Limiting Distribution of Lempel Ziv'78 Redundancy

Philippe Jacquet
INRIA,
Rocquencourt,
78153 Le Chesnay Cedex, France
Email: Philippe.Jacquet@inria.fr

Wojciech Szpankowski
Department of Computer Science,
Purdue University,
West Lafayette, IN 47907-2066 U.S.A.,
Email: spa@cs.purdue.edu

***Abstract*—We show that the Lempel Ziv'78 redundancy rate tends to a Gaussian distribution for memoryless sources. We accomplish it by extending findings from our 1995 paper [3]. We present a new simplified proof of the Central Limit Theorem for the number of phrases in the LZ'78 algorithm. As in our 1995 paper, here we first analyze the asymptotic behavior of the total path length in a digital search tree (a DST) built from independent sequences. Then we present simplified proofs and extend our analysis of LZ'78 algorithm to include new results on the convergence of moments, moderate and large deviations, and redundancy analysis.**

I. Introduction

The Lempel-Ziv compression algorithm [12] is a universal compression scheme. It partitions the text to be compressed into consecutive phrases such that the next phrase is the unique largest prefix of the uncompressed text not seen before in the compressed text. The compression code for a word w over the alphabet $\mathcal{A}$ we denote as $C(w)$. It is known that for a large class of sources the average compression rate $\rho(w) = \frac{|C(w)|}{|w|}$ tends to the source entropy rate h when $|w| \to \infty$. Our goal is to prove that the redundancy rate $r(w) = \rho(w) - h$ tends in probability and in moments to a normal distribution. In particular, we prove that

$$\mathbf{E}(r(w)) = O\left(\frac{1}{\log n}\right), \qquad \mathrm{Var}(r(w)) = O\left(\frac{1}{n}\right)$$

when $|w| \to \infty$.

It is convenient to organize the phrases (dictionary) of the Lempel-Ziv scheme in a *digital search tree* (DST) [6], [11] which is really a parsing tree. The root represents an empty phrase. The first phrase is the first symbol, say "a" which is stored in a node appended to the root. The next phrase is either "aa" stored in another node that branches out from the node containing the first phrase "a" or a new symbol that is stored in a node attached to the root. This process repeats recursively until the text is parsed into full phrases (last incomplete phrase is ignored). A detailed description can be found in [3], [11].

Let a text w be generated over an alphabet $\mathcal{A}$, and let $\mathcal{T}(w)$ be the associated digital search tree constructed by the algorithm. Each node in $\mathcal{T}(w)$ corresponds to a phrase in the parsing algorithm. Let $L(w)$ be the (total) path length in $\mathcal{T}(w)$, that is, the sum of all paths from the root to all nodes. We should have $L(w) = |w|$ (if all phrases are full). If we know the order of nodes creation in the tree $\mathcal{T}(w)$, then we can reconstruct the original text w.

The compression code $C(w)$ is a description of $\mathcal{T}(w)$, node by node in the order of creation; each node being identified by a pointer to its parent node in the tree and the symbol that labels the edge linking it to the parent node. The pointer to the kth node requires at most $\lceil \log k \rceil$ nats, and the next symbol costs $\lceil \log |\mathcal{A}| \rceil$ nats. We just assume that the total cost is $\lceil \log(k|\mathcal{A}|) \rceil$. The compressed code length is

$$|C(w)| = \sum_{k=1}^{M(w)} \lceil \log(k|\mathcal{A}|) \rceil, \tag{1}$$

where $M(w)$ is the number of full phrases needed to parse w. Clearly, $M(w)$ is also the number of nodes in the associated tree $\mathcal{T}(w)$. Notice that the code is self-consistent and does not need *a priori* knowledge of the text length, since the length is a simple function of the nodes sequence. We conclude from (1) that

$$|C(w)| \le M(w)\left(\log(M(w)) + \lceil \log(|\mathcal{A}|) \rceil\right). \tag{2}$$

In fact, different implementation may add $O(M(w)$ to the code length. Throughout, we shall assume that $|C(w)| = M(w)\left(\log(M(w)) + \log(|\mathcal{A}|)\right)$.

