Literature:

12. · · ·
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

\[
p_i := P(X = x_i) = P(x_i).
\]

Thus, for every \( A \subset \mathcal{X} \)

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

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

\[
Eg(X) = \sum p_i g(x_i) = \sum_{x \in \mathcal{X}} g(x) P(x) = \sum_{x \in \mathcal{X}_P} g(x) P(x).
\]  

\(1.1\)

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)

\[
H(X) = E(-\log P(X)) = E \log \frac{1}{P(X)}.
\]

- 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 \(X\), 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.
\[
H_b(X) = -\sum p_i \log_b p_i = -\sum_{x \in X} P(x) \log_b P(x).
\]
since \(\log_b p = \log_a a \log_a p\), it holds
\[
H_b(X) = (\log_a a)H_a(X),
\]
so that \(H_b(X) = (\log_a 2)H_a(X)\) and \(H_e(X) = (\ln 2)H(X)\). In information theory, typically, \(\log_2\) is used and such entropy is measured in \textit{bits}. The entropy defined with \(\ln\) is measured in \textit{nats}, the entropy defined with \(\log_{10}\) is measured in \textit{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 \(X = \{0, 1\}, p = P(X = 1)\), 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 \textbf{binary entropy function}. The function \(h(p)\) is concave, symmetric around \(\frac{1}{2}\) and has maximum at \(p = \frac{1}{2}\):
\[
h\left(\frac{1}{2}\right) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = \log 2 = 1.
\]

2 Consider the distributions
\[
P: \begin{array}{c|c|c|c|c|c}
a & b & c & d & e \\
\hline
\frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \\
\end{array}
\quad Q: \begin{array}{c|c|c|c|c}
a & b & c & d \\
\hline
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\end{array}
\]
\[
H(P) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{16} \log \frac{1}{16} - \frac{1}{16} \log \frac{1}{16} = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \frac{4}{16} = \frac{15}{8}
\]
\[
H(Q) = \log 4 = 2.
\]
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

\[ H(p_1, p_2, p_3, \ldots) = H\left(\sum_{i=1}^{k} p_i, p_{k+1}, p_{k+2}, \ldots\right) + \left(\sum_{i=1}^{k} p_i\right) H\left(\frac{p_1}{\sum_{i=1}^{k} p_i}, \ldots, \frac{p_k}{\sum_{i=1}^{k} p_i}\right). \]  \hfill (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 \( X \) 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.

\[ \mathcal{P}^m := \left\{ (p_1, \ldots, p_m) : p_i \geq 0, \sum_{i=1}^{m} p_i = 1 \right\}. \]

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 \( 1 < k < m \),

\[ f_m(p_1, p_2, \ldots, p_m) = f_{m-k+1}(\sum_{i=1}^{k} p_i, p_{k+1}, \ldots, p_m) + \left(\sum_{i=1}^{k} p_i\right) f_k\left(\frac{p_1}{\sum_{i=1}^{k} p_i}, \ldots, \frac{p_k}{\sum_{i=1}^{k} p_i}\right). \]

- **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 \( m \),

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

**Proof.** Let, for every \( m \),

\[ g(m) := f_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right). \]

By symmetry and applying **A3** \( m \) times, we obtain

\[ g(mn) = f_{nm}\left(\frac{1}{mn}, \ldots, \frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}\right) \]

\[ = f_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) + f_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = g(m) + g(n). \]

5
Hence, for integers \( n \) and \( k \), \( g(n^k) = kg(n) \) and by A1, \( g(2^k) = kg(2) = k \) i.e. \( g(2^k) = \log(2^k), \; \forall k. \)

Using A4, it is possible to show that the equality above holds for every integer \( n \), i.e. 
\[
g(n) = \log n, \; \forall n \in \mathbb{N}.
\]

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,
\[
g(n) = f_m\left(\frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)
\]
\[
= f_m\left(\frac{k_1}{n}, \ldots, \frac{k_m}{n}\right) + \sum_{i=1}^{m} \frac{k_i}{n} f_{k_i}\left(\frac{1}{k_i}, \ldots, \frac{1}{k_i}\right)
\]
\[
= f_m(p_1, \ldots, p_m) + \sum_{i=1}^{m} \frac{k_i}{n} g(k_i) = f_m(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i \log(k_i).
\]

Therefore,
\[
f_m(p_1, \ldots, p_m) = \log(n) - \sum_{i=1}^{m} p_i \log(k_i) = - \sum_{i=1}^{m} p_i \log\left(\frac{k_i}{n}\right) = - \sum_{i=1}^{m} p_i \log p_i
\]
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 
\[
Eg(X) \geq g(EX).
\]

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

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

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

$$-\lambda P_1(x) \log P_1(x) - (1 - \lambda) P_2(x) \log P_2(x) = \lambda f(P_1(x)) + (1 - \lambda) f(P_2(x))$$

$$\leq f(\lambda P_1(x) + (1 - \lambda) P_2(x)) = -Q(x) \log Q(x).$$

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

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

Joint distribution is often represented by the following table
When all random variables are independent, then expectation is linear.

\[ E[x] = E[y] = \cdots = E[z] = \cdots \sum \]

\[ P(x_1, y_1) = \sum_{x_2} P(x_2, y_1) = \cdots = \sum_{x_n} P(x_n, y_1) = \cdots \sum \]

\[ P(x_1, y_2) = \sum_{x_2} P(x_2, y_2) = \cdots = \sum_{x_n} P(x_n, y_2) = \cdots \sum \]

\[ p_{ij} = \sum_{x_1} p_{1j} = \cdots = \sum_{x_n} p_{nj} = \cdots \sum \]

\[ \sum_{i} p_{ii} = P(y_1) \sum_{i} p_{ij} = P(y_j) \cdots 1 \]

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

\[ P(x) := P(X = x) \quad \text{and} \quad P(y) := P(Y = y) \]

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

\[ P(x, y) = P(x)P(y) \quad \forall x \in X, y \in Y. \]

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

\[ H(X, Y) = - \sum_{i,j} p_{ij} \log p_{ij} = - \sum_{(x,y) \in X \times Y} P(x,y) \log P(x,y) = E \left( - \log P(X,Y) \right). \quad (1.6) \]

**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

\[ H(X, Y) = - \sum_{(x,y) \in X \times Y} P(x,y) \log P(x,y) = - \sum_{x \in X} \sum_{y \in Y} P(x)P(y)(\log P(x) + \log P(y)) \]

\[ = - \sum_{x \in X} P(x) \log P(x) - \sum_{y \in Y} P(y) \log P(y) = H(X) + H(Y). \]

The argument above can be restate as follows. For every \( x \in X \) and \( y \in Y \) 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

\[ H(X, Y) = -E \left( \log P(X,Y) \right) = -E \left( \log P(X) + \log P(Y) \right) \]

\[ = -E \log P(X) - E \log P(Y) = H(X) + H(Y). \]

**The joint entropy of several random variables.** By analogy, the joint entropy of several random variables \( X_1, \ldots, X_n \) is defined

\[ H(X_1, \ldots, X_n) := -E \log P(X_1, \ldots, X_n). \]

When all random variables are independent, then

\[ H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i). \]
### 1.3 Conditional entropy

#### 1.3.1 Definition

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

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

The conditional distribution of \( Y \) given \( X = x \) is

\[
\begin{array}{c|c|c|c}
  y_1 & y_2 & y_3 & \cdots \\
  P(y_1|x) & P(y_2|x) & P(y_3|x) & \cdots
\end{array}
\]

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

\[
\begin{array}{c|c|c|c}
  H(Y|x_1) & H(Y|x_2) & H(Y|x_3) & \cdots \\
  P(x_1) & P(x_2) & P(x_3) & \cdots
\end{array}
\]

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

\[
\begin{align*}
H(Y|X) := & \sum_{x \in \mathcal{X}_P} H(Y|x)P(x) = - \sum_{x \in \mathcal{X}_P} P(x) \sum_{y \in \mathcal{Y}} \log P(y|x)P(y|x) \\
= & - \sum_{x \in \mathcal{X}_P} \sum_{y \in \mathcal{Y}} \log P(y|x)P(x, y) = -E\left( \log P(Y|X) \right).
\end{align*}
\]

**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 \( P(Y = f(x)|X = x) = 1 \) 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

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

$$= -\sum_{(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}} \log P(x, y|z) P(x, y, z) = -E \log P(X, Y|Z)$$

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

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

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

$$H(X_1, \ldots, X_n) = -E \log P(X_1, \ldots, X_n)$$

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

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

\[ \blacksquare \]

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 \((X, Y, Z)\)
\[
\]

1.4 Kullback-Leibler distance

1.4.1 Definition
NB! In what follows,
\[
0 \log(0) := 0, \text{ if } q \geq 0 \text{ and } p \log(p) := \infty \text{ if } p > 0.
\]

Def 1.6 Let \(P\) and \(Q\) two distributions on \(X\). The **Kullback-Leibler distance** (Kullback-Leibler divergence, relative entropy, informational divergence) between probability distributions \(P\) and \(Q\) is defined as
\[
D(P||Q) := \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}.
\]
Where \(X \sim P\), then
\[
D(P||Q) = E \left( \log \frac{P(X)}{Q(X)} \right).
\]
When \(X \sim P\) and \(Y \sim Q\), then
\[
D(X||Y) := D(P||Q).
\]

Def 1.7 Let, for any \(x \in X\), \(P(y|x)\) and \(Q(y|x)\) be two (conditional) probability distributions on \(Y\). 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

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

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

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

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 $\infty$.

- 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

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

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

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

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

$$D(P||Q) = \frac{1}{2} \log \left( \frac{1}{2q} \right) + \frac{1}{2} \log \left( \frac{1}{2(1-q)} \right) = -\frac{1}{2} \log (4q(1-q)) \to \infty, \text{ if } q \to 0$$  

$$D(Q||P) = q \log (2q) + (1-q) \log(2(1-q)) \to 1 \text{ if } q \to 0.$$  

12
1.4.2 K-L distance is non-negative: Gibbs inequality and its consequences

**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 $X$ 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 exist

$$E|g(Y)| = \sum_{x \in X} |\log \frac{Q(x)}{P(x)}| P(x) = \sum_{x \in X} |\log \frac{P(x)}{Q(x)}| P(x) < \infty,$$

$$E|Y| = EY = \sum_{x \in X} \frac{Q(x)}{P(x)} P(x) = 1.$$

By Jensen’s inequality

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

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

**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

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

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

**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

$$0 \leq \sum a_i' \log \frac{a_i'}{b_i'} = \sum a_i \log \frac{\sum_j a_j}{\sum_j b_j} = \frac{1}{\sum_j a_j} \left[ \sum a_i \log \frac{a_i}{b_i} - (\sum a_i) \log \frac{\sum a_i}{\sum b_i} \right].$$

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)

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

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 $B \subseteq \mathcal{X}$, $P_n(B) \to P(B)$. 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

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

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

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

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

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

**Corollary 1.3**

$$D(\lambda P_1 + (1 - \lambda)P_2 || \lambda Q_1 + (1 - \lambda)Q_2) \leq \lambda D(P_1 || Q_1) + (1 - \lambda)D(P_2 || Q_2). \quad (1.11)$$

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

$$\lambda P_1(x) \log \frac{\lambda P_1(x)}{\lambda Q_1(x)} + (1 - \lambda)P_2(x) \log \frac{(1 - \lambda)P_2(x)}{(1 - \lambda)Q_2(x)}$$

$$\geq \left( \lambda P_1(x) + (1 - \lambda)P_2(x) \right) \log \frac{\lambda P_1(x) + (1 - \lambda)P_2(x)}{\lambda Q_1(x) + (1 - \lambda)Q_2(x)}.$$

Sum over $\mathcal{X}$. \qed

Take $Q_1 = Q_2 = Q$. Then from (5.31), 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:

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

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

$$D(P_1(y|x) || P_2(y|x)) \geq D(P_1 || P_2). \quad (1.13)$$

Proof of (1.13) is Exercise 16.
1.5 Mutual information

Let \((X, Y)\) be random vector with distribution \(P(x, y), (x, y) \in X \times Y\). 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 \(P(x)P(y)\)

\[
I(X; Y) := \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} = D(P(x, y)||P(x)P(y)) = E \left( \log \frac{P(X, Y)}{P(X)P(Y)} \right).
\]

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 \(P(x, y)\).
- \(0 \leq I(X; Y)\).
- mutual information is symmetric \(I(X; Y) = I(Y; X)\).
- \(I(X; Y) = 0\) iff \(X, Y\) are independent.
- The following relation is important:

\[
I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).
\] (1.14)

For the proof, note

\[
I(X; Y) = E \log \frac{P(X, Y)}{P(X)P(Y)} = E \log \frac{P(X|Y)P(Y)}{P(X)P(Y)} = E \log \frac{P(X|Y)}{P(X)}
\]

\[
= E \log P(X|Y) - E \log P(X) = H(X) - H(X|Y).
\]

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*. 

16
• Recall chain rule: $H(X|Y) = H(X, Y) - H(Y)$. Hence

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$  \hfill (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!)


<table>
<thead>
<tr>
<th>Y \setminus X</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>v</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

• 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 $Z$ be the support of $Z$.

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

$$I(X; Y|Z) := H(X|Z) - H(X|Y, Z) = E \log \frac{P(X|Y, Z)}{P(X|Z)}$$

$$= E \log \frac{P(X|Y, Z)P(Y|Z)}{P(X|Z)P(Y|Z)} = E \log \frac{P(X, Y|Z)}{P(X|Z)P(Y|Z)}$$

$$= \sum_{x, y, z} P(x, y, z) \log \frac{P(x, y|z)}{P(x|z)P(y|z)}$$

$$= \sum_z P(z) \sum_{y, x} P(x, y|z) \log \frac{P(x, y|z)}{P(x|z)P(y|z)}$$

$$= \sum_z D(P(x, y|z)||P(x|z)P(y|z)) P(z).$$
Properties:

• 
  \[ I(X; Y|Z) \geq 0, \]
  with equality iff \( X \) and \( Y \) are conditionally independent:
  \[
P(x, y|z) = P(x|z)P(y|z), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}.
  \]

For proof note that \( I(X; Y|Z) = 0 \) iff for every \( z \in \mathcal{Z} \), it holds
  \[
  D \left( P(x, y|z) \left\| P(x,z)P(y|z) \right. \right) = 0.
  \]
This means conditional independence.

• The proof of following equalities is Exercise 18
  \[
  \begin{align*}
  I(X; X|Z) & = H(X|Z) \\
  I(X; Y|Z) & = H(Y|Z) - H(Y|X, Z) \\
  I(X; Y|Z) & = H(X|Z) + H(Y|Z) - H(X, Y|Z).
  \end{align*}
  \]

In addition, the following equality holds
  \[
  I(X; Y|Z) = H(X; Z) + H(Y; Z) - H(X, Y, Z) - H(Z).
  \] (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:
  \[
  I(X_1, \ldots, X_n; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n|Y) \\
  = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \ldots, X_{n-1}) \\
  - H(X_1|Y) - H(X_2|X_1, Y) - \cdots - H(X_n|X_1, \ldots, X_{n-1}, Y).
  \]

• 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

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

be the probability of mistake made by estimation. If $P_e = 0$, then $X = \hat{X}$ a.s. so that $H(X|X) = 0$. 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

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

(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:

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

(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

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

$$+ \sum_{x \in \mathcal{X}} P(\hat{X} = x, E = 0) H(X|\hat{X} = x, E = 0).$$

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

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

If $E = 1$ and $\hat{X} = x$, then $X \in \mathcal{X}\setminus x$, 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

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

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 \(P_e = P(X > 0) = P(X = Y) = P(Z = 1) = p\). 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 \(H(X|\hat{X}) \neq 0\), when \(P_e \searrow 0\).

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

\[ H(X|\hat{X} = x, E = 1) = \log(|\mathcal{X}| - 1) \quad (1.20) \]

and

\[ H(E|\hat{X}) = H(E). \quad (1.21) \]

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

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

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

<table>
<thead>
<tr>
<th>(\hat{X}\backslash X)</th>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(\cdots)</th>
<th>(x_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_1)</td>
<td>(P(\hat{X} = x_1, X = x_1))</td>
<td>(P(\hat{X} = x_1, X = x_2))</td>
<td>(\cdots)</td>
<td>(P(\hat{X} = x_1, X = x_n))</td>
</tr>
<tr>
<td>(x_2)</td>
<td>(P(\hat{X} = x_2, X = x_1))</td>
<td>(P(\hat{X} = x_2, X = x_2))</td>
<td>(\cdots)</td>
<td>(P(\hat{X} = x_2, X = x_n))</td>
</tr>
<tr>
<td>(\cdots)</td>
<td>(\cdots)</td>
<td>(\cdots)</td>
<td>(\cdots)</td>
<td>(\cdots)</td>
</tr>
<tr>
<td>(x_n)</td>
<td>(P(\hat{X} = x_n, X = x_1))</td>
<td>(\cdots)</td>
<td>(\cdots)</td>
<td>(P(\hat{X} = x_n, X = x_n))</td>
</tr>
</tbody>
</table>

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,

\[
\begin{array}{c|ccc}
\hat{X} \setminus \mathcal{X} & a & b & c \\
\hline
a & \frac{3}{10} & \frac{1}{10} & \frac{1}{10} \\
b & \frac{1}{25} & \frac{3}{25} & \frac{1}{25} \\
c & \frac{3}{50} & \frac{3}{50} & \frac{1}{50}
\end{array}
\]

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 Markov 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 $m = 2, \ldots, n - 1$

\[
P(X_{m+1} = x_{m+1}|X_m = x_m, \ldots, X_1 = x_1) = P(X_{m+1} = x_{m+1}|X_m = x_m).
\]

(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

\[X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n.\]

Thus $X \rightarrow Y \rightarrow Z$ 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 \( X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \), then \( X_n \rightarrow X_{n-1} \rightarrow \cdots \rightarrow X_1 \) (reversed MC is also a MC).

- Every sub-chain Markov chain is a Markov chain: if \( X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \), then \( X_{n_1} \rightarrow X_{n_2} \rightarrow \cdots \rightarrow X_{n_k} \).

- If \( X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \), then for every \( m < n \) and \( x_i \in X_i \)

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

- \( X_1 \rightarrow \cdots \rightarrow X_n \) 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 \( x_m \in X_m \),

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

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 \( X \rightarrow Y \rightarrow Z \), then
\[
H(X|Z) \geq H(X|Y).
\]

**Proof.** Exercise 23. ■

Corollary 1.6 When \( X \rightarrow Y \rightarrow Z \), 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 \( P_\theta \). Recall that \( n \)-elemental random sample can always be considered as a random variable taking values in \( X^n \). Clearly the sample depends on chosen distribution \( P_\theta \) or, equivalently, on its index – parameter \( \theta \). Let \( T(X) \) be any statistic (function of the sample) giving an estimate to unknown parameter \( \theta \). Let us consider the Bayesian approach, where \( \theta \) is a random variable with (prior) distribution \( \pi \). Then \( \theta \rightarrow X \rightarrow T(X) \) 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 \( \theta \) as \( X \) and we know that the equality implies \( \theta \rightarrow T(X) \rightarrow X \). By definition of Markov chain, then for every sample \( x \in X^n \)
\[
P(X = x|T(X) = t, \theta) = P(X = x|T(X) = t)
\]
or given the value of \( T(X) \), the distribution of sample is independent of \( \theta \). 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 \( \pi \) of \( \theta \) 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
\[
P(X_1 = x_1, \ldots, X_i = x_i|T(X) = t, \theta) = \begin{cases} 0 & \text{if } \sum_i x_i \neq t, \\ \frac{1}{(\binom{n}{t})} & \text{if } \sum_i x_i = t. \end{cases}
\]

Indeed: if \( \sum_i x_i = t \), then
\[
P(X_1 = x_1, \ldots, X_n = x_n|T(X) = t, \theta) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, T(X) = t, \theta)}{P(T(X) = t, \theta)}
\]
\[
= \frac{\theta^t(1-\theta)^{n-t}\pi(\theta)}{\sum_{x_1,\ldots,x_n: \sum_i x_i = t} \theta^t(1-\theta)^{n-t}\pi(\theta)} = \frac{1}{\binom{n}{t}},
\]

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 \( X = \mathbb{Z} \). 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

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

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 \((X_1, X_2), (X_2, X_3), \ldots\) have the same distribution, the vectors \((X_1, X_2, X_3), (X_2, X_3, X_4), \ldots\) have the same distribution etc.

**Proposition 1.3** When \( \{X_n\}_{n=1}^{\infty} \) is stationary, then the limit \( H' \) 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 \( H' \). 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

\[
\frac{1}{b_n} \sum_{i=1}^n a_ix_i \to x, \text{ when } n \to \infty.
\]

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 \( P(X_m = x_m, \ldots, X_1 = x_1) > 0 \), (1.22) holds, i.e.

\[
P(X_{m+1} = x_{m+1} | X_m = x_m, \ldots, X_1 = x_1) = P(X_{m+1} = x_{m+1} | X_m = x_m).
\] (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 \( x_i, x_j \in \mathcal{X} \)

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

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

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

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 \( \pi \). Markov chain \( \{X_n\} \) is stationary iff \( \pi \) 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

\[
\left( \frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \right).
\]

**Theorem 1.17** Let \( \{X_n\} \) be stationary time-homogenous Markov chain with transition matrix \( (P_{ij}) \) and (stationary) initial distribution \( \pi \). 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 $n$ $H(X_n|X_{n-1}, \ldots, X_1) = H(X_n|X_{n-1})$. Since chain is stationary, we get $H(X_n|X_{n-1}) = H(X_2|X_1)$ and by Theorem 1.15,

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

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 first head. Let $X$ be the number tosses needed. Find $H(X)$, if the probability of head is $p$.

2. Prove grouping property

$$H(p_1, p_2, p_3, \ldots) = H(p_1 + p_2, p_3, \ldots) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

and deduce (1.2).

