Ecole Normale Superieure de Lyon

Computer Science Department

Master's Internship Report

# Asymptotics on Lempel-Ziv 78 compression

*Intern :*
Guillaume Duboc

*Supervisor :*
Wojtek Szpankowski

Center for Science of Information - Purdue University

May 21st - August 8th, 2018

École Normale Supérieure de Lyon

Purdue University

# Asymptotics on Lempel-Ziv 78 compression

## Table des matières

# I Introduction

*Data compression*, *source coding*, or *bit-rate reduction* involve encoding information using fewer bits than the original representation.

## 1 Introduction to the problem

Data compression is achieved by different types of algorithms, and it can either be lossy or lossless. In order to study it precisely, we define a data compression scheme. First, we define an algorithm which operates on words. Then, we introduce a probabilitic model for the data to be compressed, which allows us to quantify the efficiency of our algorithm.

### 1.1 Probabilistic models

**Definition 1** *Let $\mathcal{A}$ be an alphabet. An* information source *is a one-sided infinite sequence of random variables $(\mathrm{X}_k)_{k=1}^{+\infty}$ with each $\mathrm{X}_k$ having values in $\mathcal{A}$.*

**Remark 1** *Each realization of an information source is called a* sequence *or* word*.*

**Remark 2** Defining *the law of the $\mathrm{X}_k$ produces out different models for data generation which can be studied mathematically and* simulated*.*

**Definition 2** *A* memoryless source *is an* information source *for which the $\mathrm{X}_k$ are mutually independent, following the uniform law on $\mathcal{A} = \{a_1, \ldots, a_\mathrm{V}\}$ :*

$$P(\mathrm{X}_k = a_k) = p_k \qquad \textit{with } \sum_{i=1}^{\mathrm{V}} p_i = 1$$

**Remark 3** *This is the simplest information source and it has been studied successfully in the past, but it is not a realistic model. We replace it with the following Markov model whenever possible.*

**Definition 3** *A* Markov source *is an* information source *with a Markov dependency between successive symbols.*

**Definition 4** *A* Markov source of order *$r$ is a Markov source for which each symbol apparition depends on the previous $r$ symbols.*

**Remark 4** *We will study Markov sources of order 1 - where each symbol simply depends on the previous one. This is general enough, as Markov sources of superior order can be simulated by expanding the alphabet and using a Markov source of order 1.*

### 1.2 Lempel-Ziv 78

In general, Lempel-Ziv algorithms exploit previously seen redundancy to save off coding space. The Lempel-Ziv 78 does so by constructing a prefix tree which allows to describe parts of a word by referring to previously seen phrases.

### 1.3 Probabilistic analysis

Under this context, we can define and conduct a thorough analysis of several random variables with different meanings regarding the effectiveness of compression.

**Notation 1** *Let $n$ be an integer - the size of the considered words. Defining $\Omega_n$ the set of words of size $n$ on alphabet $\mathcal{A}$. Each word being an event, a natural probability space is given by considering the output of size $n$ of a Markov source.*

**Remark 5** *We will now study random variables defined on this probability space.*

**Notation 2** *The output of a Markov process is denoted by the random variable* $\mathrm{W}$.

**Notation 3** *The* number of phrases *used to compress a word* $\mathrm{W}$ *with LZ78 is given by* $\mathrm{M}_n(\mathrm{W})$ *or simply* $\mathrm{M}_n$*.*

**Remark 6** *This is one of the most important variables to consider because, as it will appear shortly, it is closely tied to the* compression ratio *of LZ78.*

**Definition 5** *The* codelength *is the* number of bits *required to encode the LZ78-compressed version of a word, denoted by* $\mathrm{C}(\mathrm{W})$*.*

**Definition 6** *The* compression ratio *is the ratio between the codelength and initial size of a word,* $\dfrac{\mathrm{C}(\mathrm{W})}{|\mathrm{W}|}$*.*

# II Numerical simulation of LZ78 for Markovian sources

## 1 Conditions of the experiment

### 1.1 Simulation details

Simulations in this report follow this experimental process:

- Generating a random Markov chain of size 2 of matrix

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

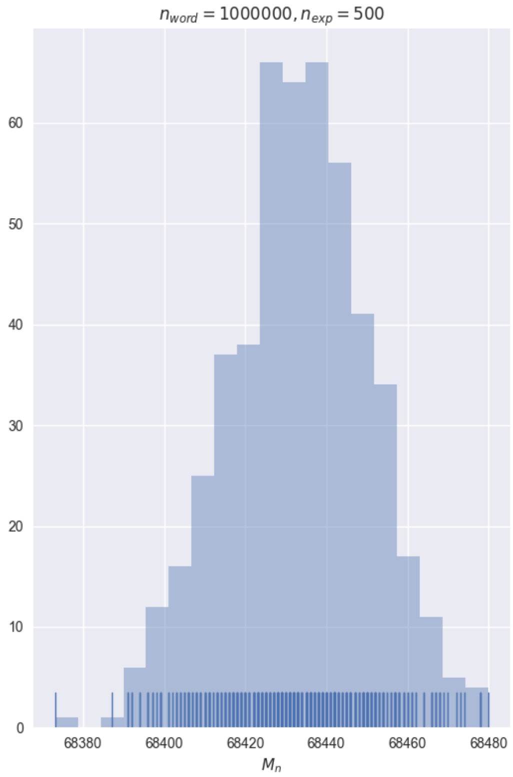
- Generating $n_{\text{exp}} \sim 10^3$ words of length $n$ (or $n_{\text{word}}$), with $n \sim 10^6$ or $10^7$
- Applying LZ78 on each of these words to estimate, for each $n$, the number of phrases $\mathrm{M}_n$. A simple histogram of these values can be seen in figure 1.
- From this sampling of the random variable $\mathrm{M}_n$ and other parameters such as the entropy of the Markov chain, computing
  - the empirical mean ($\mu$) and the empirical variance ($\sigma^2$)
  - different theoretical expressions for the variance
- Using these expressions to standardize $\mathrm{M}_n$ in different ways, plotting
  - the probability distribution of $\mathrm{M}_n$ (standardized)
  - the cumulative distribution function of $\mathrm{M}_n$ (standardized)
- Finally, comparing the different theoretical expressions for the variance by plotting their differences for a large range of values of $n$, and a constant number of experiments $n_{\text{exp}}$.

## 1.2 Implementation

## 1.3 Code overview

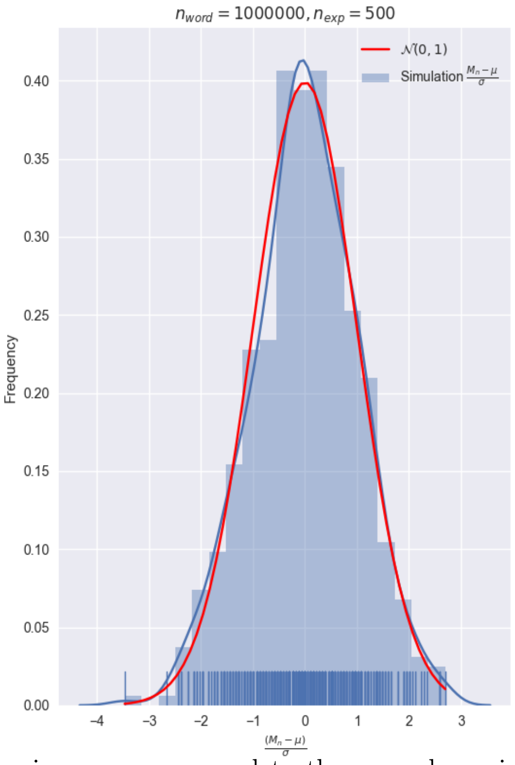
## 1.4 Example histogram

This histogram represents the counts of the different values taken by $\mathrm{M}_n$ for $n = 10^6$. Each tick on the x-axis is a data point.

