Mathematics Interdisciplinary Research **6** (2021) 215 – 223

Original Scientific Paper

A Note on the Lempel-Ziv Parsing Algorithm under Asymmetric Bernoulli Model

Hojjat Naeini, Ramin Kazemi⋆ and Mohammad Hasan Behzadi

Abstract

In this paper, by applying analytic combinatorics, we obtain an asymptotics for the t-th moment of the number of phrases of length ℓ in the Lempel-Ziv parsing algorithms built over a string generated by an asymmetric Bernoulli model. We show that the t-th moment is approximated by its Poisson transform.

Keywords: Lempel-Ziv parsing algorithm, phrases, digital search tree, moment.

2010 Mathematics Subject Classification: Primary 05C05, Secondary 60F05.

How to cite this article
H. Naeini, R. Kazemi and M. H. Behzadi, A note on the Lempel -Ziv parsing algorithm under asymmetric Bernoulli model, *Math. Interdisc. Res.* **6** (2021) 215–223.

1. Introduction

The Lempel-Ziv (LZ) algorithm is an algorithm for lossless data compression. These algorithms are used in compression utilities such as GIF image compression and gzip [1, 15].

The idea of the LZ parsing algorithm is to partition a sequence over a finite alphabet (here $\Sigma = \{0, 1\}$) into phrases (or blocks) of variable sizes such that a new phrase is the shortest substring not seen in the past as a phrase. For example,

⋆Corresponding author (E-mail: r.kazemi@sci.ikiu.ac.ir)
Academic Editor: Abdolrahman Razani
Received 10 January 2021, Accepted 21 July 2021
DOI: 10.22052/MIR.2021.240429.1263

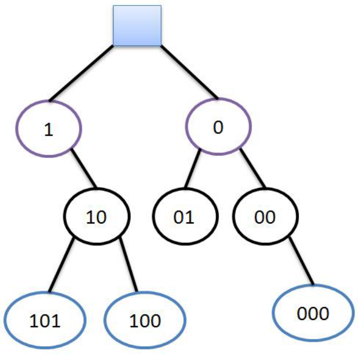

Figure 1: A digital tree representation of string 110010100010001000.

the string 110010100010001000 is parsed into

$$1 - 10 - 0 - 101 - 00 - 01 - 000 - 100.$$

See Figure 1 for the digital tree representation of LZ's parsing for the above string.

These algorithms play a crucial role in a universal data compression scheme [4, 5, 8, 9, 10, 12, 13, 14, 15]. Here, we discuss on the t-th moment of the number of phrases in this algorithm. Let $X_{n,\ell}$ be the random number of phrases of length ℓ in the LZ algorithm built over n phrases for an asymmetric Bernoulli model (each string is a binary i.i.d. sequence with p being the probability of a "1" $(0 < p < q < 1)$). First, we show the Poisson generating function of $\mathbb{E}(X^2_{n,\ell})$ (namely, $D_{\ell,2}(x)$) satisfies the following functional-differential equation

$$D^{'}_{\ell,2}(x) + D_{\ell,2}(x) = D_{\ell-1,2}(px) + D_{\ell-1,2}(qx) + 2D_{\ell-1,1}(px)D_{\ell-1,1}(qx), \qquad (1)$$

with $D_{0,2}(x) = 1 - e^{-x}$. The equation (1) translates into a new equation that we solve it by introducing two appropriate operators. Then we prove Theorem 2.1 that is crucial for the solution of our problem. Finally, we show that $\mu_{n,\ell,t} = \mathbb{E}(X^t_{n,\ell})$ is asymptotically equal to $D_{\ell,t}(n)$ for $t = 2, 3, \ldots$.

