INFORMATION THEORY AND STATISTICS

Lecture notes and exercises

Spring 2013

Jüri Lember

Literature:

1. **T.M. Cover, J.A. Thomas "Elements of information theory"**, Wiley, 1991 ja 2006;
2. Yeung, Raymond W. "A first course of information theory", Kluwer, 2002;
3. Te Sun Han, Kingo Kobayashi "Mathematics of information and coding", AMS, 1994;
4. Csiszar, I., Shields, P. "Information theory and statistics : a tutorial", MA 2004;
5. Mackay, D. "Information theory, inference and learning algorithms", Cambridge 2004;
6. McEliece, R. "Information and coding", Cambridge 2004;
7. Gray, R. "Entropy and information theory", Springer 1990;
8. Gray, R. "Entropy and information theory", Springer 1990;
9. Gray, R. "Source coding theory", Kluwer, 1990;
10. Shields, P. "The ergodic theory of discrete sample paths", AMS 1996;
11. Dembo, A., Zeitouni, O. "Large deviation techniques and Applications", Springer 2010.
12. $\cdots$

Lecture notes:

https://noppa.aalto.fi/noppa/kurssi/mat-1.c/information_theory_and_statistics

1 Main concepts

1.1 (Shannon) entropy

In what follows, let $\mathcal{X} = \{x_1, x_2, \ldots\}$ be a discrete (finite or countably infinite) *alphabet*. Let X be a random variable taking values on $\mathcal{X}$ with distribution P. We shall denote

$$p_i := \mathbf{P}(X = x_i) = P(x_i).$$

Thus, for every $A \subset \mathcal{X}$

$$P(A) = \mathbf{P}(X \in A) = \sum_{i: x_i \in A} p_i = \sum_{x \in A} P(x).$$

Since $\mathcal{X}$ is fixed, the distribution P can be uniquely represented via the probabilities p_i i.e.

$$P = (p_1, p_2, \ldots).$$

Recall that the *support* of P, denoted via $\mathcal{X}_P$ is the set of letters having positive probability (atoms), i.e.

$$\mathcal{X}_P := \{x \in \mathcal{X} : P(x) > 0\}.$$

Also recall that for any $g : \mathcal{X} \to \mathbb{R}$ such that $\sum p_i |g(x_i)| < \infty$, the expectation of $g(X)$ is defines as follows

$$Eg(X) = \sum p_i g(x_i) = \sum_{x \in \mathcal{X}} g(x) P(x) = \sum_{x \in \mathcal{X}_P} g(x) P(x). \tag{1.1}$$

NB! In what follows $\log := \log_2$ and $0 \log 0 := 0$.

1.1.1 Definition and elementary properties

Def 1.1 *The* **(Shannon) entropy** *of random variable X (distribution P) $H(X)$ is*

$$H(X) = -\sum p_i \log p_i = -\sum_{x \in \mathcal{X}} P(x) \log P(x).$$

Remarks:

- $H(X)$ depends on X via P, only.
- By (1.1)
$$H(X) = E\big(-\log P(X)\big) = E \log \frac{1}{P(X)}.$$
- The sum $\sum -p_i \log p_i$ is always defined (since $-p_i \log p_i \geq 0$), but can be infinite. Hence
$$0 \leq H(X) \leq \infty,$$
and $H(X) = 0$ iff for a letter x, $X = x$, a.s..

- Entropy does not depend on the alphabet $\mathcal{X}$, it only depends on probabilities p_i. Hence, we can also write
$$H(p_1, p_2, \ldots).$$

- In principle, any other logarithm $\log_b$ can be used in the definition of entropy. Such entropy is denoted by H_b i.e.
$$H_b(X) = -\sum p_i \log_b p_i = -\sum_{x\in\mathcal{X}} P(x) \log_b P(x).$$
since $\log_b p = \log_b a \log_a p$, it holds
$$H_b(X) = (\log_b a) H_a(X),$$
so that $H_b(X) = (\log_b 2)H(X)$ and $H_e(X) = (\ln 2)H(X)$. In information theory, typically, $\log_2$ is used and such entropy is measured in *bits*. The entropy defined with ln is measured in *nats*, the entropy defined with $\log_{10}$ is measured in *dits*.

The number $-\log p(x_i)$ can be interpreted as the amount of information one gets if X takes x_i. The smaller $p(x_i)$, the bigger is the amount of information. The entropy is thus the average amount of information or randomness X contains – the bigger $H(X)$, the more random is X. The concept of entropy was introduced by C. Shannon in his seminal paper "A mathematical theory of communication" (1948).

Examples:

1 Let $\mathcal{X} = \{0, 1\}$, $p = \mathbf{P}(X = 1)$, i.e. $X \sim B(1, p)$. Then
$$H(X) = -p \log p - (1-p)\log(1-p) =: h(p).$$
The function $h(p)$ is called the **binary entropy function**. The function $h(p)$ is concave, symmetric around $\frac{1}{2}$ and has maximum at $p = \frac{1}{2}$:
$$h(\frac{1}{2}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 1.$$

2 Consider the distributions

P:	a	b	c	d	e
	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

Q:	a	b	c	d
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$H(P) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - \frac{1}{16}\log\frac{1}{16} = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} = \frac{15}{8}$$
$$H(Q) = \log 4 = 2.$$

Thus P is "less random ", although the number of atoms (the letters with positive probability) is bigger.

1.1.2 Axiomatic approach

The entropy has the property of *grouping*

$$H(p_1, p_2, p_3, \ldots) = H(\Sigma_{i=1}^k p_i, p_{k+1}, p_{k+2}, \ldots) + \big(\Sigma_{i=1}^k p_i\big) H\Big(\frac{p_1}{\Sigma_{i=1}^k p_i}, \ldots, \frac{p_k}{\Sigma_{i=1}^k p_i}\Big). \quad (1.2)$$

The proof of (1.2) is Exercise 2. In a sense, grouping is a natural "additivity" property that a measure of information should have. It turns out that when $\mathcal{X}$ is finite, then grouping together with symmetry and continuity implies entropy.

More precisely, let for any m, $\mathcal{P}^m$ be the set all probability measures in m-dimensional alphabet, i.e.

$$\mathcal{P}^m := \Big\{(p_1, \ldots, p_m) : p_i \geq 0, \sum_{i=1}^m p_i = 1\Big\}.$$

Suppose, for every m we have a function $f_m : \mathcal{P}^m \to [0, \infty)$ that is a candidate for a measure of information. The function f_m is continuous if it is continuous with respect to all coordinates, and it is symmetric, if it value is independent of the order of the arguments.

Theorem 1.2 *Let, for every m, $f_m : \mathcal{P}^m \to [0, \infty)$ be symmetric functions satisfying the following axioms:*

A1 *f_2 is normalized, i.e. $f_2(\frac{1}{2}, \frac{1}{2}) = 1$;*

A2 *f_m is continuous for every $m = 2, 3, \ldots$;*

A3 *it has the grouping property: for every $1 < k < m$,*

$$f_m(p_1, p_2, \ldots, p_m) = f_{m-k+1}(\Sigma_{i=1}^k p_i, p_{k+1}, \ldots, p_m) + \big(\Sigma_{i=1}^k p_i\big) f_k\Big(\frac{p_1}{\Sigma_{i=1}^k p_i}, \ldots, \frac{p_k}{\Sigma_{i=1}^k p_i}\Big).$$

A4 *for every $m < n$, it holds $f_m(\frac{1}{m}, \ldots, \frac{1}{m}) \leq f_n(\frac{1}{n}, \ldots, \frac{1}{n})$.*

Then for every m,

$$f_m(p_1, \ldots, p_m) = -\sum_{i=1}^m p_i \log p_i. \quad (1.3)$$

Proof. Let, for every m,

$$g(m) := f_m(\frac{1}{m}, \ldots, \frac{1}{m}).$$

By symmetry and applying **A3** m times, we obtain

$$\begin{aligned} g(mn) &= f_{nm}\Big(\underbrace{\frac{1}{nm}, \ldots, \frac{1}{nm}}_{n}, \ldots, \underbrace{\frac{1}{nm}, \ldots, \frac{1}{nm}}_{n}\Big) \\ &= f_m(\frac{1}{m} \ldots, \frac{1}{m}) + f_n\big(\frac{1}{n}, \ldots, \frac{1}{n}\big) = g(m) + g(n). \end{aligned}$$

Hence, for integers n and k, $g(n^k) = kg(n)$ and by **A1**, $g(2^k) = kg(2) = k$ i.e.

$$g(2^k) = \log(2^k), \quad \forall k.$$

Using **A4**, it is possible to show that the equality above holds for every integer n, i.e.

$$g(n) = \log n, \quad \forall n \in \mathbb{N}.$$

Fix an arbitrary m and consider $(p_1, \ldots, p_m)$, where all components are rational. Then, there exist integers $k_1, \ldots, k_m$ and common denominator n such that $p_i = \frac{k_i}{n}$, $i = 1, \ldots, m$. In this case,

$$\begin{aligned} g(n) &= f_n\big(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{k_1}, \underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{k_2}, \ldots, \underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{k_m}\big) \\ &= f_m(\frac{k_1}{n}, \ldots, \frac{k_m}{n}) + \sum_{i=1}^{m} \frac{k_i}{n} f_{k_i}(\frac{1}{k_i}, \ldots, \frac{1}{k_i}) \\ &= f_m(p_1, \ldots, p_m) + \sum_{i=1}^{m} \frac{k_i}{n} g(k_i) = f_m(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i \log(k_i). \end{aligned}$$

Therefore,

$$f_m(p_1, \ldots, p_m) = \log(n) - \sum_{i=1}^{m} p_i \log(k_i) = -\sum_{i=1}^{m} p_i \log(\frac{k_i}{n}) = -\sum_{i=1}^{m} p_i \log p_i$$

so that (1.3) holds when all p_i are rational. Now use continuity of f_m to deduce that (1.3) always holds. ■

Remark: One can drop the axiom **A4**.

1.1.3 Entropy is strictly concave

Jensen's inequality. We shall often use Jensen's inequality. Recall that a function $g : \mathbb{R} \to \mathbb{R}$ is *convex*, if for every x_1, x_2 and $\lambda \in [0, 1]$, it holds

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2).$$

A function g is strictly convex, if equality holds only for $\lambda = 1$ or $\lambda = 0$. A function g is *concave*, if $-g$ is convex.

Theorem 1.3 (Jensen's inequality). *Let g be convex function and X a random variable such that $E|g(X)| < \infty$ and $E|X| < \infty$. Then*

$$Eg(X) \geq g(EX). \tag{1.4}$$

If g is strictly convex, then (1.4) is equality if and only if $X = EX$ a.s..

Mixture of distributions and the concavity of entropy. Let P_1 and P_2 be two distributions given in $\mathcal{X}$. (Note that any two discrete distributions can be defined in a common alphabet like the union of their supports). The *mixture of* P_1 *and* P_2 is their convex combination:

$$Q = \lambda P_1 + (1-\lambda)P_2, \quad \lambda \in (0,1).$$

When $X_1 \sim P_1$, $X_2 \sim P_2$ and $Z \sim B(1,\lambda)$, then the following random variable has the mixture distribution Q:

$$Y = \begin{cases} X_1 & \text{if } Z = 1, \\ X_2 & \text{if } Z = 0. \end{cases}$$

Clearly Q contains the randomness of P_1 and P_2. In addition, Z is random.

Proposition 1.1 *Entropy is strictly concave i.e.*

$$H(Q) \geq \lambda H(P_1) + (1-\lambda)H(P_2)$$

and the inequality is strict except when $P_1 = P_2$.
When $\mathcal{X}_{P_1}$ *and* $\mathcal{X}_{P_2}$ *are disjoint, then*

$$H(Q) = \lambda H(P_1) + (1-\lambda)H(P_2) + h(\lambda). \tag{1.5}$$

Proof. The function $f(y) = -y\log y$ is strictly concave ($y \geq 0$). Thus, for every $x \in \mathcal{X}$

$$\begin{aligned} -\lambda P_1(x)\log P_1(x) - (1-\lambda)P_2(x)\log P_2(x) &= \lambda f\big(P_1(x)\big) + (1-\lambda) f\big(P_2(x)\big) \\ &\leq f\Big(\lambda P_1(x) + (1-\lambda)P_2(x)\Big) = -Q(x)\log Q(x). \end{aligned}$$

Sum over $\mathcal{X}$ to get

$$\lambda H(P_1) + (1-\lambda)H(P_2) \leq H(Q).$$

The inequality is strict, when there is at least one $x \in \mathcal{X}$ so that $P_1(x) \neq P_2(x)$.
The proof of (1.5) is Exercise 5. ■

Example: Let $P_1 = B(1,p_1)$ and $P_2 = B(1,p_2)$ (both Bernoulli distributions). Then the mixture $\lambda P_1 + (1-\lambda)P_2$ is $B(1, \lambda p_1 + (1-\lambda)p_2)$. The concavity of entropy implies that binary entropy function $h(p)$ is strictly concave: $h(\lambda p_1 + (1-\lambda)p_2) \geq \lambda h(p_1) + (1-\lambda)h(p_2)$.

1.2 Joint entropy

Let X and Y be random variables taking values in discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then (X,Y) is random vector with support in

$$\mathcal{X} \times \mathcal{Y} = \{(x,y) : x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

Let P be the (joint) distribution of (X,Y), a probability measure on $\mathcal{X} \times \mathcal{Y}$. Denote

$$p_{ij} := P(x_i, y_j) = \mathbf{P}\big((X,Y) = (x_i, y_j)\big) = \mathbf{P}(X = x_i, Y = y_j).$$

Joint distribution is often represented by the following table

$\mathcal{X}\backslash\mathcal{Y}$	y_1	y_2	$\dots$	y_j	$\dots$	$\sum$
x_1	$P(x_1,y_1)=p_{11}$	$P(x_1,y_2)=p_{12}$	$\dots$	p_{1j}	$\dots$	$\sum_j p_{1j}=P(x_1)$
x_2	$P(x_2,y_1)=p_{21}$	$P(x_1,y_2)=p_{22}$	$\dots$	p_{2j}	$\dots$	$\sum_j p_{2j}=P(x_2)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$	$\sum_j p_{ij}=P(x_i)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\sum$	$\sum_i p_{i1}=P(y_1)$	$\sum_i p_{i2}=P(y_2)$	$\dots$	$\sum_i p_{ij}=P(y_j)$	$\dots$	1

In the table and in what follows (with some abuse of notation),

$$P(x) := \mathbf{P}(X=x) \quad \text{and} \quad P(y) := \mathbf{P}(Y=y)$$

denote marginal laws. The random variables X and Y are independent if and only if

$$P(x,y) = P(x)P(y) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

The random vector (X,Y) can be considered as a random variable in a product alphabet $\mathcal{X}\times\mathcal{Y}$, and the entropy of such a random variable is

$$H(X,Y) = -\sum_{ij} p_{ij}\log p_{ij} = -\sum_{(x,y)\in\mathcal{X}\times Y} P(x,y)\log P(x,y) = E\Big(-\log P(X,Y)\Big). \quad (1.6)$$

Def 1.4 *The entropy $H(X,Y)$ as defined in (1.6) is called the* **joint entropy** *of (X,Y).*

Independent X and Y. When X and Y are independent, then

$$\begin{aligned} H(X,Y) &= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x,y)\log P(x,y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x)P(y)(\log P(x)+\log P(y)) \\ &= -\sum_{x\in\mathcal{X}} P(x)\log P(x) - \sum_{y\in\mathcal{Y}} P(y)\log P(y) = H(X)+H(Y). \end{aligned}$$

The argument above can be restate as follows. For every $x\in\mathcal{X}$ and $y\in\mathcal{Y}$ it holds $\log P(x,y) = \log P(x)+\log P(y)$ so that

$$\log P(X,Y) = \log P(X) + \log P(Y).$$

Expectation is linear

$$\begin{aligned} H(X,Y) &= -E\big(\log P(X,Y)\big) = -E\big(\log P(X)+\log P(Y)\big) \\ &= -E\log P(X) - E\log P(Y) = H(X)+H(Y). \end{aligned}$$

The joint entropy of several random variables. By analogy, the joint entropy of several random variables $X_1,\dots,X_n$ is defined

$$H(X_1,\dots,X_n) := -E\log P(X_1,\dots,X_n).$$

When all random variables are independent, then

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

1.3 Conditional entropy

1.3.1 Definition

Let x be such that $P(x) > 0$. Then define the conditional probabilities

$$P(y|x) := \mathbf{P}(Y = y|X = x) = \frac{P(x,y)}{P(x)}.$$

The conditional distribution of Y given $X = x$ is

y_1	y_2	y_3	$\dots$
$P(y_1\|x)$	$P(y_2\|x)$	$P(y_2\|x)$	$\dots$

.

The entropy of that distribution is

$$H(Y|x) :=: H(Y|X = x) := -\sum_{y\in\mathcal{Y}} P(y|x)\log P(y|x).$$

Consider the function $x \mapsto H(Y|x)$. Applying it to the random variable $X \sim P$, we get a new random variable (the function of X) with distribution

$H(Y\|x_1)$	$H(Y\|x_2)$	$H(Y\|x_3)$	$\dots$
$P(x_1)$	$P(x_2)$	$P(x_3)$	$\dots$

.

and expectation

$$\sum_{x\in\mathcal{X}_P} H(Y|x)P(x).$$

Def 1.5 *The* **conditional entropy** *of Y given $X \sim P$ is*

$$\begin{aligned} H(Y|X) &:= \sum_{x\in\mathcal{X}_P} H(Y|x)P(x) = -\sum_{x\in\mathcal{X}_P} P(x)\sum_{y\in\mathcal{Y}} \log P(y|x)P(y|x) \\ &= -\sum_{x\in\mathcal{X}_P}\sum_{y\in\mathcal{Y}} \log P(y|x)P(x,y) = -E\Big(\log P(Y|X)\Big). \end{aligned}$$

Remarks:

- When X and Y are independent, then $P(y|x) = P(y)$ $\forall x \in \mathcal{X}_P, y \in \mathcal{Y}$ so that $H(Y|X) = H(Y)$.
- In general $H(X|Y) \neq H(Y|X)$ (take independent X, Y such that $H(X) \neq H(Y)$).
- $H(Y|X) = 0$ iff for a function f, $Y = f(X)$. Indeed, $H(Y|X) = 0$ iff

  $$H(Y|X = x) = 0 \text{ for every } x \in \mathcal{X}_P.$$

  Hence, there exists $f(x)$ such that $\mathbf{P}(Y = f(x)|X = x) = 1$ or $Y = f(X)$.

Joint entropy for more than two random variables. Let X, Y, Z be random variables with supports $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$. Considering the vector (X, Y) (or the vector (Y, Z)) as a random variable, we have

$$\begin{aligned} H(X,Y|Z) &:= -\sum_{z\in\mathcal{Z}} P(z) \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x,y|z)\log P(x,y|z) \\ &= -\sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} \log P(x,y|z)P(x,y,z) = -E\log P(X,Y|Z) \\ H(X|Y,Z) &:= -\sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} P(y,z)\sum_{x\in\mathcal{X}} P(x|y,z)\log P(x|y,z) \\ &= -\sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} \log P(x|y,z)P(x,y,z) = -E\log P(X|Y,Z). \end{aligned}$$

Moreover, given any set $X_1, \ldots, X_n$ of random variables, one can similarly define conditional entropies

$$H(X_n, X_{n-1}, \ldots, X_j | X_{j-1}, \ldots, X_1).$$

1.3.2 Chain rules for entropy

Lemma 1.1 (Chain rule) *Let $X_1, \ldots, X_n$ be random variables. Then*

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}).$$

Proof. For any $(x_1, \ldots, x_n)$ such that $P(x_1, \ldots, x_n) > 0$, it holds

$$P(x_1, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\cdots P(x_n|x_1, \ldots, x_n),$$

so that

$$\begin{aligned} H(X_1, \ldots, X_n) &= -E\log P(X_1, \ldots, X_n) \\ &= -E\log P(X_1) - E\log P(X_2|X_1) - \cdots - E\log P(X_n|X_1, \ldots, X_{n-1}) \\ &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \ldots, X_{n-1}). \end{aligned}$$

■

In particular, for any random vector (X, Y)

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

Lemma 1.2 (Chain rule for conditional entropy) *Let $X_1, \ldots, X_n, Z$ be random variables. Then*

$$H(X_1, \ldots, X_n|Z) = H(X_1|Z) + H(X_2|X_1, Z) + H(X_3|X_1, X_2, Z) + \cdots + H(X_n|X_1, \ldots, X_{n-1}, Z).$$

Proof. For every $(x_1, \ldots, x_n, z)$ such that $P(x_1, \ldots, x_n, z) > 0$, it holds

$$P(x_1, \ldots, x_n|z) = P(x_1|z)P(x_2|x_1, z)P(x_3|x_2, x_1, z) \cdots P(x_n|x_1, \ldots, x_{n-1}, z)$$

so that

$$\log P(X_1, \ldots, X_n|Z) = \log P(X_1|Z) + \log P(X_2|X_1, Z) + \cdots + P(X_n|X_1, \ldots, X_{n-1}, Z).$$

Now take expectation. ■

In particular, for any random vector (X, Y, Z)

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

1.4 Kullback-Leibler distance

1.4.1 Definition

NB! In what follows,

$$0 \log(\frac{0}{q}) := 0, \text{ if } q \geq 0 \text{ and } p \log(\frac{p}{0}) := \infty \text{ if } p > 0.$$

Def 1.6 *Let P and Q two distributions on $\mathcal{X}$. The* **Kullback-Leibler distance** *(Kullback-Leibler divergence, relative entropy, informational divergence) between probability distributions P and Q is defined as*

$$D(P||Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{1.7}$$

Where $X \sim P$, then

$$D(P||Q) = E\Big(\log \frac{P(X)}{Q(X)}\Big).$$

When $X \sim P$ and $Y \sim Q$, then

$$D(X||Y) := D(P||Q).$$

Def 1.7 *Let, for any $x \in \mathcal{X}$, $P(y|x)$ and $Q(y|x)$ be two (conditional) probability distributions on $\mathcal{Y}$. Let $P(x)$ be a probability distribution on X. The* **conditional Kullback-Leibler distance** *is the K-L distance of $P(y|x)$ and $Q(y|x)$ averaged over P*

$$D(P(y|x)||Q(y|x)) = \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} = \sum_x \sum_y P(x, y) \log \frac{P(y|x)}{Q(y|x)} = E \log \frac{P(Y|X)}{Q(Y|X)},$$

where $P(x, y) := P(y|x)P(x)$ and $(X, Y) \sim P(x, y)$.