In this paper we study the limiting distribution, large deviations, and moments of the number of phrases $M(w)$ and the redundancy when the text of length $|w| = n$ is generated by a memoryless source. We prove the Central Limit Theorem (CLT) for the number of phrases and establish the LZ'78 code redundancy (excess of the code length over the optimal length). The former result was already proved in our 1995 paper [3] while the latter was presented in [5]. However, the proof of the CLT in our 1995 paper was quite complicated; it involves a generalized analytic depoissonization over convex cones in the complex plane. In this paper we simplified and generalized it to present new comprehensive large deviations results. It should be pointed out that since our 1995 paper [3] no simpler, in fact, no new proof of CLT was presented except the one by Neininger and Ruschendorf [8] but only for *unbiased* memoryless sources (as in [1]). The proof of [8] applies the so called *contraction method* and should generalize to biased memoryless sources.

II. Main Results

Let n be a nonnegative integer. We denote by M_n the number of phrases $M(w)$ when the original text w is of fixed

length n. We shall assume throughout that the text is generated by a memoryless source over $\mathcal{A}$ such that the entropy rate is $h = -\sum_{a\in\mathcal{A}} p_a \log p_a > 0$ where p_a is the probability of symbol a. We also define $h_2 = \sum_{a\in\mathcal{A}} p_a (\log p_a)^2$ and

$$\eta = -\sum_{k\geq 2} \frac{\sum_{a\in\mathcal{A}} p_a^k \log p_a}{1-\sum_{a\in\mathcal{A}} p_a^k} \ . \tag{3}$$

Finally, we introduce two functions

$$\begin{aligned}\beta(m) &= \frac{h_2}{2h} + \gamma - 1 - \eta + \Delta_1(\log m) \\ &\quad + \frac{1}{m}\left(\log m + \frac{h_2}{2h} + \gamma - \eta - \sum_{a\in\mathcal{A}} \log p_a - \frac{1}{2}\right) \\ v(m) &= \frac{m}{h}\left((\frac{h_2}{h^2} - 1)\log m + c_2 + \Delta_2(\log m)\right)\end{aligned}$$

We prove the following theorem which improves our previous result from [3] by adding the convergence of moments.

Theorem 1. *Consider the LZ'78 algorithm over a sequence of length n generated by a memoryless source. Let*

$$\ell(m) = \frac{m}{h}\left(\log m + \beta(m)\right).$$

The number of phrases M_n has mean $\mathbf{E}[M_n]$ and variance $\mathrm{Var}(M_n)$ satisfying for all $\frac{1}{2} < \delta < 1$

$$\begin{aligned}\mathbf{E}(M_n) &= \ell^{-1}(n) + O(n^\delta) \qquad (4)\\ &= \frac{nh}{\log \ell^{-1}(n) + \beta(\ell^{-1}(n))} + O(n^\delta) \sim \frac{nh}{\log n},\\ \mathrm{Var}(M_n) &\sim \frac{v(\ell^{-1}(n))}{(\ell'(\ell^{-1}(n)))^2} \sim \frac{(h_2-h^2)n}{\log^2 n}. \qquad (5)\end{aligned}$$

Furthermore, the normalized number of phrases converges in distribution and moments to the the standard normal distribution $N(0,1)$. More precisely, for any given real x:

$$\lim_{n\to\infty} P(M_n < \mathbf{E}(M_n) + x\sqrt{\mathrm{Var}(M_n)}) = \Phi(x), \tag{6}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2} dt.$$

In addition, for all nonnegative k

$$\lim_{n\to\infty} \mathbf{E}\left(\left(\frac{M_n - \mathbf{E}(M_n)}{\sqrt{\mathrm{Var}(M_n)}}\right)^k\right) = \mu_k \tag{7}$$

where $\mu_k = 0$ for k is odd, and $\mu_k = \frac{k!}{2^{k/2}(\frac{k}{2})!}$ for k even.

We also have large and moderate deviations results. To the best of our knowledge these results are new (see also [3], [7]).

