Asymptotics on the Lempel-Ziv 78 compression of Markov sources

Exploring analytic information theory : from Markov source sampling to combinatorial analysis proofs

Guillaume Duboc

Computer Science Department
Ecole Normale Supérieure de Lyon

M1 Internship, 2018

ENS DE LYON

Table of contents

Table des matières

ENS DE LYON

Words or sequences, and memoryless sources

Definition : word or sequence or string

Given an alphabet $\mathcal{A}$, a ***word*** or ***sequence*** or ***string*** is an infinite sequence of random variables $X = (X_k)_{k\mathbb{N}^*}$, each X_k representing a symbol in $\mathcal{A}$.

Definition : Bernoulli or Memoryless source

A source of information is a ***Bernoulli*** or ***memoryless source*** when all the symbols of $\mathcal{A}$ occur independently with a fixed probability. The word can be seen as an *infinite sequence of Bernoulli trials*.

Markov sources definition

Definition : Markov source

An information source is a ***Markov source*** when there is a ***Markov dependency*** between the consecutive symbols of a string.

Definition : order of a Markov source

Let $V = |\mathcal{A}|$. A ***Markov source*** is of ***order*** r when the dependency can be encoded in a transition matrix of size $V^r \times V$, with coefficients :

$$P(c|w) \qquad \forall\,(w,c) \in \mathcal{A}^r \times \mathcal{A}$$

Informally : *the probability that a symbols occurs depends on the previous r symbols.*

Table des matières

Description of the LZ78 algorithm

Algorithm

Given a word w.

- Initialize an empty dictionary
- While it is possible :

    Find longest prefix of w that is not in the dictionary

    Add it to the dictionary, cut it from w

Elements description

The data representation is (dictionary_reference, symbol).

Remarks

The LZ78 algorithm builds a prefix tree from which the original word can be reconstructed.

Definition : number of phrases

After compressing a word w, the number of phrases in the dictinary is noted $M(w)$.
For words of size n, we write $M_n(w)$.

Code length

$$C(w) = \sum_{k=0}^{M(w)} \left(\lceil \log_2(k) \rceil + \lceil \log_2(\mathcal{A}) \rceil \right)$$

ENS DE LYON

Table des matières

ENS DE LYON

Definition : compression ratio

Let w a word, and $C(w)$ its *encoding* by a compression algorithm. The ***compression ratio*** of w is $\dfrac{|C(w)|}{|w|}$.

Main goals of compression algorithms

- Improving the compression ratio
- Fast compression/decompression speed in Mb/s

T

he tradeoff between these two goals is a sensitive research problem. Different compression standards :

- Google (Brotli, 2015)
- Facebook (Zstandard, 2016)

Optimal encoding

Entropy of a Markov source

Let π be a stationary distribution. The entropy of a Markov chain is

$$h = -\sum_{i=1}^{V} \pi \sum_{j=1}^{V} p_{ij} \log(p_{ij})$$

Optimality of LZ78

Considering words of length n.

$$\frac{|C(w)|}{|w|} - h \text{ goes to zero for } n \to +\infty$$

Table des matières

Optimal parsing

Flexible parsing

Table des matières

ENS DE LYON

Markov Independent Model

$$
\begin{aligned}
X(1) &= 0000000\ldots \\
X(2) &= 1010101\ldots \\
X(3) &= 1001101\ldots \\
X(4) &= 001100111\ldots
\end{aligned}
$$

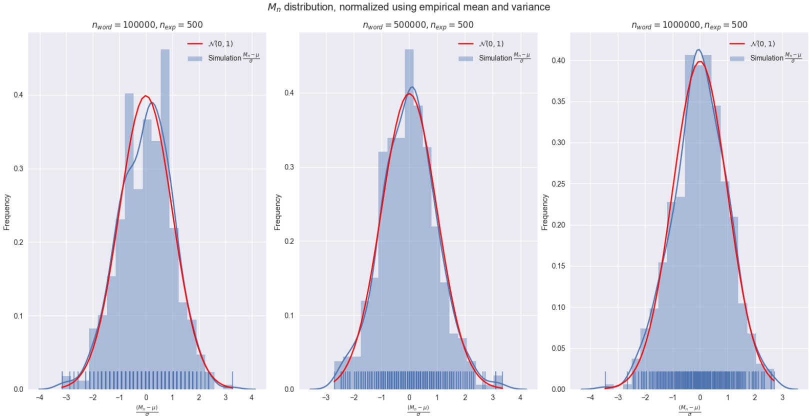
Table des matières

Coding details

- Python code $\sim$ 2000 lines
- Markov source sampling
- Optimized datastructure (digital search tree)
- Parallelization
- Reproducibility of datasets

