Information Theory
Data compression (2)

CoCoNut, 2016

Emmanuela Orsini

All the slides are available here

`https://www.cs.bris.ac.uk/home/cseao/CoCoNut.html`

- A code C is said to be **uniquely decodable** if $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{A}^+$, such that $\mathbf{x} \neq \mathbf{y}$, we have $c^+(\mathbf{x}) \neq c^+(\mathbf{y})$.
- A code C is called **instantaneous** if, for each transmitted codeword c, c can be interpreted as a codeword as soon it is received.
  - Note that an instantaneous code is also uniquely decodable, but not the other way around.
- A code C is called **prefix code** if no codeword is a prefix of any other codeword.
  - A code is instantaneous if and only if it is a prefix code.

Let X be an m-ary source and $\{p_1, \ldots, p_m\}$ its probability distribution.

- The **expected length** of a code C for X is

$$L(C, X) = \sum_{i=1}^{m} p_i \cdot l_i,$$

- $H(X) \geq L(X, C)$
- We want u.d./prefix codes and also we want l_i as small as possible
- Being uniquely decodable/prefix strictly depends on the codewords and it seems that we cannot have many short codewords length

<table>
<tr><td rowspan="8">0</td><td rowspan="4">00</td><td rowspan="2">000</td><td>0000</td></tr>
<tr><td>0001</td></tr>
<tr><td rowspan="2">001</td><td>0010</td></tr>
<tr><td>0011</td></tr>
<tr><td rowspan="4">01</td><td rowspan="2">010</td><td>0100</td></tr>
<tr><td>0101</td></tr>
<tr><td rowspan="2">011</td><td>0110</td></tr>
<tr><td>0111</td></tr>
<tr><td rowspan="8">1</td><td rowspan="4">10</td><td rowspan="2">100</td><td>1000</td></tr>
<tr><td>1001</td></tr>
<tr><td rowspan="2">101</td><td>1010</td></tr>
<tr><td>1011</td></tr>
<tr><td rowspan="4">11</td><td rowspan="2">110</td><td>1100</td></tr>
<tr><td>1101</td></tr>
<tr><td rowspan="2">111</td><td>1110</td></tr>
<tr><td>1111</td></tr>
</table>

<table>
<tr><td rowspan="8">0</td><td rowspan="4">00</td><td rowspan="2">000</td><td>0000</td></tr>
<tr><td>0001</td></tr>
<tr><td rowspan="2">001</td><td>0010</td></tr>
<tr><td>0011</td></tr>
<tr><td rowspan="4">01</td><td rowspan="2">010</td><td>0100</td></tr>
<tr><td>0101</td></tr>
<tr><td rowspan="2">011</td><td>0110</td></tr>
<tr><td>0111</td></tr>
<tr><td rowspan="8">1</td><td rowspan="4">10</td><td rowspan="2">100</td><td>1000</td></tr>
<tr><td>1001</td></tr>
<tr><td rowspan="2">101</td><td>1010</td></tr>
<tr><td>1011</td></tr>
<tr><td rowspan="4">11</td><td rowspan="2">110</td><td>1100</td></tr>
<tr><td>1101</td></tr>
<tr><td rowspan="2">111</td><td>1110</td></tr>
<tr><td>1111</td></tr>
</table>

0	00	000	0000
			0001
		001	0010
			0011
	01	010	0100
			0101
		011	0110
			0111
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
		111	1110
			1111

<table>
<tr><td rowspan="8">0</td><td rowspan="4">00</td><td rowspan="2">000</td><td>0000</td></tr>
<tr><td>0001</td></tr>
<tr><td rowspan="2">001</td><td>0010</td></tr>
<tr><td>0011</td></tr>
<tr><td rowspan="4">01</td><td rowspan="2">010</td><td>0100</td></tr>
<tr><td>0101</td></tr>
<tr><td rowspan="2">011</td><td>0110</td></tr>
<tr><td>0111</td></tr>
<tr><td rowspan="8">1</td><td rowspan="4">10</td><td rowspan="2">100</td><td>1000</td></tr>
<tr><td>1001</td></tr>
<tr><td rowspan="2">101</td><td>1010</td></tr>
<tr><td>1011</td></tr>
<tr><td rowspan="4">11</td><td rowspan="2">110</td><td>1100</td></tr>
<tr><td>1101</td></tr>
<tr><td rowspan="2">111</td><td>1110</td></tr>
<tr><td>1111</td></tr>
</table>

0	00	000	0000
			0001
		001	0010
			0011
	01	010	0100
			0101
		011	0110
			0111
1	10	100	1000
			1001
		101	1010
			1011
	11	110	1100
			1101
		111	1110
			1111

Kraft-McMillan

1. For each uniquely decodable binary code $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$, the codeword lengths must satisfy

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

This inequality is usually called **Kraft inequality**.

