Shannon Theory

1. Perfect Secrecy

We define three random variables:

$P = (\mathcal{P}, p_P)$ – plaintexts

$C = (\mathcal{C}, p_C)$ - ciphertexts

$K = (\mathcal{K}, p_K)$ -keys

We assume that P and K are independent random variables.

This means that the choice of the key does not depend on the plaintext.

Denote by $\mathcal{C}(k)$ the set of all ciphertexts obtained by enciphering plaintetxts with the key k:

$$\mathcal{C}(k) = \{E_k(x) \mid x \in \mathcal{P}\}.$$

Clearly the distribution of C is determined by the distributions of P and K. We have

$$(*) \qquad p_C(C = c) = \sum_{k \in \mathcal{K}} p_K(K = k) p_P(P = D_k(c))$$

Consider the following cryptosystem:

- $\mathcal{P} = \{a, b, c, d\}$
- $\mathcal{C} = \{1, 2, 3, 4\}$
- $\mathcal{K} = \{k_1, k_2, k_3\}$

The probability distributions of P and K are given by:

$$p_P(P = a) = \frac{1}{4}, p_P(P = b) = \frac{3}{10}, p_P(P = c) = \frac{3}{20}, p_P(D = d) = \frac{3}{10}$$

$$p_K(K = k_1) = \frac{1}{4}, p_K(K = k_2) = \frac{1}{2}, p_K(K = k_3) = \frac{1}{4}$$

Further we shall write $p(a), p(b), \ldots, p(k_1), p(k_2), \ldots$.

The enciphering and deciphering transformations are given by the following table:

	a	b	c	d
k_1	3	4	2	1
k_2	3	1	4	2
k_3	4	3	1	2

Using (*) we can compute the probabilities $p_C(.)$.

$$
\begin{aligned}
p_C(1) &= p_K(k_1)p_P(d) + p_K(k_2)p_P(b) + p_K(k_3)p_P(c) \\
&= \frac{1}{4}\cdot\frac{3}{10} + \frac{1}{2}\cdot\frac{3}{10} + \frac{1}{4}\cdot\frac{3}{20} \\
&= \frac{21}{80} = 0.2625 \\
p_C(2) &= p_K(k_1)p_P(c) + p_K(k_2)p_P(d) + p_K(k_3)p_P(d) \\
&= 0.2625 \\
p_C(3) &= p_K(k_1)p_P(a) + p_K(k_2)p_P(a) + p_K(k_3)p_P(b) \\
&= 0.2625 \\
p_C(4) &= p_K(k_1)p_P(a) + p_K(k_2)p_P(c) + p_K(k_3)p_P(a) \\
&= 0.2125
\end{aligned}
$$

Now given a ciphertext $y \in \mathcal{C}$ and a plaintext $x \in \mathcal{P}$, we can compute the conditional probability $p_{C|P}(y|x)$, i.e. the probability to receive the ciphertext y provided x was enciphered. This probability is obtained from

$$p_{C|P}(y|x) = \sum_{k:x=D_k(y)} p_K(k).$$

Here the sum is over all keys that decipher y to x.

Thus we obtain the following conditional probabilities:

$p_{C\|P}(1\|a) = 0$	$p_{C\|P}(1\|b) = 0.5$
$p_{C\|P}(2\|a) = 0$	$p_{C\|P}(2\|b) = 0$
$p_{C\|P}(3\|a) = 0.75$	$p_{C\|P}(3\|b) = 0.25$
$p_{C\|P}(4\|a) = 0.25$	$p_{C\|P}(4\|b) = 0.25$
$p_{C\|P}(1\|c) = 0.25$	$p_{C\|P}(1\|d) = 0.25$
$p_{C\|P}(2\|c) = 0.25$	$p_{C\|P}(2\|d) = 0.75$
$p_{C\|P}(3\|c) = 0$	$p_{C\|P}(3\|d) = 0$
$p_{C\|P}(4\|c) = 0.5$	$p_{C\|P}(4\|d) = 0$

In cryptanalysis we observe y, i.e. the random variable C.

A natural question is: what is the most probable plaintext x given the received ciphertext y?

In other words: What is $p(x|y)$?

We can use Bayes' formula:

$$p_{P|C}(x|y) = \frac{p_P(x)p_{C|P}(y|x)}{p_C(y)}.$$

All the probabilities on the RHS have been already obtained.