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

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

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

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

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

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

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 \( n \),

\[ P_n = \left( (1 - \frac{1}{n})p_1, \ldots, (1 - \frac{1}{n})p_m, \frac{1}{nM_n}, \ldots, \frac{1}{nM_n}, 0, \ldots \right), \quad (1.27) \]

where

\[ M_n = \lceil 2^{nc} \rceil, \quad c > 0. \]

show that

\[ H(P_n) = (1 - \frac{1}{n})H(P) + \frac{1}{n} \log_2 M_n + h\left( \frac{1}{n} \right) \to H(P) + c. \]

10. Let \( X \) infinite. Define

\[ P_n = (1 - \frac{\alpha}{\log n}, \frac{\alpha}{n \log n}, \ldots, \frac{\alpha}{n \log n}, 0, \ldots), \]

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

\[ 0 \leq H(X) - H(R) \leq \min_i H(X_i). \]

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 \( H(X) \leq H(Z) \) and \( H(Y) \leq H(Z) \).
- 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 \( \rho \) 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

\[
D(X||Y|Z) = -H(X|Z) + D(X||Y) + H(X) \leq H(Z) + D(X||Y).
\]


17. (a) Let \( X_1 \) and \( X_2 \) have the same distribution. Let

\[
\rho(X_1, X_2) := 1 - \frac{H(X_2|X_1)}{H(X_1)}.
\]

(1.28)

Prove that \( \rho \) is symmetric, \( \rho \in [0, 1] \). When \( \rho = 0? \) When \( \rho = 1? \)

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

\[
\begin{array}{|c|c|c|c|c|}
\hline
Y \backslash X & -n & -1 & 1 & n \\
\hline
n & 0 & 0 & 0 & \epsilon \\
\hline
1 & 0 & \frac{1}{3} - \epsilon & \frac{1}{3} & 0 \\
\hline
-1 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\
\hline
-n & \epsilon & 0 & 0 & 0 \\
\hline
\end{array}
\]

Find \( I(X;Y) \) and \( \rho \) (like in (1.28)). Find \( \text{cov}(X,Y) \) 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

\[
\begin{array}{|c|c|c|c|c|}
\hline
Y \backslash X & -n & -1 & 1 & n \\
\hline
n & 0 & 0 & \frac{1}{3} & 0 \\
\hline
1 & \frac{1}{3} & 0 & 0 & 0 \\
\hline
-1 & 0 & 0 & 0 & \frac{1}{3} \\
\hline
-n & 0 & \frac{1}{3} & 0 & 0 \\
\hline
\end{array}
\]

Find \( I(X;Y) \) and \( \rho \) (like in (1.28)). Find \( \text{cov}(X,Y) \) and the correlation coefficient of \( X \) and \( Y \).
18. Prove

\[ I(X; X|Z) = H(X|Z) \]
\[ I(X; Y|Z) = H(Y|Z) - H(Y|X, Z) \]
\[ I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z) \]
\[ I(X; Y|Z) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z). \]

19. Prove

\[ H(X, Y|Z) \geq H(X|Z) \]
\[ I(X, Y; Z) \geq I(X; Z) \]
\[ H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X) \]
\[ I(X; Y|Z) \geq I(Y; Z|X) - I(Y; Z) + I(X; Y). \]

When the inequalities are equalities?

20. Find \( X, Y, Z \) such that

\[ I(X; Y|Z) > I(X; Y) = 0 \]
\[ 0 = I(X; Y|Z) < I(X; Y). \]

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 \((0 \text{ or } 1 \text{ valued})\) 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, \( \pi \) is a prior distribution of \( \theta \). 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.

30
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 \(|X_1| = n\), \(|X_2| = k\), \(|X_3| = m\), \( k < n \) 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 \(|X| = m\) and let \( X \) be a random variable taking values on \( X \). Find a non-random estimate \( \hat{X} \) to \( X \) with smallest error probability. Let \( P_e = P(X \neq \hat{X}) \). Find \( X \) such that Fano's inequality is an equality
\[
H(X) = P_e \log(|X| - 1) + h(P_e).
\]

29. Let \( P \) be a probability distribution with support \( X_P = \{1, 2, \ldots\} \). Let \( \mu \) 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 $X$, we have another finite coding alphabet $D$. In what follows, $|D| = D$ so that alphabet $D$ will be referred to as $D$-ary alphabet and without loss of generality we take

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

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

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

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

Def 2.1 A code is mapping

$$C : X \rightarrow D^*.$$ 

There are different codes. A classical example of a code is Morse alphabet, where $D$ 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 $X$ 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 codewords. An then a stronger property is needed.

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

$$C^* : X^* \rightarrow D^*, \quad C^*(x_1 \cdots x_n) := C(x_1) \cdots C(x_n).$$ 

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.

<table>
<thead>
<tr>
<th>( \mathcal{X} )</th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>( b )</td>
<td>0</td>
<td>010</td>
<td>00</td>
<td>10</td>
</tr>
<tr>
<td>( c )</td>
<td>1</td>
<td>01</td>
<td>11</td>
<td>110</td>
</tr>
<tr>
<td>( d )</td>
<td>0</td>
<td>10</td>
<td>110</td>
<td>111</td>
</tr>
</tbody>
</table>

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...0 is a codeword of \( cbb...b \) or \( ddb...b \), 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:

<table>
<thead>
<tr>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( d )</th>
<th>( e )</th>
<th>( f )</th>
<th>( g )</th>
<th>( h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>010</td>
<td>012</td>
<td>02</td>
<td>000</td>
<td>001</td>
<td>002</td>
</tr>
</tbody>
</table>

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.

34
Theorem 2.6 (Kraft inequality) Let \( C : \mathcal{X} \to D^* \) 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 \( C : \mathcal{X} \to D^* \) 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 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 \( D^{l^*} \):

\[
\sum_{i=1}^m D^{l^*-l_i} \leq D^{l^*} \iff \sum_{i=1}^m D^{-l_i} \leq 1.
\]

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 \( d_1d_2\cdots d_i \). Let \( 0.d_1d_2\cdots d_i \) be the real number having the \( D \)-ary expansion \( 0.d_1d_2\cdots d_i \), i.e.

\[
0.d_1d_2\cdots d_i = \sum_{j=1}^{l_i} \frac{d_j}{D^j}. \tag{2.2}
\]

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

\[
[0.d_1d_2\cdots d_i, \ 0.d_1d_2\cdots d_i + D^{-l_i}]
\]

corresponding to the codeword \( d_1d_2\cdots d_i \). To this interval belong all real numbers whose \( D \)-ary expansion begins with \( 0.d_1d_2\cdots d_i \). 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

\[
0.\underbrace{0\cdots0}_{l_1}
\]

35
The first interval ends with $D^{-l_1}$ with $D$-ary expansion being

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

Clearly the first interval consists of these real numbers, whose $D$-ary expansion begins with 0.0·0 (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

$$0, \underbrace{0 \cdots 0}_{l_1} \underbrace{0 \cdots 0}_{l_2}.$$

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:

$$[0, \frac{1}{2}), \quad [\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4}), \quad [\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} + \frac{1}{8}), \quad [\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{8}, 1).$$
With binary representation these intervals (recall the numbers of figures after 0. must be $l_i$) are

$[0.\overline{0}, \ 0.1)$, $[0.\overline{10}, \ 0.11)$, $[0.\overline{110}, \ 0.111)$, $[0.\overline{111}, \ 1)$.

**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

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

With binary expansion these intervals are

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

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

\[
L(C) - H_D(P) = \sum_x l(x)P(x) - \sum_x P(x) \log_D \frac{1}{P(x)} \\
= -\sum_x P(x) \log_D D^{-l(x)} + \sum_x P(x) \log_D P(x).
\]

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 if and only if every \( x \in \mathcal{X} \) it holds \( P(x) = D^{-l(x)} \). Necessary condition is that \( -\log_D P(x) \) is integer for every \( x \in \mathcal{X} \).

\[\blacksquare\]

**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

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/32</td>
<td>1/32</td>
<td>1/16</td>
<td>1/16</td>
<td>1/8</td>
<td>1/8</td>
<td>1/4</td>
<td>1/4</td>
<td>1/16</td>
</tr>
</tbody>
</table>

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 (5.31) 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:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>11111</td>
<td>11110</td>
<td>1110</td>
<td>1101</td>
<td>1100</td>
<td>101</td>
<td>100</td>
<td>01</td>
<td>00</td>
</tr>
</tbody>
</table>

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

\[
l(x) = [\log_D \frac{1}{P(x)}]. \tag{2.5}
\]

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

\[
[\log_D \frac{1}{P(x)}] < \log_D \frac{1}{P(x)} + 1.
\]

Therefore

\[
L(C) = \sum_x P(x) [\log_D \frac{1}{P(x)}] < \sum_x P(x) \log_D \frac{1}{P(x)} + 1 = H_D(P) + 1.
\]

39
Corollary 2.1 For every distribution, there exist a prefix code \( C: \mathcal{X} \rightarrow \mathcal{D}^* \) 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:

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( x_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>001</td>
<td>010</td>
<td>011</td>
<td>110</td>
</tr>
</tbody>
</table>

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:
\[
\sum_x l_Q(x)P(x) = \sum_x \lceil \log \frac{1}{Q(x)} \rceil P(x) < \sum_x P(x) \left( \log \frac{1}{Q(x)} + 1 \right)
\]
\[
= \sum_x P(x) \left( \log \frac{P(x)}{Q(x)} + \log \frac{1}{P(x)} + 1 \right)
\]
\[
= D(P||Q) + H(P) + 1.
\]

To find the lower bound is Exercise 2.

**Remark:** The statement obviously holds 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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.35</td>
<td>0.1</td>
<td>0.15</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>{b, c}</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.35</td>
<td>0.25</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>{b, c}</th>
<th>{d, e}</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.35</td>
<td>0.25</td>
<td>0.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>{a, b, c}</th>
<th>{d, e}</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>01</td>
<td>02</td>
<td>000</td>
<td>001</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>*</th>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.35</td>
<td>0.1</td>
<td>0.15</td>
<td>0.2</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
In the first step the letters $d, e, *$ will be merged. Then the rest.

A Huffman code:

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>30</td>
<td>31</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

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

2.4.2 Huffman code is optimal

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

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

since $|X| < \infty$, 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

$$L(C) - L(C^*) = P(x_i)l(x_i) + P(x_j)l(x_j) - (P(x_i)l(x_j) + P(x_j)l(x_i))$$

$$= (P(x_i) - P(x_j))(l(x_i) - l(x_j)) > 0;$$

we obtain that $C$ cannot be optimal. $\blacksquare$

From Proposition 2.1 it follows that for every optimal code, there is an ordering $X = \{x_i\}$ 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 $d', d'' \in D^*$ 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 losing 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 Huffman 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

<table>
<thead>
<tr>
<th>letter</th>
<th>code $C_1$</th>
<th>code $C_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>11</td>
<td>111</td>
</tr>
<tr>
<td>b</td>
<td>101</td>
<td>110</td>
</tr>
<tr>
<td>c</td>
<td>100</td>
<td>101</td>
</tr>
<tr>
<td>d</td>
<td>011</td>
<td>100</td>
</tr>
<tr>
<td>e</td>
<td>010</td>
<td>01</td>
</tr>
<tr>
<td>f</td>
<td>00</td>
<td>00</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>letter</th>
<th>$\frac{a}{k}$</th>
<th>$\frac{b}{k}$</th>
<th>$\frac{c}{k}$</th>
<th>$\frac{d}{k}$</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1/12</td>
</tr>
</tbody>
</table>

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 \( |X| = \infty \), 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 \( X \) 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)

\[ p_1, p_2, \ldots, p_{m_1}, p^*_1. \tag{2.10} \]

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 \( p^*_1 \) can be constructed. For that, we consider the distribution (proportional to)

\[ p_{m_1+1}, p_{m_1+2}, \ldots, p_{m_2}, p^*_2. \tag{2.11} \]

The numbers (2.11) are not probability distribution, since their sum is \( p^*_1 < 1 \). 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 \( p^*_1 \). Now the tree (code) with leaves (letters)

\[ p_1, p_2, \ldots, p_{m_1}, p_{m_1+1}, p_{m_1+2}, \ldots, p_{m_2}, p^*_2 \]

is constructed and the next step is to build the subtree starting from \( p^*_2 \). For that, again, we construct the Huffman tree for the atoms

\[ p_{m_2+1}, p_{m_2+2}, \ldots, p_{m_3}, p^*_3. \tag{2.12} \]
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 uniquely decodable code with codeword lengths $\{l(x)\}$. Then Kraft inequality holds

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

**Proof.** At first, we consider special case $|X| < \infty$.

Let $C^k$ be the $k$-extension of $C$, i.e.

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

$$\left( \sum_x D^{-l(x)} \right)^k = \sum_{x_1 \in X} \sum_{x_2 \in X} \cdots \sum_{x_k \in X} D^{-l(x_1)} D^{-l(x_2)} \cdots D^{-l(x_k)}$$

$$= \sum_{x_1 x_2 \cdots x_k \in X^k} D^{-l(x_1)} D^{-l(x_2)} \cdots D^{-l(x_k)}$$

$$= \sum_{x^k \in X^k} D^{-l(x^k)},$$
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
\[
\max_{x^k \in \mathcal{X}^k} l(x^k) = kl_{\max}.
\]
Thus
\[
\left( \sum_{x} D^{-l(x)} \right)^{k} = \sum_{x^k \in \mathcal{X}^k} D^{-l(x^k)} = \sum_{m=k}^{kl_{\max}} a(m)D^{-m}.
\]
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
\[
\left( \sum_{x} D^{-l(x)} \right)^{k} = \sum_{m=k}^{kl_{\max}} a(m)D^{-m} \leq \sum_{m=1}^{kl_{\max}} D^m D^{-m} = kl_{\max}
\]
or
\[
\sum_{x} D^{-l(x)} \leq (kl_{\max})^{\frac{1}{k}}.
\]
The left hand side is independent of \( k \). Therefore
\[
\sum_{x} D^{-l(x)} \leq \lim_{k \to \infty} (kl_{\max})^{\frac{1}{k}} = 1.
\]
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

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

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 $L(C^k) = k L(C)$ 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

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

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

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

hence

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

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

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

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
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 $C_k : \mathcal{X}^k \rightarrow \mathcal{D}^*$ be optimal. Recall that a stationary process always has an entropy rate

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

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 $X$ 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 $X^k$ instead of the letters, since $X^k$ 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 $E : \{1, 2, \ldots\} \rightarrow \mathcal{D}^e$ such that

$$|E(n)| = \log_D n + o(\log_D n) \quad (2.17)$$

**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

$$|w(n)| = \lceil \log_D(n + 1) \rceil.$$

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

$$|u(n)| = |v(n)| = \lceil \log_D(1 + \lceil \log_D(n + 1) \rceil) \rceil$$

and

$$|E(n)| = \lceil \log_D(n + 1) \rceil + 2\lceil \log_D(1 + \lceil \log_D(n + 1) \rceil) \rceil = \log_D n + o(\log_D n).$$

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 $C_k : \mathcal{X}^k \rightarrow \mathcal{D}^*$ 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 $E(12) = 001001100$, we get

$$C_k^*(x^k) = 001001100001001100111.$$  

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

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

With Elias prefix we can define a general prefix code

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

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

2.7 Exercises

1. Consider the distribution

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/32</td>
<td>1/32</td>
<td>1/16</td>
<td>1/16</td>
<td>1/16</td>
<td>1/8</td>
<td>1/8</td>
<td>1/4</td>
<td>1/4</td>
</tr>
</tbody>
</table>

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

2. Find the lower bound in Theorem 2.8.

3. Let $P$

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.05</td>
<td>0.1</td>
<td>0.13</td>
<td>0.2</td>
<td>0.12</td>
<td>0.08</td>
<td>0.07</td>
</tr>
</tbody>
</table>

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 $C$ be the set of prefix codes. Prove that
   
   $$P(C \in C) = \prod_{k=1}^{m} \left(1 - \sum_{j=1}^{k-1} 2^{-l_j}\right)^+.$$  

   Prove that $P(C \in C) > 0$ 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 $P^m$.

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 $q$. 

53
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 \( D = \{\alpha, \beta, \gamma, \delta\} \) are coded into binary codewords as follows:

\[
\begin{align*}
\alpha &\mapsto 00, \beta \mapsto 01, \gamma \mapsto 10, \delta \mapsto 11.
\end{align*}
\]

Let us call this process \textit{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

\[
\min_{l_1, \ldots, l_m} \sum_{i=1}^{m} u_i l_i
\]

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 \( X \), and let \( C^k \) be its \( k \)-extension. Prove \( L(C^k) = kL(C) \).

16. Let \( Y \) be a stationary MC on alphabet \( X \) with transition matrix

\[
\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{pmatrix}
\]

Find the entropy rate of the process. Let \( C_1 \), \( C_2 \) and \( C_3 \) three binary codes on \( X \). 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 $P$ be a distribution on $\{a, b, c\}$.

$\begin{array}{ccc}
a & b & c \\ 0.5 & 0.25 & 0.25 \\
\end{array}$

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

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

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 $W^n_\epsilon \subset \mathcal{X}^n$ as follows: $x^n \in \mathcal{X}^n$ belongs to the set $W^n_\epsilon$ if and only if

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

The elements of $W^n_\epsilon$ are called weakly $\epsilon$-typical words.

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

1. If $x^n \in W^n_\epsilon$, then

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

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

$$P(W^n_\epsilon) > 1 - \epsilon.$$ (3.3)

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

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

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

$$-\frac{1}{n} \log P(X_1, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^n \log P(X_i) \overset{P}{\to} -E \log P(X_1) = H.$$ (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 $n_\epsilon$ (depending on $\epsilon$) so that

$$\mathbb{P}\left( \left| -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i) - H \right| \leq \epsilon \right) \geq 1 - \epsilon,$$

(3.6) provided $n > n_\epsilon$.

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

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

so that

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

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

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

From this

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

Therefore, if $n$ is big, the probability of $W^n_\epsilon$ is almost one. This means that most likely a realization of $X_1, \ldots, X_n$ is a weakly $\epsilon$-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).

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

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 $p$. 57
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 $\epsilon$-typical words $W_\epsilon^n$. Let us order the elements of $W_\epsilon^n$. Since $|W_\epsilon^n| \leq 2^n(H + \epsilon)$, then we can represent every index as a binary word with length $n(H + \epsilon)$.

To every such binary word add prefix 0 to show that the codeword corresponds to a weakly typical word. Hence, $l(x) \leq n(H + \epsilon) + 2, \forall x^n \in W_\epsilon^n$.

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:

$$L = \sum_{x^n \in \mathcal{X}^n} l(x^n)P(x^n) = \sum_{x^n \in W_\epsilon^n} l(x^n)P(x^n) + \sum_{x^n \notin W_\epsilon^n} l(x^n)P(x^n)$$

$$\leq \sum_{x^n \in W_\epsilon^n} (n(H + \epsilon) + 2)P(x^n) + \sum_{x^n \notin W_\epsilon^n} (n \log |\mathcal{X}| + 2)P(x^n)$$

$$= P(W_\epsilon^n)(n(H + \epsilon) + 2) + (1 - P(W_\epsilon^n))(n \log |\mathcal{X}| + 2).$$

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 $\epsilon$ small enough and $n$ big enough, $\epsilon'$ can be made arbitrary small.

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

$$H \leq L_n(C) < H + \epsilon,$$  \hspace{1cm} (3.7)  

where $C : \mathcal{X}^n \rightarrow \{0,1\}^*$ 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 \( W_\epsilon^n \) such that \( P(W_\epsilon^n) \geq 1 - \epsilon \), but the number of elements in \( W_\epsilon^n \) is relatively small. However \( W_\epsilon^n \) is not the smallest (in terms of the number of elements) set having probability at least \( 1 - \epsilon \). Let \( B_\epsilon^n \) 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 \( \epsilon_1 \), in addition let

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

Define

\[
S := W_\epsilon^n \cap B_\epsilon^n.
\]

By (3.3) and (3.4), it holds

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

Therefore

\[
\log |B_\epsilon^n| \geq \log|1 - (\epsilon + \epsilon_1)| + n(H - \epsilon_1) = n\left(\frac{\log(1 - (\epsilon + \epsilon_1))}{n}\right) + H - \epsilon_1 \geq n(H - \delta).
\]

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

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

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

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

Therefore $W_{25}^{0.2}$ 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:

\[
\begin{bmatrix}
0.8 - 0.7635908 \\
0.0000109419
\end{bmatrix} = 3328
\]

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 \( W_{0.2}^{25} \) consist 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 a joint distribution \( P(x, y) \) on \( \mathcal{X} \times \mathcal{Y} \). In what follows, let \( P_x \) and \( P_y \) be the marginal distributions of \( P \). Since the pairs \((X_i, Y_i)\) are independent, 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 \( W^n_\epsilon \) consist of pairs \((x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \) satisfying the following conditions:

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

The pairs in set \( W^n_\epsilon \) are called (weakly) jointly \( \epsilon \)-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. Recall \( P = P(x, y) \) is the joint distribution on \( \mathcal{X} \times \mathcal{Y} \), and \( P_x \) and \( P_y \) are marginal distributions. Then **product measure** \( P_x \times P_y \) is a probability measure on \( \mathcal{X} \times \mathcal{Y} \) defined as follows

\[
P_x \times P_y(x, y) = P_x(x)P_y(y).
\]

61
Hence $P_x \times P_y$ has the same marginal distributions but the joint distribution corresponds to the independence. Let us denote

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

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

1. If $n$ is big enough, then
   $$P(W^n_\epsilon) > 1 - \epsilon.$$  \hfill (3.10)

2. If $n$ is big enough, then
   $$(1 - \epsilon)2^{n(H(X,Y) - \epsilon)} \leq |W^n_\epsilon| \leq 2^{n(H(X,Y) + \epsilon)}.$$  \hfill (3.11)

3. If $n$ is big enough, then
   $$(1 - \epsilon)2^{-n(I(X;Y) + 3\epsilon)} \leq P_x \times P_y(W^n_\epsilon) \leq 2^{-n(I(X;Y) - 3\epsilon)}.$$  \hfill (3.12)

**Proof.** Proof follows that of Theorem 3.2.

1: From the weak law of large numbers

$$-\frac{1}{n} \log P_x(X_1, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log P_x(X_i) \xrightarrow{P} H(X)$$

$$-\frac{1}{n} \log P_y(Y_1, \ldots, Y_n) = -\frac{1}{n} \sum_{i=1}^{n} \log P_y(Y_i) \xrightarrow{P} H(Y)$$

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

Proving 1 is now Exercise 1.

2:

$$1 \geq P(W^n_\epsilon) = \sum_{(x^n, y^n) \in W^n_\epsilon} P(x^n, y^n) \geq |W^n_\epsilon|2^{-n(H(X,Y) + \epsilon)},$$

$$1 - \epsilon \leq P(W^n_\epsilon) \leq \sum_{(x^n, y^n) \in W^n_\epsilon} P(x^n, y^n) \leq |W^n_\epsilon|2^{-n(H(X,Y) - \epsilon)},$$

implying

$$(1 - \epsilon)2^{n(H(X,Y) - \epsilon)} \leq |W^n_\epsilon| \leq 2^{n(H(X,Y) + \epsilon)}.$$
3: Applying 2, we get

\[ P_x \times P_y(W^n) = \sum_{(x^n,y^n) \in W^n} P_x(x^n)P_y(y^n) \]

\[ \leq \sum_{(x^n,y^n) \in W^n} 2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \]

\[ \leq 2^{n(H(X)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \]

\[ = 2^{-n(I(X;Y)-3\epsilon)} \]

\[ P_x \times P_y(W^n) \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)} \]

\[ = (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}. \]

The interpretation of first two statements of Theorem 3.4 is 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

<table>
<thead>
<tr>
<th>(\mathcal{X}\backslash\mathcal{Y})</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(\frac{7}{80})</td>
<td>(\frac{3}{80})</td>
</tr>
<tr>
<td>0</td>
<td>(\frac{9}{80})</td>
<td>(\frac{63}{80})</td>
</tr>
</tbody>
</table>
Thus $X \sim B(1,0.1)$, $Y \sim B(1,0.2)$. Joint entropy

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

The words $x^n = 1000000000$ and $y^n = 0110000000$ are both weakly typical with respect to any $\epsilon$ 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

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

$$\frac{1}{n} \log P(x^n, y^n) = \frac{3}{10} \log q + \frac{7}{10} \log(1-q) + \frac{9}{10} \log(1-p) + \frac{1}{10} \log p$$

$$= q \log q + \frac{7}{40} \log q - \frac{7}{40} \log(1-q) + (1-q) \log(1-q) + (1-p) \log(1-p) + p \log p$$

$$= -h(q) - h(p) + \frac{7}{40} \log\left(\frac{q}{1-q}\right).$$

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) $X = \{X_n\}_{n=1}^{\infty}$:

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

where $H_X$ is the entropy of $X$, 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 $X$. Consider the likelihood ratio

$$\frac{Q(X_1) \cdot \cdots \cdot Q(X_n)}{P(X_1) \cdot \cdots \cdot P(X_n)}.$$ 

Prove that there exists a set $A^n \subset X^n$ and a constant $A$ such that

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

2. if $n$ is big enough, then
   $$P(A^n) > 1 - \epsilon;$$

3. if $n$ is big enough, then
   $$(1 - \epsilon)2^{n(A-\epsilon)} \leq |A^n| \leq 2^{n(A+\epsilon)}.$$

3. Let $X_1, X_2, \ldots$ be stationary MC with finite number of states ($|X| < \infty$) 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 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 \( X \) be a finite input alphabet and \( Y \) a finite output alphabet. In a noisy channel, every input character \( x \) is transmitted into a output character \( y \) with fixed probability \( P(y|x) \). The system consisting on \( X, Y \) and the transition matrix

\[
(P(y|x))_{x \in X, y \in Y}
\]

is called discrete channel. The channel is said to be memoryless if the distribution of output depends only on the input at that time and is conditionally independent of previous inputs or outputs.

