Source modeling and source coding

Louis Wehenkel

Institut Montefiore, University of Liège, Belgium

ELEN060-2
Information and coding theory
March 2023

Outline

- Stochastic processes and models for information sources
- First Shannon theorem: data compression limit
- (Overview of state of the art in data compression)
- (Relations between automatic learning and data compression)

The objectives of this part of the course are the following:

- Understand the relationship between discrete information sources and stochastic processes
- Define the notions of entropy for stochastic processes (infinitely long sequences of not necessarily independent symbols)
- Talk a little more about an important class of stochastic processes (Markov processes/chains)
- How can we compress sequences of symbols in a reversible way
- What are the ultimate limits of data compression
- Review state of the art data compression algorithms
- Look a little bit at irreversible data compression (analog signals, sound, images. . .)

Do not expect to become a specialist in data compression in one day. . .

How and why can we compress information?

Simple example:

A source S which emits messages as sequences of the 3-symbol alphabet $\{a, b, c\}$.

Let's assume successive symbols are i.i.d. with: $P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{4}$

Source entropy: $H(S) = \frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{4} \log 4 = 1.5$ Shannon/ source symbol

Not that there is redundancy: $H(S) < \log 3 = 1.585$.

How to code "optimaly" using a binary code ?

1. Associate to each symbol of the source a binary word.
2. Make sure that the code is unambiguous (can decode correctly).
3. Optimality: minimal average encoded message length.

Solutions(s)

Since there are 3 symbols: not enough binary words of length 1. ⇒ Let's use words of length 2, for example

a	0	0
b	0	1
c	1	0

Average length: $P(a)\ell(a) + P(b)\ell(b) + P(c)\ell(c) = 2.$ (bits/source symbol)

Can we do better ?

Idea: let's associate one bit to a (most frequent symbol) and two bits to b and c, for example

a	0	
b	0	1
c	1	0

BADLUCK: the two source messages ba et ac result both in 010 once encoded.

There is not a unique way to decode: **code is not uniquely decodable.**

Here, the codeword of a is a **prefix** of the codeword of b.

Another idea: let's try out without prefixes:

a	0	
b	1	0
c	1	1

Does it work out ?

Indeed: code is uniquely decodable.

Average length: $P(a)\ell(a) + P(b)\ell(b) + P(c)\ell(c) =$
$1\frac{1}{2} + 2\frac{1}{4} + 2\frac{1}{4} = 1.5$ bit/symbol $= H(S)$

Conclusion.

By using short codewords for the most probable source symbols we were able to reach an average length equal to $H(S)$ (source entropy).

Questions

Does this idea work in general ?

What if $P(a) = \frac{4}{5}, P(b) = \frac{1}{10}, P(c) = \frac{1}{10}$ **?**

What if successive symbols are not independent ?

Would it be possible to do better in terms of average length ?

Relation between the "prefix-less" condition and unique decodability

How to build a good code automatiquely ?...

Before attempting to answer all these questions, we need first to see how we can model in a realistic way some real sources $\Rightarrow$ two real examples:

Telecopy (Fax):
- Sequences of 0 and 1 resulting from a black and white scanner.

Telegraph:
- Sequences of letters of the roman alphabet (say 26 letters + the space symbol).

FAX: sequences of 0 and 1

A. Most simple model: (order 0)

Let's scan a few hundred fax messages with the appropriate resolution and quantize the grey levels into two levels 0 (white), 1 (black). Let's count the proportion of 0 and 1 in the resulting file, say: $p_0 = 0.9$ et $p_1 = 0.1$

Let's model the source as a sequence of independent and identically distributed (i.i.d.) random variables according to this law.

Source entropy: $H(S) = 0.469$ Shannon/symbol

B. Slightly better model: (order 1)

Instead of assuming that successive symbols are independent, we suppose that for a sequence of symbols $s_1, s_2, \ldots, s_n$, we have

$$P(s_n|s_{n-1}, s_{n-2}, \ldots, s_1) = P(s_n|s_{n-1})$$

which we suppose is independent of n. This model is characterized by 6 numbers: intial probabilities (first symbol) and transition probabilities $p(0|0), p(1|0), p(0|1), p(1|1)$.

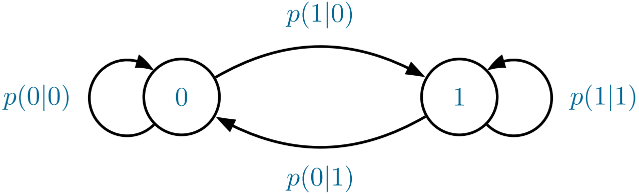
Estimates of the source model probabilities:

Initial probs.: again $p_0 = 0.9$ and $p_1 = 0.1$

And $p(0|0) = 0.93$, $p(1|0) = 0.07$, $p(0|1) = 0.63$ et $p(1|1) = 0.37$

With respect to the zero order model, we have reinforced the probability to have sequences of identical symbols $\Rightarrow$ plausible.