Thus we compute:

$p_{P\|C}(a\|1) = 0$	$p_{P\|C}(a\|2) = 0$
$p_{P\|C}(b\|1) = 0.571$	$p_{P\|C}(b\|2) = 0.143$
$p_{P\|C}(c\|1) = 0.143$	$p_{P\|C}(c\|2) = 0$
$p_{P\|C}(d\|1) = 0.286$	$p_{P\|C}(d\|2) = 0.857$
$p_{P\|C}(a\|3) = 0.714$	$p_{P\|C}(a\|4) = 0.294$
$p_{P\|C}(b\|3) = 0$	$p_{P\|C}(b\|4) = 0.352$
$p_{P\|C}(c\|3) = 0.296$	$p_{P\|C}(c\|4) = 0.352$
$p_{P\|C}(d\|3) = 0$	$p_{P\|C}(d\|4) = 0$

Thus observing C we can get some information on P. For instance:

- If the received text is 1 then it is impossible that the plaintext is a.
- In this case the most probable plaintext is b.
- Given the received ciphertext 2, the most probable plaintext is d.
- In this case plaintexts a and c are impossible and so on.

The ideal situation is when the opponent learns nothing about the plaintext.

Dfn. A cryptosystem is said to be perfect or perfectly secure or to possess perfect secrecy if for every $x \in \mathcal{P}$ and every $y \in \mathcal{C}$ it holds

$$p_{P|C}(x|y) = p_P(x).$$

By the obvious equality

$$p_P(x)p_{C|P}(y|x) = p_C(y)p_{P|C}(x|y)$$

we have that a cryptosystem is perfect if for every $x \in \mathcal{P}$ and every $y \in \mathcal{C}$, one has

$$p_{C|P}(y|x) = p_C(y).$$

Thm 1. Let $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ be perfect cryptosystem. Then

$$|\mathcal{K}| \geq |\mathcal{C}| \geq |\mathcal{P}|.$$

Proof. We have obviously

$$|\mathcal{C}| \geq |\mathcal{P}|$$

since the cryptosystem is injective.

We assume wlog that all ciphertexts are possible, i.e. $p_C(y) > 0$ for all $y \in \mathcal{C}$.

For every message $x \in \mathcal{P}$ and every ciphertext $y \in \mathcal{C}$ it holds

$$p_{C|P}(y|x) = p_C(y) > 0.$$

Hence for every pair (x, y) there exists a key k, for which $p_K(k) > 0$ and $x = D_k(y)$.

Let us fix the plaintext $x = x_0$. Now for every ciphertext y we can find a key $k(y)$, for which

$$x_0 = D_{k(y)}(y).$$

Of course, different ciphertexts correspond to different keys. Hence

$$|\mathcal{K}| \geq |\mathcal{C}|.$$

□

Thm 2. Let $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ with

$$|\mathcal{K}| = |\mathcal{C}| = |\mathcal{P}|.$$

This cryptosystem is perfect if and only if

- all keys are equally probable, i.e.

  $$p_K(k) = \frac{1}{|\mathcal{K}|}$$

  for all $k \in \mathcal{K}$.

- for every pair $x \in \mathcal{P}$, $y \in \mathcal{C}$ there exists a single key k for which

  $$E_k(x) = y.$$

Proof. 1) Assume that the cryptosystem is perfect.

Then for each pair $x \in \mathcal{P}$, $y \in \mathcal{C}$ there exists a key $k \in \mathcal{K}$

$$p_{C|P}(y|x) = p_K(k).$$

We are going to prove the first part, i.e. that all the keys are uniformly distributed.

Let $n = |\mathcal{K}|$ and $\mathcal{P} = \{x_i \mid i = 1, \ldots, n\}$. Let us fix $y \in \mathcal{C}$ and let us number the keys in such way that $E_{k_i}(x_i) = y$:

	x_1	x_2	$\dots$	x_n
k_1	y	$*$	$\dots$	$*$
k_2	$*$	y	$\dots$	$*$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
k_n	$*$	$*$	$\dots$	y

Since the cryptosystem is perfect we have $p(x_i|y) = p(x_i)$. Hence

$$\begin{aligned} p(x_i|y) = p(x_i) &= \frac{p(y|x_i)p(x_i)}{p(y)} \\ &= \frac{p(k_i)p(x_i)}{p(y)}. \end{aligned}$$

therefore $p(y) = p(k_i)$. By this equality and by the fact that y is fixed , all keys are used with the same probability

$$p(k) = p(y) = \frac{1}{|\mathcal{K}|}.$$

2) Let us now assume that

- $|\mathcal{K}| = |\mathcal{C}| = |\mathcal{P}|$;
- all keys are used with the same probability;
- for each pair $x \in \mathcal{P}$, $y \in \mathcal{C}$ there exists a unique $k \in \mathcal{K}$ such that $E_k(x) = y$.