Remarks:

- Note that $\log \frac{P(x)}{Q(x)}$ is not always non-negative so that in case of infinite $\mathcal{X}$, we have to show that the sum of the series in (1.7) is defined. Let us do it. Define

$$\mathcal{X}^+ := \Big\{x \in \mathcal{X} : \frac{P(x)}{Q(x)} > 1\Big\}, \quad \mathcal{X}^- := \Big\{x \in \mathcal{X} : \frac{P(x)}{Q(x)} \leq 1\Big\}.$$

  The series over $\mathcal{X}^-$ is absolutely convergent:

$$\sum_{x\in\mathcal{X}^-} |P(x) \log \frac{P(x)}{Q(x)}| = \sum_{x\in\mathcal{X}^-} P(x) \log \frac{Q(x)}{P(x)} \leq \sum_{x\in\mathcal{X}^-} P(x) \frac{Q(x)}{P(x)} \leq 1.$$

  If

$$\sum_{x\in\mathcal{X}^+} P(x) \log \frac{P(x)}{Q(x)} < \infty.$$

  the series (1.7) is convergent, otherwise its sum is ∞.

- As we shall show below, $D(P||Q) \geq 0$ with equality only if $P = Q$. However, in general $D(P||Q) \neq D(Q||P)$. Hence K-L distance is not a metric (true "distance"). Moreover, it does not satisfy triangular inequality (Exercise 7).

K-L distance measures the amount of "average surprise", that a distribution P provides us, when we believe that the distribution is Q. If there is a $x' \in \mathcal{X}$ such that $Q(x') = 0$ (we believe x' never occurs), but $P(x') > 0$ (it still happens sometimes), then

$$P(x') \log\Big(\frac{P(x')}{Q(x')}\Big) = \infty$$

implying that $D(P||Q) = \infty$. This matches with intuition – seeing an impossible event to happen is extremely surprising (a miracle). On the other hand, if there is a letter $x'' \in \mathcal{X}$ such that $Q(x'') > 0$ (we believe it might happen), but $P(x'') = 0$ (it actually never happens), then

$$P(x'') \log\Big(\frac{P(x'')}{Q(x'')}\Big) = 0.$$

also this matches with the intuition – we are not largely surprised if something that might happen actually never does. In this point of view the asymmetry of K-L distance is rather natural.

Example: Let $P = B(1, \frac{1}{2})$, $Q = B(1, q)$. Then

$$\begin{aligned} D(P||Q) =& \frac{1}{2}\log(\frac{1}{2q}) + \frac{1}{2}\log(\frac{1}{2(1-q)}) = -\frac{1}{2}\log(4q(1-q)) \to \infty, \text{ if } q \to 0 \\ D(Q||P) =& q\log(2q) + (1-q)\log(2(1-q)) \to 1 \text{ if } q \to 0. \end{aligned}$$

1.4.2 K-L distance is non-negative: Gibbs inequality and its consquences

Proposition 1.2 (Gibbs inequality) *$D(P||Q) \geq 0$, with equality iff $P = Q$.*

Proof. When $D(P||Q) = \infty$, then inequality trivially holds. Hence consider the situation $D(P||Q) < \infty$ i.e., series (1.7) converges absolutely (when $\mathcal{X}$ infinite).
Let $X \sim P$. Define

$$Y := \frac{Q(X)}{P(X)}$$

and let $g(x) := -\log(x)$. Note that g is strictly convex. We shall apply Jensen's inequality. Let us first convince that all expectations exists

$$E|g(Y)| = \sum_{x\in\mathcal{X}} |-\log \frac{Q(x)}{P(x)}|P(x) = \sum_{x\in\mathcal{X}} |\log \frac{P(x)}{Q(x)}|P(x) < \infty, \quad E|Y| = EY = \sum_{x\in\mathcal{X}} \frac{Q(x)}{P(x)}P(x) = 1.$$

By Jensen's inequality

$$D(P||Q) = E\Big(\log \frac{P(X)}{Q(X)}\Big) = E\Big(-\log \frac{Q(X)}{P(X)}\Big) = Eg(Y) \geq g(EY) = -\log(1) = 0,$$

with $D(P||Q) = 0$ if and only if $Y = 1$ a.s. or $Q(x) = P(x)$ for every $x \in \mathcal{X}_P$. This implies $Q(x) = P(x)$ for every $x \in \mathcal{X}$. ■

Corollary 1.1 (log-sum inequality) *Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ nonnegative numbers so that $\sum a_i < \infty$ and $0 < \sum b_i < \infty$. Then*

$$\sum a_i \log \frac{a_i}{b_i} \geq (\sum a_i) \log \frac{\sum a_i}{\sum b_i}, \tag{1.8}$$

with equality iff $\frac{a_i}{b_i} = c \quad \forall i$.

Proof. Let

$$a'_i = \frac{a_i}{\sum_j a_j}, \quad b'_i = \frac{b_i}{\sum_j b_j}.$$

Hence $(a'_1, a'_2, \ldots)$ and $(b'_1, b'_2, \ldots)$ are probability measures so that from Gibbs inequality, it follows

$$0 \leq \sum a'_i \log \frac{a'_i}{b'_i} = \sum \frac{a_i}{\sum_j a_j} \log \frac{\frac{a_i}{\sum_j a_j}}{\frac{b_i}{\sum_j b_j}} = \frac{1}{\sum_j a_j}\Big[\sum a_i \log \frac{a_i}{b_i} - (\sum a_i) \log \frac{\sum a_j}{\sum b_j}\Big].$$

Since

$$\sum a_i \log \frac{\sum a_j}{\sum b_j} < \infty,$$

the inequality (1.8) follows. We know that $D((a'_1, a'_2, \ldots)||(b'_1, b'_2, \ldots)) = 0$ iff $a'_i = b'_i$. This, however, implies that

$$\frac{a_i}{b_i} = \frac{\sum_j a_j}{\sum_j b_j} =: c, \quad \forall i.$$

■

Remark: Note that log-sum inequality and Gibbs inequality are equivalent.

From Gibbs (or log-sum) inequality, it also follows that for finite $\mathcal{X}$, the distribution with the biggest entropy is uniform. Note that if U is uniform distribution over $\mathcal{X}$, then $H(U) = \log|\mathcal{X}|$.

Corollary 1.2 *Let $|\mathcal{X}| < \infty$. Then, for any distribution P, it holds $H(P) \leq \log|\mathcal{X}|$, with equality iff P is uniform over $|\mathcal{X}|$.*

Proof. Let U be uniform distribution over $\mathcal{X}$, i.e. $U(x) = |\mathcal{X}|^{-1}\ \forall x \in \mathcal{X}$. Then

$$D(P||U) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{U(x)} = \log|\mathcal{X}| - H(P) \geq 0.$$

The equality holds iff $U(x) = P(x)$ for every $x \in \mathcal{X}$, i.e. $P = U$. ■

Pinsker inequality. There are several ways to measure the distance between different probability measures on $\mathcal{X}$. In statistics, a common measure is so-called l_1 or **total variation distance**: for any two probability measures P_1 and P_2 on $\mathcal{X}$:

$$\|P_1 - P_2\| := \sum_{x\in\mathcal{X}} |P_1(x) - P_2(x)|.$$

It is easy to see (Exercise 8)

$$\|P_1 - P_2\| = 2\sup_{B\subseteq\mathcal{X}} |P_1(B) - P_2(B)| = 2|P_1(A) - P_2(A)| \leq 2, \tag{1.9}$$

where

$$A := \{x \in \mathcal{X} : P_1(x) \geq P_2(x)\}.$$

The convergence in total variation, i.e. $\|P_n - P\| \to 0$ implies that for every $B \subset \mathcal{X}$, $P_n(B) \to P(B)$. In particular, for any $x \in \mathcal{X}$, $P_n(x) \to P(x)$. On the other hand, it is possible to show (Sheffe's theorem) that the convergence $P_n(x) \to P(x)$ for every x implies $\|P_n - P\| \to 0$. Thus

$$\|P_n - P\| \to 0 \quad \Leftrightarrow \quad P_n(x) \to P(x), \quad \forall x \in \mathcal{X}.$$

In what follows, the convergence $P_n \to P$ is always meant in total variation. Note that for finite $\mathcal{X}$ this is equivalent to the convergence in usual (Euclidian) distance. Pinsker inequality implies that convergence in K-L distance i.e. $D(P_n||P) \to 0$ or $D(P||P_n) \to 0$ implies $P_n \to P$.

Theorem 1.8 (Pinsker inequality) *For every two probability measures P_1 and P_2 on $\mathcal{X}$, it holds*

$$D(P_1||P_2) \geq \frac{1}{2\ln 2}\|P_1 - P_2\|^2. \tag{1.10}$$

The proof of Pinsker inequality is based on log-sum inequality.

Convexity of K-L distance. Let P_1, P_2, Q_1, Q_2 be the distributions on $\mathcal{X}$. consider the mixtures

$$\lambda P_1 + (1-\lambda)P_2 \quad \text{ja } \lambda Q_1 + (1-\lambda)Q_2.$$

Corollary 1.3

$$D\big(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2\big) \leq \lambda D(P_1||Q_1) + (1-\lambda)D(P_2||Q_2). \tag{1.11}$$

Proof. Fix $x \in \mathcal{X}$. Log-sum inequality:

$$\begin{aligned} &\lambda P_1(x) \log \frac{\lambda P_1(x)}{\lambda Q_1(x)} + (1-\lambda)P_2(x) \log \frac{(1-\lambda)P_2(x)}{(1-\lambda)Q_2(x)} \\ &\geq \Big(\lambda P_1(x) + (1-\lambda)P_2(x)\Big) \log \frac{\lambda P_1(x) + (1-\lambda)P_2(x)}{\lambda Q_1(x) + (1-\lambda)Q_2(x)}. \end{aligned}$$

Sum over $\mathcal{X}$. ■

Take $Q_1 = Q_2 = Q$. Then from (2.2), it follows that the function $P \mapsto D(P||Q)$ is convex. Similarly one gets that $Q \mapsto D(P||Q)$ is convex. When they are finite, then both functions are also strictly convex. Indeed:

$$D(P||Q) = \sum P(x) \log P(x) - \sum P(x) \log Q(x) = -\sum P(x) \log Q(x) - H(P). \tag{1.12}$$

The function $P \mapsto \sum P(x) \log Q(x)$ is linear, $P \mapsto H(P)$ strictly concave. The difference is, thus, strictly convex (when finite). From (1.12) also the strict convexity of $Q \mapsto D(P||Q)$ follows.

Continuity of K-L distance for finite $\mathcal{X}$. In finite-dimensional space, a finite convex function is continuous. Hence if $|\mathcal{X}| < \infty$ and the function $P \mapsto D(P||Q)$ is finite (in an open set), then it is continuous (in that set). The same holds for the function $Q \mapsto D(P||Q)$.

Example: The finiteness is important. Let $\mathcal{X} = \{a, b\}$, and let for every n the measure P_n be such that $P_n(a) = p_n$, where $p_n > 0$ and $p_n \to 0$. Let $P(a) = 0$. Clearly, $P_n \to P$, but for every n

$$\infty = D(P_n||P) \not\to D(P||P) = 0.$$

Conditioning increases K-L distance. Let, for every $x \in \mathcal{X}$, $P_1(y|x)$ and $P_2(y|x)$ be conditional probability distributions, and let $P(x)$ a probability measure on $\mathcal{X}$. Let

$$P_i(y) := \sum_x P_i(y|x)P(x), \quad \text{where } i = 1, 2.$$

Then

$$D(P_1(y|x)||P_2(y|x)) \geq D(P_1||P_2). \tag{1.13}$$

Proof of (1.13) is Exercise 16.

1.5 Mutual information

Let (X, Y) be random vector with distribution $P(x, y)$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$. As usually, let $P(x)$ and $P(y)$ be the marginal distributions, i.e. $P(x)$ is distribution of X and $P(y)$ is distribution of Y.

Def 1.9 *The* **mutual information** $I(X;Y)$ *of* X *and* Y *is K-L distance between the joint distribution* $P(x, y)$ *and the product distribution* $P(x)P(y)$

$$I(X;Y) := \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} = D\big(P(x,y)||P(x)P(y)\big) = E\Big(\log \frac{P(X,Y)}{P(X)P(Y)}\Big).$$

Hence $I(X;Y)$ is K-L distance between (X, Y) and a vector (X', Y'), where X' and Y' are distributed as X and Y, but unlike X and Y, the random variables X' and Y' are independent.

Properties:

- $I(X;Y)$ depends on joint distribution $P(x, y)$.
- $0 \le I(X;Y)$.
- mutual information is symmetric $I(X;Y) = I(Y;X)$.
- $I(X;Y) = 0$ iff X, Y are independent.
- The following relation is important:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{1.14}$$

For the proof, note

$$\begin{aligned} I(X;Y) &= E \log \frac{P(X,Y)}{P(X)P(Y)} = E \log \frac{P(X|Y)P(Y)}{P(X)P(Y)} = E \log \frac{P(X|Y)}{P(X)} \\ &= E \log P(X|Y) - E \log P(X) = H(X) - H(X|Y). \end{aligned}$$

By symmetry, the roles of X and Y can be changed.

Hence the mutual information is the reduction of randomness of X due to the knowledge of Y. When X and Y are independent, then $H(X|Y) = H(X)$, and $I(X;Y) = 0$. On the other hand, when $X = f(Y)$, then $H(X|Y) = 0$ so that $I(X;Y) = H(X)$. In particular

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

Therefore, sometimes entropy is referred to as *self-information*.

- Recall chain rule: $H(X|Y) = H(X,Y) - H(Y)$. Hence

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{1.15}$$

- Conditioning reduces entropy

$$H(X|Y) \leq H(X),$$

because $H(X) - H(X|Y) = I(X;Y) \geq 0$.

Recall $H(X|Y) = \sum_y H(X|Y=y)P(y)$. The fact that sum is smaller than $H(X)$ does not imply that $H(X|Y=y) \leq H(X)$ for every y. As the following little counterexample shows, it need not to be case (check!)

$\mathcal{Y}\backslash\mathcal{X}$	a	b
u	0	$\frac{3}{4}$
v	$\frac{1}{8}$	$\frac{1}{8}$

- For any random vector $(X_1, \ldots, X_n)$, it holds

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

with equality iff all components are independent. For the proof use chain rule

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1})$$

and apply the fact that conditioning reduces entropy.

Conditional mutual information. Let X, Y, Z be random variables, let $\mathcal{Z}$ be the support of Z.

Def 1.10 **The conditional mutual information of X, Y given Z** *is*

$$\begin{aligned}
I(X;Y|Z) :=& H(X|Z) - H(X|Y,Z) = E \log \frac{P(X|Y,Z)}{P(X|Z)} \\
=& E \log \frac{P(X|Y,Z)P(Y|Z)}{P(X|Z)P(Y|Z)} = E \log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)} \\
=& \sum_{x,y,z} P(x,y,z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)} \\
=& \sum_{z} P(z) \sum_{y,x} P(x,y|z) \log \frac{P(x,y|z)}{P(x|z)P(y|z)} \\
=& \sum_{z} D\big(P(x,y|z)||P(x|z)P(y|z)\big)P(z).
\end{aligned}$$

Properties:

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

  with equality iff X and Y are conditionally independent:

  $$P(x,y|z) = P(x|z)P(y|z), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}. \tag{1.16}$$

  For proof note that $I(X;Y|Z) = 0$ iff for every $z \in \mathcal{Z}$, it holds

  $$D\Big(P(x,y|z)||P(x|z)P(y|z)\Big) = 0.$$

  This means conditional independence.

- The proof of following equalities is Exercise 18

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

  In addition, the following equality holds

  $$I(X;Y|Z) = H(X;Z) + H(Y;Z) - H(X,Y,Z) - H(Z). \tag{1.17}$$

- Chain rule for mutual information

  $$I(X_1,\ldots,X_n;Y) = I(X_1;Y)+I(X_2;Y|X_1)+I(X_3;Y|X_1,X_2)+\cdots+I(X_n;Y|X_1,\ldots,X_{n-1}).$$

  For proof use chain rule for entropy and conditional entropy:

  $$\begin{aligned} I(X_1,\ldots,X_n;Y) =& H(X_1,\ldots,X_n) - H(X_1,\ldots,X_n|Y) \\ =& H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1,\ldots,X_{n-1}) \\ & - H(X_1|Y) - H(X_2|X_1,Y) - \cdots - H(X_n|X_1,\ldots,X_{n-1},Y). \end{aligned}$$

- Chain rule for conditional mutual information:

  $$I(X_1,\ldots,X_n;Y|Z) = I(X_1;Y|Z)+I(X_2;Y|X_1,Z)+\cdots+I(X_n;Y|X_1,\ldots,X_{n-1},Z).$$

  Proof is similar.

1.6 Fano's inequality

Let X be a (unknown) random variable and $\hat{X}$ a related random variable – an estimate of X. Let

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

be the probability of mistake made by estimation. If $P_e = 0$, then $X = \hat{X}$ a.s. so that $H(X|\hat{X}) = 0$. Therefore, it is natural to expect that when P_e is small, then $H(X|\hat{X})$ should also be small. Fano's inequality quantifies that idea.

Theorem 1.11 (Fano's inequality) *Let X and $\hat{X}$ be random variables on $\mathcal{X}$. Then*

$$H(X|\hat{X}) \leq h(P_e) + P_e \log(|\mathcal{X}| - 1), \tag{1.18}$$

where h is binary entropy function.

Proof. Let

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

Hence

$$E = I_{\{\hat{X} \neq X\}}, \quad E \sim B(1, P_e).$$

Chain rule for entropy:

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(X|\hat{X}), \tag{1.19}$$

because $H(E|X, \hat{X}) = 0$ (why?)
On the other hand,

$$H(E, X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \leq H(E) + H(X|E, \hat{X}) = h(P_e) + H(X|E, \hat{X}).$$

Note

$$\begin{aligned} H(X|E, \hat{X}) = &\sum_{x \in \mathcal{X}} \mathbf{P}(\hat{X} = x, E = 1) H(X|\hat{X} = x, E = 1) \\ &+ \sum_{x \in \mathcal{X}} \mathbf{P}(\hat{X} = x, E = 0) H(X|\hat{X} = x, E = 0). \end{aligned}$$

Given $\hat{X} = x$ and $E = 0$, we have $X = x$ and then $H(X|\hat{X} = x, E = 0) = 0$ or

$$H(X|E, \hat{X}) = \sum_{x \in \mathcal{X}} \mathbf{P}(\hat{X} = x, E = 1) H(X|\hat{X} = x, E = 1).$$

If $E = 1$ and $\hat{X} = x$, then $X \in \mathcal{X} \backslash x$, so that $H(X|\hat{X} = x, E = 1) \leq \log(|\mathcal{X}| - 1)$. To summarize:

$$H(X|E, \hat{X}) \leq P_e \log(|\mathcal{X}| - 1).$$

Form (1.19) we obtain

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

■

Corollary 1.4

$$H(X|\hat{X}) \leq 1 + P_e \log |\mathcal{X}|, \quad ehk \quad P_e \geq \frac{H(X|\hat{X}) - 1}{\log |\mathcal{X}|}.$$

If $|\mathcal{X}| < \infty$, then Fano's inequality implies: if $P_e \to 0$, then $H(X|\hat{X}) \to 0$. When $|\mathcal{X}| = \infty$, then Fano's inequality is trivial and such an implication might not exists.