This model: **Two state time-invariant Markov process (chain)**



This model is sufficiently general for our needs: we will study it in “detail”

Telegraph: sequences of letters and spaces (English)

Order 0: simply statistics of letters and spaces in English text

$H(S) \leq \log 27 = 4.76$ Shannon/symbol

$H(S) \approx 4.03$ Shannon/symbol (taking into account statistics of English text)

Higher orders

Markov process of order 1: $H(S) \approx 3.42$ Shannon/symbol

Markov process of order 15: $H(S) \approx 2.1$ Shannon/symbol

Markov process of order 100: $H(S) \approx 1.3$ Shannon/symbol

Compare the rate $\frac{4.76}{1.3}$ to "gzip" compression rates.

Alternative: model English grammar... and possibilities of misspellings

NB: Markov was a linguist.

What we have just introduced intuitively are classical models of discrete time and discrete state stochastic processes.

A discrete time stochastic process is a sequence of (time indexed) random variables. Typically (as in our context) time starts at 0 and increases indefinitely by steps of 1. To each time index i corresponds a random variable $\mathcal{S}_i$, and all the random variables share the same set of possible values which are called the states of the process. Because we are mainly concerned about digital sources of information, we will mainly look at the case where the state space is finite (i.e. discrete). Such models are appropriate tools to represent sources which send message of finite but unspecified length, as it is the case in practice.

The most simple case of such a stochastic process corresponds to the situation where all the random variables are independent and identically distributed. This is also the so-called random sampling model used in statistics to represent the collection of data from a random experiment.

We will pursue our introduction to data compression by considering this example for the time being. As we will see, the general theoretical results of information theory are valide only asymptotically, i.e. in the limit when we study very long messages sent by a source. The asymptotic equipartition property is the main result which characterizes in terms of source entropy the nature of the set of long messages produced by a source. We will formulate and study this property in the case of the i.i.d. source model, but the result holds for a much larger class of stochastic processes. However, we will have to study these models and in particular provide a sound definition of source entropy.

Exercise.

Consider the zero order fax model. What is the joint entropy of the first two symbols ? What is the joint entropy of the second and third symbol ? What is the entropy of the n first symbols ?

What is entropy rate (entropy per symbol) of the first n symbols ?

Examples of English language models

Uniform distribution of letters: DM QASCJDGFOZYNX ZSDZLXIKUD...

Zero-order model: OR L RW NILI E NNSBATEI...

First-order MARKOV model: OUCTIE IN ARE AMYST TE TUSE SOBE CTUSE...

Second-order MARKOV model:

HE AREAT BEIS HEDE THAT WISHBOUT SEED DAY OFTE AND HE IS FOR THAT MINUMB LOOTS WILL AND GIRLS A DOLL WILL IS FRIECE ABOARICE STRED SAYS...

Third-order French model: DU PARUT SE NE VIENNER PERDENT LA TET...

Second-order WORD model:

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHOEVER TOLD THE PROBLEM FOR AN UNEXPECTED...

Asymptotic Equipartition Property (the clue of information theory)

Objective: study the structure of the set of long messages produced by discrete sources

Answer to the question: **What if** $P(a) = \frac{4}{5}, P(b) = \frac{1}{10}, P(c) = \frac{1}{10}$?

Possibility to reach the limit $H(S)$ for data compression.

First:

We consider a sequence of discrete r.vars. $\mathcal{X}_1, \mathcal{X}_2, \ldots$ independent and identically distributed according to the distribution $P(\mathcal{X})$ and look at the distribution of long messages.

Corresponds to the most simple non trivial source model we can immagine.

Afterwards:

We will discuss the generalization (anticipating on the sequel)

In information theory, the Asymptotic Equipartition Property is the analog of the law of large numbers.

- The (weak) law of large numbers states that for i.i.d. real random variables, $\frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i$ is closed to its expected value for large values of n.
- The AEP states that $\frac{1}{n}\log\frac{1}{P(\mathcal{X}_1,\ldots,\mathcal{X}_n)}$ is close to the entropy $H(\mathcal{X})$.

Thus the probability of observing a particular sequence $X_1,\ldots,X_n$ must be close to 2^{-nH}.

This has dramatic consequences on the structure of so-called typical messages of a source.

Reminder: the (weak) law of large numbers (wLLN)

If $\mathcal{X}_i, \forall i = 1, \dots n$ independent and of mean μ_i and finite variance σ_i, then

If $\frac{1}{n}\sum_{i=1}^{n}\mu_i \to \mu$ *and* $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \to 0$, *then* $\overline{\mathcal{X}}_n \overset{\triangle}{=} \frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i$ *is such that*

$$(\overline{\mathcal{X}}_n) \xrightarrow{P} \mu.$$

NB. The sequence $(\mathcal{X}_n)$ of r.v. is said to converge in probability towards a r.v. a (notation $(\mathcal{X}_n) \xrightarrow{P} a$), if $\forall \epsilon$ et η (arbitrarily small), $\exists n_0$ such that $n > n_0$ implies

$$P(|\mathcal{X}_n - a| > \epsilon) < \eta.$$

Particular case: (which we will use below)

The r.v. $\mathcal{X}_i$ are i.i.d. μ, σ. Then $\frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu$ and $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{\sigma^2}{n}$.

