Numerical simulations of LZ78 for Markovian sources

Simulation

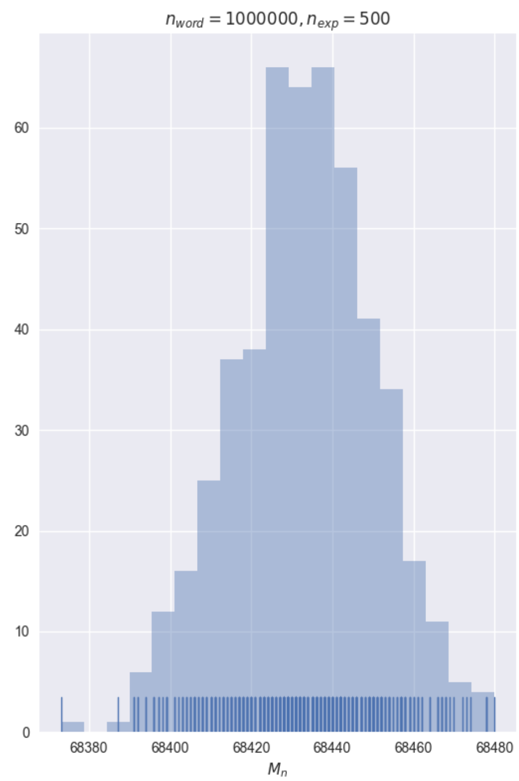
This document presents the different graphics I obtained during the following experimental process :

- Generating a random Markov chain of size 2 of matrix

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

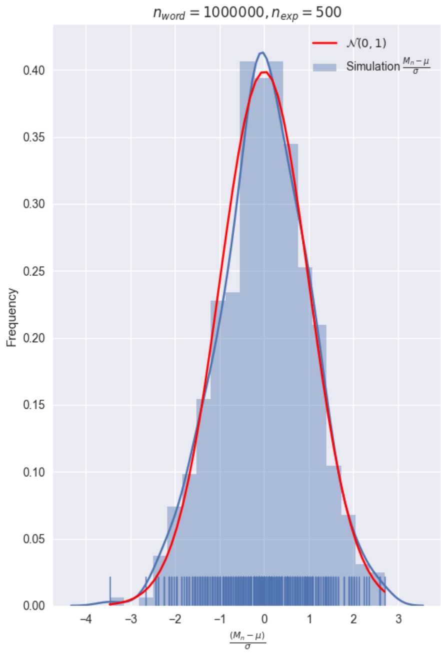
- Generating $n_{\text{exp}} \sim 10^3$ words of length n (or n_{word}), with $n \leqslant 10^6$
- Applying LZ78 on each of these words to estimate, for each n, the number of phrases M_n. A simple histogram of these values can be seen in figure 1.
- From this sampling of the random variable M_n and other 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.

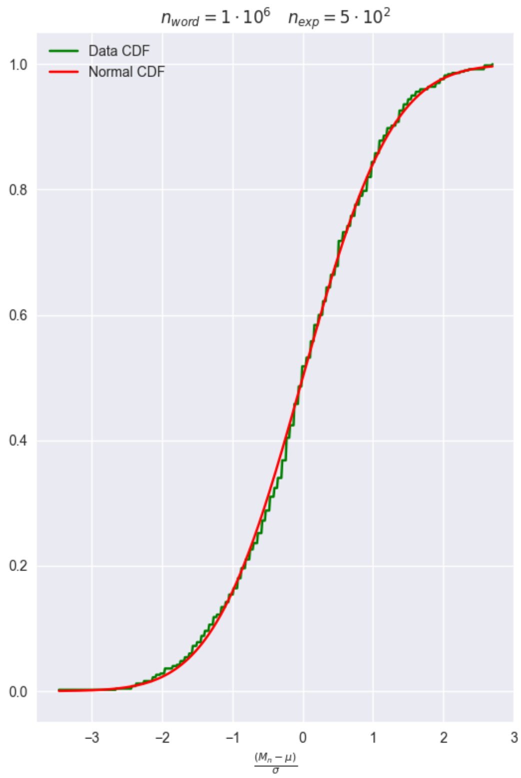


Empirical normalization

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 :



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

1 A first expression for the variance

As it will be used in the next section, this is the detail of the expression

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

from K. Lecket, N. Wormald and R. Neininger's paper *Probabilistic Analysis of Lempel-Ziv 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 $$\mathrm{H}_i = -p_{i0}\log_2(p_{i0}) - p_{i1}\log_2(p_{i1}) \qquad \mathrm{H} = \pi_0\mathrm{H}_0 + \pi_1\mathrm{H}_1$$