Kraft-McMillan

1. For each uniquely decodable binary code $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$, the codeword lengths must satisfy

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

This inequality is usually called **Kraft inequality**.

2. Given a set of codeword lengths $\{l_1, \ldots, l_m\}$, there exits a binary prefix code with these codeword lengths if and only if $l_i, i = 1, \ldots, m$, satisfy the Kraft inequality

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

- We want to minimize the expected length code $L(C, X)$
- The entropy is a lower bound:

$$L(C, X) \geq H(X)$$

- **Optimal source codelengths**: $L(C, X)$ is minimized and is equal to $H(X)$ only if the codelengths are equal to the *Shannon information content*:

$$l_i = \log_2(1/p_i)$$

- **Source coding theorem**: For a source (random variable) X, there exists a prefix code C with expected length satisfying:

$$H(X) \leq L(C, X) \leq H(X) + 1.$$

Huffman coding algorithm

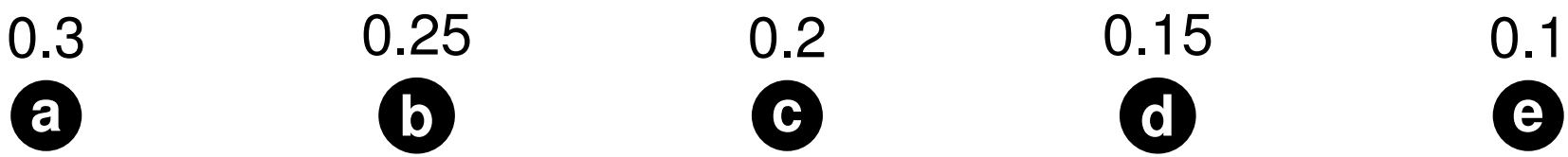
Huffman coding algorithm

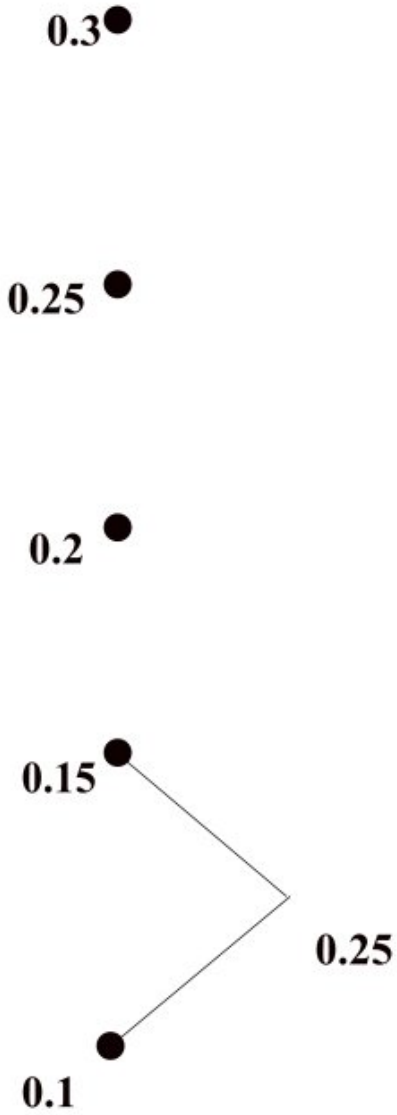
1. Take the two least probable symbols in the alphabet. These two symbols will be given the longest codewords, which will have equal length, and differ only in the last digit.

