Lecture Notes
Information Theory I

HS2022

Ravi Francesco Srinivasan*

*Special thanks to Dario Mylonopoulos for helping with reviewing and finalizing the manuscript.

Contents

List of definitions

List of theorems

List of lemmas

Notation

$\mathcal{X}$	Set (alphabet, always finite)
x	Generic element of $\mathcal{X}$
X	Chance variable taking values in $\mathcal{X}$
$P_X\left(\cdot\right)$	Function that maps $\mathcal{X} \rightarrow [0,1]$, such that $P_X\left(x\right) = \Pr\left(X = x\right)$
X^n	A sequence of n chance variables: $X^n = \left(X_1, X_2, ..., X_n\right)$
$H_b\left(p\right)$	Entropy of a p-Bernoulli random variable
$\ell\left(x\right)$	Length of string $\mathcal{C}\left(x\right)$
$\underline{x}$	A sequence of observations of X
$P^{x^n}\left(\mathcal{A}_\epsilon^{(n)}(P)\right)$	Equivalent to : $\Pr\left(\underline{x} \in \mathcal{A}_\epsilon^{(n)}(P)\right)$
$\mathcal{Q}$	Input distribution for a channel: $Q \in P\left(\mathcal{X}\right)$
$\left(QW\right)\left(Y\right)$	Equivalent to: $\sum_{x\in\mathcal{X}} Q\left(x\right) W\left(Y \mid X = x\right)$
$\left(Q \circ W\right)\left(X, Y\right)$	Equivalent to: $Q\left(X\right) W\left(Y \mid X\right)$

1 Introduction

In this course, we will develop the theory to study **Digital Communication Systems** (**DCS**), as shown in Fig. 1. To be exact, we will consider DCS that have a **Discrete Memory Source** (**DMS**), such that the input is a finite alphabet.

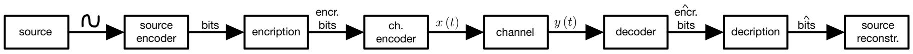


Figure 1: Discrete Memory System schematics

We will now define some of the main concepts that we will provide as a base for developing the theory.

1.1 Entropy

Definition 1.1 (Entropy)**.** We define entropy as:

$$\begin{aligned} H\left(X\right) &= \sum_{x \in \mathcal{X}} P_X\left(x\right) \log\left(\frac{1}{P_X\left(x\right)}\right) \text{ [bits]} \\ \text{where} \quad 0 \log\left(\frac{1}{0}\right) &= 0 \end{aligned}$$

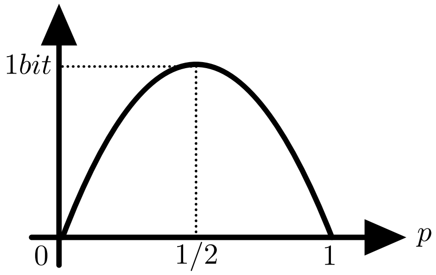
We also note that $H\left(X\right)$ is not a function of X, but it is a function of $P\left(X\right)$: we may also write $H\left(P_X\right)$

Example 1.1

Let's consider the example of tossing a coin: $\mathcal{X} = \{\mathrm{H}, \mathrm{T}\}$. In this case we have that: $P_X\left(H\right) = p = 1 - P_X(T)$ s.t. $0 \leq p \leq 1$. We can compute the entropy as follows:

$$\begin{aligned} H\left(X\right) &= P_X(H) \log\left(\frac{1}{P_X(H)}\right) + P_X(T) \log\left(\frac{1}{P_X(T)}\right) \\ &= p \log\left(\frac{1}{p}\right) + (1-p) \log\left(\frac{1}{1-p}\right) \end{aligned}$$

We have just defined the entropy of a Bernoulli variable H_b. If we observe the value of $H_b\left(\cdot\right)$ we obtain the following graph:



Value of $H_b\left(\cdot\right)$

From the definition, we observe that we can write the definition of entropy as follows:

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

Definition 1.2 (Self information)**.** Given a chance variable X which takes value in $\mathcal{X}$, we define $\wr_x$: $\mathcal{X} \to \mathbb{R}, \wr_x = \log\left(\frac{1}{P_X(a)}\right)$ as **self information**.

Remark. If we set $g = \wr_x$, we observe:

$$\begin{aligned} \mathbb{E}_X\left[g\left(x\right)\right] &= \sum_{x\in\mathcal{X}} P_X(x) g\left(x\right) = \sum_{x\in\mathcal{X}} P_X(x) \wr(x) = \sum_{x\in\mathcal{X}} P_X(x) \log\left(\frac{1}{P_X(x)}\right) \\ &= \mathbb{E}_X\left[\wr(x)\right] = H\left(X\right) \end{aligned}$$

We observe that the entropy is always positive:

Theorem 1.1 (Positivity of entropy)**.** *Given a chance variable X which takes value in $\mathcal{X}$, it always holds that $H\left(X\right) \geq 0$. Moreover, $H\left(X\right) = 0$ if and only if $P_X(\cdot)$ is deterministic.*

Proof

First, we observe that:

$$H\left(X\right) = \sum_{x\in\mathcal{X}} P_X(x) \log\left(\frac{1}{P_X(x)}\right) \geq 0$$

As $P_X(x) \geq 0$ by definition of probability and $\log\left(\frac{1}{P_X(x)}\right) \geq 0$.

We also observe that $P_X(x) \log\left(\frac{1}{P_X(x)}\right) = 0$ only if $P_X(x) = 0$ or $P_X(x) = 1$. As $\sum_{x\in\mathcal{X}} P_X(x) = 1$, we obtain that the equality holds if and only if $P_X(\cdot)$ is deterministic.

Later, we will also prove the following theorem:

Theorem 1.2 (Upper bound of entropy)**.** *Given a chance variable X which takes value in $\mathcal{X}$, it always holds that $H(X) \leq \log|\mathcal{X}|$. Moreover, $H(X) = \log|\mathcal{X}|$ if and only if $P_X(\cdot)$ is uniform.*

1.1.1 Relative entropy

We can now define the relative entropy between two chance variables:

Definition 1.3 (Relative entropy)**.** Given two PMFs P, Q on $\mathcal{X}$, we define the **relative entropy** between them as follows:

$$D(P||Q) = \sum_{x \in \mathcal{X}} P_X(x) \log\left(\frac{P(x)}{Q(x)}\right) \neq D(Q||P)$$

Remark. We need a precise definition that covers the cases where P or Q are 0:

$$D(P||Q) = \sum_{x \in \mathcal{X}, P(x)>0} P_X(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Where $D(P||Q) = +\infty$ if it exists $x \in \mathcal{X}$ such that $P(x) > 0$ and $Q(x) = 0$.

We can also observe that $D(P||Q) = \mathbb{E}_P\left[\log\left(\frac{P(X)}{Q(X)}\right)\right]$.

Example 1.2

Let's consider two PMFs P_X, P_U defined on the alphabet $\mathcal{X}$, where $P_U = \frac{1}{|\mathcal{X}|}$ is the uniform distribution on $\mathcal{X}$. We observe that:

$$\begin{aligned}
D(P_X||P_U) &= \sum_{x \in \mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{P_U(x)}\right) \\
&= \sum_{x \in \mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{\frac{1}{|\mathcal{X}|}}\right) \\
&= \sum_{x \in \mathcal{X}} P_X(x) \log\left(P_X(x)|\mathcal{X}|\right) \\
&= \sum_{x \in \mathcal{X}} P_X(x) \log|\mathcal{X}| + \sum_{x \in \mathcal{X}} P_X(x) \log\left(P_X(x)\right) \\
&= \log|\mathcal{X}| - \sum_{x \in \mathcal{X}} P_X(x) \log\left(\frac{1}{P_X(x)}\right) \\
&= \log|\mathcal{X}| - H(P_X)
\end{aligned}$$

1.1.2 Jensen's inequality

Definition 1.4 (Concave function)**.** A function $f(\cdot)$ is **concave** if $\forall x, y, \lambda \in (0,1)$:

$$f(\lambda x + (1-\lambda) y) \geq \lambda f(x) + (1-\lambda) f(y)$$

If the function is twice differentiable, it is concave if and only if $f'' \leq 0$.

Definition 1.5 (Convex function)**.** A function $f(\cdot)$ is **convex** if $-f(\cdot)$ is concave.

Having the definition of a concave function, we can now introduce Jensen's inequality:

Theorem 1.3 (Jensen's inequality)**.** *Given a random variable X and a concave function $f(\cdot)$, it holds that:*

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

Moreover, if f is strictly concave, then it holds with equality if and only if X is deterministic.

Proof

Given the function f, we can write the second-order Taylor expansion:

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(\xi)$$

For some $\xi \in [x, x_0]$. Concavity implies that $(x - x_0)^2 f''(\xi) \leq 0$. It follows that:

$$f(x) \leq f(x_0) + (x - x_0) f'(x_0)$$

Then, we can write:

$$\mathbb{E}[f(x)] \leq \mathbb{E}[f(x_0)] + \mathbb{E}[(x - x_0) f'(x_0)]$$

If we choose $x_0 = \mathbb{E}[x]$, then:

$$\mathbb{E}[f(x)] \leq \mathbb{E}[f(x_0)] + \cancelto{0}{\mathbb{E}[x - x_0]} f'(x_0) \qquad \Longrightarrow \mathbb{E}[f(x)] \leq \mathbb{E}[f(x_0)]$$

This result allows us to prove the non-negativity of the relative entropy:

Theorem 1.4 (Positivity of relative entropy)**.** *Given two PMFs P, Q on $\mathcal{X}$, it always holds that:*

$$D(P||Q) \geq 0$$

Equality holds if $P = Q$.

Proof

First, we note that $log\left(\cdot\right)$ is a concave function, thus:

$$\begin{aligned}
-D\left(P||Q\right) &= \sum_{x\in\mathcal{X}} P(X)\log\left(\frac{Q\left(X\right)}{P\left(X\right)}\right) \\
&= \mathbb{E}_P\left[\log\left(\frac{Q\left(X\right)}{P\left(X\right)}\right)\right] \\
&\leq \log\mathbb{E}_P\left[\frac{Q\left(X\right)}{P\left(X\right)}\right] && \text{(Jensen's ineq.)} \\
&= \log\left(\sum_{x\in\mathcal{X}}\cancel{P\left(X\right)}\frac{Q\left(X\right)}{\cancel{P\left(X\right)}}\right) \\
&= \log\left(1\right) = 0 \implies D\left(P||Q\right) \geq 0
\end{aligned}$$

In **Example 1.2** we obtained that $D\left(P_X||P_U\right) = \log\left|\mathcal{X}\right| - H\left(P_X\right)$. We obtain that:

- Since $D\left(P||Q\right) \geq 0$, it holds that $H\left(P_X\right) \leq log|\mathcal{X}|$ (Theorem 1.2)
- If $D\left(P_X||P_U\right) = 0$, then we can conclude $X \sim$ Unif: the uniform distribution is the maximizer of H

Another useful theorem is the Log-sum inequality:

Theorem 1.5 (Log-sum inequality)**.** *Given two sequences* $a_i \geq 0, b_i \geq 0$ *for* $i = 1, ..., m$*, it holds that:*

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

Proof

Let's consider $a = \sum a_i$ and $b = \sum b_i$, we first note that $\frac{a_i}{a} \geq 0, \frac{b_i}{b} \geq 0$ and $\sum \frac{a_i}{a} = 1, \sum \frac{b_i}{b} = 1$. For this reason we can build two PMFs P, Q, with probabilities $\frac{a_i}{a}, \frac{b_i}{b}$. From Theorem 1.4 we know that:

$$D\left(P||Q\right) \geq 0$$

Thus, we derive that:

$$\begin{aligned}
&\sum \frac{a_i}{a} \log\left(\frac{a_i/a}{b_i/b}\right) \geq 0 \\
\Longrightarrow &\sum \frac{a_i}{a} \left(\log \frac{a_i}{b_i} + \log \frac{b}{a}\right) \geq 0 \\
\Longrightarrow &\sum \frac{a_i}{a} \log \frac{a_i}{b_i} \geq \log \frac{a}{b} \\
\Longrightarrow &\sum a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b} \\
\Longrightarrow &\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right)
\end{aligned}$$

■

1.1.3 Joint Entropy and Conditional Entropy

Another fundamental element of information theory is joint entropy:

Definition 1.6 (Joint entropy)**.** Given two chance variables X, Y which take value in $\mathcal{X}, \mathcal{Y}$, with jointly distributed according to $(X,Y) \sim P_{X,Y} : P_{X,Y}(a,b) = \Pr(X = a, Y = b)$, we define the **joint entropy** of X and Y, $H(X,Y)$, or $H(P_{X,Y})$), as:

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) \log\left(\frac{1}{P_{X,Y}(x,y)}\right)$$

We also note that $H(X,Y) = H(Y,X)$. This can be extended to $n \geq 2$ chance variables.

Remark. Because the joint entropy is entropy, it takes all the previously stated properties of entropy.

Another important element is conditional entropy:

Definition 1.7 (Conditional entropy)**.** Given two chance variables X, Y which take value in $\mathcal{X}, \mathcal{Y}$, with jointly distributed according to $(X,Y) \sim P_{X,Y} : P_{X,Y}(a,b) = \Pr(X = a, Y = b)$, we define the **conditional entropy** of $X|Y = y$, $H(X|Y = y)$ as:

$$H(X|Y = y) = \sum_{x\in\mathcal{X}} P_{X|Y=y}(x) \log\left(\frac{1}{P_{X|Y=y}(x)}\right)$$

From this, we can derive the general definition of conditional entropy $H(X|Y)$

as the expectation of $H\left(X|y=y\right)$ taken over y:

$$\begin{aligned} H\left(X|Y\right) &= \mathbb{E}_{P_Y}\left[H\left(X|Y=y\right)\right] \\ &= \sum_{y\in\mathcal{Y}} P_Y\left(y\right) H\left(X|Y=y\right) \\ &= \sum_{y\in\mathcal{Y}} P_Y\left(y\right) \sum_{x\in\mathcal{X}} P_{X|Y=y}\left(x\right)\log\left(\frac{1}{P_{X|Y=y}\left(x\right)}\right) \\ &= \sum_{y\in\mathcal{Y}} \cancel{P_Y\left(y\right)} \sum_{x\in\mathcal{X}} \frac{P_{X,Y}\left(x,y\right)}{\cancel{P_Y\left(y\right)}}\log\left(\frac{1}{P_{X|Y=y}\left(x\right)}\right) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}\left(x,y\right)\log\left(\frac{1}{P_{X|Y=y}\left(x\right)}\right) \end{aligned}$$

We can now derive a very important property of joint entropy:

Theorem 1.6 (Chain rule for joint entropy)**.** *Given the joint entropy between two chance variables $H\left(X,Y\right)$, it holds that:*

$$H\left(X,Y\right) = H\left(X\right) + H\left(Y|X\right) = H\left(Y\right) + H\left(X|Y\right)$$

In general, given $n \geq 2$ chance variables, it holds that:

$$\begin{aligned} H\left(X^n\right) &= H\left(X_1\right) + H\left(X_2|X_1\right) + \cdots + H\left(X_n|X_1,...,X_{n-1}\right) \\ &= \sum_{i=1}^{n} H\left(X_i|X^{i-1}\right) \end{aligned}$$

This means that the uncertainty of two experiments is the uncertainty of one, plus the uncertainty of the second one given the result of the first. We are going to prove this theorem for two chance variables, but the proof is analogous for n chance variables:

Proof

$$\begin{aligned} H\left(X,Y\right) &= \mathbb{E}_{P_{X,Y}}\left[\log\frac{1}{P_{X,Y}\left(x,y\right)}\right] \\ &= \mathbb{E}_{P_{X,Y}}\left[\log\frac{1}{P_Y\left(y\right)P_{X|Y=y}\left(X\right)}\right] \\ &= \mathbb{E}_{P_{X,Y}}\left[\log\frac{1}{P_Y\left(y\right)} + \log\frac{1}{P_{X|Y=y}\left(x\right)}\right] = H\left(Y\right) + H\left(X|Y\right) \end{aligned}$$

■

Remark. If $X \perp Y$, then $P_{X,Y} = P_X P_Y$. It follows that

$$H\left(X|Y=y\right) = H\left(X\right) \implies H\left(X|Y\right) = H\left(X\right)$$

1.2 Mutual information

Using entropy, we can define another important element of information theory:

Definition 1.8 (Mutual information)**.** Given two chance variables X, Y which take value in $\mathcal{X}, \mathcal{Y}$, we define the **mutual information** between X and Y, $I\left(X;Y\right)$ as:

$$\begin{aligned} I\left(X;Y\right) &= H\left(X\right) - H\left(X|Y\right) && (1) \\ &= H\left(Y\right) - H\left(Y|X\right) && (2) \\ &= H\left(X\right) + H\left(Y\right) - H\left(X,Y\right) && (3) \\ &= D\left(P_{X,Y}||P_X P_Y\right) && (4) \end{aligned}$$

The mutual information can be interpreted as the uncertainty of X minus the uncertainty of X once Y is revealed (or how much information Y conveys about X), and vice-versa.

