N, \boldsymbol{Y}_{e-1} = \boldsymbol{b}_{e-1}) &= \frac{P(\boldsymbol{Y}_e = \boldsymbol{b}_e \mid \boldsymbol{X}_N = \boldsymbol{a}_N)}{P(\boldsymbol{Y}_{e-1} = \boldsymbol{b}_{e-1} \mid \boldsymbol{X}_N = \boldsymbol{a}_N)} \\ &= P(Y_1 = b_l \mid X_1 = a_l) \end{aligned}$$

$\{(X_n, Y_n)\}$ is a sequence of independent random variable system then $(\boldsymbol{X}_N, \boldsymbol{Y}_N)$ forms a DMC

□

4.8 The Noisy Coding Theorem

Theorem 4.17. *(Shannon 1949)*

Given some discrete memoryless channel of capacity C. Let $0 < R < C$ and $M_N \in$ be a sequence of integers such that

$$\frac{\log M_N}{N} < R.$$

There exists a sequence of (N, M_N)-codes with M_N codewords of length N and a constant $a > 0$ such that

$$\hat{e}(\mathcal{C}_N) \le e^{-Na}.$$

Hence, the maximum error probability tends to zero exponentially fast as the block length N tends to infinity.

Example 4.18. *Consider the BSC with $\varepsilon = 0.03$.*

$$C = 1 + (1-\varepsilon)\log_2(1-\varepsilon) + \varepsilon \log_2 \varepsilon = 0.8056$$

Choose $R = 0.8$

$$\frac{\log_2 M_N}{N} < R \Leftrightarrow M_N < 2^{NR}$$

hence choose

$$M_N = \lfloor 2^{0.8N} \rfloor$$

N	*10*	*20*	*30*
$\lvert\mathcal{X}^N\rvert = 2^N$	$1\,024$	$1\,048\,576$	$1.0737 \cdot 10^9$
$M_N = \lfloor 2^{0.8N} \rfloor$	256	$65\,536$	$16.777 \cdot 10^6$
Percentage of used codewords	25%	6.25%	1.56%

4.9 Converse of the Noisy Coding Theorem

Theorem 4.19. *(Wolfowitz 1957)*

Given some discrete memoryless channel of capacity C. Let $R > C$ and $M_N \in$ be a sequence of integers such that

$$\frac{\log M_N}{N} > R.$$

For any sequence of (N, M_N)-codes with M_N codewords of length N it holds that that

$$\lim_{N \to \infty} e(\mathcal{C}_N) = 1.$$

Hence, such codes tend to be fully unreliable.

Theorem 4.20. *(Only the outline of the proof) Use random coding, i.e r.v., $C_1 \dots C_M \in \mathcal{X}^N, C_i = (C_{i1}, \dots, C_{iN}) i = 1, .., M$ with $C_{ij} \in \mathcal{X}, i.i.d \sim p(x), i = 1, ..., M, j = 1, .., N$.*

(a)

Theorem 4.21. *For a DMC with ML-decoding, it holds for all $0 \leq \gamma \leq 1, j = 1, ...M$*

$$E(e_j(C_1, \dots, C_M)) \leq (M-1)^\gamma (\sum_{j=1}^{d} (\sum_{i=1}^{m} p_i p_1(y_j \mid x_i)^{\frac{1}{1+\gamma}})^{1+\gamma})^N$$

Proof. Set

$$G(\gamma, \boldsymbol{p}) = -\ln(\sum_{j=1}^{d} (\sum_{i=1}^{m} p_i p_1(y_j \mid x_i)^{\frac{1}{1+\gamma}})^{1+\gamma})$$

and $R = \frac{\ln M}{N}$.

$$E(e_j(C_1, \dots, c_M)) \leq exp(-N(G(\gamma, \boldsymbol{p}) - \gamma R)$$

Set $G^*(R) = \max_{0 \leq \gamma \leq 1} \max_{\boldsymbol{p}} \{G(\gamma, \boldsymbol{p}) - \gamma R\}$ □