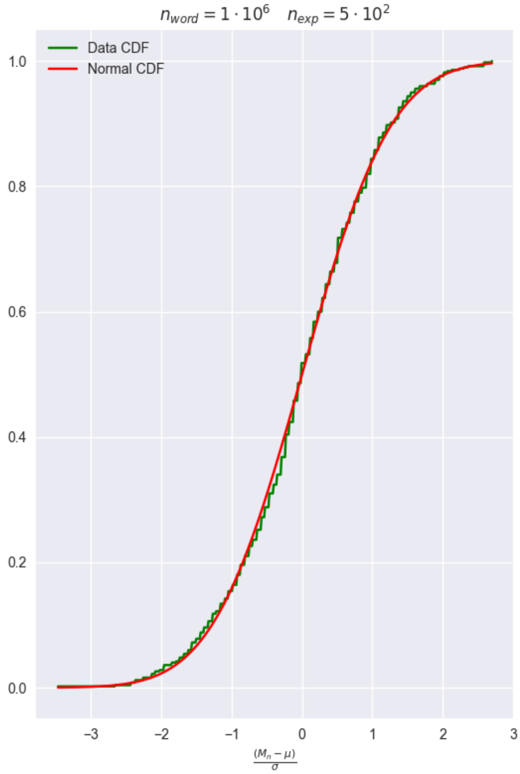


## 1.5 Empirical normalization

Using the empirical mean ($\mu$) and variance ($\sigma^2$) of the dataset to normalize $\mathrm{M}_n$, this is a plot of the normalized distribution, compared to the normal distribution in red :



and its cumulative distribution function in green, compared to the normal one in red :



These simulations and figures strongly indicate that the general distribution of $\mathrm{M}_n$ respects the central limit theorem. We now experiment with candidates for the variance of $\mathrm{M}_n : \mathrm{V}_n$

# 2 Validating variance candidates

## 2.1 A first expression

As it will be used in the next section, this is the detail of the expression from [?] :

$$\mathrm{V}_n = \frac{\mathrm{H}^3 \sigma^2 n}{\log_2^2(n)}$$

$$\sigma^2 = \sigma_0^2 + \sigma_1^2$$

where $$\sigma_i^2 = \frac{\pi_i p_{i0} p_{i1}}{\mathrm{H}^3} \left( \log_2 \left( \frac{p_{i0}}{p_{i1}} \right) + \frac{\mathrm{H}_1 - \mathrm{H}_0}{p_{01} + p_{10}} \right)^2$$

with $$\pi_0 = \frac{p_{10}}{p_{10} + p_{01}} \qquad \pi_1 = \frac{p_{01}}{p_{10} + p_{01}}$$

and $$\mathrm{H}_i = -p_{i0} \log_2(p_{i0}) - p_{i1} \log_2(p_{i1}) \qquad \mathrm{H} = \pi_0 \mathrm{H}_0 + \pi_1 \mathrm{H}_1$$

## 2.2 Using the Frobenius eigenvalue of $\mathrm{P}(s)$

An expression which seems to be succesful for the variance is :

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

Let's compute $\ddot{\lambda}(-1)$ with a Markov chain of order 1.

In the paper, $\ddot{\lambda}(-1) = \pi\ddot{\mathrm{P}}(-1)\psi + 2\dot{\pi}(-1)\dot{\mathrm{P}}(-1)\psi - 2\dot{\lambda}(-1)\dot{\pi}(-1)\psi$

However, the relations defining $\pi(s)$ :

$$\begin{cases} \pi(s)\mathrm{P}(s) & = \lambda(s)\pi(s) \\ \mathrm{P}(s)\psi(s) & = \lambda(s)\psi(s) \\ \pi(s)\psi(s) & = \lambda(s) \end{cases}$$

did not seem to allow me to directly compute $\dot{\pi}(s)$ (it seemed like I need one more). Therefore, I computed $\lambda(s)$ as the greatest eigenvalue of $\mathrm{P}(s)$. Let $\chi$ the characteristic polynomial of $\mathrm{P}(s)$, and $\Delta$ its discrimant

$$\chi = (\mathrm{X} - {p_{00}}^{-s})(\mathrm{X} - {p_{11}}^{-s}) - (p_{01}\,p_{10})^{-s}$$

and
$$\begin{aligned}\Delta &= ({p_{00}}^{-s} + {p_{11}}^{-s})^2 - 4[(p_{00}\,p_{11})^{-s} - (p_{01}\,p_{10})^{-s}] \\ &= {p_{00}}^{-2s} + {p_{11}}^{-s} - 2(p_{00}\,p_{11})^{-s} + 4(p_{01}\,p_{10})^{-s}\end{aligned}$$

Informally, we have this expression for $\lambda(s)$ where we need to decide which sign is the correct one :

$$\boxed{\lambda(s) = \frac{{p_{00}}^{-s} + {p_{11}}^{-s} \pm \sqrt{\Delta(s)}}{2}}$$

Since $\Delta(-1) = (p_{00} + p_{11})^2 - 2p_{00}p_{11} + 4p_{01}p_{10} = (p_{00} + p_{11} - 2)^2$

then $\sqrt{\Delta(-1)} = 2 - p_{00} - p_{11} = p_{01} + p_{10}$. Thus, picking the + sign in the former expression, we verify that

$$\lambda(-1) = \frac{p_{00} + p_{11} + \sqrt{\Delta(-1)}}{2} = 1$$

Derivating $\dot{\lambda}(s) = \dfrac{1}{2}\left(-\ln p_{00}\,{p_{00}}^{-s} - \ln p_{11}\,{p_{11}}^{-s} + \dfrac{\Delta'(s)}{2\sqrt{\Delta(s)}}\right)$

and the expression for $\Delta'(s)$

$$\Delta'(s) = -2\ln p_{00}\,{p_{00}}^{-2s} - 2\ln p_{11}\,{p_{11}}^{-2s} + 2\ln(p_{00}p_{11})\,(p_{00}\,p_{11})^{-s} - 4\ln(p_{01}p_{10})\,(p_{01}\,p_{10})^{-s}$$

gives

$$\Delta'(-1) = -2\ln p_{00}\,{p_{00}}^2 - 2\ln p_{11}\,{p_{11}}^2 + 2\ln(p_{00}p_{11})\,(p_{00}p_{11}) - 4\ln(p_{01}p_{10})\,(p_{01}p_{10})$$

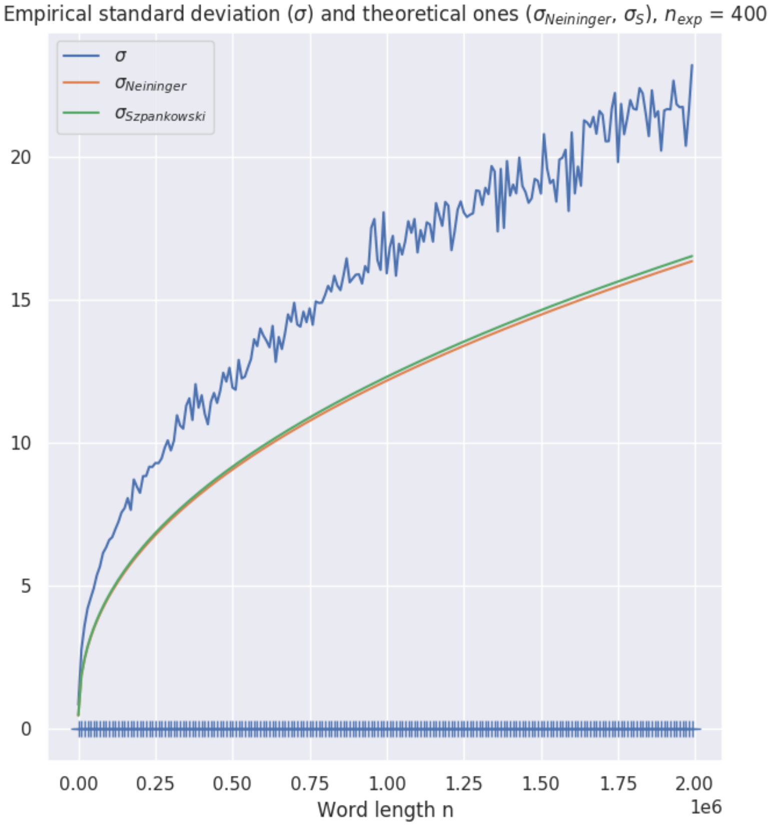
allowing to compute $\dot{\lambda}(-1)$. Numerically, we verified that $\dot{\lambda}(-1) = h$. Derivating again yields