Proof

We are now going to prove the equivalency of the definitions, taking (1) as the main definition. We start by proving (3):

$$\begin{aligned} I\left(X;Y\right) &= H\left(X\right) - H\left(X|Y\right) \\ &= H\left(X\right) - \left[H\left(X,Y\right) - H\left(Y\right)\right] && \text{(Chain rule)} \\ &= H\left(X\right) + H\left(Y\right) - H\left(X,Y\right) \end{aligned}$$

As $H\left(X,Y\right) = H\left(Y,X\right)$, we conclude that (2) is valid. To prove (4) we write:

$$\begin{aligned} D\left(P_{X,Y}||P_X P_Y\right) &= \sum_{x,y} P_{X,Y}\left(x,y\right) \log\left(\frac{P_{X,Y}\left(x,y\right)}{P_X\left(x\right) P_Y\left(y\right)}\right) \\ &= \sum_{x,y} P_{X,Y}\left(x,y\right) \log\left(P_{X,Y}\left(x,y\right)\right) \\ &\quad + \sum_{x,y} P_{X,Y}\left(x,y\right) \log\left(\frac{1}{P_X\left(x\right)}\right) \\ &\quad + \sum_{x,y} P_{X,Y}\left(x,y\right) \log\left(\frac{1}{P_Y\left(y\right)}\right) \\ &= -E_{X,Y}\left[\log\frac{1}{P_{X,Y}\left(x,y\right)}\right] + E_{X,Y}\left[\log\frac{1}{P_X\left(x\right)}\right] + E_{X,Y}\left[\log\frac{1}{P_X\left(x\right)}\right] \\ &\overset{(i)}{=} -H\left(X,Y\right) + H\left(X\right) + H\left(Y\right) = I\left(X;Y\right) \end{aligned}$$

Where (i) follows from $E_X\left[f\left(x\right)\right] = E_{X,Y}\left[f\left(x\right)\right]$:

$$E_X\left[f\left(x\right)\right] = \sum_{x} P_X(x) f\left(x\right) = \sum_{x,y} P_{X,Y}\left(x,y\right) f\left(x\right) = \mathbb{E}_{X,Y}\left[f\left(x\right)\right]$$

Remembering that $P_X(x) = \sum_y P_{X,Y}(x,y)$.

These different definitions reveal some interesting properties of mutual information:

Theorem 1.7 (Positivity on mutual information)**.** *From* (4) *we obtain that* $I(X;Y) \geq 0$*, with equality if and only if* $X \perp Y$

Theorem 1.8 (Upper bound on mutual information)**.** *From* (1) *and* (2) *we obtain that* $I(X;Y) \leq \log|\mathcal{X}|$

Theorem 1.9 (Symmetry mutual information)**.** *Because* $H(X,Y) = H(Y,X)$*, from* (3) *we obtain that* $I(X;Y) = I(Y;X)$

Theorem 1.10 (Conditioning reduces entropy)**.** *Because* $I(X;Y) \geq 0$ *combined with* (1)*, we obtain that* $H(X|Y) \leq H(X)$*. Equality holds if and only if* $X \perp Y$*.*

Also, we state and prove the following theorem:

Theorem 1.11 (Concavity of entropy)**.** *Entropy is concave, i.e. given two PMFs* P, Q *and* $\lambda \in [0,1]$*, it holds that:*

$$H(\lambda P + (1-\lambda)Q) \geq \lambda H(P) + (1-\lambda)H(Q)$$

Where $(\lambda P + (1-\lambda)Q)(x) = \lambda P(x) + (1-\lambda)Q(x)$*. Equality holds if and only if* $P \equiv Q$*.*

Proof

First, we prove that $(\lambda P + (1-\lambda)Q)(x)$ is a PMF. We easily prove that it is always positive:

$$(\lambda P + (1-\lambda)Q)(x) = \lambda P(x) + (1-\lambda)Q(x) \geq 0$$

Then:

$$\sum_x (\lambda P + (1-\lambda)Q)(x) = \sum_x \lambda P(x) + \sum_x (1-\lambda)Q(x) = \lambda + (1-\lambda) = 1$$

To prove the theorem we define a new chance variable E taking values in $\{0,1\}$, such that $\Pr(E=0) = \lambda$ and $\Pr(E=1) = 1-\lambda$. We can now write:

$$P_{X,E}(x) = \lambda P(x) + (1-\lambda)Q(x) = \Pr(E=0)P(x) + \Pr(E=1)Q(x)$$

Because of Theorem 1.10, we can write:

$$\begin{aligned} H(X|E) &= \Pr(E=0)H(X|E=0) + \Pr(E=1)H(X|E=1) \\ &= \lambda H(P) + (1-\lambda)H(Q) \end{aligned}$$

Because of Theorem 1.10:

$$\lambda H\left(P\right)+\left(1-\lambda\right)H\left(Q\right)=H\left(X|E\right)\leq H\left(X\right)=H\left(\lambda P+\left(1-\lambda\right)Q\right)$$

We note that equality holds if and only if $X\perp E$. We observe that this is true if $P_{X,E}\left(x\right)=P_{X}P_{E}$. We can write:

$$P_{X,E}\left(x\right)=\lambda P\left(x\right)+\left(1-\lambda\right)Q\left(x\right)=P_{X}$$

Which hold if and only if $P\equiv Q$.

■

We can now define conditional mutual information:

Definition 1.9 (Conditional mutual information)**.** Given three chance variables X,Y,Z, we define the conditional mutual information $I\left(X;Y|Z\right)$ as:

$$I\left(X;Y|Z\right)=\sum_{z}P_{Z}\left(z\right)I\left(X;Y|Z=z\right)=\sum_{z}P_{Z}\left(z\right)I\left(P_{X;Y|Z=z}\right)$$

Remark. Because of the positivity of the mutual information, it holds that $I\left(X;Y|Z\right)\geq 0$, with equality if and only if $I\left(X;Y|Z=z\right)=0\ \forall z$. In this case, we say that $X\perp Y|Z$ (X and Y are conditionally independent on Z), and $X -\!\!\circ\!\!- Z -\!\!\circ\!\!- Y$ form a **Markov Chain**.

We also note that:

$$\begin{aligned}I\left(X;Y|Z\right)&=\sum_{z}P_{Z}\left(z\right)I\left(X;Y|Z=z\right)\\&=\sum_{z}P_{Z}\left(z\right)\left(H\left(X|Z=z\right)-H\left(X|Y,Z=z\right)\right)\\&=H\left(X|Z\right)-H\left(X|Y,Z\right)\end{aligned}$$

Similarly to conditional entropy, a chain rule exists for mutual information:

Theorem 1.12 (Chain rule for mutual information)**.** *Given the mutual information entropy between $n+1$ chance variables X^{n},Y, it holds that:*

$$I\left(X^{n};Y\right)=\sum_{i=1}^{n}I\left(X_{i};Y|X^{i-1}\right)$$

1.3 Fano's inequality

Fano's inequality is a fundamental inequality that ties the quantity from the probability of errors and HP tests to information theory. Informally, it states that if you can guess X from Y ”reliably”, then $H\left(X|Y\right)$ must be small. This means that the best guess is the one that guesses the most probable outcome. Formally:

Theorem 1.13 (Fano's inequality)**.** *Given two chance variables X, Y that take values respectively in $\mathcal{X}, \mathcal{Y}$, distributed according to the joint probability $P_{X,Y}$ over $\mathcal{X} \times \mathcal{Y}$, a guessing function $g : \mathcal{Y} \to \mathcal{X}$ and a probability of error with value $p_e = Pr(X \neq g(Y)) = \sum P_X(x) \, Pr(err|X = x)$, then it holds that:*

$$H(X|Y) \leq H_b(p_e) + p_e \log(|\mathcal{X}| - 1)$$

Remark. A looser version of Fano's inequality states that

$$H(X|Y) \leq 1 + p_e \log(|\mathcal{X}|)$$

Proof

We define a chance variable E:

$$E = \begin{cases} 1 & \text{if} \quad g(y) \neq x \\ 0 & \text{otherwise} \end{cases}$$

Using the chain rule, we can now write:

$$H(X, E|Y) \overset{(i)}{=} H(X|Y) + \cancelto{0}{H(E|X, Y)}$$

$$H(X, E|Y) \overset{(ii)}{=} H(E|Y) + \cancelto{0}{H(X|E, Y)}$$

It follows:

$$\begin{aligned}
&H(X|Y) = H(E|Y) + H(X|E, Y) \\
&\overset{(iii)}{\leq} H(E) + H(X|E, Y) \\
&= H_b(p_e) + H(X|E, Y) \\
&\overset{(iv)}{=} H_b(p_e) + \Pr(E = 0) \cancelto{0}{H(X|E = 0, Y)} + \cancelto{p_e}{\Pr(E = 1)} H(X|E = 1, Y) \\
&= H_b(p_e) + p_e H(X|E = 1, Y) \\
&\overset{(v)}{\leq} H_b(p_e) + p_e \log(|\mathcal{X}| - 1)
\end{aligned}$$

Where $(i), (ii), (iv)$ hold because knowing two out of X, Y, E implies knowing the third; (iii) holds because conditioning reduces entropy and (v) hold because $H(X) \leq \log(|\mathcal{X}|)$, and if $E = 1$ is fixed, we can rule out one value of $\mathcal{X}$, obtaining $H(X|E = 1, Y) \leq \log(|\mathcal{X}| - 1)$.

∎

If we have $X_i, ..., X_n \overset{IID}{\sim} \text{Ber}(p)$, the following holds:

$$\begin{aligned}
H\left(X_i|\hat{X}_i\right) &\leq H_b\left(\Pr\left[X_i \neq \hat{X}_i\right]\right) + \Pr\left[X_i \neq \hat{X}_i\right] \log(\underbrace{|\mathcal{X}_i|}_{=2} - 1) \\
&= H_b\left(\Pr\left[X_i \neq \hat{X}_i\right]\right)
\end{aligned}$$

We also note that $H_b(p_e) + p_e \log(m-1)$ is a non-decreasing function in p_e for $p_e \in \left[0, \frac{m-1}{m}\right]$. We prove this using concavity of H:

$$\begin{aligned}
H_b(p_e) + p_e \log(m-1) &= p_e \log \frac{1}{p_e} + (1-p_e) \log \frac{1}{1-p_e} + p_e \log(m-1) \\
&= p_e \log \frac{m-1}{p_e} + (1-p_e) \log \frac{1}{1-p_e} \\
&= \log m + p_e \log \frac{(m-1)/m}{p_e} + (1-p_e) \log \frac{1/m}{1-p_e} \\
&= \log m - D\left((p_e, 1-p_e) \,||\, \left(\frac{m-1}{m}, \frac{1}{m}\right)\right)
\end{aligned}$$

Relative entropy is non-negative and equal to zero if and only if the arguments coincide. It follows that we obtain the maximum of $H_b(p_e) + p_e \log(m-1)$ at $p_e = (m-1)/m$. Since $H_b(p_e) + p_e \log(m-1)$ is a sum of entropy and a linear function (which are both concave), and the sum of concave functions is concave, $H_b(p_e) + p_e \log(m-1)$ is non-decreasing in $p_e \in \left[0, \frac{m-1}{m}\right]$, and is non-increasing for $p_e \in \left[\frac{m-1}{m}, 1\right]$.

2 Source coding

Source coding refers to solving the problem of representing outcomes using bits.

2.1 Codes

To represent outcomes using bits, we use codes:

Definition 2.1 (Code)**.** Given a chance variable X that take value on a finite input alphabet $\mathcal{X}$, and a PMF P_X on $\mathcal{X}$ and a finite output alphabet $\mathcal{Y}$, a **code** is a mapping from outcomes to strings:

$$\mathcal{C} : \mathcal{X} \to \mathcal{Y}$$

Example 2.1

For example, we can consider the finite input alphabet $\mathcal{X} = \{a, b, c, d\}$, and the binary output alphabet $0, 1^+ = \{0, 1, 00, 01, 11, ...\}$ and the following code:

- $a \to 0$
- $b \to 1$
- $c \to 1$
- $d \to 10$

We will work mainly with the binary output alphabet. We immediately note that the code in **Example 2.1** is a bad code, as multiple input elements are mapped to the same output string. For this reason, we need a way to identify good and bad codes:

Definition 2.2 (Non singular code)**.** A code $\mathcal{C} : \mathcal{X} \to \{0, 1\}^+$ is **non singular** if there are not two outcomes of $\mathcal{X}$ mapped to the same string:

$$x \neq x' \implies \mathcal{C}(x) \neq \mathcal{C}(x')$$

From this, we can define:

Definition 2.3 (Expected description length)**.** Given a chance variable X that take value on a finite input alphabet $\mathcal{X}$ and a code $\mathcal{C}$, we define the **expected description length** of code $\mathcal{C}$, $L(\mathcal{C})$ as:

$$L(\mathcal{C}) = \sum_{x \in \mathcal{X}} P_X(x)\, \ell(x)$$

A good question to start with is: how do we minimize $L(\mathcal{C})$? The main idea is to give the most probable outcomes the shortest code. But what happens if we have a sequence of input outcomes?

Example 2.2

Continuing from example **Example 2.1**, we modify our code s.t. it is non singular:

- $a \to 0$
- $b \to 1$
- $c \to 10$
- $d \to 01$

What happens if we try to decode the input sequence *abcd*?

$$abcd \to 0101\ ?\ \begin{matrix} \nearrow abcd \\ \searrow dba \end{matrix}$$

It is not enough to ask for a non-singular code, we need a code s.t. every sequence has a unique description: a **uniquely decodable code**. To define such code, we need the following:

Definition 2.4 (Code extension)**.** Given a code $\mathcal{C}$, the extension of code $\mathcal{C}$, denoted with $\mathcal{C}^+$ is a code that maps strings of characters from input alphabet $\mathcal{X}^n$ to $\{0,1\}^+$ by concatenation:

$$\mathcal{C}^+ (x_1 x_2 ... x_n) \to \mathcal{C}(x_1) \mathcal{C}(x_2) ... \mathcal{C}(x_n) \qquad \text{s.t.} \quad \mathcal{C}^+ : \mathcal{X}^n \to \{0,1\}^+$$

We can now define a uniquely decodable code:

Definition 2.5 (Uniquely decodable code)**.** Given a code $\mathcal{C}$, it is a **uniquely decodable code** if $\mathcal{C}^+$ is non-singular.

We also define:

Definition 2.6 (Prefix-free code)**.** Given a code $\mathcal{C}$, it is a **prefix-free code** (or **instantaneous code**) if no coded word is the prefix of another

From this, it follows:

Theorem 2.1 (Every prefix-free code is uniquely decodable)**.** *Given code $\mathcal{C}$, if $\mathcal{C}$ is prefix-free then it is uniquely decodable.*

Proof

Consider the input sequence $x_1 x_2 ... x_n$ and the encoding $\mathcal{C}(x_1) \mathcal{C}(x_2) ... \mathcal{C}(x_n)$, with $\mathcal{C}$ prefix-free. We can use the following decoding strategy:

1. Ask if the bit is a code:
   - Yes: decode and go to the next bit
   - No: include next bit
2. Continue

Using this strategy, if the code is prefix-free we ensure that the encoding will be correct. If the bit(s) we are considering leads to multiple decoding options, it would mean that the code is not prefix-free.

What is the relationship between the different types of codes? It is described in Fig. 2. This leads us to the **Kraft's inequality**. Let's say we are trying to minimize the expected code length given a chance variable X with PMF P_X:

$$\underset{\mathcal{C} \in \text{u.d. codes}}{\arg\min} \sum_{x \in \mathcal{X}} P_X(x) \ell(x)$$

Which sets of lengths correspond to uniquely decodable codes? We cannot use calculus, as we are in the integer domain.

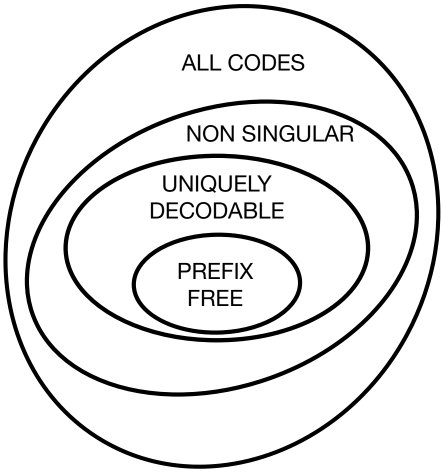


Figure 2: Relationship between types of codes

2.2 Kraft's inequality

The **Kraft's** inequality provides a characterization of the lengths of uniquely decodable codes:

Theorem 2.2 (Kraft's inequality)**.** *The* ***Kraft inequality*** *states that:*

1. *If $\ell_1, \ell_2, ..., \ell_n$ are positive integers such that $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$, then it exists a prefix free code of these lengths*
2. *If $\ell_1, \ell_2, ..., \ell_n$ are lengths of the code-words of a uniquely decodable code $\mathcal{C}$, then it holds that $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$ and ℓ_i are positive integers*

We will prove this theorem later. For now, we will continue with the question we asked in the previous chapter:

$$\underset{\mathcal{C} \in \text{u.d. codes}}{\arg\min} \sum_{x \in \mathcal{X}} P_X(x) \, \ell(x)$$

2.2.1 Bounds on L^*

Given that we are in the domain of uniquely decodable codes, thanks to Kraft's inequality we can rewrite the problem:

$$L^* = \min_{\ell_1, \ldots, \ell_n \in \mathbb{N}} \sum_{x \in \mathcal{X}} P_X(x) \, \ell(x) \qquad \text{s.t.} \quad \sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

First, we can limit ourselves to prefix-free codes, as in general the wider the class the lower the minimum (Fig. 2 reminds us that the family of prefix-free codes is contained in the family of uniquely decodable codes). Furthermore, if we find an amazing uniquely decodable code such that $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$, we know that there exists a prefix-free code with the same lengths (part 2 of Kraft's inequality). Second, we consider a relaxation of L^*. Instead of searching for integer lengths, we search in $\mathbb{R}^+$:

$$L^* \geq \min_{\ell_1,\ldots,\ell_n \in \mathbb{R}^+} \sum_{x \in \mathcal{X}} P_X(x)\,\ell(x) \qquad\qquad \text{s.t.} \quad \sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

We claim that we can write the equality in the constraint, as we can reduce ℓ_i to achieve it. We note that $L^*_{\mathbb{R}^+} \leq L^*$, as we are searching in a bigger domain. We can now treat this as a calculus problem:

Theorem 2.3 (Lower bound on L^*)**.** *Given a chance variable X, it holds that $L^* \geq H(X)$*

To show this, we will prove that $L^*_{\mathbb{R}^+} = H(X)$, from which it follows that $L^* \geq H(X)$:

Proof

First, we claim we can solve this problem using the Lagrange multipliers:

$$\begin{aligned}
\frac{\delta}{\delta \ell_i}\left(\sum_{i=1}^{n} p_i \ell_i - \lambda \sum_{i=1}^{n} 2^{-\ell_i}\right) &= p_i - \lambda \frac{\delta}{\delta \ell_i} 2^{-\ell_i} \\
&= p_i - \lambda \frac{\delta}{\delta \ell_i} e^{-\ell_i \ln 2} \\
&= p_i - \lambda \ln 2 \cdot 2^{-\ell_i} = 0 \\
\implies p_i = \lambda' 2^{-\ell_i} \implies \lambda' \overset{(i)}{=} 1 \implies \ell_i &= \log \frac{1}{p_i}
\end{aligned}$$

(i) holds because $\sum p_i = \sum 2^{-\ell_i} = 1$. We now show that $\ell_i = \log \frac{1}{p_i}$ is optimal: suppose that $\ell_i \in \mathbb{R}^+$ satisfies $\sum 2^{-\ell_i} = 1$. Remembering that $H(X) = \sum p_i \log \frac{1}{p_i}$, we write:

$$\begin{aligned}
L^*_{\mathbb{R}^+} - H(X) &= \sum p_i \ell_i - \sum p_i \log \frac{1}{p_i} \\
&= -\sum p_i \log 2^{-\ell_i} - \sum p_i \log \frac{1}{p_i} \\
&= \sum p_i \log \left(\frac{p_i}{2^{-\ell_i}}\right) \\
&= D\left(P || \{2^{-\ell_i}\}\right) \geq 0
\end{aligned}$$

First, we observe that $\{2^{-\ell_i}\}$ describes a probability distribution, as $2^{-\ell_i} \geq 0, \sum 2^{-\ell_i} = 1$. We also note that this equality holds if and only if $\log\left(\frac{p_i}{2^{-\ell_i}}\right) = 0$, meaning $\ell_i = \log \frac{1}{p_i}$. This implies that $L^*_{\mathbb{R}^+} = H(X)$ with the optimal solution $\ell_i = \log \frac{1}{p_i}$. Thus, we can conclude that:

$$L^* \geq H(X)$$

We can also define an upper bound on L^*:

Theorem 2.4 (Upper bound on L^*)**.** *Given a chance variable X, it holds that $L^* < H(X) + 1$*

Proof

First, we show that if we define $\ell_i = \lceil \log \frac{1}{p_i} \rceil$, the Kraft's inequality is satisfied:

$$\sum 2^{-\ell_i} = \sum 2^{-\left\lceil \log \frac{1}{p_i} \right\rceil} \leq \sum 2^{-\log \frac{1}{p_i}} = \sum p_i = 1$$

Thus, the expected length becomes:

$$\sum p_i \ell_i = \sum p_i \left\lceil \log \frac{1}{p_i} \right\rceil < \sum p_i \left(\log \frac{1}{p_i} + 1 \right) = H(X) + 1$$

With these two theorems we have bounded L^*:

$$H(X) \leq L^* < H(X) + 1$$

What happens if we describe two symbols at the time instead of one? We normalize the length of the number of symbols described, and we obtain:

$$\frac{1}{2} L^*_2 < \frac{1}{2} \left(H(X_1, X_2) + 1 \right) = \frac{1}{2} \left(2H(P) + 1 \right) = H(P) + \frac{1}{2}$$

Remark. We cannot do better than $H(P)$, but it is possible to get arbitrarily close: for $n \to \infty$:

$$\frac{1}{n} L^* \leq H(P) + \frac{1}{n} \to H(P)$$

Example 2.3

If instead of using $\ell_i = \left\lceil \log \frac{1}{p_i} \right\rceil$ we use a different distribution Q, and $\ell_i = \left\lceil \log \frac{1}{q_i} \right\rceil$ we can write:

$$\begin{aligned} L(\mathcal{C}) &= \sum p_i \log \frac{1}{q_i} \\ &= \sum p_i \left(\log \frac{p_i}{q_i} + \log \frac{1}{p_i} \right) \\ &= D(P||Q) + H(P) \end{aligned}$$

In this case, $D(P||Q)$ represents a penalty we pay by using the wrong distribution instead of the true input distribution.

2.2.2 Proving Kraft's inequality

We start by proving the first part, i.e. that if $\ell_1, \ell_2, ..., \ell_n$ are positive integers such that $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$, then it exists a prefix free code of these lengths.

Proof

First, we note that given a prefix-free code, we can write it as a rooted binary tree, as shown in Fig. 3.

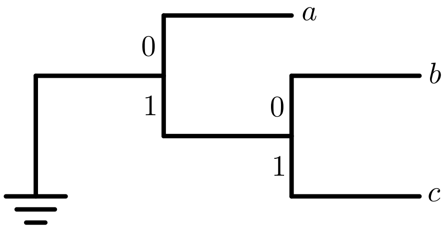


Figure 3: Tree construction for p.f. code $a \to 0, b \to 10, c \to 11$

We order the lengths so that $\ell_i \leq \ell_2 \leq ... \leq \ell_n$. To build the code, we proceed in the following way:

1. Starting from the root we build a tree of depth ℓ_1, and assign a leaf
2. Starting from the root we build a tree of depth ℓ_2, and assign a leaf

⋮

n. Starting from the root we build a tree of depth ℓ_n, and assign a leaf

When we extend the depth to ℓ_{k+1}, the total number of leaves of the tree will be $2^{\ell_{k+1}}$. Some sub-trees of the tree will be unavailable, as every assigned node of the tree (the red ones in Fig. 4) will block its sub-tree (dotted sub-tree in Fig. 4). Every node at level ℓ_i will have $2^{\ell_n - \ell_i}$ descendants at level ℓ_n, and it must hold that the total number of descendants of all nodes is less or equal to 2^{ℓ_n}. Hence, summing for all nodes we obtain:

$$\sum_{i=1}^{n} 2^{\ell_n - \ell_i} \leq 2^{\ell_n} \implies \sum_{i=1}^{n} 2^{-\ell_i} \leq 1$$

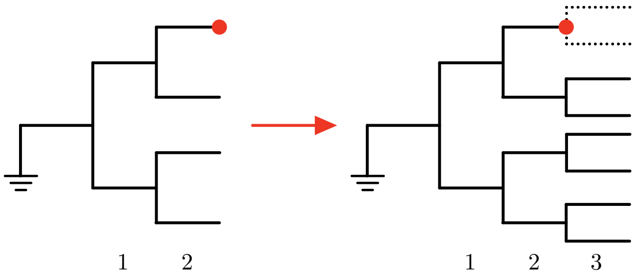


Figure 4: Extension of the tree from depth $\ell_1 = 2$ to depth $\ell_2 = 3$. The number of unavailable leafs is $2^{\ell_2 - \ell_1} = 2^{3-2} = 2$.

We can now prove part 2, i.e. that if $\ell_1, \ell_2, ..., \ell_n$ are lengths of the code-words of a uniquely decodable code $\mathcal{C}$, then it holds that $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$ and ℓ_i are positive integers

Proof

We are given a uniquely decodable code with lengths $\ell_1, ..., \ell_n$. We can write:

$$\sum_{i=1}^{n} 2^{-\ell_i} = \sum_{x\in\mathcal{X}} 2^{-\ell(x)}$$

We can now raise to the kth power:

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} 2^{-\ell(x)}\right)^k &= \left(\sum_{x_1\in\mathcal{X}} 2^{-l(x_1)}\right)\left(\sum_{x_2\in\mathcal{X}} 2^{-l(x_2)}\right)\cdots\left(\sum_{x_k\in\mathcal{X}} 2^{-l(x_k)}\right)\\
&= \sum_{x_1\in\mathcal{X}, x_2\in\mathcal{X}, \ldots, x_k\in\mathcal{X}} 2^{-l(x_1)} 2^{-l(x_2)} \cdots 2^{-l(x_k)}\\
&\overset{(i)}{=} \sum_{x\in\mathcal{X}^k} 2^{-l(\underline{x})}\\
&\overset{(ii)}{=} \sum_{\mathcal{V}=1}^{k\cdot\ell_{\max}} a(\mathcal{V})\, 2^{-\mathcal{V}} \overset{(iii)}{\leq} k\cdot\ell_{\max}
\end{aligned}$$

Where (i) holds if we define $l(\underline{x}) = l(x_1) + \cdots + l(x_k)$ as the length of the concatenation; (ii) if we define $a(\mathcal{V}) = |\{\underline{x} \in \mathcal{X}^k : l(\underline{x}) = \mathcal{V}\}| \leq 2^{\mathcal{V}}$ and (iii) holds because $a(\mathcal{V})\, 2^{-\mathcal{V}} \leq 2^{\mathcal{V}} 2^{-\mathcal{V}} = 1$. Now we take the kth root and obtain:

$$\sum_{x\in\mathcal{X}} 2^{-\ell(x)} \leq (k\cdot\ell_{\max})^{1/k}$$

As this holds for every positive value of k, we make $k \to \infty$ and obtain:

$$\sum_{x\in\mathcal{X}} 2^{-\ell(x)} \leq 1$$

2.2.3 Huffman's procedure

Now that we know what is the minimum expected length, how do we find a good code given an input alphabet? The solution is to use Huffman's procedure.

Example 2.4

Let's consider the input alphabet $\mathcal{X} = \{a, b, c, d, e, f\}$. The chance variable X has the following PMF: $P_X = (0.3, 0.3, 0.15, 0.2, 0.05)$. The idea behind Huffman's procedure is to create a tree: we combine the least likely symbols and form a sub-tree with these symbols as leaves. This node will represent a new symbol with a probability equal to the sum of the probability of these symbols. We continue the process iteratively until we reach the root, as illustrated in Fig. 5.

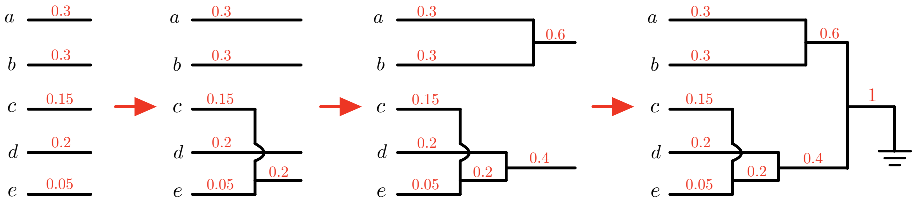


Figure 5: An example of Huffman's procedure

In this case, we would obtain the following code: $a \to 00, b \to 01, c \to 001, d \to 01, e \to 000$.

How can we be sure that this leads to the optimal code in terms of expected length?

Theorem 2.5 (Optimality of Huffman's procedure)**.** *Given an input alphabet $\mathcal{X}$, a chance variable X with PMF P_X on $\mathcal{X}$, when applying Huffman's procedure we can restrict our search without loss of optimality to trees where the two least likely symbols are siblings.*

Proof

We note that:

- In any optimal tree, $pi > p_j \implies \ell_i \leq \ell_j$
- There is no loss of optimality at having the two least likely symbols at maximum length
- The number of leaves at maximum depth is even

Thus, we can conclude that there exists an optimal tree where the two least likely symbols are siblings.

Huffman's procedure can be extended to any d-ary alphabet. The idea is to add symbols with probability 0 until, for some integer $\mathcal{V}$, it holds that:

$$\mathcal{V}\left(d-1\right) = |\tilde{\mathcal{X}}| - 1$$

where $\tilde{\mathcal{X}}$ is the alphabet $\mathcal{X}$ extended with the symbols with probability 0. Once this is done, Huffman's procedure is the same as the binary case, but we consider the d least likely symbols at each step. In the end, we just remove the codes for the symbols with probability 0.

2.3 Asymptotic equipartition property

Before introducing the asymptotic equipartition property, we introduce the notion of typicality.

2.3.1 Typicality

Let's consider the experiment of tossing a coin 250 times, and we are interested in the number of heads and tails we obtain. Let's say we obtain $\#H = 150$ heads and $\#T = 100$ tails. In this case, we have obtained an empirical probability of:

$$\hat{P}_X\left(H\right) = \frac{150}{200} \quad \hat{P}_X\left(T\right) = \frac{100}{200}$$

Is this the result we would have expected?

Definition 2.7 (Exact typicality)**.** Given an alphabet $\mathcal{X}$, a sequence $\underline{x} \in \mathcal{X}^n$ and a PMF P, we define:

$$\frac{1}{n}N\left(a|\underline{x}\right) = \frac{1}{n}\sum_{i=1}^{n} I\{x_i = a\}$$

We say that $\underline{x}$ is of exact type P if

$$P\left(a\right) = \frac{1}{n}N\left(a|\underline{x}\right)$$

In this case, the results we obtained are clearly not strongly typical to $P_X = (0.5, 0.5)$. A more loose definition is:

Definition 2.8 (Strong typical set)**.** Given an alphabet $\mathcal{X}$, a sequence $\underline{x} \in \mathcal{X}^n$, a PMF P and $\epsilon > 0$, the strong typical set $\mathcal{T}_\epsilon^{(n)}\left(P\right) \subseteq \mathcal{X}^n$ is defined as:

$$\mathcal{T}_\epsilon^{(n)}\left(P\right) = \left\{\underline{x} \in \mathcal{X}^n : \left|\frac{1}{n}N\left(a|\underline{x}\right) - P\left(a\right)\right| \leq \epsilon P\left(a\right)\right\}$$

We say that sequence $\underline{x}$ is strongly typical to P if $\underline{x} \in \mathcal{T}_\epsilon^{(n)}\left(P\right)$.

We can see here that ϵ allows for a non exact solution. There is an even looser version of typicality:

Definition 2.9 (Weak typical set)**.** Given an alphabet $\mathcal{X}$, a sequence $\underline{x} \in \mathcal{X}^n$, a PMF P and $\epsilon > 0$, the weak typical set $\mathcal{A}_\epsilon^{(n)}(P) \subseteq \mathcal{X}^n$ is defined as:

$$\mathcal{A}_\epsilon^{(n)}(P) = \left\{ \xi \in \mathcal{X}^n : 2^{-n(H(P)+\epsilon)} \leq \prod_{i=1}^{n} P(\xi_i) \leq 2^{-n(H(P)-\epsilon)} \right\}$$

$\prod_{i=1}^{n} P(\xi_i)$ is the probability of the sequence. We say that sequence $\underline{x}$ is weakly typical to P if $\underline{x} \in \mathcal{A}_\epsilon^{(n)}(P)$.

Remark. We note that $\mathcal{T}_\epsilon^{(n)}(P) \subseteq \mathcal{A}_\epsilon^{(n)}(P)$, meaning that strong typicality implies weak typicality.

2.3.2 AEP

We can now state and prove the **asymptotic equipartition property** (**AEP**):

Theorem 2.6 (Weak asymptotic equipartition property)**.** *Given $X_1, X_2, ..., X_n \overset{IID}{\sim} P$, it holds that:*

$$\lim_{n \to \infty} Pr\left(\underline{x} \in \mathcal{A}_\epsilon^{(n)}(P)\right) \to 1 \quad \forall \epsilon > 0$$

Proof

We need to prove that $\Pr\left(2^{-n(H(P)+\epsilon)} \leq \prod_{i=1}^{n} P(x_i) \leq 2^{-n(H(P)-\epsilon)}\right) \to 1$ for $n \to \infty$. This is equivalent of writing:

$$\begin{aligned}
&\Pr\left(\log\left(2^{-n(H(P)+\epsilon)}\right) \leq \log\left(\prod_{i=1}^{n} P(x_i)\right) \leq \log\left(2^{-n(H(P)-\epsilon)}\right)\right) = \\
&\quad = \Pr\left(-\cancel{n}(H(P)+\epsilon) \leq \frac{1}{n}\left(\sum_{i=1}^{n} \log P(x_i)\right) \leq -\cancel{n}(H(P)-\epsilon)\right) \\
&\quad = \Pr\left(H(P)-\epsilon \leq \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{P(x_i)}\right) \leq H(P)+\epsilon\right)
\end{aligned}$$

We notice that $\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{P(x_i)}\right) \approx \mathbb{E}\left[\log \frac{1}{P(x_i)}\right] = H(X)$. Using the law of large numbers, for $n \to \infty$ we conclude that:

$$\Pr\left(H(P)-\epsilon \leq \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{P(x_i)}\right) \leq H(P)+\epsilon\right) \to 1$$

It also holds that:

Theorem 2.7 (Strong asymptotic equipartition property)**.** *Given* $X_1, X_2, ..., X_n \overset{IID}{\sim} P$*, it holds that:*

$$\lim_{n\to\infty} Pr\Big(\underline{x} \in \mathcal{T}_\epsilon^{(n)}(P)\Big) \to 1 \quad \forall \epsilon > 0$$

We can now prove these two lemmas:

Lemma 2.1. *It always holds that* $|\mathcal{A}_\epsilon^{(n)}(P)| \leq 2^{n(H(P)+\epsilon)}$

Proof

$$\begin{aligned}
1 &\geq P^{x^n}\left(\mathcal{A}_\epsilon^{(n)}(P)\right) \\
&= \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} P(\xi) \\
&= \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} \prod_{i=1}^{n} P(\xi_i) \\
&\geq \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} 2^{-n(H(P)+\epsilon)} && \text{(Weak typicality)} \\
&= |\mathcal{A}_\epsilon^{(n)}(P)| 2^{-n(H(P)+\epsilon)}
\end{aligned}$$

This implies that $|\mathcal{A}_\epsilon^{(n)}(P)| \leq 2^{n(H(P)+\epsilon)}$.

■

Lemma 2.2. *For n large enough, it holds that* $|\mathcal{A}_\epsilon^{(n)}(P)| \geq 2^{n(H(P)-\epsilon)}$

Proof

$$\begin{aligned}
1-\epsilon &\leq P^{x^n}\left(\mathcal{A}_\epsilon^{(n)}(P)\right) && \text{(For } n \text{ large enough, weak AEP)} \\
&= \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} P^{x^n}(\xi) \\
&\leq \sum_{\xi \in \mathcal{A}_\epsilon^{(n)}(P)} 2^{-n(H(P)-\epsilon)} && \text{(Weak typicality)} \\
&= |\mathcal{A}_\epsilon^{(n)}(P)| 2^{-n(H(P)-\epsilon)}
\end{aligned}$$

This implies that $|\mathcal{A}_\epsilon^{(n)}(P)| \geq 2^{n(H(P)-\epsilon)}$.