The AEP

If the $\mathcal{X}_1, \mathcal{X}_2, \ldots$ *are i.i.d. according to* $P(\mathcal{X})$, *then*

$$-\frac{1}{n}\log P(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) \xrightarrow{P} H(\mathcal{X}), \tag{1}$$

Proof: (almost) trivial application of the law of large numbers

Indeed. First notice that

$$-\frac{1}{n}\log P(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) = -\frac{1}{n}\sum_{i=1}^{n}\log P(\mathcal{X}_i), \tag{2}$$

$\Rightarrow$ sum of i.i.d. random variables. wLLN can be applied and

$$-\frac{1}{n}\sum_{i=1}^{n}\log P(\mathcal{X}_i) \xrightarrow{P} -E\{\log P(\mathcal{X})\} = H(\mathcal{X}),$$

which proves the AEP. □

Interpretation:

Let $X_1, \ldots, X_{|\mathcal{X}|}$ be the $|\mathcal{X}|$ possible values of each $\mathcal{X}_i$. For any message of length n we can decompose the right hand part of (2) as follows

$$-\frac{1}{n}\sum_{i=1}^{n}\log P(\mathcal{X}_i) = -\frac{1}{n}\sum_{j=1}^{|\mathcal{X}|} n_j \log P(X_j) = -\sum_{j=1}^{|\mathcal{X}|}\frac{n_j}{n}\log P(X_j), \tag{3}$$

where we count by n_j the number of occurrences of the j-th value of $\mathcal{X}$ in the message.

The meaning of the AEP is that almost surely the source message will be such that the right hand converges towards

$$H(\mathcal{X}) = -\sum_{j=1}^{|\mathcal{X}|} P(X_j)\log P(X_j),$$

which is a simple consequence of the well known (statistical fact) that $\frac{n_j}{n} \xrightarrow{P} P(X_j)$ (relative frequencies provide "good" estimates of probabilities).

The quantity $-\log P(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$ is the self-information provided by a message of length n of our source. This is a random variable because the message itself is random.

However, the AEP states that

$$-\frac{1}{n} \log P(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$$

which we can call the “per symbol sample entropy”, will converge towards the source entropy (in probability).

Thus, we can divide the set of all possible messages of length n into two subsets: those for which the sample entropy is close (up to ϵ) to the source entropy (we will call them typical messages) and the other messages (the atypical ones).

Hence, long messages almost surely will be typical: if we need to prove a property we need only to focus on typical messages.

The second important consequence is that we can derive an upper bound on the *number* of typical messages. Because the probability of these messages is lower bounded by $2^{-n(H+\epsilon)}$ their number is upper bounded by $2^{n(H+\epsilon)}$. This number is often much smaller than the number of possible messages (equal to $2^{n \log |\mathcal{X}|}$).

In turn this latter fact has a direct consequence in terms of data compression rates which are achievable with long sequences of source messages.

The AEP was first stated by Shannon in his original 1948 paper, where he proved the result for i.i.d. processes and stated the result for stationary ergodic processes. McMillan (1953), Breiman (1957), Chung (1961), Barron (1985), Algoet and Cover (1988) provided proofs of generalized version of the theorem, including continuous ergodic processes.

Typical sets $A_\epsilon^{(n)}$ (of ϵ-typical messages of length n)

$\mathcal{X}^n = \mathcal{X} \times \cdots \times \mathcal{X} \equiv$ set of all possible messages of length n.

$A_\epsilon^{(n)}$ w.r.t. $P(\mathcal{X}^n)$ is defined as the set of messages $(X_1, X_2, \dots, X_n)$ of $\mathcal{X}^n$ which satisfy:

$$2^{-n(H(\mathcal{X})+\epsilon)} \leq P(X_1, X_2, \dots X_n) \leq 2^{-n(H(\mathcal{X})-\epsilon)}. \tag{4}$$

It is thus a subset of messages which probability of occurrence is constrained to be close to these two limits. The following properties hold for this set:

1. $(X_1, X_2, \dots X_n) \in A_\epsilon^{(n)}$
   $\Rightarrow H(\mathcal{X}) - \epsilon \leq -\frac{1}{n} \log P(X_1, X_2, \dots X_n) \leq H(\mathcal{X}) + \epsilon$, $\forall n$.
2. $P(A_\epsilon^{(n)}) > 1 - \epsilon$, for sufficiently large n.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(\mathcal{X})+\epsilon)}$, $\forall n$.
   ($|A|$ denotes the number of elements of the set A)
4. $|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(\mathcal{X})-\epsilon)}$, for sufficiently large n.

Comments

Typical set $A_\epsilon^{(n)}$:

1. In probability close to 1.
2. In number close to 2^{nH}.

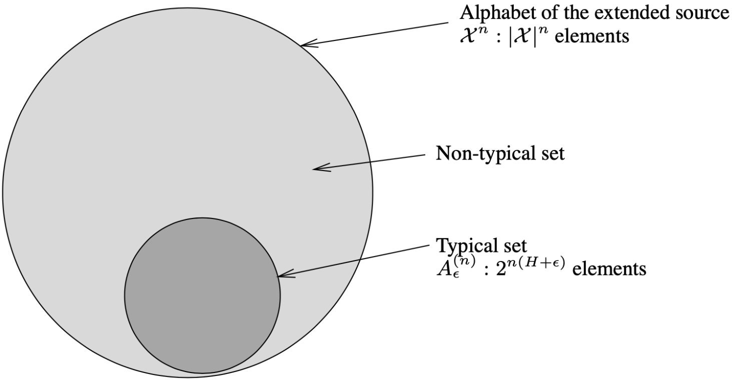
Set of all possible messages $\mathcal{X}^n$:

1. In probability = 1.
2. In number = $|\mathcal{X}|^n = 2^{n \log |\mathcal{X}|}$

If $P(\mathcal{X})$ uniform:

Typical messages = all messages. Why ?