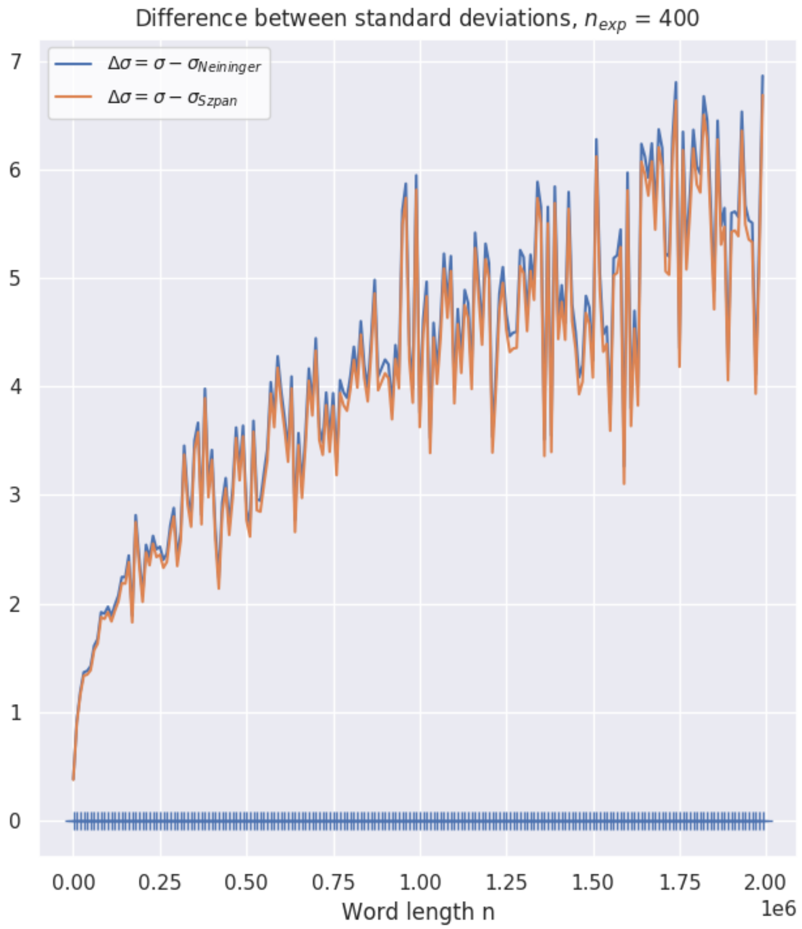
$$\ddot{\lambda}(s) = \frac{1}{2}\left(\ln^2 p_{00}{p_{00}}^{-s} + \ln^2 p_{11}{p_{11}}^{-s} + \frac{\Delta''(s)\sqrt{\Delta(s)} - \Delta'(s)\cdot\Delta'(s)/2\sqrt{\Delta(s)}}{2\Delta(s)}\right)$$

with $\Delta''(s) = 4\ln^2 p_{00}\,{p_{00}}^{-2s} + 4\ln^2 p_{11}\,{p_{11}}^{-2s} - 2\ln^2(p_{00}p_{11})\,(p_{00}\,p_{11})^{-s} + 4\ln^2(p_{01}p_{10})\,(p_{01}\,p_{10})^{-s}$

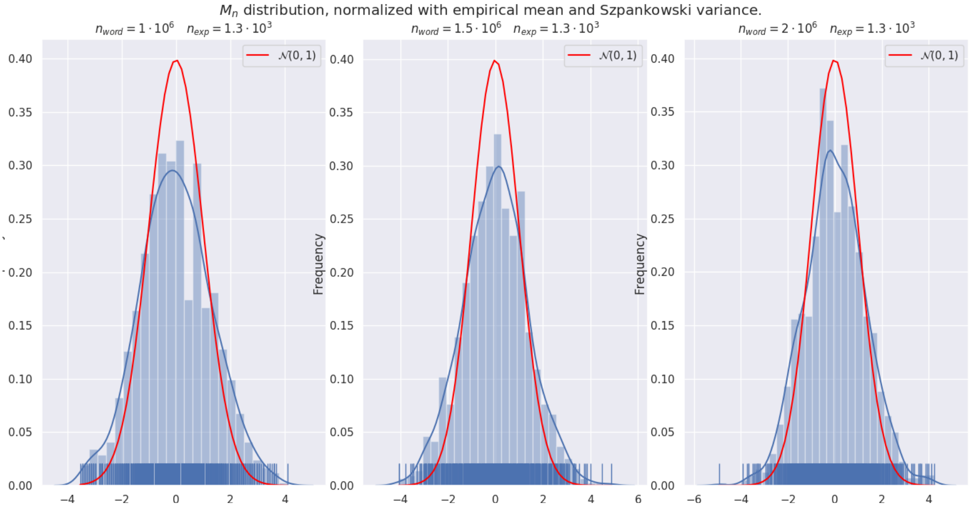
Finally
$$\boxed{\ddot{\lambda}(-1) = \frac{1}{2}\left(\ln^2 p_{00}\,p_{00} + \ln^2 p_{11}\,p_{11} + \frac{\Delta''(-1)\sqrt{\Delta(-1)} - \Delta'(-1)^2/2\sqrt{\Delta(-1)}}{2\Delta(-1)}\right)}$$

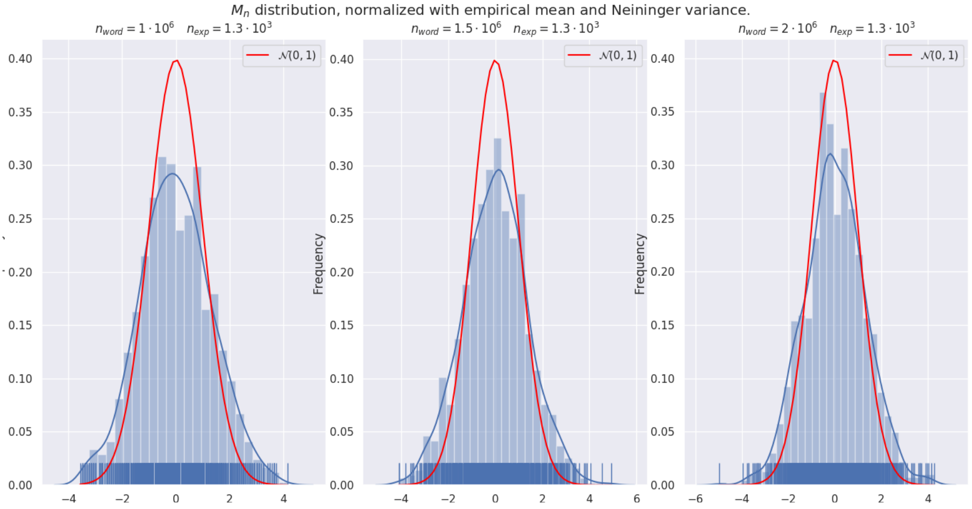
The simulations using this coefficient for the variance are quite good. It also seems that this formula for the variance is equivalent to the one used in the unpublished paper *Probabilistic Analysis of Lempel-Ziv Parsing for Markov Sources* by Leckey, Wormald and Neininger, but our two ways of deriving it differs. Numerical instability might account for the tiny differences found for high $n$ values ($10^7$), although this hasn't been verified.