We are going to prove that $p(x|y) = p(x)$.

By the fact that all keys are used with the same probability, we get

$$\begin{aligned} p(y) &= \sum_k p(k)p(D_k(y)) \\ &= \frac{1}{|\mathcal{K}|} \sum_k p(D_k(y)). \end{aligned}$$

Since for each pair x, y there is a unique key transforming x into y, we have

$$\sum_k p(D_k(y)) = \sum_x p(x) = 1.$$

Therefore $p(y) = \frac{1}{|\mathcal{K}|}$. If $y = E_k(x)$ then

$$p(y|x) = p(k) = \frac{1}{|\mathcal{K}|}.$$

Now by Bayes' theorem

$$\begin{aligned} p(x|y) &= \frac{p(x)p(y|x)}{p(y)} \\ &= \frac{p(x) \cdot \frac{1}{|\mathcal{K}|}}{\frac{1}{|\mathcal{K}|}} \end{aligned}$$

□

Cor 1. The one-time pad is a perfect cryptosystem.

Shannon Theory

Let X be a random variable, defined on the set $\{x_1, x_2, \ldots, x_n\}$ via

$$Pr_X\{X = x_i\} = p_i, 1 \leq i \leq n.$$

Measure for the amount of information, obtained when the event x_i occurs is

$$J(p_i) = -\log_2 Pr\{X = x_i\} = -\log_2 p_i.$$

The base of the logarithm is immaterial; if it is 2 the information unit is *bit*.

- an event occuring with probability 1 gives no information: $J(1)$=0;
- an event occuring with probability $\frac{1}{2}$ gives 1 bit of information: $J(\frac{1}{2}) = 1$;
- generaly, $J(\frac{1}{2^k}) = k$.

Def 1. The expected value of $J\left(Pr_X\{X = x\}\right)$ is called the entropy of X and is denoted by $H(X)$, or $H(\boldsymbol{p}), \boldsymbol{P} = (p_1, \ldots, p_n)$:

$$H(\boldsymbol{p}) = H(X) = E\left(J(Pr_X\{X = x\})\right) = \sum_{i=1}^{n} p_i J(p_i) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Interpretations of $H(X)$:

- the expected amount of information obtained by observing X;

- our uncertainty about X;

- the expected number of bits needed to describe the outcome of X.

Remark 1. One can expect (based on our interpretations) that $H(X)$ has the following properties:

(P1) $H(p_1, \ldots, p_n) = H(p_1, \ldots, p_n, 0)$;

(P2) $H(p_1, \ldots, p_n) = H(p_{\sigma(1)}, \ldots, p_{\sigma(n)}), \quad \sigma \in \mathrm{Sym}(n)$;

(P3) $0 \leq H(p_1, \ldots, p_n) \leq H(\frac{1}{n}, \ldots, \frac{1}{n}) = \log_2 n$;

(P4) $H(\frac{1}{2}, \frac{1}{2}) < H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) < H(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}) < \ldots$;

(P5) $H(p_1, \ldots, p_n) = H(p_1, \ldots, p_{n-2}, p_{n-1} + p_n) + (p_{n-1} + p_n)H(\frac{p_{n-1}}{p_{n-1}+p_n}, \frac{p_n}{p_{n-1}+p_n})$;

It turns out that $J(p) = -\log_2 p$ is the only continous function satisfying (P1)–(P5) and the additional condition $J(\frac{1}{2^k}) = k$.

Remark 2. When $n = 2$ we write $h(p)$ instead of $H(p, 1-p)$. Thus

$$h(p) = -p\log_2 p - (1-p)\log_2(1-p), \quad 0 \leq p \leq 1. \tag{1}$$

We set $h(0) = h(1) = 0$ in order to make $h(p)$ continuous in the interval $[0, 1]$.