Otherwise, the relative size of the two sets $\downarrow$ exponentially, when $n \uparrow$.


Alphabet of the extended source
$\mathcal{X}^n : |\mathcal{X}|^n$ elements
Non-typical set
Typical set
$A_\epsilon^{(n)} : 2^{n(H+\epsilon)}$ elements

Typical and non-typical messages

Data compression

We can exploit the AEP and typical sets to encode efficiently source messages.

Let us fix a value of ϵ and then n sufficiently large, so that $P(A_\epsilon^{(n)}) > 1 - \epsilon$.

Let us "construct" $A_\epsilon^{(n)}$: it contains at most $2^{n(H(\mathcal{X})+\epsilon)}$ different messages.

Let us also construct its complement $\neg A_\epsilon^{(n)}$: this set is included in $\mathcal{X}^n$. Its number of messages is therefore upper bounded by $|\mathcal{X}|^n$.

Binary code for a set containing M messages

It is possible to do it with at most $1 + \log M$ bits (actually $\lceil \log M \rceil$).
(extra bit because $\log M$ is not necessarily integer)

$\Rightarrow$ for $A_\epsilon^{(n)}$: $1 + \log\{2^{n(H(\mathcal{X})+\epsilon)}\} = n(H(\mathcal{X}) + \epsilon) + 1$ bits

$\Rightarrow$ for $\neg A_\epsilon^{(n)}$: $1 + n \log |\mathcal{X}|$ bits.

To distinguish the two kind of messages we need to add an extra bit:

0 if message is in $A_\epsilon^{(n)}$, 1 if in $\neg A_\epsilon^{(n)}$: **uniquely decodable code.**

Let us compute the expected message length (average number of bits)

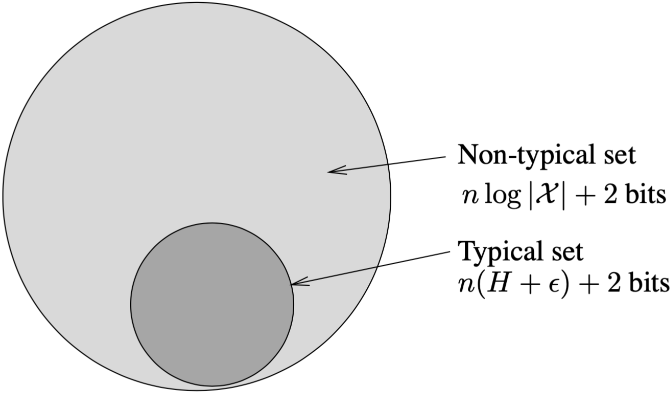
$$
\begin{aligned}
E\{\ell(\mathcal{X}^n)\} &= \sum_{X^n} P(X^n)\ell(X^n) \\
&= \sum_{X^n \in A_\epsilon^{(n)}} P(X^n)\ell(X^n) + \sum_{X^n \notin A_\epsilon^{(n)}} P(X^n)\ell(X^n) \\
&\leq \sum_{X^n \in A_\epsilon^{(n)}} P(X^n)(n(H(\mathcal{X}) + \epsilon) + 2) + \sum_{X^n \notin A_\epsilon^{(n)}} P(X^n)(n \log |\mathcal{X}| + 2) \\
&= P(A_\epsilon^{(n)})(n(H(\mathcal{X}) + \epsilon) + 2) + (1 - P(A_\epsilon^{(n)})(n \log |\mathcal{X}| + 2) \\
&\leq n(H(\mathcal{X}) + \epsilon) + \epsilon n \log |\mathcal{X}| + 2(P(A_\epsilon^{(n)}) + (1 - P(A_\epsilon^{(n)})) \\
&= n(H + \epsilon')
\end{aligned}
$$

where $\quad \epsilon' = \epsilon + \epsilon \log |\mathcal{X}| + \frac{2}{n}$

Conclusion

If we take n sufficiently large, it is possible to make ϵ' arbitrarily small.

We have thus shown that it is possible, by encoding long source messages (possibly very long messages), to build a simple code of variable word length, which requires in the average $H(\mathcal{X})$ bits/source symbol.



Questions: (i) generalization; (ii) can we do better? (iii) practical feasibility

2. Entropy rates of a stochastic process

General model for discrete sources

Indexed sequence of r.vars. $\mathcal{X}_i$, $i = 1, 2 \ldots$, *with identical (and finite) sets of possible values* $\mathcal{X}_i = \mathcal{X}$ (i represents in general time, or position in a string.)

Let's say that $\mathcal{X} = \{1, \ldots, q\}$: finite q-ary alphabet.

The process is characterized by the joint distributions: $P(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ defined on $\mathcal{X}^n$, $\forall n = 1, 2, \ldots$.

