Information and complexity

Source coding

Dimitri Petritis

Institut de recherche mathématique de Rennes
Université de Rennes 1 et CNRS (UMR 6625)

September 2022

Information and complexity

Source coding

Dimitri Petritis

Institut de recherche mathématique de Rennes
Université de Rennes 1 et CNRS (UMR 6625)

September 2022

Sources
Different classes

Definition

$(\Omega, \mathcal{F}, \mathbb{P})$ appropriate probability space.

1. A **discrete source** is an emitter of sequence $(X_n)_{n\in\mathbb{N}}$ of random variables taking values in finite alphabet $\mathbb{X}$ (the set of symbols emitted by source.)
2. Source totally determined by $((\Omega, \mathcal{F}, \mathbb{P}), \mathbb{X})$ shortened to $(\mathbb{X}, \mathbb{P})$.
3. Discrete source $(\mathbb{X}, \mathbb{P})$ is **stationary** if for all $N \in \mathbb{N}$, all $n \geq 1$, and all n-tuples $(x_1, \ldots, x_n) \in \mathbb{X}^n$,

$$\mathbb{P}(X_{N+1} = x_1, \ldots, X_{N+n} = x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

4. Discrete source $(\mathbb{X}, \mathbb{P})$ is **Markovian** if sequence $(X_n)_{n\in\mathbb{N}}$ is a Markov chain $\mathrm{MC}(\mathbb{X}, P, \cdot)$.
5. Discrete source $(\mathbb{X}, \mathbb{P})$ is **memoryless** if sequence $(X_n)_{n\in\mathbb{N}}$ is independent.
6. Memoryless discrete source $(\mathbb{X}, \mathbb{P})$ is **homogeneous** if independent sequence $(X_n)_{n\in\mathbb{N}}$ is identically distributed with $\mathbb{P}(X_n = x) = p(x)$, $x \in \mathbb{X}$, and $n \in \mathbb{N}$, where $\mathbf{p} = (p(x))_{x\in\mathbb{X}} \in \mathrm{PV}_{\mathbb{X}}$ law of X_1.

Sources
Coding

Definition

$(\mathbb{X}, \mathbb{P})$ arbitrary discrete source and $\mathbb{A}$ finite alphabet with card$\mathbb{A} \geq 2$.

- A **coding** of emitted symbols is a map (the code) $C : \mathbb{X} \to \mathbb{A}^+ := \cup_{n \geq 1} \mathbb{A}^n$.
- **Glossary of the code** the set $C(\mathbb{X}) \subset \mathbb{A}^+$.
- $|C(x)|$ the **length** of word $\alpha \in \mathbb{A}^+$ coding symbol x.
- When $\exists n$ such that for all $x \in \mathbb{X}$, $C(x) \in \mathbb{A}^n$ **fixed length code** else variable length.
- A **block code** (m-block code), with $m \geq 1$, is a code $C : \mathbb{X}^m \to \mathbb{A}^+$.
- A **code extended by concatenation** is a map $\tilde{C} : \mathbb{X}^+ \to \mathbb{A}^+$, defined for all $\xi \in \mathbb{X}^+$ by

$$\tilde{C}(\xi) := C(\xi_1) \cdots C(\xi_{|\xi|}).$$

Omit distinctive tilda $\tilde{C}$ from extension in the sequel.

$\mathbb{A} = \{0,1\}$

$\mathbb{A}^0 = \{()\} = \{\varepsilon\}$

$\mathbb{A}^1 \simeq \mathbb{A}$

$\mathbb{A}^2 = \{00, 01, 10, 11\}$

$\mathbb{A}^n = \{a_1 \cdots a_n,\ a_i \in \mathbb{A}\}$

$\mathbb{A}^* = \bigcup_{n \geq 0} \mathbb{A}^n$

$\mathbb{A}^+ = \bigcup_{n \geq 1} \mathbb{A}^n$

2-A: Examples.