Example 1. (coin flipping)

$$Pr(\text{face}) = p, \quad Pr(\text{tail}) = 1 - p.$$

The entropy is given by 1. If the game is fair, i.e. $p = 1 - p = \frac{1}{2}$, we have $h(\frac{1}{2}) = 1$, hwich means that 1 bit is enough to represent the outcome of one flipping. Let the game be "unfair", i.e $p \neq 1 - p$ ($p = \frac{1}{4}$, say). Then we have $h(\frac{1}{4}) \approx 0.8113$. This means that we can get arbitrarily close to $h(\frac{1}{4})$ considering and encoding a sequence of coin flippings. For instance, considering two consecutive flippings, we can have

two consecutive outcomes	probability	representation
f f	1/16	111
f t	3/16	110
t f	3/16	10
t t	9/16	0

The expected length of the output representation is

$$3 \cdot \frac{1}{16} + 3 \cdot \frac{3}{16} + 2 \cdot \frac{3}{16} + 1 \cdot \frac{9}{16} = \frac{27}{16}.$$

Each sequence of 0' and 1's reperesents the outcome of two coin flippings, so on the average this scheme requires $\frac{27}{32} \approx 0.843$ bits per flipping. If we encode in one group the outcome of 3,4 etc. numbers of experiments, we can still get better approximations. It is easily checked that each sequence of 0's and 1's obtained by a concatenation of the strings 111, 110, 10, 0 can be split uniquely in strings of this type.

Example 2. The 26 letters of the English alphabet can be represented with $\log_2 26 \approx 4.70$ bits per letter if we only encode sufficiently long strings of letters into binary strings. On the other hand, the entropy of the 1-grams (using the frequency of each letter in an English text) is 4.15 bits/letter. Similar computations can be made for $n = 2, 3$ etc.:

H(1-grams)≈4.15 bits/letter;

H(2-grams)≈3.62 bits/letter;

H(3-grams)≈3.22 bits/letter.

Tests imply that the asymtotic value for $n \to \infty$ is 1.5 bits/letter.
Def 2. Let $(X_0, X_1, \ldots, X_{n-1}), n \geq 1$ be a plain text generated by the source S over the alphabet $\mathbb{Z}_q$. The *redundancy* D_n of $(X_0, \ldots, X_{n-1})$ is defined by

$$D_n = n \log_2 q - H(X_0, \ldots, X_{n-1}).$$

The average redundancy of a letter from the plaintext is defined by

$$\delta_n = D_n/n.$$

The redundancy of a given text measures by how much the text length (transformed into a binary string) exceeds the necessary length needed to carry the information of the text.

Def 3. Assume a ciphertext attack has been made against a cryptosystem with a set of keys K and a source of plaintexts S. The *unicity distance* of the plaintext is defined by

$$\min\{n \in \mathbb{N} \mid D_n \geq H(K)\}$$

- When the redundancy of the plaintext exceeds our unsertainty about the key, a cryptanalyst with enough recources should be able to determine the plaintext by the ciphertext only.

- The unicity point determines the moment when the user should change the key in order to keep the system safe.

Example 3. (continued)

For a simple substitution in an English text, we have (provided all keys are equally probable)

$$H(K) = -\sum_{i=1}^{26!} \frac{1}{26!} \log_2 \frac{1}{26!} = \log_2 26! \approx 88.382\text{bits}.$$

The redundancy for a text of length n is

$$D_n = (4.7 - 1.5)\, n = 3.2n.$$

Thus from $3.2n \geq 88.382$ we get a critical distance of approximately 28.

Mutual Information

Let X and Y be two random variables. The joint distribution of X and Y is denoted by

$$Pr_{X,Y}\{X = x, Y = y\} = p_{X,Y}(x, y).$$

The conditional probability for $X = x$ given $Y = y$ is denoted by

$$Pr_{X|Y}\{X = x \mid Y = y\} = p_{X|Y}(x|y).$$

Our uncertainty about X given $Y = y$ is defined by

$$H(X|Y = y) = -\sum_{x} p_{X|Y}(x|y) \cdot \log_2 p_{X|Y}(x|y).$$

This is interpreted as the expected amount of information obtained from the realization of X when it is known that $Y = y$ has occured.

Def 4. *Equivocation* $H(X|Y)$ or *conditional entropy* of X given Y is the expected value of $H(X|Y = y)$:

$$\begin{aligned} H(X|Y) &= \sum_y p_Y(y) H(X|Y = y) \\ &= -\sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log_2 p_{X|Y}(x|y) \\ &= -\sum_x \sum_y p_Y(y) p_{X|Y}(x|y) \log_2 p_{X|Y}(x|y) \\ &= -\sum_x \sum_y p_{X,Y}(x,y) \cdot \log_2 p_{X|Y}(x|y). \end{aligned}$$

Def 5.

$$H(X,Y) = -\sum_x \sum_y p_{X,Y}(x,y) \log_2 p_{X,Y}(x,y).$$

Thm 3.