(b)

Theorem 4.22. *For a DMC with MC decoding there exists a code $c_1, \dots, c_M \in \mathcal{X}^N$s.t $\hat{e}(c_1, \dots, c_M) \leq 4e^{-NG^*(R)}$*

Proof. Use $2M$ random codewords. Then

$$\frac{1}{2M} \sum_{j=1}^{2M} E(e_j(C_1, \dots, C_{2M})) \leq e^{-NG^*(\frac{\ln 2M}{N})}$$

There exists a sample $c_1, \dots, c_{2M}$ s.t

$$\frac{1}{2M} \sum_{j=1}^{2M} e_j(C_1, \dots, C_{2M}) \leq e^{-NG^*(\frac{\ln 2M}{N})} \quad (*).$$

Remove M codewords, particularly with

$$e_k(c_1,\dots,c_{2M}) > 2e^{-NG^*(\frac{\ln 2M}{N})}$$

There are at most M,other wise $(*)$ would be violated. For the remaining ones

$$e_j(c_{i1},\dots,c_{iM}) \leq 4e^{-NG^*(R)} \;\; \forall_j = 1\dots M.$$

□

(c)

Theorem 4.23. *If* $R = \frac{lnM}{N} < C, then$

$$\begin{aligned} G^*(R) &= \max_p \max_{0\leq\gamma\leq 1}\{G(\gamma,\boldsymbol{p}) - \gamma R\} \\ &\geq \max_{0\leq\gamma\leq 1}\{G(\gamma,\boldsymbol{p}^*) - \gamma R\} > 0 \end{aligned}$$

where $\mathbf{p}^*$ *denotes the capacity-achieving distribution. For detailed proof please refer RM p 103-114)*

5 Rate Distortion Theory

Motivation:

a) By the source coding theorem(Th 3.7 and 3.9): error free / loss less encoding needs at least on average $H(X)$ bits per symbol.

b) ASignal is represented by bits. What is the min no of bits needed not to exceed a certain maximum distortion?.

Example 5.1. *a) Representing a real number by k bits:* $\mathcal{X} = \mathrm{I\!R}$ *,*$\hat{\mathcal{X}} = \{(b_1, \ldots, b_k) \mid b_i \in \{0, 1\}\}$

b) 1-bit quantization $\mathcal{X} = \mathrm{I\!R}$ *,*$\hat{\mathcal{X}} = \{0, 1\}$

Definition 5.2. *A distortion function measure is a mapping* $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$.

Examples:

a) Hamming distance , $\mathcal{X} = \hat{\mathcal{X}} = \{0, 1\}$

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & \textit{otherwise} \end{cases}$$

b) Squared error :$d(\hat{x}, x) = (x - \hat{x})^2$.

Definition 5.3. *The distortion measure between sequence* $x^n, \hat{x}^n$ *is defined as*

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)\,.$$

Definition 5.4. *A* $(2^{nR}, n)$ *rate distortion code of rate R and block length n consists an encoder*

$$f_n : \mathcal{X}^n \to \{1, 2 \ldots 2^{nR}\}$$

and a decoders

$$g_n : \{1, 2 \ldots 2^{nR}\} \to \hat{\mathcal{X}}^n\,.$$

The expected distortion of the (f_n, g_n) *is*

$$D = Ed(X^n, \hat{X}^n) = Ed(X^n, g_n(f_n(x^n))).$$

Remarks:

a) $\mathcal{X}, \hat{\mathcal{X}}^n$ are assumed to be finite

b) 2^{nR} means $\lceil 2^{nR} \rceil$,if it is not integer

c) f_n yields 2^{nR} different values. We need $\approx nR$ bits to represent each. Hence, $R =$ number of bits per source symbol needed to represent $f_n(X^n)$

d)

$$D = Ed(X^n, \hat{X}^n) \quad = Ed(X_n, g_n(f_n(X^n))) = \sum_{x^n \in \hat{\mathcal{X}}^n} p(x^n) d(x_n, g_n(f_n(x_n)))$$

e) $\{g_n(1), \ldots, g_n(2^{nR})\}$ is called codebook, while $f_n^{-1}(1), \ldots, f_n^{-1}(2^{nR})$ are called assignment regions.