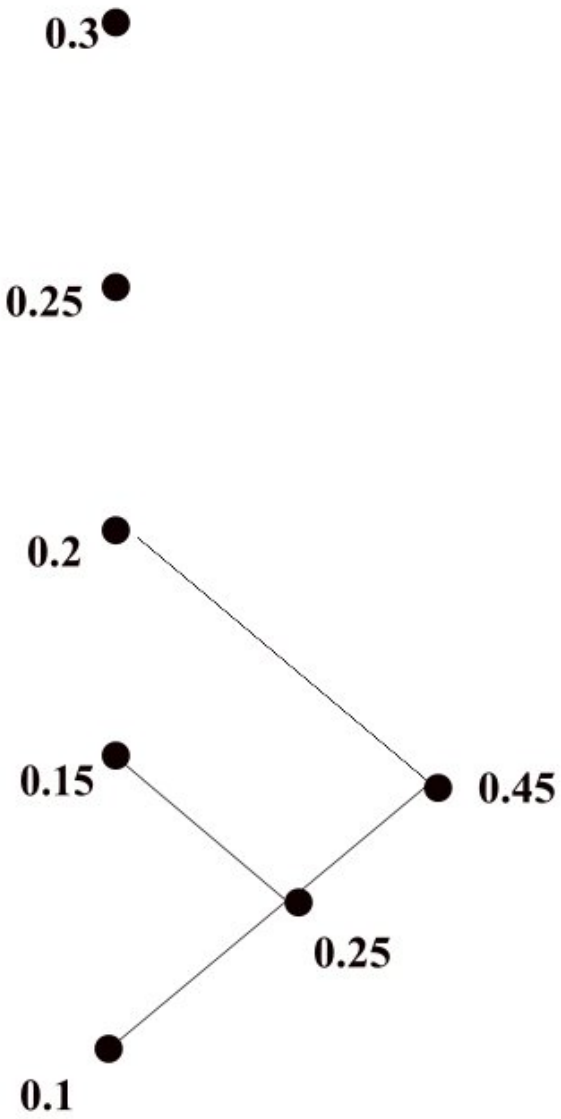
Huffman coding algorithm

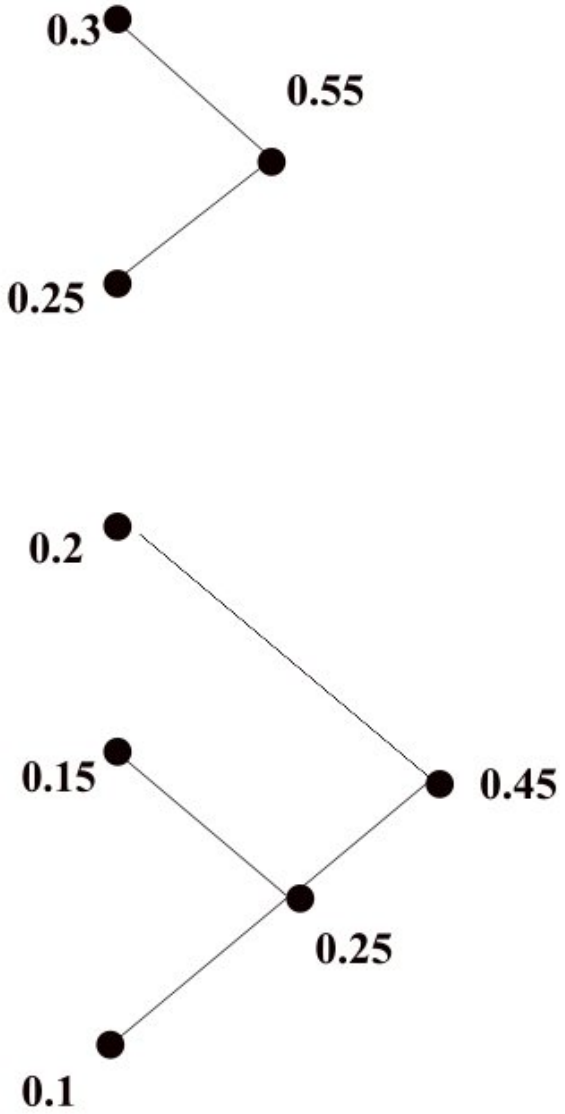
1. Take the two least probable symbols in the alphabet. These two symbols will be given the longest codewords, which will have equal length, and differ only in the last digit.
2. Combine these two symbols into a single symbol, and repeat.

