Numerical simulations of LZ78 for Markovian sources

Simulation

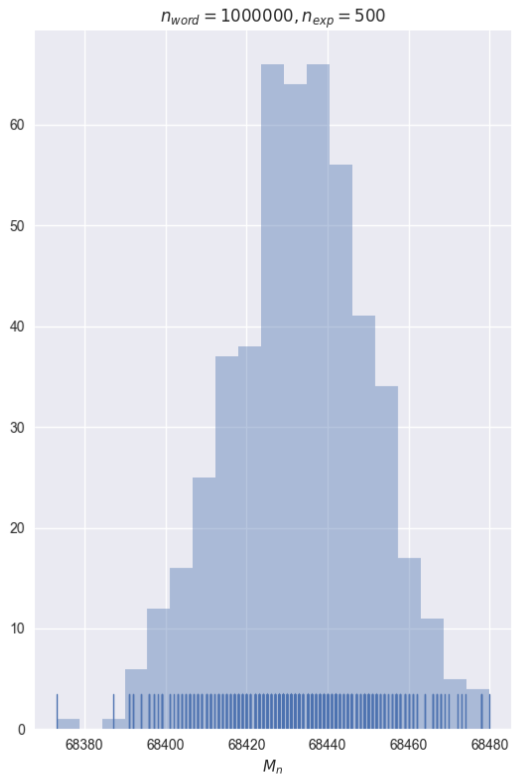
This document presents the different graphics I obtained during the following experimental process:

- Generating a random Markov chain of size 2 of matrix

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

- Generating $n_{\text{exp}} \sim 500$ words of length n (or n_{word}), with $n \leqslant 10^6$
- Applying LZ78 on each of these words to estimate, for each n, the number of phrases M_n. A simple histogram of these values can be seen in figure 1.
- From this sampling of the random variable M_n and other parameters such as the entropy of the Markov chain, computing
  - the empirical mean (μ) and the empirical variance (σ^2)
  - different theoretical expressions of the mean and variance
- Using these expressions to standardize M_n in different ways, plotting
  - the probability distribution of M_n (standardized)
  - the cumulative distribution function of M_n (standardized)
- Finally, comparing the different theoretical expressions for the mean and variance by plotting their differences for a large range of values of n, and a constant number of experiments n_{exp}.

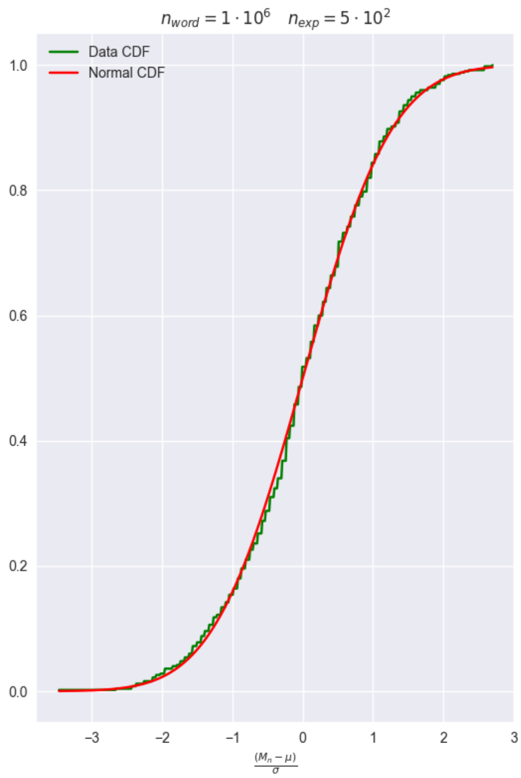
This histogram represents the counts of the different values taken by M_n for $n = 10^6$. Each tick on the x-axis is a data point.



Empirical normalization

Using the empirical mean (μ) and variance (σ^2) of the dataset to normalize 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 :

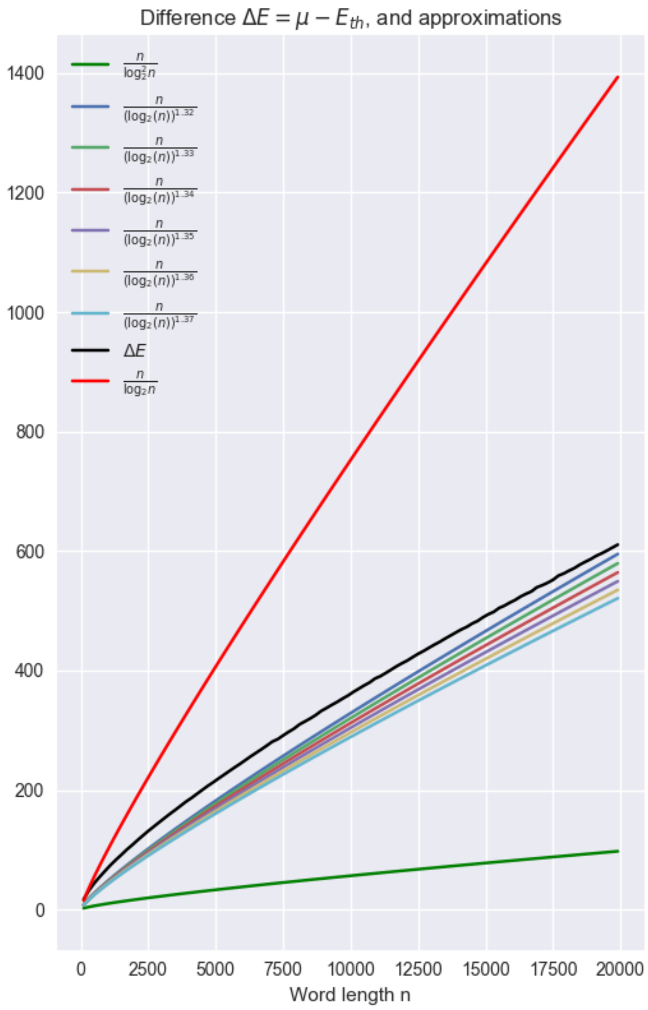


Theoretical mean

I also tried to normalize M_n using theoretical expressions of the mean and variance. For the mean, the first order expression

$$\mathrm{E}_n \sim \frac{nh}{\log_2(n)}$$

is, under $n \leqslant 10^6$, not sufficient to center the distribution. I conducted a numerical analysis of the difference between this expression and the empirical mean for growing values of n. In particular, here is how their difference, in black, compares with different approximation functions



This is not troubling as it was already predicted in the formula :

$$\mathrm{E}_n = \frac{nh}{\log_2(n)} + \mathcal{O}\left(\frac{n}{\log_2(n)}\right)$$

Theoretical variance

For the variance, I tried to use the expression of $\dfrac{\mathrm{H}^3\sigma^2}{n\log_2^2(n)}$ from K. Lecket, N. Wormald and R. Neininger's paper *Probabilistic Analysis of Lempel-Zil Parsing for Markov Sources* :

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

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

> The first term in the squared part of σ_i^2 accounts for the expression of the variance for memoryless sources :
>
> $$\begin{aligned} p_{i0}\,p_{i1}\log_2^2\left(\frac{p_{i0}}{p_{i1}}\right) &= p_{i0}\log_2^2(p_{i0}) + p_{i1}\log_2^2(p_{i1}) - (-p_{i0}\log_2(p_{i0}) - p_{i1}\log_2(p_{i1}))^2 \\ &= h_2 - h^2 \end{aligned}$$

It seems, from simulations, that this variance is too small and doesn't catch up with the empirical variance. Here is how they compare when plotted together :