Ultimate goal of lossy source coding is to

- minimise R for a given D or
- minimise D for a given R.

Definition 5.5. *A rate distortion pair (R, D) is called achievable if there exists a sequence of $(2^{nR}, n)$ rate distortion codes such that,*

$$\lim_{n\to\infty} Ed(X_n, g_n(f_n(X_n))) \leq D.$$

Definition 5.6. *The rate distortion function is defined as*

$$R(D) = \inf_R (R, D)$$

is achievable.

Definition 5.7. *The informatin distortion function $R_I(D)$ is defined as fallows:*

$$\begin{aligned} R_I(D) &= \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x,\hat{x}) d(x,\hat{x}) \leq D} I(X, \hat{X}) \\ &= \min_{p(\hat{x}|x): E[d(X,\hat{X})] \leq D} I(X, \hat{X}) \,. \end{aligned}$$

Compare with capacity:

- *C: given $p(\hat{x} \mid x), \max I(X, \hat{X})$ over the input distribution,*
- *$R_I(D)$: given $p(x)$, $\min(X, \hat{X})$ over "channels" s.t. the expected distortion does exceed D.*

Theorem 5.8. *a) $R_I(D)$ is a convex non increasing function of D.*

b) $R_I(D)$ =0 if $D > D^*$,$D^* = min_{\hat{x} \in \hat{\mathcal{X}}} Ed(X, \hat{x})$.

c) $R_I(0) \leq H(X)$.

Proof. Yeung p.198 ff, Cover and Thomas p.316 ff. □

Theorem 5.9. *$X \in \{0, 1\}, P(X = 0) = 1 - p, P(X = 1) = p, 0 \leq p \leq 1$ and d is the Hamming distance.*

$$R_I(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\} \\ 0, & \textit{otherwise} \end{cases}$$

Proof. w.l.o.g assume $p < \frac{1}{2}$, otherwise interchange 0 and 1

$$\min_{p(\hat{x}|x):E[d(X,\hat{X})]\leq D} I(X,\hat{X})\,.$$

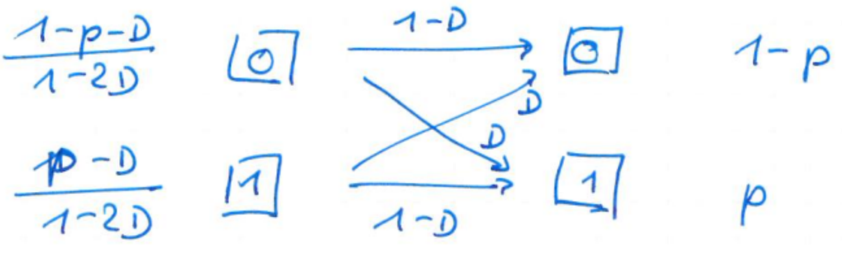
Assume $D \leq p < \frac{1}{2}$, then

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X \mid \hat{X}) \\ &= H(X) - H(X \oplus \hat{X} \mid \hat{X}) \\ &\geq H(X) - H(X \oplus \hat{X}) \\ &= H(p) - H(P(X \neq \hat{X})) \\ &\geq H(p) - H(D). \end{aligned}$$

This lower bound is attained by the following joint distribution of $(X, \hat{X})$.