■

2.3.3 Data compression

Given a sequence, $\underline{x}$, how can we compress it? The idea is to use a flag 0/1, where 0 indicates typicality and 1 the opposite:

- For typical sequences append the address in $\mathcal{A}_\epsilon^{(n)}(P)$ using $n\left(H\left(P\right)+\epsilon\right)$ bits
- For non typical sequences append the address in $\mathcal{X}^n$ using $n\log|\mathcal{X}|$ bits

We derive that the expected code length $L\left(\mathcal{C}\right)$ to describe n symbols is:

$$L\left(\mathcal{C}\right)=\Pr\left(\text{typical}\right)\left[1+n\left(H\left(P\right)+\epsilon\right)\right]+\Pr\left(\text{non typical}\right)\left[1+n\log|\mathcal{X}|\right]$$

If $n\to\infty$:

$$L\left(\mathcal{C}\right)=\underbrace{\Pr\left(\text{typical}\right)}_{\to1}\left[1+n\left(H\left(P\right)+\epsilon\right)\right]+\underbrace{\Pr\left(\text{non typical}\right)}_{\to0}\left[1+n\log|\mathcal{X}|\right]$$

Which implies the average length to describe one symbol is:

$$\frac{1}{n}L\left(\mathcal{C}\right)=\underbrace{\frac{1}{n}+\left(H\left(P\right)+\epsilon\right)}_{\to H(P)}$$

So, how much can we compress?

Theorem 2.8 (Source coding theorem)**.** *Consider an encoder $f:\mathcal{X}^n\to\{0,1\}^k$ and a decoder $\phi:\{0,1\}^k\to\mathcal{X}$. We define the probability of error as:*

$$Pr\left(error\right)=Pr\left(\phi(f\left(\underline{x}\right))\neq\underline{x}\right)$$

Then it holds that:

1. *If $k/n\geq H\left(P\right)+\epsilon$, then $P\left(error\right)\to0$ as $n\to\infty$*
2. *If $k/n<H\left(P\right)-2\epsilon$, then $P\left(error\right)\to1$ as $n\to\infty$*

We are going to prove the second part:

Proof

We choose $k/n<H\left(P\right)-2\epsilon$. Given any f,ϕ, the probability of success is defined as:

$$\Pr\left(\text{success}\right)=\Pr\left(\text{success and }\underline{x}\text{ is typical}\right)+\Pr\left(\text{success and }\underline{x}\text{ is atypical}\right)$$

We note that $\Pr\left(\text{success and }\underline{x}\text{ is atypical}\right)\to0$ as $n\to\infty$. At most, 2^k typical sequences can be correctly described, and each one has probability $\leq2^{-n(H(P)-\epsilon)}$. This means that:

$$\begin{aligned}\Pr\left(\text{success}\right)\leq2^k2^{-n(H(P)-\epsilon)}&=2^{-n(H(P)-\epsilon-k/n)}\\&\leq2^{-n(H(P)-\epsilon-H(P)+2\epsilon)}\\&=2^{-n\epsilon}\end{aligned}$$

We note that $2^{-n\epsilon}\to0$ as $n\to\infty$

■

3 Channel coding

Before introducing the notion of channel coding, we are going to look at Markov kernels.

3.1 Markov kernel

Definition 3.1 (Markov kernel)**.** Given an input alphabet $\mathcal{X}$ and an output alphabet $\mathcal{Y}$, a **Markov kernel** (or **Markov channel**) is defined as a transition probability $W(y|x)$, where $W(y|x)$ is a PMF on $\mathcal{Y}$ i.e. for all $x \in \mathcal{X}$, it holds that $W(y|x) \geq 0$ and $\sum_{y \in \mathcal{Y}} W(y|x) = 1$.

We can also introduce noiseless channels:

Definition 3.2 (Noiseless channel)**.** A channel $W(y|x)$ is said to be **noiseless** if it can be written as:

$$W(y|x) = \begin{cases} 1 & \text{if} \quad x = y \\ 0 & \text{otherwise} \end{cases}$$

Markov kernels can be written in matrix form:

$$W(y|x) = \begin{pmatrix} W(y_1|x_1) & W(y_2|x_1) & \cdots \\ W(y_2|x_1) & \ddots & \\ \vdots & & \end{pmatrix}$$

Let's look at some examples:

Example 3.1

The **binary symmetric channel** (**BSC**(p)) is a channel defined on input and output alphabets $\mathcal{X} \equiv \mathcal{Y} = \{0, 1\}$, and the transition probability follows the structure of Fig. 6 with **crossover probability** $p \in [0, 1]$.

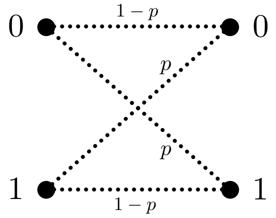


Figure 6: Binary symmetric channel

In matrix form:

$$W(y|x) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

Example 3.2

The **Z channel** is a channel defined on input and output alphabets $\mathcal{X} \equiv \mathcal{Y} = \{0, 1\}$, and the transition probability follows the structure of Fig. 7 with **crossover probability** $p \in [0, 1]$.

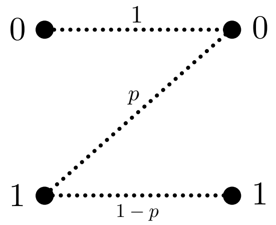


Figure 7: Z channel

In matrix form:

$$W(y|x) = \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}$$

Given the input PMF P_X, we can compute the output probability:

$$P_Y(Y = y) = \sum_{x \in \mathcal{X}} P_X(x) W(y|x)$$

Moreover, P_X in row vector form $P_X = [P_X(x_1), P_X(x_2), ..., P_X(x_n)]$ and W in matrix form, then:

$$P_Y = P_X W$$

3.2 Block code

Let's consider a BSC with crossover probability $p = 0.1$. If we send the inputs as they are through the channel, it is easy to see that the crossover probability will cause a 10% probability of error. How can we improve?

- What if we repeat every bit two times? It wouldn't really help
- What if we repeat every bit three times and do majority voting?

$$P_e = \binom{3}{2}(1-p)p^2 + \binom{3}{3}p^3 \approx 3\%$$

- What if we repeat every bit five times and do majority voting?

We see a reduction in the error probability, but every bit requires multiple channel uses and becomes very expensive to send. How can we solve this? The idea is to use **block codes**: we buffer k bits together and we send them over n channel uses. Our encoder will be in the form of $\{0,1\}^k \to \mathcal{X}^n$. We would send at rate of $\frac{k}{n}\left[\frac{\text{bits}}{\text{ch. uses}}\right]$. Our decoder would be in the form of $\phi : \mathcal{Y}^n \to \{0,1\}^k$. Furthermore, if we have the message set $\mathcal{M} \subseteq \{0,1\}^k$, we notice that we can index the messages:

$$\frac{k}{n} = \frac{\log \mathcal{M}}{n} = R \implies M = \{1, 2, ..., 2^{nR}\}$$

3.3 Rate and capacity

How do we define the **reliability of the rate R**? We could talk about:

- The probability of error of message m:

$$\lambda_m = \sum_{\underline{y}:\phi(\underline{y})\neq m} \Pr\left(Y^n = \underline{y}|M = m\right) = \sum_{\underline{y}:\phi(\underline{y})\neq m} \prod_{i=1}^{n} W\left(y_i|x_i\left(m\right)\right)$$

- The maximal probability of error $\lambda_{\max} = \max_{m\in\mathcal{M}} \lambda_m$
- The average probability of error $P_e^{(n)} = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \lambda_m$

We can now define:

Definition 3.3 (Rate achievability)**.** Given a channel, rate R is **achievable** if for every $\epsilon > 0$ there exists some $n_0, f^{(n)}, \phi^{(n)}$ of rate R such that for all $n > n_0$, $\lambda_{\max} < \epsilon$, or, equivalently, $\lim_{n\to\infty} \lambda_{\max} = 0$

Definition 3.4 (Channel capacity)**.** Given a channel, its **capacity** C is the supremum of all achievable rates ($C = \sup R$). I.e., for every $\delta > 0$, rate $C - \delta$ is achievable, while $C + \delta$ is not.

Remark. $C \geq 0$ for every channel: it always exists an achievable rate.

Definition 3.5 (Discrete memory-less channel)**.** A **discrete memory-less channel** (**DMC**) is a channel such that the distribution of $Y_i|X_i$ does not depend on the past values of $y_i|x_i$:

$$P\left(y_i|x_i, x^{i-1}, y_{i-1}\right) = P\left(y_i|x_i\right)$$

We can now state a fundamental theorem regarding channel capacity:

Theorem 3.1 (Channel coding theorem (preview))**.** *Given a DMC W, its capacity is $C = \max_{Q\in\mathcal{Q}} I\left(Q, W\right)$:*

- ***Direct:*** *$R < \max_Q I\left(Q, W\right) \implies R$ is achievable*

- ***Converse:*** *$R > \max_Q I(Q, W) \implies$ R is not achievable*

How do we compute this? Let's look at some examples:

Example 3.3

Let's consider a BSC(p). We can proceed with the following steps:

1. **Upper bound $\mathbf{I}(\mathbf{Q}, \mathbf{W})$** for every Q:

$$\begin{aligned} I(Q, W) &= I(Y; X) \\ &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x \in \{0,1\}} Q(x) H(Y|X = x) \\ &\leq \log 2 - \sum_{x \in \{0,1\}} Q(x) H(Y|X = x) && (Y \text{ takes two values}) \\ &= \log 2 - H_b(p) \sum_{x \in \{0,1\}} Q(x) && (H(Y|X = x) = H_b(p)) \\ &= 1 - H_b(p) \text{ bits} \quad \forall Q \end{aligned}$$

2. Can we find a Q such that the previously found inequality $I(Q, W) \leq 1 - H_b(p)$ holds with equality? If the input distribution Q is uniform, then the output distribution is also uniform, meaning that $H(Y) = \log 2$, meaning that $I(Q, W) = 1 - H_b(p)$

Example 3.4

Let's consider a **weakly symmetric channels**, i.e. a channel where all rows of W are permutations of the first row and all columns have the same sum:

1.

$$\begin{aligned} I(Q, W) &= I(Y; X) \\ &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x \in \{0,1\}} Q(x) H(\text{row}) \\ &= H(Y) - H(\text{row}) \leq \log |\mathcal{Y}| - H(\text{row}) \end{aligned}$$

2. If we try the uniform distribution, being that all the columns have the same sum:

$$P(Y = y) = \sum_{x \in \{0,1\}} Q(x) W(Y|X = x) = \frac{1}{|\mathcal{X}|} \cdot \text{sum of column}$$

Then the distribution of Y is also uniform, meaning that $I(Q, W) = \log |\mathcal{Y}| - H(\text{row})$

Example 3.5

Let's now consider a **binary erasure channel**, i.e. a channel with a transition matrix:

$$W\left(Y|X\right) = \begin{pmatrix} 1-p & 0 & p \\ 0 & 1-p & p \end{pmatrix}$$

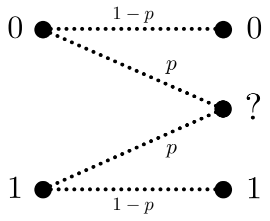


Figure 8: Binary erasure channel

These types of channels are very typical of packet communication, as the ? can represent a lost packet. We cannot use the previous step because no Q achieves the upper bound with equality.

The first way to solve this is by using symmetry. As 0s and 1s are treated in the same way, we suppose that $Q^*\left(0\right) = 1/2$ and $Q^*\left(1\right) = 1/2$. To verify that this holds we write two possible distributions $Q_0 = \left(\alpha, 1-\alpha\right), Q_1 = \left(1-\alpha, \alpha\right)$ and suppose that both achieve capacity, i.e. $C = I\left(Q_0, W\right), C = I\left(Q_1, W\right)$. By concavity:

$$I\left(\frac{1}{2}Q_0 + \frac{1}{2}Q_1, W\right) \geq \frac{1}{2}I\left(Q_0, W\right) + \frac{1}{2}I\left(Q_1, W\right) = C$$

Equality holds if $Q_0 = Q_1 = \left(1/2, 1/2\right)$ **The second way** is to proceed as follows:

$$\begin{aligned} I\left(X;Y\right) &= H\left(X\right) - H\left(X|Y\right) \\ &= H\left(X\right) - \sum_{y\in\mathcal{Y}} \left(QW\right)\left(y\right) H\left(X|Y=y\right) \\ &= H\left(X\right) - \left(QW\right)\left(?\right) H\left(X|Y=?\right) && \left(H\left(X|Y\in\{0,1\}\right) = 0\right) \\ &= H\left(X\right) - \left(Q\left(0\right)p + Q\left(1\right)p\right) H\left(X\right) \\ &= H\left(X\right) - pH\left(X\right) \\ &= \left(1-p\right) H\left(X\right) \end{aligned}$$

It is easy to see that the maximum is obtained by having $Q = \left(1/2, 1/2\right)$, in which case $H\left(X\right) = H_b\left(1/2\right) = 1$, thus $I\left(X;Y\right) = 1-p$

3.4 Karush–Kuhn–Tucker conditions

As we have not proven the Capacity of a DMC yet, we will write $C^{(I)} = \max_Q I(Q, W)$. We now state a very important theorem regarding capacity, which we will prove later:

Theorem 3.2 (Karush–Kuhn–Tucker conditions)**.** *Consider a channel $W(Y|X)$, the* ***Karush–Kuhn–Tucker conditions (KKT conditions)*** *state that:*

- ***Sufficiency****: if $Q \in \mathcal{P}(\mathcal{X})$ and $\lambda \in \mathbb{R}$ are such that:*

$$\begin{aligned} &D\left(W(\cdot|x)\,||\,(QW)(\cdot)\right) \leq \lambda \quad \forall x \in \mathcal{X} \\ &D\left(W(\cdot|x)\,||\,(QW)(\cdot)\right) = \lambda \quad \forall x \in \mathcal{X} : Q(x) > 0 \end{aligned}$$

  Then Q achieves capacity and $\lambda = C^{(I)} = \max_Q I(Q, W)$

- ***Necessity****: if Q^* achieves $\max_Q I(Q, W)$, then:*

$$\begin{aligned} &D\left(W(\cdot|x)\,||\,(QW)(\cdot)\right) \leq C^{(I)} \quad \forall x \in \mathcal{X} \\ &D\left(W(\cdot|x)\,||\,(QW)(\cdot)\right) = C^{(I)} \quad \forall x \in \mathcal{X} : Q^*(x) > 0 \end{aligned}$$

Why is this theorem important? Let's look at **Example 3.5**:

Example 3.6

Previously we found that the optimal Q^* for the BEC is $Q^* = (1/2, 1/2)$, can we verify that it is optimal? We note that $(Q^*W)(0) = \frac{1}{2}(1-p), (Q^*W)(1) = \frac{1}{2}(1-p), (Q^*W)(?) = p$:

	$Y = 0$	$Y = 1$	$Y =?$
$W(\cdot\|X = 0)$	$1-p$	0	p
$W(\cdot\|X = 1)$	0	$1-p$	p
$(QW)(\cdot)$	$\frac{1-p}{2}$	$\frac{1-p}{2}$	p

It follows:

$$\begin{aligned} &D\left(W(\cdot|X=0)\,||\,(Q^*W)(\cdot)\right) \\ &\quad = (1-p)\log\left(\frac{1-p}{\frac{1}{2}(1-p)}\right) + p\log\left(\frac{p}{p}\right) + 0\log\left(\frac{0}{\frac{1}{2}(1-p)}\right) \\ &\quad = (1-p)\log 2 = 1-p \end{aligned}$$

Analogously $D\left(W(\cdot|X=1)\,||\,(Q^*W)(\cdot)\right) = 1-p$. We have found that the sufficiency conditions hold, thus we can conclude that Q^* is indeed optimal.

We are going to prove the conditions, but we first need to introduce some more theorems:

3.4.1 Properties of $I(Q,W)$

We start by proving the following:

Theorem 3.3 (Concavity of $I(Q,W)$ w.r.t. Q). *$I(Q,W)$ is concave w.r.t. Q, i.e. for all $\lambda \in [0,1], Q_1 \in \mathcal{Q}, Q_2 \in \mathcal{Q}$ and channel W:*

$$I(\lambda Q_1 + (1-\lambda) Q_2, W) \geq \lambda I(Q_1, W) + (1-\lambda) I(Q_2, W)$$

Proof

We note $\overline{\lambda} = 1 - \lambda$. We first write:

$$I(Q,W) = H(Y) - H(Y|X) = H(QW) - \sum_{x\in\mathcal{X}} Q(x) H(Y|X=x)$$

$$\Longrightarrow I\left(\lambda Q_1 + \overline{\lambda} Q_2\right) =$$

$$= H\left(\left(\lambda Q_1 + \overline{\lambda} Q_2\right) W\right) - \sum_{x\in\mathcal{X}} \left(\lambda Q_1 + \overline{\lambda} Q_2\right)(x) H(Y|X=x)$$

It follows:

$$\begin{aligned} & I\left(\lambda Q_1 + \overline{\lambda} Q_2\right) \\ & \quad = H\left(\lambda Q_1 W + \overline{\lambda} Q_2 W\right) - \sum_{x\in\mathcal{X}} \left[\lambda Q_1(x) H(Y|X=x) + \overline{\lambda} Q_2(x) H(Y|X=x)\right] \\ & \quad \geq \lambda H(Q_1 W) + \overline{\lambda} H(Q_2 W) - \sum_{x\in\mathcal{X}} \left[\lambda Q_1(x) H(Y|X=x) + \overline{\lambda} Q_2(x) H(Y|X=x)\right] \\ & \quad = \lambda I(Q_1, W) + \overline{\lambda} I(Q_2, W) \end{aligned}$$

■

3.4.2 Concave maximization

How do we find the maximum of a concave function? If we are in a constrained problem, we also need ot check the edges of our constraint. Without the constraint, to check if ξ^* achieves the maximum of $f(\xi)$ we would need to check:

1. $\lim_{k\to 0^-} \frac{f(\xi^*+k)-f(\xi^*)}{k} = f'_+(\xi^*) \leq 0$
2. $\lim_{k\to 0^+} \frac{f(\xi^*+k)-f(\xi^*)}{k} = f'_-(\xi^*) \geq 0$

But if we are constrained and the optimal point lies on one edge, we cannot do this. For this reason, the KKT conditions differentiate the cases where $Q = 0$ and $Q > 0$:

Theorem 3.4 (Concave maximization). *Let $f(\alpha)$ be a concave function of $\alpha = (\alpha_1, ..., \alpha_n)$ over the probability simplex $\mathcal{R} \triangleq \{\alpha : \alpha_i \geq 0 \; \forall i. \sum_{i=1}^n \alpha_i = 1\}$. Assume that the partial derivatives, $\frac{\delta f(\alpha)}{\delta a_k}$ are defined and continuous over the* *simplex $\mathcal{R}$ with the possible exception that $\lim_{\alpha_k \downarrow 0} \frac{\delta f(\alpha)}{\delta a_k}$ may be $+\infty$. Then for $\lambda' \in \mathbb{R}$:*

$$\left.\frac{\delta f(\alpha)}{\delta a_k}\right|_{\alpha=\alpha^*} = \lambda' \quad \forall k \text{ s.t. } \quad a_k^* > 0$$
$$\left.\frac{\delta f(\alpha)}{\delta a_k}\right|_{\alpha=\alpha^*} \leq \lambda' \quad \forall k \text{ s.t. } \quad a_k^* = 0$$

are necessary and sufficient conditions on a probability vector α^* to maximize $f(\alpha)$ over $\mathcal{R}$.