Theorem 2. *Consider the LZ'78 algorithm over a sequence of length n generated by a memoryless source.*
(i) [Large Deviations]. *For all $\frac{1}{2} < \delta < 1$ there exist $\varepsilon > 0$, $B > 0$ and $\beta > 0$ such that for all $y > 0$*

$$P(|M_n - \mathbf{E}(M_n)| > yn^\delta) \leq A\exp(-\beta n^\varepsilon \frac{y}{(1+n^{-\varepsilon}y)^\delta}). \tag{8}$$

(ii) [Moderate Deviation]. *Let $\delta < \frac{1}{6}$ and $A > 0$. There exists $B > 0$ such that for all non-negative real number $x < An^\delta$:*

$$P(|M_n - \mathbf{E}(M_n)| \geq x\sqrt{\mathrm{Var}(M_n)}) \leq Be^{-\frac{x^2}{2}}.$$

As direct consequence of these large deviations result, we conclude that the average compression rate converges to the entropy rate. Furthermore, our large deviation results allow us also to estimate the average redundancy $\mathbf{E}[r_n] = \mathbf{E}[C(w)]/|w| - h$ and its limiting distribution when $|w| = n$.

Corollary 1. *The redundancy rate r_n for sequences of length n satisfies for all $\frac{1}{2} < \delta < 1$:*

$$\begin{aligned}\mathbf{E}(r_n) &= \frac{\mathbf{E}(M_n \log M_n + M_n \log(|\mathcal{A}|))}{n} - h\\ &= h\frac{\log(|\mathcal{A}|) - \beta(\ell^{-1}(n))}{\log \ell^{-1}(n) + \beta(\ell^{-1}(n))} + O(n^{\delta-1})\\ &\sim h\frac{\log(|\mathcal{A}|) - \beta\left(h\frac{n}{\log n}\right)}{\log n},\end{aligned}$$

and

$$\mathrm{Var}(r_n) \sim \frac{(h_2 - h^2)}{n}\ .$$

Furthermore, the limiting distribution $\frac{r_n - \mathbf{E}(r_n)}{\sqrt{\mathrm{Var}(r_n)}}$ converges in distribution and moments to $N(0,1)$.

The redundancy average estimate was first proved in [5], [10] but we provide a new proof. The limiting distribution of the redundancy is new.

III. From Lempel-Ziv to Digital Search Tree

In this section we make a connection between the Lempel-Ziv algorithm and digital search trees using a renewal argument [2].

Our goal is to derive an estimate on the probability distribution of M_n. We assume that our original text is a prefix of an infinite sequence X generated by a memoryless source over the alphabet $\mathcal{A}$. We build a Digital Search Tree (DST) by parsing the infinite sequence X up to the mth phrase. Thus the DST is constructed over m strings (phrases).

Let L_m be the total path length in the associated DST after inserting m (independent) strings. The quantity M_n is exactly the number of strings needed to be inserted to increase the path length of the associated DST to n. This observation leads to the following identity valid for all integers n and m:

$$P(M_n > m) = P(L_m < n)\ . \tag{9}$$

We now use generating functions to find a functional equation for the distribution of L_m. Let $L_m(u) = \mathbf{E}(u^{L_m})$ be the moment generating function of L_m. In the following, $\mathbf{k}$ is a tuple in $\mathbb{N}^{|\mathcal{A}|}$ and k_a for $a \in \mathcal{A}$ is the component of $\mathbf{k}$ for symbol a. Since inserted strings are independent, we conclude that

$$L_{m+1}(u) = u^m \sum_{\mathbf{k}\in\mathbb{N}^{|\mathcal{A}|}} \binom{m}{\mathbf{k}} \prod_{a\in\mathcal{A}} p_a^{k_a} L_{k_a}(u), \tag{10}$$

where $\binom{m}{\mathbf{k}} = \frac{m!}{\prod_{a\in\mathcal{A}} k_a!}$. Next, we introduce the exponential generating function $L(z,u) = \sum_m \frac{z^m}{m!} L_m(u)$ leading to

$$\frac{\partial}{\partial z} L(z,u) = \prod_{a\in\mathcal{A}} L(p_a u z, u) \ . \tag{11}$$

It is clear from the construction that $L(z,1) = e^z$, since $L_m(1) = 1$ for all integer m. Via the cumulant formula, we also know that for all integers m and for t complex sufficiently small for which $\log(L_m(e^t)$ exists, we have

$$\log(L_m(e^t)) = t\mathbf{E}(L_m) + \frac{t^2}{2}\mathrm{Var}(L_m) + O(t^3) \ . \tag{12}$$

Notice that the term $O(t^3)$ is not uniform in m. In passing we remark that $\mathbf{E}(L_m) = L'_m(1)$ and $\mathrm{Var}(L_m) = L''_m(1) + L'_m(1) - (L'_m(1))^2$.