Example: Let $Z \sim B(1, p)$ and let Y be such a random variable that $Y > 0$ and $H(Y) = \infty$. Define X as follows

$$X = \begin{cases} 0 & \text{if } Z = 0, \\ Y & \text{if } Z = 1. \end{cases}$$

Let $\hat{X} = 0$ a.s.. Then $P_e = \mathbf{P}(X > 0) = \mathbf{P}(X = Y) = \mathbf{P}(Z = 1) = p$. But

$$H(X|\hat{X}) = H(X) \geq H(X|Z) = pH(Y) = \infty.$$

Then for every $p > 0$, clearly $H(X|\hat{X}) = \infty$ and therefore $H(X|\hat{X}) \not\to 0$, when $P_e \searrow 0$.

When Fano's inequality is an equality? Inspecting the proof reveals that equality holds iff for every $x \in \mathcal{X}$,

$$H(X|\hat{X} = x, E = 1) = \log(|\mathcal{X}| - 1) \tag{1.20}$$

and

$$H(E|\hat{X}) = H(E). \tag{1.21}$$

The equality (1.20) means that the conditional distribution of X given $X \neq \hat{X} = x$ is uniform over all remaining alphabet $\mathcal{X} \backslash x$. That, in turn, means that to every $x_i \in \mathcal{X}$ corresponds p_i so that

$$\mathbf{P}(\hat{X} = x_i, X = x_j) = p_i, \quad \forall j \neq i.$$

In other words, the joint distribution of $(\hat{X}, X)$

$\hat{X} \backslash X$	x_1	x_2	$\cdots$	x_n
x_1	$\mathbf{P}(\hat{X} = x_1, X = x_1)$	$\mathbf{P}(\hat{X} = x_1, X = x_2)$	$\cdots$	$\mathbf{P}(\hat{X} = x_1, X = x_n)$
x_2	$\mathbf{P}(\hat{X} = x_2, X = x_1)$	$\mathbf{P}(\hat{X} = x_2, X = x_2)$	$\cdots$	$\mathbf{P}(\hat{X} = x_2, X = x_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_n	$\mathbf{P}(\hat{X} = x_n, X = x_1)$	$\cdots$	$\cdots$	$\mathbf{P}(\hat{X} = x_n, X = x_n)$

is such that in every row, all elements outside the main diagonal are equal (to a constant depending on the row). The relation (1.21) means that for every $x \in \mathcal{X}$, it holds that $P(X = x|\hat{X} = x) = 1 - P_e$ (in every row the probability in main diagonal divided by the

sum of the whole row equals to $1 - P_e$. A joint distribution satisfying both requirements (1.20) and (1.21) is, for example,

$\hat{X}\backslash\mathcal{X}$	a	b	c
a	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
b	$\frac{1}{25}$	$\frac{3}{25}$	$\frac{1}{25}$
c	$\frac{3}{50}$	$\frac{3}{50}$	$\frac{9}{50}$

with this distribution, $P_e = \frac{2}{5}$, $\log(|\mathcal{X}| - 1) = 1$ so that

$$P_e \log(|\mathcal{X}| - 1) + h(P_e) = \frac{2}{5} + \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log\frac{5}{2} = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log 5.$$

On the other hand

$$H(X|\hat{X} = a) = H(X|\hat{X} = b) = H(X|\hat{X} = c) = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log 5,$$

implying that

$$H(X|\hat{X}) = \frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log 5.$$

Therefore, Fano's inequality is an equality.

1.7 Data processing inequality

1.7.1 Finite Markovi chain

Def 1.12 *The random variables $X_1, \ldots, X_n$ with supports $\mathcal{X}_1, \ldots, \mathcal{X}_n$ form a* **Markov chain** *when for every $x_i \in \mathcal{X}_i$ and $m = 2, \ldots, n-1$*

$$\mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m, \ldots, X_1 = x_1) = \mathbf{P}(X_{m+1} = x_{m+1}|X_m = x_m). \tag{1.22}$$

Then $X_1, \ldots, X_n$ is Markov chain iff for every $x_1, \ldots, x_n$ such that $x_i \in \mathcal{X}_i$

$$P(x_1, \ldots, x_n) = P(x_1, x_2)P(x_3|x_2)\cdots P(x_n|x_{n-1}).$$

The fact that $X_1, \ldots, X_n$ form a Markov chain is in information theory denoted as

$$X_1 \to X_2 \to \cdots \to X_n.$$

Thus $X \to Y \to Z$ iff

$$P(x, y, z) = P(x)P(y|x)P(z|y).$$

We shall now list (without proofs) some elementary properties of Markov chains.

Properties:

- If $X_1 \to X_2 \to \cdots \to X_n$, then $X_n \to X_{n-1} \to \cdots \to X_1$ (reversed MC is also a MC).
- Every sub-chain Markov chain is a Markov chain: if $X_1 \to X_2 \to \cdots \to X_n$, then $X_{n_1} \to X_{n_2} \to \cdots \to X_{n_k}$.
- If $X_1 \to X_2 \to \cdots \to X_n$, then for every $m < n$ and $x_i \in \mathcal{X}_i$

$$P(x_n, \ldots, x_{m+1}|x_m, \ldots, x_1) = P(x_n, \ldots, x_{m+1}|x_m). \tag{1.23}$$

- $X_1 \to \cdots \to X_n$ iff for every $m = 2, \ldots, n-1$ the random variables $X_1, \ldots, X_{m-1}$ and $X_{m+1}, \ldots, X_n$ are conditionally independent given X_m: for every $x_m \in \mathcal{X}_m$,

$$P(x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n|x_m) = P(x_1, \ldots, x_{m-1}|x_m)P(x_{m+1}, \ldots, x_n|x_m). \tag{1.24}$$

1.7.2 Data processing inequality

Lemma 1.3 (Data processing inequality) *When $X \to Y \to Z$, then*

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

with equality iff $X \to Z \to Y$.

Proof. From $X \to Y \to Z$ it follows that X and Z are conditionally independent given Y. This implies $I(X;Z|Y) = 0$ and from the chain rule for entropy, it follows

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y) = I(X;Y). \tag{1.25}$$

Since $I(X;Y|Z) \geq 0$,we obtain $I(X;Z) \leq I(X;Y)$ and the equality holds iff $I(X;Y|Z) = 0$ or the random variables X and Y are conditionally independent given Z. That means $X \to Z \to Y$. ■

Let X be an unknown random variable we are interested in. Instead of X, we know Y (data) giving us $I(X;Y)$ bits of information. Would it be possible to process the data so that the amount of information about X increases? The data are possible to process deterministically applying a deterministic function g, obtaining $g(Y)$. Hence we have Markov chain $X \to Y \to g(Y)$ and from data processing inequality $I(X;Y) \geq I(X;g(Y))$ it follows that $g(Y)$ does not give more information about X as Y. Another possibility is to process Y by applying additional randomness independent of X. Since this additional randomness is independent of X, then $X \to Y \to Z$ is still Markov chain and from data processing inequality $I(X;Y) \geq I(X;Z)$. Hence, the data processing inequality postulates well-known fact: it is not possible to increase information by processing the data.

Corollary 1.5 *When $X \to Y \to Z$, then*

$$H(X|Z) \geq H(X|Y).$$

Proof. Exercise 23. ■

Corollary 1.6 *When $X \to Y \to Z$, then*

$$I(X;Z) \leq I(Y;Z), \quad I(X;Y|Z) \leq I(X;Y).$$

Proof. Exercise 23. ■

1.7.3 Sufficient statistics

Let $\{P_\theta\}$ be a family of probability distributions – *model*. Let X be a random sample from the distribution P_θ. Recall that n-elemental random sample can always be considered as a random variable taking values in $\mathcal{X}^n$. Clearly the sample depends on chosen distribution P_θ or, equivalently, on its index — *parameter* – θ. Let $T(X)$ be any statistic (function of the sample) giving an estimate to unknown parameter θ. Let us consider the Bayesian approach, where θ is a random variable with (prior) distribution π. Then $\theta \to X \to T(X)$ is Markov chain and from data processing inequality

$$I(\theta; T(X)) \leq I(\theta; X).$$

When the inequality above is an equality, then $T(X)$ gives as much information about θ as X and we know that the equality implies $\theta \to T(X) \to X$. By definition of Markov chain, then for every sample $x \in \mathcal{X}^n$

$$\mathbf{P}(X = x|T(X) = t, \theta) = \mathbf{P}(X = x|T(X) = t)$$

or given the value of $T(X)$, the distribution of sample is independent of θ. In statistics, a statistic $T(X)$ having such a property is called **sufficient**.

Corollary 1.7 *A statistic T is sufficient iff for every distribution π of θ the following equality holds true*

$$I(\theta; T(X)) = I(\theta; X).$$

Example: Let $\{P_\theta\}$ the family of all Bernoulli distributions. A statistic $T(X) = \sum_{i=1}^n X_i$ is sufficient, because

$$\mathbf{P}(X_1 = x_1, \ldots, X_i = x_i|T(X) = t, \theta) = \begin{cases} 0 & \text{if } \sum_i x_i \neq t, \\ \frac{1}{\binom{n}{t}} & \text{if } \sum_i x_i = t. \end{cases}$$

Indeed: if $\sum_i x_i = t$, then

$$\begin{aligned} \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n|T(X) = t, \theta) &= \frac{\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, T(X) = t, \theta)}{\mathbf{P}(T(X) = t, \theta)} \\ &= \frac{\theta^t(1-\theta)^{n-t}\pi(\theta)}{\sum_{x_1,\ldots,x_n:\sum_i x_i = t} \theta^t(1-\theta)^{n-t}\pi(\theta)} = \frac{1}{\binom{n}{t}}, \end{aligned}$$

because given sum t (the number of ones) there are exactly $\binom{n}{t}$ possibilities for different samples.

1.8 Entropy rate of a stochastic process

Let us consider a *stochastic process* $\{X_n\}_{n=1}^{\infty}$.

Def 1.13 **The entropy rate of a stochastic process** $\{X_n\}_{n=1}^{\infty}$ *is*

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

provided the limit exists.

Examples:

- let $\{X_n\}_{n=1}^{\infty}$ i.i.d. random variables from the distribution P, i.e. $X_i \sim P$. then

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

  Thus, in i.i.d. case the entropy rate of the process equals to the entropy of X_1.

- Let $\{X_n\}_{n=1}^{\infty}$ be independent random variables

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

  The limit need not always exists so that the entropy rate is not always defined for that process.

- Let $X_1, X_2, \ldots$ i.i.d. random variables $X_i \sim P$. Let $\mathcal{X} = \mathbb{Z}$. Consider *random walk* $\{S_n\}_{n=0}^{\infty}$, s.t.

$$S_0 = 0,\ S_1 = X_1,\ S_2 = X_1 + X_2, \ldots, S_n = X_1 + \cdots + X_n.$$

  The entropy rate of random walk is $H_S = H(P)$. The proof of that is Exercise 32.

The limit H'_X. Consider the limit (when exists)

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

We shall now show that for a large class of stochastic processes, called stationary processes, the limit H'_X always exists.

Def 1.14 *A stochastic process* $\{X_n\}_{n=1}^{\infty}$ *is* **stationary**, *if for every* $n \geq 1$ *and every* $k \geq 1$ *the random vectors*

$$(X_1, \ldots, X_n) \text{ and } (X_{k+1}, \ldots, X_{k+n})$$

have the same distributions.

Hence, when $\{X_n\}_{n=1}^{\infty}$ is stationary, then all random variables $X_1, X_2, \ldots$ have the same distributions, all two-dimensional random vectors (X_1, X_2), (X_2, X_3), ... have the same distribution, the vectors (X_1, X_2, X_3), (X_2, X_3, X_4), ... have the same distribution etc.

Proposition 1.3 *When $\{X_n\}_{n=1}^{\infty}$ is stationary, then the limit H'_X always exists.*

Proof. Since $\{X_n\}_{n=1}^{\infty}$ is stationary, then for every n the random vectors $(X_1, \ldots, X_n)$ and $(X_2, \ldots, X_{n+1})$ have the same distributions. Hence, for every n

$$H(X_n|X_1, \ldots, X_{n-1}) = H(X_{n+1}|X_2, \ldots, X_n).$$

Therefore

$$H(X_{n+1}|X_1, \ldots, X_n) \leq H(X_{n+1}|X_2, \ldots, X_n) = H(X_n|X_1, \ldots, X_{n-1}),$$

so that the sequence $\{H(X_n|X_1, \ldots, X_{n-1})\}$ is non-negative and non-increasing. Such a sequence has always a limit. ■

Next, we show that for a stationary process the entropy rate is always defined and equals to H'_X. We need Cesaro's lemma

Lemma 1.4 (Cesaro) *Let $\{a_n\}$ non-negative real numbers with $a_1 > 0$ and $\sum_n a_n = \infty$. Denote $b_n := \sum_{i=1}^{n} a_i$. Let $x_n \to x$ be arbitrary convergent sequence. Then*

$$\frac{1}{b_n} \sum_{i=1}^{n} a_i x_i \to x, \quad \text{when} \quad n \to \infty.$$

In a special case $a_n = 1$, we obtain

$$\frac{x_1 + \ldots + x_n}{n} \to x.$$

Theorem 1.15 *When $\{X_n\}_{n=1}^{\infty}$ is a stationary process, then H_X always exists and $H'_X = H_X$.*

Proof. From the chain rule for entropy:

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

Use $H(X_k|X_1, \ldots, X_{k-1}) \to H'_X$, together with Cesaro lemma to obtain

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} H(X_k|X_1, \ldots, X_{k-1}) = H'_X.$$

■

Hence, every stationary process has an entropy rate that equals to H'_X. It might be 0 even if X is still random (can you find an example of such process?). On the other hand, also a non-stationary processes might have an entropy rate (which of the examples above was non-stationary).

1.8.1 Entropy rate of Markov chain

Determining a entropy rate of a stochastic process is, in general, not an easy task. In this sub-subsection, we find the entropy rate of stationary Markov chain.

Let $\{X_n\}_{n=1}^{\infty}$ be a random process where all random variables X_i are taking the values on discrete alphabet $\mathcal{X}$.

Def 1.16 *The random process* $\{X_n\}_{n=1}^{\infty}$ *is* **Markov chain**, *if for every* $m \geq 1$ *and* $x_1, \ldots, x_m \in \mathcal{X}$ *such that* $\mathbf{P}(X_m = x_m, \ldots, X_1 = x_1) > 0$, *(1.22) holds, i.e.*

$$\mathbf{P}(X_{m+1} = x_{m+1} | X_m = x_m, \ldots, X_1 = x_1) = \mathbf{P}(X_{m+1} = x_{m+1} | X_m = x_m). \tag{1.26}$$

In therminology of Markov chains, the elements of $\mathcal{X}$ are called *states*, and the chain is called *time homogenous*, if the the right hand side of equality (1.26) is independent of m. In this case, for every m and $x_i, x_j \in \mathcal{X}$

$$\mathbf{P}(X_{m+1} = x_j | X_m = x_i) = P(X_2 = x_j | X_1 = x_i) =: P_{ij}.$$

The matrix $P = (P_{ij})$ is *transition matrix* of time-homogenous MC $\{X_n\}$. Let $\pi(i) = \pi(x_i)$ – *initial distribution* – be the distribution of X_1. Then

$$\mathbf{P}(X_2 = x_j) = \sum_{x_i \in \mathcal{X}} \mathbf{P}(X_2 = x_j | X_1 = x_i) \mathbf{P}(X_1 = x_i) = \sum_i P_{ij} \pi(i)$$

so that the distribution of X_2 is $\pi^T P$. Similarly, the distribution of X_k is $\pi^T P^k$. Now, it is not hard to see that the distribution of any finite vector $(X_k, \ldots, X_{k+l})$ is fully determined by transition matrix P and initial distribution π. Markov chain $\{X_n\}$ is stationary iff π is such that $\pi^T P = \pi$ or $\pi(j) = \sum_i \pi(i) P_{ij} \ \forall \ j$. Such initial distribution (when exists) is called *stationary initial distribution*. Whether it exists and is unique, depends on the transition matrix P.

Example: Let $|\mathcal{X}| = 2$ and let the transition matrix be

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

Unique stationary initial distribution corresponding to that transition matrix is

$$(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}).$$

Theorem 1.17 *Let* $\{X_n\}$ *be stationary time-homogenous Markov chain with transition matrix* (P_{ij}) *and (stationary) initial distribution* π. *Then*

$$H_X = H(X_2 | X_1) = -\sum_i \pi(i) \sum_j P_{ij} \log P_{ij}.$$

Proof. From (1.26), we obtain that for every n $H(X_n|X_{n-1},\dots,X_1) = H(X_n|X_{n-1})$. Since chain is stationary, we get $H(X_n|X_{n-1}) = H(X_2|X_1)$ and by Theorem 1.15,

$$H_X = H'_X = \lim_{n\to\infty} H(X_n|X_{n-1},\dots,X_1) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1).$$

The equation

$$H(X_2|X_1) = -\sum_i \pi(i) \sum_j P_{ij} \log P_{ij}$$

is Exercise 31. ■

1.9 Exercises

1. Let us toss until the firs head. Let X be the number tosses needed. Find $H(X)$, if the probability of head is p.

2. Prove *grouping property*

$$H(p_1, p_2, p_3, \dots) = H(p_1 + p_2, p_3, \dots) + (p_1 + p_2)H(\frac{p_1}{(p_1 + p_2)}, \frac{p_2}{(p_1 + p_2)})$$

   and deduce (1.2).

3. Let $g : \mathcal{X} \to \mathcal{X}$ a function. Prove that

$$H(g(X)) \le H(X), \quad H(g(X)|Y) \le H(X|Y).$$

4. Find P such that $H(P) = \infty$.

5. let X_1 and X_2 random variables with disjoint supports. Let X have mixture distribution, i.e.

$$X = \begin{cases} X_1 & \text{if } Z = 1, \\ X_2 & \text{if } Z = 0, \end{cases}$$

   where $Z \sim B(1, p)$. Find $H(X)$. Show that

$$2^{H(X)} \le 2^{H(X_1)} + 2^{H(X_2)}.$$

6. Let $X \sim P$. Show that

$$\mathbf{P}\big(P(X) \le d\big)(\log \frac{1}{d}) \le H(X).$$

7. Find distributions P, Q and R show that

$$D(P\|Q) > D(P\|R) + D(R\|Q).$$

8. Prove (1.9).

9. Let
$$P = (p_1, p_2, \ldots, p_m, 0, 0, \ldots)$$
and for every n,
$$P_n = \big((1-\frac{1}{n})p_1, \ldots, (1-\frac{1}{n})p_m, \underbrace{\frac{1}{nM_n}, \ldots \frac{1}{nM_n}}_{M_n}, 0, \ldots), \tag{1.27}$$
where
$$M_n = \lceil 2^{nc} \rceil, \quad c > 0.$$
show that
$$H(P_n) = (1-\frac{1}{n})H(P) + \frac{1}{n}\log_2 M_n + h(\frac{1}{n}) \to H(P) + c.$$

10. Let $\mathcal{X}$ infinite. Define
$$P_n = (1 - \frac{\alpha}{\log n}, \underbrace{\frac{\alpha}{n \log n}, \ldots, \frac{\alpha}{n \log n}}_{n}, 0, \cdots),$$
where $\alpha > 0$. Show that $P_n \to P$, where $P = (1, 0, \ldots)$, but $H(P_n) \to \alpha$. Let
$$Q = (q_1, q_2, q_3, \ldots),$$
where $q_i = (1-q)q^{i-1}$. Show that $D(P||Q) < \infty$, but
$$D(P_n||Q) \to \infty.$$

11. Let $X = (X_1, \ldots, X_n)$ random vector, where X_i has Bernoulli distribution for every i. The random variables X_i are neither independent nor identically distributed. Let $R = (R_1, \ldots, R_n)$ be the run lengths of X. For example, if $X = (1, 0, 0, 0, 1, 1, 0)$, then $R = (1, 3, 2, 1)$. Show that
$$0 \le H(X) - H(R) \le \min_i H(X_i).$$

12. Let X, Y be random variables, let $Z = X + Y$.
    - Show that $H(Z|X) = H(Y|X)$.
    - Show that when X and Y are independent, then $H(X) \le H(Z)$ and $H(Y) \le H(Z)$.
    - Find X and Y such that $H(X) > H(Z)$ and $H(Y) > H(Z)$.
    - When $H(Z) = H(X) + H(Y)$?