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 discrete memoryless 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

\[
C = \max_{P(x)} I(X; Y) \leq \max_{P(x)} H(X) \leq \log |\mathcal{X}|, \quad C = \max_{P(x)} I(X; Y) \leq \max_{P(x)} H(Y) \leq \log |\mathcal{Y}|,
\]

so that the following inequality holds: \( C \leq \log \min\{ |\mathcal{X}|, |\mathcal{Y}| \} \).
### 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 $C \leq \log \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, the maximum possible channel capacity for every channel with binary input alphabet is 1.

**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 channel, 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

$$C = \max_{P(x)} \left( H(Y) - H(Y|X) \right) = \max_{P(x)} H(Y) - 1 = \log 26 - 1 = \log 13,$$

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 (BSC).** 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 $X = \{0, 1\}$ and $Y = \{0, 1, e\}$. 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

$$C = \max_{P(x)} (H(Y) - H(Y|X)) = \max_{P(x)} H(Y) - h(p).$$

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 $\pi = P(X = 1)$. Then $P(Y = 1|Y \neq e) = \pi$ and $P(Y = 0|Y \neq e) = (1 - \pi)$ and

$$H(Y|E) = H(Y|Y \neq e)P(Y \neq e) = h(\pi)(1 - p).$$

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:

\[
\begin{pmatrix}
\frac{1}{3} & \frac{1}{6} & \frac{1}{2} & \frac{1}{6}
\end{pmatrix}.
\]

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

\[
C : \{1, 2, \ldots, M\} \mapsto \mathcal{X}^n,
\]

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 of length $n$ 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
  $$C : \{1, \ldots, M\} \to \mathcal{X}^n.$$  
  The set of codewords $C(1), \ldots, C(M)$ is called the codebook.
- A decoding function
  $$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}.$$  

**Error probabilities.** Let $\lambda_i$ be the conditional probability of error of $(M, n)$ code given that the index $i$ was sent. Thus

$$\lambda_i := P(\hat{W} \neq i | W = i) = P(g(Y^n) \neq i | W = i) = \sum_{y^n : g(y^n) \neq i} P(y^n | C(i)).$$

Let

$$\lambda_{\text{max}} := \max_i \lambda_i$$

and let $P_e$ be the error of mistake provided that the distribution of $W$ us uniform on $\{1, \ldots, M\}$. Thus

$$P_e = P(\hat{W} \neq W) = \sum_i P(\hat{W} \neq i | W = i)P(W = i) = \frac{1}{M} \sum_i P(\hat{W} \neq i | W = i) = \frac{1}{M} \sum_i \lambda_i.$$  

Obviously

$$P_e \leq \lambda_{\text{max}}.$$
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 big. Indeed, if \(|\mathcal{X}| = 2\), then the smallest codeword length for fixed-length non-singular code \(C\) is \([\log M]\). Thus, for any meaningful \((M, n)\) code (with non-singular code \(C\)) the rate cannot be bigger 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 \([\log M]\). Let call this code **uniform**. When the channel is noiseless, then uniform code works just fine and \(\lambda_{\text{max}} = 0\). However, using the code with binary symmetric channel the error probability \(\lambda_{\text{max}}\) increases with \(n\):

\[
1 - \lambda_i = P(\hat{W} = i|W = i) = P(Y^n = C(i)) = (1 - p)^n.
\]

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 \([\log M]m\). 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 \(\lambda_{\text{max}} < \epsilon\). 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 \((r 2^{nR\epsilon}, n)\) codes such that \(\lambda_{\text{max}} \to 0\) 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 \((r 2^{nR\epsilon}, n)\) code such that \(\lambda_{\text{max}} < \epsilon\). When \(\lambda_{\text{max}} < \epsilon\), then for any distribution of \(W\), the probability of error is at most \(\epsilon\).

**NB!** In what follows, we shall denote \(r 2^{nR\epsilon}\) by \(2^{nR}\).
4.3.2 Channel coding theorem

The following theorem, sometimes called Shannon’s 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 \((2^n R, n)\) codes so that \( \lambda_{\max} \to 0 \) as \( n \) grows. Conversely, any sequence of \((2^n R, n)\) 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 at least one non-random code that must work also well.

More precisely: let \( R < C \). A random \((2^n R, n)\) 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 \( 2^n R \) random \( n \)-dimensional vectors, each of them is i.i.d vector with components distributed as \( P(x) \). Obtained words \( x^n(1), \ldots, x^n(2^n R) \) form (random) codebook:

   \[ C : \{1, \ldots, 2^n R\} \to \mathcal{X}^n, \quad C(i) = x^n(i). \]

   This code is revealed to both sender and receiver.

3. A message \( W \) is chosen from \( \{1, \ldots, 2^n R\} \) 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

   \[ P\left(y^n|x^n(w)\right) = \prod_i^n P\left(y_i|x_i(w)\right). \]

6. Receiver decodes obtained word \( y^n \) according to the rule:

   \[ g(y^n) = \begin{cases} k & \text{if } (x^n(k), y^n) \in W^n_\epsilon \text{ and for every } i \neq k, (x^n(i), y^n) \notin W^n_\epsilon, \\ * & \text{else.} \end{cases} \]

   Since \( * \notin \mathcal{Y} \), the output \( * \) is always a mistake. Here \( \epsilon > 0 \) is so small that \( C - R - 3\epsilon > 0 \) and \( W^n_\epsilon \) is the set of jointly typical words. The receiver knows \( P(x)P(y|x) \), hence he also knows the set \( W^n_\epsilon \).
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 $2\epsilon$, provided $n$ is big enough. Thus, for $n$ big enough

$$\sum_C P(C) P_e(C) = \sum_C P(C) \frac{1}{2^{nR}} \sum \lambda_j(C) \leq 2\epsilon,$$

where $P(C)$ is the probability of obtaining a particular code $C$. Since the average probability of error is smaller than $2\epsilon$, there must be at least one deterministic code $C^*$ so that

$$P_e(C^*) = \frac{1}{2^{nR}} \sum \lambda_i \leq 2\epsilon,$$

where $\lambda_i := \lambda_i(C^*)$. 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 $4\epsilon$. 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 $|\mathcal{X}|^{2^{nR}}$ 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 indicates 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."

Nowadays: Turbo codes and Low Density Parity Check codes near capacity.

4.3.3 The proof of converse

**Lemma 4.1** Let \( X^n = C(W) \) random codeword, let \( Y^n = (Y_1, \ldots, Y_n) \) be its output and \( C \) channel capacity. Then

\[
I(X^n; Y^n) \leq nC.
\]

**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

\[
I(X^n; Y^n) = H(Y^n) - H(Y^n | X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i | X_i)
\]

\[
\leq \sum_{i=1}^{n} (H(Y_i) - H(Y_i | X_i)) = \sum_{i=1}^{n} I(X_i; Y_i) \leq nC.
\]

\[\]
Zero-error codes. The second claim of channel coding theorem is the following: if there is a \((2^{nR}, n)\) code having small maximal error, then \(R \leq C\). The following claim is weaker.

Proposition 4.1 If there is an \((2^{nR}, n)\) code so that \(\lambda_{\text{max}} = 0\), then \(R \leq C\).

Proof. Let \((2^{nR}, n)\) be such code. Then there exists a decoding function \(g\) so that \(g(Y^n) = W\) a.s. so that \(H(W|Y^n) = 0\) implying that

\[
I(W; Y^n) = H(W) - H(W|Y^n) = H(W).
\] (4.3)

Recall that \(X^n = C(W)\) is random codeword. Since \(W \rightarrow X^n \rightarrow Y^n\),

the data processing inequality implies

\[
I(W; Y^n) \leq I(X^n; Y^n).
\] (4.4)

Let \(W\) have uniform distribution. Then \(H(W) = nR\) and from (4.4), (4.3) and Lemma 4.1, it follows that

\[
nR = H(W) = I(W; Y^n) \leq I(X^n; Y^n) \leq \sum_{i=1}^{n} I(X_i; Y_i) \leq nC.
\]

What are necessary properties for \(C\) to be zero-error code? If \(\lambda_{\text{max}} = 0\), then \(W = g(Y^n)\) so that

\[
W \rightarrow X^n \rightarrow Y^n \rightarrow W.
\]

Now, it is easy to see that

\[
I(W; Y^n) = I(W; X^n) = I(X^n; Y^n) = I(W; W) = H(W) = H(X^n) = H(Y^n).
\]

Indeed, applying data processing inequality to \(W \rightarrow X^n \rightarrow Y^n\) and \(X^n \rightarrow Y^n \rightarrow W\), we get the first equality. The data processing inequality applied to \(Y^n \rightarrow W \rightarrow X^n\) together with (4.4) and the first equality implies the second equality. The forth inequality is obvious and the third follows from the forth together with (4.3). Since \(X^n = C(W)\), it follows that \(H(X^n|W) = 0\) so that \(H(W) = I(X^n; W) = H(X^n) - H(X^n|W) = H(X^n)\) and so we have the fifth equality. The last inequality follows similarly from the fact that \(W = g(Y^n)\) implying \(H(W|Y^n) = 0\) and \(H(W) = I(Y^n; W) = H(Y^n) - H(W|Y^n)\).

From \(I(X^n; W) = H(W) - H(W|X^n) = H(W)\), it follows that

\[
H(W|X^n) = H(W|C(W)) = 0
\]
i.e. \( \mathcal{C} \) is non-singular.

Suppose now that the rate of \( \mathcal{C} \) equals to \( C \) and \( \lambda_{\text{max}} = 0 \). Then the inequalities in the proof of Proposition 4.1 must be equalities. The first equality is

\[
I(W;Y^n) = I(X^n;Y^n)
\]

which is a consequence of \( \lambda_{\text{max}} = 0 \) and then \( \mathcal{C} \) is non-singular. The second equality is

\[
H(Y^n) = \sum_{i=1}^{n} H(Y_i)
\]

implying that the random variables \( Y_i \) are independent. The third equality

\[
\sum_{i=1}^{n} I(X_i;Y_i) = nC
\]

holds when \( I(X_i;Y_i) = C \) for every \( i \) or, equivalently, the distribution of \( X_i \) is the one achieving the capacity of the channel for every \( i \).

Hence, if \( (2^nR, n) \) is a code such that \( P_e = 0 \) and \( R = C \), then:

- \( \mathcal{C} \) is one-to-one (non-singular);
- for uniformly distributed \( W \), the random variables \( X_i \) have distribution \( P^*(x) \) that achieves the capacity of the channel;
- for uniformly distributed \( W \), the random variables \( Y_i \) are i.i.d. random variables with distribution

\[
P(y) = \sum_x P(y|x)P^*(x).
\] (4.5)

Any code having small maximal error and rate close to \( C \) must have similar properties.

**Examples:**

- Noisy keyboard. In this case the capacity of channel is easy to achieve. Indeed, let \( M = 2^{13n} \) and let \( \mathcal{C} \) be any non-singular code taking values on \( \{1, 3, 5, \ldots , 25\}^n \). The rate of this code is \( R = (\log M)/n = 13 = C \), where \( C \) is the capacity of channel. Clearly in this case \( \lambda_{\text{max}} = 0 \) is achievable. If \( W \) is uniform, then random codeword \( X^n = X_1, \ldots , X_n \) be uniform on \( \{1, 3, 5, \ldots , 25\}^n \). Then \( X_1, \ldots , X_n \) are i.i.d. and \( X_i \) is uniform on \( \{1, 3, 5, \ldots , 25\} \). This is also the distribution \( P^* \). With this input, \( Y^n = (Y_1, \ldots , Y_n) \) are i.i.d. random variables with distribution (4.5).

- BSC. For this channel \( P_e > 0 \). However, in order \( P_e \) being small, the best code should be such that for uniform \( W \), the distribution of \( Y^n \) is close to i.i.d. Bernoulli 1/2 distribution. This is certainly not so by repetition code.
Fano’s inequality revisited.

Lemma 4.2 (Fano’s inequality) Let \( W \) be a random word. Then

\[
H(W|Y^n) \leq 1 + P(W \neq \hat{W})nR. \tag{4.6}
\]

Proof. Recall Fano’s inequality:

\[
H(W|\hat{W}) \leq h(P(W \neq \hat{W})) + P(W \neq \hat{W}) \log(2^{nR} - 1) \leq 1 + P(W \neq \hat{W})nR.
\]

Since \( \hat{W} = g(Y^n) \), from data processing inequality \((W \to Y^n \to \hat{W})\), it follows

\[
I(W;Y^n) \geq I(W,\hat{W})
\]

so that

\[
H(W|\hat{W}) = H(W|g(Y^n)) \geq H(W|Y^n).
\]

Proof of converse to the channel coding theorem. Let \((2^nR, n)\) be a sequence of codes so that \( \lambda_{max} \to 0 \). The objective is to show that then \( R \leq C \).

Since \( \lambda_{max} \to 0 \), then

\[
P_e = \frac{1}{2^{nR}} \sum_{i=1}^{2^nR} \lambda_i \to 0.
\]

We show that the convergence \( P_e \to 0 \) implies \( R \leq C \). Recall \( P_e = P(\hat{W} \neq W) \) given \( W \) has uniform distribution. Recall the proof of Proposition 4.1 based on the equalities

\[
nR = H(W) = H(W) - H(W|Y^n) + H(W|Y^n) = I(W;Y^n) + H(W|Y^n) = I(W;Y^n),
\]

since by zero-error code \( H(W|Y^n) = 0 \). In the present case \( H(W|Y^n) \neq 0 \), but by Fano’s inequality it can be estimated from above: \( H(W|Y^n) \leq 1 + P_e nR \). Thus

\[
nR = H(W) = H(W|Y^n) + I(W;Y^n) \leq 1 + P_e nR + I(W;Y^n)
\]

\[
\leq 1 + P_e nR + I(X^n;Y^n) \leq 1 + P_e nR + nC,
\]

where the two last inequalities follows from data processing inequality (4.4) and Lemma (4.1), respectively. Now

\[
R \leq P_e R + \frac{1}{n} + C. \tag{4.7}
\]

Since \( P_e R + \frac{1}{n} \to 0 \) as \( n \) grows, we get \( R \leq C \). 

77
Remarks:

1. From (4.7):

\[ P_e \geq 1 - \frac{C}{R} - \frac{1}{nR}, \]

where \( asn \to \infty \) \[ 1 - \frac{C}{R} - \frac{1}{nR} \to 1 - \frac{C}{R} > 0. \]

Thus, when \( C < R \), then \( \frac{C}{R} < 1 \) so that for a \( \delta' > 0 \), \( P_e > \delta' \), provided \( n \) is big enough. That implies \( P_e > 0 \) also for small \( n \) (because, if for a small \( n \) \( P_e = 0 \), it holds also for several big \( n \)). Thus, there exists \( \delta > 0 \) so that \( P_e > \delta \) for any \( n \).

2. It can be shown that \( R > C \) actually implies that \( P_e \to 1 \) exponentially fast (strong converse).

4.4 Feedback capacity

Channel with feedback is the following: after transmitting the \( i \)-th element of codeword \( x^n \), the received \( y_i \) is sent back to sender without errors. The sender takes this information into consideration when sending the next bit. Hence, code \( C \) is replaced by sequence \( C_i \), where besides \( W \), the arguments of \( C_i \) are also \( y_1, \ldots, y_{i-1} \).

**Def 4.6** Let \( \{P(y|x)\}_{x \in \mathcal{X}, y \in \mathcal{Y}} \) be discrete memoryless channel. A \((M, n)\) feedback code consist of:

- set \( \{1, \ldots, M\} \);
- sequence of codes \( C_i : \{1, \ldots, M\} \times \mathcal{Y}^{i-1} \to \mathcal{X} \);
- a decoding function \( g : \mathcal{Y}^n \to \{1, 2, \ldots, M\} \).

**Example:** Recall binary erasure channel. Here feedback is useful: whenever \( e \) is received, it is sent back to sender so that the corresponding letter can be sent afresh.

Channel without feedback is a special case of the channel with feedback. Hence if a rate \( R \) is achievable without feedback then it is clearly achievable with feedback. However, one expects that with feedback a higher rate as \( C \) can be achieved. Surprisingly, this is not the case.
**Theorem 4.7 (Feedback capacity)** Assume feedback channel. If $R$ is an achievable rate, then $R \leq C$.

**Proof.** The proof similar to the one without feedback. Let $(2^nR, n)$ be such that $\lambda_{\text{max}} \to 0$. The objective is to show that $R \leq C$.

Let $W$ have uniform distribution. Then $P_e = P(\hat{W} \neq W) \to 0$. Fano's inequality

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) \leq 1 + P_enR + I(W; Y^n).$$

Let us bound

$$I(W; Y^n) = H(Y^n) - H(Y^n|W) = H(Y^n) - H(Y_1|W) - H(Y_2|Y_1, W) - \cdots - H(Y_n|Y_1, \ldots, Y_{n-1}, W)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W, X_i).$$

The last inequality holds since $X_i = C_i(Y_1, \ldots, Y_{i-1}, W)$. The channel is memoryless. Thus $Y_i$ depends solely on $X_i$ and so

$$P(y_i|y_1, \ldots, y_{i-1}, w, x_i) = P(y_i|x_i) \quad \text{and} \quad H(Y_i|Y_1, \ldots, Y_{i-1}, W, X_i) = H(Y_i|X_i).$$

Now argue as previously:

$$I(W; Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W, X_i) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)$$

$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) = \sum_{i} I(X_i, Y_i) \leq nC.$$

Therefore, $nR \leq P_enR + 1 + nC$ or $R \leq P_eR + \frac{1}{n} + C \to C$, provided $P_e \to 0$. ■

As previously, if $R > C$ then

$$P_e \geq 1 - \frac{R}{C} - \frac{1}{nR}$$

so that for $\delta' > 0$, $P_e > \delta'$, provided $n$ is big enough and that implies that there exists a $\delta > 0$ so that $P_e > \delta$ for any $n$. 

79
4.5 Source-channel separation

Let $V_1, V_2, \ldots$ be a stationary source with entropy rate $H$ taking values on finite alphabet $\mathcal{V}$. Can $V^n = (V_1, \ldots, V_n)$ be transmitted with $n$ symbols though the channel such that error is small?

The vector $V^n$ can be regarded as a random word taking values on $\mathcal{V}^n$ and apply channel coding theorem. Then $M = |\mathcal{V}|^n$ and since the aim is to transmit with $n$ transmission, the rate of the code has to be $R = (\log M)/n = \log |\mathcal{V}|$. If $\log |\mathcal{V}| < C$, where $C$ is the channel capacity, then by channel coding theorem, a good transmission with this rate is possible.

The following theorem states more: if the process $V_1, V_2, \ldots$ has AEP property with and $H < C$, then good transmission is possible no matter how large is $\mathcal{V}$.

**Theorem 4.8** Let $V = V_1, V_2, \ldots$ be a stationary process satisfying AEP and the the condition $H < C$. Then $V^n$ is possible transmit with $n$ transmission so that

$$P(\hat{V}^n \neq V^n) \to 0.$$

**Proof.** Take $\epsilon > 0$ so small that $H + 2\epsilon < C$. By AEP: $\forall \epsilon > 0$ there exists $W^n_e$ such that $P(W^n_e) > \epsilon$ and $|W^n_e| \leq 2^{n(H+\epsilon)}$. Hence the set $W^n_e$ (weakly typical words) can be considered as a vocabulary consisting of $2^{nR}$ words, where $R \leq H + \epsilon < C$. Formally, there exists a one-to-one function (source code)

$$f : W^n_e \to \{1, \ldots, 2^{nR}\}.$$

Since $R \leq H + \epsilon < C$, by channel coding theorem there exists a $(2^{nR}, n)$ code so that $\lambda_{max} \to 0$. Let

$$g : \mathcal{Y}^n \to \mathcal{V}^n$$

be the joint decoder for channel and source. Hence, if $n$ is big enough, then

$$P(\hat{V}^n \neq V^n) \leq P(V^n \notin W^n_e) + P(g(Y^n) \neq V^n|V^n \in W^n_e) \leq 2\epsilon.$$

In this proof, the two-stage coding was used: at first the source $V^n$ was compressed into its most efficient representation (weakly typical words are the optimal compression when $n$ is big) and then using an appropriate channel code $\mathcal{C}$ (that depends on channel but not on source distribution), it was sent over the channel. On the other hand these two-stages can be combined and design a code that maps the source directly into the input of the channel. This approach is called joint source-channel coding. In the presence of feedback, the joint source-channel code is the sequence of codes $\mathcal{C}_i$, where

$$\mathcal{C}_i : \mathcal{V}^n \times \mathcal{Y}^{i-1} \to \mathcal{X}.$$
Clearly the two-stage coding is a special kind of joint source-channel coding so that best joint source-channel coding achieves at least the same rate as the best two-stage coding. By Theorem 4.8, thus, every rate \( R < C \) is achievable. But can a good joint source-channel coding do better? The answer is, again, no.

**Theorem 4.9 (Separation theorem, source-channel coding theorem)** Assume feedback channel with capacity \( C \). Let \( V = V_1, V_2, \ldots \) be a stationary process on finite alphabet \( \mathcal{V} \) satisfying AEP and having entropy rate \( H \). Let \( \hat{V}^n \) be the output of \( V^n \) after \( n \) transmissions with joint source-channel code. If \( H > C \), then there exists \( \delta > 0 \) so that \( P(\hat{V} \neq V) > \delta \) for every \( n \). Hence \( P(\hat{V} \neq V) \to 0 \) implies \( R \leq C \).

**Proof.** Let

\[ C_i : \mathcal{Y}^n \times \mathcal{Y}^{i-1} \to \mathcal{X}, \quad i = 1, \ldots, n \]

and

\[ g : \mathcal{Y}^n \to \mathcal{V}^n, \quad \hat{V} = g(Y^n). \]

Since source is stationary,

\[ H \leq \frac{H(V_1, \ldots, V_n)}{n} = \frac{1}{n} H(V^n) = \frac{1}{n} H(V^n | \hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n). \]

The first inequality holds, since \( H(V^n | V_1, \ldots, V_{n-1}) \leq H \) and by stationarity

\[ H(V_1, \ldots, V_n) = H(V_1) + \cdots + H(V_n | V_1, \ldots, V_{n-1}) = H(V_n) + H(V_n | V_{n-1}) + \cdots + H(V_n | V_1, \ldots, V_{n-1}) \geq nH(V_n | V_1, \ldots, V_{n-1}). \]

Fano’s inequality (recall \( |\mathcal{V}| < \infty \))

\[ H(V^n | \hat{V}^n) \leq 1 + P(\hat{V}^n \neq V^n) \log |\mathcal{V}| = 1 + P(\hat{V}^n \neq V^n) n \log |\mathcal{V}|. \]

From data processing inequality \( (V^n \to Y^n \to \hat{V}^n) \),

\[ I(V^n; \hat{V}^n) \leq I(V^n; Y^n). \]

From the proof of Theorem 4.7 we know

\[ I(V^n; Y^n) \leq nC. \]

Thus

\[ H \leq \frac{1}{n} + P(\hat{V}^n \neq V^n) \log |\mathcal{V}| + C. \]

If \( P(\hat{V}^n \neq V^n) \to 0 \), then \( H \leq C \).