In [3] we proved the following result that we adopt here.

Theorem 3. *Consider a digital search tree built over m independent strings. Then*

$$\mathbf{E}(L_m) = \ell(m) + O(1), \qquad \mathrm{Var}(L_m) = v(m).$$

We are aiming at showing that the limiting distribution of the path length is normal for $m \to \infty$. In order to accomplish it, we need one technical result presented next to be proved only in the final version of this paper.

Theorem 4. *For all $\delta > 0$ and for all $\delta' < \delta$ there exists $\varepsilon > 0$ such that for $|t| \le \varepsilon$: $\log L_m(e^{tm^{-\delta}})$ exists, and*

$$\log L_m(e^{tm^{-\delta}}) = O(m),$$

$$\log L_m(e^{tm^{-\delta}}) = \frac{t}{m^\delta}\mathbf{E}(L_m) + \frac{t^2}{2m^{2\delta}}\mathrm{Var}(L_m) + t^3 O(m^{1-3\delta'}) \ .$$

Provided Theorem 4 is true, we are ready to prove our main results concerning the path length L_m.

Theorem 5. *Consider a digital search tree built over m sequences generated by a memoryless source. The random variable $\frac{L_m - \mathbf{E}(L_m)}{\sqrt{\mathrm{Var} L_m}}$ tends to a normal distribution with mean 0 and variance 1 in probability and in moments. More precisely, for any given real number x:*

$$\lim_{m\to\infty} P(L_m < \mathbf{E}(L_m) + x\sqrt{\mathrm{Var}(L_m)}) = \Phi(x), \tag{13}$$

and for all nonnegative integer k and $\varepsilon > 0$

$$\mathbf{E}\left(\left(\frac{L_m - \mathbf{E}(L_m)}{\sqrt{\mathrm{Var} L_m}}\right)^k\right) = \mu_k + O(m^{-\frac{1}{2}+\varepsilon}) \tag{14}$$

where $\mu_k = 0$ for k odd and $\mu_k = \frac{k!}{2^{k/2}(\frac{k}{2})!}$ for k even.

Proof: We apply Levy's continuity theorem or equivalently Goncharov's result [11] asserting that $\frac{L_m - \mathbf{E}(L_m)}{\sqrt{\mathrm{Var} L_m}}$ tends to the standard normal distribution if for complex τ

$$L_m\left(\exp\left(\frac{\tau}{\sqrt{\mathrm{Var}(L_m)}}\right)\right) e^{-\tau E(L_m)/\sqrt{\mathrm{Var}(L_m)}} \to e^{\tau^2/2}. \tag{15}$$

To prove it we apply several times our main technical result Theorem 4 with $t = \frac{\tau m^\delta}{\sqrt{\mathrm{Var} L_m}} = O(\varepsilon m^{-\delta})$. For some $\frac{1}{2} < \delta < 1$ and $\delta' > \delta$ such that $1 - 3\delta' < 0$, we obtain

$$\log L_m\left(\exp\left(\frac{\tau}{\sqrt{\mathrm{Var}(L_m)}}\right)\right) = \frac{\tau E[L_m]}{\sqrt{\mathrm{Var} L_m}} + \frac{\tau^2}{2} + O(m^{-\frac{1}{2}+\varepsilon'}) \tag{16}$$

for some $\varepsilon' > 0$. Thus by (15) the normality result follows.