13. Let
$$\rho(X,Y) = H(X|Y) + H(Y|X).$$
Show that ρ is semi-metric. When $\rho(X,Y) = 0$?
Show that
$$\rho(X,Y) = H(X)+H(Y)-2I(X;Y) = H(X,Y)-I(X;Y) = 2H(X,Y)-H(X)-H(Y).$$

14. Prove that for every $n \geq 2$
$$H(X_1,\ldots,X_n) \geq \sum_{i=1}^{n} H(X_i|X_j, j \neq i).$$
Show that
$$\frac{1}{2}[H(X_1,X_2) + H(X_3,X_2) + H(X_1,X_3)] \geq H(X_1,X_2,X_3).$$

15. Let X, Y, Z be random variables, with Y and Z being independent. Show that
$$D(X||Y|Z) = -H(X|Z) + D(X||Y) + H(X) \leq H(Z) + D(X||Y).$$

16. Using log-sum inequality prove (1.13).

17. (a) Let X_1 and X_2 have the same distribution. Let
$$\rho(X_1,X_2) := 1 - \frac{H(X_2|X_1)}{H(X_1)}. \tag{1.28}$$
Prove that ρ is symmetric, $\rho \in [0,1]$. When $\rho = 0$? When $\rho = 1$?

(b) Let (X,Y) have the following joint distribution, where $\epsilon \in (0, \frac{1}{4}]$:

$Y\backslash X$	$-n$	-1	1	n
n	0	0	0	ϵ
1	0	$\frac{1}{4}-\epsilon$	$\frac{1}{4}$	0
-1	0	$\frac{1}{4}$	$\frac{1}{4}-\epsilon$	0
$-n$	ϵ	0	0	0

Find $I(X;Y)$ and ρ (like in (1.28)). Find $\mathrm{cov}(X,Y)$ and the correlation coefficient of X and Y. Note that when $n \to \infty$, then the limit of correlation coefficient is 1 for every $\epsilon > 0$.

(c) Let (X,Y) have the following joint distribution

$Y\backslash X$	$-n$	-1	1	n
n	0	0	$\frac{1}{4}$	0
1	$\frac{1}{4}$	0	0	0
-1	0	0	0	$\frac{1}{4}$
$-n$	0	$\frac{1}{4}$	0	0

Find $I(X;Y)$ and ρ (like in (1.28)). Find $\mathrm{cov}(X,Y)$ and the correlation coefficient of X and Y.

18. Prove

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

19. Prove

$$\begin{aligned} H(X,Y|Z) &\geq H(X|Z) \\ I(X,Y;Z) &\geq I(X;Z) \\ H(X,Y,Z) - H(X,Y) &\leq H(X,Z) - H(X) \\ I(X;Y|Z) &\geq I(Y;Z|X) - I(Y;Z) + I(X;Y). \end{aligned}$$

When the inequalities are equalities?

20. find X, Y, Z such that

$$\begin{aligned} I(X;Y|Z) > I(X;Y) = 0 \\ 0 = I(X;Y|Z) < I(X;Y). \end{aligned}$$

21. Prove that

$$H(X|g(Y)) \geq H(X|Y).$$

find (X, Y) such that X and Y are depending, g is not one-to-one, but the inequality is an equality.

22. Let $X = (X_1, \ldots, X_n)$ be a random vector with binary (0 or 1 valued) components having the following distribution:

$$P(x_1, \ldots, x_n) = \begin{cases} 2^{-(n-1)} & \text{when } \sum_i x_i \text{ is even;} \\ 0, & \text{when } \sum_i x_i \text{ is odd.} \end{cases}$$

Find the distribution of X_i. Find the distribution of (X_i, X_{i+1}). Find

$$I(X_1;X_2), I(X_2;X_3|X_1), I(X_4;X_3|X_1,X_2), \ldots, I(X_n;X_{n-1}|X_1,X_2,\ldots,X_{n-2}).$$

23. Prove that if $X \to Y \to Z$, then $H(X|Z) \geq H(X|Y)$, $I(X;Z) \leq I(Y;Z)$ and $I(X;Y|Z) \leq I(X;Y)$.

24. Let $\{P_\theta\}$ be a set of Bernoulli distributions, $\theta \in \Theta$, where $\Theta$is discrete set, π is a prior distribution of θ. Let X be a random sample and $T(X) = \sum_{i=1}^n X_i$. Find $H(\theta|T(X))$ and $H(\theta|X)$. Show that data processing inequality is an equality.

25. Let $X_1 \to X_2 \to X_3 \to X_4$. Prove

$$I(X_1; X_4) \leq I(X_2; X_3).$$

26. let $X_1 \to X_2 \to \cdots \to X_n$. Find $I(X_1; X_2, X_3, \ldots, X_n)$.

27. Let $X_1 \to X_2 \to X_3$ be Markov chain, where $|\mathcal{X}_1| = n$, $|\mathcal{X}_2| = k$, $|\mathcal{X}_3| = m$, $k < n$ and $k < m$. Show that "bottleneck" decreases mutual information between X_1 and X_3 i.e. $I(X_1; X_3) \leq \log k$. Show that when $k = 1$, then X_1 and X_3 are independent.

28. Let $|\mathcal{X}| = m$ and let X be a random variable taking values on $\mathcal{X}$. Find a non-random estimate $\hat{X}$ to X with smallest error probability. Let $P_e = \mathbf{P}(X \neq \hat{X})$. find X such that Fano's inequality is an equality

$$H(X) = P_e \log(|\mathcal{X}| - 1) + h(P_e)?$$

29. Let P be a probability distribution with support $\mathcal{X}_P = \{1, 2, \ldots\}$. Let μ be the mean of P. Prove that

$$H(P) \leq \mu \log \mu + (1 - \mu) \log(\mu - 1),$$

with equality iff P has geometric distribution. Hence, amongst such distributions, the geometric distribution has the biggest entropy.

30. Let $\{X_n\}_{n=1}^{\infty}$ be a stationary random process. Prove

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

31. Prove that for stationary MC,

$$H(X_2|X_1) = -\sum_i \pi(i) \sum_j P_{ij} \log P_{ij}.$$

32. Let $X_1, X_2, \ldots$ be i.i.d. random variables $X_i \sim P$. Consider random walk $\{S_n\}_{n=0}^{\infty}$, s.t.

$$S_0 = 0,\ S_1 = X_1,\ S_2 = X_1 + X_2, \ldots, S_n = X_1 + \cdots + X_n.$$

Prove that the entropy rate of random walk is $H_S = H(P)$.

33. A dog walks on the integers: at time 0 is it on position 0. Then it start to move, with probability 0.5 to left and with the same probability to right. Then it continues moving in the same direction, possibly reversing direction with probability 0.1. A typical walk might look like

$$(X_0, X_1, \ldots) = (0, -1, -2, -3, -4, -3, -2, -1, 0, 1, 2, 3, \ldots).$$

Find H_X.

34. Consider random walk on ring $(0, 1, \ldots, l)$, i.e. l is followed by 0. Let

$$S_n = \sum_{i=1}^{n} X_i,$$

where X_1 has uniform distribution on $(0, 1, \ldots, l)$ and $X_2, X_3, \ldots$ are i.i.d. random variables $P(X_2 = 1) = P(X_2 = 2) = 0.5$. Find H_S.

2 Zero-error data compression

2.1 Codes

In this section, we suppose that besides our original alphabet $\mathcal{X}$, we have another finite *coding alphabet* $\mathcal{D}$. In what follows, $|\mathcal{D}| =: D$ so that alphabet $\mathcal{D}$ will be referred to as D-ary alphabet and without loss of generality we take

$$\mathcal{D} = \{0, \ldots, D-1\}.$$

In case $D = 2$, thus, we speak about binary alphabet $\{0, 1\}$ etc. The alphabet $\mathcal{D}$ is used in data transmission. Typically $D < |\mathcal{X}|$, hence to transmit a letter x it should be represented as a finite string of letters from $\mathcal{D}$ - a *codeword*.

In what follows, let $\mathcal{D}^*$ be the set of all finite length strings (codewords) from $\mathcal{D}$. Formally, thus

$$\mathcal{D}^* := \cup_{n=1}^{\infty} \mathcal{D}^n, \quad \mathcal{X}^* := \cup_{n=1}^{\infty} \mathcal{X}^n.$$

Def 2.1 *A* **code** *is mapping*

$$C : \mathcal{X} \to \mathcal{D}^*.$$

There are different codes. A classical example of a code is Morse alphabet, where $\mathcal{D}$ consists of three elements: a dot, a dash and a letter space. Actually there is also a word space but when coding letters only, it will not be needed. In Morse code, short letters represent frequent letters (in English) and long sequences represent infrequent letters. This makes Morse code reasonably efficient but, as we shall see, this is not the most efficient (optimal) code. One can see this immediately by noticing that one of the three code-letters – space – is used in the end of the word, only.

Def 2.2 *A code C is* **non-singular**, *when it is injective i.e. every element of $\mathcal{X}$ is mapped into a different codeword: if $x_i \neq x_j$ then $C(x_i) \neq C(x_j)$.*

Non-singularity is sufficient to decode uniquely letters, but typically one need to code-words. An then a stronger property is needed.

Def 2.3 *An* **extension** *of a code C is a mapping C^* from $\mathcal{X}^*$ into $\mathcal{D}^*$ defined as follows*

$$C^* : \mathcal{X}^* \to \mathcal{D}^*, \quad C^*(x_1 \cdots x_n) := C(x_1) \cdots C(x_n).$$

Hence the extension of a code C is a concatenation of codewords of letters to obtain a codeword for word.

Def 2.4 *A code C is* **uniquely decodable**, *if its extension is non-singular.*

Hence, if C is uniquely decodable, then to every codeword $C(x_1) \cdots C(x_n)$ corresponds only one original word (source string) $x_1 \cdots x_n$. However, one may have to look at the entire string to determine even the first symbol in the corresponding source string. It is natural to expect that the first letter x_1 can be decoded as soon as $C(x_1)$ has been observed – decoding can be performed "on-line". This means that $C(x_1)$ cannot be the beginning (prefix) of any other codeword.

Def 2.5 *A code C is* **prefix code (prefix-free code, instantaneous code)** *if no codeword is a prefix of any other codeword i.e. there are no different letters x_i and x_j such that $C(x_i)$ is a prefix of $C(x_j)$.*

Clearly prefix codes are uniquely decodable and uniquely decodable codes are non-singular.

Examples:

- Morse code is prefix code, since every codeword ends with space.
- Let $\mathcal{X} = \{a, b, c, d\}$ and consider binary codes C_1, C_2, C_3 and C_4, represented in the table.

$\mathcal{X}$	C_1	C_2	C_3	C_4
a	0	0	10	0
b	0	010	00	10
c	1	01	11	110
d	0	10	110	111

  Code C_1 is not non-singular; C_2 is non-singular but not uniquely decodable, since 010 could stand for the letter b as well as for the words ad and ca. Code C_3 is uniquely decodable but not prefix code. Indeed, to figure out whether $1100\ldots0$ is a codeword of $cbb\ldots b$ or $dbb\ldots b$, one has to count all 0's. Thus, one cannot decode the first letter before the whole codeword is read. This is so, because the codeword $C(c) = 11$ is a prefix of the codeword $C(d) = 110$. Code C_4 is prefix code, hence every letter can be decoded as soon as it codeword has observed. Decode "on-line" the word 01011111010.

2.2 Kraft inequality

Prefix code as a tree. Every prefix code can be represented as D-ary tree, where every node has at most D children. To every branch of a tree correspond a letter from $\mathcal{D}$, to every leave corresponds a letter from $\mathcal{X}$ and the path from the root to the letter is the codeword of that letter (leave). The length of that codeword is the length (or level) of that leave.

Example: Let $D = 3$. Let us construct a code tree of the following prefix code:

a	b	c	d	e	f	g	h
1	2	010	012	02	000	001	002

In what follows, given a code C, we shall denote by $l(x) := |C(x)|$ the length of the codeword. In the example above, $|\mathcal{X}| = 8$ and the lengths of codewords in increasing order are

$$l_1 = l_2 = 1, \quad l_3 = 2, \quad l_4 = l_5 = l_6 = l_7 = l_8 = 3.$$

It is clear that when C is a prefix code and can be represented as a tree, then the codeword lengths cannot be arbitrary small. Kraft inequality gives a nice bound.

Theorem 2.6 (Kraft inequality) *Let $C : \mathcal{X} \to \mathcal{D}^*$ be a prefix code $l_i = l(x_i)$. Then*

$$\sum_i D^{-l_i} \leq 1. \tag{2.1}$$

Conversely, let $\{l_i\}_{i=1}^{|\mathcal{X}|}$ integers that satisfy (2.1). Then there exist prefix code $C : \mathcal{X} \to \mathcal{D}^*$ such that $l_i = l(x_i)$ $\forall x_i \in \mathcal{X}$.

Proof. Let us start with proving the first claim for the case $|\mathcal{X}| = m < \infty$. Let $l^* := \max\{l_1, \ldots, l_m\} < \infty$. Organize the set $\{l_1, \ldots, l_m\}$ (code) as a D-ary tree. A codeword at level l_i has $D^{l^*-l_i}$ descendants at level l^*. All the the descendant sets (corresponding to different l_i) must be disjoint. Therefore the total number of nodes in these sets (over all codewords) must be less than or equal to D^{l^*}:

$$\sum_{i=1}^{m} D^{l^*-l_i} \leq D^{l^*} \quad \Leftrightarrow \quad \sum_{i=1}^{m} D^{-l_i} \leq 1.$$

Let us now prove the same claim for general case, where $|\mathcal{X}| \leq \infty$. Recall

$$\mathcal{D} = \{0, \ldots, D-1\}$$

and consider the codeword $d_1 d_2 \cdots d_{l_i}$. Let $0.d_1 d_2 \cdots d_{l_i}$ be the real number having the D-ary expansion $0.d_1 d_2 \cdots d_{l_i}$, i.e.

$$0.d_1 d_2 \cdots d_{l_i} = \sum_{j=1}^{l_i} \frac{d_j}{D^j}. \tag{2.2}$$

Consider the interval (sub-interval of $[0, 1]$)

$$[0.d_1 d_2 \cdots d_{l_i}, \quad 0.d_1 d_2 \cdots d_{l_i} + D^{-l_i}).$$

corresponding to the codeword $d_1 d_2 \cdots d_{l_i}$. To this interval belong all real numbers whose D-ary expansion begins with $0.d_1 d_2 \cdots d_{l_i}$. Clearly the length of that interval is D^{-l_i}. Since C is prefix code the intervals corresponding to different codewords are disjoint. Since they are all sub-intervals of $[0, 1]$, their lengths sum up something less than or equal to 1. This means that (2.1) holds.

Let us prove the second statement: we are given the set $\{l_i\}_{i=1}^{|\mathcal{X}|}$ satisfying (2.1). We aim to construct a prefix code so that the codewords have lengths $\{l_i\}$. Since (2.1) holds, it is possible to divide unit interval into disjoint subintervals with lengths D^{l_i}. Indeed, order $l_1 \leq l_2 \leq \cdots$. Let the first interval be $[0, D^{-l_1})$, second $[D^{-l_1}, D^{-l_1} + D^{-l_2})$ and so on.
Thus the first interval corresponds to l_1. It begins with 0 that can be represented as

$$0.\underbrace{0 \cdots 0}_{l_1}$$

The first interval ends with D^{-l_1} with D-ary expansion being

$$0.\underbrace{0\cdots01}_{l_1}.$$

Clearly the first interval consists of these real numbers, whose D-ary expansion begins with $0.0\cdots0$ (with l_1 zeros).
Second interval corresponds to l_2. We represent both D^{-l_1} as well as $D^{-l_1}+D^{-l_2}$ as D-ary real numbers with l_2 numbers after 0.. Recall that $l_2 \geq l_1$. If $l_2 = l_1$, then the D^{-l_1} will be represented just like previously, otherwise it will be represented as

$$0.\overbrace{\underbrace{0\cdots01}_{l_1}0\cdots0}^{l_2}. \tag{2.3}$$

Clearly one needs at most l_2 figures after 0. to expand $D^{-l_1}+D^{-l_2}$: To this interval belong all these real numbers whose D-ary expansion begins with (2.3). The beginning of the third interval (corresponding to l_3) can be represented as D-ary number $0.d_1d_2\cdots d_{l_3}$. Again, recall $l_3 \geq l_2$ and if $l_3 > l_2$, then the last $l_3 - l_2$ elements of that representation are zero. The D-ary expansion of the endpoint of that interval $D^{-l_1}+D^{-l_2}+D^{-l_3}$ has obviously at most l_3 elements after 0.. We proceed similarly: the interval corresponding to l_i begins with $D^{-l_1}+\cdots+D^{-l_{i-1}}$. The D-ary expansion of that number has at most l_{i-1} elements after 0. and we use l_i elements which is possible because $l_i \geq l_{i-1}$. Hence, the D-ary representation is $0.d_1\cdots d_{l_i}$. To this interval belong real numbers whose D-ary expansion begins with that representation.
To construct the code, take to every l_i (to letter x_i) the word $d_1\cdots d_{l_i}$ from the D-ary expansion of $D^{-l_1}+\cdots+D^{-l_{i-1}}$ (beginning of the interval). Since different codewords belong to different intervals, the obtained code is a prefix code. ■

Examples:

- Consider the code C_4. Then $l_1 = 1$, $l_2 = 2$, $l_3 = l_4 = 3$. Let us find the real numbers whose D-ary representations are $0.d_1d_2\cdots d_{l_i}$. We obtain

  $$0.0_2 = 0, \quad 0.10_2 = 0.1_2 = 0.5, \quad 0.110_2 = 0.11_2 = \frac{1}{2}+\frac{1}{4} = 0.75, \quad 0.111_2 = \frac{1}{2}+\frac{1}{4}+\frac{1}{8} = 0.875.$$

  Hence the intervals used in the first part of the proof are

  $$[0, 0+\frac{1}{2}), \quad [0.5, 0.5+0.25), \quad [0.75, 0.75+0.125), \quad [0.875, 0.875+0.125).$$

  In this example, the Kraft inequality is an equality.

- The converse: Let $\{1,2,3,3\}$ be the lengths of the codewords. The easiest way to construct the corresponding code is to construct a tree. The procedure used in the proof is as follows. Let us construct the intervals:

  $$[0,\frac{1}{2}), \quad [\frac{1}{2},\frac{1}{2}+\frac{1}{4}), \quad [\frac{1}{2}+\frac{1}{4},\frac{1}{2}+\frac{1}{4}+\frac{1}{8}), \quad [\frac{1}{2}+\frac{1}{4}+\frac{1}{8},1).$$

With binary representation these intervals (recall the numbers of figures after 0. must be l_i) are

$$[0.\overbrace{0}^{1}, \quad 0.1), \quad [0.\overbrace{10}^{2}, \quad 0.11), \quad [0.\overbrace{110}^{3}, \quad 0.111), \quad [0.\overbrace{111}^{3}, \quad 1).$$

Codewords: 0, 10, 110, 111.

- Let the lengths of the codewords be $\{2, 2, 3, 3\}$. Note that Kraft inequality is strict: $2^{-2} + 2^{-2} + 2^{-3} + 2^{-3} = \frac{3}{4} < 1$. Intervals

$$[0, \frac{1}{4}), \quad [\frac{1}{4}, \frac{1}{2}), \quad [\frac{1}{2}, \frac{1}{2} + \frac{1}{8}), \quad [\frac{1}{2} + \frac{1}{8}, \frac{1}{2} + \frac{1}{8} + \frac{1}{8}).$$

With binary expansion these intervals are

$$[0.00, 0.01), \quad [0.01, 0.10), \quad [0.100, 0.101), \quad [0.101, 0.110).$$

Codewords: 00, 01, 100, 101.

2.3 Expected length and entropy

Let us consider the case where letters are chosen randomly according to a distribution P on $\mathcal{X}$. In other words, we consider a random variable $X \sim P$. Given a code C we are interested in the expected length of a codeword. Since $l(x)$ is the length of codeword $C(x)$, the *expected length of the code* C is

$$L(C) = \sum_x l(x)P(x).$$

Example: Consider the code C_4. Let $P(a) = \frac{1}{2}$, $P(b) = \frac{1}{4}$, $P(c) = P(d) = \frac{1}{8}$. Then

$$L(C_4) = \frac{1}{2} + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = \frac{7}{4}.$$

Note that $H(P) = \frac{7}{4}$.

Hence L is the average number of symbols we need to describe the outcome of X when the code C is used. Clearly, the smaller the expected length, the better code. The expected length is obviously small when all codeword are small i.e. $l(x)$ is small for every x. On the other hand, we know that for prefix code the lengths $l(x)$ cannot be arbitrary small, since they have to satisfy Kraft inequality. But given the lengths $l(x)$ that satisfy Kraft equality, how to choose the code with minimal expected length? We know how to find the codewords, but how to assign these words to letters x? The intuition correctly suggest that the expected length is small if the frequent (high probability) letters have small codewords and infrequent letters longer. Also the Morse code follows the same principle, but the symbol "space" is only used to mark the end of the word, hence one can figure out a three letter prefix code with smaller expected length.