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

Proof. We get

$$\begin{aligned}
H(X,Y) &= -\sum_x \sum_y p_{X,Y}(x,y)\log_2 p_{X,Y}(x,y) \\
&= -\sum_x \sum_y p_{X,Y}(x,y)\log_2 \big(p_Y(y)p_{X|Y}(x|y)\big) \\
&= -\sum_x \sum_y p_{X,Y}(x,y)\log_2 p_Y(y) - \sum_x \sum_y p_{X,Y}(x,y)\log_2 p_{X|Y}(x|y) \\
&= -\sum_y \log_2 p_Y(y) \sum_x p_{X,Y}(x,y) + H(X|Y) \\
&= -\sum_y \log_2 p_Y(y)\cdot p_Y(y) + H(X|Y) \\
&= H(Y) + H(X|Y).
\end{aligned}$$

$\square$

Cor 2. Let X and Y be two independent random variables then

(a) $H(X, Y) = H(X) + H(Y)$;

(b) $H(X|Y) = H(X)$;

(c) $H(Y|X) = H(Y)$.

Proof. Same as that of Theorem 3, using $p_{X,Y}(x, y) = p_X(x)p_Y(y)$. □

The amount of information $I_{X,Y}(x,y)$, which $Y=y$ gives about the realization of $X=x$ is equal to amount of information, which the occurence of x gives minus the amount provided by $X=x$ given that $Y=y$ is already known:

$$\begin{aligned} I_{X,Y}(x,y) &= -\log_2 p_X(x) + \log_2 p_{X|Y}(x|y) \\ &= -\log_2 \frac{p_X(x)}{p_{X|Y}(x|y)} \\ &= -\log_2 \frac{p_X(x)p_Y(y)}{p_{X,Y}(x,y)} \\ &= I_{Y,X}(y,x). \end{aligned}$$

Def 6. The *mutual infromation* $I(X;Y)$ of X and Y is the expected value of $I_{X,Y}(x,y)$, i.e.

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p_{X,Y}(x,y) I_{X,Y}(x,y) \\ &= -\sum_x \sum_y p_{X,Y}(x,y) \cdot \log_2 \frac{p_X(x)p_Y(y)}{p_{X,Y}(x,y)} \\ &= -\sum_x \sum_y p_{X,Y}(x,y) \cdot \log_2 p_X(x) p_{X|Y}(x|y). \end{aligned}$$

$I(X;Y)$ is interpreted as the expected amount of information, which Y gives about X (or X gives about Y).

Thm 4. $I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Proof.

$$\begin{aligned}
I(X;Y) &= -\sum_x \sum_y p_{X,Y}(x,y) \cdot \log_2 \frac{p_X(x)}{p_{X|Y}(x|y)} \\
&= -\sum_x \sum_y p_{X,Y}(x,y) \cdot \log_2 p_X(x) + \sum_x \sum_y p_{X,Y} \cdot \log_2 p_{X|Y}(x|y) \\
&= -\sum_x p_X(x) \cdot \log_2 p_X(x) + \sum_x \sum_y p_{X,Y}(x,y) \cdot \log_2 p_{X|Y}(x|y) \\
&= H(X) - H(X|Y).
\end{aligned}$$

□

Example 4. (binary symmetric channel)

A source sends $X = 0$ or $X = 1$ both with probability $\frac{1}{2}$. The receiver gets $Y = X$ with probability $1 - p$ and $Y = 1 - X$ with probability p. So,

$$p_Y(0) = p_{Y|X}(0|0)p_X(0) + p_{Y|X}(0|1)p_X(1) = (1-p)\frac{1}{2} + p \cdot \frac{1}{2} = \frac{1}{2}.$$

Similarly, $p_Y(1) = \frac{1}{2}$. We have also

$$p_{X,Y}(0,0) = p_{X,Y}(1,1) = \frac{1-p}{2}, \quad p_{X,Y}(0,1) = p_{X,Y}(1,0) = \frac{p}{2}.$$

Hence

$$
\begin{aligned}
I(X;Y) &= -2\left\{\frac{1-p}{2}\cdot\log_2\frac{\frac{1}{2}}{1-p}+\frac{p}{2}\cdot\log_2\frac{\frac{1}{2}}{p}\right\}\\
&= 1+p\log_2 p+(1-p)\log_2(1-p)\\
&= 1-H(p).
\end{aligned}
$$

The receiver gets $1-H(p)$ bits of information about X for every received symbol Y. For $p=\frac{1}{2}$ he obtains no information about X. How tocome close to $1-H(p)$ is a main problem in coding theory.