To establish the convergence in moments, we use (16) in the Cauchy formula applied on a circle of radius R encircling the origin, that is,

$$\mathbf{E}\left(\left(\frac{L_m - \mathbf{E}(L_m)}{\sqrt{\mathrm{Var} L_m}}\right)^k\right)$$

$$= \frac{1}{2i\pi}\oint \frac{d\tau}{\tau^{k+1}} L_m(\exp(\frac{\tau}{\sqrt{\mathrm{Var}(L_m)}})) e^{-\tau/\sqrt{\mathrm{Var}(L_m)}}$$

$$= \frac{1}{2i\pi}\oint \frac{d\tau}{\tau^{k+1}} \exp(\frac{\tau^2}{2})(1 + O(m^{-\frac{1}{2}+\varepsilon'})$$

$$= \mu_k + O(R^{-k}\exp(\frac{R^2}{2}) m^{-\frac{1}{2}+\varepsilon'}).$$

This completes the proof. ∎

We also have large deviation results for the path length presented next.

Theorem 6. *Consider a digital search tree built over m sequences generated by a memoryless source.*
(i) [Large deviation]. *Let $\frac{1}{2} < \delta < 1$. Then there exist $\varepsilon > 0$, $B > 0$, and $\beta > 0$ such that for all $x \ge 0$:*

$$P(|L_m - \mathbf{E}(L_m)| > x m^\delta) \le B\exp(-\beta m^\varepsilon x). \tag{17}$$

(ii) [Moderate deviation]. *Let $\delta < \frac{1}{6}$ and $A > 0$. Then there exists $B > 0$ such that for non-negative real number $x < An^\delta$:*

$$P(|L_m - \mathbf{E}(L_m)| \ge x\sqrt{\mathrm{Var}(L_m)}) \le Be^{-\frac{x^2}{2}} \tag{18}$$

as $m \to \infty$.

Proof: We apply the Chernov bound. We take t as being a non negative real number. We have the identity:

$$P\left(L_m > E(L_m) + xm^\delta\right) = P\left(e^{tL_m} > e^{(E(L_m)+xm^\delta)t}\right) \ . \tag{19}$$

Using Markov's inequality we find

$$P(e^{tL_m} > e^{(E(L_m)+xm^\delta)t}) \le \frac{\mathbf{E}(e^{tL_m})}{e^{(E(L_m)+xm^\delta)t}}$$
$$= \quad L_m(e^t)\exp(-t\mathbf{E}(L_m) - xm^\delta t) \ .$$

Here we take $\delta' = \frac{\delta+1/2}{2}$ and $\varepsilon = \delta' - \frac{1}{2}$. Let fix $t = \beta m^{-\delta'}$ such that the estimate $\log L_m(e^t) = tE(L_m) + O(t^2 m^{1+\varepsilon})$ is valid. Therefore we have

$$\log L_m(e^t) - t\mathbf{E}[L_m] = O(m^{-\varepsilon}), \tag{20}$$

which tends to zero. We complete the proof by noticing that $tm^\delta x = \beta m^\varepsilon x$.

To obtain an upper bound we follow the same route only considering $-t$ instead of t. Indeed,

$$P\left(L_m < E(L_m) - xm^\delta\right) = P\left(e^{-tL_m} > e^{-(E(L_m)-xm^\delta)t}\right)$$
$$\leq L_m(e^{-t})\exp(t\mathbf{E}(L_m) - xm^\delta t).$$

To prove part (ii) of **moderate deviation**, we apply again Theorem 4 with $t = \frac{xm^\delta}{\sqrt{\text{Var}L_m}}$ leading to

$$\log L_m\left(\exp(\frac{x}{\sqrt{\text{Var}(L_m)}}\right) - \mathbf{E}(L_m)\frac{x}{\sqrt{\text{Var}(L_m)}} = \frac{x^2}{2} + o(1). \tag{21}$$

Indeed, from Theorem 4 with $\delta < \frac{1}{6}$ and $\delta' < \delta$

$$\log L_m(\exp(\frac{x}{\sqrt{\text{Var}(L_m)}})) = \log L_m(e^{tm^{-\delta}})$$
$$= \mathbf{E}(L_m)\frac{x}{\sqrt{\text{Var}(L_m)}} + \frac{x^2}{2\text{Var}(L_m)}\text{Var}(L_m)$$
$$+\frac{x^3 m^{3\delta}}{(\text{Var}(L_m))^{\frac{3}{2}}}O(m^{1-3\delta'}).$$