Finally, if \( H > C \), then

\[ P(\hat{V}^n \neq V^n) \geq \frac{H - C}{\log |\mathcal{V}|} - \frac{1}{n \log |\mathcal{V}|}. \]
and so $H > C$ implies the existence of $\delta' > 0$ such that $\mathbf{P}(\hat{V} \neq V^n) > \delta'$, provided $n$ is big enough. Then, for some $\delta$, the inequality holds for every $n$ i.e. $\mathbf{P}(\hat{V} \neq V^n) > \epsilon$ for every $n$. ■

The theorem implies that the best source code (depending on the source $V$) and the best channel code (depending on the channel) can be designed separately: The combination will be (at least asymptotically) as efficient as anything we could design by considering both problems together. Has practical implications.

### 4.6 Exercises

1. Let $\mathcal{X} = \{0, 1\}$. Consider channel where to input $X$ an independent noise $aZ$ is added. Let $Z \sim B(1, 0.5)$. Find the capacity of this channel.

2. Let $\mathcal{X} = \{0, \ldots, 10\}$. Consider the channel, where $Y = X + Z \pmod{11}$, $X$ is input, $Y$ is output and $Z$ is independent of $X$. Let the distribution of $Z$ be

<table>
<thead>
<tr>
<th>$Z$</th>
<th>1/3</th>
<th>2/3</th>
<th>3/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>1/3</td>
<td>2/3</td>
<td>3/3</td>
</tr>
</tbody>
</table>

Find the capacity of this channel and distribution $P^*$.

3. Let $(\mathcal{X}_1, P_1(y|x), \mathcal{Y}_1)$ and $(\mathcal{X}_2, P_2(y|x), \mathcal{Y}_2)$ two channels with capacities $C_1$ and $C_2$, resp. Define the product channel

$$(\mathcal{X}_1 \times \mathcal{X}_2, P_1(y_1|x_1)P_2(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2).$$

Find its capacity.

4. Let $K(\epsilon)$ be a BSC with error probability $\epsilon$ and capacity $C(\epsilon)$. $\epsilon$. Let $K(\epsilon_1) \rightarrow K(\epsilon_2)$ be a cascade.

- Find the capacity of the cascade $C$.
- Prove that $C \leq C(K(\epsilon_1)) \wedge C(K(\epsilon_2))$.
- Prove that the $n$-fold cascade

$$X \rightarrow K(\epsilon) \rightarrow K(\epsilon) \rightarrow \cdots \rightarrow K(\epsilon) \rightarrow Y(n)$$

is the same as $K(\frac{1}{2}(1 - (1 - 2\epsilon)^n))$, and hence $\lim_n I(X; Y(n)) = 0$.

5. Find the capacity and $P^*$ of the following $Z$-channel

$$
\begin{pmatrix}
1 & 0 \\
0.5 & 0.5
\end{pmatrix}
$$
6. Channels with memory. Consider binary symmetric channels \( Y_i = X_i + Z_i \) (mod 2), where \( X = Y = \{0,1\} \). Let \( Z_1, \ldots, Z_n \) have the same distribution but not independent, \( Z_i \sim B(1, \epsilon) \). Assume that vector \( Z^n = (Z_1, \ldots, Z_n) \) is independent of vector \( X^n = (X_1, \ldots, X_n) \). Hence we have \( n \) BSCs with probability error \( \epsilon \) and capacity \( C(\epsilon) \). When \( Z_i \) are dependent, these channels have memory.

- Prove that \( I(X^n;Y^n) \leq n - h(\epsilon) \). Find the distribution of \( X^n \) and \( Z^n \) so that the inequality is an equality.
- Show that memory increases the capacity of channel:

\[
\max_{P(x^n)} I(X^n, Y^n) > nC(\epsilon).
\]

7. Let \( (X, P_1, X) \) and \( (X, P_2, X) \) be two channels with capacities \( C_1 \) and \( C_2 \). Let \( C \) the capacity of cascade \( P_1 \to P_2 \). Prove

\[
C \leq C_1 \wedge C_2.
\]

8. Let \( x^n(1), \ldots, x^n(2^{nR}) \) be a codebook. Let the decoder \( g \) be the maximal likelihood decoder:

\[
g(y^n) := \arg \max_i P(y^n|x^n(i)) = \arg \max_i P(Y^n = y^n|W = i).
\]

Assume \( W \) has uniform distribution.

- Prove that \( g \) has minimal error probability over all possible decoders:

\[
P_e = P(g(Y^n) \neq W) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} P(g(Y^n) \neq i|W = i) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i.
\]

- Find a counterexample to show that \( g \) does not minimize \( \lambda_{max} \) over all decoding functions.

**Hint:** Show that

\[
\arg \max_i P(Y^n = y^n|W = i) = \arg \max_i P(W = i|Y^n = y^n) =: g^*(y^n).
\]

Then show that for every other decoding function \( g \)

\[
P(W \neq g^*(y^n)|Y^n = y^n) \leq P(W \neq g(y^n)|Y^n = y^n), \quad \forall y^n.
\]

9. Let \( K(\epsilon) \) BSC with \( \epsilon < \frac{1}{2} \). Let \( x^n(1), \ldots, x^n(2^{nR}) \) a codebook. For every two vectors \( x^n, y^n \in \{0,1\} \) define Hamming distance

\[
d(x^n, y^n) = \sum_{i=1}^{n} |x_i - y_i|.
\]
Let the decoding function be

\[ g(y^n) = \arg \min_i d(y^n, x^n(i)). \]

Prove that \( g \) is the maximal likelihood decoder as in the previous exercise.

10. Let \( \mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3, 4\} \). Let the channel be given by the following transition matrix:

\[
\begin{pmatrix}
    0 & 1 & 0 & 0 & 1 \\
    1 & 0 & 1 & 0 & 0 \\
    0 & 1 & 0 & 1 & 0 \\
    0 & 0 & 1 & 0 & 1 \\
    1 & 0 & 0 & 1 & 0
\end{pmatrix}
\]

Find the codebook \( x^2(1), \ldots, x^2(5) \) so that every word can be transmitted without errors: there exists \( g \) such that \( P(g(Y^2) = i | W = i) = 0 \) for every \( i = 1, \ldots, 5 \).
5 Information theory and statistics

The method of types considered in this section was mainly developed in 1980's by I. Csiszar and J. Körner.

NB! Throughout the chapter we assume finite alphabet, i.e. $|\mathcal{X}| < \infty$.

5.1 Types

Let $\mathcal{X} = \{a_1, \ldots, a_m\}$ be a finite alphabet. We consider a sequence $x^n \in \mathcal{X}^n$ and we shall denote by $N_{x^n}(a)$ the number of times the symbol $a$ occurs in $x^n$ (the frequency of $a$). Thus

$$N_{x^n}(a) := \sum_{i=1}^{n} I_{x_i}(a).$$

**Def 5.1** The type $P_{x^n}$ of a sequence $x^n$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$, i.e

$$P_{x^n}(a) := \frac{N_{x^n}(a)}{n}.$$  

In other words, the type of $x^n$ is the empirical distribution of $x^n$.

Note that $P_{x^n}(a)$ is a probability distribution on $\mathcal{X}$ with the following property: the probability of each letter is on form $\frac{k}{n}$, where $k \in \mathbb{Z}^+$. On the other hand, every distribution with this property is a type for a sequence with length $n$. These probability distributions will be called $n$-types

**Def 5.2** Let $\mathcal{P}_n$ be set of all $n$-types.

Examples:

1. If $\mathcal{X} = \{0, 1\}$, then

$$\mathcal{P}_n = \{(0, 1), \left(\frac{1}{n}, \frac{n-1}{n}\right), \ldots, \left(\frac{n-1}{n}, \frac{1}{n}\right), (1, 0)\},$$

all together $n + 1$ types.

2. If $\mathcal{X} = \{a, b, c\}$, then

$$\mathcal{P}_n = \{(0, 0, 1), (0, \frac{1}{n}, \frac{n-1}{n}), \ldots, (0, 1, 0), \ldots, (\frac{1}{n}, \frac{n-1}{n}, 0), \ldots, (\frac{n-1}{n}, \frac{1}{n}, 0), (1, 0, 0)\},$$

the number of types:

$$\frac{(n + 2)(n + 1)}{2}.$$
The usefulness of the method of types arises from the fact that the number of \( n \)-types is at most polynomial in \( n \): 

\[
|\mathcal{P}_n| \leq (n + 1)^{|\mathcal{X}| - 1}. \tag{5.1}
\]

To see this bound, notice that for a \( n \)-type, the probability \( P_x^n(a) \) is one of \( (n + 1) \) numbers: \( 0/n, \ldots, 1/n \). Hence for the first \( |\mathcal{X}| - 1 \) letters there are at most \( (n + 1)^{|\mathcal{X}| - 1} \) different choices. Since the probabilities sum up to one, the last one is determined by others.

The exact number of \( n \)-types is

\[
|\mathcal{P}_n| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1}.
\]

**Random i.i.d. sequences and types.** We now consider i.i.d. random variables \( X_1, \ldots, X_n \) from alphabet \( \mathcal{X} \) having distribution \( Q \). Then, for any sequence \( x^n \), the probability is

\[
Q^n(x^n) := \prod_{i=1}^n Q(x_i).
\]

**Lemma 5.1** Let \( X_1, \ldots, X_n \) i.i.d. sequences on \( \mathcal{X} \) having distribution \( Q \). Then its probability depends only on type \( P_{x^n} \), namely

\[
Q^n(x^n) = 2^{-n(H(P_{x^n}) + D(P_{x^n} \| Q))}. \tag{5.2}
\]

The proof of the equality is Exercise 1. The equality (5.2) holds for infinite alphabet as well.

Let \( x^n \) be given (maximum likelihood estimator). Let \( \mathcal{P} \) be the set of all probabilities on \( \mathcal{X} \). Let us find the probability under which the probability of obtaining \( x^n \) is maximal? In other words, let us find the maximal likelihood estimator (MLE) \( \hat{Q} \) of unknown \( Q \)? From Lemma 5.1, it follows that the the MLE of \( Q \) is \( P_{x^n} \). Indeed,

\[
\hat{Q} := \arg\max_{Q \in \mathcal{P}} Q^n(x^n) = \arg\min_{Q \in \mathcal{P}} (H(P_{x^n}) + D(P_{x^n} \| Q)) = \arg\min_{Q \in \mathcal{P}} D(P_{x^n} \| Q) = P_{x^n}.
\]

If the maximum likelihood estimator is searched from a subset (model) \( \mathcal{P}_0 \subset \mathcal{P} \), then

\[
\hat{Q} = \arg\min_{Q \in \mathcal{P}_0} D(P_{x^n} \| Q),
\]

i.e. MLE is the best approximation of \( P_{x^n} \) from \( \mathcal{P}_0 \) in K-L sense. If \( \mathcal{P}_0 \) contains \( P_{x^n} \), then \( \hat{Q} = P_{x^n} \), otherwise \( \hat{Q} \) is the distribution from \( \mathcal{P}_0 \) that in a sense (K-L) is closest to \( P_{x^n} \).

Let \( Q \in \mathcal{P}_n \) be fixed (maximal likelihood realization). When source distribution \( Q \) is fixed, then the realization \( x^n \) with highest probability is not necessarily the one with type \( Q \) even when \( Q \in \mathcal{P}_n \). Indeed, from Lemma 5.1, it follows that \( x^n \) with largest probability has type

\[
\arg\min_{P \in \mathcal{P}_n} (H(P) + D(P \| Q))
\]

86
that need not be $Q$. For instance, if $\mathcal{X} = \{0, 1\}$ and $Q = B(1, q)$, $q > 0.5$, then the realization with highest probability consists of ones, only. Indeed, $B(1, 1)$ is the distribution minimizing $P \mapsto H(P) + D(P \| Q)$. To show that is Exercise 2.

**Type classes.** All sequences $x^n$ can be divided into equivalence classes according to their types.

**Def 5.3** Let $P \in \mathcal{P}_n$. The **type class of $P$** is the set of $n$-elemental sequences

$$T(P) := \{ x^n \in \mathcal{X}^n : P_x^n = P \}.$$ 

**Example:** Let $\mathcal{X} = \{0, 1, 2\}$, $x^5 = 00210$. Type $P_{x^5}$ is

$$\begin{array}{ccc}
0 & 1 & 2 \\
\frac{3}{5} & \frac{1}{5} & \frac{1}{5}
\end{array}$$

The type class $T(P_{x^5})$ is

$$T(P_{x^5}) = \{(00012), (00021), (00102), \ldots, (12000)\}.$$ 

Hence the size of the type class of $P_{x^n}$ is

$$|T(P_{x^n})| = \frac{5!}{3!1!1!} = 20.$$ 

It is important to note that the number of types increases polynomially, whilst the number of sequence increases exponentially. It means that at least one class has exponentially many sequences. The exact number of elements in every type class is rather easy to find: for every $P \in \mathcal{P}_n$:

$$|T(P)| = \frac{n!}{(nP(a_1))!(nP(a_2))! \cdots (nP(a_m))!}.$$ 

This number is not easy to deal with, so we shall give simple estimates to $|T(P)|$.

**Theorem 5.4** For every $P \in \mathcal{P}_n$, it holds

$$1 \geq P^n(T(P)) := \sum_{x^n \in T(P)} P^n(x^n) = |T(P)| 2^{-nH(P)},$$ 

so that

$$T(P) \leq 2^{nH(P)}.$$ 

**Proof.** Let $P \in \mathcal{P}_n$. From (5.2), it follows

$$1 \geq P^n(T(P)) := \sum_{x^n \in T(P)} P^n(x^n) = |T(P)| 2^{-nH(P)},$$ 

so that

$$T(P) \leq 2^{nH(P)}.$$ 

Lower bound. For every $R \in \mathcal{P}_n$, it holds

$$P^n(T(P)) \geq P^n(T(R)).$$ (5.4)
Inequality (5.4) states: if $X_1, \ldots, X_n$ are i.i.d. random variables with distribution $P \in \mathcal{P}_n$, the class $T(P)$ has biggest probability.

We shall now prove (5.4). Recall that $P, R \in \mathcal{P}_n$.

$$\frac{P^n(T(P))}{P^n(T(R))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(R)| \prod_{a \in \mathcal{X}} P(a)^{nR(a)}}.$$ 

It holds

$$\frac{|T(P)|}{|T(R)|} = \frac{n!}{(nP(a_1))!(nP(a_2))! \cdots (nP(a_m))!} \div \frac{n!}{(nR(a_1))!(nR(a_2))! \cdots (nR(a_m))!} = \prod_{a \in \mathcal{X}} \frac{(nR(a))!}{(nP(a))!}.$$ 

For every $n, m$

$$\frac{m!}{n!} \geq n^{m-n}.$$ 

Hence

$$\frac{(nR(a))!}{(nP(a))!} \geq (nP(a))^{nR(a)-nP(a)}$$

or

$$\frac{P^n(T(P))}{P^n(T(R))} = \prod_{a \in \mathcal{X}} \frac{(nR(a))!}{(nP(a))!} \frac{P(a)^{nP(a)-R(a)}}{P(a)^{nP(a)-P(a)}} \geq \prod_{a \in \mathcal{X}} (nP(a))^{n(R(a)-P(a))} \frac{P(a)^{nP(a)-R(a)}}{P(a)^{nP(a)-P(a)}} = \prod_{a \in \mathcal{X}} n^{n(R(a)-P(a))} = n^{n(\sum a R(a) - \sum a P(a))} = n^{n(1-1)} = 1.$$ 

Now we obtain the lower bound to $|T(P)|$:

$$1 = \sum_{R \in \mathcal{P}_n} P^n(T(R)) \leq \sum_{R \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) = \sum_{R \in \mathcal{P}_n} P^n(T(P))$$

$$\leq (n + 1)^{|\mathcal{X}|} P^n(T(P)) = (n + 1)^{|\mathcal{X}|} \sum_{x^n \in T(P)} P^n(x^n)$$

$$= (n + 1)^{|\mathcal{X}|} \sum_{x^n \in T(P)} 2^{-nH(P)} = (n + 1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)},$$

so that

$$|T(P)| \geq \frac{2^{nH(P)}}{(n + 1)^{|\mathcal{X}|}}.$$ 

The second last equality again follows from (5.2). ■

**Remark:** Obviously the lower bound in (5.3) can be improved by replacing the bound $(n + 1)^{|\mathcal{X}|}$ by better estimate $(n + 1)^{|\mathcal{X}|-1}$ or even by the exact number of type classes.
Example: Let $X = \{a, b\}$ and $P \in \mathcal{P}_n$. Then there exists $k$ so that $P(a) = \frac{k}{n}$, $P(b) = \frac{n-k}{n}$ and $H(P) = h(\frac{k}{n})$. By two-letter alphabet

$$|T(P)| = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$  

The number of type classes is $(n+1)$. Hence, from Theorem 5.4 we get

$$\frac{1}{n+1} 2^{nh(\frac{k}{n})} \leq \binom{n}{k} \leq 2^{nh(\frac{k}{n})}.$$  

This inequality is sometimes very useful. When $k = \alpha n$, then from the inequality above, it follows

$$\log \binom{n}{k} \sim nh(\alpha).$$

Theorem 5.5 Let $X_1, \ldots, X_n$ i.i.d. random variables having distribution $Q$. Then for every $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|X|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}. \quad (5.5)$$

Proof. From Lemma 5.1

$$Q^n(T(P)) = \sum_{x^n \in T(P)} Q^n(x^n) = \sum_{x^n \in T(P)} 2^{-n(D(P||Q)+H(P))} = |T(P)| 2^{-n(D(P||Q)+H(P))}.$$  

From (5.3)

$$Q^n(T(P)) \leq 2^{nH(P)} 2^{-n(D(P||Q)+H(P))} = 2^{-nD(P||Q)}$$

and

$$Q^n(T(P)) \geq (n+1)^{-|X|} 2^{nH(P)} 2^{-n(D(P||Q)+H(P))} = (n+1)^{-|X|} 2^{-nD(P||Q)}.$$  

To recapitulate:

1. The set $X^n$ can be divided into type classes, the number of classes is bounded above by $(n+1)^{|X|}$.

2. For every $x^n \in T(P)$ with $P \in \mathcal{P}_n$

$$Q^n(x^n) = 2^{-n(H(P)+D(P||Q))}.$$  

3. The capacity of type class $T(P)$ is about $2^{nH(P)}$.

4. For every $T(P)$, with $P \in \mathcal{P}_n$, the probability $Q^n(T(P))$ is about $2^{-nD(P||Q)}$. If $Q \in \mathcal{P}_n$, then

$$Q^n(T(Q)) \geq Q^n(T(P)) \quad \forall P \in \mathcal{P}_n.$$  

89
5.2 Sanov’s theorem

5.2.1 Large deviation inequalities

Using the method of types, it is easy to prove very important results in probability – *large deviation inequalities*. In their most common form, large deviation inequalities measure how well the sample average of an i.i.d. sample is concentrated around the mean $EX_i$. Therefore, such inequalities are also called *concentration inequalities*.

**Some well-known large deviation (concentration) inequalities.** Let $X_1, \ldots, X_n$ be i.i.d. random variables. Let us recall some well-known large deviation inequalities. In these inequalities $\mu := EX_i$, $\epsilon > 0$ and

$$S_n := X_1 + \cdots + X_n.$$

**Markov inequality:**

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{E|X_1 - \mu|}{\epsilon}.$$  

**Chebyshev inequality:**

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\text{Var}(X_1)}{\epsilon^2 n}.$$  

**Hoeffding inequality:** If, for some constants $a < b$, $a \leq X_i \leq b$, then

$$P\left(\frac{S_n}{n} - \mu \geq \epsilon\right) \leq \exp\left[-\frac{2\epsilon^2}{(b-a)^2 n}\right]. \quad (5.6)$$

From this, it immediately follows that

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq 2 \exp\left[-\frac{2\epsilon^2}{(b-a)^2 n}\right].$$

**Large deviation bound:** If, for $t > 0$, it holds

$$M(t) := E \exp[tX_1] < \infty$$

(the moments generating function of $X_1$ is finite in a neighborhood of zero), then

$$P\left(\frac{S_n}{n} - \mu \geq \epsilon\right) \leq \exp[-n\Lambda^*(\mu + \epsilon)], \quad (5.7)$$

where for every $x \in \mathbb{R}$,

$$\Lambda^*(x) := \sup_{t \in \mathbb{R}} (xt - \Lambda(t)), \quad \Lambda(t) := \ln M(t).$$
From Chebyshev inequality, the weak law of large numbers follows; from Hoeffding’s inequality or large deviation inequality the strong law of large number (via Borel-Cantelli lemma) follows. The exponent in the last inequality (5.7) cannot be improved, because the following large deviation principle:

$$\lim_{n \to \infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} - \mu \geq \epsilon\right) = -\Lambda^*(\mu + \epsilon).$$

(5.8)

We shall now prove the convergence (5.8) via method of types giving another interpretation to the limit $\Lambda^*(\mu + \epsilon)$.

**Sanov’s theorem.** Recall $|\mathcal{X}| < \infty$. Let $\mathcal{P}$ be the set of all distributions on $\mathcal{X}$ and $\mathcal{P}_n$ the set of types. For any distribution $Q$ and subset $E \subset \mathcal{P}$, let

$$d(E, Q) := \inf_{P \in E} D(P||Q).$$

Our main object of interest is the probability

$$Q^n(E) := \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)).$$

(5.9)

Hence, if $X_1, \ldots, X_n$ are i.i.d. random variables with distribution $Q$, then $Q^n(E)$ is the probability that a realization belongs to a type class from the set $E \cap \mathcal{P}_n$.

Recall that for any distribution $P \in \mathcal{P}$, the support of $P$ is

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

Define $E_Q$ as the measures from $E$ with support belonging to the support of $Q$:

$$E_Q := \{P \in E : \mathcal{X}_P \subseteq \mathcal{X}_Q\}.$$

If $\mathcal{X}_Q = \mathcal{X}$, then $E_Q = E$. Notice that when a letter $a$ has zero probability, i.e. $Q(a) = 0$, then for any type $P$ with $P(a) > 0$, it holds $Q^n(T(P)) = 0$. Therefore $Q^n(E) = Q^n(E_Q)$ so that the set of interest is actually $E_Q$. For any distribution $P$, following implications hold

$$D(P||Q) < \infty \quad \iff \quad \mathcal{X}_P \subseteq \mathcal{X}_Q.$$

(5.10)

Thus, if $d(E, Q) < \infty$, then $d(E, Q) = d(E_Q, Q)$. Finally notice the following facts:

1. when $E$ is closed in $\mathbb{R}^{|\mathcal{X}|}$, then so is $E_Q$;
2. for every $E$, the function $P \mapsto D(P||Q)$ is continuous on $E_Q$.