Table des matières

Central Limit Theorem confirmation



ENS DE LYON

Hypothesis testing for the variance

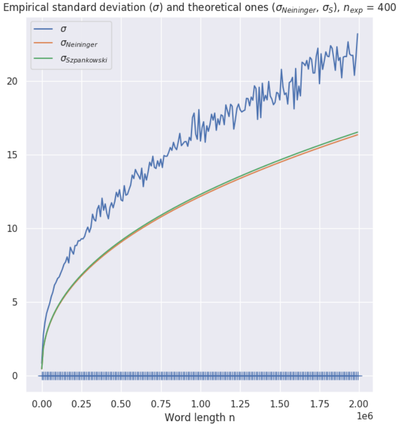
Complex matrix

Defining $P(s)$ as $\begin{matrix} p_{11}^{-s} & p_{12}^{-s} \\ p_{21}^{-s} & p_{22}^{-s} \end{matrix}$

Variance expression

$$V_n = \left(\ddot{\lambda}(-1) - \dot{\lambda}(-1)^2\right) \frac{n}{\ln^2 n}$$

ENS DE LYON


Empirical standard deviation (σ) and theoretical ones (σNeininger, σS), nexp = 400
σ
σNeininger
σSzpankowski
0
5
10
15
20
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e6
Word length n

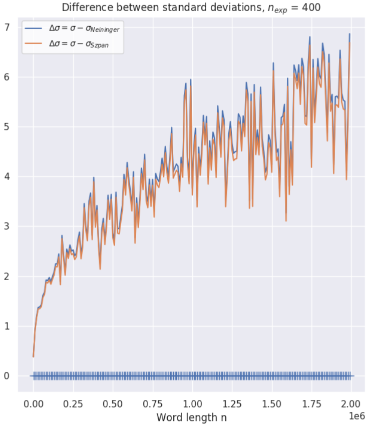

Difference between standard deviations, nexp = 400
Δσ = σ − σNeininger
Δσ = σ − σSzpan
0
1
2
3
4
5
6
7
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e6
Word length n

Table des matières

Definition, usage

Definition

$$A(z) = \sum_{n \geqslant 0} a_n z^n$$

Remarks

- Used as an algebraic item with the convolution product
- No convergence problems

ENS DE LYON

Table des matières

Poissonization and Depoissonization

$$\widetilde{G}(z) = \sum_{n \geqslant 0} a_n \frac{z^n}{n!} \mathrm{e}^{-z}$$

Mellin transform

Make recurrence relation between random variables become linear in order to solve them more easily.

Table des matières

Tail symbols

Illustration

$$
\begin{aligned}
X(1) &= 0000000\ldots\\
X(2) &= 1010101\ldots\\
X(3) &= 1001101\ldots\\
X(4) &= 001100111\ldots
\end{aligned}
$$

Definition

Let c be a character from our alphabet $\{a, b\}$. In the case when all the sequences start with a c, we define T_n^c the *number of times a is a tail symbol in the experiment*.

ENS DE LYON

Definition and relation

Recurrence

For $n \geqslant 0$, we have :

$$\boxed{T^c_{n+1} = \delta_a + \widetilde{T}^a_{N_a} + \widetilde{T}^b_{N_b}}$$

Notations

- $\delta_a = \begin{cases} 1 & \text{if } a \text{ is the tail symbol of the first sequence} \\ 0 & \text{else} \end{cases}$
- N_a is the random variable giving *the size of the left subtree which contains phrases whose second letter is a*
- $\widetilde{T}^a_{N_a}$ is the number of times a is a tail symbol for the sequences that were used to build the subtree with

Total path lenght

Definition

Defining L_n^c as the *total path length of the nodes of the DST that was built with MI model with n sequences starting with letter c*. It is the sum of the lengths of all the prefix phrases.

Recurrence relation

For all $n \geqslant 0$:

$$\boxed{L_{n+1}^c = n + \widetilde{L}_{N_a}^a + \widetilde{L}_{N_b}^b}$$

Table des matières

ENS DE LYON

Inconclusive, but informative

Table des matières

Recurrence

$$\boxed{\mathrm{Cov}(T^c_{n+1}, L^c_{n+1}) = \mathrm{Cov}(\widetilde{T}^a_{N_a}, \widetilde{L}^a_{N_a}) + \mathrm{Cov}(\widetilde{T}^b_{N_b}, \widetilde{L}^b_{N_b})}$$