Difference between standard deviations, $n_{exp}$ = 400
$\Delta\sigma = \sigma - \sigma_{Neininger}$
$\Delta\sigma = \sigma - \sigma_{Szpan}$
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

Now, for some distributions of very long words that were normalized using our theoretical standard deviations, and empirical means. The blue plot is a gaussian fit for the simulation results, which also appear as a blue histogram. The two sets of figures are identical but obtained using different expressions.





## 3 Conclusion

Similar results were obtained for a variety of randomly generated Markov sources, which seem to indicate that this formula for the variance could be proven theoretically correct.

> **Limits of this work**
>
> The figures suffer from imprecision over the computation of the empirical variance : this is due to the difficulties encountered in computing large amounts of long words (of size over $10^6$). The figure in appendix A is an example of this limitation : with only 100 experiments, the empirical variance varies a lot. Possible ideas of improvement might come from parallelization, rewriting functions in a computationnal language such as Julia, or using/devising a datastructure specific to the task of building very long words.

> Another problem is that the space of random Markov chains (here : stochastic matrices of size 2) is not sampled thoroughly. Sampling a small number of Markov chains uniformly according to their entropy might be interesting as a representation of the space, because otherwise it would be hard to sample a large number of stochastic matrices due to the necessity of computing large words for each of them.
>
> Finally, the difference between empirical $\mathrm{V}_n$ and our expression seems to be growing very slowly with $n$. This might be a term that is negligible (*i.e.* of order lower than $n/\log_2^2(n)$), or a small detail in the formula of $\mathrm{V}_n$. For example, the natural base logarithm (in $n/\ln^2(n)$) in the eigenvalue expression works slightly better than the one in base 2 (in $n/\log_2^2(n)$), with the inverse situation happening for the first expression ($\sigma_{\text{Neininger}}$). Anyway, the figures obtained seem to indicate that we are close to the exact solution.

The code for these experiments, with detailed procedures for reproducibility, is hosted on GitHub in one of my private repositories (which I plan to make public eventually). Another (probably bad) way of computing $\lambda(s)$ is in appendix B.

# III Tail symbols analysis

# Covariance asymptotics

## 1 Markov Independent Model

Let M a Markov source and $n$ an integer. The Markov Independent Model (MI) considers $n$ infinite words generated by M. The choice of the starting symbol of each sequence is a parameter of the model. For example, all the sequences might start with the same letter of the alphabet $c \in \mathcal{A}$. Or the first symbol could be initialized using the stationary distribution of the Markov chain related to M.

This is an example for $n = 4$. These are the sequences :

$$\begin{aligned} \mathrm{X}(1) &= 00000\ldots \\ \mathrm{X}(2) &= 1010101\ldots \\ \mathrm{X}(3) &= 1001101\ldots \\ \mathrm{X}(4) &= 001100111\ldots \end{aligned}$$

These sequences are used to build a digital search tree by considering the shortest prefix of each sequence that has not appeared yet in the previously considered sequences.

On our example, it yields the parsed word :

$$()(0)(1)(10)(00)$$

which can be read as the DST :

INSERT TREE

# 2 Tail symbols

## 2.1 Definition

Each of the $n$ sequences possess a tail symbol. For each sequence, its tail symbol is the character that immediately follows the prefix phrase inserted into the DST. Therefore the tail symbol is a specific character of this sequence. If we only have the DST containing the prefix phrases, we cannot recover the tail symbols.
Visually, with the prefix phrases in red and the tail symbol in green :

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

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

## 2.2 Recurrence relation

For $n \geqslant 0$, we have :

$$\boxed{\mathrm{T}_{n+1}^c = \delta_a + \widetilde{\mathrm{T}}_{\mathrm{N}_a}^a + \widetilde{\mathrm{T}}_{\mathrm{N}_b}^b}$$

where :

- $\delta_a = \begin{cases} 1 & \text{if } a \text{ is the tail symbol of the first sequence} \\ 0 & \text{else} \end{cases}$
- $\mathrm{N}_a$ is the random variable giving *the size of the left subtree which contains phrases whose second letter is a*
- $\widetilde{\mathrm{T}}_{\mathrm{N}_a}^a$ is the number of times $a$ is a tail symbol for the sequences that were used to build the subtree with phrases having $a$ as second symbol.
- $\mathrm{T}_0^c$ for all $c$ by convention.

If we take $\{\mathrm{N}_a = k\}$, then $\{\mathrm{N}_b = n - k\}$, then the count of the tail symbols on the left tree is independent of the one on the right tree : *i.e.* these quantities are conditionnaly independent.

# 3 Total path length

## 3.1 Definition

Defining $\mathrm{L}_n^c$ 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.

## 3.2 Recurrence relation

There is another recursive stochastic relation for this quantity, which is, for all $n \geqslant 0$ :

$$\boxed{\mathrm{L}_{n+1}^c = n + \widetilde{\mathrm{L}}_{\mathrm{N}_a}^a + \widetilde{\mathrm{L}}_{\mathrm{N}_b}^b}$$

with the convention that $\mathrm{L}_0^c = 0$ for all $c$.
Same as for the number of tail symbols, this relation is found by considering the DST and its two main subtrees. Except that this time, we count the number of times an

edge contributes to the path length. The root with its two nodes contributes as $n$. The two subtrees contribute respectively for $\widetilde{\mathrm{L}}^a_{\mathrm{N}_a}$ and $\widetilde{\mathrm{L}}^b_{\mathrm{N}_b}$.
It is convenient that these two quantities are conditionnaly independent in the same way as previously seen for the tail symbols.

# 4 Poisson tranform differential equation

## 4.1 Covariance recurrence relation

Using the previous recurrence relations, we have for all $n \geqslant 0$:

$$\mathrm{Cov}(\mathrm{T}^c_{n+1}, \mathrm{L}^c_{n+1}) = \mathrm{Cov}(\delta_a + \widetilde{\mathrm{T}}^a_{\mathrm{N}_a} + \widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, n + \widetilde{\mathrm{L}}^a_{\mathrm{N}_a} + \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})$$

Since the covariance is a bilinear function which is equal to zero if its two terms are independent or if one is constant, we can ignore the term $n$ and expand this quantity into six terms:

$$\begin{aligned}\mathrm{Cov}(\mathrm{T}^c_{n+1}, \mathrm{L}^c_{n+1}) &= \mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) \\ &+ \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})\end{aligned}$$

Since $\delta_a$ is given by the tail symbol of the first sequence, which is independent from the rest of the process: $\mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) = \mathrm{Cov}(\delta_a, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) = 0$
Now, it is not obvious if the pairs $(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b})$ and $(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a})$ are independent or uncorrelated, because the random variable $\mathrm{N}_a$ is not fixed. However they are conditionnaly independent, therefore:

$$\begin{aligned}\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b}) &= \sum_{k=0}^{n} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^b_{\mathrm{N}_b} \,|\, \mathrm{N}_a = k)\mathrm{P}(\mathrm{N}_a = k) \\ &= \sum_{k=0}^{n} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_k, \widetilde{\mathrm{L}}^b_{n-k})\mathrm{P}(\mathrm{N}_a = k) \\ &= \sum_{k=0}^{n} 0 \cdot \mathrm{P}(\mathrm{N}_a = k) \\ &= 0\end{aligned}$$

Samely, $\mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) = 0$. Yielding:

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

## 4.2 Poisson transform

Defining

$$\boxed{\mathrm{C}_c(z) = \sum_{n \geqslant 0} \mathrm{Cov}(\mathrm{T}^c_n, \mathrm{L}^c_n)\frac{z^n}{n!}\mathrm{e}^{-z}}$$