Values: $P(\mathcal{X}_1 = i^1, \ldots, \mathcal{X}_n = i^n), \forall n, \forall i^j \in \{1, \ldots, q\}$.

These distributions allow us to compute all other joint and conditional distributions of any finite combination of r.v. of the process.

For example $P(\mathcal{X}_n, \ldots, \mathcal{X}_{n+k}), \forall n > 0, k \geq 0$ can be computed by marginalization.

NB: w.r.t. Bayesian networks we extend here to an infinite number of variables...

Stationarity of a stochastic process

A stochastic process is said to be stationary if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index, i.e., $\forall n, \ell > 0, \forall i^1, \ldots, i^n \in \mathcal{X}$

$$P(\mathcal{X}_1 = i^1, \ldots, \mathcal{X}_n = i^n) = P(\mathcal{X}_{1+\ell} = i^1, \ldots, \mathcal{X}_{n+\ell} = i^n)$$

Example: memoryless stationary source. . .

Ergodicity of a stationary stochastic process

Informally:

A stationary process is said to be ergodic if temporal statistics along a trajectory converge to the ensemble *probabilities.*

Example: memoryless stationary source. . .

Markov chains

A discrete stochastic process is said to be a (first-order) Markov chain *or* Markov process, *if* $\forall n = 1, 2, \ldots,$ et $\forall i^{n+1}, i^n, \ldots i^1 \in \mathcal{X}$

$$P(\mathcal{X}_{n+1} = i^{n+1} | \mathcal{X}_n = i^n, \ldots, \mathcal{X}_1 = i^1) = P(\mathcal{X}_{n+1} = i^{n+1} | \mathcal{X}_n = i^n), \tag{5}$$

In this case the joint probability distribution may be written as

$$P(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n) = P(\mathcal{X}_1) P(\mathcal{X}_2 | \mathcal{X}_1), \ldots, P(\mathcal{X}_n | \mathcal{X}_{n-1}). \tag{6}$$

States: values $\{1, \ldots, q\}$

Time invariance:

The Markov chain is said to be time invariant if $\forall n = 1, 2, \ldots,$

$$P(\mathcal{X}_{n+1} = a | \mathcal{X}_n = b) = P(\mathcal{X}_2 = a | \mathcal{X}_1 = b), \quad \forall a, b \in \mathcal{X}. \tag{7}$$

($\neq$ stationarity)

Time invariant Markov chains

Characterized by: π *and* $\mathbf{\Pi}$*:* $\pi_i = P(\mathcal{X}_1 = i)$ *and* $\mathbf{\Pi}_{i,j} = P(\mathcal{X}_{k+1} = j | \mathcal{X}_k = i)$.

Irreducibility: *all states "communicate" (finite number of steps, two directions)*

Periodic states: *possible return times are multiple of an integer* > 1.

Recurrent states: *we always come back again to such a state*

Transient states: *with prob. =1, we transit only a finite number of times through it*

Stationary distribution: $(p_1, \ldots, p_q)$ *which satisfies* $p_j = \sum_{i=1}^{q} \Pi_{ij} p_i$

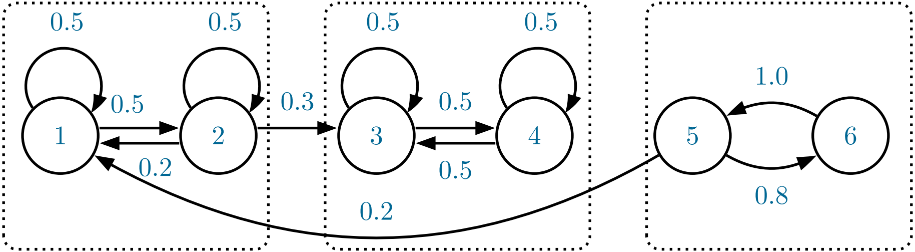
Steady state behavior: *If* $\lim_{n\to\infty} P(\mathcal{X}_n = i) = p_i$ *exists for every* $i = 1, \ldots, q$.

If this limit exists, it satisfies $p_j = \sum_{i=1}^{q} \Pi_{ij} p_i$

If the chain is initialized with a stationary distribution, it forms a stationary process

Illustration: communication classes, irreducibility. . .

a. Reducible



b. Irreducible

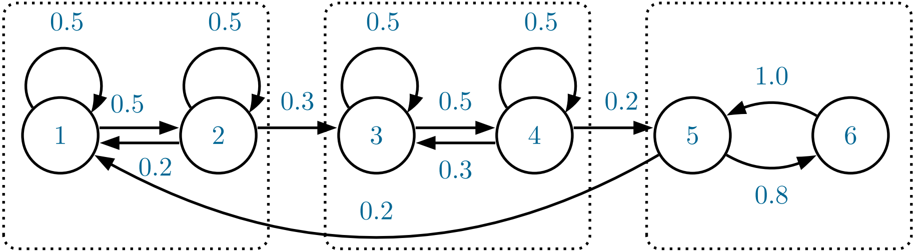
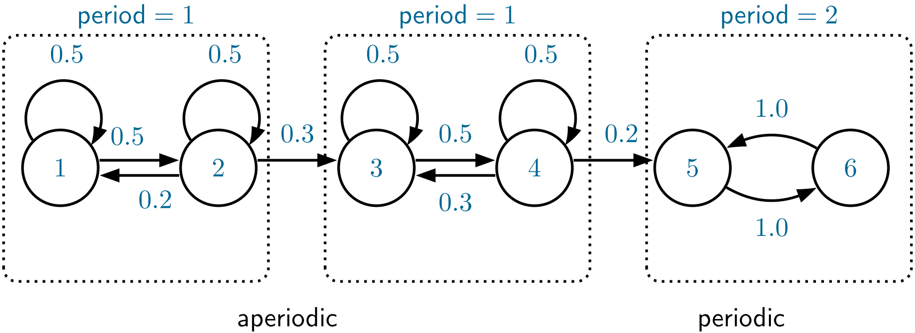