Proof of 1. and 2. is Exercise 3.
Proposition 5.1 Let $E$ be closed. Then there exists $P^* \in E$ such that

$$d(E, Q) = D(P^*||Q).$$

If $d(E, Q) < \infty$, then $P^* \in E_Q$.

Proof. A closed set is non-empty. If $d(E, Q) = \infty$, then the existence of $P^*$ is trivial (why?). Thus consider the case $d(E, Q) < \infty$. Let $P_n \subset E$ be a sequence satisfying $D(P_n||Q) \rightarrow d(E, Q)$. Clearly such a sequence exists (why?). Since $d(E, Q) < \infty$, then also $D(P_n||Q) < \infty$. By (5.10), this means that for every $n$, $P_n \in E_Q$. The set $E$ is closed and therefore also $E_Q$ is closed (claim 1 above). Consequently $E_Q$ is compact on $\mathbb{R}^{[X]}$. Compactness implies the existence of a subsequence $P_{n_k}$ so that $P_{n_k} \rightarrow P^*$ where $P^* \in E_Q$. Since $P \mapsto D(P||Q)$ is continuous on $E_Q$ (claim 2 above), then $D(P_{n_k}||Q) \rightarrow D(P^*||Q)$. Hence $D(P^*||Q) = d(E, Q)$. \[\blacksquare\]

Theorem 5.6 (Sanov’s theorem) Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution $Q$. Let $E \subset \mathcal{P}$ be a set of probability distributions. Then

$$Q^n(E) \leq (n + 1)^{|X|}2^{-nd(E, Q)}.$$ (5.11)

If, in addition, the set $E_Q$ is the closure of its interior, then there exists $P^* \in E_Q$ such that $d(E, Q) = D(P^*||Q)$ and

$$\frac{1}{n} \log Q^n(E) \rightarrow -d(E, Q).$$ (5.12)

Proof. The proof is based on inequalities (5.5) and $|\mathcal{P}_n| \leq (n + 1)^{|X|}$:

$$Q^n(E) = \sum_{P \in \mathcal{P} \cap P_n} Q^n(T(P))$$
$$\leq \sum_{P \in \mathcal{P} \cap P_n} 2^{-D(P||Q)n}$$
$$\leq \sum_{P \in \mathcal{P} \cap P_n} \max_{P \in \mathcal{P} \cap P_n} 2^{-D(P||Q)n}$$
$$\leq (n + 1)^{|X|}2^{-d(E \cap P_n, Q)n}$$
$$\leq (n + 1)^{|X|}2^{-d(E, Q)n},$$

because

$$d(E \cap P_n, Q) = \inf_{P \in E \cap P_n} D(P||Q) \geq \inf_{P \in E} D(P||Q) = d(E, Q).$$

Proof of converse. By assumption, the interior of $E_Q$ is non-empty so that $d(E_Q, Q) = d(E, Q) < \infty$. Since $E_Q$ is closed, by Proposition 5.1, there exists $P^* \in E_Q$ such that $d(E, Q) = D(P^*||Q)$. The proof is based on the fact that $\cup_n \mathcal{P}_n$ is dense in $\mathcal{P}$. Since the interior of $E_Q$ is non-empty, we obtain that the set

$$E_Q \cap (\cup_n \mathcal{P}_n)$$

92
is dense in $E_Q$ meaning that for every $P \in E_Q$ there exists a sequence $P_n \in (\cup_{n} P_n) \cap E_Q$ such that $P_n \to P$. W.l.o.g. choose $P_n \in P_n$ (why this is possible?). Hence, there exists $P_n \in P_n \cap E_Q$ such that $P_n \to P^*$ and since $P \mapsto D(P||Q)$ is continuous on $E_Q$, it follows $D(P_n||Q) \to D(P^*||Q) = d(E, Q)$. Now

$$Q^n(E) \geq Q^n(T(P_n)) \geq (n + 1)^{-|X|}2^{-D(P_n||Q)},$$

so that

$$\liminf \frac{1}{n} \log Q^n(E) \geq \liminf \left( -\frac{|X| \log(n + 1)}{n} - D(P_n||Q) \right) = -D(P^*||Q) = -d(E, Q).$$

Corollary 5.1  Let $E \subset \mathcal{P}$ be such that $E_Q$ belongs to the closure of its interior. Then the convergence (5.12) holds.


Corollary 5.2  If $E \subset \mathcal{P}$ is non-empty and open, then the convergence (5.12) holds.

Proof. A non-empty open set is always contained in the closure of its interior. Hence Corollary 5.1 applies.

Large deviation principle. From (5.11), it follows that for every set $E$,

$$\limsup \frac{1}{n} \log Q^n(E) \leq -d(E, Q).$$

If, in addition $E_Q$ belongs to the closure of its interior, then

$$\lim \frac{1}{n} \log Q^n(E) = -d(E, Q).$$

It turns out that Theorem 5.6 can generalize as follows.

Theorem 5.7  (Large deviation principle) For every $E \subset \mathcal{P}$,

$$-d(E^o, Q) \leq \liminf \frac{1}{n} \log Q^n(E) \leq \limsup \frac{1}{n} \log Q^n(E) \leq -d(E, Q),$$

where $E^o$ is the interior of $E$ and $d(\emptyset, Q) := \infty$.

The right hand side follows from Sanov’s theorem (Theorem 5.6) and the left hand side can be proven similarly as in Theorem 5.1. The complete proof can be found in (Dembo, Zeitouni, Thm 2.10.).

Examples. The rigorous proofs of all examples is Exercise 5.
1. Let $|\mathcal{X}| = 4$, $E := \{P\}$, where $P = (\frac{1}{2}, \frac{1}{2}, 0, 0)$. Let $Q = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Show that

$$Q^n(E) = \begin{cases} (\frac{n}{2})\left(\frac{1}{2}\right)^n & \text{when } n \text{ even;} \\ 0, & \text{when } n \text{ odd.} \end{cases}$$

Thus $E = E_Q$ is not contained in the closure its interior. Prove that

$$-\infty = -d(E^o, Q) = \lim \inf_n \frac{1}{n} \log Q^n(E) < \lim \sup_n \frac{1}{n} \log Q^n(E) = -d(E, Q) = -1,$$

so that the right-most and left-most inequalities in (5.13) are equalities.

2. Let $|\mathcal{X}| = 3$, $Q = (\frac{1}{2}, \frac{1}{2}, 0)$, $P_1 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $P_2 = (0, \frac{1}{2}, \frac{1}{2})$, $P_3 = Q$ and

$$E = \text{conv}\{P_1, P_2, P_3\} \text{ i.e. } E = \left\{\lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3 : \lambda_i \geq 0, \sum \lambda_i = 1\right\}.$$

The set $E$ is the closure its interior, but $E_Q = \{Q\}$ and hence $E_Q$ is not the closure it interior. Show that

$$Q^n(E) = \begin{cases} (\frac{n}{2})\left(\frac{1}{2}\right)^n & \text{when } n \text{ even;} \\ 0, & \text{when } n \text{ odd.} \end{cases}$$

and

$$-\infty = -d(E^o, Q) = \lim \inf_n \frac{1}{n} \log Q^n(E) < \lim \sup_n \frac{1}{n} \log Q^n(E) = -d(E, Q) = 0.$$

3. Let $|\mathcal{X}| = 3$, $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $P_1 = (\frac{1}{2}, \frac{1}{2}, 0)$, $P_2 = (0, \frac{1}{2}, \frac{1}{2})$, $P_3 = Q$ and

$$E = \text{conv}\{P_1, P_2, P_3\}.$$

Show that $E = E_Q$ and therefore the assumptions of converse of Theorem 5.6 hold. This implies the convergence

$$\lim \frac{1}{n} \log Q^n(E) = 0.$$

On the other hand $d(E^o; Q) = 0$ so the convergence above follows also from large deviation principle (5.13).

4. Let $|\mathcal{X}| = 4$,

$$E = \{P \in \mathcal{P} : P(a_1) = P(a_2) = 0\}.$$ 

Show that $E$ is the closure of its interior. Let $Q = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Find $d(E, Q)$, $d(E^o, Q)$ and $P^*$. Show that $Q^n(E) = 2^{-d(E,Q)n}$ and deduce the convergence (5.12). Show that $d(E, Q) = d(E^o, Q)$ and the convergence (5.12) follows from large deviation principle (5.13).

5. Let $|\mathcal{X}| = 4$, $E = \{P \in \mathcal{P} : P(a_1) = P(a_2) = 0\}$, $Q = (q_1, q_2, q_3, 0)$, where $q_i > 0$, $i = 1, 2, 3$. Find $E_Q$ and $P^*$. Show that $d(E, Q) > 0$ but $d(E^o, Q) = -\infty$. Show that $Q^n(E) = 2^{-d(E,Q)n}$ and deduce the convergence (5.12).
5.2.2 Sanov’s theorem and large deviation inequalities

How from Sanov’s theorem the large deviation inequalities follow? Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{X}$-valued random variables $X_i \sim Q$. Let $g : \mathcal{X} \to \mathbb{R}$ be a function. Large deviation inequalities bound the probabilities

$$P\left( \frac{1}{n} \sum_{i}^{n} g(X_i) \geq c \right). \quad (5.14)$$

If $\mathcal{X} \subset \mathbb{R}$, $g(x) = x$, $\mu := EX_1$ and $c = \mu + \epsilon$, then (5.14) is the probability

$$P\left( \frac{S_n}{n} - \mu \geq \epsilon \right).$$

Clearly the probability (5.14) can be written as

$$Q^n\left( \{x^n \in \mathcal{X}^n : \frac{1}{n} \sum_{j=1}^{n} g(x_j) \geq c \} \right). \quad (5.15)$$

Sanov’s theorem is applicable, because whether $x^n$ belongs to the set

$$\{ x^n \in \mathcal{X}^n : \frac{1}{n} \sum_{j=1}^{n} g(x_j) \geq c \}$$

or not, depends on the type of $x^n$, only. Indeed:

$$\frac{1}{n} \sum_{j=1}^{n} g(x_j) = \sum_{a \in \mathcal{X}} g(a) P_{x^n}(a),$$

so

$$\frac{1}{n} \sum_{j=1}^{n} g(x_j) \geq c \iff \sum_{a \in \mathcal{X}} g(a) P_{x^n}(a) \geq c \iff P_{x^n} \in E \cap \mathcal{P}_n,$$

where

$$E := \{ P : \sum_{a \in \mathcal{X}} g(a) P(a) \geq c \}. \quad (5.16)$$

Hence, with $E$ being as in (5.16), the probability (5.15) is

$$\sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) = Q^n(E).$$

Let

$$\overline{g} = \max_{a \in \mathcal{X}_Q} g(a).$$
If $c < \overline{g}$, then clearly the set $E_Q$ with $E$ as in (5.16) is the closure of its interior so that from Sanov’s theorem, it follows that

$$\lim_{n} \frac{1}{n} \log P \left( \frac{1}{n} \sum_{i} g(X_{i}) \geq c \right) = -d(E, Q). \quad (5.17)$$

When $c = \overline{g}$, then (5.17) holds as well (Exercise 6) and when $c > \overline{g}$, then the probability(5.14) equals to zero and $d(E, Q) = \infty$, so that the convergence (5.17) holds for every $c$.

When $c \leq \mathbb{E} g(X_{1}) = \sum_{a \in X} g(a) Q(a)$, then $Q \in E$ so that $d(E, Q) = 0$ (as it obviously also follows from CLT).

Finally notice that when $X \subset \mathbb{R}$, $g(x) = x$ and $c = \mu + \epsilon$ from (5.8) and (5.17), it follows that for any $\epsilon > 0$,

$$\Lambda^{*} (\mu + \epsilon) = \ln 2 d(E, Q).$$

**Finding $d(E, Q)$**. Let $E$ be as in (5.16). Since $E$ is closed, there exists $P^{*} \in E$ such that

$$D(P^{*}||Q) = d(E, Q) = \min_{P \in E} D(P||Q).$$

Knowing $P^{*}$, the exponent $d(E, Q)$ is easy to find. The following lemma helps to find $P^{*}$.

**Lemma 5.2** Let $P^{*}$ a distribution on $\mathcal{X}$ in the form

$$P^{*}(a) = \frac{Q(a)2^{\lambda g(a)}}{\sum_{a \in \mathcal{X}} Q(a)2^{\lambda g(a)}}, \quad (5.18)$$

where $\lambda \geq 0$ is such that

$$\sum_{a \in \mathcal{X}} g(a) P^{*}(a) = c. \quad (5.19)$$

If $E$ is as in (5.16), then

$$D(P^{*}||Q) = \min_{P \in E} D(P||Q) = d(E, Q).$$

**Proof.** Let $P \in E$ be arbitrary. Let us show that

$$D(P||Q) \geq D(P^{*}||Q) + D(P||P^{*}) \geq D(P^{*}||Q). \quad (5.20)$$

Let $P \in E$, i.e. $\sum_{a \in \mathcal{X}} g(a) P(a) \geq c$. Find

$$D(P||Q) = \sum_{a} P(a) \log \frac{P(a)}{Q(a)}$$

$$= \sum_{a} P(a) \log \frac{P(a)}{P^{*}(a)} + \sum_{a} P(a) \log \frac{P^{*}(a)}{Q(a)}$$

$$= D(P||P^{*}) + \sum_{a} P(a) \log \frac{P^{*}(a)}{Q(a)}. \quad (5.20)$$
Estimate
\[ \sum_a P(a) \log \frac{P^*(a)}{Q(a)}. \]

Let
\[ C := \sum_a Q(a) 2^{\lambda g(a)}, \]
so that
\[ \frac{P^*(a)}{Q(a)} = \frac{2^{\lambda g(a)}}{C}. \]

Then, since \( \sum_a g(a)P(a) \geq c \), we have
\[
\sum_a P(a) \log \frac{P^*(a)}{Q(a)} = \sum_a \left( \lambda g(a) - \log C \right) P(a) \\
= \lambda \sum_a g(a)P(a) - \log C \\
\geq \lambda c - \log C \\
= \sum_a \left( \lambda g(a) - \log C \right) P^*(a) \\
= \sum_a P^*(a) \log \frac{P^*(a)}{Q(a)} \\
= D(P^*||Q).
\]

\[ \square \]

Remark: The inequality (5.20): for any \( P \in E \)

\[ D(P||Q) \geq D(P^*||Q) + D(P||P^*) \]

is essentially Pythagorean theorem. It holds for any closed convex set not just for \( E \) as in (5.16) (T. Cover, J. Thomas, Thm 12.6.1). Note that when \( d(E, Q) < \infty \), then by strict convexity of \( P \mapsto D(P||Q) \), for convex \( E \), the best approximation \( P^* \) is unique.
Examples:

**Coins (Sanov):** Let $\mathcal{X} = \{0, 1\}$, $X_1, \ldots, X_n$ be i.i.d. with distribution $Q = B(1, q)$.

Estimate

$$P\left(\frac{S_n}{n} \geq c\right),$$

where $S_n = X_1 + \cdots + X_n$ and $q < c < 1$. Set

$$E = \{P : \sum_a P(a) a = P(1) \geq c\} = \{B(1, p) : p \geq c\}.$$

Hence the only distribution satisfying $\sum_a g(a) P(a) = c$ is $P^* = B(1, c)$ and

$$D(P^*||Q) = c \log \frac{c}{q} + (1 - c) \log \frac{1 - c}{1 - q}.$$

For example, if $q = \frac{1}{2}$, then $D(P^*||Q) = 1 - h(c)$. When $c = 0.7$, then $1 - h(c) = 1 - h(0.7) = 0.119$. Thus the probability of 700 or more heads in 1000 trials is approximately

$$2^{-1000(1-h(0.7))} = 2^{-119}.$$

The exact bound from inequality (5.11):

$$P(S_{1000} \geq 700) \leq (1001)^2 2^{-119} < 2^{10 - 119} = 2^{-109}.$$

**Coins (Hoeffding):** Let us combine the obtained bound with the one via Hoeffding inequality. Then $a = 0$, $b = 1$ and the inequality is

$$P(S_n - ES_n \geq c) \leq \exp[-\frac{2c^2}{n}]. \quad (5.21)$$

Thus the probability of 700 or more heads in 1000 trials is bounded above by

$$P(S_{1000} \geq 700) = P(S_{1000} - 500 \geq 200) \leq e^{-\frac{8000}{1000}} = e^{-80} = 2^{-\frac{80}{\ln 2}} \leq 2^{-115}.$$

Compare: from Sanov’s theorem the best possible exponent is -119. Thus in this example, the exponent from Hoeffding inequality is almost the best one. Moreover, Hoeffding inequality gives the exact bound:

$$P(S_{1000} \geq 700) \leq 2^{-115}$$

and that is even better then the exact bound obtained by (5.11):

$$P(S_{1000} \geq 700) \leq 2^{-109}.$$
Dice (Sanov): Suppose we toss a fair die $n$ times. What is the probability that the average is at least 4?

The set $E$:

$$E = \{ P : \sum_{i=1}^{6} iP(i) \geq 4 \},$$

the distribution $Q$ is uniform over $\{1, 2, 3, 4, 5, 6\}$. Sanov’s theorem:

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q),$$

where

$$P^*(i) \propto 2^{\lambda i}, \quad \sum_{i=1}^{6} iP^*(i) = 4.$$  

Solving numerically, we obtain $\lambda = 0.2519$, hence $P^*$ is

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_i$</td>
<td>0.1031</td>
<td>0.1227</td>
<td>0.1467</td>
<td>0.174</td>
<td>0.2072</td>
<td>0.2468</td>
</tr>
</tbody>
</table>

Now $D(P^*||Q) = \sum_{i=1}^{6} P^*(i) \log 6P^*(i) = \log 6 - H(P^*) = 0.0624$. Thus

$$\frac{1}{n} \log Q^n(E) \to -0.0624.$$  

When $n = 10000$, then

$$Q^n(E) \approx 2^{-624}.$$  

Exact bound

$$Q^n(E) \leq (n + 1)^{52nD(P^*||Q)} \leq 2^{67-624}.$$  

Dice (Hoeffding): Let us compare the obtained bounds with the Hoeffding’s ones. Here $a = 1, b = 6$ and

$$P(S_n \geq (3.5 + 0.5)n) \leq e^{-\frac{(0.5)^2}{2n}}.$$  

Hence the probability that the average of 1000 tosses is at least 4 can be bounded above by

$$P\left(\frac{S_n}{n} \geq 3.5 + 0.5\right) \leq e^{-\frac{5000}{2n}} = e^{-200} = 2^{\frac{-200}{\ln 2}} \approx 2^{-288}.$$  

Recall that the best bound, obtained via Sanov’s theorem is -624. Hence in this case Hoeffding inequality is far from being optimal and also the exact bound obtained by (5.11) is much better:

$$2^{67-624} \ll 2^{-288}.$$
5.3 Hypothesis testing

5.3.1 Simple hypotheses: Neyman-Pearson lemma

Let \( X_1, \ldots, X_n \) be i.i.d. random variables with distribution \( Q \). Consider simple hypotheses:

\[
H_0 : Q = P_0 \\
H_1 : Q = P_1.
\]

In what follows, we assume that \( P_0 \) and \( P_1 \) have the same support and w.l.o.g. it is \( \mathcal{X} \):

\[
X_{P_0} = X_{P_1} = \mathcal{X}.
\]

Let \( x^n = x_1, \ldots, x_n \) be a sample and \( g : \mathcal{X}^n \to \{0, 1\} \) the test: the hypotheses \( H_i \) is accepted iff \( g(x^n) = 1 \). Hence test \( g \) is \( I_{A_n} \), where \( A_n \subseteq \mathcal{X}^n \) is critical region of rejecting \( H_0 \) and accepting \( H_1 \). In other words: \( H_1 \) is accepted iff \( x^n \in A_n \), otherwise \( H_0 \) is accepted.

Two types of errors:

\[
\alpha := P(g(X_1, \ldots, X_n) = 1 | H_0) = P_0^n(A_n), \quad \beta := P(g(X_1, \ldots, X_n) = 0 | H_1) = P_1^n(A_n^c).
\]

The probability \( \alpha \) is the probability of false positive, often called the probability of type I error; the probability \( \beta \) is the probability of false negative and is called the probability of type II error. One aims to design test so that both probabilities were minimal, but there is a trade-off: decreasing one of them implies increasing the another. Since \( P_0 \) and \( P_1 \) have equal supports, then \( \alpha = 0 \) iff \( \beta = 1 \).

The next lemma states that optimal test is obtained by likelihood ratio.

**Theorem 5.8 (Neyman-Pearson lemma)** Let \( T > 0 \),

\[
A_n(T) := \left\{ x^n : \frac{P_1^n(x^n)}{P_0^n(x^n)} > T \right\}
\]

and

\[
\alpha^* = P_0^n(A_n(T)), \quad \beta^* = P_1^n(A_n^c(T)).
\]

Let \( B \subseteq \mathcal{X}^n \) another test with error probabilities \( \alpha = P_0^n(B), \beta = P_1^n(B^c) \). Then the following inequality holds:

\[
(\beta - \beta^*) + T(\alpha - \alpha^*) \geq 0.
\]

**Proof.** For every \( x^n \),

\[
(I_{A_n}(x^n) - I_{B^c}(x^n))(TP_0(x^n) - P_1(x^n)) \geq 0.
\]

Sum over \( x^n \) to obtain

\[
T(P_0^n(A_n) - P_0^n(B^c)) - P_1^n(A_n^c) + P_0(B^c) = T((1 - \alpha^*) - (1 - \alpha)) - \beta^* + \beta \geq 0.
\]
From Neyman-Pearson lemma, it follows: if there exists a test $B$ with type I error not bigger than $\alpha^*$, i.e. $\alpha \leq \alpha^*$ (resp. $\alpha < \alpha^*$), then corresponding type II error cannot be smaller than $\beta^*$, i.e. $\beta \geq \beta^*$ (resp. $\beta > \beta^*$). And vice versa. In terminology of hypotheses, a test having that property is uniformly most powerful. Hence Neyman-Pearson lemma states that likelihood ratio test is uniformly most powerful test for simple hypotheses.