$X = \{a, b, c\}$

$X^+ = \{aa, ab, ac, \ldots \quad ca\}$

$\Sigma = \{0, 1\}$

x	$C(x)$
a	0
b	10
c	11

$x_1 x_2$	$\tilde{C}(x_1 x_2) = C(x_1)C(x_2)$
aa	00
ab	010
ac	011
ba	100
bb	1010
bc	1011
ca	110
cb	1110
cc	1111

Sources
Classes of codes

Definition

Code C is

1. **non-singular** if $C : \mathbb{X} \to \mathbb{A}^+$ is injective, i.e. $x \neq x' \Rightarrow C(x) \neq C(x')$,
2. **uniquely decodable** if its extension to $\mathbb{X}^+$ is injective,
3. **instantaneous** if no word $C(x), x \in \mathbb{X}$ of the gloassary is prefix of another word of the glossary.

Families of different codes — sub-families of $\mathcal{C}$ — denoted $\mathcal{C}_{\text{inst}}$, $\mathcal{C}_{\text{ud}}$, $\mathcal{C}_{\text{ns}}$.
Obvious inclusions

$$\mathcal{C}_{\text{inst}} \subseteq \mathcal{C}_{\text{ud}} \subseteq \mathcal{C}_{\text{ns}} \subseteq \mathcal{C}.$$

Example

x	$C_1(x)$	$C_2(x)$	$C_3(x)$	$C_4(x)$
1	0	0	10	0
2	0	010	00	10
3	1	01	11	110
4	1	10	110	111

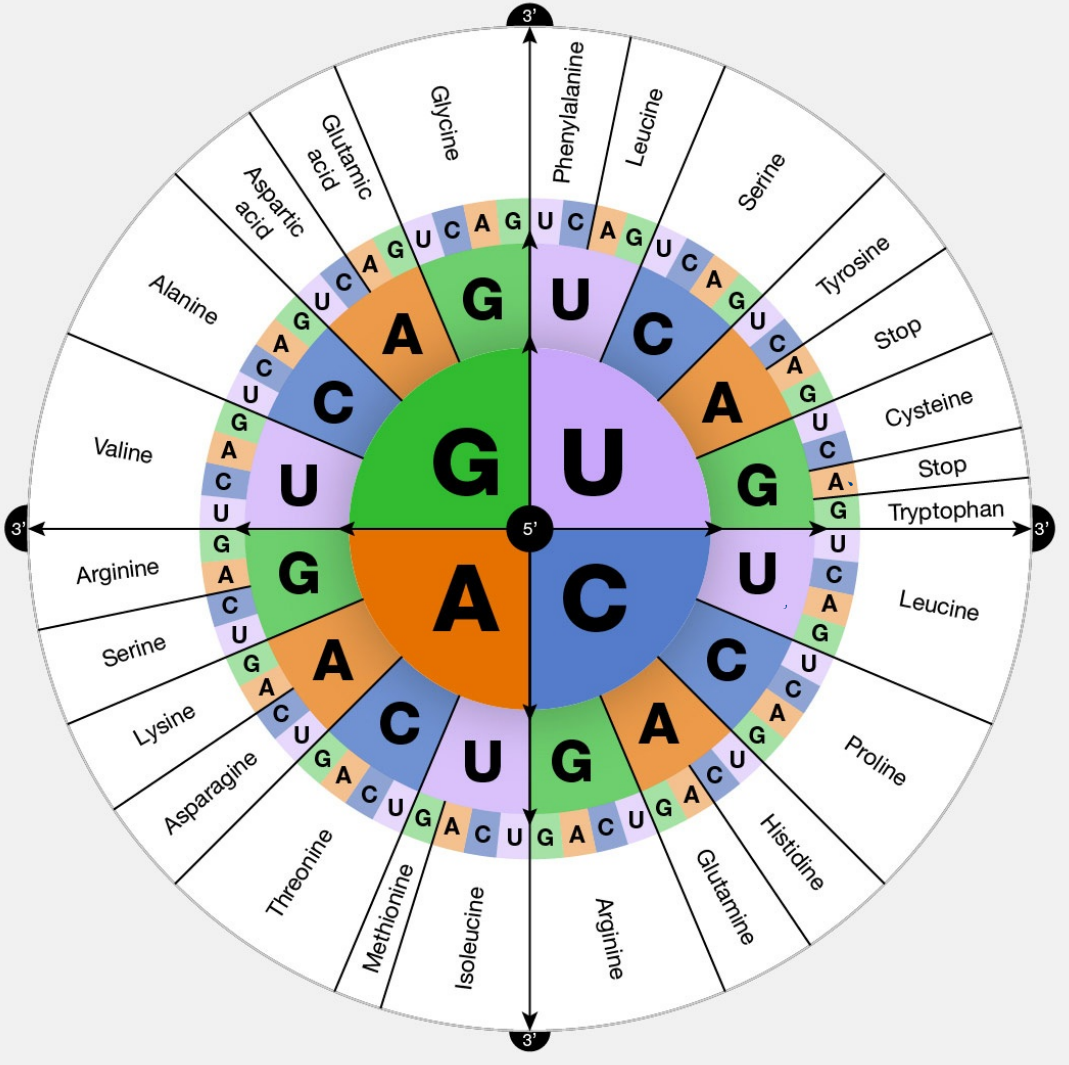
C_1 is singular since non injective, C_2 non-singular but not uniquely decodable since $C_2^{-1}(010) = \{31, 2, 14\}$, C_3 is uniquely decodable but not instantaneous since the codeword 11 is prefix of codeword 110, C_4 is instantaneous.