Illustration: periodicity vs aperiodicity



Existence and unicity of stationary distribution:
Garanteed for all irreducible and aperiodic chains.

Ergodicity:
Garanteed for any irreducible and aperiodic chain, if initialized with the stationary distribution.

Entropy rates of stochastic processes

NB: the joint entropy of the first n states, $H(\mathcal{X}_1, \ldots, \mathcal{X}_n)$, may stabilize or grow indefinitely and more or less quickly when n goes to infinity.

Definition: ("Average entropy per symbol")

The (asymptotic) entropy rate of a process (denoted by $H(S)$) is defined by

$$H(S) = \lim_{n \to \infty} \frac{H(\mathcal{X}_1, \ldots, \mathcal{X}_n)}{n}$$

whenever this limit exists.

Examples:

Monkey on the typewriter.
Stationary memoryless source.
Sequence of independent symbols, but not identically distributed

Alternative plausible definition: $H'(S) = \lim_{n \to \infty} H(\mathcal{X}_n | \mathcal{X}_{n-1} \ldots, \mathcal{X}_1)$.
$\Rightarrow$ *new information*/symbol

(NB: If memoryless stationary source:$H'(S) = H(S)$)

Theorem: relation between $H(S)$ and $H'(S)$

$\boxed{\textit{If } H'(S) \textit{ exists then } H(S) \textit{ exists and } H(S) = H'(S)}$

Proof:

We use the “Cesaro mean” theorem, which states

$\boxed{\textit{If } a_n \to a \textit{ and } b_n = \frac{1}{n}\sum_{i=1}^{n} a_i \textit{ then } b_n \to a.}$

Let $a_n = H(\mathcal{X}_n|\mathcal{X}_{n-1},\ldots,\mathcal{X}_1)$.

We thus have $a_n \to H'(S)$, which implies that

$$\frac{1}{n}\sum_{i=1}^{n} H(\mathcal{X}_i|\mathcal{X}_{i-1}\ldots,\mathcal{X}_1) \to H'(S)$$

But, $\forall n$: $H(\mathcal{X}_1,\ldots,\mathcal{X}_n) = \sum_{i=1}^{n} H(\mathcal{X}_i|\mathcal{X}_{i-1}\ldots,\mathcal{X}_1)$

(entropy chain rule).

Theorem: entropy rate of stationary processes

For any stationary process $H'(S)$ and hence $H(S)$ exist.

Indeed, in general we have

$$H(\mathcal{X}_{n+1}|\mathcal{X}_n, \dots, \mathcal{X}_1) \leq H(\mathcal{X}_{n+1}|\mathcal{X}_n, \dots, \mathcal{X}_2).$$

Stationarity implies also

$$H(\mathcal{X}_{n+1}|\mathcal{X}_n, \dots, \mathcal{X}_2) = H(\mathcal{X}_n|\mathcal{X}_{n-1}, \dots, \mathcal{X}_1).$$

Conclusion: for a stationary process: $H(\mathcal{X}_{n+1}|\mathcal{X}_n, \dots, \mathcal{X}_1)$ decreases in n.

Since this sequence is also lower bounded (entropies can not be negative) it must converge.

Thus $H'(S)$ exists and therefore also $H(S)$, (because of the preceeding theorem).

Bounds and monotonicity

1. $\forall n : H(S) \leq H(\mathcal{X}_n|\mathcal{X}_{n-1},\ldots,\mathcal{X}_1)$. (follows immediately from what precedes)
2. $H(S) \leq n^{-1}H(\mathcal{X}_n,\mathcal{X}_{n-1},\ldots,\mathcal{X}_1)$. (from 1 and chain rule)
3. $\frac{H(\mathcal{X}_n,\mathcal{X}_{n-1},\ldots,\mathcal{X}_1)}{n}$ is a decreasing sequence.

Indeed: $H(\mathcal{X}_n,\mathcal{X}_{n-1},\ldots,\mathcal{X}_1) = \sum_{i=1}^{n} H(\mathcal{X}_i|\mathcal{X}_{i-1},\ldots,\mathcal{X}_1)$, and $H(\mathcal{X}_i|\mathcal{X}_{i-1},\ldots,\mathcal{X}_1)$ decreases, thus average must decrease.

AEP for general stochastic processes

For a stationary ergodic process,

$$-\frac{1}{n}\log P(\mathcal{X}_1,\ldots,\mathcal{X}_n) \longrightarrow H(S)$$

with probability 1.

Entropy rate of a Markov chain.