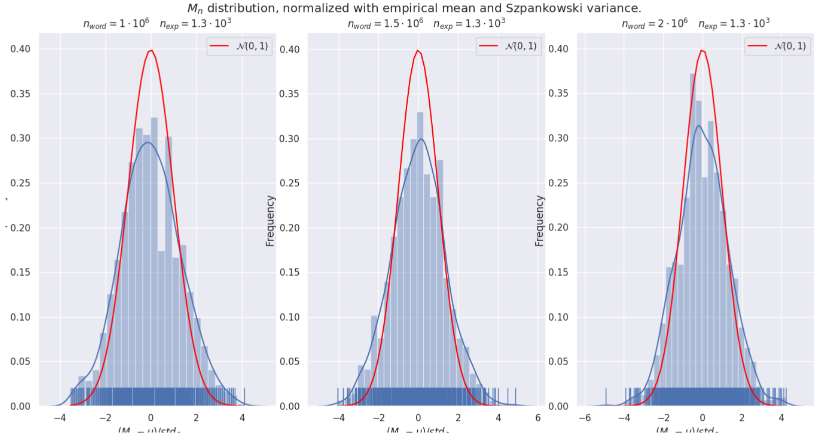
Poisson transform

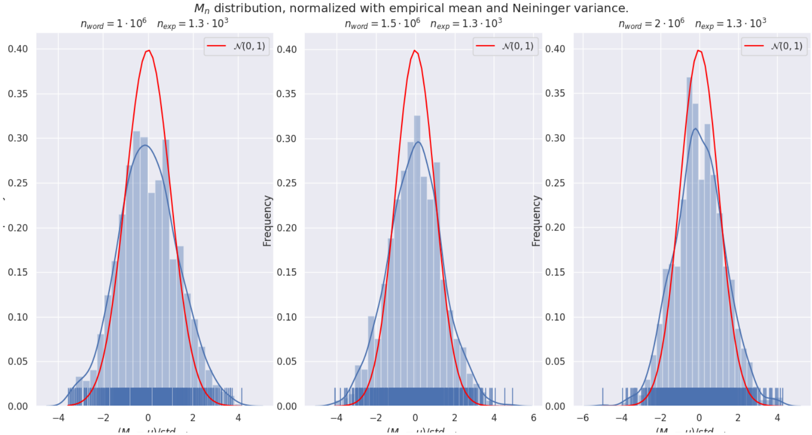
Defining

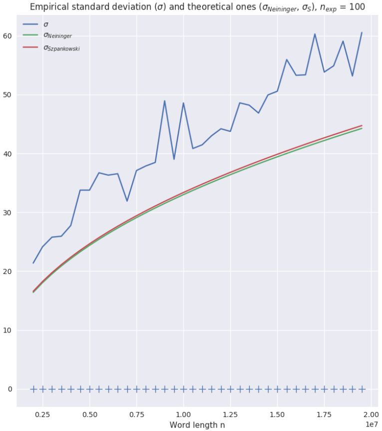
$$\boxed{C_c(z) = \sum_{n \geqslant 0} \mathrm{Cov}(T_n^c, L_n^c) \frac{z^n}{n!} \mathrm{e}^{-z}}$$

Differential equation

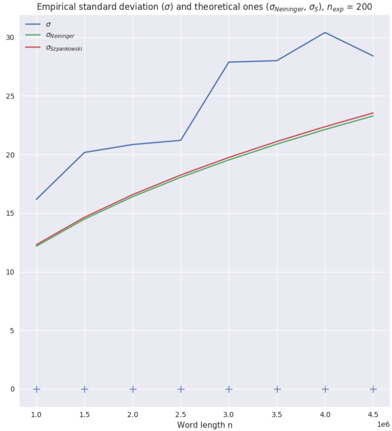
$$\boxed{\partial_z C_c(z) + C_c(z) = C_a(zp) + C_b(zq)}$$

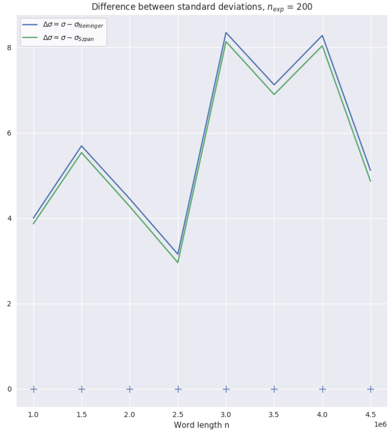

M_n distribution, normalized with empirical mean and Szpankowski variance.
$n_{word} = 1 \cdot 10^6$ $n_{exp} = 1.3 \cdot 10^3$
$n_{word} = 1.5 \cdot 10^6$ $n_{exp} = 1.3 \cdot 10^3$
$n_{word} = 2 \cdot 10^6$ $n_{exp} = 1.3 \cdot 10^3$
$\mathcal{N}(0, 1)$
Frequency


M_n distribution, normalized with empirical mean and Neininger variance.
$n_{word} = 1 \cdot 10^6$ $n_{exp} = 1.3 \cdot 10^3$
$n_{word} = 1.5 \cdot 10^6$ $n_{exp} = 1.3 \cdot 10^3$
$n_{word} = 2 \cdot 10^6$ $n_{exp} = 1.3 \cdot 10^3$
$\mathcal{N}(0, 1)$
Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6


Empirical standard deviation (σ) and theoretical ones ($\sigma_{Neininger}$, σ_S), n_{exp} = 100
σ
$\sigma_{Neininger}$
$\sigma_{Szpankowski}$
0
10
20
30
40
50
60
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e7
Word length n


Difference between standard deviations, n_{exp} = 100
$\Delta\sigma = \sigma - \sigma_{Neininger}$
$\Delta\sigma = \sigma - \sigma_{Szpan}$
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e7
Word length n


Empirical standard deviation (σ) and theoretical ones ($\sigma_{Neininger}$, σ_S), n_{exp} = 200
σ
$\sigma_{Neininger}$
$\sigma_{Szpankowski}$
Word length n
1e6


Difference between standard deviations, n_{exp} = 200
$\Delta\sigma = \sigma - \sigma_{Neininger}$
$\Delta\sigma = \sigma - \sigma_{Szpan}$
Word length n
1e6

ENS DE LYON