Optimal parsing of LZ78 for memoryless sources

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.

Notations

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

Definition 1 *For all $n \in \mathbb{N}$, calling Ω_n the set of words of length n.*

Definition 2 *Defining $\mathrm{W} \in \Omega_n$ to be a random variable which outputs words of length n from a memoryless source.*

Definition 3 *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 J with W being :*

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

Remark 1 *We might choose another randomness for J, but this one seems more natural.*

Definition 4 *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 5 *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 6 *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 W and J. If $i \leqslant 0$, we might assume that L_i will be the empty word of size 0.*

Notation 1 *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^{th} phrase of the flexible parsing of the word w.

Definition 7 *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 $\max_{0 \leqslant k \leqslant g_w(\mathrm{J})} \left\{ \mathrm{B}^k_{\mathrm{J},w} - k \right\}$, with J a random variable and w fixed, which is the length of a uniformly randomly selected phrase of the flexible parsing of w.

Definition 8 *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 2 *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 2 *We denote* ${b_n}^\delta$ *as*

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

We will study

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

$$\boxed{\textbf{Computations}}$$

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

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

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 ${\rm D_N}^{\rm LZ}$ isn't really known.
- If we can identify N and, let's say, condition our probability with $\{{\rm N}=n'\}$, it is not obvious that the choice of a phrase at position $g_{\rm W}({\rm J})-k$ knowing $\{{\rm N}=n'\}$ is the same as the uniform choice that operates when choosing a random phrase from a word of size n', in ${{\rm D}_{n'}}^{\rm LZ}$.
- As for (3), this result is true on average, but not in all cases. Indeed, since ${{\rm D}_n}^{\rm LZ}$ (resp. ${{\rm D}'_n}^{\rm LZ}$) 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

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

and

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

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 upperbound probability to converge to zero.

Upperbound proof

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

$$\sum_{k=0}^{+\infty} \mathrm{P}\mathrm{D}_n^{\rm LZ}(\mathrm{W}) > k + b_n^\delta \leqslant \mathrm{A}\alpha^{(c_3 x_n)^{\delta-1/2}} \sum_{i=0}^{+\infty} \alpha^{i/\sqrt{c_3 x_n}} \tag{26}$$

We can prove directly that the upperbound term goes to 0 for n going to $+\infty$, without resorting to another majoration. Since $\sqrt{c_3 x_n}$ goes to infinity for $n \to +\infty$, we can pick n such that $\sqrt{c_3 x_n} > 1$. Therefore $\alpha^{1/\sqrt{c_3 x_n}} < 1$ and the geometric sum gives

$$\sum_{i=0}^{+\infty} \alpha^{i/\sqrt{c_3 x_n}} = \frac{1}{1-\alpha^{1/\sqrt{c_3 x_n}}}$$

It remains to prove that

$$\lim_{n\to+\infty} \frac{\alpha^{(c_3 x_n)^{\delta-1/2}}}{1-\alpha^{1/\sqrt{c_3 x_n}}} = 0$$

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 $\frac{f(x)}{g(x)} \underset{x\to+\infty}{\longrightarrow} 0$.