Observe that the error term is at most $O(m^{1-\frac{3}{2}+3\delta-3\delta'}(\log m)^{-3}) = o(1)$, as needed. Therefore, by Markov inequality for all $t > 0$,

$$P(L_m < \mathbf{E}(L_m) + x\sqrt{\text{Var}(L_m)}) \leq$$
$$\leq \exp(\log L_m(e^t) - t\mathbf{E}(L_m) - xt\sqrt{\text{Var}(L_m)}).$$

Taking t as above and applying the estimate (21), we find

$$P(L_m < \mathbf{E}(L_m) + x\sqrt{\text{Var}(L_m)}) \leq$$
$$\exp(\log L_m(e^t) - t\mathbf{E}(L_m) - xt\sqrt{\text{Var}(L_m)})$$
$$= \exp\left(\frac{x^2}{2} + o(1) - xt\sqrt{\text{Var}(L_m)}\right) \sim \exp(-\frac{x^2}{2}) .$$

This completes the proof. ∎

IV. Proofs of Main Theorems 1 and 2

In this section we prove our main results, namely Theorems 1 and 2. We start with the large deviation results.

A. Proof of Theorem 2

We start with Theorem 2(i). By (9) we have

$$P(M_n > \ell^{-1}(n) + yn^\delta) = P(M_n > \lfloor \ell^{-1}(n) + yn^\delta \rfloor)$$
$$= P(L_{\lfloor \ell^{-1}(n)+yn^\delta \rfloor} < n) .$$

Observe that $\mathbf{E}(L_m) = \ell(m) + O(1)$, hence

$$\mathbf{E}(L_{\lfloor \ell^{-1}(n)+yn^\delta \rfloor}) = \ell(\ell^{-1}(n) + yn^\delta) + O(1). \tag{22}$$

Since the function $\ell(.)$ is convex and $\ell(0) = 0$, we have for all real numbers $a > 0$ and $b > 0$

$$\ell(a+b) \geq \ell(a) + \frac{\ell(a)}{a}b, \tag{23}$$
$$\ell(a-b) \leq \ell(a) - \frac{\ell(a)}{a}b. \tag{24}$$

Applying inequality (23) to $a = \ell^{-1}(n)$ and $b = yn^\delta$ we arrive at

$$n - \mathbf{E}(L_{\lfloor \ell^{-1}(n)+yn^\delta \rfloor}) \leq -y\frac{n}{\ell^{-1}(n)}n^\delta + O(1). \tag{25}$$

Thus

$$P(L_{\lfloor \ell^{-1}(n)+yn^\delta \rfloor} < n) \leq P(L_m - \mathbf{E}(L_m) < -xm^\delta + O(1)) , \tag{26}$$

by identifying

$$m = \lfloor \ell^{-1}(n) + yn^\delta \rfloor, \qquad x = \frac{n}{\ell^{-1}(n)}\frac{n^\delta}{m^\delta}y.$$

We now apply several times Theorem 6 from the previous section regarding the path length L_m. That is, for all $x > 0$ and for all m, there exist $\varepsilon > 0$ and A such that

$$P(L_m - \mathbf{E}(L_m) < xm^\delta) < Ae^{-\beta x m^\varepsilon} . \tag{27}$$

In other words,

$$P(L_m - \mathbf{E}(L_m) < xm^\delta + O(1)) \leq Ae^{-\beta x m^\varepsilon + O(n^{\varepsilon-\delta)})} \leq A'e^{-\beta x m^\varepsilon}$$

for some $A' > A$ we find

$$P(M_n > \ell^{-1}(n) + yn^\delta) \leq A'\exp(-\beta x m^\varepsilon). \tag{28}$$

We know that $\ell^{-1}(n) = \Omega(\frac{n}{\log n})$. Thus $x = O((\log n)^{1-\delta})\frac{y}{(1+n^{\delta-1}\log ny)^\delta} \leq \beta'\frac{n^{\varepsilon'}y}{(1+yn^{-\varepsilon'})^\delta}$ for some $0 < \varepsilon' < \varepsilon$ and $\beta' > 0$.