Computing, with $p = \mathrm{P}(a|c)$ and $q = 1 - p$:

$$\begin{aligned}\sum_{n \geqslant 0} \mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a})\frac{z^n}{n!}\mathrm{e}^{-z} &= \sum_{n \geqslant 0}\sum_{k=0}^{n} \mathrm{P}(\mathrm{N}_a = k)\,\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a}, \widetilde{\mathrm{L}}^a_{\mathrm{N}_a} \,|\, \mathrm{N}_a = k)\frac{z^n}{n!}\mathrm{e}^{-z} \\ &= \sum_{n \geqslant 0}\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}\,\mathrm{Cov}(\widetilde{\mathrm{T}}^a_k, \widetilde{\mathrm{L}}^a_k)\frac{z^n}{n!}\mathrm{e}^{-z}\end{aligned}$$

In this case, $\widetilde{\mathrm{T}}^a_k$ and $\mathrm{T}^a_k$ as well as $\widetilde{\mathrm{L}}^a_k$ and $\mathrm{L}^a_k$ have the same distribution, hence :

$$\begin{aligned}
\sum_{n\geqslant 0}\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a},\widetilde{\mathrm{L}}^a_{\mathrm{N}_a})\frac{z^n}{n!}\mathrm{e}^{-z} &= \sum_{n\geqslant 0}\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\,\mathrm{Cov}(\mathrm{T}^a_k,\mathrm{L}^a_k)\frac{z^n}{n!}\mathrm{e}^{-z}\\
&= \sum_{n\geqslant 0}\sum_{k=0}^{n}\left(\frac{(zp)^k}{k!}\,\mathrm{Cov}(\mathrm{T}^a_k,\mathrm{L}^a_k)\mathrm{e}^{-zp}\right)\left(\frac{(zq)^{n-k}}{(n-k)!}\mathrm{e}^{-zq}\right)\\
&= \underbrace{\left(\sum_{n\geqslant 0}\frac{(zp)^n}{n!}\,\mathrm{Cov}(\mathrm{T}^a_n,\mathrm{L}^a_n)\mathrm{e}^{-zp}\right)}_{=\mathrm{C}_a(zp)}\underbrace{\left(\sum_{n\geqslant 0}\frac{(zq)^n}{n!}\mathrm{e}^{-zq}\right)}_{=1}\\
&= \mathrm{C}_a(zp)
\end{aligned}$$

A similar computation gives $\sum_{n\geqslant 0}\mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b},\widetilde{\mathrm{L}}^b_{\mathrm{N}_b})\frac{z^n}{n!}\mathrm{e}^{-z} = \mathrm{C}_b(zq)$, this time conditionning on $\mathrm{P}(\mathrm{N}_b = k) = \binom{n}{k}q^k p^{n-k}$.
From what we've seen, when derivating $\mathrm{C}_c(z)$ we get :

$$\begin{aligned}
\partial_z\mathrm{C}_c(z) &= \sum_{n\geqslant 0}\mathrm{Cov}(\mathrm{T}^c_n,\mathrm{L}^c_n)n\frac{z^{n-1}}{n!}\mathrm{e}^{-z} - \mathrm{C}_c(z)\\
&= \sum_{n\geqslant 0}\mathrm{Cov}(\mathrm{T}^c_{n+1},\mathrm{L}^c_{n+1})\frac{z^n}{n!}\mathrm{e}^{-z} - \mathrm{C}_c(z)\\
&= \sum_{n\geqslant 0}\left[\mathrm{Cov}(\widetilde{\mathrm{T}}^a_{\mathrm{N}_a},\widetilde{\mathrm{L}}^a_{\mathrm{N}_a}) + \mathrm{Cov}(\widetilde{\mathrm{T}}^b_{\mathrm{N}_b},\widetilde{\mathrm{L}}^b_{\mathrm{N}_b})\right]\frac{z^n}{n!}\mathrm{e}^{-z} - \mathrm{C}_c(z)\\
&= \mathrm{C}_a(zp) + \mathrm{C}_b(zq) - \mathrm{C}_c(z)
\end{aligned}$$

Finally the equation for $\mathrm{C}_c(z)$ is :

$$\boxed{\partial_z\mathrm{C}_c(z) + \mathrm{C}_c(z) = \mathrm{C}_a(zp) + \mathrm{C}_b(zq)}$$

# IV Optimal parsing

## 1 LZ77 and LZ78 comparison

### 1.1 Practical uses of LZ77

Although it is based on the same approach of identifying previously seen phrases - but storing them in a fixed size dictionary instead of an evergrowing parse tree like in LZ78 - the LZ77 algorithm is the most used in practical implementations of universal compression algorithms.
Let us dive in on a few of the most recent ones :

- DivANS
- zStandard
-

### 1.2

## 2 Observations regarding the results and the proof

**Corrections**

- The formal definition of $\mathrm{D}_n{}^{\mathrm{LZ}}$ in (**3**) is misleading because it contradicts the previous verbal definition. For $w$ a word of size $n$, the formula

$$\frac{1}{\mathrm{M}_n(w)} \sum_{j=0}^{\mathrm{M}_n(w)-1} |u_j{}^{\mathrm{LZ}}|$$

  would rather be the empirical average length of a phrase in the Lempel-Ziv parsing of a word $\overline{\mathrm{D}}_n{}^{\mathrm{LZ}}$, whereas $\mathrm{D}_n{}^{\mathrm{LZ}}$ is used in the rest of the paper as the length of a randomly selected phrase. These two aren't equal : if we build a Lempel-Ziv DST from a word, then $\mathrm{D}_n{}^{\mathrm{LZ}}$ can be seen as the depth of a random node, which is different from the average path length computed on all the nodes.
- In *Remark 2*, I think the definition of $v$ and $t$ should rather be $v = a_{i'} \dots a_i$ and $t = a_{i+1} \dots a_n$.
- The result (**14**) should be an equality, and it is one indeed because of the flexible parsing algorithm. A proof by contradiction can show this. However, since we upperbound $g(j)$ by $+\infty$ in (**23**) there might be no purpose to using (**14**) instead of just (**13**).
- The proof around *Theorem 1* has several flaws. The notation X for a sequence depending on $n$ and not a random variable is misleading. On the other hand, it should appear that both $g$ and $j$ are random variables, as the randomness of $j$ is used in the end of the proof. I wrote some possible definitions here, and applied them to make some computations that seemed otherwise incorrect because of their use of randomness outside of a probability measure (here).
- As for the arguments that link $|\mathrm{L}_{g(j)-k}|$ to $\mathrm{D}_n{}^{\mathrm{LZ}}$, I have indicated how I think they could be developped in this part. These arguments are the most controversial part right now I think.
- *Theorem 2* is false as stated : we proved *Theorem 1* using a random $j$. The randomness remains, so the quantifier 'for any $j < \mathrm{M}_n$' should be removed. This would be true for *Theorem 1* too.
- The proof of *Theorem 2* may stop at (**26**) since we can directly prove this upperbound goes to 0. This yields a tighter upperbound for *Theorem 2*. I detailed this analysis in the last part of this report.
- In that same proof, the step between (**25**) and (**26**) relies heavily on a result from [6]. A bit more context on this result (and why it does apply here) would make things clearer.
- The conclusion claims to use *Cramer's* theorem to link

$$\max_{0 \leqslant k \leqslant g_{\mathrm{W}}(\mathrm{J})} \left\{ \mathrm{L}_{g_{\mathrm{W}}(\mathrm{J})-k} - k \right\}$$

  to $\mathrm{D}_n{}^{\mathrm{FLEX}}$, which is a sum of random variables. Since *Cramer* applies to independent random variables and the lengths of successive phrases are not independent, something must be missing there.

# 3 Definitions

## 3.1 Definitions

**Notations**

These are definitions and notations in order to write the proof of *Theorem 1*.