3.4.3 Proof of KTT conditions

We can now prove the KKT conditions:

Proof

We first write:

$$\begin{aligned} I(Q,W) &= \sum_x \sum_y Q(x) W(y|x) \ln \left(\frac{W(y|x) Q(x)}{Q(x)(QW)(y)} \right) \\ &= \sum_x \sum_y Q(x) W(y|x) \ln \left(\frac{W(y|x)}{\sum_{x'} Q(x') W(y|x')} \right) \end{aligned}$$

(We use the natural logarithms without loss of generality). We can now compute the derivatives, using $Q_k = Q(x_k)$:

$$\begin{aligned} \frac{\delta I(Q,W)}{\delta Q_k} =& \sum_x \sum_y I\{x = x_k\} W(y|x) \ln \left(\frac{W(y|x)}{\sum_{x'} Q(x') W(y|x')} \right) \\ &+ \sum_x \sum_y Q(x) W(y|x) \cdot \frac{\sum_{x'} Q(x') W(y|x')}{W(y|x)} \cdot \frac{-W(y|x) W(y|x_k)}{\left(\sum_{x'} Q(x') W(y|x')\right)^2} \\ =& \sum_y W(y|x_k) \ln \frac{W(y|x_k)}{\sum_{x'} Q(x') W(y|x')} - \sum_y \frac{W(y|x_k)}{\sum_{x'} Q(x') W(y|x')} \sum_x Q(x) W(y|x) \\ =& \sum_y W(y|x_k) \ln \frac{W(y|x_k)}{\sum_{x'} Q(x') W(y|x')} - \sum_y W(y|X_k) \\ =& \sum_y W(y|x_k) \ln \frac{W(y|x_k)}{\sum_{x'} Q(x') W(y|x')} - 1 \\ =& D(W(\cdot|x_k) \,||\, (QW)(\cdot)) - 1 \end{aligned}$$

Sufficiency condition: from Theorem 3.4, by setting $\lambda' = \lambda - 1$ we conclude that Q maximizes $I(\cdot|W)$ over all input distributions, i.e. Q achieves capacity.

Then:

$$\begin{aligned}
C \overset{(i)}{=} I(Q,W) &= \sum_x \sum_y Q(x) W(y|x) \ln\left(\frac{W(y|x)}{\sum_{x'} Q(x') W(y|x')}\right) \\
&= \sum_x Q(x) \sum_y W(y|x) \ln\left(\frac{W(y|x)}{(QW)(y)}\right) \\
&= \sum_x Q(x) D\left(W(\cdot|x) \,||\, (QW)(\cdot)\right) \\
&\overset{(ii)}{=} \lambda
\end{aligned}$$

where (i) holds because Q achieves capacity and (ii) follows from the hypothesis of the condition.

Necessary condition: because Q maximizes $I(\cdot, W)$ over all input distributions, we know by Theorem 3.4 that there exists a λ' such that:

$$\begin{aligned}
D\left(W(\cdot|x) \,||\, (QW)(\cdot)\right) = \lambda' + 1 \quad \forall x \in \mathcal{X} : Q(x) > 0 \\
D\left(W(\cdot|x) \,||\, (QW)(\cdot)\right) \leq \lambda' + 1 \quad \forall x \in \mathcal{X} : Q(x) = 0
\end{aligned}$$

from the same computations as before, we obtain $C = I(Q,W) = \lambda' + 1$, it follows that $\lambda' + 1 = C$

3.4.4 Convexity of relative entropy

We can now prove that:

Theorem 3.5 (Convexity of relative entropy)**.** *The relative entropy $D(\cdot|\cdot)$ is a convex function, i.e. given four distributions P_1, P_2, Q_1, Q_2, for every $\lambda \in [0,1]$ it holds that:*

$$D\left(\lambda P_1 + (1-\lambda) P_2 || \lambda Q_1 + (1-\lambda) Q_2\right) \leq \lambda D(P_1||Q_1) + (1-\lambda) D(P_2||Q_2)$$

Proof

We consider $\overline{\lambda} = 1 - \lambda$. We can write:

$$\begin{aligned}
&D\left(\lambda P_1 + \overline{\lambda} P_2 || \lambda Q_1 + \overline{\lambda} Q_2\right) = \\
&\quad = \sum_{x \in \mathcal{X}} \left(\lambda P_1(x) + \overline{\lambda} P_2(x)\right) \log \frac{\lambda P_1(x) + \overline{\lambda} P_2(x)}{\lambda Q_1(x) + \overline{\lambda} Q_2(x)} \\
&\quad \overset{(i)}{\leq} \sum_{x \in \mathcal{X}} \left[\lambda P_1(x) \log \frac{\lambda P_1(x)}{\lambda Q_1(x)} + \overline{\lambda} P_2(x) \log \frac{\lambda P_2(x)}{\lambda Q_2(x)}\right] \\
&= \lambda D(P_1||Q_1) + \overline{\lambda} D(P_2||Q_2)
\end{aligned}$$

Where (i) holds because $\log(\cdot)$ is concave and non-decreasing.

3.4.5 Convexity of $I(Q, W)$ w.r.t. W

We can now prove that:

Theorem 3.6 (Convexity of $I(Q, W)$ w.r.t. W)**.** *$I(Q, W)$ is convex w.r.t. W, i.e. for all $\lambda \in [0, 1], W_1, W_2, Q \in \mathcal{Q}$:*

$$I(Q; \lambda W_1 + (1-\lambda) W_2) \leq \lambda I(Q, W_1) + (1-\lambda) I(Q, W_2)$$

Proof

We can write:

$$\begin{aligned} I(Q, W) &= D(Q \circ W || Q(X)(QW)(y)) \\ &= \sum_{x,y} Q(x) W(y|x) \log \frac{Q(x) W(y|x)}{Q(x)(QW)(y)} \end{aligned}$$

We know that:

$$\begin{aligned} &Q \circ \left(\lambda W_1 + \overline{\lambda} W_2\right) = \lambda (Q \circ W_1) + \overline{\lambda} (Q \circ W_2) \\ &Q \left(\lambda W_1 + \overline{\lambda} W_2\right) = \lambda (QW_1) + \overline{\lambda} (QW_2) \end{aligned}$$

Thus:

$$\begin{aligned} I\left(Q, \lambda W_1 + \overline{\lambda} W_2\right) &= D\left(Q \circ \left(\lambda W_1 + \overline{\lambda} W_2\right) || Q(X) \left(Q\left(\lambda W_1 + \overline{\lambda} W_2\right)\right)(Y)\right) \\ &= D\left(\lambda (Q \circ W_1) + \overline{\lambda} (Q \circ W_2) || \lambda (QW_1) + \overline{\lambda} (QW_2)\right) \end{aligned}$$

By applying the convexity of relative entropy we prove the theorem.

3.5 Data processing inequality

The **data processing inequality** (**DPI**) is another important theorem regarding channel coding.

Theorem 3.7 (Data processing inequality for $D(P||Q)$)**.** *$D(PW||QW) \leq D(P||Q)$*

Proof

We can write:

$$\begin{aligned}
D\left(PW \| QW\right) &= \sum_{y\in\mathcal{Y}} \left(PW\right)\left(y\right) \log \frac{\left(PW\right)\left(y\right)}{\left(PW\right)\left(y\right)} \\
&= \sum_{y\in\mathcal{Y}} \left(\sum_{x\in\mathcal{X}} P\left(x\right) W\left(y|x\right) \right) \log \frac{\sum_{x\in\mathcal{X}} P\left(x\right) W\left(y|x\right)}{\sum_{x\in\mathcal{X}} Q\left(x\right) W\left(y|x\right)} \\
&\leq \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} P\left(x\right) W\left(y|x\right) \log \frac{P\left(x\right) \cancel{W\left(y|x\right)}}{Q\left(x\right) \cancel{W\left(y|x\right)}} && \text{(Log-sum)} \\
&= \sum_{x\in\mathcal{X}} P\left(x\right) \log \frac{P\left(x\right)}{Q\left(x\right)} = D\left(P \| Q\right)
\end{aligned}$$

Let's now consider a recording of a Luis Armstrong track X produced Y in the 1930s. The track Y is remastered on Z. There is no way to improve the quality of the sound given the recording Y, and this is what the data processing inequality for the mutual information:

Theorem 3.8 (Data processing inequality for mutual information)**.** *Given three input variables X, Y, Z such that $X \multimap\!\!- Y \multimap\!\!- Z$, then $I\left(X;Z\right) \leq I\left(X;Y\right)$.*

$\boxed{\textbf{Proof}}$

We can write:

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

It follows:

$$\begin{aligned}
&I\left(X;Y\right) + \cancelto{0}{I\left(X;Z|Y\right)} = I\left(X;Z\right) + \underbrace{I\left(X;Y|Z\right)}_{\geq 0} \qquad \left(X \multimap\!\!- Y \multimap\!\!- Z\right) \\
&\implies I\left(X;Z\right) \leq I\left(X;Y\right)
\end{aligned}$$

3.6 Channel coding theorem

We will now prove the following theorem:

Theorem 3.9 (Channel coding theorem)**.** *Given a DMC W, it holds that:*

$$C = C^{(I)} = \max_{Q\in\mathcal{Q}} I\left(Q, W\right)$$

To prove this theorem, we will split it in two parts:

1. **Direct part**: $C \geq C^{(I)}$
2. **Converse part**: $C \leq C^{(I)}$

3.6.1 Converse part

To prove the converse part, we will first state and prove the following lemma:

Lemma 3.1. *Given a arbitrary $X^n \sim P_{X^n}$, and Y_i being the output of input X_i passed through channel W, it holds that $I\left(X^n;Y^n\right) \leq nC^{(I)}$*

Proof

Considering the fact that the channel is a DMC, it implies that $Y_i \perp Y_j$ and $Y_i \perp X_j$ for $i \neq j$. It follows

$$\begin{aligned}
I\left(X^n;Y^n\right) &= H\left(Y^n\right) - H\left(Y^n|X^n\right) \\
&= \sum_{i=1}^{n} H\left(Y_i|Y^{i-1}\right) - \sum_{i=1}^{n} H\left(Y_i|Y^{i-1},X^n\right) \\
&= \sum_{i=1}^{n} H\left(Y_i\right) - \sum_{i=1}^{n} H\left(Y_i|X_i\right) \\
&= \sum_{i=1}^{n} \underbrace{I\left(X_i;Y_i\right)}_{\leq C^{(I)}} \leq nC^{(I)}
\end{aligned}$$

■

We can now prove the converse part, i.e. $C \leq C^{(I)}$:

Proof

To prove this theorem we will use the following tools:

- **Fano's inequality**: $H\left(W|Y\right) \leq p_e \log(|W|) + 1$
- **DPI**: $W -\!\!\circ\!\!- X -\!\!\circ\!\!- Y \implies I\left(W;Y\right) \leq I\left(W;X\right)$
- **Lemma 3.1**

First, we consider the following mapping:

$$\begin{aligned}
&f:\mathcal{M} \to \mathcal{X}^n, \quad \mathcal{M} = \{1,...,2^{nR}\}, \quad f\left(m\right) = x\left(m\right) = \left(x_1\left(m\right),...,x_n\left(m\right)\right) \\
&\phi:\mathcal{Y}^n \to \mathcal{M} \\
&\lambda_m = \sum_{\underline{y}:\phi\left(\underline{y}\neq m\right)} \prod_{i=1}^{n} W\left(y_i|x_i\left(m\right)\right)
\end{aligned}$$

We need to prove that $\lambda_{\max} = \max_{m\in\mathcal{M}} \lambda_m \to 0$ implies $R \leq C^{(I)}$, meaning that if a rate is achievable it must be smaller than $C^{(I)}$. Without loss of generality, we will consider that the messages are drawn uniformly at random. To makes things easier, we will consider the case in which:

$$\begin{aligned}
P_e^{(n)} &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \lambda_m \\
&= \sum_{m \in \mathcal{M}} \underbrace{\Pr\left(M = m\right)}_{1/|\mathcal{M}|} \underbrace{\Pr\left(\hat{M} \neq m | M = m\right)}_{\lambda_m} \\
&= \Pr\left(\hat{M} \neq M\right) = p_e \to 0
\end{aligned}$$

(note that this implies the original assumption). We have that $M -\!\!\circ\!\!- X^n\left(M\right) -\!\!\circ\!\!- Y^n -\!\!\circ\!\!- \hat{M}$, thus:

$$\begin{aligned}
nR = H\left(M\right) &= H\left(M|\hat{M}\right) + I\left(M;\hat{M}\right) && \left(\text{Def of } I\left(M;\hat{M}\right)\right) \\
&\leq p_e \log|\mathcal{M}| + 1 + I\left(M;\hat{M}\right) && \text{(Fano's inequality)} \\
&= P_e^{(n)} \log|\mathcal{M}| + 1 + I\left(M;\hat{M}\right) && \left(p_e = P_e^{(n)}\right) \\
&\leq P_e^{(n)} nR + 1 + I\left(X^n;Y^n\right) && \text{(DPI)} \\
&= n\left(P_e^{(n)} R + \frac{1}{n}\right) + I\left(X^n;Y^n\right) \\
&\leq n\left(P_e^{(n)} R + \frac{1}{n}\right) + nC^{(I)} && \text{(Lemma 3.1)}
\end{aligned}$$

Now we can compute the limit for $n \to \infty$:

$$\begin{aligned}
\cancel{n}R \leq \cancel{n}\left(\cancelto{0}{P_e^{(n)}} R + \cancelto{0}{\frac{1}{n}}\right) + \cancel{n}C^{(I)} \\
\implies R \leq C^{(I)}
\end{aligned}$$

This implies that:

$$C \leq C^{(I)}$$

3.6.2 Feedback channel

What would happen if we introduce feedback from the output to the input? Could this improve the capacity? The idea here is that the encoder takes into account also the previous outputs, i.e.

$$x_i = f_i(m, y^{i-1})$$

Theorem 3.10 (Capacity of feedback channel)**.** *In a feedback channel, even with perfect feedback, if R is achievable, R must be smaller than* $C^{(I)}$

Proof

As in the previous proof, we can write:

$$\begin{aligned} I\left(M;Y^n\right) &= H\left(Y^n\right) - H\left(Y^n|M\right) \\ &= \sum_{i=1}^{n} H\left(Y_i|Y^{i-1}\right) - \sum_{i=1}^{n} H(Y_i| \underbrace{Y^{i-1},M}_{x_i=f(y^{i-1},m)}) \\ &\leq \sum_{i=1}^{n} H\left(Y_i\right) - \sum_{i=1}^{n} H\left(Y_i|X_i\right) && \text{(Theorem 1.10)} \\ &= \sum_{i=1}^{n} I\left(X_i;Y_i\right) \leq nC^{(I)} \end{aligned}$$

We can then proceed as in the previous proof

3.6.3 Direct part

To prove this part, we need to introduce some important notions:

3.6.3.1 Joint typicality

The first one is **joint weak typicality**

Definition 3.6 (Joint strong typicality)**.** Consider two chance variables X, Y defined over alphabets $\mathcal{X}, \mathcal{Y}$, with joint distribution $P_{X,Y} \in \mathcal{P}\left(\mathcal{X} \times \mathcal{Y}\right)$. Given an integer n and $\epsilon > 0$ we define the **joint strong typical set** as:

$$\begin{aligned} \mathcal{T}_{\epsilon}^{(n)}\left(P_{X,Y}\right) = \{(\xi,\eta) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ \left|\frac{1}{n} N\left(\left(a,b\right)|\underline{\xi},\underline{\eta}\right) - P_{X,Y}\left(a,b\right)\right| \leq \epsilon P_{X,Y}\left(a,b\right), \forall \left(a,b\right) \in \mathcal{X} \times \mathcal{Y}\bigg\} \end{aligned}$$

Definition 3.7 (Joint weak typicality)**.** Consider two chance variables X, Y defined over alphabets $\mathcal{X}, \mathcal{Y}$, with joint distribution $P_{X,Y} \in \mathcal{P}\left(\mathcal{X} \times \mathcal{Y}\right)$. Given an integer n and $\epsilon > 0$ we define the **joint weak typical set** as:

$$\begin{aligned} \mathcal{A}_{\epsilon}^{(n)}\left(P_{X,Y}\right) = \big\{\left(\underline{\xi},\underline{\eta}\right) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ 2^{-n(H(X,Y)+\epsilon)} < \prod_{i=1}^{n} P_{X,Y}\left(\xi_i,\eta_i\right) < 2^{-n(H(X,Y)-\epsilon)}, \\ 2^{-n(H(X)+\epsilon)} < \prod_{i=1}^{n} P_{X}\left(\xi_i\right) < 2^{-n(H(X)-\epsilon)}, \\ 2^{-n(H(Y)+\epsilon)} < \prod_{i=1}^{n} P_{Y}\left(\eta_i\right) < 2^{-n(H(Y)-\epsilon)}\Bigg\} \end{aligned}$$