2 Variance candidate using the Frobenius eigenvalue of $\mathrm{P}(s)$

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

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

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 χ the characteristic polynomial of $\mathrm{P}(s)$, and Δ 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 $$\dot{\lambda}(s) = \frac{1}{2}\left(-\ln p_{00}\,{p_{00}}^{-s} - \ln p_{11}\,{p_{11}}^{-s} + \frac{\Delta'(s)}{2\sqrt{\Delta(s)}}\right)$$

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

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

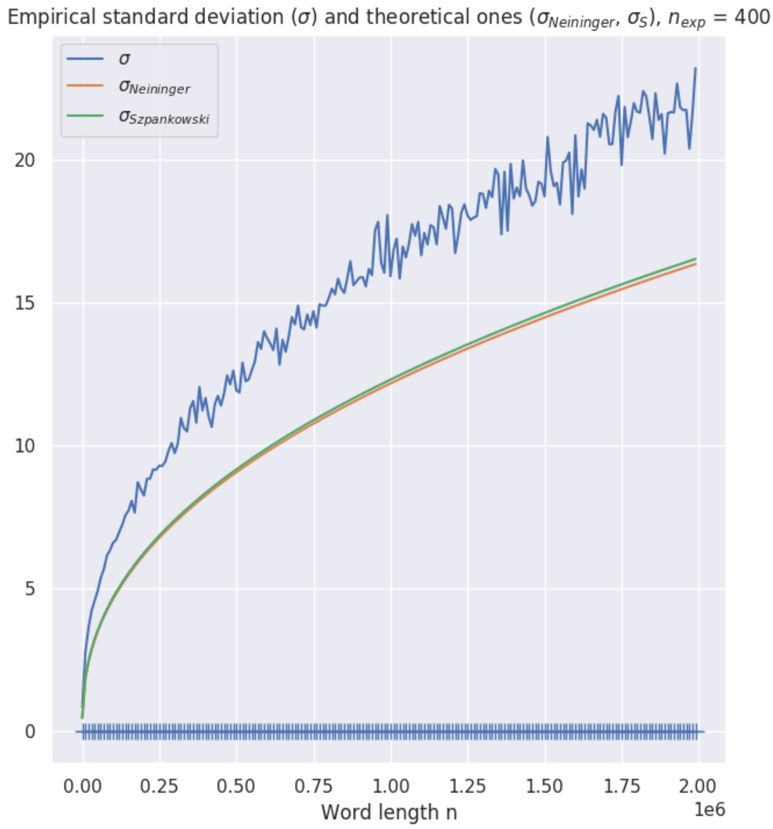
allowing to compute $\dot{\lambda}(-1)$. Numerically, we verified that $\dot{\lambda}(-1) = h$. Derivating again yields

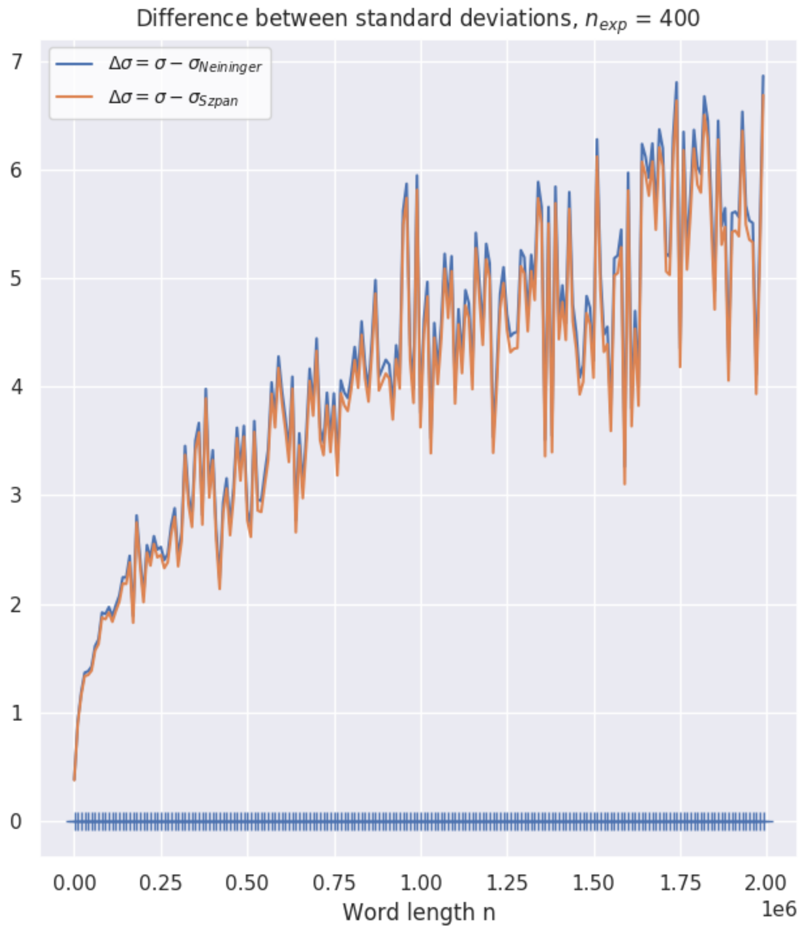
$$\ddot{\lambda}(s) = \frac{1}{2}\left(\ln^2 p_{00}{p_{00}}^{-s} + \ln^2 p_{11}{p_{11}}^{-s} + \frac{\Delta''(s)\sqrt{\Delta(s)} - \Delta'(s)\cdot\Delta'(s)/2\sqrt{\Delta(s)}}{2\Delta(s)}\right)$$