Codons = strings of 3 letters (bases)

{A, C, G, T}

64 codons

No	Name	3-abbr	1-abbr
1	Alanine	Ala	A
2	Arginine	Arg	R
3	Asparagine	Asn	N
4	Aspartate	Asp	D
5	Cysteine	Cys	C
6	Glutamate	Glu	E
7	Glutamine	Gln	Q
8	Glycine	Gly	G
9	Histidine	His	H
10	Isoleucine	Ile	I
11	Leucine	Leu	L
12	Lysine	Lys	K
13	Methionine	Met	M
14	Phenylalanine	Phe	F
15	Proline	Pro	P
16	Serine	Ser	S
17	Threonine	Thr	T
18	Tryptophan	Trp	W
19	Tyrosine	Tyr	Y
20	Valine	Val	V
21	Selenocysteine	Sec	U
22	Pyrrolysine	Pyl	O



U = T

4 codons → Leucine

C U V C A G

Unique decodability
Kraft's inequality

Definition

$\mathbb{A}$ finite alphabet with $A = \text{card}\mathbb{A}$ and $\ell := (\ell_i)_{i\in I}$ family of integers $\ell_i \geq 1$ for $i \in I$. Family ℓ filfils **Kraft's inequality** if

$$\sum_{i\in I} A^{-\ell_i} \leq 1.$$

Theorem (Kraft1949)

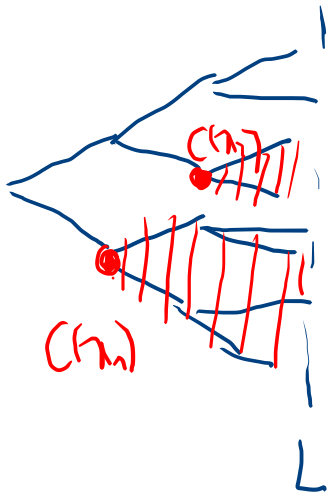
1. *If $C : \mathbb{X} \to \mathbb{A}^+$ is instantaneous, the family $(|C(x)|)_{x\in\mathbb{X}}$ of the lengths of the codewords must fulfil Kraft's inequality.*
2. *Conversely, if $(\ell_x)_{x\in\mathbb{X}}$ sequence of integers ≥ 1 fulfilling Kraft's inequality, there exists instantaneous $C : \mathbb{X} \to \mathbb{A}^+$ such that $|C(x)| = \ell_x, \forall x \in \mathbb{X}$.*

2-B: Proof of Kraft's theorem.