We also introduce some properties:

Lemma 3.2. *If* $(x_1, y_1), (x_2, y_2), ..., \overset{IID}{\sim} P_{X,Y}$ *then it holds that for* $n \to \infty$*:*

$$Pr\Big((X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})\Big) \to 1$$

Lemma 3.3. *It holds that* $\left|\mathcal{A}_\epsilon^{(n)}(P_{X,Y})\right| \leq 2^{n(H(X,Y)+\epsilon)}$

Lemma 3.4. *It holds that* $\left|\mathcal{A}_\epsilon^{(n)}(P_{X,Y})\right| \geq (1-\epsilon)\, 2^{n(H(X,Y)-\epsilon)}$ *for* n *large enough.*

These lemmas derive directly from the properties that we previously derived for $\mathcal{A}_\epsilon^{(n)}(P)$ by setting $P = P_{X,Y}$. Another lemma that we need is the following:

Lemma 3.5. *If* $(x_1, y_1), (x_2, y_2), ..., \overset{IID}{\sim} P_X \times P_Y$*, i.e.:*

- $X^n \perp Y^n$
- $X^n \overset{IID}{\sim} P_X$
- $Y^n \overset{IID}{\sim} P_Y$

Then it holds that:

$$Pr\Big((X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})\Big) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

We are also going to prove this last lemma:

Proof

We can write:

$$\begin{aligned}
\Pr\Big((X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})\Big) &= \sum_{(\underline{\xi}, \underline{\eta}) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})} \Pr\left(X^n = \underline{\xi}, Y^n = \underline{\eta}\right) \\
&\overset{(i)}{=} \sum_{(\underline{\xi}, \underline{\eta}) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})} \Pr\left(X^n = \underline{\xi}\right) \Pr\left(Y^n = \underline{\eta}\right) \\
&\overset{(ii)}{\leq} \sum_{(\underline{\xi}, \underline{\eta}) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\
&\overset{(iii)}{\leq} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} 2^{n(H(X,Y)+\epsilon)} \\
&= 2^{-n(H(X)+H(Y)-H(X,Y)-3\epsilon)} \\
&= 2^{-n(I(X;Y)-3\epsilon)}
\end{aligned}$$

Where (i) holds because $X_n \perp Y_n$, (ii) holds because $\underline{\xi} \in \mathcal{A}_\epsilon^{(n)}(P_X), \underline{\eta} \in \mathcal{A}_\epsilon^{(n)}(P_Y)$ (definition of joint weak typical set) and (iii) from **Lemma 3.3**.

3.6.3.2 Proof

We can now prove the direct (achievability) part, i.e. $C \geq C^{(I)}$:

$\boxed{\textbf{Proof}}$

We need to show that given a DMC W, if $R < C^{(I)}$, then R is achievable, and this implies that $C \geq C^{(I)}$. To do so, we will proceed in two steps:

1. We show that $\exists\{\mathcal{C}^{(n)}\} \quad \text{s.t.} \quad P_e^{(n)}\left(\mathcal{C}\right) \overset{n\to\infty}{\to} 0$
2. We show that $\exists\{\mathcal{C}^{(n)}\} \quad \text{s.t.} \quad \lambda_{\max} \overset{n\to\infty}{\to} 0$

Step 1: we build a **weak joint typicality decoder** determined by $P_{X,Y}, \epsilon, n$. Given an output sequence $\underline{y}$, it checks:

$$\begin{gathered}\left(\underline{x}\left(1\right), \underline{y}\right) \overset{?}{\in} \mathcal{A}_\epsilon^{(n)}(P_{X,Y})\\ \vdots \\ \left(\underline{x}\left(2^{nR}\right), \underline{y}\right) \overset{?}{\in} \mathcal{A}_\epsilon^{(n)}(P_{X,Y})\end{gathered}$$

If it finds only one $x\left(\tilde{m}\right)$ such that $\left(x\left(\tilde{m}\right), \underline{y}\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})$ then it concludes that the correct decoding is $\hat{m} = \tilde{m}$. If it finds none or more than one it concludes that it cannot decode $\underline{y}$ correctly. Our objective is to show that:

$$\begin{aligned}\sum_{\mathcal{C}} P\left(\mathcal{C}\right) P_e^{(n)}\left(\mathcal{C}\right) &< \text{small}\left(\epsilon\right)\\ \implies \exists\mathcal{C}^* \quad \text{s.t.} \quad P_e^{(n)} &< \text{small}\left(\epsilon\right)\end{aligned}$$

So how do we build our encoder? Our encoder will encode every message with n bits. We notice that the total number of bits in the encoder is the number of total messages 2^{nR} times the number of bits n. We sample these bits from a distribution Q: $2^{nR}n$ entries $\overset{IID}{\sim} Q$. From now on we will use this notation:

- $\overline{\lambda_m} = \sum_{\mathcal{C}} P\left(\mathcal{C}\right) \lambda_m\left(\mathcal{C}\right), \; m \in \mathcal{M}$
- $\overline{P}_e^{(n)} = \sum_{\mathcal{C}} P\left(\mathcal{C}\right) P_e^{(n)}(\mathcal{C})$
- $P_e^{(n)} = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \lambda_m$

Because the code-words are chosen IID, we can write that $\overline{\lambda}_m = \overline{\lambda}_1 \; \forall m \in \mathcal{M}$:

$$\begin{aligned}
\overline{P}_e^{(n)} &= \sum_{\mathcal{C}} P\left(\mathcal{C}\right) P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} P\left(\mathcal{C}\right) \frac{1}{2^{nR}} \sum_{m \in \mathcal{M}} \lambda_m\left(\mathcal{C}\right) \\
&= \frac{1}{2^{nR}} \sum_{m \in \mathcal{M}} \underbrace{\sum_{\mathcal{C}} P\left(\mathcal{C}\right) \lambda_m\left(\mathcal{C}\right)}_{\overline{\lambda}_m} = \overline{\lambda}_1
\end{aligned}$$

Now, if we show that $\overline{\lambda_1} \to 0$, then we also show that $\overline{\lambda_m} \to 0 \implies \overline{P}_e^{(n)} \to 0$. Now we ask, what is the source of randomness in this experiment? We are sending m_1 over the channel W which is a source of randomness, and the encoded $x\left(m_1\right)$ randomness depends on the distribution Q. We can write:

$$\begin{aligned}
&Q\left(X\right) W\left(Y|X\right) \equiv Q \circ W \equiv P_{X,Y} \\
&\left(x_1\left(m_1\right), y_1\right), \left(x_2\left(m_1\right), y_2\right), ..., \left(x_n\left(m_1\right), y_n\right) \overset{IID}{\sim} Q \circ W
\end{aligned}$$

How do we compute λ_1? If we denote with E_i the event of the output $\underline{y}$ being jointly typical with $x\left(m_i\right)$ (i.e. $\left(x\left(m_i\right), y\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})$), and with $\overline{E}_i$ the complementary event, we have:

$$\begin{aligned}
\overline{\lambda_1} &= \Pr\left(\overline{E}_1 \cup \bigcup_{j=2}^{2^{nR}} E_j \,\middle|\, m = 1 \right) \\
&\leq \Pr\left(\overline{E}_1 | m = 1\right) + \sum_{i=2}^{2^{nR}} \Pr\left(E_j | m = 1\right) \qquad \text{(Union bound)}
\end{aligned}$$

From the complement of the **asymptotic equipartition property** we know that for $n \to \infty$, $\Pr\left(\overline{E}_1 | m = 1\right) \to 0$. Regarding the other term, we know that $\underline{y}$ is generated according to $\underline{y} \sim W$, and is dependent on row 1 (we are sending message 1, meaning that we encoded it with the first row of our code book). We also know that $x\left(m_j\right) \overset{IID}{\sim} Q$, and for $j \neq 1$, x_j is independent of $\underline{y}$. This means that we are in the situation described in **Lemma 3.5**, and we can write that:

$$\Pr\left(E_j | m = 1\right) \leq 2^{-n(I(Q,W) - 3\epsilon)} \qquad \text{for} \quad j \neq 1$$

We have established the values of the two terms, this means that:

$$\begin{aligned}
&\overline{\lambda}_1 \leq \Pr\left(\overline{E}_1\right) + \sum_{i=2}^{2^{nR}} \Pr\left(E_j | m = 1\right) \\
&\overset{n \to \infty}{\Longrightarrow} \overline{\lambda}_1 \leq \left(2^{nR} - 1\right) 2^{-n(I(Q,W) - 3\epsilon)} \approx 2^{-n(I(Q,W) - 3\epsilon - R)}
\end{aligned}$$

This implies that for $R \leq I(Q,W) - 3\epsilon$ we have that $P_e^{(n)} \to 0$. Hence, for $R < I(Q,W)$ the probability of error goes to 0, and have proven Part 1.

Part 2: if we have $R < I(Q,W)$ we have proven that for all ϵ, we can find a code $\mathcal{C}^*$ such that for a large enough n, $P_e^{(n)}(\mathcal{C}^*) < \epsilon$. Let's now order the errors for all the messages in ascending order:

$$\lambda_1(\mathcal{C}^*) \leq \lambda_2(\mathcal{C}^*) \leq \cdots \leq \lambda_{2^{nR}}(\mathcal{C}^*)$$

What happens if we throw away the bad half of the code-words? First we note that given a sequence $a_1 \leq a_2 \leq \cdots \leq a_{2\nu}$, such that $\sum_{i=1}^{2\nu} a_i < 2\nu\epsilon$ and $a_i \geq 0$, it holds that:

$$2\nu\epsilon > \sum_{i=1}^{2\nu} a_i \geq \sum_{i=\nu+1}^{2\nu} a_i \geq a_\nu \nu \implies a_\nu < 2\epsilon$$

We note that $P_e^{(n)} = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_m < \epsilon$, which implies $\sum_{i=1}^{2^{nR}} \lambda_m < 2^{nR}\epsilon$. Thus, if we throw away half of the code, we bound $\overline{\lambda}_{\max} < 2\epsilon$, meaning that for $n \to \infty$, $\overline{\lambda}_{\max} \to 0$. But what happens to the rate? The reduction in the rate we obtain does not influence the rate. The rate R' we would obtain if we throw away half of the messages we obtain:

$$R' = \frac{\log(|\mathcal{M}|/2)}{n} = \frac{\log|\mathcal{M}|}{n} - \frac{\log 2}{n} \overset{n\to\infty}{=} \frac{\log|\mathcal{M}|}{n} = R$$

This implies that for $n \to \infty$, if $P_e^{(n)} \to 0$ then $\overline{\lambda}_{\max} \to 0$. Thus, we have proven **Part 2**.

Conclusion: we have proven that to have $\overline{\lambda}_{\max} \overset{n\to\infty}{\to} 0$, it must hold that $R \leq I(Q,W) - 3\epsilon$. As this holds for every $\epsilon > 0$, we have proven that if $R < I(Q,W)$ (which implies $R < \max_Q I(Q,W)$), then R is achievable. This implies that $C \geq C^{(I)}$.

3.7 Fano's inequality for sequences

We are going to state and prove the following:

Theorem 3.11 (Fano's inequality for sequences)**.** *Given a k-long sequence U^k, observations Y^n and decoded observations $\hat{U}^k(Y^n)$, if we define:*

$$P_{e,i} = Pr\left(U_i \neq \hat{U}_i\right)$$

$$P_{avg} = \frac{1}{k}\sum_{i=1}^{k} P_{e,i}$$

Then it holds that:

$$\frac{1}{k}H\left(U^k|\hat{U}^k\right) \leq H_b\left(P_{avg}\right) + P_{avg}\log\left(|\mathcal{U}|-1\right)$$

Proof

We can write:

$$\begin{aligned}
\frac{1}{k}H\left(U^k|\hat{U}\right) &= \sum_{i=1}^{k}\frac{1}{k}H\left(U_i|U^{i-1},\hat{U}\right) \\
&\leq \frac{1}{k}\sum_{i=1}^{k}H\left(U_i|\hat{U}_i\right) && \text{(Theorem 1.10)} \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\left(H_b\left(P_{e,i}\right) + P_{e,i}\log\left(|\mathcal{U}|-1\right)\right) && \text{(Fano's inequality)} \\
&= \frac{1}{k}\sum_{i=1}^{k}H_b\left(P_{e,i}\right) + \frac{1}{k}\sum_{i=1}^{k}P_{e,i}\log\left(|\mathcal{U}|-1\right) \\
&= \frac{1}{k}\sum_{i=1}^{k}H_b\left(P_{e,i}\right) + P_{\text{avg}}\log\left(|\mathcal{U}|-1\right) \\
&\leq H_b\left(P_{\text{avg}}\right) + P_{\text{avg}}\log\left(|\mathcal{U}|-1\right) && \text{(Concavity)}
\end{aligned}$$

■

3.8 Source-channel separation

In digital communications, usually, the source and the channel are separated. Does this influence the amount of information we can send?

Theorem 3.12 (Source-channel separation theorem)**.** *Given a source U that produces symbols at rate $\rho = k/n$ [src symb. / ch. uses], and a channel with capacity C it holds that:*

1. *If $H(U)\cdot\rho < C$ then the source can be communicated reliably with the channel separation approach*
2. *If $H(U)\cdot\rho > C$ then the source cannot be communicated reliably with the channel separation approach*

With reliably we mean that $\Pr\left(\hat{U}^k \neq U^k\right) \overset{n\to\infty}{\to} 0$.

Proof

Part 1: the proof is really simple. We can use the weak typical set $\mathcal{A}_\epsilon^{(n)}(P_U)$. If the sequence produced belongs to $\mathcal{A}_\epsilon^{(n)}(P_U)$, we send the tuple containing the address of the sequence. We know that:

$$|\mathcal{A}_\epsilon^{(n)}(P_U)| \leq 2^{k(H(U)+\epsilon)} \tag{5}$$

This means that we will need $k\left(H\left(U\right)+\epsilon\right)$ bits to describe one address. We have that the source produces ρ source symbols per channel use, so if we find a code book such that $\rho\left(H\left(U\right)+\epsilon\right) < C$ we are done. We note that we have errors if the source sequence does not belong to $\mathcal{A}_\epsilon^{(n)}(P_U)$. From the AEP we know that the probability of this goes to 0 with $n \to \infty$, and we have proven the first part.

Part 2: we consider that the source can be reliably communicated with the channel-separation approach. As the source is memory-less, we note that $H\left(U^k\right) = kH\left(U\right)$:

$$\begin{aligned} H\left(U^k\right) &= \frac{1}{k}H\left(U^k\right) - \frac{1}{k}H\left(U^k|\hat{U}^k\right) + \frac{1}{k}H\left(U^k|\hat{U}^k\right) \\ &\leq \frac{1}{k}I\left(U^k;\hat{U}^k\right) + H_b\left(P_{\text{avg}}\right) + P_{\text{avg}}\log\left(|\mathcal{U}|-1\right) && \text{(Fano's for seq.)} \\ &\leq \frac{1}{k}I\left(X^n;Y^n\right) + H_b\left(P_{\text{avg}}\right) + P_{\text{avg}}\log|\mu| && \text{(DPI)} \\ &\leq \frac{1}{k}nC + H_b\left(P_{\text{avg}}\right) + P_{\text{avg}}\log|\mu| && \text{(Lemma 3.1)} \end{aligned}$$

For $n \to \infty$ we notice that the second and third terms go to 0 by hypothesis. This implies that:

$$H\left(U\right) \leq \frac{n}{k}C \implies H\left(U\right)\rho \leq C$$

This condition is necessary for reliable communication, implying that if

$$H\left(U\right)\rho > C$$

the source cannot be reliably communicated via channel-separation.

4 Rate-distortion theory

For now, we have considered almost lossless coding. We now introduce rate-distortion theory, which takes into account the cases in which we allow some distortion. Let's consider a memory-less source on alphabet $\mathcal{X}$ with PMF P_X. We pass the input to our channel and obtain an output in alphabet $\hat{\mathcal{X}}$ (we usually consider $\mathcal{X} \equiv \hat{\mathcal{X}}$):

Definition 4.1 (Distortion function)**.** Given an input alphabet $\mathcal{X}$ and an output alphabet $\hat{\mathcal{X}}$, the distortion function is a function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$.

Example 4.1

Let's consider the input alphabet $\mathcal{X} = \{0,1\}$ and the output alphabet $\hat{\mathcal{X}} = \{0,1\}$. The **Hamming distortion** (or **Hamming distance**) is the function:

$$d_H(x, \hat{x}) = \begin{cases} 1 & \text{if} \quad x \neq \hat{x} \\ 0 & \text{otherwise} \end{cases}$$

We can also extend the definition:

Definition 4.2 (Distortion function for sequences)**.** Given a distortion function d, an input sequence X^n and an output sequence $\hat{X}^n$, the distortion of the sequence is:

$$d\left(X^n, \hat{X}^n\right) = \frac{1}{n}\sum_{i=1}^{n} d\left(X_i, \hat{X}_i\right)$$

.

4.1 Rate

Let's now consider having an encoder/decoder scheme:

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

We want to allow a maximal distortion D:

Definition 4.3 (Achievable rate-distortion)**.** We say that (R, D) is achievable if for any $\delta > 0$ and large enough n, it exists an encoder and decoder $f : \mathcal{X}^n \to \{1, ..., 2^{nR}\}, \phi : \{1, ..., 2^{nR}\} \to \hat{\mathcal{X}}^n$, such that:

$$\mathbb{E}\left[d\left(X^n, \phi\left(f\left(X^n\right)\right)\right)\right] \leq D + \delta$$

Definition 4.4 ($R(D)$)**.** $R(D)$ is the minimum of all R such that (R, D) is achievable.