**Definition 7** *For all $n \in \mathbb{N}$, calling $\Omega_n$ the set of words of length $n$.*

**Definition 8** *Defining $\mathrm{W} \in \Omega_n$ to be a random variable which outputs words of length $n$ from a memoryless source.*

**Definition 9** *Considering $\mathrm{J} \in \mathbb{N}^\star$ to be a random variable which, in the event $\{\mathrm{W} = w\}$, uniformly randomly picks the index of one of the phrases of $w$. The joint law of $\mathrm{J}$ with $\mathrm{W}$ being:*

$$P\mathrm{W} = w, \mathrm{J} = j = \begin{cases} 1/\mathrm{M}_n(w) & \textit{if } j \leqslant \mathrm{M}_n(w) \\ 0 & \textit{else} \end{cases}$$

**Remark 7** *We might choose another randomness for $\mathrm{J}$, but this one seems more natural.*

**Definition 10** *For a given word $w \in \Omega_n$, and for all $i \in \mathbb{N}^\star$, we consider $g_w(i)$ defined by*

$$g_w(i) = f_w(i) + |\mathrm{L}_{f(i)}|$$

*where $f_w(i)$ is the starting index of the $i^{th}$ phrase of the flexible parsing of $w$, and $|\mathrm{L}_{f_w(i)}|$ is the length of the longest greedy phrase given by the Lempel-Ziv parsing of this same word.*

**Definition 11** *We define $g_{\mathrm{W}}(\mathrm{J})$ to be the random variable which outputs $g_w(j)$ during the events $\{\mathrm{W} = w\}$ and $\{\mathrm{J} = j\}$.*

**Definition 12** *For all $k \in \mathbb{N}$, let $\mathrm{L}_{g_{\mathrm{W}}(\mathrm{J})-k}$ be the random variables which gives the $(k+1)^{th}$ possible phrase for the flexible parsing at index $g_{\mathrm{W}}(\mathrm{J})-k$. Its only randomness comes from $\mathrm{W}$ and $\mathrm{J}$. If $i \leqslant 0$, we might assume that $\mathrm{L}_i$ will be the empty word of size 0.*

**Notation 4** *Denoting by*

$$\mathrm{B}^k_{\mathrm{J},\mathrm{W}} = |\mathrm{L}_{g_{\mathrm{W}}(\mathrm{J})-k}|$$

*the length of this $(k+1)^{th}$ candidate.*

We can now study the random variable

$$\max_{0 \leqslant k \leqslant g_{\mathrm{W}}(\mathrm{J})} \left\{\mathrm{B}^k_{\mathrm{J},\mathrm{W}} - k\right\}$$

Given any $(j, w) \in \mathbb{N}^\star \times \Omega_n$, under the events $\{\mathrm{J} = j\}$, $\{\mathrm{W} = w\}$ and $\{\mathrm{J} \leqslant \mathrm{M}_n(\mathrm{W})\}$, this random variable is the length of the $j^{\text{th}}$ phrase of the flexible parsing of the word $w$.

**Definition 13** *Defining the random variable $\mathrm{D}_n^{\mathrm{FLEX}}$ as*

$$\begin{aligned} \mathrm{D}_n^{\mathrm{FLEX}}(w) &= \frac{1}{\mathrm{M}_n(w)} \sum_{j=0}^{\mathrm{M}_n(w)-1} |u_j^{\mathrm{FLEX}}(w)| \\ &= \frac{1}{\mathrm{M}_n(w)} \sum_{j=0}^{\mathrm{M}_n(w)-1} \max_{0 \leqslant k \leqslant g_w(j)} \left\{\mathrm{B}^k_{j,w} - k\right\} \end{aligned}$$

*where $\mathrm{D}_n^{\mathrm{FLEX}}(w)$ is the empirical average of the lengths of the phrases of the flexible parsing of a word $w$, contrary to $\max\limits_{0 \leqslant k \leqslant g_w(\mathrm{J})} \left\{\mathrm{B}^k_{\mathrm{J},w} - k\right\}$, with $\mathrm{J}$ a random variable and $w$ fixed, which is the length of a uniformly randomly selected phrase of the flexible parsing of $w$.*

**Definition 14** *We denote $x_n$ the average value of* ${\mathrm{D}_n}^{LZ}$ *:*

$$x_n = \frac{1}{h}\log_2\left(\frac{nh}{\log_2(n)}\right)$$

**Remark 8** *The random variable* ${\mathrm{D}_n}^{LZ}$ *is the length of a phrase randomly taken from the Lempel-Ziv parsing of a memoryless-generated word. It is not the same as the empirical average length of a phrase, which we can denote, for* $w \in \Omega_n$ *:*

$$\overline{\mathrm{D}_n}^{LZ}(w) = \frac{1}{\mathrm{M}_n(w)} \sum_{j=0}^{\mathrm{M}_n(w)-1} \left|u_j^{FLEX}\right|$$

**Notation 5** *We denote* ${b_n}^{\delta}$ *as*

$${b_n}^{\delta} = x_n + (c_3 x_n)^{\delta}$$

We will study

$$\mathrm{P} \max_{0 \leqslant k \leqslant g_{\mathrm{W}}(\mathrm{J})} \left\{\mathrm{B}^k_{\mathrm{J,W}} - k\right\} > {b_n}^{\delta}$$

### 3.2 Clues in proving the result

Let $k \in \mathbb{N}$. Currently, the proof to show this stands on three arguments:

(1) The first is that $\mathrm{L}_{g_{\mathrm{W}}(\mathrm{J})-k}$ is a random phrase from the Lempel-Ziv parsing of a word of length N, where $\mathrm{N} \leqslant g_{\mathrm{W}}(\mathrm{J}) \leqslant n$. In the event $\{\mathrm{N} = n'\}$, we consider ${\mathrm{D}_{n'}}^{\mathrm{LZ}}$.

(2) The second, is that $|\mathrm{L}_{g_{\mathrm{W}}(\mathrm{J})-k}|$ can therefore be considered the same as ${\mathrm{D}_{\mathrm{N}}}^{\mathrm{LZ}}$ *i.e* at least equal in law.

(3) The third is that, for all $n' \leqslant n$, ${\mathrm{D}'_n}^{\mathrm{LZ}} \leqslant {\mathrm{D}_n}^{\mathrm{LZ}}$.

Although they seem generally true, there are different problems with each of these arguments:

- N isn't clearly established, so ${\mathrm{D}_{\mathrm{N}}}^{\mathrm{LZ}}$ isn't really known.
- If we can identify N and, let's say, condition our probability with $\{\mathrm{N} = n'\}$, it is not obvious that the choice of a phrase at position $g_{\mathrm{W}}(\mathrm{J}) - k$ knowing $\{\mathrm{N} = n'\}$ is the same as the uniform choice that operates when choosing a random phrase from a word of size $n'$, in ${\mathrm{D}_{n'}}^{\mathrm{LZ}}$.
- As for (3), this result is true on average, but not in all cases. Indeed, since ${\mathrm{D}_n}^{\mathrm{LZ}}$ (resp. ${\mathrm{D}'_n}^{\mathrm{LZ}}$) is concentrated around $x_n$ (resp. $x'_n$), and $x'_n < x_n$ since $n' < n$, we can show that this result holds with high probability. To write the proof, we may condition using events of the type

  $$\{|{\mathrm{D}_n}^{\mathrm{LZ}} - x_n| \leqslant {k_n}^{(1)} v_n\}$$

  and $$\{|{\mathrm{D}_{n'}}^{\mathrm{LZ}} - x_{n'}| \leqslant {k_n}^{(2)} v_{n'}\}$$

  where

  $$v_n = \sqrt{\log(nh/\log(n))}$$

  A sketch of the proof is to apply concentration inequalities to these events while picking $({k_n}^{(1)}, {k_n}^{(2)})$ such that

  $${k_n}^{(1)}\, v_n + {k_n}^{(2)}\, v_{n'} < x_n - x_{n'}$$

  and having ${k_n}^{(1)}$ and ${k_n}^{(2)}$ go to $+\infty$ for $n$ going to $+\infty$ in order for the upper-bound probability to converge to zero.