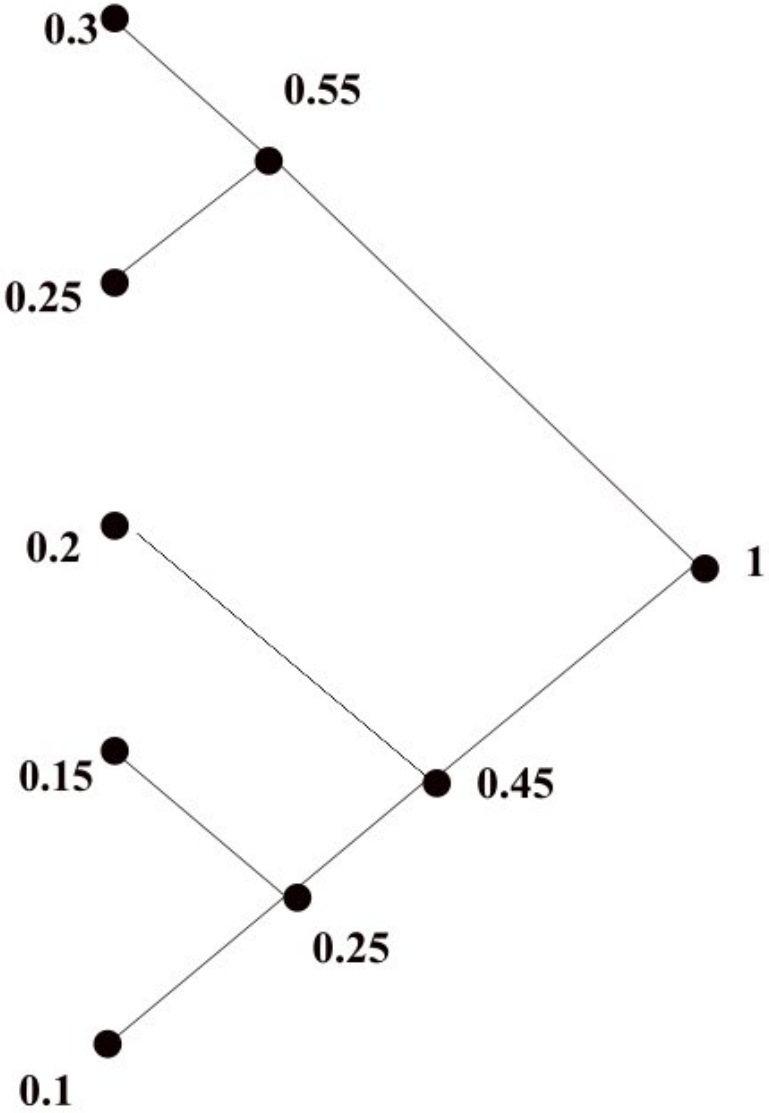
symbols	a	b	c	d	e
p_i	0.3	0.25	0.2	0.15	0.1