**Likelihood ratio and K-L distance.** It is easy to see that likelihood ratio test is nothing but choosing the best approximation to $P_{x^n}$ in K-L sense. To see that consider log-likelihood ratio:

\[
L(x^n) = \log \frac{P_1^n(x^n)}{P_0^n(x^n)} = \log \frac{P_1^n(x_1, \ldots, x_n)}{P_0^n(x_1, \ldots, x_n)} = \sum_{i=1}^n \log \frac{P_1^n(x_i)}{P_0^n(x_i)}
\]

\[
= \sum_{a \in X'} \log \left( \frac{P_1(a)}{P_0(a)} \right) P_{x^n}(a) n = n \sum_{a \in X'} P_{x^n}(a) \log \left( \frac{P_1(a) P_{x^n}(a)}{P_0(a) P_{x^n}(a)} \right)
\]

\[
= n \sum_{a \in X'} P_{x^n}(a) \left( \log \frac{P_{x^n}(a)}{P_0(a)} - \log \frac{P_{x^n}(a)}{P_1(a)} \right) = n \left( D(P_{x^n}||P_0) - D(P_{x^n}||P_1) \right).
\]

Consequently

\[
\frac{P_1^n(x^n)}{P_0^n(x^n)} > T \iff \log \frac{P_1^n(x^n)}{P_0^n(x^n)} = n \left( D(P_{x^n}||P_0) - D(P_{x^n}||P_1) \right) > \log T
\]

\[
\iff D(P_{x^n}||P_1) < D(P_{x^n}||P_0) - \frac{\log T}{n}.
\]

When $T = 1$, we get

\[
H_0 \iff D(P_{x^n}||P_0) \leq D(P_{x^n}||P_1)
\]

\[
H_1 \iff D(P_{x^n}||P_1) < D(P_{x^n}||P_0).
\]

**5.3.2 Convergence of error probabilities for likelihood ratio test.**

**Upper bound to probabilities of first and second type of errors.** We consider the asymptotic behavior of error probabilities of likelihood ratio test. Let these errors be (we shall drop "*" and add $n$)

\[
\alpha_n(T) := P_0^n(A_n(T)), \quad \beta_n(T) := P_1^n(A_n^c(T)),
\]

where

\[
A_n(T) = \left\{ x^n : \frac{1}{n} L(x^n) = D(P_{x^n}||P_0) - D(P_{x^n}||P_1) > \frac{\log T}{n} \right\}.
\]
We consider $T$ to be fixed and left out of notation. Since $X_1, \ldots, X_n$ are i.i.d random variables (distributed either as $P_1$ or $P_0$), then from weak LLN, it immediately follows that both error probabilities converge to zero (Exercise 7):

$$\alpha_n \to 0, \quad \beta_n \to 0.$$  

(5.24)

Whether $x^n \in A_n(T)$ or not, depends on on the type $P_{x^n}$, only:

$$P_{x^n} \in E_n \iff x^n \in A_n,$$

where

$$E_n = \left\{ P \in \mathcal{P} : \sum_a P(a) \log \frac{P_1(a)}{P_0(a)} = D(P||P_0) - D(P||P_1) > \frac{\log T}{n} \right\}.$$  

Therefore

$$\alpha_n = P^n_0(A_n(T)) = P^n_0(E_n), \quad \beta_n = P^n_1(A'_n(T)) = P^n_1(E'_n)$$

and Sanov’s theorem is applicable. From the first part of it (inequality (5.11)) the upper bounds to the first type and second type of error probabilities immediately follows:

$$\alpha_n = P^n_0(E_n) \leq (n + 1)^{|X|} 2^{-d(E_n||P_0)}, \quad \beta_n = P^n_1(E'_n) \leq (n + 1)^{|X|} 2^{-d(E'_n||P_1)}.$$  

Recall that (5.11) holds for any kind of sets. To apply second part of Sanov’s theorem (5.12), we have to have a closer look at the sets $E_n$ and $E'_n$. Clearly $E_n$ is open and $E'_n$ a closed set. It is easy to see (check!) that both $E_n$ and $E'_n$ are convex with their common boundary being

$$\partial E_n = \partial E'_n = \left\{ P \in \mathcal{P} : \sum_a P(a) \log \frac{P_1(a)}{P_0(a)} = \frac{\log T}{n} \right\}.$$  

By Proposition 5.1, there exists the best approximation of $P_1$ from $E'_n$ and the best approximation of $P_0$ from $E_n$ (since both sets are closed). Let these best approximations be $P^*_1$ and $P^*_0$, respectively. Since both sets are convex, these best approximations are unique but since the sets depend on $n$, so do $P^*_1$ and $P^*_0$. Thus

$$D(P^*_1||P_1) = d(E'_n, P_1) \quad \text{and} \quad D(P^*_0||P_0) = d(E_n||P_0) \leq d(E_n||P_0).$$

Since every element in $\partial E_n$ is approximated by a sequence from $E_n$ i.e. $\bar{E}_n$ is the closure of its interior, then

$$d(\bar{E}_n||P_0) = d(E_n||P_0) = D(P^*_0||P_0).$$  

(5.25)
Measures $P^*_0$ and $P^*_1$. Taking account the form of $E_n$ and $E^c_n$, it is possible to show (using Lagrange method, see Thomas and Cover, 12.7) that

$$P^*_0 = P^*_1 \propto P_0^\lambda P_1^{1-\lambda} =: P_\lambda,$$

where $\lambda \geq 0$ is such that

$$D(P_\lambda||P_0) - D(P_\lambda||P_1) = \frac{\log T}{n}. \quad (5.26)$$

Thus, for every $a \in \mathcal{X},$

$$P^*_0(a) = P^*_1(a) = \frac{P_0(a)^\lambda P_1(a)^{1-\lambda}}{C(\lambda)},$$

where the normalizing constant is

$$C(\lambda) = \sum_{a \in \mathcal{X}} P_0(a)^\lambda P_1(a)^{1-\lambda}. \quad (5.27)$$

The set $E_n$ is independent of $n$. Everything stated above holds also when the constant $T$ depends on $n$, thus $T$ will be replaced by $T_n$. This is the case, when the likelihood test is as follows:

$$A_n = \{x^n : \frac{L(x^n)}{n} > \gamma \} = \{x^n : D(P_{x^n}||P_0) - D(P_{x^n}||P_1) > \gamma \}, \quad \gamma \in (-D(P_0||P_1), D(P_1||P_0)).$$

In this case $T_n = 2^\gamma n$. Now the sets $E_n$ are independent of $n:$$$

$$E_n = E = \{P \in \mathcal{P} : D(P||P_0) > D(P||P_1) + \gamma \}$$

and the second part of Sanov’s theorem is directly applicable. Indeed, since $E$ is open and $E^c$ is the closure of its interior, from Corollary 5.2 (together with (5.25)) and the second part of Sanov’s theorem (5.12), it follows

$$\frac{1}{n} \log \alpha_n \to -d(E, P_0) = -D(P_\lambda||P_0) \quad (5.28)$$

$$\frac{1}{n} \log \beta_n \to -d(E^c, P_1) = -D(P_\lambda||P_1) = \gamma - D(P_\lambda||P_0), \quad (5.29)$$

where $\lambda(\gamma)$ is such that

$$D(P_\lambda||P_0) = D(P_\lambda||P_1) + \gamma.$$

Clearly, when $\gamma$ increases, then $d(E, P_0)$ increases meaning that $\alpha_n$ converges faster, whilst $d(E^c, P_1)$ decreases meaning that $\beta_n$ converges slower.

A special case $T = 1$. Then $\gamma = 0$, so

$$E = \{P \in \mathcal{P} : D(P||P_0) > D(P||P_1)\}$$

103
and
\[ d(E^c, P_1) = D(P_\lambda || P_1) = D(P_\lambda || P_0) = d(E, P_0), \]
where \( \lambda(0) =: \lambda^* \) is such that
\[ D(P_\lambda || P_1) = D(P_\lambda || P_0) =: D. \]
Thus, denoting the error probabilities of log-likelihood tests corresponding to \( T = 1 \) as \( \alpha_n^* \) and \( \beta_n^* \), we get
\[
\frac{1}{n} \log \alpha_n^* \to -D, \quad \frac{1}{n} \log \beta_n^* \to -D.
\]
(5.30)

**Chernoff’s information.** The number \( D \) defined above is called Chernoff’s information of the measures \( P_0 \) and \( P_1 \). It can be shown that
\[
D = -\min_{0 \leq \lambda \leq 1} \log C(\lambda),
\]
(5.31)
where \( C(\lambda) \) is as in (5.27). The proof of (5.31) is Exercise 8.

Thus, Chernoff’s information \( D \) is the rate of convergence of error-probabilities of log-likelihood test in case \( T = 1 \). In what follows, we shall see that \( D \) is also the best possible rate of convergence of (weighted) average of \( \alpha_n \) and \( \beta_n \).

### 5.3.3 The (weighted) average of \( \alpha_n \) and \( \beta_n \)

We saw that by log-likelihood tests, one can speed up the convergence of \( \alpha_n \) (resp. \( \beta_n \)) by increasing (resp. decreasing) \( \gamma \). The price would be slower convergence of \( \beta_n \) (resp. \( \alpha_n \)). Suppose now that both errors are equal and we want to design the test so that the average of both error probabilities \( \frac{1}{2}(\alpha_n + \beta_n) \) converges as fast as possible. And, given such a test, what would be the corresponding (best) rate of convergence? To be more general, we consider the weighted average \( \pi \alpha_n + (1 - \pi) \beta_n \), and we ask the same questions: what would be the best (in terms of convergence rate) test and what would be the corresponding rate? Clearly the convergence rate is fastest, if for every \( n \), the weighted average \( \pi \alpha_n + (1 - \pi) \beta_n \) is minimal over all possible tests. The following Proposition states that a test with minimal weighted average is the log-likelihood test with \( T = 1/\pi \).

**Proposition 5.2** Let \( \pi \in (0, 1) \), and let \( A_n \) be the likelihood ratio test with \( \frac{\pi}{1-\pi} \), i.e., \( A_n = A_n(\frac{\pi}{1-\pi}) \). Then for any \( B \subset X^n \), it holds
\[
\pi P_0^n(B) + (1 - \pi) P_1^n(B^c) \geq \pi P_0^n(A_n) + (1 - \pi) P_1^n(A_n^c).
\]

The proof of the proposition is Exercise 9.

Hence, we now consider the likelihood ratio test \( A_n \) as defined in Proposition 5.2 and we shall now study the speed (rate) of convergence of \( \pi \alpha_n + (1 - \pi) \beta_n \), where
\[
\alpha_n = P_0^n(A_n), \quad \beta_n = P_1^n(A_n^c).
\]
It turns out that independently of \( \pi \), the rate of convergence is Chernoff information \( D \).
Theorem 5.9 (Chernoff’s bound) Let \( \pi \in (0, 1) \) and \( \alpha_n \) and \( \beta_n \) be the error probabilities of log-likelihood test with \( T = \frac{\pi}{1 - \pi} \). Then, for every \( n \) the following inequality holds:

\[
\pi \alpha_n + (1 - \pi) \beta_n \leq 2^{-Dn}. \tag{5.32}
\]

Moreover the constant \( D \) cannot be improved, since the following convergence holds:

\[
\frac{1}{n} \log(\pi \alpha_n + (1 - \pi) \beta_n) \to -D. \tag{5.33}
\]

Proof. Use (5.31) to see that (5.32) is proven when we show that

\[
\pi \alpha_n + (1 - \pi) \beta_n \leq \min_{0 \leq \lambda \leq 1} \left( \sum_{a \in X} P^0_\lambda(a) P^{1-\lambda}_1(a) \right)^n. \tag{5.34}
\]

Since \( x^n \in A_n \left( \frac{\pi}{1 - \pi} \right) \) if and only if \( \pi P^n_0(x^n) < (1 - \pi) P^n_1(x^n) \), it holds

\[
\pi \alpha_n + (1 - \pi) \beta_n = \pi P^n_0(A_n) + (1 - \pi) P^n_1(A_n^c)
\]

\[
= \sum_{x^n} \min\{\pi P^n_0(x^n), (1 - \pi) P^n_1(x^n)\}. \nonumber
\]

For every \( a, b \geq 0 \) and \( \lambda \in (0, 1) \) it holds \( \min\{a, b\} \leq a^\lambda b^{1-\lambda} \), implying that for every \( \lambda \in (0, 1) \) it holds

\[
\sum_{x^n} \min\{\pi P^n_0(x^n), (1 - \pi) P^n_1(x^n)\} \leq \sum_{x^n} \left( \pi P^n_0(x^n) \right)^\lambda \left( (1 - \pi) P^n_1(x^n) \right)^{1-\lambda}
\]

\[
\leq \sum_{x^n} \left( P^n_0(x^n) \right)^\lambda \left( P^n_1(x^n) \right)^{1-\lambda}
\]

\[
= \prod_{i=1}^n P^\lambda_0(x_i) \prod_{i=1}^n P^{1-\lambda}_1(x_i)
\]

\[
= \prod_{i=1}^n \sum_{a \in X} P^\lambda_0(a) P^{1-\lambda}_1(a)
\]

So we have proven (5.34). Let us now prove the convergence (5.33). Let \( \alpha_n^* \) and \( \beta_n^* \) be the error probabilities of likelihood ratio test where \( T = 1 \). From Neyman-Pearson lemma, it follows that \( \alpha_n + \beta_n \geq \alpha_n^* + \beta_n^* \) so that

\[
\pi \alpha_n + (1 - \pi) \beta_n \geq \min\{\pi, 1 - \pi\}(\alpha_n + \beta_n) \geq \min\{\pi, 1 - \pi\}(\alpha_n^* + \beta_n^*) \geq \min\{\pi, 1 - \pi\} \alpha_n^*.
\]

Thus

\[
\frac{1}{n} \log(\pi \alpha_n + (1 - \pi) \beta_n) \geq \frac{1}{n} \log(\min\{\pi, 1 - \pi\}) + \frac{1}{n} \log \alpha_n^*.
\]
From the convergence (5.30), we obtain
\[
\frac{1}{n} \log \alpha_n^* \rightarrow -D,
\]
thus
\[
\lim \inf \frac{1}{n} \log (\pi \alpha + (1 - \pi) \beta) \geq -D.
\]
The convergence (5.33) follows now from (5.32). ■

Bayes approach. Let \(X_1, \ldots, X_n\) be i.i.d. random variables with distribution \(Q\), where
\[
P(Q = P_0) = \pi, \quad P(Q = P_1) = 1 - \pi.
\]
The aim is to estimate \(Q\). Let the test be \(g_n = I_{A_n}\) and the error
\[
P_e = P(g_n(X_1, \ldots, X_n) \neq Q) = P(g_n(X_1, \ldots, X_n) = 1|Q = 0)\pi + P(g_n(X_1, \ldots, X_n) = 0|Q = 1)(1 - \pi) = P_{0n}(A_n)\pi + P_{1n}(A_n^c)(1 - \pi).
\]

From Theorem 5.9, it follows that independently of \(\pi\), the convergence of \(P_e\) cannot be faster than (5.33) and this speed can be obtained by likelihood ratio tests.

To recapitulate: The rate of convergence of averaged error probabilities of the best possible test is Chernoff information \(D\) that also equals to the rate of convergence (of both) error probabilities for likelihood ratio test with \(T = 1\). Therefore Chernoff information measures how well measures \(P_0\) and \(P_1\) are separated or, equivalently, how well they can be tested against each other.

5.3.4 Stein’s lemma

The averaged error probability corresponds to the case when both errors are equally important. Suppose this is not the case, and we are primary interested in the fast convergence of \(\beta_n\) (the probability of the second type of error). We also know that the convergence of \(\beta_n\) can be speeded up, but this entails slower convergence of \(\alpha_n\). As an extreme case, it is possible to construct a test (how?) so that \(\beta_n = 0\) i.e. the rate of convergence is maximal (equals to \(\infty\)), but in this case \(\alpha_n = 1\). In practice, it is desirable to look for a test so that \(\beta_n\) would converge possible fast, but \(\alpha_n\) is controlled: \(\alpha_n \leq \epsilon\), where \(\epsilon > 0\) is a fixed critical value (often 0.05 or 0.01). What is in then the best possible rate of convergence of \(\beta_n\)? Does it depend on \(\epsilon\)? In other words, we are interested in the rate of convergence of \(\beta_n^\epsilon\), where
\[
\beta_n^\epsilon = \min_{A_n \subset A_n: \alpha_n \leq \epsilon} \beta_n.
\] (5.35)

Note that for \(n\) big enough, \(\beta_n^\epsilon \leq 2^{-Dn}\) (why?), hence the rate is at least \(D\), but can one do faster? Following result (Stein’s lemma) shows that in this case, for any \(\epsilon > 0\), the best possible convergence rate is \(D(P_0||P_1)\). Again, this speed is achieved using likelihood ratio tests.
Lemma 5.3 (Stein’s lemma) Let $P_0$ and $P_1$ two probability measures on $X$. Then the following statement hold:

1. Let for every $n$, $B_n \subset X^n$ be such that $P^n_0(B_n) \leq \epsilon$. Then
   \[
   \lim \inf_n \frac{1}{n} \log \left( P^n_1(B^n_n) \right) \geq -D(P_0||P_1). \tag{5.36}
   \]

2. There exists a sequence $A_n \subset X^n$ so that
   \[
   P^n_0(A_n) \to 0, \quad \lim \sup_n \frac{1}{n} \log \left( P^n_1(A^n_n) \right) \leq -D(P_0||P_1). \tag{5.37}
   \]

Proof. Let for every $n$, the sets $B_n \subset X^n$ be such that
\[
\begin{align*}
\alpha_n & := P^n_0(B_n) \leq \epsilon, \\
\beta_n & := P^n_1(B^n_n).
\end{align*}
\]

We shall show that for every $\delta > 0$,
\[
\lim \inf_n \frac{1}{n} \log \left( P^n_1(B^n_n) \right) \geq -D(P_0||P_1) - \delta. \tag{5.38}
\]

Then (5.36) follows.

Fix $\delta > 0$ and construct $C_n(\delta) \subset X^n$ so that:
\[
C_n(\delta) := \{ x^n : -D(P_0||P_1) - \delta \leq \frac{L(x^n)}{n} \leq -D(P_0||P_1) + \delta \}.
\]

Form weak LLN, it follows that
\[
P^n_0(C_n) \to 1. \tag{5.39}
\]

The proof of (5.39) is Exercise 7. Since
\[
P^n_0(x^n)2^{L(x^n)} = P^n_1(x^n),
\]
then
\[
P^n_1(C_n) = \sum_{x^n \in C_n} P^n_1(x^n) = \sum_{x^n \in C_n} P^n_0(x^n)2^{L(x^n)} \leq \sum_{x^n \in C_n} P^n_0(x^n)2^{-n(D(P_0||P_1) - \delta)}
\]
\[
= 2^{-n(D(P_0||P_1) - \delta)} \sum_{x^n \in C_n} P_0(x^n) = 2^{-n(D(P_0||P_1) - \delta)} P^n_0(C_n).
\]

Thus
\[
\frac{1}{n} \log \left( P^n_1(C_n) \right) \leq -D(P_0||P_1) + \delta + \frac{1}{n} \log \left( P^n_0(C_n) \right),
\]
implying that (recall (5.39))
\[
\lim \sup_n \frac{1}{n} \log \left( P^n_1(C_n) \right) \leq -D(P_0||P_1) + \delta. \tag{5.40}
\]
Sets $C_n(\delta)$ are used in proving (5.38):

$$\beta_n = P^n(B^n_n \cap C_n) = \sum_{x^n \in B^n_n \cap C_n} P^n(x^n) \geq \sum_{x^n \in B^n_n \cap C_n} P^n(x^n)2^{-n(D(P_0||P_1)+\delta)} \geq 2^{-n(D(P_0||P_1)+\delta)} P^n(B^n_n \cap C_n) \geq 2^{-n(D(P_0||P_1)+\delta)}(1 - P^n_0(B_n) - P^n_0(C_n)).$$

Thus

$$\frac{1}{n} \log \beta_n \geq -D(P_0||P_1) - \delta \geq \frac{1}{n} \log(1 - P^n_0(B_n) - P^n_0(C_n)).$$

Since $1 - P^n_0(C_n) = P^n_0(C_n) \to 1$ and $P^n_0(B_n) \leq \epsilon < 1$, for every $n$ big enough

$$1 - P^n_0(B_n) - P^n_0(C_n) \geq 1 - \epsilon > 0.$$

From this, it follows

$$\frac{1}{n} \log(1 - P^n_0(B_n) - P^n_0(C_n)) \to 0$$

and (5.38) holds.

Let us now show the existence of $A_n$ satisfying (5.37). From (5.39) and (5.40), it follows that for every $\delta > 0$ there exists $C_n(\delta) \subset X^n$ such that

$$P^n_0(C_n(\delta)) \to 1, \quad \limsup_{n} \frac{1}{n} \log \left( P^n_1(C_n(\delta)) \right) \leq -D(P_0||P_1) + \delta.$$

Since $C_n(\delta_1) \subseteq C_n(\delta_2)$, when $\delta_1 < \delta_2$, there exist sets $C^*_n$ so that

$$P^n_0(C^*_n) \to 1, \quad \limsup_{n} \frac{1}{n} \log \left( P^n_1(C^*_n) \right) \leq -D(P_0||P_1). \quad (5.41)$$

(Exercise 10). Now take $A_n := C^*_n$. ■

**Corollary 5.3** Let $\epsilon < 1$ and $\beta^*_n$ as in (5.35). Then

$$\lim_{n} \frac{1}{n} \log \beta^*_n = -D(P_0||P_1) \quad (5.42)$$

**Proof.** Corollary immediately from Stein’s lemma. Indeed, (5.36) implies that for every test satisfying $\alpha_n \leq \epsilon$, it holds that

$$\liminf_{n} \frac{1}{n} \log \beta_n \geq -D(P_0||P_1).$$

This implies

$$\liminf_{n} \frac{1}{n} \log \beta^*_n \geq -D(P_0||P_1).$$

On the other hand, by (5.37) there exists a test $A_n$ such that $\alpha_n \to 0$, but

$$\lim_{n} \frac{1}{n} \log \beta_n = -D(P_0||P_1).$$

Together they imply (5.42). ■
5.3.5 Composed hypotheses

**Fixed** $P_0$. Let $P_0$ be fixed (*a priori* known) as previously, but the law under hypotheses $H_1$ is not exactly known and assumed to belong to a set $\mathcal{P}_1$. Hence $X_1, \ldots, X_n$ are i.i.d. with distribution $Q$ and the hypotheses are:

- $H_0 : Q = P_0$
- $H_1 : Q \in \mathcal{P}_1$.

For example, the set $\mathcal{P}_1$ can be $\mathcal{P} \setminus P_0$ and in this case $H_1 : Q \neq P_0$.

Now, instead of likelihood ratio (which is not applicable), we consider **generalized likelihood ratio**:

$$\frac{P^n_{x^n}(x^n)}{P^n_0(x^n)}.$$ 

Taking logarithms and applying Lemma 5.1, we obtain

$$\log \frac{P^n_{x^n}(x^n)}{P^n_0(x^n)} = nD(P_{x^n}||P_0),$$

so that the **generalized likelihood ratio test** is

$$A_n := \left\{ x^n : \frac{P^n_{x^n}(x^n)}{P^n_0(x^n)} > T_n \right\} = \left\{ x^n : D(P_{x^n}||P_0) > \gamma_n \right\}, \text{ where } \gamma_n = \frac{\log T_n}{n}.$$  \hspace{1cm} (5.43)

Using empirical distribution (type) makes sense: by the strong law of large numbers $P_{x^n} \to Q$ a.s. so that for big $n$, we have $P_{x^n} \approx Q$. When $Q \in \mathcal{P}_1$, i.e there exists $P_1 \in \mathcal{P}_1$ such that $Q = P_1$ (hypotheses $H_1$ holds), then generalized likelihood ratio behaves just like likelihood ratio (for that $P_1$). When $H_0$ holds, then $D(P_{x^n}||P_0) \approx 0$.

**Remark:** In the literature sometimes generalized likelihood ratio is defined as

$$\sup_{P \in \mathcal{P}_1} P^n_{x^n}(x^n)$$ \hspace{1cm} (5.44)

If there exists the (restricted) maximum likelihood estimator:

$$\hat{P}_1 = \arg \sup_{P \in \mathcal{P}_1} P^n_{x^n}(x^n),$$

then (5.44) is $\hat{P}_1(x^n)/P_0(x^n)$. If $P_{x^n} \in \mathcal{P}_1$, then $\hat{P}_1 = P_{x^n}$ and the ratio (5.44) is exactly generalized likelihood ratio defined above. In particular, this is so when (and this is often the case) $\mathcal{P}_1 = \mathcal{P} \setminus P_0$. Note that in this case (5.44) equals to generalized likelihood ratio even if $P_{x^n} = P_0$, because $\inf_{P \in \mathcal{P}_1} D(P||P_0) = 0$. If $\mathcal{P}_1 \neq \mathcal{P} \setminus P_0$, then these two statistics are not necessarily the same. However, one can always replace $H_1 : Q \in \mathcal{P}_1$ by more general $H_1 : Q \neq P_0$.