### 3.3 Working with these probabilistic objects

**Computations**

With these definitions, we can do the formal computations at the beginning of the proof of *Theorem 1.* By conditionning on W and J, we obtain

$$\begin{aligned}
\mathrm{P}\max_{0\leqslant k\leqslant g_{\mathrm{W}}(\mathrm{J})}\left\{\mathrm{B}^k_{\mathrm{J,W}}-k\right\}>{b_n}^\delta &= \sum_{w\in\Omega_n}\sum_{j\in\mathbb{N}^\star}\mathrm{P}\max_{0\leqslant k\leqslant g_{\mathrm{W}}(\mathrm{J})}\left\{\mathrm{B}^k_{\mathrm{J,W}}-k\right\}>{b_n}^\delta,\mathrm{W}=w,\mathrm{J}=j\\
&= \sum_{w\in\Omega_n}\sum_{j\in\mathbb{N}^\star}\mathrm{P}\bigcup_{k=0}^{g_w(j)}\left\{\mathrm{B}^k_{w,j}>k+{b_n}^\delta\right\},\mathrm{W}=w,\mathrm{J}=j\\
&\leqslant \sum_{w\in\Omega_n}\sum_{j\in\mathbb{N}^\star}\sum_{k=0}^{g_w(j)}\mathrm{P}\mathrm{B}^k_{w,j}>k+{b_n}^\delta,\mathrm{W}=w,\mathrm{J}=j\\
&\leqslant \sum_{w\in\Omega_n}\sum_{j\in\mathbb{N}^\star}\sum_{k=0}^{+\infty}\mathrm{P}\mathrm{B}^k_{w,j}>k+{b_n}^\delta,\mathrm{W}=w,\mathrm{J}=j\\
&= \sum_{k=0}^{+\infty}\sum_{w\in\Omega_n}\sum_{j\in\mathbb{N}^\star}\mathrm{P}\mathrm{B}^k_{w,j}>k+{b_n}^\delta,\mathrm{W}=w,\mathrm{J}=j\\
&= \sum_{k=0}^{+\infty}\mathrm{P}\mathrm{B}^k_{\mathrm{W,J}}>k+{b_n}^\delta
\end{aligned}$$

For all $k\in\mathbb{N}$, we may now prove that

$$\mathrm{P}\mathrm{B}^k_{\mathrm{J,W}}>k+{b_n}^\delta\leqslant\mathrm{P}{\mathrm{D}_n}^{\mathrm{LZ}}>k+{b_n}^\delta$$

## 4 Proof of an upperbound for the paper

**Upperbound proof**

This is a proof that the upperbound in (**26**) goes to 0. Assuming

$$\sum_{k=0}^{+\infty}\mathrm{PD}_n^{\mathrm{LZ}}(\mathrm{W})>k+b_n^\delta\leqslant\mathrm{A}\alpha^{(c_3x_n)^{\delta-1/2}}\sum_{i=0}^{+\infty}\alpha^{i/\sqrt{c_3x_n}}\tag{\textbf{26}}$$

We can prove directly that the upperbound term goes to 0 for $n$ going to $+\infty$, without resorting to another majoration. Since $\sqrt{c_3x_n}$ goes to infinity for $n\to+\infty$, we can pick $n$ such that $\sqrt{c_3x_n}>1$. Therefore $\alpha^{1/\sqrt{c_3x_n}}<1$ and the geometric sum gives

$$\sum_{i=0}^{+\infty}\alpha^{i/\sqrt{c_3x_n}}=\frac{1}{1-\alpha^{1/\sqrt{c_3x_n}}}$$

It remains to prove that

$$\lim_{n\to+\infty}\frac{\alpha^{(c_3x_n)^{\delta-1/2}}}{1-\alpha^{1/\sqrt{c_3x_n}}}=0$$

which is done by using L'Hospital's rule. Define :

$$f\colon\begin{cases}\mathbb{R}_+^*\longrightarrow\mathbb{R}\\ x\longmapsto\alpha^{x^{\delta-1/2}}\end{cases}\qquad\text{and}\qquad g\colon\begin{cases}\mathbb{R}_+^*\longrightarrow\mathbb{R}\\ x\longmapsto 1-\alpha^{\frac{1}{\sqrt{x}}}\end{cases}$$

Let $x\in\,]\,0\,;+\infty\,[$. Derivating yields :

$$f'(x) = \ln\alpha \left(\delta - \frac{1}{2}\right) x^{\delta-3/2} f(x) \qquad \text{and} \qquad g'(x) = \ln\alpha \, \frac{1}{2x\sqrt{x}} \, \alpha^{\frac{1}{\sqrt{x}}}$$

We proceed to show that $\dfrac{f'(x)}{g'(x)}$ goes to 0 as $x$ goes to $+\infty$. We have

$$\frac{f'(x)}{g'(x)} = \frac{\left(\delta - \frac{1}{2}\right) x^{\delta-3/2} \alpha^{x^{\delta-1/2}}}{\frac{1}{2x\sqrt{x}} \alpha^{\frac{1}{\sqrt{x}}}} = 2\left(\delta - \frac{1}{2}\right) x^{\delta} \cdot \frac{\alpha^{x^{\delta-1/2}}}{\alpha^{\frac{1}{\sqrt{x}}}}$$

Since $\alpha^{\frac{1}{\sqrt{x}}} \underset{x\to+\infty}{\longrightarrow} 1$, we are left to study $x^{\delta}\alpha^{x^{\delta-1/2}}$.

Writing $$x^{\delta}\alpha^{x^{\delta-1/2}} = \mathrm{e}^{\delta \log x + \log\alpha \cdot x^{\delta-1/2}}$$

and taking the log, since $\delta > 1/2$ and $\log\alpha < 0$ we see that

$$\delta \log x + \log\alpha \cdot x^{\delta-1/2} \underset{x\to+\infty}{\longrightarrow} -\infty$$

Therefore $x^{\delta}\alpha^{x^{\delta-1/2}} \underset{x\to+\infty}{\to} 0$, and given that $f(0) = g(0) = 0$, L'Hospital's rule applies, proving that $\dfrac{f(x)}{g(x)} \underset{x\to+\infty}{\longrightarrow} 0$.

# Références

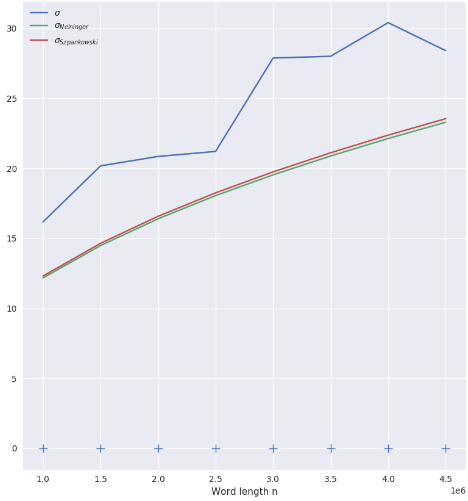
[1] Daniel S Hirschberg, Michael J Pazzani, and Kamal M Ali. *Average Case Analysis of.* Number 1990. 1994.

[2] Yonghui Wu, Stefano Lonardi, and Wojciech Szpankowski. Error-resilient LZW data compression. *Data Compression Conference Proceedings*, pages 193–202, 2006.