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

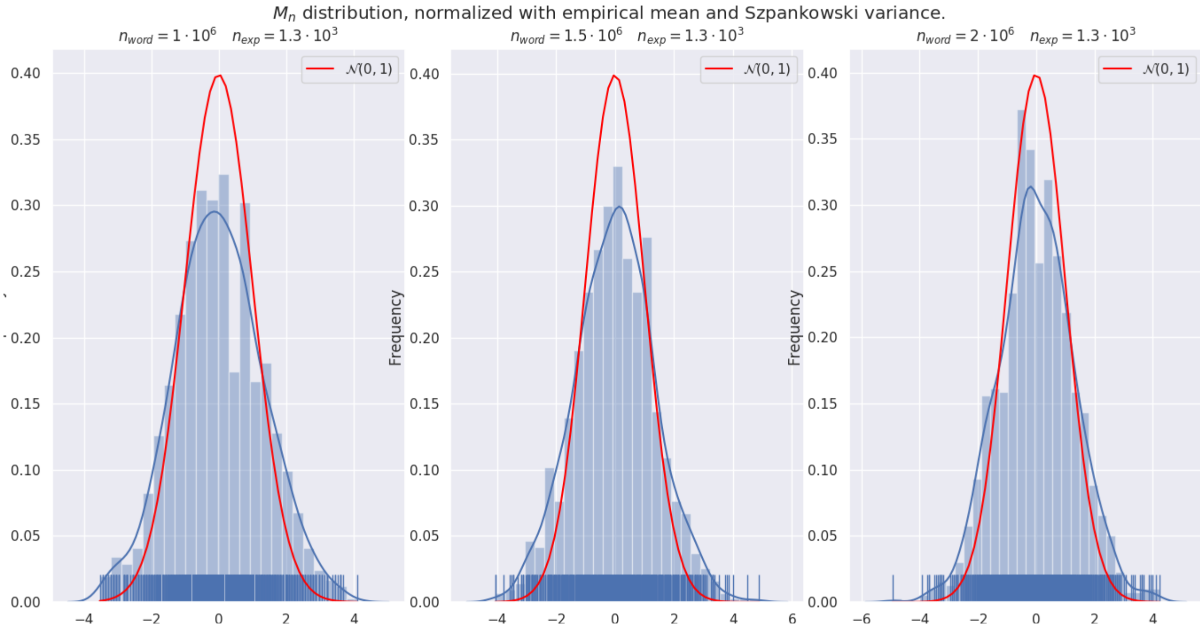
Finally $$\boxed{\ddot{\lambda}(-1) = \frac{1}{2}\left(\ln^2 p_{00}\, p_{00} + \ln^2 p_{11}\, p_{11} + \frac{\Delta''(-1)\sqrt{\Delta(-1)} - \Delta'(-1)^2/2\sqrt{\Delta(-1)}}{2\Delta(-1)}\right)}$$