Why do we require the minimum? The idea is that we want succinct descriptions with low distortion (we are talking about source coding, there is no channel involved). There is a theorem equivalent to the channel coding theorem but for rate-distortion:

Theorem 4.1 (Rate distortion theorem (preview))**.** *Given a distortion function d, a maximum distortion D, input and output variables $X, \hat{X}$ jointly distributed according to $P_{X,\hat{X}}$ it holds that:*

$$R^{(I)}(D) = \min_{P_{X,\hat{X}}} I\left(X; \hat{X}\right) \quad s.t. \quad \mathbb{E}_{P_{X,\hat{X}}}\left[d\left(X, \hat{X}\right)\right] \leq D$$

This is equivalent to writing:

$$R^{(I)}(D) = \min_{P_{\hat{X}|X}} \underbrace{I\left(X; \hat{X}\right)}_{P_X P_{\hat{X}|X}} \quad \text{s.t.} \quad \mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] \leq D$$

We can now look at an example on how to compute $R^{(I)}(D)$:

Example 4.2

Let's consider a Bernoulli variable $P_X \sim \text{Ber}(p)$, the Hamming distance d_H and $\mathcal{X} = \hat{\mathcal{X}} = \{0, 1\}$:

If $\mathbf{D} \geq \min\{\mathbf{p}, \mathbf{1} - \mathbf{p}\}$ we can show that $R^{(I)}(D) = 0$. First we observe that $R^{(I)}(D) \geq 0$. If $p \geq 1/2$ we can choose $P_{\hat{X}|X}$ to be deterministically 0 (in the case $p < 1/2$ deterministically 1). We can show that this achieves the constrained minimum. Let's consider the case in which $p \geq 1/2$. We consider the expected distortion:

$$\begin{aligned}
\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] &= \sum_x \sum_{\hat{x}} P_X(x) P_{\hat{X}|X=x}(\hat{x}) d(x, \hat{x}) \\
&= p \cdot 1 \cdot 0 + p \cdot 0 \cdot 1 + (1-p) \cdot 1 \cdot 1 + (1-p) \cdot 0 \cdot 0 \\
&= 1 - p \leq D
\end{aligned}$$

We can proceed analogously for the other case, and we can conclude that with this encoding/decoding scheme, we have $\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] \leq D$ for $D \geq \min\{p, 1-p\}$. As we have chosen a deterministic output, we have that $I\left(X; \hat{X}\right) = 0$ which implies that $R^{(I)}(D) = 0$.

If $\mathbf{D} < \min\{\mathbf{p}, \mathbf{1} - \mathbf{p}\}$. First we not that we can express the distortion as $d = \mathbb{E}_{P_X P_{\hat{X}|X}}\left[X \oplus \hat{X}\right]$. We can now write:

$$\begin{aligned}
I\left(X; \hat{X}\right) &= H(X) - H\left(X|\hat{X}\right) \\
&= H(X) - H\left(X \oplus \hat{X}|\hat{X}\right) && \left(P_{X \oplus \hat{X}|\hat{X}}(X) = P_{X|\hat{X}}(X)\right) \\
&\overset{(i)}{\geq} H(X) - H\left(X \oplus \hat{X}\right) && \text{(Theorem 1.10)} \\
&\overset{(ii)}{\geq} H_b(p) - H_b(D)
\end{aligned}$$

The last inequality holds as $X \oplus \hat{X}$ is a Bernoulli variable with probability of success $\leq D$. We now have to find a probability distribution that achieves the inequality with equality:

(i) holds with equality if $X \oplus \hat{X} \perp \hat{X}$. This holds if $X = \hat{X} \oplus Z$ with $Z \perp \hat{X}$

(ii) holds with equality if $E\left[X \oplus \hat{X}\right] = D$, which holds if $Z \sim \text{Ber}\left(D\right)$

Can we find a distribution $P_{\hat{X}|X}$ for which these two conditions hold? The answer is yes, if we define $\hat{X} \sim \text{Ber}\left(q\right)$ and $Z \sim \text{Ber}\left(D\right)$, then $X = \hat{X} \oplus Z \sim \text{Ber}\left(p\right)$:

$$q = \frac{p-D}{1-2D} \qquad 1-q = \frac{1-D-p}{1-2D}$$

We have found a distribution for which the minimum is achieved. We can conclude that for a Bernoulli experiment:

$$R^{(I)}\left(D\right) = \begin{cases} H_b\left(p\right) - H_b\left(D\right) & \text{if} \quad D \le \min\{p, 1-p\} \\ 0 & \text{otherwise} \end{cases}$$

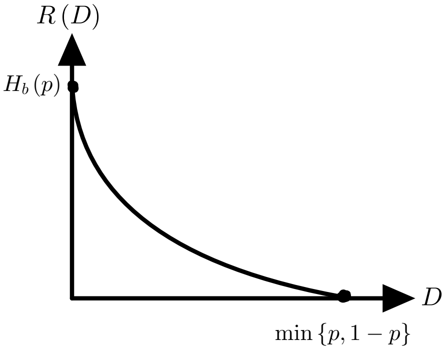


Figure 9: $R\left(D\right)$ for a Bernoulli experiment

4.2 Rate distortion theorem

We will now prove the rate-distortion theorem:

Theorem 4.2 (Rate distortion theorem (preview))**.** *Given a distortion function d, a maximum distortion D, input and output variables $X, \hat{X}$ jointly distributed according to $P_{X,\hat{X}}$ it holds that:*

$$R\left(D\right) = R^{(I)}\left(D\right) = \min_{P_{X,\hat{X}}} I\left(X;\hat{X}\right) \quad s.t. \quad \mathbb{E}_{P_{X,\hat{X}}}\left[d\left(X,\hat{X}\right)\right] \le D$$

Will prove this in two steps, as we did for the channel coding theorem:

1. **Direct**: for any $\tilde{\epsilon} > 0, \delta > 0$, if it exists $P_{\hat{X}|X}$ is such that

$$\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X,\hat{X}\right)\right] \le D$$

then it exists a code f, ϕ of rate $I\left(X; \hat{X}\right) + \tilde{\epsilon}$ such that

$$\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X, \phi\left(f\left(X\right)\right)\right)\right] \leq D + \delta$$

2. **Converse**: for any rate-R scheme f, ϕ, it holds that

$$\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] \leq D \implies R\left(D\right) \geq R^{(I)}\left(D\right)$$

4.2.1 Proof of the converse part

4.2.1.1 Properties of $R^{(I)}\left(D\right)$

To prove the converse part we need the following theorem:

Theorem 4.3 ($R^{(I)}\left(D\right)$ is monotonically decreasing, convex and continuous)**.** *$R^{(I)}\left(D\right)$ is monotonically decreasing, convex and continuous, i.e. for all $\delta > 0$ it holds that:*

$$R^{(I)}\left(D+\delta\right) \leq R^{(I)}\left(D\right)$$

Proof

We show the following: given $D^{(0)}, D^{(1)}, \lambda \in [0,1]$ and $\overline{\lambda} = 1 - \lambda$, it holds that:

$$R^{(I)}\left(\lambda D^{(0)} + \overline{\lambda} D^{(I)}\right) \leq \lambda \underbrace{R^{(I)}\left(D^{(0)}\right)}_{P^{(0)}_{\hat{X}|X}} + \overline{\lambda} \underbrace{R^{(I)}\left(D^{(I)}\right)}_{P^{(1)}_{\hat{X}|X}} \quad \text{with}$$

$$\sum_x \sum_{\hat{x}} P_X\left(x\right) P^{(0)}_{\hat{X}|X=x}\left(\hat{x}|X\right) d\left(x, \hat{x}\right) \leq D^{(0)}$$

$$\sum_x \sum_{\hat{x}} P_X\left(x\right) P^{(1)}_{\hat{X}|X=x}\left(\hat{x}|X\right) d\left(x, \hat{x}\right) \leq D^{(1)}$$

We consider $\lambda P^{(0)}_{\hat{X}|X} + \overline{\lambda} P^{(1)}_{\hat{X}|X}$. We are going to prove that this distribution satisfies the distortion constraint $\leq \lambda D^{(0)} + \overline{\lambda} D^{(1)}$. By linearity of the expectation we can write:

$$\begin{aligned}
&\mathbb{E}_{\lambda P^{(0)}_{\hat{X}|X} + \overline{\lambda} P^{(1)}_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] = \\
&\quad = \sum_x \sum_{\hat{x}} P_x\left(x\right)\left(\lambda P^{(0)}_{\hat{X}|X=x} + \overline{\lambda} P^{(1)}_{\hat{X}|X=x}\right) d\left(x, \hat{x}\right) \\
&\quad = \lambda \underbrace{\sum_x \sum_{\hat{x}} P_x\left(x\right)\left(P^{(0)}_{\hat{X}|X=x}\right) d\left(x, \hat{x}\right)}_{\leq D^{(0)}} + \overline{\lambda} \underbrace{\sum_x \sum_{\hat{x}} P_x\left(x\right)\left(P^{(1)}_{\hat{X}|X=x}\right) d\left(x, \hat{x}\right)}_{\leq D^{(1)}} \\
&\quad \leq \lambda D^{(0)} + \overline{\lambda} D^{(1)}
\end{aligned}$$

Now we claim that

$$\underbrace{I\left(X;\hat{X}\right)}_{\lambda P^{(0)}_{\hat{X}|X}+\overline{\lambda}P^{(1)}_{\hat{X}|X}} \leq \lambda R^{(I)}\left(D^{(0)}\right)+\overline{\lambda}R^{(I)}\left(D^{(1)}\right)$$

By convexity of $I\left(Q,W\right)$ w.r.t. W:

$$\underbrace{I\left(X;\hat{X}\right)}_{\lambda P^{(0)}_{\hat{X}|X}+\overline{\lambda}P^{(1)}_{\hat{X}|X}} \leq \lambda\underbrace{I\left(X;\hat{X}\right)}_{P_X P^{(0)}_{\hat{X}|X}}+\overline{\lambda}\underbrace{I\left(X;\hat{X}\right)}_{P_X P^{(1)}_{\hat{X}|X}} \leq \lambda R^{(I)}\left(D^{(0)}\right)+\overline{\lambda}R^{(I)}\left(D^{(1)}\right)$$

4.2.1.2 Proof

We can now prove the converse part, i.e. for any rate-R scheme f,ϕ, it holds that

$$\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X,\hat{X}\right)\right] \leq D \implies R\left(D\right) \geq R^{(I)}\left(D\right)$$

Proof

We consider $X^n \overset{IID}{\sim} P_X$ and an encoder decoder scheme f,ϕ. We first show that $nR \geq I\left(X^n;\hat{X}^n\right)$. The encoder f given an input sequence X^n produces a message $J \in \{1,...,2^{nR}\}$, it follows:

$$\begin{aligned}
I\left(X^n,\hat{X}^n\right) &\leq I\left(X^n;J\right) && \text{(DPI)}\\
&= H\left(J\right)-\cancelto{0}{H\left(J|X^n\right)} && \left(J=f\left(X^n\right)\right)\\
&\leq \log\left(2^{nR}\right)=nR
\end{aligned}$$

We define $D_i=\mathbb{E}\left[d\left(x_i,\hat{x}_i\right)\right]$. Since $I\left(X^n;\hat{X}^n\right)=H\left(X^n\right)-H\left(X^n|\hat{X}^n\right)$, we can write:

$$\begin{aligned}
nR &\geq H\left(X^n\right)-H\left(X^n|\hat{X}^n\right)\\
&= \sum_{i=1}^{n}H\left(X_i\right)-\sum_{i=1}^{n}H\left(X_i|X^{i-1},\hat{X}^n\right)\\
&\geq \sum_{i=1}^{n}H\left(X_i\right)-\sum_{i=1}^{n}H\left(X_i|\hat{X}_i\right) && \text{(Theorem 1.10)}\\
&= \sum_{i=1}^{n}I\left(X_i;\hat{X}_i\right)\\
&\geq \sum_{i=1}^{n}R^{(I)}\left(D_i\right) && \left(R^{(I)}\left(D_i\right)=\min I\left(X_i;\hat{X}_i\right)\right)
\end{aligned}$$

Our hypothesis is that:

$$\frac{1}{n}\sum_{i=1}^{n} D_i \leq D$$

It follows:

$$\begin{aligned} R &\geq \frac{1}{n}\sum_{i=1}^{n} R^{(I)}\left(D_i\right) \\ &\geq R^{(I)}\left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) && \text{(Convexity + Jensen's ineq.)} \\ &\geq R^{(I)}\left(D\right) && \text{(Monotonicity)} \end{aligned}$$

We have proven the converse part of the theorem.

4.2.2 Proof of the direct part

To prove this part we need three lemmas:

4.2.2.1 Lemmas

Lemma 4.1. *Given n independent $Ber(p)$ random variables, it holds that:*

$$\lim_{np\to\infty} Pr(success) \to 1$$

Proof

We can write:

$$\begin{aligned} \Pr\left(\text{failure}\right) &= 1 - \Pr\left(\text{success}\right) \\ &= \left(1-p\right)^n \\ &\leq \left(e^{-p}\right)^n && \left(\xi \geq 0 \implies 1 - \xi \leq e^{-\xi}\right) \\ &= e^{-np} \overset{np\to\infty}{\to} 0 \end{aligned}$$

Lemma 4.2. *Given a non negative function $g : \mathcal{X} \to \mathbb{R}_+$, a sequence strongly typical w.r.t. $X \sim P$, i.e. $\underline{x} \in \mathcal{T}_\epsilon^{(n)}$, it holds that:*

$$\frac{1}{n}\sum_{i=1}^{n} g\left(x_i\right) \leq \left(1+\epsilon\right)\mathbb{E}_P\left[g\left(X\right)\right]$$

Proof

The strong typical set $\mathcal{T}_\epsilon^{(n)}(P)$ is defined as:

$$\begin{aligned}\mathcal{T}_\epsilon^{(n)}(P) &= \left\{\underline{x} \in \mathcal{X}^n : \left|\frac{1}{n} N(a|\underline{x}) - P(a)\right| \leq \epsilon P(a)\right\} \\ \implies \frac{1}{n} N(a|\underline{x}) &\leq (1+\epsilon) P(a)\end{aligned}$$

Thus, we can write:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} g(x_i) &= \frac{1}{n}\sum_{a\in\mathcal{X}} N(a|\underline{x})\, g(a) \\ &\leq \sum_{a\in\mathcal{X}} P(a)(1+\epsilon)\, g(a) \\ &= (1+\epsilon)\,\mathbb{E}_P[g(X)]\end{aligned}$$

Lemma 4.3. *With $0 < \epsilon' < \epsilon$ fixed, PMF $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$ for sufficiently large n, if $\underline{x} \in \mathcal{T}_{\epsilon'}^{(n)}(P_X)$ and $\{y_i\} \overset{IID}{\sim} P_Y$, it holds that:*

$$Pr\left((\underline{x}, \underline{y}) \in T_\epsilon^{(n)}(P_{X,Y})\right) \geq 2^{-n(I(X;Y)+4\delta_{X,Y})}$$

where $I(X;Y)$ is computer w.r.t. $P_{X,Y}$ and $\delta_{X,Y} \overset{\epsilon\to 0}{\to} 0$

4.2.2.2 Proof

We can now prove the direct part, i.e. that for any $\tilde{\epsilon} > 0, \delta > 0$ if it exists $P_{\hat{X}|X}$ is such that

$$\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] \leq D$$

then it exists a code f, ϕ of rate $I\left(X; \hat{X}\right) + \tilde{\epsilon}$ such that

$$\mathbb{E}_{P_X P_{\hat{X}|X}}\left[d(X, \phi(f(X)))\right] \leq D + \delta$$

Proof

We consider a random code book of 2^{nR} code-words, and denote $\hat{x}(j)$ the j^{th} code-word of this code book. Let's consider a source X. We use a strongly typical encoder $f : \mathcal{X}^n \to \{1, ..., 2^{^nR}\}$ w.r.t. $P_X P_{\hat{X}|X}$. The encoder look for $\hat{x}(j)$ such that:

$$(\underline{x}, \hat{x}(j)) \in \mathcal{T}_{\epsilon'}^{(n)}\left(P_X P_{\hat{X}|X}\right)$$

If one or more $\hat{x}(j)$ is found, the encoder encodes the code-word with the lowest index in the set among the indexes of those $\hat{x}$, otherwise it fails and maps it to 1. Based on the encoded index j, the decoder ϕ produces the j^{th} sequence from $\mathcal{T}_{\epsilon'}^{(n)}\left(P_X P_{\hat{X}|X}\right)$. We know that in case of

- **Success**: the distortion is upper bounded by:

$$\begin{aligned} d\left(\underline{x}, \phi\left(f\left(\underline{x}\right)\right)\right) &= \frac{1}{n}\sum_{i=1}^{n} d\left(x_i, \hat{x}_i\right) \\ &\leq \left(1+\epsilon\right) E_{P_X P_{\hat{X}|X}}\left[d\left(X, \hat{X}\right)\right] && \text{(Lemma 4.2)} \\ &= \left(1+\epsilon\right) D && \text{(Hypothesis)} \end{aligned}$$

- **Failure**: the distortion is upper-bounded by $d_{\max}$.

This means that the expected distortion is upper bounded by $D + \underbrace{\epsilon D + \delta' d_{\max}}_{\delta}$.