$(\Rightarrow)$ Words of A^+ arising as codewords

$$L = \max_l \{ l : |C(x)| = l, \ x \in X \}$$

order X $\quad x_1 \le x_2 \le \cdots$

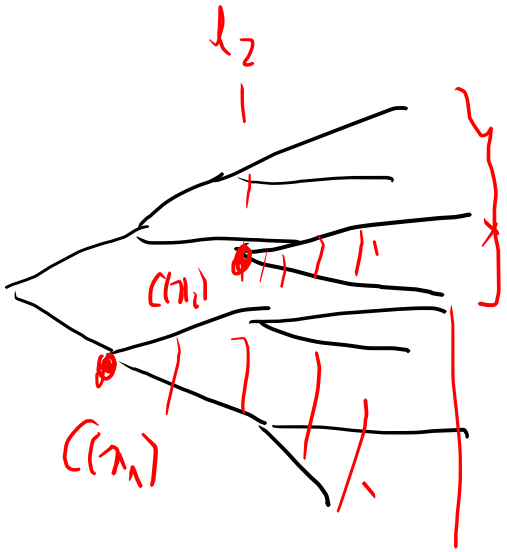


$A^{L - l_{x_1}}$ excluded

$A^{L - l_{x_2}}$ -"-

$$\sum_{i \in I} A^{L - l_{x_i}} \le A^L \quad \Rightarrow \quad \sum_{x \in X} A^{-l_{x_i}} \le 1$$

$(\Leftarrow)$ $l_1 \le l_2 \cdots$


l_2
$C(\gamma_2)$
$C(\gamma_1)$

Unique decodability
Optimal codes

- Kraft's inequality $\Leftrightarrow$ existence instantaneous code with lengths of codewords (ℓ_x).
- Nevertheless, every permutation (ℓ_x) acceptable.
- If ℓ_x not constant and underlying probability on $\mathbb{X}$ not uniform, there exists code better than the others.
- Hence Kraft's inequality, existential result not constructive.
- To get optimal code, must minimise $\mathbb{E}|C(X)| = \sum_{x\in\mathbb{X}} p(x)\ell_x$ under contraint $\sum_{x\in\mathbb{X}} A^{-\ell_x} \leq 1$.
- Use traditional method of Lagrange multipliers. Consider functional

$$J(\boldsymbol{\ell}) = \sum_x p(x)\ell_x + \lambda\left(\sum_x A^{-\ell_x} - 1\right);$$

on deriving we get

$$\forall x, \frac{\partial J}{\partial \ell_x} = p(x) - \lambda A^{-\ell_x}\log A = 0,$$

admitting solution $\boldsymbol{\ell}^* = (\ell^*(x))_{x\in\mathbb{X}}$ where $A^{-\ell_x^*} = \frac{p(x)}{\lambda \log A}$.

- But $\ell^*(x), x \in \mathbb{X}$ are not necessarily integers.

Unique decodability
Optimal codes (cont'd)

Theorem (Shannon1948)

For every instantaneous code $C : \mathbb{X} \to \mathbb{A}^+$, with $A = \mathrm{card}\mathbb{A}$, the minorisation

$$\mathbb{E}|C(X)| \geq H_A(X),$$

holds with equality iff $A^{-|C(x)|} = p(x), \forall x$.

Proof.

Introduce new probability vector $r(x) = \frac{A^{-|C(x)|}}{\sum_y A^{-|C(y)|}}$ on $\mathbb{X}$.

$$\begin{aligned}
\mathbb{E}|C(X)| - H_A(X) &= \sum_x p(x)|C(x)| + \sum_x p(x) \log_A p(x) \\
&= -\sum_x p(x) \log_A A^{-|C(x)|} + \sum_x p(x) \log_A p(x) \\
&= \sum_x p(x) \log_A \frac{p(x)}{r(x)} - \log\Big(\sum_y A^{-|C(y)|}\Big) \\
&= D(\mathbf{p}\|\mathbf{r}) + \log \frac{1}{\sum_y A^{-|C(y)|}} \\
&\geq 0 \quad \text{because} \quad D \geq 0 \text{ and } \sum_y A^{-|C(y)|} \leq 1.
\end{aligned}$$

Unique decodability
Optimal codes (cont'd)

Theorem

Let X random variable with law $\mathbf{p}$. There exists instantaneous code C, such that

$$H_A(\mathbf{p}) \leq \mathbb{E}|C(X)| < H_A(\mathbf{p}) + 1.$$

Proof.