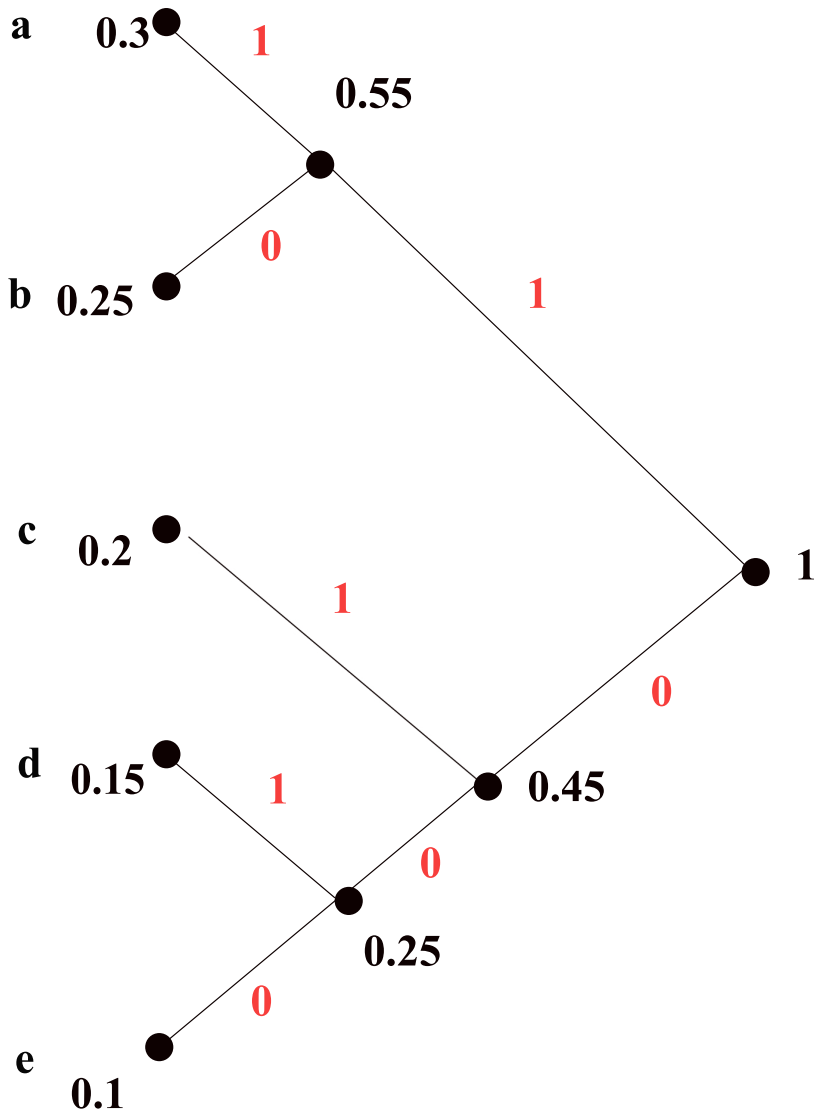



0.3
0.25
0.2
0.15
0.25
0.1


0.3
0.25
0.2
0.15
0.45
0.25
0.1


0.3
0.55
0.25
0.2
0.15
0.45
0.25
0.1


0.3
0.55
0.25
0.2
1
0.15
0.45
0.25
0.1


a
0.3
1
0.55
0
b
0.25
1
c
0.2
1
1
0
d
0.15
1
0.45
0
0.25
0
e
0.1

A loaded dice with outcome $x_i \in \{1, \ldots, 6\}$ has the following probability distribution:

$$p_1 = 1/12 \quad p_2 = 1/9 \quad p_3 = 1/18 \quad p_4 = 1/6 \quad p_5 = 1/12 \quad p_6 = 1/2$$

where $p_i = Pr(X = x_i)$.

a) Compute the entropy $H(X)$.

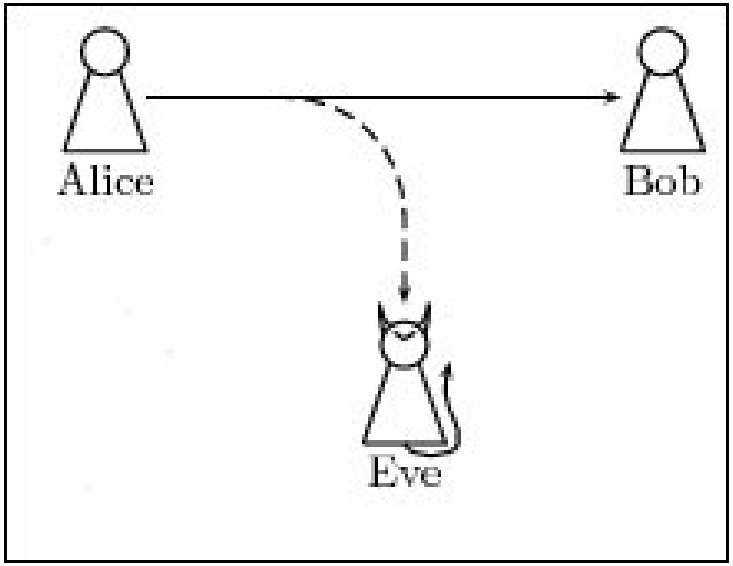
b) You are allowed to ask questions of the form

"Is X contained in the set S?"

and you will receive binary answers (i.e., yes or no).

Design the sequence of questions to ask in order to guess X with the minimum number of questions on average.