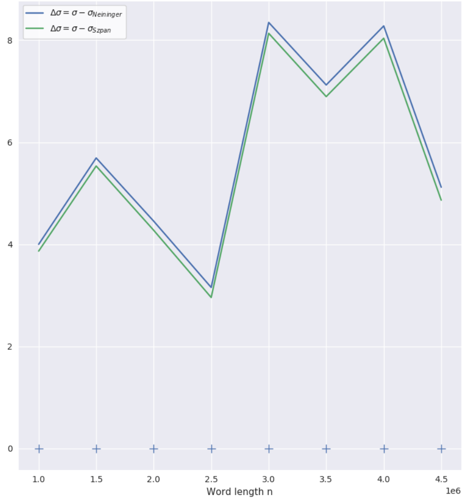
# Appendices

## A Much longer words

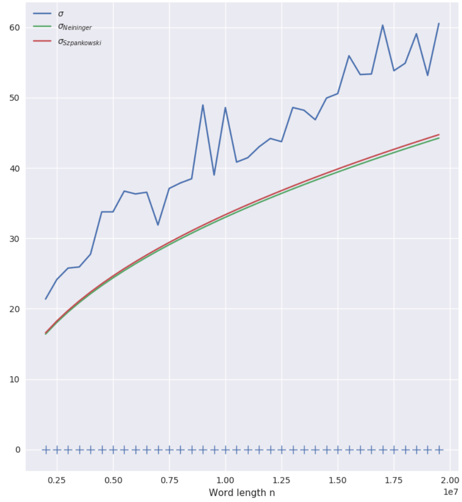
Empirical standard deviation ($\sigma$) and theoretical ones ($\sigma_{Neininger}$, $\sigma_S$), $n_{exp}$ = 200



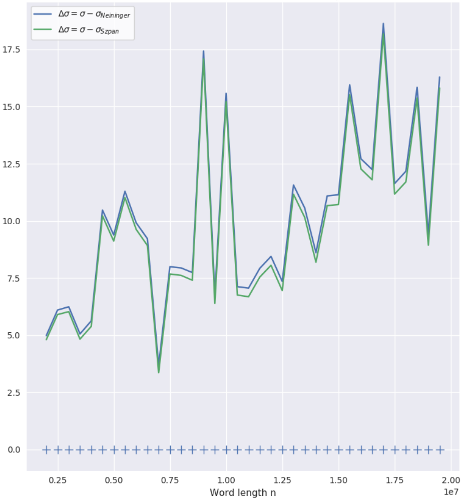
Difference between standard deviations, $n_{exp}$ = 200



Empirical standard deviation ($\sigma$) and theoretical ones ($\sigma_{Neininger}$, $\sigma_S$), $n_{exp}$ = 100



Difference between standard deviations, $n_{exp}$ = 100



# B Another (more complicated) computation of $\ddot{\lambda}(-1)$

This expression gives the sames numerical results as the first one, but is more complex to compute for no apparent gain other than having yet another similar way of computing $\ddot{\lambda}(-1)$. Computing $\delta(s)$, a complex root of $\Delta(s)$, writing $\Delta$ as :

$$\begin{aligned}\Delta = {} & \underbrace{p_{00}^{-2\,\mathrm{Re}\,(s)}\cos(2\ln(p_{00})\,\mathrm{Im}\,(s))}_{a_0(s)} \\ & + \underbrace{p_{11}^{-2\,\mathrm{Re}\,(s)}\cos(2\ln(p_{11})\,\mathrm{Im}\,(s))}_{a_1(s)} \\ & \underbrace{-2(p_{00}\,p_{11})^{-\,\mathrm{Re}\,(s)}\cos(\ln(p_{00}\,p_{11})\,\mathrm{Im}\,(s))}_{a_2(s)} \\ & + \underbrace{4(p_{01}\,p_{10})^{-\,\mathrm{Re}\,(s)}\cos(\ln(p_{01}\,p_{10})\,\mathrm{Im}\,(s))}_{a_3(s)} \\ & + i\,\mathrm{Im}\,(\Delta)\end{aligned}$$

where $\mathrm{Im}\,(\Delta) = b_0(s) + b_1(s) + b_2(s) + b_3(s)$, with each $b_i(s)$ being the same term as $a_i(s)$ with cos replaced by sin. Writing

$$\Delta = \alpha(s) + i\beta(s)$$

and searching for $\delta = x(s) + iy(s)$, meaning that

$$\begin{cases} x^2 - y^2 & = \alpha \\ 2\,x\,y & = \beta \\ x^2 + y^2 & = \sqrt{\alpha^2 + \beta^2} \end{cases}$$

This yields

$$\begin{cases} x & = \pm\sqrt{\frac{1}{2}(\sqrt{\alpha^2+\beta^2} + \alpha)} \\ y & = \pm\sqrt{\frac{1}{2}(\sqrt{\alpha^2+\beta^2} - \alpha)} \end{cases}$$

and since $2xy = \beta$, there is $\varepsilon \in \{-1, 1\}$ such that

$$\delta = \pm(x + \mathrm{i}\varepsilon y)$$

so
$$\lambda(s) = \frac{{p_{00}}^{-s} + {p_{11}}^{-s} \pm (x + \mathrm{i}\varepsilon y)}{2}$$

i.e.
$$\ddot{\lambda}(-1) = \frac{p_{00}\ln^2(p_{00}) + p_{11}\ln^2(p_{11}) \pm (\ddot{x}(-1) + \mathrm{i}\varepsilon\ddot{y}(-1))}{2}$$

where we'll have to find what is $\varepsilon$ and which sign to pick.
But first, computing the derivatives of $x(s) = \sqrt{f(s)}$ :

$$\dot{x}(s) = \frac{f'(s)}{2x(s)}$$

and
$$\ddot{x}(s) = \frac{f''(s)x(s) - f'(s)\cdot\dfrac{f'(s)}{2x(s)}}{2x^2(s)}$$

and then computing $f(s)$ :

$$f(s) = \frac{1}{2}(\sqrt{\alpha^2+\beta^2}+\alpha)$$

$$f'(s) = \frac{1}{2}\left[\frac{\overbrace{\dot\alpha\alpha+\dot\beta\beta}^{\gamma(s)}}{\underbrace{\sqrt{\alpha^2+\beta^2}}_{\kappa(s)}}+\dot\alpha\right]$$

with $$\dot\alpha = \dot{a}_0+\dot{a}_1+\dot{a}_2+\dot{a}_3$$

As for $f''(s)$, it is $$f''(s) = \frac{1}{2}\left[\frac{\dot\gamma(s)\kappa(s)-\gamma(s)\dot\kappa(s)}{\kappa^2(s)}+\ddot\alpha(s)\right]$$

with $$\dot\gamma(s) = \ddot\alpha\alpha+\dot\alpha^2+\ddot\beta\beta+\dot\beta^2$$

$$\dot\kappa(s) = \frac{2\alpha\dot\alpha+2\beta\dot\beta}{2\sqrt{\alpha^2+\beta^2}}$$

Derivating according to $s$ amounts to derivating according to $\mathrm{Re}\,(s)$, so in $s=-1$ :

$$\dot\alpha(-1) = -2\ln p_{00}a_0(-1)-2\ln p_{11}a_1(-1)-\ln q_0a_2(-1)-\ln q_1a_3(-1)$$

and $$\ddot\alpha(-1) = 4\ln^2 p_{00}a_0(-1)+4\ln^2 p_{11}a_1(-1)+\ln^2 q_0a_2(-1)+\ln^2 q_1a_3(-1)$$

At this point we have fully determined $\ddot{x}(s)$, and we realize two things :

1. In $s=-1$, since $\mathrm{I}m(-1)=0$ and because of the sinus function, all the $\beta$ terms, including derivatives, are equal to 0. This will simplify the expression for $\ddot{x}(-1)$.

2. Furthermore, it also means that $\ddot{y}(-1)=0$, so

$$\boxed{\ddot\lambda(-1) = \frac{p_{00}\ln^2(p_{00})+p_{11}\ln^2(p_{11})+\ddot{x}(-1)}{2}}$$

   where the $+$ comes from the fact that $\lambda(s)$ is the highest eigenvalue (and $\ddot{x}(-1)>0$, so by continuity the expression around $s=-1$ retained the same sign)

The final expression of $\ddot\lambda$ (as well as $\dot\lambda(-1)$)) can be fully expressed with $\alpha(-1), \dot\alpha(-1)$ and $\ddot\alpha(-1)$. I empirically verified that $\dot\lambda(-1)=h$, and the final result is the same as with the first method of computation.