The next theorem provides a fundamental lower bound on the expected length of any prefix code. It turns out than the for D-ary code the expected length cannot be lower then $H_D(P)$.

Theorem 2.7 *Let $C : \mathcal{X} \to \mathcal{D}^*$ be a prefix code. Then*

$$L(C) \geq H_D(P),$$

with the equality if and only if $l(x) = -\log_D P(x)$, $\forall x \in \mathcal{X}$.

Proof.

$$\begin{aligned} L(C) - H_D(P) &= \sum_x l(x)P(x) - \sum_x P(x) \log_D \frac{1}{P(x)} \\ &= -\sum_x P(x) \log_D D^{-l(x)} + \sum_x P(x) \log_D P(x). \end{aligned}$$

Let

$$c := \sum_x D^{-l(x)}, \quad R(x) := \frac{D^{-l(x)}}{c}.$$

Then

$$L(C) - H_D(P) = \sum_x P(x) \log_D \frac{P(x)}{R(x)} - \log_D c = D(P||R) + \log_D \frac{1}{c} \geq 0,$$

because $D(P||R) \geq 0$ and from Kraft inequality, it follows $\log_D \frac{1}{c} \geq 0$.
The inequality is an equality only if $P = R$ and $c = 1$. This holds iff for every $x \in \mathcal{X}$ it holds $P(x) = D^{-l(x)}$. Necessary condition is that $-\log_D P(x)$ is integer for every $x \in \mathcal{X}$.
■

Optimal codes for D-adic distribution. The code with minimum expected length is called *optimal*. From the preceding theorem, it follows that if P satisfies the following condition:

$$\log_D \frac{1}{P(x)} \in \mathbb{Z}, \quad \forall x \in \mathcal{X}, \tag{2.4}$$

(sometimes such distributions are called *D-adic*), then optimal prefix code is easy to construct: take

$$l(x) = \log_D \frac{1}{P(x)}.$$

The lengths $l(x)$ satisfy Kraft inequality (with equality) and the corresponding optimal code can be constructed via constructing the tree or using the interval as in the proof of Kraft inequality. The expected length of such code is $H_D(P)$ and from the preceding theorem we know that it must be then optimal.

Example: A distribution satisfying (2.4) is

a	b	c	d	e	f	g	h	i
$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$

The lengths of codewords are $\{l(x)\}_{x\in\mathcal{X}} = \{5,5,4,4,4,3,3,2,2\}$. The optimal code can be constructed by constructing a full binary tree at depth 5 and reduce or prune it according to the word lengths (Exercise 1).
Second option is to use intervals as in the proof of Kraft equality. Then the intervals are

$[0, 2^{-2}), [2^{-2}, 2^{-2}+2^{-2}), [2^{-1}, 2^{-1}+2^{-3}), [2^{-1}+2^{-3}, 2^{-1}+2^{-3}+2^{-3}),$
$[2^{-1}+2^{-2}, 2^{-1}+2^{-2}+2^{-4}), [2^{-1}+2^{-2}+2^{-4}, 2^{-1}+2^{-2}+2^{-3}),$
$[2^{-1}+2^{-2}+2^{-3}, 2^{-1}+2^{-2}+2^{-3}+2^{-4}), [2^{-1}+2^{-2}+2^{-3}+2^{-4}, 2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5})$
$[2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5}, 1).$

These intervals in binary expansion (2.2) are

$[0.00, 0.01), [0.01, 0.10), [0.100, 0.101), [0.101, 0.110), [0.1100, 0.1101), [0.1101, 0.1110),$
$[0.1110, 0.1111), [0.11110, 0.11111), [0.11111, 1).$

The code:

a	b	c	d	e	f	g	h	i
11111	11110	1110	1101	1100	101	100	01	00

Shannon-Fano code. Unfortunately not all distributions satisfy (2.4) and then the above-described easy procedure cannot be applied. We can modify it as follows: replace $\log_D \frac{1}{P(x)}$ (not necessary an integer) with

$$l(x) = \lceil \log_D \frac{1}{P(x)} \rceil. \tag{2.5}$$

The lengths $l(x)$ obtained by (2.5) clearly satisfy Kraft inequality, hence a prefix code with (codeword) lengths $l(x)$ exists. Such a code is called **Shannon-Fano** code. In other words, a code C is Shannon-Fano code iff for every $x \in \mathcal{X}$ the relation (2.5) holds.

Clearly the rounding makes the code longer, hence in general (unless distribution is D-aric) the expected length of Shannon-Fano code is larger than $H_D(P)$. This does not necessarily imply that Shannon-Fano code is not optimal prefix code, but typically it is the case. How much do we loose by rounding? Note

$$\lceil \log_D \frac{1}{P(x)} \rceil < \log_D \frac{1}{P(x)} + 1.$$

Therefore

$$L(C) = \sum_x P(x) \lceil \log_D \frac{1}{P(x)} \rceil < \sum_x P(x) \log_D \frac{1}{P(x)} + 1 = H_D(P) + 1.$$

Corollary 2.1 *For every distribution, there exist a prefix code $C : \mathcal{X} \to \mathcal{D}^*$ such that*

$$H_D(P) \leq L(C) < H_D(P) + 1.$$

Example: Let P uniform over 5 letter: $P(x_i) = \frac{1}{5}$, $i = 1, \ldots, 5$. then

$$l(x) = \log \frac{1}{P(x)} = \log 5 \text{ ja } \lceil \log \frac{1}{P(x)} \rceil = 3.$$

A Shannon-Fano code:

x_1	x_2	x_3	x_4	x_5
000	001	010	011	110

The expected length of that code is 3. Hence

$$H(P) = \log 5 < L(C) = 3 < \log 10 = H(P) + 1.$$

It is possible to construct a prefix code with lengths $\{3, 3, 2, 2, 2\}$ (how?). The expected length of that code is $\frac{12}{5} = 2.4$, hence (for that P) Shannon-Fano code is not optimal.

Wrong distribution. In order to construct Shannon - Fano code, the distribution of P has to be known. Suppose that by constructing the code, instead of true distribution P, one uses wrong distribution Q. Clearly the obtained code might not be (close to) optimal, on the other hand, if $P \approx Q$, then one could expect also that the obtained codes have similar length. The following theorem shows that for binary codes the increase of the expected length is about $D(P||Q)$.

Theorem 2.8 *Assume $D = 2$. Let P be the true distribution of letters and let*

$$l_Q(x) := \lceil \log \frac{1}{Q(x)} \rceil.$$

Then

$$H(P) + D(P\|Q) \leq \sum_x l_Q(x)P(x) < H(P) + D(P\|Q) + 1. \tag{2.6}$$

Proof. The upper bound:

$$\begin{aligned}
\sum_x l_Q(x)P(x) &= \sum_x \lceil \log \frac{1}{Q(x)} \rceil P(x) < \sum_x P(x)\Big(\log \frac{1}{Q(x)} + 1\Big) \\
&= \sum_x P(x)\Big(\log \frac{P(x)}{Q(x)} + \log \frac{1}{P(x)} + 1\Big) \\
&= D(P\|Q) + H(P) + 1.
\end{aligned}$$

To find the lower bound is Exercise 2 ■

Remark: The statement obviously hods for $D > 2$ provided entropy as well as K-L distance are defined using $\log_D$ instated of $\log_2$.

2.4 Huffman code

Shannon-Fano code is optimal (with shortest expected length), if P is D-adic i.e. satisfies (2.4). We shall now describe a relatively simple procedure that gives optimal prefix code for any distribution. The procedure is called **Huffman procedure** and resulting codes **Huffman codes**. Recall that every prefix code is represented by a code tree, with each leaf in the tree corresponding to a codeword. The Huffman procedure is to form a tree such that the expected length is minimum.

NB! Assume $|\mathcal{X}| < \infty$.

2.4.1 Huffman procedure

The easiest way to understand the procedure is by example.

Example: Let $\mathcal{X} = \{a, b, c, d, e\}$ and let P be

a	b	c	d	e
0.35	0.1	0.15	0.2	0.2

Huffman procedure for $D = 2$. Huffman procedure for binary tree is: find two letters with smallest probability and merge them to form an internal node. In the example, thus, join the letters b, c. Sum the corresponding probabilities $0.1 + 0.15 = 0.25$ and consider reduced alphabet $\{a, \{b, c\}, d, e\}$ with probabilities $0.35, 0.25, 0.2, 0.2$. Hence, we obtain the so-called reduced distribution

a	$\{b, c\}$	d	e
0.35	0.25	0.2	0.2

Now find two letters with smallest probability in reduced alphabet and merge them. In this example, merge the letters d and e (sum the probabilities) and form another internal node in the tree. After merging d and e, one ends up with the following reduced alphabet

a	$\{b, c\}$	$\{d, e\}$
0.35	0.25	0.4

Now, again find in the distribution above two letters with smallest probability and merge them in the tree. We get once more reduced alphabet $\{a, b, c\}, \{d, e\}$ and new distribution

$\{a, b, c\}$	$\{d, e\}$
0.6	0.4

In this alphabet, there are only two letters which should be merged in the first level. A code tree is then formed. Upon assigning 0 and 1 (in any convenient way) to each pair of branches at an internal node, we obtain a codeword assigned to each x. For example the obtained Huffman code C can be as follows.

A Huffman code:

a	b	c	d	e
00	010	011	10	11

The expected length is $L(C) = 2\frac{3}{4} + 3\frac{1}{4} = \frac{9}{4} = 2.25$. Compare it with

$$H(P) = -0.35\log(0.35) - 0.1\log(0.1) - 0.15\log(0.15) - 0.4\log(0.4) = 2.202.$$

If in a step, there are more than one pairs to merge (with smallest probability) pick any of them. All choices guarantee optimality.

Huffman procedure for $D > 2$. Huffman procedure for constructing D-ary code (tree) is essentially the same: the smallest D probability masses are merged in each step. If the resulting tree is formed in $k+1$ steps, then there will be $k+1$ internal nodes and $D + k(D-1)$ leaves. Hence the alphabet contains $D + k(D-1)$ letters (for an integer k), then the Huffman procedure can be applied directly. Otherwise, we need to add a few dummy symbols with probability 0 to make the total number of symbols have the form $D + k(D-1)$. Adding those dummy variables will not change the distribution, but they ensure that in the last step of the procedure D letters can be merged.

Examples:

- Let P be as follows

a	b	c	d	e	f
0.25	0.25	0.2	0.1	0.1	0.1

  Let $D = 3$. since $6 \neq 3 + k(3-1)$, one dummy variable should be added.

a	b	c	d	e	f	$*$
0.25	0.25	0.2	0.1	0.1	0.1	0

  Huffman procedure: in the first level e, f and $*$ will be merged; next $\{e, f, *\}$, d and c will be merged; in the last step $\{c, d, e, f, *\}$, b and a will be merged.

  A Huffman code:

a	b	c	d	e	f
1	2	01	02	000	001

- Consider once again the first example. Let $D = 4$. Since $|\mathcal{X}| = 5$, two dummy variables should be added: $7 = (D-1)+D$. With dummy variables, the distribution is

a	b	c	d	e	$*$	$*$
0.35	0.1	0.15	0.2	0.2	0	0

In the first step the letters $d, e, *, *$ will be merged. Then the rest.

A Huffman code:

a	b	c	d	e
0	30	31	2	1

Remark: Note that Huffman procedure can be applied for finite alphabet, only.

2.4.2 Huffman code is optimal

Let $\mathcal{X} = \{x_1, \ldots, x_m\}$ and w.l.o.g. assume

$$P(x_1) \geq P(x_2) \geq \cdots \geq P(x_m). \tag{2.7}$$

since $|\mathcal{X}| < \infty$, we know that there exists at least one optimal code. We shall now study the properties of optimal codes. The first property states that every optimal code assigns longer codewords to the less probably letters.

Proposition 2.1 *Let C be an optimal code. Then $l(x_i) > l(x_j)$ only if $P(x_i) \leq P(x_j)$.*

Proof. Assume that there exist x_i and x_j such that $P(x_i) > P(x_j)$ and $l(x_i) > l(x_j)$. Define a new code C^* by changing the codewords $C(x_i)$ and $C(x_j)$. Since

$$\begin{aligned} L(C) - L(C^*) &= P(x_i)l(x_i) + P(x_j)l(x_j) - (P(x_i)l(x_j) + P(x_j)l(x_i)) \\ &= (P(x_i) - P(x_j))(l(x_i) - l(x_j)) > 0, \end{aligned}$$

we obtain that C cannot be optimal. ■

From Proposition 2.1 it follows that for every optimal code, there is an ordering $\mathcal{X} = \{x_i\}$ such that (2.7) holds and

$$l(x_1) \leq l(x_2) \leq \cdots \leq l(x_m). \tag{2.8}$$

Def 2.9 *The codewords $d', d'' \in \mathcal{D}^*$ are* **siblings**, *when they have the same length and differ only in the last symbol.*

Binary Huffman codes ($D = 2$)**.** Let us, for simplicity, consider the binary codes and proof the optimality of binary Huffman codes. In case of binary codes, every codeword has only one sibling. At first, we show that there exists an optimal code C so that the codewords associated to the words with smallest probabilities are siblings.

Proposition 2.2 *There exists optimal code C so that $C(x_{m-1})$ and $C(x_m)$ are siblings.*

Proof. Let C be an optimal code such that equalities (2.7) and (2.8) both hold. This means that $C(x_m)$ is the longest codeword. Since $C(x_m)$ is the longest, its sibling $C(x_m)$ cannot be prefix of any other codeword. Also it is clear that the sibling of $C(x_m)$ has to be a codeword – if not, we could reduce the length of $C(x_m)$ by one (replace $C(x_m)$ by its

parent) and that would contradict the optimality of C. Hence, there is a letter x_j so that $C(x_j)$ and $C(x_m)$ are siblings. If $j = m-1$, then the statement holds. If $j < m-1$, then from (2.8) we obtain $l(x_j) = l(x_{m-1}) = l(x_m)$, hence we can change $C(x_j)$ and $C(x_{m-1})$ without loosing the optimality. ■

Theorem 2.10 *Binary Huffman code is optimal.*

Proof. By Proposition 2.2, there exists an optimal code C so that $C(x_{m-1})$ and $C(x_m)$ are siblings. Note that Huffman code has the same property. If we replace these codeword by a common codeword at their parent, then we obtain a reduced code C' (reduced tree), corresponding to the reduced distribution where x_m and x_{m-1} are merged into one letter, say y, having the probability $p_m + p_{m-1}$. The code C' is in average shorter than C, their difference is

$$L(C) - L(C') = lp_m + lp_{m-1} - (p_m + p_{m-1})(l-1) = p_m + p_{m-1},$$

where $l = l(x_m) = l(x_{m-1})$. It is important to notice that the difference does not depend on the structure of the tree (code) C. Hence C is optimal iff C' is optimal on reduced alphabet and from any optimal code on reduced alphabet, we can easily obtain (by replacing the node y by two descendants) optimal original code. In other words, after finding an optimal tree (code) in reduced alphabet, we obtain an optimal tree in original alphabet by attaching to y a subtree that is created with Huffman procedure
By Proposition 2.2, again, we know that there is an optimal code on the reduced alphabet so that the codewords corresponding to the two smallest probabilities are siblings. Merging these letters, just like in Huffman procedure, we get more reduced alphabet. Just like previously, we see that from any optimal tree on more reduced alphabet, we get an optimal tree on original alphabet by growing it according to Huffmann procedure.
Proceeding with Huffman procedure, we eventually end up with reduced alphabet consisting on two (merged) letters. Each of these two letters is a root of a subtree obtained by Huffman procedure. Moreover, we know that with these subtrees an optimal tree on two letters can be extended to an optimal tree for original alphabet. Obviously there is only one optimal tree on two letter alphabet – joining these letters on first level – and, therefore Huffman procedure produces an optimal tree. ■

The case $D > 2$. Let us briefly sketch the proof for the case $D > 2$. W.l.o.g. assume that the size of the alphabet is $D + k(D-1)$, where k is an integer (otherwise add dummy letters). Recall that a D-ary tree with $D + k(D-1)$ leaves is called *complete*, if every internal node has exactly D children. Complete tree satisfies Kraft inequality with equality. It is not hard to see that every optimal D-ary tree with $D + k(D-1)$ leaves has to be complete. After seeing that, the proof of the optimality of Huffman D-ary tree is almost the same as for binary tree. Indeed, Proposition 2.1 holds for every D. Therefore, there is an optimal code C so that equalities (2.7) and (2.8) both hold. Hence $C(x_m)$ has to be the longest codeword and since C corresponds to the complete tree, all siblings of $C(x_m)$ must be codewords as well. The arguing just like in the proof of Proposition

2.2, we see that there exists an optimal code such that the codewords corresponding to $x_{m-D+1}, x_{m-D+2}, \ldots, x_m$ are siblings. Now the proof of Theorem 2.10 directly applies.

Remarks:

- Not all optimal codas are Huffman ones, i.e. there exist optimal codes that cannot be constructed by Huffman procedure. For example, let $\mathcal{X} = \{a, b, c, d, e, f\}$ and let P be uniform. Consider two binary codes C_1 and C_2 given as follows

letter \ code	C_1	C_2
a	11	111
b	101	110
c	100	101
d	011	100
e	010	01
f	00	00

  The code C_2 is a Huffman code, but C_1 cannot be constructed by Huffman procedure, both are optimal (Exercise 6).

- The expected length of an optimal code is not always $H_D(P)$. Indeed, in the example above
$$L = L(C_1) = L(C_2) = \frac{8}{3} > \log 6 = H(P).$$

- We know that the expected length of optimal code L always satisfies inequalities
$$H_D(P) \leq L < H_D(P) + 1,$$
  Where the first inequality can be strict or equality. Can the second inequality be improved, i.e. would it possible to replace the number 1 in the second inequality be something smaller like 0.5? Let us see that this is not possible meaning that L can be arbitrary close to $H_D(P) + 1$. To see that consider the distribution (k is large enough integer)

a	b	c	d
$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	$1 - \frac{3}{k}$

  The lengths of Huffman binary codewords are $l(a) = l(b) = 3$ $l(c) = 2$ $l(d) = 1$ (provided k is large enough), hence $L = \frac{8}{k} + 1 - \frac{3}{k} \to 1$, if $k \to \infty$. On the other hand
$$H(P) = \frac{3}{k} \log k - (1 - \frac{3}{k}) \log(1 - \frac{3}{k}) \to 0, \text{ if } k \to \infty.$$
  Hence $H(P) + 1 - L \to 0$, if $k \to \infty$.
  What is the Shannon-Fano code in this case?

- It is not true that the codeword lengths of Shannon-Fano code are always at least as long as the ones of any optimal code. As an counterexample consider the distribution

a	b	c	d
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{12}$

  Huffman codeword lengths are $(2, 2, 2, 2)$ or $(1, 2, 3, 3)$. Hence, there exists an optimal code so that $l(c) = 3$. By Shannon-Fano code, however, $l(c) = 2$.

2.4.3 Huffman procedure with infinite alphabet

When $|\mathcal{X}| = \infty$, then Huffman procedure cannot be, in general, applied. However, under some additional assumptions the code can be constructed "piecewise", from up to down. For simplicity assume that $D = 2$.
Let the probabilities be arranged in decreasing order

$$p_1 \geq p_2 \geq \cdots .$$

Suppose, there are infinitely many atoms p_m, satisfying the following condition

$$p_m \geq \sum_{i>m} p_i =: p_m^*. \tag{2.9}$$

Suppose, for a moment that alphabet is finite but very large. Let $p_{m_1}, p_{m_2}, \ldots$ satisfy (2.9). Since p_{m_1} satisfies (2.9), it is clear that applying Huffman procedure (since $\mathcal{X}$ is finite, it is possible), all letters corresponding to p_j, where $j > m_1$, will be joined before p_{m_1}. Hence, at some point Huffman procedure reaches to the restricted distribution (alphabet)

$$p_1, p_2, \ldots, p_{m_1}, p_{m_1}^*. \tag{2.10}$$

Now it is clear that one can start with constructing first the optimal tree corresponding to (2.10). After that the subtree starting from the node $p_{m_1}^*$ can be constructed. For that, we consider the distribution (proportional to)

$$p_{m_1+1}, p_{m_1+2}, \ldots, p_{m_2}, p_{m_2}^*. \tag{2.11}$$

The numbers (2.11) are not probability distribution, since their sum is $p_{m_1}^* < 1$. From the point of view of Huffman procedure, the total sum is not important. Therefore, we construct the Huffman tree for (2.11), the root of that subtree is $p_{m_1}^*$. Now the tree (code) with leaves (letters)

$$p_1, p_2, \ldots, p_{m_1}, p_{m_1+1}, p_{m_1+2}, \ldots, p_{m_2}, p_{m_2}^*$$

is constructed and the next step is to build the subtree starting from $p_{m_2}^*$. For that, again, we construct the Huffman tree for the atoms

$$p_{m_2+1}, p_{m_2+2}, \ldots, p_{m_3}, p_{m_3}^*, \tag{2.12}$$

so that the root of that tree is $p^*_{m_2}$ and now the tree corresponding to

$$p_1, p_2, \ldots, p_{m_3}, p^*_{m_3}$$

is constructed. Clearly such a piecewise procedure is independent of the number of letters and – given that there are infinitely many atoms p_m satisfying (2.9) – can also be applied for the case of infinite alphabet.