- In general, minimisers $\ell_x^* = -\log_A p(x)$ are not integers but every interval $[-\log_A p(x), -\log_A p(x) + 1[$ contains necessarily an (unique) integer ℓ_x.
- Family $(\ell_x)_{x \in \mathbb{X}}$ fulfills Kraft's inequality because $A^{-\ell_x} \leq A^{-\ell_x^*}$, for all x.
- Hence, by Karft's theorem, exists instantaneous code C admitting ℓ as lengths of codewords.
- From $\ell_x^* \leq \ell_x \leq \ell_x^* + 1$ follows $H_A(\mathbf{p}) \leq \mathbb{E}|C(X)| < H_A(\mathbf{p}) + 1$.

$-\log p(x)$

□

Theorem (McMillan1953)

- *Every $C \in \mathcal{C}_{ud}$ fulfills $\sum_x A^{-|C(x)|} \leq 1$.*
- *Conversely, for every family of integers (ℓ_x) fulfilling Kraft, exists code $C \in \mathcal{C}_{ud}$ s.t. $|C(x)| = \ell_x, \forall x$.*

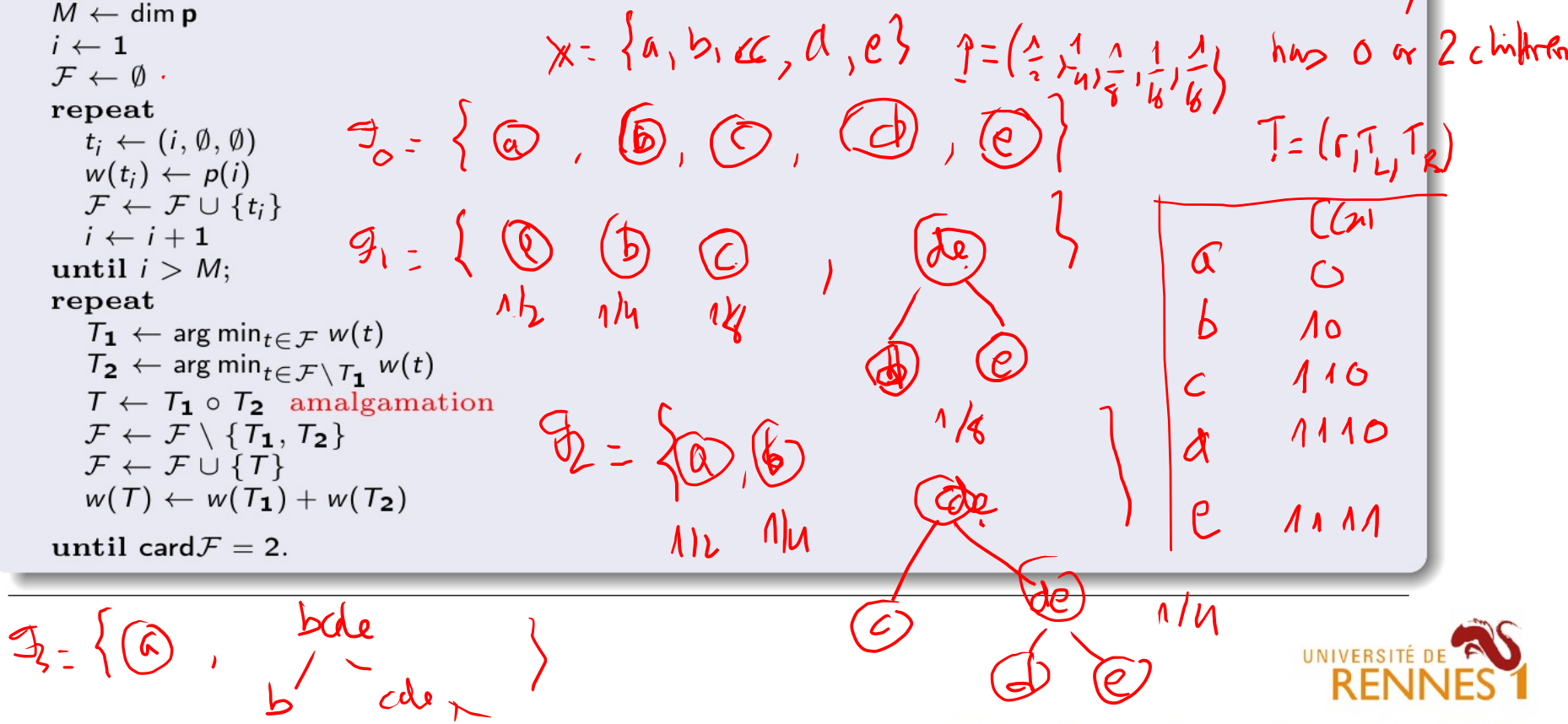
Unique decodability

Huffman's algorithmic construction of optimal codes

Algorithm

Require: Probability vector $\mathbf{p}$.
Ensure: Forest $\mathcal{F} = \{T_1, T_2\}$ composed of two binary non empty trees T_1 and T_2.

$M \leftarrow \dim \mathbf{p}$
$i \leftarrow 1$
$\mathcal{F} \leftarrow \emptyset$.
repeat
  $t_i \leftarrow (i, \emptyset, \emptyset)$
  $w(t_i) \leftarrow p(i)$
  $\mathcal{F} \leftarrow \mathcal{F} \cup \{t_i\}$
  $i \leftarrow i + 1$
until $i > M$;
repeat
  $T_1 \leftarrow \arg\min_{t \in \mathcal{F}} w(t)$
  $T_2 \leftarrow \arg\min_{t \in \mathcal{F} \setminus T_1} w(t)$
  $T \leftarrow T_1 \circ T_2$ amalgamation