For a stationary Markov chain, the entropy rate is given by $H(\mathcal{X}_2|\mathcal{X}_1)$.

Indeed:

$$\begin{aligned} H(S) = H'(S) &= \lim_{n\to\infty} H(\mathcal{X}_n|\mathcal{X}_{n-1},\ldots,\mathcal{X}_1) \\ &= \lim_{n\to\infty} H(\mathcal{X}_n|\mathcal{X}_{n-1}) = H(\mathcal{X}_2|\mathcal{X}_1), \end{aligned} \tag{8}$$

NB: to compute $H(\mathcal{X}_2|\mathcal{X}_1)$ we use a stationary disribution for $P(\mathcal{X}_1)$ and Π.

Question :

What if the Markov chain is not intitialized with the stationary distribution ?

3. Data compression theory

Problem statement (source coding)

Given:

A source S using Q-ary alphabet (e.g. $\mathcal{S} = \{a, b, c, \ldots, z\}$)

Model of S as a stochastic process (probabilities)

A q-ary code alphabet (e.g. $\{0, 1\}$)

Objectives:

Build a *code*: a mapping of source messages into sequences of code symbols.

Minimise encoded message length.

Be able to reconstruct the source message from the coded sequence.

NB: = compact representation of stochastic process trajectories.

Let us particularize.

A code: rule which associates to each source *symbol* s_i of S a word m_i, composed of a certain number (say n_i) code symbols.

Notation: $s_i \xrightarrow{C} m_i$

Extension of the source:

Nothing prevents us from coding blocs of n successive source symbols.

We merely apply the preceding ideas to the extended source S^n.

Code for S^n: $s_{i_1} \cdots s_{i_n} \xrightarrow{C} m_{i_1,\ldots,i_n}$

Extension of a code:

Let C be a code for S.

Order n extension of C: $s_{i_1} \cdots s_{i_n} \xrightarrow{C^n} m_{i_1} \cdots m_{i_n}$

$\Rightarrow$ a special kind of code for S^n.

Properties (desired) of codes

Focus on decodability.

Regular codes. $m_i \neq m_j$ (or non-singularity)

Uniquely decodable codes ($\equiv$ non ambiguous)

E.g. constant word length. Code-words with separators. Prefix-less code.

Instantaneous. possibility to decode "on-line".

Relations:

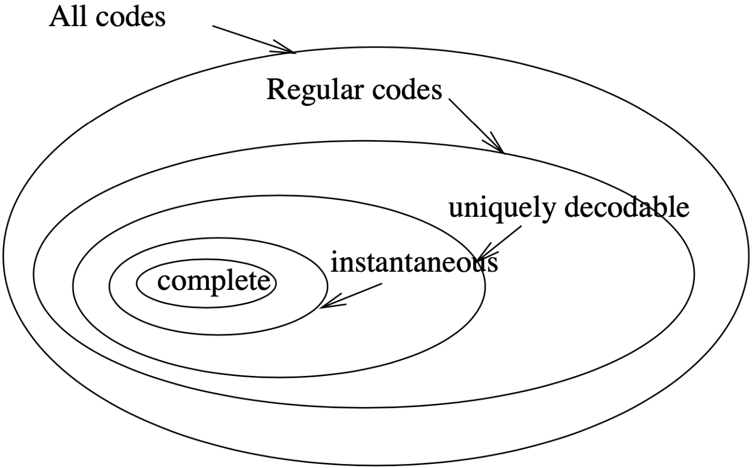
Uniquely decodable codes $\Leftrightarrow$ all extensions are regular.

Prefix free $\Rightarrow$ uniquely decodable (but $\not\Leftarrow$)

Prefix free $\Leftrightarrow$ instantaneous.

Proofs: almost trivial...

Example of uniquely decodable code, but not prefix-free ?


All codes
Regular codes
uniquely decodable
instantaneous
complete

Definition: A code is prefix-free (or prefix code, or instantaneous code) if no codeword is a prefix of any other codeword.

Example

s_i	Singular	Non-singular, but not uniquely decodable	Uniquely decodable, but not instantaneous	Instantaneous
s_1	0	0	10	0
s_2	0	010	00	10
s_3	0	01	11	110
s_4	0	10	110	111

Kraft inequality: the $\boxed{\text{FUNDAMENTAL}}$ result for codes

What can we say about the existence of a code, given word lengths n_i and alphabet sizes Q and q ?

McMillan condition. If $n_1, \ldots, n_Q$ denote candidate word lengths, then

$$\sum_{i=1}^{Q} q^{-n_i} \leq 1 \tag{9}$$

$\Leftrightarrow$ **there exists** a uniquely decodable code having such word lengths.

Instantaneous codes. *The Kraft inequality is also a necessary and sufficient condition of existence of an instantaneous code.*

Conclusion.

If there is a uniquely decodable code using some word lengths, there is also an instantaneous code using the same word lengths.

The **necessary condition** means that:

- If a code is uniquely decodable (and hence a fortiori if it is instantaneous), the word lengths must satisfy (9).
- Means also that if the condition is not satisfied by a given code, it can not be uniquely decodable (and certainly not instantaneous).

The **sufficient condition** means (more subtly) that: if we specify a set of word lengths which satisfy (9), then it is possible to define a uniquely decodable (and even instantaneous) code using these lengths.