Part II

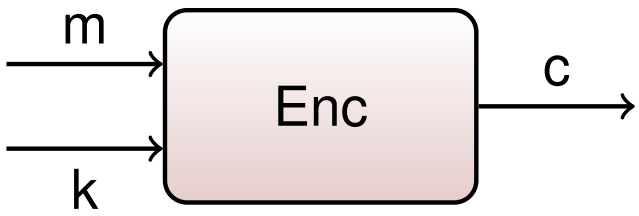


- ⋆ *Unconditional security*: secure against adv with unbounded computational power
- ⋆ All the algorithms are publicly known

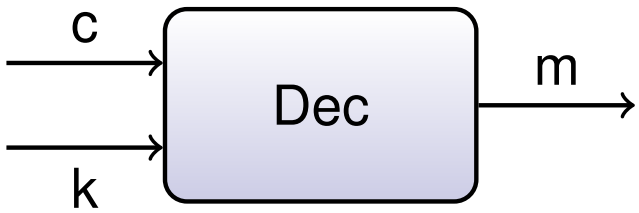
We show under what conditions **perfect secrecy** can and cannot be achieved and why it is the case.

Given a message space $\mathcal{M}$, a key space $\mathcal{K}$, and a ciphertext space $\mathcal{C}$, an *encryption scheme* is composed by three algorithms:

- A *key generation algorithm* Gen: it outputs a key k in $\mathcal{K}$;
  - It is a probabilistic algorithm that outputs a key k according to some distribution
- An *encryption algorithm* $\mathrm{Enc}_k(m)$: given a key k and a message $m \in \mathcal{M}$, it outputs a ciphertext $c \in \mathcal{C}$;



- A *decryption algorithm* $\mathrm{Dec}_k(c)$: given a key k and a ciphertext $c \in \mathcal{C}$, it outputs a message $m \in \mathcal{M}$.



- For every $k \in \mathcal{K}$ and $m \in \mathcal{M}$, it should be that

$$Dec_k(Enc_k(m)) = m.$$

- $|\mathcal{C}| \geq |\mathcal{M}|$
- Probability distribution over $\mathcal{K}$, $\mathcal{M}$, $\mathcal{C}$.
- The distribution over $\mathcal{K}$, $\mathcal{M}$ are independent.
- $\Pr[K = k]$, $\Pr[M = m]$, $\Pr[C = c]$
- K, M, C random variables over $\mathcal{K}, \mathcal{M}, \mathcal{C}$

WHEN AN ENCRYPTION SCHEME IS SECURE?

WHAT "SECURE" MEANS?

INTUITION:

- Suppose that Adv knows the probability distribution over $\mathcal{M}$
- Then Adv observes the ciphertexts: ideally these ciphertexts should have no effect on the knowledge of the adversary

$\Rightarrow$ the a POSTERIORI likelihood that some msg m was sent should be no different from the a PRIORI probability that m would be sent.

Consider the following encryption scheme with $\mathcal{M} = \{m_1, m_2, m_3\}$, s.t.

$$\Pr(M = m_1) = 1/2, \quad \Pr(M = m_2) = 1/3, \quad \Pr(M = m_3) = 1/6;$$

$\mathcal{K} = \{k_1, k_2, k_3\}$, with $\Pr(K = k_i) = 1/3, \ \forall i = 1, 2, 3$;
$\mathcal{C} = \{1, 2, 3, 4\}$, with encryption table

	m_1	m_2	m_3
k_1	1	2	3
k_2	2	3	4
k_3	3	4	1

Using Bayes' rule

$$\Pr(M = m_1|C = 1) = 3/4 \quad \Pr(M = m_2|C = 1) = 0 \quad \Pr(M = m_3|C = 1) = 1/4$$
$$\Pr(M = m_1|C = 2) = 3/5 \quad \Pr(M = m_2|C = 2) = 2/5 \quad \Pr(M = m_3|C = 2) = 0$$
$$\Pr(M = m_1|C = 3) = 1/2 \quad \Pr(M = m_2|C = 3) = 1/3 \quad \Pr(M = m_3|C = 3) = 1/6$$
$$\Pr(M = m_1|C = 4) = 0 \quad \Pr(M = m_2|C = 4) = 2/3 \quad \Pr(M = m_3|C = 4) = 1/4$$

Definition (Shannon)

An encryption scheme (Gen, Enc, Dec) over a message space $\mathcal{M}$ is **perfectly secret** if for every probability distribution over $\mathcal{M}$, every message $m \in \mathcal{M}$, and every ciphertext $c \in \mathcal{C}$, for which $\Pr[C = c] > 0$:

$$\Pr[M = m \mid C = c] = \Pr[M = m]$$

Definition (Shannon)