  $\mathcal{F} \leftarrow \mathcal{F} \setminus \{T_1, T_2\}$
  $\mathcal{F} \leftarrow \mathcal{F} \cup \{T\}$
  $w(T) \leftarrow w(T_1) + w(T_2)$
until card$\mathcal{F} = 2$.



Unique decodability

Huffman's algorithmic construction of optimal codes (Problem)

Exercise

Let X be a $\mathbb{X}$-valued random variable with $\mathbb{X} = \{a, b, c, d, e\}$, distributed according to law with probability vector $\mathbf{p} = (1/2, 1/4, 1/8, 1/16, 1/16)$.

1. Compute $H(\mathbf{p})$.
2. Determine code $C : \mathbb{X} \to \mathbb{A}^+$ with $\mathbb{A} = \{0, 1\}$ by previous algorithm.
3. For $x \in \mathbb{X}$, denote by $|C(x)|$ length of word $C(x)$ and by $\nu_1(C(x))$ the number of bits 1 within this word. Compute $\mathbb{E}(|C(X)|)$ and $\mathbb{E}(\nu_1(C(X)))$.
4. Let $\xi \in \mathbb{X}^+$ be an arbitrary word on alphabet $\mathbb{X}$, obtained by concatenation of independent realisations of X, denote by $L = |\xi|$ its length and by the same letter C the extension obtained by concatenation of the code C, i.e. if $\xi = x_1 \cdots x_L$ then $C(\xi) = C(x_1) \cdots C(x_L)$. Let $\alpha := \alpha(\xi) = C(\xi) \in \mathbb{A}^+$ the word coding ξ. Compute the probability that a random bit in α is equal to 1, asymptotically, when $L \to \infty$.
5. Should it be possible to have predicted the last result from first principles, without any computation?

2-C: Some hints to solve exercise