Usefulness of the result ?

Tells us that the n_i are sufficient to decide about decodability existence. We can now focus on instantaneous codes without loosing anything in terms of data compression possibilities.

Proofs

We will not provide the proofs of these results. See e.g. Cover and Thomas for the proofs.

Remarks

If $\sum_{k=1}^{s} r_k q^{-k} < 1$ it is possible to add one more word (or several) to the code
(Choose n_{i+1} sufficiently large.)

If $\sum_{k=1}^{s} r_k q^{-k} = 1$ it is impossible to complete the code in such a way.

One says that the code is **complete**

First Shannon theorem

Rem.: average length/symbol of messages of a memoryless stationary S: $\sum_{i=1}^{Q} P(s_i)n_i$
$=$ expected length of the codeword of the first symbol S.

A. Memoryless stationary source, coded symbol per symbol:

1. **Lower bound on average length** $\overline{n} = \sum_{i=1}^{Q} P(s_i)n_i \geq \frac{H(S)}{\log q}$
2. **It is possible to reach** $\overline{n} < \frac{H(S)}{\log q} + 1$.

B. Order k extension of this source

S^k $=$ memoryless source of entropy $kH(S)$:

Optimal code must satisfy: $\frac{kH(S)}{\log q} \leq \overline{n^k} < \frac{kH(S)}{\log q} + 1$

Per symbol of the original source S: divide by k: $k \to \infty$

C. General stationary source

For any stationary source, there exists a uniquely decodable code, such that the average length $\overline{n}$ per symbol is as close as possible to its lower bound $H(S)/\log q$.

In fact: take a stationary source, and messages of length s of this source and an optimal code for these;

Then: we have

$$\frac{H_s(S)}{\log q} \leq \overline{n^s} < \frac{H_s(S)}{\log q} + 1,$$

where $H_s(S) = H(S_1, \ldots, S_s)$ denotes the joint entropy of the first s source symbols.

This implies

$$\lim_{s\to\infty} \frac{\overline{n^s}}{s} = \frac{H(S)}{\log q}.$$

Proof

1. Lower bound: $\frac{H(S)}{\log q}$

Take a uniquely decodable code $\Rightarrow$ satisfies Kraft inequality.

Thus $0 < \mathcal{Q} = \sum_{i=1}^{Q} q^{-n_i} \leq 1$.

Let $q_i = \frac{q^{-n_i}}{\mathcal{Q}}$: a kind of probability distribution.

$\boxed{\textbf{Gibbs}}$ $\Rightarrow$: $\sum_{i=1}^{n} P(s_i) \log \frac{q_i}{P(s_i)} \leq 0$

$\Leftrightarrow -\sum_{i=1}^{n} P(s_i) \log P(s_i) \leq -\sum_{i=1}^{n} P(s_i) \log \frac{q^{-n_i}}{\mathcal{Q}}$

But: $-\sum_{i=1}^{n} P(s_i) \log \frac{q^{-n_i}}{\mathcal{Q}} = \log q \sum_{i=1}^{n} P(s_i) n_i + \log \mathcal{Q}$

Since $\mathcal{Q} \leq 1, \log \mathcal{Q} \leq 0$.

What about equality ?

2. Is it possible to reach $\overline{n} < \frac{H(S)}{\log q} + 1$.

We only need to find one code for which it works. Ideas ?

Shannon code: word lenghts: $n_i = \lceil -\log_q P(s_i) \rceil$

$\Rightarrow \frac{-\log P(s_i)}{\log q} \leq n_i = \lceil -\log_q P(s_i) \rceil < \frac{-\log P(s_i)}{\log q} + 1$

Average length: multiply by $P(s_i)$ and sum over i

$\frac{H(S)}{\log q} \leq \overline{n} < \frac{H(S)}{\log q} + 1$.

But can we prove existence of such a code ?

Yes, since the n_i satisfy Kraft inequality.

Indeed: $\frac{-\log P(s_i)}{\log q} \leq n_i \Leftrightarrow q^{-n_i} \leq P(s_i)$

Sum over i: $\sum_{i=1}^{Q} q^{-n_i} \leq \sum_{i=1}^{Q} P(s_i) = 1$

Further reading

- D. MacKay, *Information theory, inference, and learning algorithms*
  - Chapters 4, 5

Frequently asked questions

- Define the notion of Markov chain over three discrete random variables. State, prove, discuss and illustrate the data processing inequality. Give an example where the data processing inequality can not be applied and where it is also not satisfied. State and discuss the corollaries of the data processing inequality.
- Define the notion of convergence in probability, formulate the AEP, and explain the principle of its proof. Define the set of ϵ-typical messages of length $n(A_\epsilon^{(n)})$, and state exactly the 4 main properties of this set. Explain the use of this notion in the context of data compression.
- Define the notion of (discrete) stationary data source. State the 2 definitions of the entropy rate of a discrete stationary source ($H(S)$ and $H'(S)$). Show mathematically why the existence of $H'(S)$ implies that of $H(S)$. Show mathematically that $H'(S)$ exists for a stationary source.
- State the first Shannon theorem and prove it mathematically. Give two examples of reversible data compression methods and explain their advantages and drawbacks.