Example: The condition (2.9) holds for any m when the distribution P is geometric with parameter $p \geq 0.5$. The proof of that is an easy exercise.

2.5 Uniquely decodable codes

Every prefix code is uniquely decodable but not vice versa. Hence the class of uniquely decodable codes is larger than the one of prefix codes and it is reasonable to ask whether the expected length of optimal uniquely decodable code can be shorter than the expected length of the optimal prefix code. Here we proof that this is not the case, since Kraft inequality also holds for uniquely decodable codes. From this follows that the expected length of optimal uniquely decodable code is the same as that one of optimal prefix code. Indeed (as we shall see) every uniquely decodable code must satisfy Kraft inequality. But from Theorem 2.6 we know that for any set of integers $\{l_i\}$ satisfying Kraft inequality corresponds at least one prefix code with codeword lengths $\{l_i\}$. Hence, to any uniquely decodable code corresponds a prefix code with exactly the same codeword lengths and, hence, with the same expected length. So as far as the codeword lengths are concerned, the uniquely decodable codes have no advantage over prefix codes.

Theorem 2.11 (McMillan) *Let C be an uniqualy decodable code with codeword lengths $\{l(x)\}$. Then Kraft inequality holds*

$$\sum_x D^{-l(x)} \leq 1. \tag{2.13}$$

Proof. At first, we consider special case $|\mathcal{X}| < \infty$.
Let C^k be the k-extension of C, i.e.

$$C^k : \mathcal{X}^k \to \mathcal{D}^*, \quad C^k(x_1 \cdots x_k) = C(x_1) \cdots C(x_k).$$

$$\begin{aligned}
\Big(\sum_x D^{-l(x)}\Big)^k &= \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \cdots \sum_{x_k \in \mathcal{X}} D^{-l(x_1)} D^{-l(x_2)} \ldots D^{-l(x_k)} \\
&= \sum_{x_1 x_2 \cdots x_k \in \mathcal{X}^k} D^{-l(x_1)} D^{-l(x_2)} \ldots D^{-l(x_k)} \\
&= \sum_{x^k \in \mathcal{X}^k} D^{-l(x^k)},
\end{aligned}$$

where $x^k := x_1 \cdots x_k$ and

$$l(x^k) := l(x_1) + \cdots + l(x_k) = |C^k(x^k)|.$$

Let $a(m)$ be the number of source sequences x^k mapping into codewords length m. Formally,

$$a(m) = |\{x^k \in \mathcal{X}^k : l(x^k) = m\}|.$$

Recall we consider now the case where $\mathcal{X}$ is finite. Let

$$l_{max} := \max_{x \in \mathcal{X}} l(x).$$

Clearly

$$\max_{x^k \in \mathcal{X}^k} l(x^k) = k l_{max}.$$

Thus

$$\Big(\sum_x D^{-l(x)}\Big)^k = \sum_{x^k \in \mathcal{X}^k} D^{-l(x^k)} = \sum_{m=k}^{k l_{max}} a(m) D^{-m}.$$

Fix m and consider the set $\{x^k \in \mathcal{X}^k : l(x^k) = m\}$. There are at most D^m codewords with length m. Since C is uniquely decodable, the extension C^k is non-singular. Therefore, to every codeword (with length m) corresponds at most one original word. Hence, $a(m) \leq D^m$. Therefore

$$\Big(\sum_x D^{-l(x)}\Big)^k = \sum_{m=k}^{k l_{max}} a(m) D^{-m} \leq \sum_{m=1}^{k l_{max}} D^m D^{-m} = k l_{max}$$

or

$$\sum_x D^{-l(x)} \leq \big(k l_{max}\big)^{\frac{1}{k}}.$$

The left hand side is independent of k. Therefore

$$\sum_x D^{-l(x)} \leq \lim_{k \to \infty} \big(k l_{max}\big)^{\frac{1}{k}} = 1.$$

Let now $|\mathcal{X}| = \infty$. The proof above does not apply since $l_{max} = \infty$. Consider finite sub-alphabet $\mathcal{X}_m = \{x_1, \ldots, x_m\} \subset \mathcal{X}$. The restriction of an uniquely decodable code to alphabet $\mathcal{X}_m$ remains uniquely decodable. Since $\mathcal{X}_m$ is finite, from the first part of the proof we obtain

$$\sum_{x \in \mathcal{X}_m} D^{-l(x)} \leq 1.$$

This holds for every m so that

$$\sum_{x \in \mathcal{X}} D^{-l(x)} = \lim_{m \to \infty} \sum_{x \in \mathcal{X}_m} D^{-l(x)} \leq 1.$$

■

Note that trivially holds the converse: if the integers $\{l_i\}$ satisfy Kraft inequality, then there exist a prefix code having with these codeword lengths. Every prefix code is uniquely decodable, hence there is also an uniquely decodable code with given lengths.

2.6 Coding words

Let $X_1, \ldots, X_k$ be random vector on alphabet $\mathcal{X}^k$. We shall denote the elements of $\mathcal{X}^k$ by $x^k := (x_1, \ldots, x_k)$. This random vector could be interpreted as a random word with length k. Let C be a code on alphabet $\mathcal{X}$. Then its k-extension C^k is a code for words. On the other hand, one can consider the set $\mathcal{X}^k$ as an alphabet and then design a code

$$C_k : \mathcal{X}^k \to \mathcal{D}^*$$

with small expected length. Which approach – to design an good code for letters and then extend it to alphabet or to design a good code directly for words – is preferable? To answer that question, the measure of goodness should be specified. Clearly the code for words has longer codewords and the expected length of C_k depends on k. Therefore, for any code C_k, it is customary to measure the expected length per input letter. More specifically, with $l(x^k)$ being the codeword lengths of C_k, we define

$$L_k := \frac{1}{k} L(C_k) = \frac{1}{k} \sum_{x^k \in \mathcal{X}^k} P(x^k) l(x^k) = \frac{1}{k} El(X_1, \ldots, X_k).$$

Identically distributed letters. Consider the case where $X_1, \ldots, X_k$ are identically distributed (with distribution P) but not necessarily independent. Let C be a code for alphabet $\mathcal{X}$ and consider the k-extension C^k. It is easy to see that $L(C^k) = kL(C)$ so that

$$L_k(C^k) = L(C). \tag{2.14}$$

The proof of (2.14) is Exercise 15. Therefore, if C is optimal letter code for P, then

$$H_D(P) \le L_k < H_D(P) + 1,$$

and the right hand side cannot be improved.

Consider now the optimal code for words. From Corollary 2.1 we know that there exists a code C_k so that

$$H_D(X_1, \ldots, X_k) \le L(C_k) < H_D(X_1, \ldots, X_k) + 1,$$

hence

$$\frac{H_D(X_1, \ldots, X_k)}{k} \le L_k \le \frac{H_D(X_1, \ldots, X_k)}{k} + \frac{1}{k}. \tag{2.15}$$

i.i.d. words. Suppose now that $X_1, \ldots, X_k$ are i.i.d. with $X_i \sim P$. Then $H_D(X_1, \ldots, X_k) = \sum_{i=1}^k H_D(X_i) = kH_D(P)$ and from (2.15), we obtain

$$H_D(P) \le L_k < H_D(P) + \frac{1}{k}. \tag{2.16}$$

The inequality (2.16) is sometimes referred to as *Shannon first theorem (source coding theorem)*. Hence, there exists a code C_k such that $L_k(C_k)$ differs from $H_D(P)$ by at most

$\frac{1}{k}$. Hence, choosing k large enough, we can find a code for k-letter words having L_k arbitrary close to $H_D(P)$. This is not the case for extended code C^k, since there exists distribution P so that for optimal letter code C, it holds that $L_k(C^k) \approx H_D(P) + 1$, $\forall k$.

Stationary process. Let $X = X_1, X_2, \ldots$ be a stationary process, $X_i \sim P$. In information theory, such a process is called *stationary source* and can be considered as a model for the language. Let, for every k the code $C_k : \mathcal{X}^k \to \mathcal{D}^*$ be optimal. Recall that a stationary process always has an entropy rate

$$H_X = \lim_k \frac{H_D(X_1, \ldots, X_k)}{k} = \lim_k H_D(X_k | X_1, \ldots, X_{k-1}) \leq H(P).$$

For $D > 2$, the entropy rate is defined just like for $D = 2$. Since D is fixed, we leave it out from the notation. From (2.15) it follows that

$$L^* := \lim_k L_k = \lim_k \frac{H_D(X_1, \ldots, X_k)}{k} = H_X.$$

Hence, the entropy rate of a stationary process is the average number of bits per symbol required to code the process.

Let us now recapitulate. If $X = X_1, X_2, \ldots$ are i.i.d (a very special case of stationary process), then $H_X = H_D(P)$ so that $L^* = H_D(P)$ and the only advantage of coding the words over coding the letters (both optimally) is that for k large enough, we could get L_k arbitrary close to $H_D(P)$.
However, if $H_X < H_D(P)$ (recall that $H_X \leq H_D(P)$), then the for k large enough, the expected code length per input symbol L_k is (arbitrary) close to H_X, hence smaller than $H_D(P)$. Therefore, if H_X is much smaller then $H_D(P)$, the advantage of coding words instead of coding letters might be remarkable.

Example: Let X be a stationary MC with transition matrix I_k (k states). Then $H(P) = \log k$, but $L_k = H_X = 0$.

2.6.1 Elias extension

To every uniquely decodable code corresponds a prefix code with the same codeword lengths. If $\mathcal{X}$ is not very large, then constructing the tree (prefix code) with given codeword lengths can be easy; in general the interval method used in the proof of Kraft inequality can be used. In practice, however, it can be still complicated especially when alphabet is very large. The alphabet, in turn, can be arbitrary large when one codes the words $\mathcal{X}^k$ instead of the letters, since $\mathcal{X}^k$ increases with k.

We shall now consider an easy method of turning an uniquely decodable code into a prefix code by adding a suitable prefix. This makes the codewords longer, but for long codewords the length of prefix is very small in comparison with codeword lengths so that when coding stationary source the limit L^* remains unchanged.

Elias delta code.

Lemma 2.1 *There exists a prefix code* $E : \{1, 2, \ldots\} \to \mathcal{D}^*$ *such that*

$$|E(n)| = \log_D n + o(\log_D n) \tag{2.17}$$

Proof. Every number will be coded in three parts

$$E(n) = u(n)v(n)w(n),$$

where $w(n)$ is D-adic representation of n. Therefore

$$|w(n)| = \ulcorner \log_D(n+1) \urcorner.$$

The second part $v(n)$ is the D-adic representation of the length $|w(n)|$ and the first part $u(n)$ consists of $|v(n)|$ zeros. Therefore

$$|u(n)| = |v(n)| = \ulcorner \log_D(1 + \ulcorner \log_D(n+1) \urcorner) \urcorner$$

and

$$|E(n)| = \ulcorner \log_D(n+1) \urcorner + 2\ulcorner \log_D(1 + \ulcorner \log_D(n+1) \urcorner) \urcorner = log_D n + o(\log_D n).$$

It is easy to see that $E(n)$ is a prefix code. Assume, on contrary, that there exist integers n and m such that $E(m)$ is the prefix of $E(n)$ i.e.

$$u(n)v(n)w(n) = u(m)v(m)w(m)w'.$$

In this case $u(n) = u(m)$, because both consist of zeros and the first symbol of $v(m)$ and $v(n)$ is not zero. That, in turn implies that $v(n) = v(m)$, since they have to be at equal length. But then it must be that $w(m) = w(n)$ so that w' is empty and $n = m$.
■

The described code is called **Elias (delta) code**.

Example. Let $D = 2$ and let us find $E(12)$. Since $12_2 = 1100$, we get $w(12) = 1100$. Since $|w(12)| = 4$, we get $v(12) = 100$. Finally $u(12) = 000$. Thus

$$E(12) = u(12)v(12)w(12) = 0001001100.$$

Remark. If $D = 2$, then Elias delta code can be shortened by two bits. Indeed, since for every $n \geq 0$, $|v(n)| \geq 1$, then instead of writing $|v(n)|$ zeros in the beginning, one can write $|v(n)| - 1$ zeros. Secondly, since every binary number begins with one, it can be left out from the code. Thus $w(n)$ is now the binary representation of with the leading bit removed. However, $v(n)$ is still the length oh the full binary representation of n. The obtained code is now two bits shorter. Let that code be E^*. Thus

$$E^*(12) = 00100100.$$

Turning uniquely decodable codes into prefix codes. Let $C_k : \mathcal{X}^k \to \mathcal{D}^*$ be an uniquely decodable code for words with codeword lengths $l(x^k)$. The **Elias extension** C_k^* of C_k is defined as follows:

$$C_k^*(x^k) = E(l(x^k))C^k(x^k).$$

This is now prefix code, since the prefix $E(l(x^k))$ determines the length of the codeword. By decoding, one first decodes $E(l(x^k))$. Since E is a prefix code, one can decode it immediately (on-line). After decoding $E(l(x^k))$, one obtains the length of the following codeword $l(x^k)$ and hence knows exactly when the word ends. Therefore, the whole word can be decoded on-line.

Example: Let $D = 2$ and $C^k(x^k) = 001001100111$. The length of that word is 12. Since $E(12) = 0001001100$, we get

$$C_k^*(x^k) = 000100110000101100111.$$

In this example, the Elias prefix is almost as long as the codeword itself, but from (2.17) we know that when the codewords lengths increase (for example k increases), then the length of the prefix increases logarithmically and becomes negligible.

Combining codes. Another application of Elias extension is to combine several codes into one. Suppose, for every $k \geq 1$, we have a prefix code

$$C^k : \mathcal{X}^k \to \mathcal{D}^*.$$

With Elias prefix we can define a general prefix code

$$C : \mathcal{X}^* \to \mathcal{D}^*, \quad C(x^k) = E(k)C_k(x^k).$$

Then the prefix determines the (index of) code and then the word is decoded.

2.7 Exercises

1. Consider the distribution

a	b	c	d	e	f	g	h	i
$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$

   find the optimal code tree directly (Shannon-Fano code) and by Huffman procedure.

2. Find the lower bound in Theorem 2.8.

3. Let P

a	b	c	d	e	f	g	h
0.25	0.05	0.1	0.13	0.2	0.12	0.08	0.07

   Find optimal code for $D = 2$ and $D = 3$. Find their expected length.

4. Let the codeword lengths be $1, 1, 2, 2, 3, 3, 3$.

   a) Is there any binary code having such lengths? If yes, find it. Is it optimal for some P?

   b) Let $D = 3$. Is there any 3-code having such lengths? If yes, find it. Is it optimal for some P?

   c) Let $D = 4$. Is there any 4-code having such lengths? If yes, find it. Is it optimal for some P?

5. Can C be a Huffman code if the codewords are

   - $\{0, 10, 11\}$
   - $\{00, 01, 10, 110\}$
   - $\{10, 01, 00, \}$?

6. Let P be uniform over 6 letters. Prove that a code C with words 11, 101, 100, 011, 010, 00 is optimal but cannot be obtained by Huffman procedure.

7. A code is suffix code, if no codeword is suffix of any other codeword. Is a suffix code uniquely decodable?

8. Let
$$l_1 \leq l_2 \leq \cdots \leq l_m$$
be integers. For every $1 \leq k \leq m$ a binary codeword with length l_k is chosen randomly amongst all codewords with length l_k. In such a way, a random code is constructed. Let $\mathcal{C}$ be the set of prefix codes. Prove that
$$\mathbf{P}(C \in \mathcal{C}) = \prod_{k=1}^{m} \Big(1 - \sum_{j=1}^{k-1} 2^{-l_j}\Big)^{+}.$$
Prove that $\mathbf{P}(C \in \mathcal{C}) > 0$ iff the integers $l_1 \leq l_2 \leq \cdots \leq l_m$ satisfy Kraft inequality

9. Let $L_D(p_1, \ldots, p_m)$ be the length of optimal code of $(p_1, \ldots, p_m)$. Prove that $L_D(p_1, \ldots, p_m)$ is a continuous function on $\mathcal{P}^m$.

10. Prove that the equality $L_D(p_1, \ldots, p_m) = H_D(p_1, \ldots, p_m)$ implies that
$$m = D + k(D - 1),$$
where k in a non-negative integer.

11. Let $q < \frac{2}{3}$. Let $p \in [0, 1]$ such that
$$L_2(1 - q, \frac{q}{2}, \frac{q}{2},) = H_2(1 - p, \frac{p}{2}, \frac{p}{2}).$$
Find the relation between p and q.

12. **a)** Find $L_2(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$ and $L_4(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$.

    b) Consider binary code obtained from four-code ($D = 4$) in the following way: Every letter of $\mathcal{D} = \{\alpha, \beta, \gamma, \delta\}$ are coded into binary codewords as follows:

    $$\alpha \mapsto 00, \beta \mapsto 01, \gamma \mapsto 10, \delta \mapsto 11.$$

    Let us call this process *doubling*. Find the optimal 4-code for $(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$ and the binary code obtained by doubling. What is the expected length of the binary code obtained in such a way?

    c) Let $L_T(P)$ be the expected length of the binary code obtained form Huffman code (for P) by doubling (depends on chosen optimal 4-code). Prove

    $$L_2(P) \leq L_T \leq L_2(P) + 1.$$

    d) Show that the inequalities can be equalities.

13. Let $u_1, u_2, \ldots, u_m$ non-negative integers. Find the solution of the following problem

    $$\min_{l_1,\ldots,l_m} \sum_{i=1}^{m} u_i l_i$$
    $$\text{such that } \sum_{i=1}^{m} D^{-l_i} \leq 1.$$

14. Let P be such that $P(x_1) > P(x_2) \geq P(x_3) \geq \cdots \geq P(x_m)$. There exists a and b such that

    - if $P(x_1) > a$, then for every binary Huffman code $l(x_1) = 1.$;
    - If $P(x_1) < b$, then for every binary Huffman code $l(x_1) \geq 2$.

    Find minimal a and maximal b.

15. Let $X_1, \ldots, X_n$ be indentically distributed random variables. Let C a code on $\mathcal{X}$, and let C^k be its k-extension. Prove $L(C^k) = kL(C)$.

16. Let Y be a stationary MC on alphabet $\mathcal{X}$ with transition matrix

    $$\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

    Find the entropy rate of the process. Let C_1 C_2 and C_3 three binary codes on $\mathcal{X}$. Consider the following coding procedure: Code Y_1 with C_1. After observing Y_n, pick the code associated to that state (if $Y_n = 3$, then C_3) and code Y_{n+1} with that code. Then pick the code associated to Y_{n+1} and code with that code Y_{n+2} and so on. Do there exist codes C_1, C_2, C_3 such that $L^* = H_Y$?

17. Let P

a	b	c
0.5	0.25	0.25

Let $X_1, X_2, \ldots$ be i.i.d. with distribution P. Let C be a binary code on $\{a, b, c\}$. Consider the process

$$Z = Z_1 Z_2 Z_3, \ldots = C(X_1)C(X_2)\ldots$$

Is Z always stationary?
Find the entropy rate of Z provided C is as follows:

(a)

$$C(x) = \begin{cases} 0, & \text{if } x = a; \\ 10, & \text{if } x = b; \\ 11, & \text{if } x = c. \end{cases}$$

(b)

$$C(x) = \begin{cases} 00, & \text{if } x = a; \\ 10, & \text{if } x = b; \\ 01, & \text{if } x = c. \end{cases}$$

(c)

$$C(x) = \begin{cases} 00, & \text{if } x = a; \\ 1, & \text{if } x = b; \\ 01, & \text{if } x = c. \end{cases}$$

18. Let $P(x_1) \geq P(x_2) \geq P(x_3) \geq \cdots P(x_m)$. Define

$$F(x_i) := \sum_{k=1}^{i-1} P(x_k).$$

Let

$$l(x_i) := \ulcorner -\log P(x_i) \urcorner.$$

For every x_i take $C(x_i)$ as the binary representation of $F(x_i)$ rounded off to $l(x_i)$ bits. Prove that C is prefix code. This code is sometimes called as *Shannon code.*

3 Asymptotic equipartition property (AEP)

3.1 Weak typicality

Let $X_1, X_2, \ldots$ i.i.d. random variables on alphabet $\mathcal{X}$, where $X_i \sim P$.

NB! Assume throughout: $H := H(P) < \infty$.