$X \backslash \hat{X}$	0	1	
0	$\frac{(1-D)(1-p-D)}{1-2D}$	$\frac{D(p-D)}{1-2D}$	$1-p$
1	$\frac{D(1-p-D)}{1-2D}$	$\frac{1-D(p-D)}{1-2D}$	p
Total	$\frac{(1-p-D)}{1-2D}$	$\frac{(p-D)}{1-2D}$	1

This corresponds to the following BSC.



It follows that

$$\begin{aligned} P(X \neq \hat{X}) &= Ed(X,\hat{X}) \\ &= \frac{D(p-D)}{1-2D} + \frac{D(1-p-D)}{1-2D} = D. \end{aligned}$$

Further,

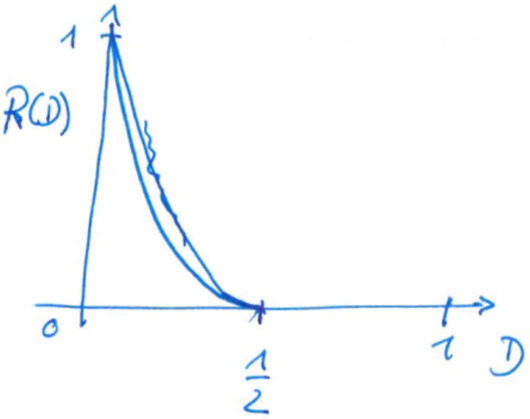
$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X \mid \hat{X}) \\ &= H(p) - [\underbrace{H(X \mid \hat{X}=0)}_{H(D)} P(\hat{X}=0) + \underbrace{H(X \mid \hat{X}=1)}_{H(1-D)=H(D)} P(\hat{X}=1)] \\ &= H(p) - H(D). \end{aligned}$$

X \ $\hat{X}$	0	1	
0	$1-p$	0	$1-p$
1	p	0	p
Total	1	0	1

such that the lower bound is attained. If $D \geq p$ set $P(\hat{X}=0)=1$ and get
Then $Ed(X;\hat{X}) = P(X \neq \hat{X}) = P(X=1) = p \leq D$ and

$$I(X;\hat{X}) = H(X) - H(X \mid \hat{X}) \quad = H(p) - H(X \mid \hat{X}=0).1 = H(p) - H(p) = 0.$$

Plot for Bin (1 ,$\frac{1}{2}$):



□

Theorem 5.10. *(converse to the rate distortion theorem)*

$$R(D) \geq R_I(D)$$

Proof. Recall the general situation $X_1, \dots X_n$ i.i.d $\sim X \sim p(x), x \in \mathcal{X}, \hat{\mathcal{X}}^n = g_n(f_n(X^n))$ has at most 2^{nR} values. Hence

$$H(\hat{X}^n) \leq \log 2^{nR} \leq nR.$$

We first show (R,D) achievable $\Rightarrow R \geq R_I(D)$ suppose (R,D) is achievable. Then

$$\begin{aligned}
nR &\geq H(\hat{X}) \\
&\geq H(\hat{X}) - H(\hat{X} \mid X^n) \\
&= I(\hat{X}^n; X^n) = I(X^n, \hat{X}^n) \\
&= H(X_n) - H(X^n \mid \hat{X}^n) \\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \hat{X}^n, (X_1, \dots X_{i=1})) \\
&\geq \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\
&\geq \sum_{i=1}^{n} R_I(Ed(X_i, \hat{X}_i)) \\
&= n \sum_{i=1}^{n} \frac{1}{n} R_I(Ed(X_i, \hat{X}_i)) \\
&\geq nR_I(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \\
&= nR_I Ed(X^n, \hat{X}^n) \\
&\geq nR_I(D)
\end{aligned}$$

that is $R = R(D) \geq R_I(D)$, hence (R, D) is achievable. □

The reverse inequality in th 5.10 also holds
Theorem 5.11.

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

Proof.

- $R(D) \geq R_I(D)$
- $R(D) \leq R_I(D)$

□

Yeung :Section 9.5, p. 206-212, Cover and thomas section 10.5 p.318-324.