Let a cryptosystem be given. On the set of keys is given a distribution $Pr_K\{K = k\}$. We denote the plaintext and the cryptotext by

$$\begin{aligned} M^n &= (M_0, M_1, \ldots, M_{n-1}), \\ C^\nu &= (C_0, C_1, \ldots, C_{\nu-1}), \end{aligned}$$

respectively. Thus $C^\nu = E_K(M^n)$. E_K is an 1-1 mapping and therefore

$$H(M^n \mid K, C^\nu) = 0. \tag{2}$$

In other words, if we know the key and the ciphertext we can decipher. The user of the cryptosystem is interested in the following question: what amount of information about M^n is contained in C^ν.

Thm 5. $I(M^n; C^\nu \geq H(M^n) - H(K)$.

Proof. We have by (2) and by Theorem 3

$$\begin{aligned} H(K|C^\nu) &= H(K|C^\nu) + H(M^n|K, C^\nu) = H(M^n, K|C^\nu) \\ &= H(M^n|C^\nu) + H(K|M^n, C^\nu) \geq H(M^n|C^\nu, \end{aligned}$$

i.e., given a ciphertext, our uncertainty about the key is at least as big as our uncertainty about the plaintext. Therefore,

$$H(M^n|C^\nu) \leq H(K|C^\nu) \leq H(K),$$

whence $I(M^n; C^\nu) = H(M^n) - H(M^n|C^\nu) \geq H(M^n) - H(K)$. □

Def 7. A cryptosystem is said to be *unconditionally secure* (or to have a *perfect security*) if for any n and ν $I(M^n;C^\nu)=0$.

Cor 3. One necessary condition for a system to be unconditionally secure is $H(M^n)\leq H(K)$.

Note that $I(M^n;c^\nu)=0$ iff $H(M^n)=H(M^n\mid C^\nu)$.

Def 8. Denote by $M=\{m_1,\ldots,m_s\}$ the set of al plaintexts, by $C=\{c_1,\ldots,c_u\}$ – the set of all cryptotexts and by $T=\{t_1,\ldots,t_k\}$ – the set of all transformations. Let further $p(m_i)$ be the a priori probability of m_i being sent and by $p_j(m_i)$ of m_i being sent provided c_j has been received. A cryptosystem is said to be unconditionally secure if for all i and all j, $p_j(m_i)=p(m_i)$.

Remark 3. Definitons 7 and 8 are equivalent.

Assume $H(M^n) = H(M^n|C^\nu)$. Then

$$\begin{aligned}
\sum p_M(M^n)\log_2 p_M(M^n) &= \sum_{M^n}\sum_{C^\nu} p_C(C^\nu)p_{M|C}(M^n|C^\nu)\log_2 p_{M|C}(M^n|C^\nu) \\
&\leq \sum_{M^n}\sum_{C^\nu} p_C(C^\nu)p_M(M^n)\log_2 p_M(M^n) \\
&= \sum_{M^n} p_M(M^n)\log_2 p_M(M^n)\sum_{C^\nu} p_C(C^\nu) \\
&= \sum_{M^n} p_M(M^n)\log_2 p_M(M^n).
\end{aligned}$$

We have equality iff for any message of length n and any cryptotext of length ν, we have $p_{M|C}(M^n|C^\nu) = p_M(M^n)$ (i.e. the inequality above is an equality).

Assume we have Definition 8, i.e.

$p(m_i) = p_M(M^n) = p_j(m_i) = p_{M|C}(M^n|C^\nu)$. Then we have

$$\begin{aligned} H(M^n|C^\nu) &= -\sum_{M^n}\sum_{C^\nu} p_C(C^\nu) p_{M|C}(M^n|C^\nu) \log_2 p_{M|C}(M^n|C^\nu) \\ &= -\sum_{M^n}\sum_{C^\nu} p_C(C^\nu) p_M(M^n) \log_2 p_M(M^n) \\ &= -\sum_{M^n} p_M(M^n) \log_2 p_M(M^n) \\ &= H(M^n). \end{aligned}$$