Let $X_1, \ldots, X_n$ be (the first) n random variables. Values on set $\mathcal{X}^n$. We shall denote the elements of $\mathcal{X}^n$ by x^n. Thus

$$x^n := (x_1, \ldots, x_n).$$

Since $X_1, \ldots, X_n$ are i.i.d., for every $x^n \in \mathcal{X}^n$, it holds

$$P(x^n) = P(x_1, \ldots, x_n) = P(x_1) \cdots P(x_n).$$

We shall investigate the random variable $P(X_1, \ldots, X_n)$ and we shall see that with high probability

$$P(X_1, \ldots, X_n) \approx 2^{nH},$$

provided n is big enough. This means that most of the outcomes of $X_1, \ldots, X_n$ have almost the same probability when n is big – *asymptotic equipartition property.*

Def 3.1 *Let $\epsilon > 0$. Define the set $W_\epsilon^n \subset \mathcal{X}^n$ as follows: $x^n \in \mathcal{X}^n$ belongs to the set $W^n(\epsilon)$ if and only if*

$$2^{-n(H+\epsilon)} \leq P(x^n) \leq 2^{-n(H-\epsilon)}. \tag{3.1}$$

The elements of $W_n(\epsilon)$ are called **weakly ϵ-typical words**.

Theorem 3.2 (Weak AEP) *For every $\epsilon > 0$ the following statements hold:*

1 *If $x^n \in W_\epsilon^n$, then*

$$2^{-n(H+\epsilon)} \leq P(x^n) \leq 2^{-n(H-\epsilon)}. \tag{3.2}$$

2 *There exists $n_o(\epsilon)$ so that for every $n > n_o$*

$$P(W_\epsilon^n) > 1 - \epsilon. \tag{3.3}$$

3 *There exists $n_1(\epsilon)$ so that for every $n > n_1$*

$$(1-\epsilon)2^{n(H-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H+\epsilon)}. \tag{3.4}$$

The proof is based on the weak law of large numbers (weak LLN). From that, it immediately follows (here $\xrightarrow{P}$ stands for the convergence in probability)

$$-\frac{1}{n}\log P(X_1, \ldots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) \xrightarrow{P} -E \log P(X_1) = H. \tag{3.5}$$

Proof. **1** is the definition (3.1).

2 follows from (3.5). Indeed, from the convergence in probability, it follows that for every $\forall \epsilon > 0$ there exists n_o (depending on ϵ) so that

$$\mathbf{P}\Big(\big| -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i) - H\big| \leq \epsilon\Big) \geq 1 - \epsilon, \tag{3.6}$$

provided $n > n_o$.

3: Since the probability of a weakly ϵ-typical word is at least $2^{-n(H+\epsilon)}$, then

$$1 \geq P(W_\epsilon^n) = \sum_{x^n \in W_\epsilon^n} P(x^n) \geq |W_\epsilon^n| 2^{-n(H+\epsilon)},$$

so that

$$|W_\epsilon^n| \leq 2^{n(H+\epsilon)}.$$

Note that the obtained bound holds for any n. On the other hand, when n is big enough, then $P(W_\epsilon^n) > 1 - \epsilon$. This bound together with the fact that the probability of a weakly ϵ-typical word is at most $2^{-n(H-\epsilon)}$ gives us the estimate

$$1 - \epsilon \leq P(W_\epsilon^n) = \sum_{x^n \in W_\epsilon^n} P(x^n) \leq |W_\epsilon^n| 2^{-n(H-\epsilon)}.$$

From this

$$|W_\epsilon^n| \geq (1 - \epsilon) 2^{n(H-\epsilon)}.$$

■

Therefore, if n is big, the probability of W_ϵ^n is almost one. This means that most likely a realization of $X_1, \ldots, X_n$ is a weakly ϵ-typical word. All weakly typical words have roughly equal probability that usually (that depends on P) is smaller than the maximum possible probability. On the other hand, the proportion of weakly typical words becomes negligible as n grows. Indeed, let $H < \log |\mathcal{X}| < \infty$ (the distribution is not uniform). Then the proportion of weakly typical words tends to zero, since (provided $\epsilon > 0$ is not too big).

$$\frac{|W_\epsilon^n|}{|\mathcal{X}|^n} \leq \frac{2^{n(H+\epsilon)}}{2^{n \log |\mathcal{X}|}} = 2^{n(H+\epsilon-\log|\mathcal{X}|)} \to 0.$$

Example: Let $X_1, \ldots, X_n$ Bernoulli $B(1, p)$. Then

$$P(x^n) = p^k (1-p)^{n-k}, \quad k = \sum_{i=1}^{n} x_i.$$

Therefore

$$-\frac{1}{n} \log P(x^n) = -\frac{k}{n} \log p - \frac{n-k}{n} \log(1-p),$$

so that x^n is weakly typical if the proportion of ones is almost p.

3.1.1 Weak AEP and coding

With weak AEP property it is easy to see that the vector $X_1, \ldots, X_n$ is indeed possible to code such that the expected length per letter L_n is arbitrary close to H provided n is close enough. We shall consider binary codes, extension to $D > 2$ is obvious.

Indeed, let $\mathcal{X}$ be finite so that $X_1, \ldots, X_n$ are i.i.d. random variables on finite alphabet $\mathcal{X}$. Let $\epsilon > 0$ fixed and consider the set of weakly ϵ-typical words W_ϵ^n. Let us order the elements of W_ϵ^n. Since $|W_\epsilon^n| \leq 2^{n(H+\epsilon)}$, then we can represent every index as a binary word with length $\lceil n(H+\epsilon) \rceil \leq n(H+\epsilon)+1$. To every such binary word add prefix 0 to show that the codeword corresponds to a weakly typical word. Hence,

$$l(x^n) \leq n(H+\epsilon)+2, \quad \forall x^n \in W_\epsilon^n.$$

For coding the rest of the words, order them and code their indexes similarly. Since the set of words that are not weakly typical is smaller than $2^{n \log |\mathcal{X}|}$, it takes at most $n \log |\mathcal{X}| + 1$ bits to code each of them. Hence, we can code every non-typical word as a binary word with length $n \log |\mathcal{X}| + 1$ (in fact, we can represent every word like that). For those words, we add prefix 1 (showing that the binary index corresponds to a word that is not weakly typical) and so we obtain the codewords for the set $\mathcal{X} \setminus W_\epsilon^n$. The code is prefix code, since the first bit shows the length of the following codeword. Obviously such a code is not optimal, since most of the words (the ones that are not weakly typical) are coded very roughly.
The expected length of obtained code:

$$\begin{aligned} L &= \sum_{x^n \in \mathcal{X}^n} l(x^n)P(x^n) = \sum_{x^n \in W_\epsilon^n} l(x^n)P(x^n) + \sum_{x^n \notin W_\epsilon^n} l(x^n)P(x^n) \\ &= \sum_{x^n \in W_\epsilon^n} (n(H+\epsilon)+2)P(x^n) + \sum_{x^n \notin W_\epsilon^n} (n \log |\mathcal{X}| + 2)P(x^n) \\ &= P(W_\epsilon^n)(n(H+\epsilon)+2) + (1 - P(W_\epsilon^n))(n \log |\mathcal{X}| + 2). \end{aligned}$$

Thus, when n is big enough, by **2** of Theorem 3.2, it holds $P(W_\epsilon^n) \leq \epsilon$ so that

$$L \leq n(H+\epsilon) + \epsilon(n \log |\mathcal{X}|) + 2 = n(H+\epsilon'),$$

where $\epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$ and choosing ϵ small enough and n big enough, ϵ' can be made arbitrary small.

To recapitulate: For every $\epsilon > 0$ and n big enough

$$H \leq L_n(C) < H + \epsilon, \tag{3.7}$$

where $C : \mathcal{X}^n \to \{0,1\}^*$ is a prefix code based on weak-AEP property as described above.

3.1.2 High probability set

The coding procedure based on weak AEP property works well because for big n, there exists a set W_ϵ^n such that $P(W_\epsilon^n) \geq 1-\epsilon$, but the number of elements in W_ϵ^n is relatively small. However W_ϵ^n is not the smallest (in terms of the number of elements) set having probability at least $1-\epsilon$. Let B_ϵ^n be the smallest set (in terms of number of elements) having the probability $1-\epsilon$. Then above-described coding scheme gives even smaller length. Is the difference essential? From (3.7) it follows that the expected length per letter cannot decrease much. Therefore, $|W_\epsilon^n|$ cannot be much larger than $|B_\epsilon^n|$ and, indeed, as the following lemma shows, also $|B_\epsilon^n| \approx 2^{nH}$.

Lemma 3.1 *For every $1 > \epsilon > 0$ and $\delta > 0$, there exists n such that*

$$|B_\epsilon^n| \geq 2^{n(H-\delta)}. \tag{3.8}$$

Proof. Take $\epsilon_1 > 0$ so small that $\epsilon_1 < \delta$ and $\epsilon_1 + \epsilon < 1$. Let n be so big that (3.3) and (3.4) hold for ϵ_1, in addition let

$$\epsilon_1 - \frac{\log(1-(\epsilon+\epsilon_1))}{n} < \delta. \tag{3.9}$$

Define

$$S := W_{\epsilon_1}^n \cap B_\epsilon^n.$$

By (3.3) and (3.4), it holds ($P(S^c) \leq P(W_{\epsilon_1}^{nc}) + P(B_\epsilon^{nc})$)

$$1-(\epsilon_1+\epsilon) \leq P(S) = \sum_{x^n \in S} P(x^n) \leq |S| 2^{-n(H-\epsilon_1)} \leq |B_\epsilon^n| 2^{-n(H-\epsilon_1)}.$$

Therefore

$$\log|B_\epsilon^n| \geq \log(1-(\epsilon+\epsilon_1)) + n(H-\epsilon_1) = n\big(\frac{\log(1-(\epsilon+\epsilon_1))}{n} + H - \epsilon_1\big) \geq n(H-\delta).$$

Last inequality follows from (3.9). ■

3.1.3 Example

Let $X_1, \ldots, X_{25}$ be i.i.d. with distribution $B(1, 0.1)$. Hence $|\mathcal{X}^n| = 2^{25}$. In the table below, all vectors x^n are distributed into different classes according to the number of ones, denoted via k. The vectors in every class are equiprobable. In the second column is the number of elements in each class and in the third column the sum of $P(x^n)$ in every class – the probability of class. In the last column are the numbers $\frac{1}{n}\log P(x^n)$, where $P(x^n)$ is the probability of every vector in a class (not the class probability).

k	$\binom{n}{k}$	$\binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}P(x^n)$	$-\frac{1}{n}\log P(x^n)$
0	1	0.0717898	0.152003
1	25	0.199416	**0.2788**
2	300	0.265888	**0.405597**
3	2300	0.226497	**0.532394**
4	12650	0.138415	**0.659191**
5	53130	0.0645937	0.785988
6	177100	0.0239236	0.912785
7	480700	0.00721505	1.03958
8	1081575	0.00180376	1.16638
9	2042975	0.000378567	1.29318
10	3268760	0.0000673009	1.41997
11	4457400	0.0000101971	1.54677
12	5200300	1.32185×10^{-6}	1.67357
13	5200300	1.46872×10^{-7}	1.80036
14	4457400	1.39878×10^{-8}	1.92716
15	3268760	≈ 0	2.05396
16	2042975	≈ 0	2.18076
17	1081575	≈ 0	2.30755
18	480700	≈ 0	2.43435
19	177100	≈ 0	2.56115
20	53130	≈ 0	2.68794
21	12650	≈ 0	2.81474
22	2300	≈ 0	2.94154
23	300	≈ 0	3.06833
24	25	≈ 0	3.19513
25	1	≈ 0	3.32193

Take $\epsilon = 0.2$. Since $h(0.1) = 0.468996$, we obtain that the set $W_{0.2}^{25}$ contains all elements in classes $k = 1, 2, 3, 4$. Hence

$$P(W_{0.2}^{25}) = 0.199416 + 0.265888 + 0.226497 + 0.138415 = 0.830216 \geq 1 - \epsilon.$$

On the other hand $|W_{0.2}^{25}| = 25 + 300 + 2300 + 12650 = 15275$, so that

$$\frac{1}{25}\log|W_{0.2}^{25}| \approx 0.556 \in (0.468996 - 0.2, 0.468996 + 0.2)$$

Therefore $W_{0.2}^{25}$ satisfies (3.3) and (3.4).

Let us find B_n^{25}. Since the probabilities are in decreasing order, we collect them starting from the one with biggest probability (consisting only on zeros), then the vectors with one "1" and so on. The total sum of first four classes is 0.7635908, hence all them belong to B_n^{25}. Then we have to take some elements from the fifth class ($k = 4$). The probability of

an element from that class is $\frac{0.138415}{12650} = 0.0000109419$, hence from that class the following number of elements has to be taken:

$$\left\lceil \frac{0.8 - 0.7635908}{0.0000109419} \right\rceil = 3328$$

Thus

$$|B_{0.2}^{25}| = 1 + 25 + 300 + 2300 + 3325 = 5951$$

and

$$\frac{1}{25} \log |B_{0.2}^{25}| \approx 0.501.$$

The sets $B_{0.2}^{25}$ and $W_{0.2}^{25}$ consists of pretty much similar vectors. However, $B_{0.2}^{25}$ is smaller since it contains only some elements from the fifth class, whilst $W_{0.2}^{25}$ contains the whole class.

3.2 Weak joint typicality

Let $P(x, y)$ be a probability distribution on $\mathcal{X} \times \mathcal{Y}$, $(X, Y) \sim P$. Consider i.i.d. random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the pairs are distributed according to P. Then for every pair $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$

$$P(x^n, y^n) = \prod_{i=1}^{n} P(x_i, y_i).$$

Def 3.3 *The set W_ϵ^n consist of pairs $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying the following conditions:*

- $2^{-n(H(X)+\epsilon)} \leq P(x^n) \leq 2^{-n(H(X)-\epsilon)}$
- $2^{-n(H(Y)+\epsilon)} \leq P(y^n) \leq 2^{-n(H(Y)-\epsilon)}$
- $2^{-n(H(X,Y)+\epsilon)} \leq P(x^n, y^n) \leq 2^{-n(H(X,Y)-\epsilon)}$.

The pairs in set W_ϵ^n are called **(weakly) jointly ϵ-typical**.

Hence (x^n, y^n) is jointly typical if both x^n and y^n are weakly typical and the probability of the pair (x^n, y^n) is approximatively $2^{-nH(X,Y)}$.

We shall now prove the two-dimensional counterpart of Theorem 3.2. Let, as previously, $P = P(x, y)$ be a joint distribution on $\mathcal{X} \times \mathcal{Y}$, and let P_x and P_y be its marginal distributions. Then *product measure* $P_x \times P_y$ is a probability measure on $\mathcal{X} \times \mathcal{Y}$ defined as follows

$$P_x \times P_y(x, y) = P_x(x) P_y(y).$$

Hence $P_x \times P_y$ has the same marginal distributions but the joint distribution corresponds to the independence. Let us denote

$$P_x \times P_y(x^n, y^n) := \prod_{i=1}^{n} P_x \times P_y(x_i, y_i).$$

Theorem 3.4 *For every $\epsilon > 0$ the following statements hold:*

1 *If n is big enough, then*

$$P(W_\epsilon^n) > 1 - \epsilon. \tag{3.10}$$

2 *If n is big enough, then*

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H(X,Y)+\epsilon)}. \tag{3.11}$$

3 *If n is big enough, then*

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq P_x \times P_y(W_\epsilon^n) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Proof. Proof follows that of Theorem 3.2.
1: From the weak law of large numbers

$$\begin{aligned}
-\frac{1}{n}\log P(X_1,\ldots,X_n) &= -\frac{1}{n}\sum_{i=1}^n \log P(X_i) \xrightarrow{P} H(X)\\
-\frac{1}{n}\log P(Y_1,\ldots,Y_n) &= -\frac{1}{n}\sum_{i=1}^n \log P(Y_i) \xrightarrow{P} H(Y)\\
-\frac{1}{n}\log P\big((X_1,Y_1),\ldots,(X_n,Y_n)\big) &= -\frac{1}{n}\sum_{i=1}^n \log P(X_i,Y_i) \xrightarrow{P} H(X,Y).
\end{aligned}$$

Proving **1** is now Exercise 1.
2:

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

implying

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |W_\epsilon^n| \leq 2^{n(H(X,Y)+\epsilon)}.$$

3: Applying **2**, we get

$$\begin{aligned}
P_x \times P_y(W_\epsilon^n) &= \sum_{(x^n,y^n)\in W_\epsilon^n} P(x^n)P(y^n)\\
&\leq \sum_{(x^n,y^n)\in W_\epsilon^n} 2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)}\\
&\leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)}\\
&= 2^{-n(I(X;Y)-3\epsilon)}\\
P_x \times P_y(W_\epsilon^n) &\geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)}\\
&= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}.
\end{aligned}$$

■

The interpretation of first two statements of Theorem 3.4 is the the same as in the case of Theorem 3.2: the probability of jointly typical words (pairs) is nearly one, all jointly typical pairs have almost equal probability and the number of those pairs is approximatively $2^{nH(X,Y)}$.
A necessary condition for a pair (x^n, y^n) to be jointly typical is that both words – x^n and y^n – are weakly typical. The number of those pairs, where both words are weakly typical is approximatively $2^{nH(X)}2^{nH(Y)}$, provided n is large enough. On the other hand, in general

$$2^{nH(X,Y)} < 2^{nH(X)}2^{nH(Y)}$$

so that amongst those pairs only a small fraction are jointly typical. To every weakly typical word x^n corresponds roughly

$$2^{n(H(X,Y)-H(X))} = 2^{nH(Y|X)}$$

jointly typical words. Therefore, if a weakly typical x^n in fixed, then choosing randomly a weakly typical y^n, the obtained pair turns out to be jointly typical with probability roughly

$$2^{nH(Y|X)-nH(Y)} = 2^{-nI(X;Y)}.$$

This is actually the third claim of Theorem 3.4: if a pair (x^n, y^n) is chosen randomly (according to P_x and P_y) and the components are chosen independently from each other, then it is jointly typical with probability close to $2^{-nI(X;Y)}$. The bigger $I(X,Y)$, the smaller the probability and the less likely is to get a jointly typical set by choosing the pairs independently. On the other hand, if $I(X;Y) = 0$ (the components are independent), then almost any such randomly chosen pair is jointly typical.

Example: Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and let

$\mathcal{X}\backslash\mathcal{Y}$	1	0
1	$\frac{7}{80}$	$\frac{1}{80}$
0	$\frac{9}{80}$	$\frac{63}{80}$

Thus $X \sim B(1, 0.1)$, $Y \sim B(1, 0.2)$. Joint entropy

$$H(X,Y) = H(X) + H(Y|X) = h(\frac{1}{10}) + h(\frac{7}{8}).$$

The words $x^n = 100000000$ and $y^n = 011000000$ are both weakly typical with respect to any ϵ so that

$$x^n \in W_\epsilon^{10}, \quad y^n \in W_\epsilon^{10}.$$

Denote $p = \frac{1}{10}, q = \frac{1}{8}$ and find

$$P(x^n, y^n) = \Big(\frac{1}{80}\Big)\Big(\frac{9}{80}\Big)^2\Big(\frac{63}{80}\Big)^7 = (pq)((1-p)q)^2((1-p)(1-q))^7 = q^3(1-q)^7(1-p)^9p.$$

$$\begin{aligned}\frac{1}{n}\log P(x^n,y^n) &= \frac{3}{10}\log q + \frac{7}{10}\log(1-q) + \frac{9}{10}\log(1-p) + \frac{1}{10}\log p \\ &= q\log q + \frac{7}{40}\log q - \frac{7}{40}\log(1-q) + (1-q)\log(1-q) + (1-p)\log(1-p) + p\log p \\ &= -h(q) - h(p) + \frac{7}{40}\log(\frac{q}{1-q}).\end{aligned}$$

therefore

$$-\frac{1}{n}\log P(x^n,y^n) - H(X,Y) = \frac{7}{40}\log(7)$$

implying

$$(x^n,y^n) \notin W_\epsilon^{10},$$

when $\epsilon < \frac{7}{40}\log(7)$.

3.3 Weak AEP processes

Weak AEP property (Theorems 3.2 and 3.4)is based on the following property of i.i.d. random variables (i.i.d. process) $X = \{X_n\}_{n=1}^{\infty}$:

$$-\frac{1}{n}\log P(X_1,\ldots,X_n) \to H_X, \quad \text{p.k.}, \tag{3.12}$$

where H_X is the entropy of X_i and, therefore, the entropy rate of i.i.d. process. In the case of i.i.d. process, the convergence (3.12) immediately follows from weak law of large numbers. However, it turns out that (3.12) holds for a large class of stationary processes rather than just i.i.d. process. And then, obviously, all claims of Theorem 3.2 hold (check!).