An encryption scheme (Gen, Enc, Dec) over a message space $\mathcal{M}$ is **perfectly secret** if for every probability distribution over $\mathcal{M}$, every message $m \in \mathcal{M}$, and every ciphertext $c \in \mathcal{C}$, for which $\Pr[C = c] > 0$:

$$\Pr[M = m \mid C = c] = \Pr[M = m]$$

Lemma

An encryption scheme (Gen, Enc, Dec) over a message space $\mathcal{M}$ is perfectly secret if and only if for every probability distribution over $\mathcal{M}$, every message $m \in \mathcal{M}$, and every ciphertext $c \in \mathcal{C}$:

$$\Pr[C = c \mid M = m] = \Pr[C = c]$$

Definition (Shannon)

An encryption scheme (Gen, Enc, Dec) over a message space $\mathcal{M}$ is **perfectly secret** if for every probability distribution over $\mathcal{M}$, every message $m \in \mathcal{M}$, and every ciphertext $c \in \mathcal{C}$, for which $\Pr[C = c] > 0$:

$$\Pr[M = m \mid C = c] = \Pr[M = m]$$

Lemma

An encryption scheme (Gen, Enc, Dec) over a message space $\mathcal{M}$ is perfectly secret if and only if for every probability distribution over $\mathcal{M}$, every message $m \in \mathcal{M}$, and every ciphertext $c \in \mathcal{C}$:

$$\Pr[C = c \mid M = m] = \Pr[C = c]$$

Equivalent formulation of previous definition

Yet another equivalent formulation:

Lemma

An encryption scheme (Gen, Enc, Dec) *over a message space* $\mathcal{M}$ *is perfectly secret if and only if for every probability distribution over* $\mathcal{M}$*, every messages* $m_1, m_2 \in \mathcal{M}$*, and every ciphertext* $c \in \mathcal{C}$*:*

$$\Pr[C = c \mid M = m_1] = \Pr[C = c \mid M = m_2]$$

It is impossible to distinguish an encryption of m_1 from an encryption of m_2, because the distributions over the ciphertext are the same.

- Let $a \oplus b$ denote the bitwise XOR

Definition (One-time pad)

Let $n \geq 1$ be an integer and $\mathcal{K} = \mathcal{M} = \mathcal{C} = \{0,1\}^n$:

- *Gen*: it outputs a uniformly random $k \in \{0,1\}^n$
- $Enc_k(m)$: given a message $m \in \{0,1\}^n$, it outputs $c = m \oplus k \in \{0,1\}^n$
- $Dec_k(c)$: it computes $m = c \oplus k \in \{0,1\}^n$

- $Dec_k(Enc_k(m)) = k \oplus k \oplus m = m$

Theorem

The one-time pad encryption scheme is perfectly secure.

- ⋆ In the one-time pad the key has to be as long as the message
- ⋆ This scheme is secure only if a key is used once

The following result is valid for any perfectly secure encryption scheme.

Theorem

Let (Gen, Enc, Dec) *be a perfectly-secret encryption scheme over a message space* $\mathcal{M}$*, and let* $\mathcal{K}$ *be the key space. Then*

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

◇ It is impossible to achieve "cheap" perfect privacy

Theorem

An encryption scheme (Gen, Enc, Dec) *over a message space* $\mathcal{M}$*, such that* $|\mathcal{M}| = |\mathcal{K}| = |\mathcal{C}|$*, is perfectly secret if and only if:*

1. *Every key* $k \in \mathcal{K}$ *is chosen with equal probability* $1/|\mathcal{K}|$ *by Gen*
2. *For every* $m \in \mathcal{M}$ *and every* $c \in \mathcal{C}$*, there exists a unique* $k \in \mathcal{K}$ *such that* $Enc_k(m)$ *outputs* c*.*

Theorem

An encryption scheme (Gen, Enc, Dec) *over a message space* $\mathcal{M}$*, such that* $|\mathcal{M}| = |\mathcal{K}| = |\mathcal{C}|$*, is perfectly secret if and only if:*

1. *Every key* $k \in \mathcal{K}$ *is chosen with equal probability* $1/|\mathcal{K}|$ *by Gen*
2. *For every* $m \in \mathcal{M}$ *and every* $c \in \mathcal{C}$*, there exists a unique* $k \in \mathcal{K}$ *such that* $Enc_k(m)$ *outputs* c*.*

It can be used for proving whether a given scheme is or is not perfectly secret.