In a similar fashion, we have

$$P(M_n < \ell^{-1}(n) - yn^\delta) = P(L_{\lfloor \ell^{-1}(n)-yn^\delta \rfloor} > n) \tag{29}$$

and

$$\mathbf{E}(L_{\lfloor \ell^{-1}(n)-yn^\delta \rfloor}) = \ell(\ell^{-1}(n) - yn^\delta) + O(1). \tag{30}$$

Using inequality (24) we obtain

$$n - \mathbf{E}(L_{\lfloor \ell^{-1}(n)-yn^\delta \rfloor}) \geq y\frac{n}{\ell^{-1}(n)}n^\delta + O(1). \tag{31}$$

In conclusion,

$$P(L_{\lfloor \ell^{-1}(n)-yn^\delta \rfloor} > n) \leq P(L_m - \mathbf{E}(L_m) > xm^\delta + O(1)) ,$$

by identifying

$$m = \lfloor \ell^{-1}(n) - yn^\delta \rfloor, \qquad x = \frac{n}{\ell^{-1}(n)}\frac{n^\delta}{m^\delta}y.$$

Observe that this case is easier since we have now $m < \ell^{-1}(n)$ and we don't need the correcting term $(1 + yn^\varepsilon)^{-\delta}$.

Now we can turn our attention to **moderate deviation** expressed in Theorem 2(ii). It is essentially the same proof except that we consider $y\frac{s_n}{\ell'(\ell^{-1}(n))}$ with $s_n = \sqrt{v(\ell^{-1}(n))}$ instead of yn^δ, and we assume $y = O(n^{\delta'})$ for some $\delta' < \frac{1}{6}$. Thus $y\frac{s_n}{\ell'(\ell^{-1}(n))} = O(n^{\frac{1}{2}+\delta}) = o(n)$. If $y > 0$, then

$$P(M_n > \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))}) = P(L_m < n) \tag{32}$$

with $m = \lfloor \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))} \rfloor$. We use the estimate

$$\ell(a+b) = \ell(a) + \ell'(a)b + o(1)$$

when $b = o(a)$ when $a \to \infty$. Thus

$$\ell(\ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))}) = n + ys_n + o(1)\ . \qquad (33)$$

Since $\sqrt{v(m)} = s_n + O(1)$ we have $n = \mathbf{E}(L_m) - y\sqrt{v(m)} + o(1)$. Referring again to Theorem 6: we know that

$$P(L_m < E(L_m) - y\sqrt{v(m)} + O(1)) \le A\exp(-\frac{y^2}{2}),$$

where the term $O(1)$ inducing a term $\exp(O(\frac{y^2}{v(m)})) = \exp(o(1))$ that is absorbed by A. The proof for $y < 0$ follows a similar path.

B. Proof of Theorem 1

We first show that for all $1 > \delta > \frac{1}{2}$

$$\mathbf{E}(M_n) = \ell^{-1}(n) + O(n^\delta).$$

Indeed, noticing that for any random variable X: $|\mathbf{E}(X)| \le \mathbf{E}(|X|) = \int_0^\infty P(|X| > y)dy$, we set $X = M_n - \ell^{-1}(n)$ to find from Theorem 2(i)

$$|\mathbf{E}(M_n) - \ell^{-1}(n)| \le$$

$$\le n^\delta + n^\delta \int_1^\infty P(|M_n - \ell^{-1}(n)| > yn^\delta)dy = O(n^\delta).$$

Let us now move our attention to **large deviations** results. For a given y, we have

$$P(M_n > \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))}) =$$

$$P(L_{\lfloor \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))} \rfloor} < n).$$