Def 3.5 *Stochastic process $X_1, X_2 \ldots$ has* **(weak) AEP property**, *if the convergence (3.12), with H_X being the entropy rate, holds.*

All *ergodic processes* have weak AEP property. Like irreducible MC.

3.4 Exercises

1. Prove **1** of Theorem 3.4.

2. Let $X_1, X_2, \ldots$ i.i.d., $X_i \sim P$. Let Q be another distribution on $\mathcal{X}$. Consider the likelihood ratio
$$\frac{Q(X_1)\cdots Q(X_n)}{P(X_1)\cdots P(X_n)}.$$
Prove that there exists a set $A_\epsilon^n \subset \mathcal{X}^n$ and a constant A such that

   1 if $x^n \in A_\epsilon^n$, then
$$2^{-n(A+\epsilon)} \le \frac{Q(x^n)}{P(x^n)} \le 2^{-n(A-\epsilon)};$$

2 if n is big enough, then

$$P(A_\epsilon^n) > 1 - \epsilon;$$

3 if n is big enough, then

$$(1-\epsilon)2^{n(A-\epsilon)} \leq |A_\epsilon^n| \leq 2^{n(A+\epsilon)}.$$

3. Let $X_1, X_2, \ldots$ be stationary MC with finite number of states ($|\mathcal{X}| < \infty$) and transition matrix I (unit). Prove (3.12).

4 Communication through channel

In this section, we briefly consider the communication through discrete (say binary) channel. This goes as follows: the source (message) is encoded using a (say binary) code. The the codewords are transmitted via a channel and the output is decoded. Such a communication system would be perfect, if the channel were noiseless. Unfortunately, this is not the case and the output sequence of the channel can be random (noise is modeled random) but has a distribution that depends on the input sequence. Then the decoded text can differ from the original one and is nothing but an estimate of the original message.

4.1 Discrete channel

Let $\mathcal{X}$ be a finite *input alphabet* and $\mathcal{Y}$ a finite *output alphabet*. In a noisy *memorless* channel, every input character x is transmitted into a output character y with fixed probability $P(y|x)$. The system consisting on $\mathcal{X}$, $\mathcal{Y}$ and the transition matrix

$$\big(P(y|x)\big)_{x\in\mathcal{X},y\in\mathcal{Y}} \tag{4.1}$$

is called **discrete (memorless) channel**.

Channel capacity. Let the channel be fixed and let $P(x)$ be a distribution on input alphabet $\mathcal{X}$, considered as a *input distribution*. With matrix (4.1), we now obtain joint distribution $P(x,y) = P(x)P(y|x)$ on $\mathcal{X}\times\mathcal{Y}$. Let $(X,Y)\sim P(x,y)$ be a random vector with this joint distribution, i.e. X is a random input (with input distribution) and Y is a random output.

Def 4.1 *The* **capacity** *of channel (4.1) is*

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

where maximum is taken over all possible input distributions on $\mathcal{X}$.

Remarks:

- It is not hard to see that when transition matrix is fixed, then the function $P(x)\to I(X;Y)$ is a concave function. Since input alphabet is finite, it is a convex function over closed convex set (simplex) in finite-dimensional space. Such a function is always continuous, hence the maximum always exists. This justifies the use of maximum instead of supremum in the definition of capacity.

- The capacity of channel can be interpreted as the maximum amount of information which can be sent through the channel. Note that the following inequality holds

  $$C \le \log\min\{|\mathcal{X}|,|\mathcal{Y}|\}.$$

  Indeed:

  $$C = \max_{P(x)} I(X;Y) \le \max_{P(x)} H(X) \le \log|\mathcal{X}|, \quad C = \max_{P(x)} I(X;Y) \le \max_{P(x)} H(Y) \le \log|\mathcal{Y}|.$$

4.2 Examples of channels

Noiseless binary channel. Here $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and the transition matrix $P(y|x)$ is unit matrix. By this channel every transmitted bit is received without error. Thus by every transmission only one error-free bit can be transmitted and the capacity of channel is also 1. Indeed, $I(X;Y) = H(X;X) = H(X)$ so that

$$C = \max_{P(x)} H(X) = 1,$$

where the maximum is achieved by using $B(1, \frac{1}{2})$ as input distribution. Note that by inequality $C \leq \log \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, 1 is the maximum possible channel capacity for every channel with binary input alphabet.

Noisy channel with non-overlapping outputs. By this channel $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, 2, 3\}$ and the transition matrix is

$$\begin{pmatrix} p & 1-p & 0 & 0 \\ 0 & 0 & q & 1-q \end{pmatrix}$$

Although the channel has noise, every input can be determined from output so the noise really does not matter. The capacity of this cannel, obviously, is also one bit per transmission, i.e. $C = 1$. Formally,

$$C = \max_{P(x)} H(X) - H(X|Y) = \max_{P(x)} H(X) = 1,$$

because $X = f(Y)$ and therefore $H(X|Y) = 0$. Thus the input distribution achieving the maximum is again uniform over two input letters.

Noisy keyboard (typewriter). Here $\mathcal{X} = \mathcal{Y}$ is (English) alphabet so $|\mathcal{X}| = 26$. By noisy keyboard, every letter is transmitted correctly with probability 0.5, but with the same probability an input letter is transmitted into next letter.
The capacity

$$C = \max_{P(x)} \big(H(Y) - H(Y|X)\big) = \max_{P(x)} H(Y) - 1 = \log 26 - 1 = \log 13,$$

where the maximum is achieved using uniform input alphabet. The obtained capacity matches with intuition – half of the letters (13) can be transmitted without errors.

Binary symmetric channel. Here $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and the transition matrix is

$$\begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

The input symbol is transmitted correctly with probability $1-p$, but with probability p it is transmitted to another symbol. Thus an output 0 can correspond to input 0 or to input 1. Let, for any input X find the mutual information

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x P(x)H(Y|X=x)$$
$$= H(Y) - \sum_x P(x)h(p) = H(Y) - h(p).$$

Hence $I(X;Y)$ is maximum if Y has uniform distribution. This is achieved when X has uniform distribution. Therefore,

$$C = \max_{P(x)} I(X;Y) = 1 - h(p).$$

In case $p = 0$, the channel is noiseless and its capacity is 1. If $p = 0.5$, then X and Y are independent. Then the channel allows no communication and its capacity is, obviously, equal to 0.

J. Thomas and T. Cover: "This is the simplest model of a channel with errors; yet it captures most of the complexity of the general problem".

Binary erasure channel. Here $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,e\}$. The character e can be interpreted as a sign that the input character is erased. Both input characters are erased with the same probability and the receiver knows which bits have been erased. Transition matrix

$$P(x|x) = 1-p, \quad P(e|x) = p, \quad x = 0,1.$$

Let us find the capacity

$$C = \max_{P(x)}\big(H(Y) - H(Y|X)\big) = \max_{P(x)} H(Y) - h(p).$$

To find $\max_{P(x)} H(Y)$, let us define $E = \{Y = e\}$. Since $E = f(Y)$, then for any input distribution $P(x)$

$$H(Y) = H(Y,E) = H(E) + H(Y|E) = h(p) + H(Y|E).$$

Let $\pi = \mathbf{P}(X=1)$. Then $\mathbf{P}(Y=1|Y\neq e) = \pi$ and $\mathbf{P}(Y=0|Y\neq e) = (1-\pi)$ and

$$H(Y|E) = H(Y|Y\neq e)\mathbf{P}(Y\neq e) = h(\pi)(1-p).$$

Therefore

$$C = \max_{P(x)} H(Y|E) = \max_{\pi} h(\pi)(1-p) = 1-p.$$

The capacity $1-p$ matches with intuition: in average, a proportion p of all input bits are erased and $1-p$ of them are transmitted correctly.

Symmetric channel. Channel is symmetric if all rows in transition matrix are permutations of each others and all columns are permutations of each other. The following channels are symmetric:

$$\begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{pmatrix} \quad \begin{pmatrix} 0.2 & 0.2 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.2 & 0.2 \end{pmatrix}.$$

Capacity is easy to find. Let H_r be the entropy of a row. Then

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H_r \leq \log|\mathcal{Y}| - H_r,$$

where the equality holds if the output distribution is uniform. Let us see that uniform output distribution holds for uniform input distribution. Indeed, if input distribution is uniform, then

$$P(y) = \sum_{x\in\mathcal{X}} P(y|x)P(x) = \frac{1}{|\mathcal{X}|}\sum_{x} P(y|x) = \frac{c}{|\mathcal{X}|},$$

where c is the sum of columns. Hence $P(y)$ is independent of y and, therefore, the output is uniform and

$$C = \log|\mathcal{Y}| - H_r.$$

The derivation above holds also when the rows of transition matrix are permutations from each others and the sum of columns are constant (but columns might not be permutations from each other). Such channels are called *weakly symmetric*. The following channel is weakly symmetric but not symmetric:

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix}.$$

J. Thomas and T. Cover: "In general, there are no closed form solution for the capacity. but for many simple channels it is possible to calculate the capacity using properties like symmetry."

4.3 The channel coding theorem

4.3.1 (M,n)-code

Let $\{1,2,\ldots,M\}$ be the index set of a vocabulary. A random word W is drawn from the index set. Using a fixed length block code

$$\mathcal{C} : \{1,2,\ldots,M\} \mapsto \mathcal{X}^n,$$

the message W is encoded, yielding a codeword $X^n(W)$. The codeword is an n-elemental random vector that is sent componentwise through the channel

$$\{P(y|x)\}_{x\in\mathcal{X},y\in\mathcal{Y}}.$$

Since the channel is memoryless, the probability of receiving the output y^n given input x^n is

$$P(y^n|x^n) = \prod_{i=1}^{n} P(y_i|x_i).$$

The output is then a random vector n that is decoded using a decoding function

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

After decoding, we obtain the index estimate $\hat{W} = g(Y^n)$ that is not necessarily the original word W.

Def 4.2 *Let $\{P(y|x)\}_{x\in\mathcal{X}, y\in\mathcal{Y}}$ be a discrete memoryless channel. An* **(M, n) code** *for the channel consists of the following:*

- *An index set $\{1, \ldots, M\}$.*
- *An encoding function*
$$\mathcal{C} : \{1, \ldots, M\} \to \mathcal{X}^n.$$
The set of codewords $\mathcal{C}(1), \ldots, \mathcal{C}(M)$ is called the codebook.
- *A decoding function*
$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}.$$

Error probabilities. Let λ_i be the conditional probability of error of (M, n) code given that the index i was sent. Thus

$$\lambda_i := \mathbf{P}(\hat{W} \neq i | W = i) = \mathbf{P}\big(g(Y^n) \neq i | W = i\big) = \sum_{y^n:\, g(y^n)\neq i} P(y^n|\mathcal{C}(i)).$$

Let

$$\lambda_{max} := \max_i \lambda_i$$

and let P_e be the error of mistake provided that the distribution of W us uniform on $\{1, \ldots, M\}$. Thus

$$P_e = \mathbf{P}(\hat{W} \neq W) = \sum_i \mathbf{P}(\hat{W} \neq i | W = i)\mathbf{P}(W = i) = \frac{1}{M}\sum_i \mathbf{P}(\hat{W} \neq i | W = i) = \frac{1}{M}\sum_i \lambda_i.$$

Obviously

$$P_e \leq \lambda_{max}.$$

Rate of (M, n) code.

Def 4.3 *The* **rate of an (M, n) code** *is*

$$R := \frac{\log M}{n}.$$

The rate of an (M, n) code measures the (maximal) proportion of information per single transmission. Indeed, suppose W has uniform distribution. Then $H(W) = \log M$ so that $\log M$ is the (maximal) amount of information contained in W. Every word (index) is represented as n dimensional codeword. Thus, the amount of information per one transmission is the rate of the code.
Formally the rate is only a property of (M, n) code and one aims to design the code so that the rate were as big as possible. On the other hand, in order the communication to be meaningful, the rate cannot be arbitrary small. Indeed, if $|\mathcal{X}| = 2$, then the smallest codeword length for fixed-length non-singular code $\mathcal{C}$ is $\lceil \log M \rceil$. Thus, for any meaningful (M, n) code (with non-singular code $\mathcal{C}$) the rate cannot be smaller than 1. Whether a code is useful or not depends on the channel – one looks for a code such that the error probability were as small as possible. And that cannot be achieved using codes with very high rate.

Example: Consider the case $|\mathcal{X}| = 2$ and the code with codeword lengths $\lceil \log M \rceil$. Let call this code *uniform*. When the channel is noiseless, then uniform code works just fine and $\lambda_{max} = 0$. However, using the code with binary symmetric channel the error probability λ_{max} increases with n:

$$1 - \lambda_i = \mathbf{P}(\hat{W} = i | W = i) = \mathbf{P}(Y^n = \mathcal{C}(i)) = (1 - p)^n.$$

Although the rate of the code is high, it is not useful. For that channel, the first obvious solution seems to be so-called *repetition code*: every bit in uniform code is repeated m times. The length of every codeword is then $\lceil \log M \rceil m$. If m is large enough and $p < 0.5$, then by LLN, majority amongst m received bits should be the right one. Thus decoding procedure is to decode any m-block as the majority of received bits (to avoid ties take m odd). For every given $\epsilon > 0$ one can find m long enough (depends on M) such that $\lambda_{max} < \epsilon$. The rate of this code is about $\frac{1}{m}$.

Def 4.4 *Let $P(y|x)$ be a discrete memoryless channel. A rate $R > 0$ is said to be* **achievable**, *if there exists a sequence of $(\ulcorner 2^{nR} \urcorner, n)$ codes such that $\lambda_{max} \to 0$ as $n \to \infty$.*

Whether R is achievable or not depends on the channel. If R is achievable, then for every $\epsilon > 0$, there is a n and a $(\ulcorner 2^{nR} \urcorner, n)$ code such that $\lambda_{max} < \epsilon$. When $\lambda_{max} < \epsilon$, then for *any distribution* of W, the probability of error is at most ϵ.

NB! In what follows, we shall denote $\ulcorner 2^{nR} \urcorner$ by 2^{nR}.

4.3.2 Channel coding theorem

The following theorem, sometimes called *Shannon second theorem* is a central result of information theory.

Theorem 4.5 (Channel coding theorem) *Let C be the capacity of a channel. Then every rate R satisfying $R < C$ is achievable, i.e. for every R there exists a sequence of $(2^{nR}, n)$ codes so that $\lambda_{max} \to 0$ as n grows.*
Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda_{max} \to 0$ must have $R \leq C$.

About the proof of the first claim. The proof is non-constructive: the code is constructed randomly. Then it is proved that in average the random code works well. Then there must be ta least one non-random code that must work also well.

More precisely: let $R < C$. A random $(2^{nR}, n)$ code is generated as follows.

1. Fix input distribution $P(x)$ that satisfies, $I(X;Y) = C$. This distribution as well as channel $\{P(y|x)\}$ are known to receiver (recall $P(x)$ depends on channel, only).

2. Generate 2^{nR} random n-dimensional vectors, each of them is i.i.d vector with components distributed as $P(x)$. Obtained words $x^n(1), \ldots, x^n(2^{nR})$ form (random) codebook:
$$\mathcal{C} : \{1, \ldots, 2^{nR}\} \to \mathcal{X}^n, \quad \mathcal{C}(i) = x^n(i).$$
This code is revealed to both sender and receiver.

3. A message W is chosen from $\{1, \ldots, 2^{nR}\}$ according to a uniform distribution.

4. The chosen word w is encoded and the corresponding codeword $x^n(w)$ is sent over the channel.

5. The receiver receives a (random) sequence Y^n according to the distribution
$$P\big(y^n|x^n(w)\big) = \prod_i^n P\big(y_i|x_i(w)\big).$$

6. Receiver decodes obtained word y^n according to the rule:
$$g(y^n) = \begin{cases} k & \text{if } (x^n(k), y^n) \in W_\epsilon^n \text{ and for every } i \neq k,\ (x^n(i), y^n) \notin W_\epsilon^n, \\ * & \text{else.} \end{cases}$$
Since $* \notin \mathcal{Y}$, the output $*$ is always a mistake. Here $\epsilon > 0$ is so small that $C - R - 3\epsilon > 0$ and W_ϵ^n is the set of jointly typical words. The receiver knows $P(x)P(y|x)$, hence he also knows the set W_ϵ^n.

With the help of AEP (Theorem 3.4), it is possible to show that the average error made by this procedure (over all possible codes) is smaller than 2ϵ, provided n is big enough. Thus, for n big enough

$$\sum_{\mathcal{C}} P(\mathcal{C})P_e(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C})\frac{1}{2^{nR}}\sum_{j}^{2^{nR}} \lambda_j(\mathcal{C}) \leq 2\epsilon,$$

where $P(\mathcal{C})$ is the probability of obtaining a particular code $\mathcal{C}$. Since the average probability of error is smaller than 2ϵ, there must be at least one deterministic code $\mathcal{C}^*$ so that

$$P_e(\mathcal{C}^*) = \frac{1}{2^{nR}}\sum_{i}^{2^{nR}} \lambda_i \leq 2\epsilon,$$

where $\lambda_i := \lambda_i(\mathcal{C}^*)$. From the inequality above, it follows that there exist at least 2^{nR-1} indexes i so that $\lambda_i \leq 4\epsilon$. Indeed, if not (i.e. the number of indexes i satisfying $\lambda_i > 4\epsilon$ would be at least $2^{nR-1}+1$), then $\sum_i^{2^{nR}} \lambda_i > 4\epsilon(2^{nR-1}+1) > 2\epsilon 2^{nR}$. Hence the best half of the codewords have maximal probability of error less than 4ϵ. We keep these codewords, only. With this (reduced) code it is possible to encode at least

$$2^{nR-1} = 2^{n(R-\frac{1}{n})}$$

words. Hence we have $(2^{n(R-\frac{1}{n})}, n)$ code such that $\lambda_{max} \leq 4\epsilon$. The rate drops from R to $R - \frac{1}{n}$, which is negligible for large n. Thus every rate R so that $R < C$ is achievable.

Remarks:

- The intuition behind the proof: a random codeword x^n is weakly typical with high probability. Then the output y^n is with high probability one of these vectors that are jointly typical with x^n. Given x^n, there are in average about $2^{nH(Y|X)}$ such outputs. The decoding procedure works if the jointly typical outputs corresponding to different codewords x^n form disjoint classes with about $2^{nH(Y|X)}$ elements in each class. Since the number of weakly typical outputs is about $2^{nH(Y)}$, it means that the number if classes must be about

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

  This is the upper limit to the number of codewords and, hence, to M.

- The proof does not provide a way of constructing the best codes. One can, obviously, find the maximal probability of error for any particular code. But to find the best $(2^{nR}, n)$ code, all possible $|\mathcal{X}|^{n2^{nR}}$ codes need to be checked and that is impossible. It is also possible to generate the code random as suggested in the proof. Such a code is likely to be good for long block lengths. The problem is decoding. Indeed, without some structure in the code, the only possibility seems to be the table lookup.

But the table is as large as $n \times 2^{Rn}$, so that method is impractical. Hence, theorem does not provide any practical coding scheme. However, it indicate when a good scheme is possible.

In practice: *turbo codes*, *parity check codes*, *error-correcting codes*, *Lempel-Ziv codes* and many more.

J. Thomas and T. Cover: "Ever since Shannon's original paper on information theory, researches have tried to develop structural codes that are easy to encode and decode. So far, they have developed many codes with interesting and useful structures, but the asymptotic rates of these codes are not yet near capacity."

4.3.3 The proof of second claim

Lemma 4.1 *Let $X^n = \mathcal{C}(W)$ random codeword and let $Y^n = (Y_1, \ldots, Y_n)$ be its output. Then*

$$I(X^n; Y^n) \leq nC,$$

where C is the channel capacity.

Proof. Chain rule

$$H(Y^n|X^n) = H(Y_1|X^n) + H(Y_2|Y_1, X^n) + \cdots + H(Y_n|Y_1, \ldots, Y_{n-1}, X^n).$$

By definition

$$H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) = -\sum_{y_i, y^{i-1}, x^n} \log P(y_i|y_1, \ldots, y_{i-1}, x_1, \ldots, x_n) P(y_1, \ldots, y_i, x_1, \ldots, x_n).$$

The channel is memoryless, i.e. for every i

$$P(y_i|y_1, \ldots, y_{i-1}, x_1, \ldots, x_n) = P(y_i|x_i)$$

and

$$P(y_1, \ldots, y_i, x_1, \ldots, x_n) = P(y_i|x_i) P(y_1, \ldots, y_{i-1}, x_1, \ldots, x_n),$$

so that

$$H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) = H(Y_i|X_i).$$

Thus

$$H(Y^n|X^n) = \sum_{i=1}^{n} H(Y_i|X_i), \tag{4.2}$$

implying that

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} (H(Y_i) - H(Y_i|X_i)) = \sum_{i=1}^{n} I(X_i; Y_i) \leq nC. \end{aligned}$$

■