Many well-known statistics like Student’s $t$-statistic are actually generalized likelihood ratios, many other well-known statistics like Pearson’s chi-square statistic are approximations of generalized likelihood ratios.
Example: Consider the hypotheses
\[ H_0 : Q = P_0 \]
\[ H_1 : Q \neq P_0. \]

The Pearson chi-square statistic is
\[ \sum_{a \in \mathcal{X}} \frac{(N_x(a) - nP_0(a))^2}{nP_0(a)} = n \frac{(P_{x^n}(a) - P_0(a))^2}{P_0(a)} = n\chi^2(P_{x^n}||P_0). \]

For any two probability measure \( P \) and \( Q \) on \( \mathcal{X} \), it holds (Exercise 11)
\[ \chi^2(P||Q) \geq (\ln 2) D(P||Q). \]

On the other hand, using Taylor approximation \( c > 0 \)
\[ x \ln \frac{x}{c} = (x - c) + \frac{(x - c)^2}{2a} - \frac{(x - c)^3}{6a^2} + \cdots, \]
we obtain that
\[ (\ln 2) D(P||Q) = \sum_a P(a) \ln \frac{P(a)}{Q(a)} = \sum_a \left( (P(a) - Q(a)) + \frac{(P(a) - Q(a))^2}{2Q(a)} + \cdots \right) \approx \frac{1}{2} \chi^2(P||Q). \]

The approximation is better when \( P(a) \approx Q(a) \) for any \( a \), i.e. \( P \) is close to \( Q \). We therefore obtain that
\[ n\chi^2(P_{x^n}||P_0) \geq n(\ln 2) D(P_{x^n}||P_0), \]
and if \( H_0 \) holds and \( n \) is sufficiently big then \( P_{x^n} \approx P_0 \) so that
\[ n\chi^2(P_{x^n}||P_0) \approx n(2 \ln 2) D(P_{x^n}||P_0). \] (5.45)

Therefore, under \( H_0 \) both test-statistics in (5.45) have (approximatively) \( \chi^2_{|\mathcal{X}|-1} \) distribution. This is a special case of Wilks’ theorem.

Error probabilities. Let us study the error probabilities of generalized likelihood ratio test (5.43). The probability of first type of error is easy to deal with. Let, for every \( n \gamma_n > 0 \) and let \( \gamma_n \rightarrow 0 \). Define
\[ E_n := \{ P : D(P||P_0) > \gamma_n \}, \]
and from the direct part of Sanov’s theorem it follows
\[ P_0^n(A_n) = P_0^n(E_n) \leq (n + 1)^{|\mathcal{X}|}2^{-n\gamma_n}. \]
Hence \( P_0^n(A_n) \rightarrow 0 \), provided \( \gamma_n \) does not converge to zero too fast. Such a sufficiently slowly convergent sequence is (why?)
\[ \gamma_n = (|\mathcal{X}| + 1)\log(n + 1) \]
\[ n. \] (5.46)
Before, we proceed with the probabilities of errors of second type, let us prove an elementary property of $E_n^c$. Recall that

$$E_n^c = \{ P : D(P||P_0) \leq \gamma_n \}.$$  

**Claim 5.10** Let, for every $n$, $\gamma_n \geq 0$ and $\gamma_n \to 0$. Then, for every $P$

$$\lim_{n} d(E_n^c, P) \to D(P_0||P_1). \quad (5.47)$$  

**Proof.** For every $n$ it holds that $P_0 \in E_n^c$ implying that $d(E_n^c, P_1) \leq D(P_0||P_1)$. Thus $\limsup_{n} d(E_n^c, P_1) \leq D(P_0||P_1)$ and the statement is proven, if we show that

$$\liminf_{n} d(E_n^c, P_1) \geq D(P_0||P_1). \quad (5.48)$$  

First of all note that that when $\liminf_{n} d(E_n^c, P_1) = \infty$, then also $D(P_0||P_1) = \infty$ and (5.47) holds. Hence we consider the case $\liminf_{n} d(E_n^c, P_1) < \infty$. If (5.48) does not hold, then $\liminf_{n} d(E_n^c, P_1) < D(P_0||P_1)$ and there exists $\epsilon > 0$ and a subsequence $\{d(E_n^c, P_1)\}$ such that $d(E_n^c, P_1) < D(P_0||P_1) - \epsilon$ for every $k = 1, 2, \ldots$. This means the existence $P_k \in E_n^c$ such that

$$D(P_k||P_1) < D(P_0||P_1) - \epsilon \quad \text{for every } k = 1, 2, \ldots \quad (5.49)$$  

On the other hand, the convergence $\gamma_{n_k} \to 0$ entails $D(P_k||P_0) \to 0$. Since $\mathcal{X}$ is finite, the sequence $\{P_k\}$ contains a subsequence converging to a probability measure $Q$, i.e. $P_{n_k} \to Q$. By $D(P_k||P_0) < \infty$, for every $k$ it holds that $\mathcal{X}_{P_k} \subset \mathcal{X}_{P_0}$ and since $D(\cdot||P_0)$ is continuous (on $\mathcal{X}_{P_0}$), we have $D(P_k||P_0) \to D(Q||P_0) = 0$. Since $D(Q||P_0) = 0$ iff $Q = P_0$, it follows that $P_{n_k} \to P_0$. Since $D(P_k||P_1) < \infty$, for every $k$, it holds $\mathcal{X}_{P_k} \subset \mathcal{X}_{P_1}$ and by the continuity of $D(\cdot||P_1)$ (on $\mathcal{X}_{P_1}$), it follows that $D(P_{n_k}||P_1) \to D(P_0||P_1)$. The obtained convergence contradicts the assumption (5.49). Therefore (5.48) holds. $\blacksquare$

In our setup, the second type of error probability depends on $P_1 \in \mathcal{P}_1$. However, as the following proposition shows, for fixed $P_1 \neq P_0$, the speed of convergence of the second type of error is exponential with rate $D(P_0||P_1)$. From Stein’s lemma we know that this is the fastest possible (optimal) rate. This means that if $Q = P_1$ (implying that $H_1$ holds), then the probability that generalized likelihood test makes mistake decreases with optimal rate. It is important to note that the test does not know the true distribution $P_1$, yet the error probability converges with the same optimal rate as in the case of simple hypotheses where $P_1$ is known. In the terminology of statistics, the generalized likelihood test adapts the best possible rate.

**Proposition 5.3** Let $A_n$ be as in (5.43) (generalized likelihood ratio test), where $\gamma_n$ is as in (5.46). Then the probability of the first type of error converges to zero: $P_n^n(A_n) \to 0$ and for every $P_1 \neq P_0$ the probability of the second type of error converges with maximal possible exponential rate:

$$\lim_{n} \frac{1}{n} \log P_n^n(A_n^c) = -D(P_0||P_1).$$  

111
Proof. We have already shown the convergence $P^n_0(A_n) \to 0$. To prove the second claim, note that the first claim of Sanov’s theorem implies

$$P^n_1(A^c_n) = P^n_1(E^c_n) \leq (n + 1)^{X} 2^{-nd(E^c_n; P_1)},$$

so that

$$\limsup_n P^n_1(A^c_n) \leq \limsup_n -d(E^c_n; P_1) = D(P_0||P_1),$$

where the last equality follows from (5.47). If $P^n_0(A_n) \to 0$, then by Stein’s lemma (inequality (5.36)), $\liminf_n P^n_1(A^c_n) \geq -D(P_0||P_1)$, implying the convergence

$$\lim_n \frac{1}{n} \log P^n_1(A^c_n) = -D(P_0||P_1).$$

The case $H_0 : Q \in \mathcal{P}_0$. Let us now relax the assumption about knowing $P_0$. Instead, we assume the existence of a set $\mathcal{P}_0$ such that $P_0 \in \mathcal{P}_0$. Thus the setup is now as follows: let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution $Q$. The hypotheses are:

$$H_0 : Q \in \mathcal{P}_0$$
$$H_1 : Q \notin \mathcal{P}_0.$$  

For example $H_0$ can be such that the for a real-valued function $g$, the expectation $\mu := Eg(X_i)$ is given, for example $0$. Then

$$H_0 : \mu = 0 \iff P \in \mathcal{P}_0,$$

where $\mathcal{P}_0 = \{P : \sum_{x \in X} g(x) P(x) = 0\}.$

The sets $\mathcal{P}_0$ and $\mathcal{P}_1$ are disjoint. Often $\mathcal{P}_1 = \mathcal{P} \setminus \mathcal{P}_0$ and in this case the alternative is

$$H_1 : Q \notin \mathcal{P}_0.$$  

Now, also the probability of first type of error depends on chosen $P_0 \in \mathcal{P}_0$. In practice, it is often the case, when the probability of the first type of error is supposed to be bounded by $\epsilon > 0$ for every $P_0 \in \mathcal{P}_0$. Hence, we consider tests $B_n$ such that

$$P^n_0(B_n) \leq \epsilon, \quad \forall P_0 \in \mathcal{P}_0 \iff \sup_{P_0 \in \mathcal{P}_0} P^n_0(B_n) \leq \epsilon.$$  

The logarithm of generalized likelihood ratio is now:

$$\log \frac{P^n_{x^n}(x^n)}{\sup_{P_0 \in \mathcal{P}_0} P^n_0(x^n)} = n \inf_{P_0 \in \mathcal{P}_0} D(P_{x^n}||P_0),$$

where the equality follows from Lemma 5.1. Hence the generalized likelihood ratio test:

$$A_n := \left\{x^n : \frac{P^n_{x^n}(x^n)}{\sup_{P_0 \in \mathcal{P}_0} P^n_0(x^n)} > T_n \right\} = \left\{x^n : \inf_{P_0 \in \mathcal{P}_0} D(P_{x^n}||P_0) > \gamma_n \right\}, \quad \gamma_n = \frac{\log T_n}{n}. \quad (5.50)$$

112
Remark: Recall that the type \( P_{x^n} \) is the maximum likelihood estimator of unknown \( Q \) over the set of all possible distributions \( \mathcal{P} \). If there exists a restricted maximum likelihood estimator \( \hat{P}_0 \in \mathcal{P}_0 \) such that
\[
\sup_{P_0 \in \mathcal{P}_0} P(x^n) = \hat{P}_0^n(x^n),
\]
then the logarithm of generalized likelihood ratio can be written as
\[
\log \frac{\max_{\mathcal{P}} P^n(x^n)}{\max_{\mathcal{P}_0} P^n(x^n)} = \log \frac{P^n(x^n)}{\hat{P}_0^n(x^n)} = nD(P_{x^n} \| \hat{P}_0).
\]
Hence, the generalized likelihood ratio has the following clear interpretation: it is the maximum probability (likelihood) over all possible distributions \( \mathcal{P} \) divided by the maximum probability over \( \mathcal{P}_0 \).

The uniform convergence of the probability of the first type of error follows just like in the case of fixed \( P_0 \). Indeed, define
\[
E_n := \{ P : \inf_{\mathcal{P}_0} D(P_{x^n} \| P_0) > \gamma_n \}.
\]
Then, for every \( P_0 \in \mathcal{P}_0 \) clearly \( d(E_n, P_0) > \gamma_n \) so that by the first part of Sanov’s theorem, for every \( P_0 \in \mathcal{P}_0 \) it holds
\[
\forall P_0 \in \mathcal{P}_0 \quad P_0^n(A_n) = P_0^n(E_n) \leq (n + 1)^{|X|}2^{-n\gamma_n} \iff \sup_{P_0 \in \mathcal{P}_0} P_0^n(A_n) \leq (n + 1)^{|X|}2^{-n\gamma_n}.
\]
(5.51)
The right side of previous equalities converges to zero provided \( \gamma_n \) does not converge to zero too fast.

The convergence of second type of error probabilities is also similar. At first note that the generalization of (5.47) holds: for every \( P_1 \),
\[
\lim_n d(E_{n}^c, P_1) = d(\mathcal{P}_0, P_1).
\]
(5.52)
The proof of (5.52) is Exercise 12. Note that when \( \mathcal{P}_0 = \{ P_0 \} \), then (5.52) is exactly (5.47). Using the first part of Sanov’s theorem, for a fixed \( P_1 \in \mathcal{P}_1 \)
\[
P_1^n(A_n^c) = P_1^n(E_n^c) \leq (n + 1)^{|X|}2^{-n d(E_n^c, P_1)}.
\]
(5.53)
With (5.52), thus
\[
\limsup_n \frac{1}{n} \log P_1^n(A_n^c) \leq -d(\mathcal{P}_0, P_1) = d(\mathcal{P}_0, P_1).
\]
Now from (5.51), it follows that if \( n \) is big enough, then for any \( \epsilon > 0 \) \( P_0(A_n) \leq \epsilon \) for every \( P_0 \in \mathcal{P}_0 \). From Stein’s lemma it now follows that also the inequality (5.36) must
hold for every \( P_0 \in \mathcal{P}_0 \). Hence
\[
\liminf_n \frac{1}{n} \log P_n^{	ext{c}}(A_n^\text{c}) \geq -D(P_0||P_1), \quad \forall P_0 \in \mathcal{P}_0 \\
\liminf_n \frac{1}{n} \log P_1^n(A_n^\text{c}) \geq - \inf_{P_0 \in \mathcal{P}_0} D(P_0||P_1) = d(P_0, P_1).
\]

Thus, if \( d(\mathcal{P}_0, P_1) > 0 \), then the generalized likelihood test again gives the optimal rate of convergence for any \( P_1 \in \mathcal{P}_1 \):
\[
\frac{1}{n} \log P_1^n(A_n^\text{c}) \rightarrow -d(P_0, P_1).
\]

The condition \( d(\mathcal{P}_0, P_1) > 0 \) is guaranteed to hold for any \( P_1 \neq P_0 \), if \( \mathcal{P}_0 \) is closed. Thus, we have proved the following theorem:

**Theorem 5.11** Let \( A_n \) be generalized likelihood test (5.50), where \( \gamma_n \) is as in (5.46). Then the probability of the first type of error converges to zero uniformly:
\[
\sup_{P_0 \in \mathcal{P}_0} P_0^n(A_n) \leq \frac{1}{n+1}.
\]

If, in addition, \( \mathcal{P}_0 \) is closed, then for every \( P_1 \notin \mathcal{P}_0 \), the probability of the second type of error converges with maximal possible rate:
\[
\lim_n \frac{1}{n} \log P_1^n(A_n^\text{c}) = -d(\mathcal{P}_0, P_1) > 0.
\]

**Remarks:**

1. In the case \( \mathcal{P}_0 = \{P_0\} \), from Theorem 5.11 follows Proposition 5.3.
2. If \( d(\mathcal{P}_0, P_1) = 0 \), then \( P_1^n(A_n) \) might not converge to 0.

### 5.4 Exercises

1. Prove (5.2).

2. Prove that if \( \mathcal{X} = \{0, 1\} \) and \( Q = B(1, q), q > \frac{1}{2} \), then
\[
B(1, 1) = \arg \min_{P \in \mathcal{P}} (H(P) + D(P||Q)).
\]

3. Prove:
   
   (a) when \( E \) is closed in \( \mathbb{R}^{||X||} \), then so is \( E_Q \);
   
   (b) for every \( E \), the function \( P \mapsto D(P||Q) \) is continuous on \( E_Q \).

4. Prove Corollary 5.1.
5. Work out all details in Examples 1–5.

6. Let $X_1, X_2, \ldots$ be i.i.d. with distribution $Q$. Let

$$
\bar{g} = \max_{a \in A_Q} g(a).
$$

Show that

$$
\lim_{n \to \infty} \frac{1}{n} \log P\left(\frac{1}{n} \sum_{i=1}^{n} g(X_i) \geq \bar{g}\right) \to -d(E, Q),
$$

where

$$
E = \{\sum_a g(a) P(a) \geq \bar{g}\}.
$$

7. a) Let $\alpha_n = P^n_0(A_n(T))$ and $\beta_n = P^n_1(A^c_n(T))$, where $A_n(T)$ is as in (5.22). Prove that $\alpha_n \to 0$ and $\beta_n \to 0$.

b) Prove that for any $\delta > 0$,

$$
P^n_0(C_n(\delta)) \to 1,
$$

where

$$
C_n(\delta) = \{x^n : -D(P_0||P_1) - \delta \leq \frac{L(x^n)}{n} \leq -D(P_0||P_1) + \delta\}.
$$

8. Prove (5.31).

9. Prove Proposition 5.2

10. Prove the existence $C^*_n \subset X^n$ such that (5.41) holds.

11. Let $P$ and $Q$ two probability distributions on $X$. Let

$$
\chi^2(P||Q) := \sum_{x \in X} \frac{(P(x) - Q(x))^2}{Q(x)}.
$$

Prove that

$$
\chi^2(P||Q) \geq (\ln 2) D(P||Q).
$$

12. Prove (5.52).
6 Differential entropy and MaxEnt principle

Let us consider how the main concepts of information theory generalize to continuous random variables. We shall see that there are some significant differences.

Hence, in what follows, let $X \sim P$ be a continuous random variable with density $f$. The support $S$ of $P$ is defined as the smallest closed set having $P(S) = 1$. It always exists and is unique.

6.1 Differential entropy

Def 6.1 The differential entropy of $X$ (distribution $P$ or density $f$) is

$$h(X) := h(P) := h(f) := \int_S -f(x) \log f(x) dx,$$

provided the integral exists.

Remarks:

- The integral (6.1) might not exist and then the differential entropy is not defined.
- The differential entropy can take any values in $[-\infty, \infty]$. In particular, it can be negative.
- As it follows from the previous remark, the equality $h(X) = 0$ does not imply that $X$ is a.s. equal to a constant (non-random).

6.1.1 Examples

Uniform distribution. Let $X \sim U(0, a)$. Then $f(x) = \frac{1}{a} I_{(0,a)}$ and

$$h(X) = \int_0^a \frac{1}{a} \log adx = \log a.$$

Hence, if $a = 1$, then $h(X) = 0$, $\lim_{a \to \infty} h(X) = \infty$ and $\lim_{a \to 0} h(X) = -\infty$.

Normal distribution. Let $X \sim N(0, \sigma^2)$. Then

$$\int_{-\infty}^{\infty} f(x) \ln f(x) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \left( -\ln \sqrt{2\pi\sigma^2} - \frac{x^2}{2\sigma^2} \right) dx$$

$$= -\ln \sqrt{2\pi\sigma^2} - \int_{-\infty}^{\infty} \frac{x^2}{2\sigma^2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= -\frac{EX^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}$$

$$= -\left( \frac{1}{2} + \ln \sqrt{2\pi\sigma^2} \right)$$

$$= -\left( \frac{1}{2} \ln e^{2\pi\sigma^2} \right).$$

116
Hence
\[ h_e(X) := -\int_{-\infty}^{\infty} f(x) \ln f(x) \, dx = \frac{1}{2} \ln(2\pi\sigma^2) \]
and since \( \ln(a) = \ln 2 \log a \), we obtain
\[ -\int_{-\infty}^{\infty} f(x) \log f(x) \, dx = \frac{1}{\ln 2} h_e(X) = \frac{1}{2} \log(2\pi\sigma^2). \]

**Exponential distribution.** Let \( X \sim E(\lambda) \) i.e.
\[ f(x) = \lambda e^{-\lambda x}, \quad x \geq 0. \]
Then
\[ \int_{0}^{\infty} f(x) \ln f(x) \, dx = \ln \lambda - \int_{0}^{\infty} \lambda x f(x) \, dx = \ln \lambda - 1, \]
implying that
\[ h_e(X) = 1 - \ln \lambda \quad \text{and} \quad h(X) = \frac{1}{\ln 2} - \log \lambda. \]

**Remark:** In the examples above the following implication holds: if the variance converges to zero, then \( h \to -\infty \). Hence, in these examples \( h(X) = -\infty \) if for a constant \( c \), \( X = c \) a.s. Is this always so? Unfortunately this is not the case and there exists random variables (distributions) so that \( h(X) = -\infty \) but \( X \) is not a.s. equal to a constant.

### 6.2 Relation of differential entropy to discrete entropy

**Quantization** is approximation of a continuous distribution (random variable) to a discrete distribution (random variable). A common example of quantization is an histogram. Given a good discrete approximation \( Q \) of a continuous distribution \( P \), is the differential entropy of \( P \) close to that of \( Q \)? It is clearly not so and the obvious reason is that the differential entropy can be negative.

Let us consider the following most common quantization: divide \( \mathbb{R} \) into intervals of lengths \( \Delta \). Thus, let
\[ I_i := (i\Delta, (i + 1)\Delta], \quad \mathbb{R} = \bigcup_{i \in \mathbb{Z}} I_i. \]
Let us assume (for simplicity) that \( f \) is continuous on every interval \( I_i \). Then, there exists \( x_i \in I_i \) such that
\[ f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx. \]
Let us now define a discrete distribution
\[ P(\Delta) = \{x_i, p_i\}, \quad \text{where} \quad p_i := \int_{i\Delta}^{(i+1)\Delta} f(x) \, dx = f(x_i)\Delta. \]
The entropy of $P(\Delta)$ is
\[
H(P(\Delta)) = - \sum_i p_i \log p_i \\
= - \sum_i f(x_i) \Delta \log(f(x_i) \Delta) \\
= - \sum_i f(x_i) \Delta \log(f(x_i)) - \log(\Delta) \sum_i f(x_i) \Delta \\
= - \sum_i f(x_i) \Delta \log(f(x_i)) - \log(\Delta),
\]
because
\[
\Delta \sum_i f(x_i) = \sum \int_{i\Delta}^{(i+1)\Delta} f(x) dx = \int f(x) dx = 1.
\]
If $f(x) \log f(x)$ is Riemann integrable, then
\[
\lim_{\Delta \to 0} - \sum_i f(x_i) \Delta \log(f(x_i)) = - \int f(x) \log f(x) dx = h(f),
\]
so that if $\Delta \to 0$, then
\[
H(P(\Delta)) + \log \Delta \to h(f). \tag{6.2}
\]
For example, if $\Delta = n^{-1}$, then for $n$ big enough we obtain from (6.2):
\[
H(P(\frac{1}{n})) - \log n \approx h(f).
\]
**Example:** Let $X \sim U(0, 1)$, $\Delta = 2^{-n}$. Then $H(P(\Delta)) = n$ and $\log \Delta = -n$, implying that
\[
H(P(\Delta)) + \log \Delta = 0 = h(f),
\]
i.e. (6.2) holds as an equality for every $n$.