We have that the two sources or randomness are the random code book and the random sequence:

$$\Pr\left(\text{success}\right) = \sum_{\mathcal{C}} P\left(\mathcal{C}\right)\Pr\left(\text{success}|\mathcal{C}\right) = \sum_{\underline{x}\in\mathcal{X}^n} P\left(\underline{x}\right)\Pr\left(\text{success}|\underline{x}\right)$$

We can write:

$$\begin{aligned} &\sum_{\underline{x}\in\mathcal{X}^n} P\left(\underline{x}\right)\Pr\left(\text{success}|\underline{x}\right) = \\ &\quad= \sum_{\underline{x}\in\mathcal{T}_{\epsilon}^{(n)}} P\left(\underline{x}\right)\Pr\left(\text{success}|\underline{x}\right) + \sum_{\underline{x}\notin\mathcal{T}_{\epsilon}^{(n)}} P\left(\underline{x}\right)\Pr\left(\text{success}|\underline{x}\right) \\ &\quad\geq \sum_{\underline{x}\in\mathcal{T}_{\epsilon}^{(n)}} P\left(\underline{x}\right)\Pr\left(\text{success}|\underline{x}\right) \end{aligned}$$

Let's consider a ϵ' such that $\epsilon' < \epsilon$. By strong AEP, we know that:

$$\lim_{n\to\infty} \Pr\left(\underline{x} \in \mathcal{T}_{\epsilon'}^{(n)}\left(P_X P_{\hat{X}|X}\right)\right) \to 1$$

We note that we are in a case where the code-words are drawn independently. We have 2^{nR} tries to find a j.t. $\hat{x}$. Together with **Lemma 4.3**, this means that we have a Bernoulli experiment with an expected number of successes lower bounded by:

$$2^{nR} 2^{-n\left(I(X;Y)+4\delta_{X,\hat{X}}\right)}$$

This tends to infinity as long as $R > I\left(X;Y\right) + \underbrace{4\delta_{X,\hat{X}}}_{\tilde{\epsilon}}$. From **Lemma 4.1** we conclude that:

$$\sum_{\underline{x}\in\mathcal{T}_\epsilon^{(n)}} P\left(\underline{x}\right)\Pr\left(\text{success}|\underline{x}\right)\to 1$$
$$\implies \Pr\left(\text{success}\right)\to 1$$

We have found that for $R > I\left(X;Y\right)+\tilde{\epsilon}$, we can find a code book which achieves expected distortion upper bounded by $D+\delta$.

5 Multi-terminal information theory

We now consider the case in which instead of one source, we have two, as illustrated in Fig. 10:

$$\left(X_1,Y_1,...,\left(X_n,Y_n\right)\right)\overset{IID}{\sim}P_{X,Y}$$

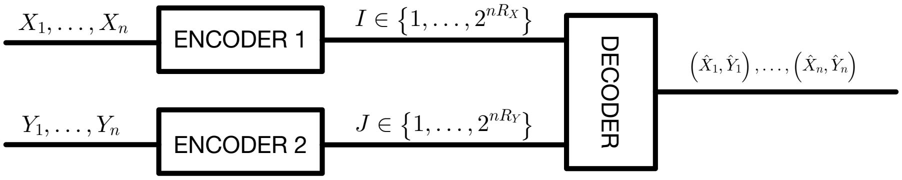

$X_1,\ldots,X_n$
ENCODER 1
$I\in\left\{1,\ldots,2^{nR_X}\right\}$
$Y_1,\ldots,Y_n$
ENCODER 2
$J\in\left\{1,\ldots,2^{nR_Y}\right\}$
DECODER
$\left(\hat{X}_1,\hat{Y}_1\right),\ldots,\left(\hat{X}_n,\hat{Y}_n\right)$

Figure 10: A 2-terminal channel

As previously, our objective is to make sure that:

$$\Pr\left(\left(\hat{X^n},\hat{Y^n}\right)\neq\left(X^n,Y^n\right)\right)\overset{n\to\infty}{\to}0$$

What rates R_X,R_Y make this possible?

5.1 Slepian-Wolf coding

Theorem 5.1 (Slepian-Wolf coding)**.** *Given two sources and a joint encoder, the bound for the lossless coding rates are:*

$$R_X\geq H\left(X|Y\right)$$
$$R_Y\geq H\left(Y|X\right)$$
$$R_X+R_Y\geq H\left(X,Y\right)$$

These bounds are illustrated in Fig. 11.

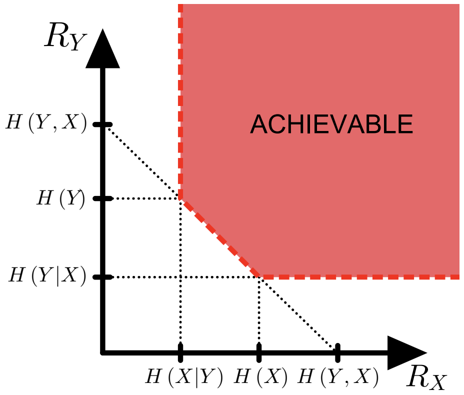


Figure 11: Rate bounds - Slepian-Wolf coding

Before proving this theorem, we state and prove the following:

Lemma 5.1. *Given a chance variable X, every rate $R > H(X)$ is achievable*

Proof

We consider $X \sim P$. We show that $R > H(X)$ is achievable by binning. We create 2^{nR} bins: our encoder $f : \mathcal{X} \to \{1, ..., 2^{nR}\}$ chooses the bin uniformly at random, and passes the address of the bin to the decoder ϕ. The decoder looks for a typical sequence w.r.t. P in the given bins and returns if it is found, otherwise it fails. We now can look at the error probability:

$$\Pr(\text{error}) = \sum_{\underline{x} \in \mathcal{X}^n} P(X^n = \underline{x}) \Pr(\text{error}|X^n = \underline{x})$$

We note that if $\underline{x} \notin \mathcal{A}_\epsilon^{(n)}(P)$ then $\Pr(\text{error}|X^n = \underline{x}) = 1$. Instead, if $\underline{x} \in \mathcal{A}_\epsilon^{(n)}(P)$ then we have an error if it exists $\underline{x}'$ such that $\underline{x}' \in \mathcal{A}_\epsilon^{(n)}(P)$ and $\underline{x}$ and $\underline{x}'$ are mapped to the same bin. We call this event E, we can write:

$$\begin{aligned} \Pr(E) &\leq \sum_{\substack{\underline{x}' \neq \underline{x} \\ \underline{x}' \in \mathcal{A}_\epsilon^{(n)}(P)}} \underbrace{\Pr(f(\underline{x}') = f(\underline{x}))}_{2^{-nR}} && \text{(Union bound)} \\ &= 2^{-nR}\left(\left|\mathcal{A}_\epsilon^{(n)}(P)\right| - 1\right) \\ &\leq 2^{-nR} 2^{n(H(X)+\epsilon)} && \text{(Lemma 2.1)} \end{aligned}$$

Thus, we can write:

$$\begin{aligned}\Pr\left(\text{error}\right) &= \sum_{X^n\in\mathcal{X}^n} P\left(X^n=\underline{x}\right)\Pr\left(\text{error}|X^n=\underline{x}\right) = \\ &= \sum_{\underline{x}'\notin\mathcal{A}_\epsilon^{(n)}(P)} P\left(X^n=\underline{x}\right)\underbrace{\Pr\left(\text{error}|X^n=\underline{x}\right)}_{=1} \\ &\quad+ \sum_{\underline{x}'\in\mathcal{A}_\epsilon^{(n)}(P)} P\left(X^n=\underline{x}\right)\underbrace{\Pr\left(\text{error}|X^n=\underline{x}\right)}_{\leq 2^{-nR}2^{n(H(X)+\epsilon)}}\end{aligned}$$

By weak AEP, we know that if $n\to\infty$ then $\Pr\left(\underline{x}\notin\mathcal{A}_\epsilon^{(n)}(P)\right)\to 0$, meaning that:

$$\Pr\left(\text{error}\right)\leq 2^{-nR}2^{n(H(X)+\epsilon)}$$

For $n\to\infty$, we notice that the RHS of the inequality goes to 0 for $R>H\left(x\right)+\epsilon$, meaning that every rate $R>H\left(X\right)$ is achievable.

We also need the following lemma:

Lemma 5.2. *Having fixed* $\underline{y}\in\mathcal{Y}^n$*, it holds that:*

$$\left|\left\{\underline{\xi}\in\mathcal{X}^n:\left(\underline{\xi},\underline{y}\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y})\right\}\right|\leq 2^{n(H(X|Y)+2\epsilon)}$$

Proof

If we consider $\left(\underline{\xi},\underline{y}\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y})$, we can write:

$$\begin{aligned}\Pr\left(X^n=\underline{\xi}|Y^n=\underline{y}\right) &= \frac{\Pr\left((X^n,Y^n)=\left(\underline{\xi},\underline{y}\right)\right)}{\Pr\left(Y^n=\underline{y}\right)} \\ &\geq \frac{2^{-n(H(X,Y)+\epsilon)}}{2^{-n(H(Y)-\epsilon)}} \qquad \left(\text{Def. } \mathcal{A}_\epsilon^{(n)}\left(P_{X,Y}\right)\right) \\ &= 2^{-n(H(X,Y)-H(Y)+2\epsilon)} \\ &= 2^{-n(H(X|Y)+2\epsilon)}\end{aligned}$$

Now we can write:

$$\begin{aligned}1 &\geq \sum_{\substack{\underline{\xi}\in\mathcal{X}^n \\ \left(\underline{\xi},\underline{y}\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y})}} \Pr\left(X^n=\underline{x}|Y^n=\underline{y}\right) \\ &\geq \sum_{(\dots)} 2^{-n(H(X|Y)+2\epsilon)} \\ &= 2^{-n(H(X|Y)+2\epsilon)}\left|\left\{\underline{\xi}\in\mathcal{X}^n:\left(\underline{\xi},\underline{y}\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y})\right\}\right|\end{aligned}$$

Which implies:

$$\left|\left\{\underline{\xi} \in \mathcal{X}^n : \left(\underline{\xi}, \underline{y}\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})\right\}\right| \leq 2^{n(H(X|Y)+2\epsilon)}$$

5.1.1 Proof of the direct part

We can now prove the direct part of the Slepian-Wolf coding theorem, i.e. that given two sources and a joint encoder such that:

$$\begin{aligned} &R_X \geq H\left(X|Y\right) \\ &R_Y \geq H\left(Y|X\right) \\ &R_X + R_Y \geq H\left(H, Y\right) \end{aligned}$$

then (R_X, R_Y) is achievable.

Proof

To prove this, we use a similar strategy as in the proof of Lemma 5.1. In this case, the encoder bins the sequences X^n, Y^n independently and sends the addresses (i, j) of the bins to the decoder. The decoder ϕ looks for jointly typical sequences w.r.t. $P_{X,Y}$ in the bins (i, j) and returns if it is found, otherwise it fails. We now try to identify the source of errors. First, we note that

$$\left(\underline{x}, \underline{y}\right) \notin \mathcal{A}_\epsilon^{(n)}(P_{X,Y}) \implies \Pr\left(\text{error}|X^n = \underline{x}, Y^n = \underline{y}\right) = 1$$

Instead, if $\left(\underline{x}, \underline{y}\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y})$, we have an error if:

$\epsilon_{1,2}$. It exist $\underline{x}', \underline{y}'$ such that $\underline{x}' \neq \underline{x}, \underline{y}' \neq \underline{y}$ and they have the same description:

$$\begin{aligned} &\exists \underline{x}' \neq \underline{x}, \underline{y}' \neq \underline{y} \quad \text{s.t.} \\ &\quad \left(\underline{x}', \underline{y}'\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y}), f_X\left(\underline{x}'\right) = f_X\left(\underline{x}\right), f_Y\left(\underline{y}'\right) = f_Y\left(\underline{y}\right) \end{aligned}$$

ϵ_1. It exists $\underline{x}'$ such that $\underline{x}' \neq \underline{x}$ and they have the same description:

$$\exists \underline{x}' \neq \underline{x} \quad \text{s.t.} \quad \left(\underline{x}', \underline{y}\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y}), f_X\left(\underline{x}'\right) = f_X\left(\underline{x}\right)$$

ϵ_2. It exists $\underline{y}'$ such that $\underline{y}' \neq \underline{y}$ and they have the same description:

$$\exists \underline{y}' \neq \underline{y} \quad \text{s.t.} \quad \left(\underline{x}, \underline{y}'\right) \in \mathcal{A}_\epsilon^{(n)}(P_{X,Y}), f_Y\left(\underline{y}'\right) = f_Y\left(\underline{y}\right)$$

$\boldsymbol{\epsilon_{1,2}}$. We can write:

$$\begin{aligned}
\Pr\left(\epsilon_{1,2}\right) &\leq \sum_{\substack{\left(\underline{x}',\underline{y}'\right):\\ \underline{x}'\neq\underline{x},\underline{y}'\neq\underline{y},\\ \left(\underline{x}',\underline{y}'\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y})}} \Pr\left(\underbrace{f_X\left(\underline{x}'\right)=f_X\left(\underline{x}\right)}_{2^{-nR_X}}\wedge\underbrace{f_Y\left(\underline{y}'\right)=f_Y\left(\underline{y}\right)}_{2^{-nR_Y}}\right)\\
&=\sum_{(\ldots)}2^{-n(R_X+R_Y)}\\
&=2^{-n(R_X+R_Y)}\left|\mathcal{A}_\epsilon^{(n)}(P_{X,Y})\right|\\
&\leq 2^{-n(R_X+R_Y)}2^{n(H(X,Y)+\epsilon)} && \text{(Lemma 2.1)}
\end{aligned}$$

which implies that $\epsilon_{1,2}\to 0$ if $R_X+R_Y>H\left(X,Y\right)+\epsilon$.

$\boldsymbol{\epsilon_1}$. We can write:

$$\begin{aligned}
\Pr\left(\epsilon_1\right) &\leq \sum_{\substack{\underline{x}'\neq\underline{x}'\\ \left(\underline{x}',\underline{y}\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y})}} \underbrace{\Pr\left(f_X\left(\underline{x}'\right)=f_X\left(\underline{x}\right)\right)}_{2^{-nR_X}}\\
&=2^{-nR_X}\left|\left\{\underline{x}':\left(\underline{x}',\underline{y}\right)\in\mathcal{A}_\epsilon^{(n)}(P_{X,Y}),\underline{x}'\neq\underline{x}\right\}\right|\\
&\leq 2^{-nR_X}2^{n(H(X|Y)+2\epsilon)} && \text{(Lemma 5.2)}
\end{aligned}$$

which implies that $\epsilon_1\to 0$ if $R_X>H\left(X|Y\right)+2\epsilon$.

The proof is analogous for ϵ_2. By the union bound, we conclude that $\Pr\left(\text{error}\right)\to 0$, and we have proven the direct part of the Slepian wolf coding.

5.1.2 Proof of the converse part

We can now prove the converse part of the Slepian-Wolf coding theorem, i.e. that given two sources and a joint encoder, $\left(R_X,R_Y\right)$ is achievable if

$$\begin{aligned}
R_X &\geq H\left(X|Y\right)\\
R_Y &\geq H\left(Y|X\right)\\
R_X+R_Y &\geq H\left(H,Y\right)
\end{aligned}$$

Proof

$\boldsymbol{\epsilon_1}$ Suppose we have a scheme that drives the probability of error to 0. We call I,J the output variables of the encoder. First, we note that because of Fano's inequality for sequences we can write:

$$H(X^n|Y^n, I, J) \leq 1 + p_e^{(n)} n \log(|\mathcal{X}|) = n \underbrace{\left(p_e^{(n)} \log(|\mathcal{X}|) + \frac{1}{n}\right)}_{\epsilon_n}$$

We note that $\epsilon_n \to 0$ as $n \to \infty$. It follows:

$$\begin{aligned}
nR_X &\geq H(I) \\
&\geq H(I|Y^n) \\
&= I(X^n; I|Y^n) + \cancelto{0}{H(I|X^n, Y^n)} && (I = f_X(X^n)) \\
&= H(X^n|Y^n) - H(X^n|I, Y^n) && \text{(chain rule)} \\
&\geq H(X^n|Y^n) - H(X^n|I, J, Y^n) && \text{(Theorem 1.10)} \\
&\geq H(X^n|Y^n) - n\epsilon_n && \text{(Fano's for seq.)} \\
&= nH(X|Y) - n\epsilon_n
\end{aligned}$$

The last equality holds because the sources are memoryless, meaning that $(X_1, Y_1), ..., (X_n, Y_n) \overset{IID}{\sim} P_{X,Y}$. If we divide by n, we obtain $R_X \geq H(X|Y) - \epsilon_n \overset{n\to\infty}{\to} H(X|Y)$. The proof is analogous for ϵ_2.

$\boldsymbol{\epsilon_{1,2}}$ We can treat (X, Y) as one chance variable Z. From part 1 of the source coding theorem, we know that if the compression is $R_Z \geq H(Z) + \epsilon$ then the probability of error is driven to 0:

$$R_Z = R_X + R_Y \geq H(X, Y) + \epsilon = H(Z) + \epsilon$$

This is a necessary condition, as stated in part 2 of the theorem.

We have proven that errors $\epsilon_1, \epsilon_2, \epsilon_{1,2}$ are driven to 0 only if the rates are bounded as stated in the Slepian-Wolf coding theorem.

Appendix A Probability

Variance $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$

Variance of sum $\mathrm{Var}(\alpha X + \beta Y) = \alpha^2 \mathrm{Var}(X) + \beta^2 \mathrm{Var}(Y) - 2\alpha\beta \mathrm{Cov}(X, Y)$

Expected value of product If $X_1, ..., X_n \overset{IID}{\sim} P \implies \mathbb{E}\left[\prod X_i\right] = \prod \mathbb{E}[X_i]$

Conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$

Law of total probability $P(A) = \sum_{i=1}^{n} P(A|B_i) P(B_i)$

Bayes theorem $P(A|B) = \frac{P(B|A)P(B)}{P(A)}$

Probability of union $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Union bound $\Pr\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \Pr(E_i)$

Bernoulli r.v. $X \sim \mathrm{Ber}(p)$: $P(1) = p, P(0) = 1 - p$. $E[X] = p$, $\mathrm{Var}(X) = p(1-p)$

Multiple Bernoulli exp. $X_1, ..., X_n \overset{IID}{\to} \mathrm{Ber}(p), Z = \sum_{i=1}^{n} X_i$, the probability of z successes is given by $P(Z = z) = \binom{n}{z} p^z (1-p)^{n-z}$

Markov inequality $\Pr[X \geq \delta] \leq \frac{\mathbb{E}[X]}{\delta}$

Chebyshev's inequality $\Pr[|Y - \mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2}$

Weak law of large numbers $\Pr\left[|\overline{Z}_n - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}$ (holds also for functions of random variables)