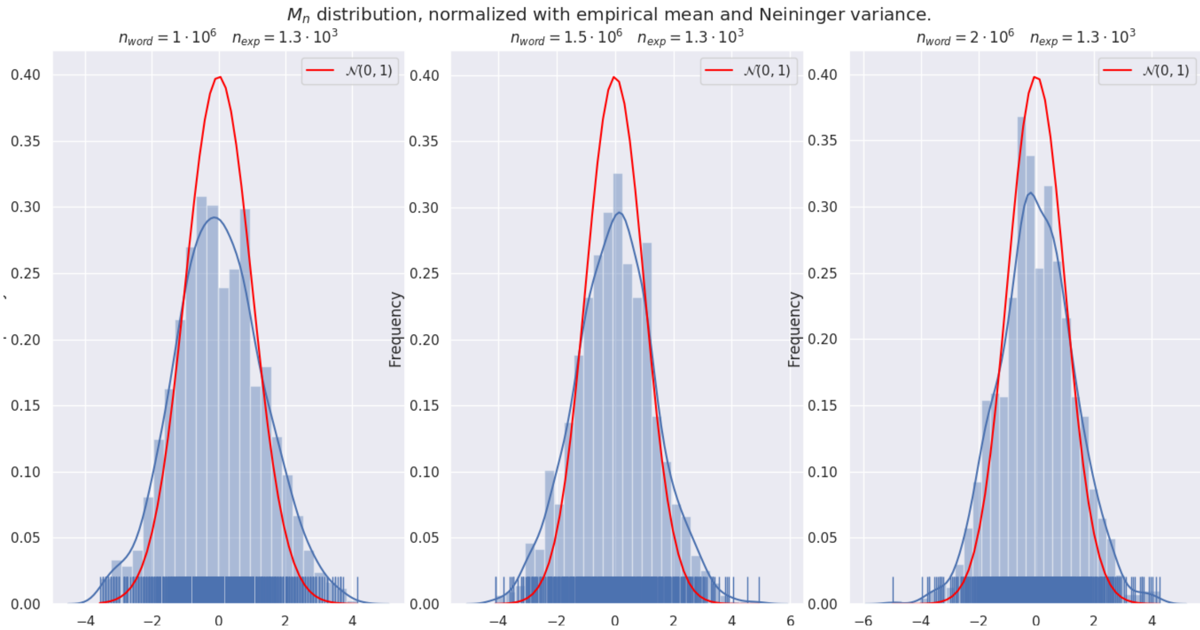
The simulations using this coefficient for the variance are quite good. It also seems that this formula for the variance is equivalent to the one used in the unpublished paper *Probabilistic Analysis of Lempel-Ziv Parsing for Markov Sources* by Leckey, Wormald and Neininger, but our two ways of deriving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




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.





I 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). 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, therefore this claim might only seem to hold for some specific Markov chains. 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.

Appendices

A First lambda computation

```
def compute_lambda2(M):

    p00 = M[0, 0]
    p01 = M[0, 1]
    p10 = M[1, 0]
    p11 = M[1, 1]

    q0 = p00 * p11
    q1 = p01 * p10

    Delta = p00 ** 2 + p11 ** 2 - 2.0 * q0 + 4.0 * q1

    sqrt_Delta = p01 + p10

    der_Delta = -2.0 * log(p00) * (p00 ** 2) \
                -2.0 * log(p11) * (p11 ** 2) \
                + 2.0 * log(q0) * q0 \
                - 4.0 * log(q1) * q1

    der2_Delta = 4.0 * (log(p00) ** 2) * (p00 ** 2) \
                 + 4.0 * (log(p11) ** 2) * (p11 ** 2) \
                 - 2.0 * (log(q0) ** 2) * q0 \
                 + 4.0 * (log(q1) ** 2) * q1

    lam = 0.5 * ( p00 + p11 + sqrt_Delta )

    assert(abs(lam - 1) < 1e-8) # verify lambda(-1) is 1

    der_lam = 0.5 * ( - log(p00) * p00 - log(p11) * p11 \
                      + der_Delta / (2 * sqrt_Delta) )

    h = entropy(M)

    assert(abs(der_lam - h) < 1e-6) # verify we find entropy h for der_lambda in -1

    der2_lam = 0.
    der2_lam += p00 * (log(p00)**2)
    der2_lam += p11 * (log(p11)**2)

    snd_part = der2_Delta * sqrt_Delta - (der_Delta ** 2) / (2.0 * sqrt_Delta)
    snd_part /= 2
    snd_part /= Delta

    der2_lam += snd_part
    der2_lam /= 2

    v_coeff = der2_lam - der_lam ** 2

    assert(v_coeff >= 0)    # verify variance is positive

    return (der2_lam - der_lam ** 2)
```

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 Δ 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

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

This yields

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

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 ε 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 β 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.

Références

[1] Jacquet, Szpankowski, Tang, *Average profile of the Lempel-Ziv parsing scheme for a Markovian source*