To recapitulate: As we all know, quantization can be used for approximating moments of a continuous distribution. For example, in the same setup as above, for any integrable function $g$
\[
\sum_i g(x_i)p_i = \sum_i g(x_i)f(x_i) \Delta \to \int g(x) f(x) dx,
\]
provided $\Delta \to 0$. However, quantization cannot be used for (direct) approximation of the entropy.
6.3 AEP and differential entropy

Recall that when \( X_1, X_2, \ldots \) is a weak AEP process (on alphabet \( \mathcal{X} \)), then for every \( \epsilon > 0 \) and \( n > n(\epsilon) \) there exists a set \( W^n_\epsilon \subset \mathcal{X}^n \) such that \( P(W^n_\epsilon) > 1 - \epsilon \),

\[
(1 - \epsilon)2^{n(\text{H} - \epsilon)} \leq |W^n_\epsilon| \leq 2^{n(\text{H} + \epsilon)} \tag{6.3}
\]

and for every \( x^n \in W^n_\epsilon \),

\[
2^{-n(\text{H} + \epsilon)} \leq P(x^n) \leq 2^{-n(\text{H} - \epsilon)}. \tag{6.4}
\]

Here \( \text{H} = \text{H}(P) \). It is easy to see that with differential entropy instead of the entropy, the weak AEP holds also for i.i.d. continuous random variables. The only difference is that instead of the capacity \( |W^n_\epsilon| \) (which now is infinite), one uses the volume of \( W^n_\epsilon \).

**Def 6.2** The **volume** of a measurable set \( A \subset \mathbb{R} \) is

\[
V(A) := \int_A dx_1 \cdots dx_n.
\]

**Theorem 6.3** Let \( X_1, X_2, \ldots \) i.i.d. continuous random variables, where the distribution of \( X_i \) is continuous with density \( f \). Let \( f \log f \) be integrable. Then, for every \( \epsilon > 0 \) and \( n > n(\epsilon) \) there exists a set \( W^n_\epsilon \subset \mathbb{R}^n \) such that the following holds:

1) \[
P^n(W^n_\epsilon) > 1 - \epsilon. \tag{6.4}
\]

2) \[
(1 - \epsilon)2^{n(\text{h} - \epsilon)} \leq V(W^n_\epsilon) \leq 2^{n(\text{h} + \epsilon)}. \tag{6.5}
\]

3) **For every** \( x^n \in W^n_\epsilon \),

\[
2^{-n(\text{h} + \epsilon)} \leq f^n(x^n) \leq 2^{-n(\text{h} - \epsilon)}, \tag{6.6}
\]

where \( h := h(f) \) and \( f^n(x^n) = f^n(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) \).

**Proof.** The proof is the same as in the case of discrete random variables. Define

\[
W^n_\epsilon := \{ x^n \in \mathbb{R}^n : 2^{-n(\text{h} + \epsilon)} \leq f^n(x^n) \leq 2^{-n(\text{h} - \epsilon)} \}.
\]

From WLLN, it follows that

\[
-\frac{\log f(X_1, \ldots, X_n)}{n} \rightarrow -E\left( \log f(X_1) \right) = h(f), \quad \text{a.s.}
\]

and so (6.6) and (6.4) follows. From

\[
1 - \epsilon \leq P^n(W^n_\epsilon) = \int_{W^n_\epsilon} f^n(x_1, \ldots, x_n) dx_1 \cdots dx_n \leq 1,
\]

we obtain (6.5). \( \blacksquare \)
6.4 Joint differential entropy

Joint differential entropy of continuous random vector \((X_1, \ldots, X_n)\) is defined just like the joint entropy.

**Def 6.4** Let \(X^n = (X_1, \ldots, X_n)\) be continuous random vector with joint density \(f\). Joint differential entropy of \(X^n\) is

\[
h(X^n) = h(X_1, \ldots, X_n) := - \int f(x^n) \log f(x^n) dx^n = - \int f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n) dx_1 \cdots dx_n,
\]

provided the integral exists.

**Example:** Let \(\phi(x^n)\) be the density of multivariate normal distribution \(N(\mu, \Sigma)\)

\[
\phi(x^n) = \frac{1}{(\sqrt{2\pi})^n|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x^n - \mu)'\Sigma^{-1}(x^n - \mu)\right].
\]

Joint entropy

\[
- \int_{-\infty}^{\infty} \phi(x^n) \ln \phi(x^n) dx^n = \int_{-\infty}^{\infty} \frac{1}{2}(x^n - \mu)'\Sigma^{-1}(x^n - \mu)\phi(x^n) dx^n + \ln[(2\pi)^{n/2}|\Sigma|^{1/2}]
\]

\[
= \frac{1}{2} E((X^n - \mu)'\Sigma^{-1}(X^n - \mu)) + \frac{1}{2} \ln[(2\pi)^{n}|\Sigma|],
\]

where \(X^n\) is a random vector with density \(\phi\). Since \(\text{tr}(AB) = \text{tr}(BA)\), we obtain

\[
(X^n - \mu)'\Sigma^{-1}(X^n - \mu) = \text{tr}((X^n - \mu)'\Sigma^{-1}(X^n - \mu)) = \text{tr}(\Sigma^{-1}(X^n - \mu)(X^n - \mu)').
\]

Now

\[
E(X^n - \mu)'\Sigma^{-1}(X^n - \mu) = E\text{tr}((X^n - \mu)'\Sigma^{-1}(X^n - \mu)) = \text{tr}\left(E(\Sigma^{-1}(X^n - \mu)(X^n - \mu)')\right)
\]

\[
= \text{tr}\left(\Sigma^{-1}E(X^n - \mu)(X^n - \mu)\right) = \text{tr}(I_n) = n.
\]

Therefore

\[
- \int_{-\infty}^{\infty} \phi(x^n) \ln \phi(x^n) dx^n = \frac{1}{2}[n + \ln((2\pi)^n|\Sigma|)] = \frac{1}{2}[n \ln e^{n} + \ln((2\pi)^n|\Sigma|)] = \frac{1}{2}\ln[(2\pi e)^{n}|\Sigma|].
\]

Thus the differential entropy is \(\frac{1}{2}\ln[(2\pi e)^{n}|\Sigma|]\) nats and

\[
\frac{1}{2}\log[(2\pi e)^{n}|\Sigma|]\] bits.
Properties:

- Let $X^n$ be a random vector with continuous distribution, $\mu \in \mathbb{R}^n$. Then

  $$h(X^n + \mu) = h(X^n).$$

- Let $X^n$ be a random vector with continuous distribution, $A$ an invertible matrix. Then

  $$h(AX^n) = h(X^n) + \log |A|,$$

  where $|A| = |\det A|$.

The proof of these properties is Exercise 6.5

### 6.5 Conditional differential entropy, Kullback-Leibler distance and mutual information

**Conditional differential entropy.** Recall: if $(X, Y)$ is a random vector $f(x, y)$ random vector, then the conditional density of $X$ given $y$ is

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

$f(x)$ and $f(y)$ are marginal densities.

**Def 6.5** Let $(X, Y)$ have joint density $f(x, y)$. **Conditional differential entropy** is

$$h(X|Y) = -\int \int f(x|y) \log f(x|y) dx dy = -\int \int f(x, y) \log f(x|y) dx dy,$$

provided the integral exists.

Chain rule for two random variables is

$$h(X, Y) = -\int \int f(x, y) \log f(x, y) dx dy = -\int \int f(x, y) \log \left( \frac{f(x, y)}{f(y)} \right) dx dy$$

$$= -\int \int f(x, y) \log f(x|y) dx dy - \int \int f(x, y) \log f(y) dx dy$$

$$= h(X|Y) + h(Y).$$

Now it follows that

$$h(X_1, \ldots, X_n) = h(X_1) + h(X_2|X_1) + \cdots + h(X_n|X_1, \ldots, X_{n-1}).$$
Kullback-Leibler distance.

**Def 6.6** Let \( f, g \) be two probability densities. Their **Kullback-Leibler distance** is

\[
D(f \| g) := \int f(x) \ln \left( \frac{f(x)}{g(x)} \right) dx.
\]

Here, as previously, \( 0 \log \frac{0}{0} = 0 \).

**Remark:** The integral in the definition of \( D(f \| g) \) is always defined (proof is the same as in discrete case), but it might be equal to \( \infty \). Thus, there is no need to add "provided the integral exists" to the definition. When \( D(f \| g) < \infty \), then the support of \( g \) is contained in that of \( f \).

**Lemma 6.1 (Gibbs inequality)** *For any two densities*

\[
D(f \| g) \geq 0,
\]

*with equality \( D(f \| g) = 0 \) if and only if \( f = g \) a.s.*

**Proof.** The same as in discrete case (check!) 

**Example:** Let \( \{f_\theta : \theta \in \Theta\} \) be a family of densities. Let \( X_1, \ldots, X_n \) be i.i.d. random variables with density \( f_\theta^* \in \Theta \) and consider the log-likelihood function

\[
l_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(X_i).
\]

By SLLN, the log-likelihood function converges a.s. to a limit

\[
l(\theta) := \int f_\theta^*(x) \ln f_\theta(x) dx.
\]

The function \( l(\theta) \) is called **likelihood contrast**. Now, from Gibbs inequality, it follows that

\[
0 \leq \int f_\theta^*(x) \ln \left( \frac{f_\theta^*(x)}{f_\theta(x)} \right) dx = \int f_\theta^*(x) \ln f_\theta^*(x) dx - \int f_\theta^*(x) \ln f_\theta(x) dx = l(\theta^*) - l(\theta).
\]

Hence the true parameter \( \theta^* \) maximizes the likelihood contrast \( l(\theta) \) over \( \Theta \). This is the basis of the consistency of maximum likelihood estimator.
Mutual information

**Def 6.7** Let \((X, Y)\) be a random vector with joint density \(f(x, y)\) and marginal densities \(f(x)\) and \(f(y)\). **Mutual information** between \(X\) and \(Y\) is

\[
I(X; Y) := D(f(x, y)||f(x)f(y)) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy.
\]

The properties of mutual information are the same as in discrete case.

**Properties:**

- The mutual information \(I(X; Y)\) depends on joint density (distribution) of \((X, Y)\).
- \(0 \leq I(X; Y)\).
- Mutual information is symmetric: \(I(X; Y) = I(Y; X)\).
- \(I(X; Y) = 0\) iff \(f(x, y) = f(x)f(y)\), i.e. \(X\) are \(Y\) independent.
- If \(h(X|Y)\) and \(h(Y|X)\) are finite, then

\[
I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \geq 0.
\]

Hence, from chain rule it follows

\[
h(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i).
\]

Applying to multivariate normal, we obtain so-called **Hadamard’s inequality**:

\[
\frac{1}{2} \log[(2\pi e)^n |\Sigma|] \leq \sum_{i=1}^{n} \frac{1}{2} \log[(2\pi e)\sigma_i^2] \quad \Leftrightarrow \quad |\Sigma| \leq \prod_{i=1}^{n} \sigma_i^2. \quad (6.7)
\]

### 6.6 MaxEnt principle

Suppose we are interested in finding (estimating) unknown distribution \(P\), given we know:

- \(\text{supp}(P) = S\) (support);
- \(\int F_idP = c_i, \ i = 1, \ldots, k,\)

where \(F_i\) are functions, often polynomials and \(F_i\)-moments \(c_i\) are typically estimated by sample.

The **method of moments**: from a given set of distributions \(\mathcal{P}\) – model – pick the one with \(F_i\)-moments \(c_i\). This approach assumes the existence of the model such that the moments determine the distribution uniquely.
MaxEnt principle: From all distributions satisfying the constraints above, pick the one (if exists and unique) with the maximal (differential) entropy – the MaxEnt distribution. When we are looking for a continuous distribution, we obtain the following optimization problem:

The MaxEnt problem for continuous distributions: \( \max h(f) \) over the functions \( f \) satisfying:

1) \( f(x) \geq 0, f(x) = 0 \Leftrightarrow x \notin S \);
2) \( \int_S f(x)dx = 1 \);
3) \( \int_S F_i(x)f(x)dx = c_i, i = 1, \ldots, k \).

The following theorem gives an easy criterion for solving the problem and finding MaxEnt distribution. Recall that for any function \( f \) and set \( S \subset \mathbb{R} \),

\[
f(x)I_S(x) = \begin{cases} f(x), & \text{if } x \in S; \\ 0, & \text{else.} \end{cases}
\]

**Theorem 6.8** If there are constants \( a_0, a_1, \ldots, a_k \) such that the function

\[
f^*(x) = \exp[a_0 + \sum_{i=1}^{k} a_i F_i(x)]I_S(x)
\]

satisfies 2), 3), then \( f^* \) is the only (Lebesgue a.s.) solution of the above-described MaxEnt optimization problem.

**Proof.** Let \( g \) be an arbitrary density satisfying 1), 2), 3). Theorem is proven if we show \( h_\epsilon(g) \leq h_\epsilon(f^*) \) and the inequality holds iff \( g = f^* \) a.s..

\[
h_\epsilon(g) = -\int_S g(x) \ln g(x)dx = -\int_S g(x) \ln \left( f^*(x) \frac{g(x)}{f^*(x)} \right)dx = -D_\epsilon(g||f^*) - \int_S g(x) \ln f^*(x)dx
\]

\[
\leq -\int_S g(x) \ln f^*(x)dx = -\int_S (a_0 + \sum_{i=1}^{k} a_i F_i(x))g(x)dx
\]

\[
= -(a_0 + \sum_{i=1}^{k} a_i c_i) = -\int_S (a_0 + \sum_{i=1}^{k} a_i F_i(x))f^*(x)dx
\]

\[
= -\int_S f^*(x) \ln f^*(x)dx = h_\epsilon(f^*).
\]

The equality \( h_\epsilon(f^*) = h_\epsilon(g) \) holds iff

\[
D_\epsilon(g||f^*) = \int_S g(x) \ln \frac{g(x)}{f^*(x)}dx = 0.
\]

From Gibbs inequality we know that this holds iff \( g = f^* \) a.s..
Remarks:

• Theorem holds also in multivariate case, where the joint (differential) entropy is maximized. The proof is the same.

• When the support $S$ is a discrete set $\mathcal{X}$, the discrete MaxEnt distribution is searched. Replacing integration with summation, it immediately follows that the proof above also holds in discrete case. Thus, in the discrete case the MaxEnt distribution (if exists) is

$$P^*(x) = \frac{\exp[\sum_{i=1}^k a_i F_i(x)]}{\sum_{x \in \mathcal{X}} \exp[\sum_{i=1}^k a_i F_i(x)]},$$

(6.9)

where $a_i$ are chosen so that the constraints are satisfied. Note that for finite alphabet, (6.9) also follows from Lemma 5.2, where the minimizer of $D(P || Q)$ over all distributions $P$ satisfying $\sum_x F(x)P(x) = c$ was proven to be

$$P^*(x) = \frac{Q(x)2^{\lambda F(x)}}{\sum_{x \in \mathcal{X}} Q(x)2^{\lambda F(x)}} = \frac{Q(x)\exp[(\lambda \ln 2)F(x)]}{\sum_{x \in \mathcal{X}} Q(x)\exp[(\lambda \ln 2)F(x)]},$$

(6.10)

where $\lambda$ is chosen such that the constraint were satisfied. The proof of this lemma can be easily generalize to the case where there are more than one constraints (see Exercise 4. To get (6.9) from (6.10) take $Q$ uniform. Then minimizing K-L distance is the same as maximizing entropy.

Examples:

Mean and variance: Let $S = \mathbb{R}$, $F_1(x) = x$, $c_1 = 0$ and $F_2(x) = x^2$, $c_2 = \sigma^2$. Thus, we look for MaxEnt distribution over real line with mean 0 and variance $\sigma^2$. The density (6.8) is

$$\exp[a_0 + a_1 x + a_2 x^2].$$

We recognize the density of normal distribution; MaxEnt distribution: $\mathcal{N}(0, \sigma^2)$.

Mean and second moment: Let $S = \mathbb{R}$, $F_1(x) = x$, $c_1 = \mu$ and $F_2(x) = x^2$, $c_2 = \alpha$. Thus, we look for MaxEnt density with given first and second moment. The density (6.8) is

$$\exp[a_0 + a_1 x + a_2 x^2].$$

Again, we recognize the normal distribution; MaxEnt distribution: $\mathcal{N}(\mu, \alpha - \mu^2)$

Mean: Let $S = \mathbb{R}$, $F_1(x) = x$, $c_1 = \mu$. Hence the only restriction is given expectation. There is no such MaxEnt distribution.

Mean and non-negativity: Let $S = [0, \infty)$, $F_1(x) = x$, $c_1 = \mu$. As in the previous example, the only restriction is mean, but we are looking for the MaxEnt density on $[0, \infty)$. That makes the difference. Indeed, there is a density on $[0, \infty)$ with given mean $\mu$ having the form (6.8):

$$\exp[a_0 + a_1 x].$$

The MaxEnt distribution is exponential: $E(\mu^{-1})$. 

125
**Mean and countable support:** Let $S = \{1, 2, \ldots\}$, $F_1(x) = x$, $c_1 = \mu$. The distribution 
(6.8):
$$\exp[a_0 + a_1 x]$$
MaxEnt distribution: Geometric $G(\frac{1}{\mu}).$

**Second moment and non-negativity:** Let $S = [0, \infty)$ and $F_1(x) = x^2$, $c_1 = 1 > 0$. The distribution (6.8):
$$\exp[a_0 + a_1 x^2]$$
For this specific $c_1$, the MaxEnt distribution is
$$f(x) = \sqrt{\frac{2}{\pi}} \exp[-\frac{x^2}{2}], \ x \geq 0 \quad (6.11)$$

**Bounded support:** Let $S = [a, b]$, no further constraints. The distribution (6.8): $\exp[a_0]$. MaxEnt distribution: uniform $U(a, b)$.

**Finite support:** Let $S = \{1, 2, 3, 4, 5, 6\}$, no further constraints. The distribution (6.8): $\exp[a_0]$. MaxEnt distribution: uniform.

**Mixed moments:** Let $S = \mathbb{R}^n$, $F_{ij} = x_i x_j$, $c_{ij} = \sigma_{ij}$, $i, j = 1, \ldots, n$. Hence the constraints are on mixed moments: $EX_i X_j = \sigma_{ij}$. The distribution (6.8):
$$f(x^n) = \exp[a_0 + \sum_{ij} a_{ij} x_i x_j].$$
MaxEnt distribution: multivariate normal $\mathcal{N}(0, \Sigma)$, where $\Sigma = (\sigma_{ij})$.

### 6.7 Change of variables

Let $X$ be a random variable with support $S$ and density $f$. Let $g : S \to S$ be a bijection such that the inverse $g^{-1}$ is differentiable. Then the density of $Y := g(X)$ is
$$f_Y(y) = f(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|. \quad (6.12)$$

**Differential entropy and MaxEnt distribution is not invariant under change of variables.** Differential entropy is not invariant under change of variable: the differential entropy of $f(x)$ is not necessarily the same as the one of $f_Y(y)$. In other words: if $X$ is a continuous random variable and $g$ is one-to-one then the differential entropy of $g(X)$ is not necessarily the same as the differential entropy of $X$. For discrete random variables this is so. For example if $a \neq 0$, then
$$h(aX) = h(X) + \log |a|.$$
Therefore, for continuous random variables also MaxEnt distribution is not invariant under the change of variables. Indeed: if $E[F(X)] = c$, then clearly $E[G(Y)] = c$, where $G = F \circ g^{-1}$. However, if $f$ is a MaxEnt distribution $S$ under constraints

$$\int F_i(x)f(x)dx = c_i,$$

then after the change of variable $y = g(x)$, the same distribution $f_Y(y)$ is not necessarily the MaxEnt distribution with respect to the constraints

$$\int G_i(y)f_Y(y)dy = c_i,$$

where $G_i := F_i \circ g^{-1}$. In other words: if $X$ has a MaxEnt law under constraints $E[F_i(X)] = c_i$, then $Y := g(X)$ satisfies the constraints $E[G_i(Y)] = c_i$ but $Y$ has not necessarily MaxEnt law under these constraints.

**Example:** Let $S = [0, \infty)$ and consider the constraint $EX^2 = 1$. Then the MaxEnt distribution is (6.11): for $x \geq 0$,

$$f(x) = \sqrt{\frac{2}{\pi}} \exp[-\frac{x^2}{2}] = 2\phi(x),$$

where $\phi(x)$ is the density of standard normal distribution.

Let $Y = X^2$. Then the density of $Y$ is

$$f_Y(y) = f(\sqrt{y}) \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} \exp[-\frac{y}{2}],$$

One can easily check that $EY = 1$. However, $f_Y(y)$ is clearly not the MaxEnt density on $[0, \infty)$ satisfying $EY = 1$. This would be exponential distribution with parameter one:

$$f_E(y) = \exp[-y]$$

Vice versa: Let the density of $Y$ is $f_E$ (exponential with parameter one, and hence MaxEnt). Let now $x = g(y) = \sqrt{y}$. Then the density of $X := \sqrt{Y}$ is

$$f(x) = f_E(x^2)2x = 2x\exp[-x^2]$$

and so $X$ does not have MaxEnt distribution under the constraint $EX^2 = 1$.

**K-L distance is invariant under change of variables.** Indeed, if $f^1$ and $f^2$ are two densities, then after change of variables $y = g(x)$ they are

$$f_Y^i(y) = f^i(g^{-1}(y))|J(y)|, \text{ where } J(y) = \frac{d}{dy}g^{-1}(y), \quad i = 1, 2.$$  

Now applying the change of variable

$$D(f_Y^1||f_Y^2) = \int f^1(g^{-1}(y))\log\left(\frac{f_Y^1(g^{-1}(y))}{f_Y^2(g^{-1}(y))}\right)|J(y)|dy = \int f^1(x)\log\left(\frac{f_Y^1}{f_Y^2}\right)dx = D(f^1||f^2).$$
6.8 Exercises

1. Find \( h(f) \), where \( f(x) = \frac{1}{2} \lambda \exp[-\lambda|x|] \) (Laplace distribution).

2. Let \( X \sim U(-\frac{1}{2}, \frac{1}{2}) \), \( Z \sim U(-\frac{a}{2}, \frac{a}{2}) \), \( a > 0 \), \( X \) and \( Z \) are independent, \( Y = X + Z \). Find \( I(X; Y) \).

3. Let \( \Pi \) be the product manifold on \((\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))\) – the set of all product measures. Let \((X, Y)\) be a continuous random vector with joint density \( f(x, y) \). Prove that

\[
I(X; Y) = \inf_{g_1(x) \times g_2(y) \in \Pi} D(f(x, y) \| g_1(x) \times g_2(y)).
\]

Show that the minimum is attained at \( f(x) \times f(y) \), where \( f(x) \) and \( f(y) \) are marginal densities of \( f(x, y) \).

4. Let \( \mathcal{X} \) be a discrete alphabet and \( \mathcal{P} \) the set of all probability distributions on \( \mathcal{X} \) satisfying

\[
\sum_j F_i(x_j)P(x_j) = c_i, \quad i = 1, \ldots, k.
\]

Let \( Q \) arbitrary. Show that if there exists constants \( a_i, i = 0, \ldots, k \) such that \( P^* \in \mathcal{P} \), where

\[
P^*(x_j) = Q(x_j) \exp[a_0 + \sum_{i=1}^k a_i F_i(x_j)],
\]

then

\[
P^* = \arg \min_{P \in \mathcal{P}} D(P || Q).
\]

5. Let \( f_o \) be arbitrary density with support \( S \). Prove the existence of \( F \) and a constant \( c \) such that \( f_o \) is MaxEnt distribution over all densities with support \( S \) and constraint \( \int F(x)f(x)dx = c \).

6. Let \( X^n \) be a random vector with continuous distribution, \( \mu \in \mathbb{R}^n \) and \( A \) is an invertible matrix. Show that

\[
h(X^n + \mu) = h(X^n), \quad h(AX^n) = h(X^n) + \log |A|,
\]

where \( |A| = |\det A| \).