Let $m = \lfloor \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))} \rfloor$. We know that

$$n - E(L_m) = ys_n + O(1)$$

and

$$s_n = \sqrt{v(\ell^{-1}(n))} = \sqrt{\text{Var}(L_m)} + o(1).$$

Therefore

$$P\left(M_n > \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))}\right) =$$

$$P\left(L_m < E(L_m) + y\sqrt{\text{Var}(L_m)} + O(1)\right).$$

Assume that the $|O(1)| \le A$

$$P\left(L_m < E(L_m) + (y + \frac{A}{\sqrt{\text{Var}(L_m)}})\sqrt{\text{Var}(L_m)}\right) \ge$$

$$P\left(M_n > \ell^{-1} + y\frac{s_n}{\ell'(\ell^{-1}(n))}\right)$$

since for all y' $\lim_{m\to\infty} P\left(L_m < E(L_m) + y'\sqrt{\text{Var}(L_m)}\right) = \Phi(y')$ and therefore by continuity of $\Phi(x)$

$$\lim_{m\to\infty} P\left(L_m < E(L_m) + (y \pm \frac{A}{\sqrt{\text{Var}(L_m)}}\sqrt{\text{Var}(L_m)}\right) = \Phi(y)\ .$$

Therefore

$$\lim_{m\to\infty} P\left(M_n > \ell^{-1}(n) + y\frac{s_n}{\ell'(\ell^{-1}(n))}\right) = 1 - \Phi(y)$$

and following the same footsteps we also establish the matching lower bound

$$\lim_{m\to\infty} P(M_n < \ell^{-1}(n) - y\frac{s_n}{\ell'(\ell^{-1}(n))}) = \Phi(y)$$

This proves two things: first that

$$(M_n - \ell^{-1}(n))\frac{\ell'(\ell(^{-1}(n))}{s_n}$$

tends to the normal distribution in probability. Second, since by the moderate deviation result the normalized random variable $(M_n - \ell^{-1}(n))\frac{\ell'(\ell(^{-1}(n))}{s_n}$ has bounded moments, and then by the virtue of the dominated convergence:

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}\left((M_n - \ell^{-1}(n))\frac{\ell'(\ell(^{-1}(n))}{s_n})\right) &= 0,\\ \lim_{n\to\infty} \mathbf{E}\left(\left((M_n - \ell^{-1}(n))\frac{\ell'(\ell(^{-1}(n))}{s_n})\right)^2\right) &= 1.\end{aligned}$$

In other words,

$$\text{Var}(M_n) \sim \frac{v(\ell^{-1}(n))}{(\ell'(\ell^{-1}(n)))^2}.$$

This completes the proof of our main result Theorem 1.

ACKNOWLEDGMENT

This work was supported by the NSF Science and Technology Center for Science of Information Grant CCF-0939370, NSF Grants DMS-0800568 and CCF-0830140, and the MNSW grant N206 369739.

REFERENCES

[1] D. Aldous, and P. Shields, A Diffusion Limit for a Class of Random-Growing Binary Trees, *Probab. Th. Rel. Fields*, 79, 509–542, 1988.
[2] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York 1968.
[3] P. Jacquet and W. Szpankowski, Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees, *Theoretical Computer Science*, 144, 161–197, 1995.
[4] P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, *Algorithmica*, 31, 318–360, 2001.
[5] G Louchard, W Szpankowski, On the average redundancy rate of the Lempel-Ziv code. *IEEE Transactions on Information Theory*, 43, 2–8, 1997.
[6] D. Knuth, *The Art of Computer Programming. Vol. III Sorting and Searching*, (Second Edition), Addison-Wesley (1998).
[7] N. Merhav, Universal Coding with Minimum Probability of Codeword Length Overflow, *IEEE Trans. Information Theory*, 37, 556–563, 1991.
[8] R. Neininger and L. Ruschendorf, A General Limit Theorem for Recursive Algorithms and Combinatorial Structures, *The Annals of Applied Probability*, 14, No. 1, 378-418, 2004.
[9] E. Plotnik, M.J. Weinberger, and J. Ziv, Upper Bounds on the Probability of Sequences Emitted by Finite-State Sources and on the Redundancy of the Lempel-Ziv Algorithm, *IEEE Trans. Information Theory*, 38, 66-72 (1992).
[10] S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. Information Theory*, 43, 9–21, 1997.
[11] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.
[12] J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Transactions on Information Theory*, 24, 530–536, 1978.