2. The Main Results

Because it is not possible to determine the probability function of the random variable $X_{n,\ell}$ by probabilistic method, we use the combinatorial method. As it is natural in enumeration problems related to labelled structures, we define the exponential generating functions

$$f_{\ell,i}(x) \quad = \quad \sum_{n\geq 0} \mathbb{E}(X^i_{n,\ell})\frac{x^n}{n!}, \qquad i \geq 1.$$

and their Poisson transforms, i.e., $D_{\ell,i}(x) = e^{-x} f_{\ell,i}(x)$. By the same method of [2, 7, 11] and the relation introduced in Section 1, for $\mathcal{P}_{n,\ell}(u) = \mathbb{E}(u^{X_{n,\ell}})$ we have

$$\mathcal{P}_{n+1,\ell}(v) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \mathcal{P}_{i,\ell-1}(v) \mathcal{P}_{n-i,\ell-1}(v),$$

with initial conditions $\mathcal{P}_{0,\ell}(v) = 1$ for $\ell \geq 1$, $\mathcal{P}_{0,0}(v) = v$, $\mathcal{P}_{n,0}(v) = 1$ for $n \geq 1$. First, we focus on the case $t = 2$. The function $\mathcal{G}_\ell(x, v)$ as

$$\mathcal{G}_\ell(x, v) = \sum_{n \geq 0} \mathcal{P}_{n,\ell}(v) \frac{x^n}{n!},$$

fulfills the following functional recurrence

$$\frac{\partial}{\partial x} \mathcal{G}_\ell(x, v) = \mathcal{G}_{\ell-1}(px, v) \mathcal{G}_{\ell-1}(qx, v), \qquad \ell \geq 1,$$

with initial conditions $\mathcal{G}_0(x, v) = v + e^x - 1$ and $\mathcal{G}_\ell(0, v) = 1$ $(\ell \geq 1)$. By taking second derivatives with respect to v (and setting $v = 1$) we obtain for $f_{\ell,2}(x)$ fulfills the following functional recurrence

$$f'_{\ell,2}(x) = e^{qx} f_{\ell-1,2}(px) + e^{px} f_{\ell-1,2}(qx) + 2 f_{\ell-1,1}(px) f_{\ell-1,1}(qx). \qquad (2)$$

Also, the Poisson transform of $D_{\ell,2}(x)$ translates recurrence (2) into

$$D'_{\ell,2}(x) + D_{\ell,2}(x) = D_{\ell-1,2}(px) + D_{\ell-1,2}(qx) + 2 D_{\ell-1,1}(px) D_{\ell-1,1}(qx), \qquad (3)$$

with initial conditions $D_{0,2}(x) = 1 - e^{-x}$ and $D_{\ell,2}(0) = 0$ $(\ell \geq 1)$. For $n \leq \ell$, $X_{n,\ell} = 0$ (as it is for the internal profiles). Thus $f_{\ell,2}(x) = \mathcal{O}(x^{\ell+1})$ as $x \to 0$. Then, the Mellin transform $D^*_{\ell,2}(s)$ actually exists for s with $-\ell - 1 < \Re(s) < 0$. By the structure of function of $D_{\ell,2}(x)$, we can express $D^*_{\ell,2}(s)$ as $-\Gamma(s)\mathcal{F}_\ell(s)$ where $\Gamma(s)$ is the gamma function. Thus $\mathcal{F}_\ell(s)$ is the finite linear combinations of functions a^{-s} with certain values of a and one can be considered as an entire function. Furthermore (3) translates into

$$\mathcal{F}_\ell(s) - \mathcal{F}_\ell(s-1) = \mathcal{S}(s)\mathcal{F}_{\ell-1}(s) + \mathcal{H}_\ell(s), \qquad \ell \geq 0, \qquad (4)$$

where

$$\mathcal{H}_\ell(s) = \int_0^\infty (\Gamma(s))^{-1} 2 D_{\ell-1,1}(px) D_{\ell-1,1}(qx) x^{s-1} dx,$$

and $\mathcal{F}_0(s) = 1$. The equation (4) holds for all s, since $\mathcal{F}_\ell(s)$ continues analytically to an entire function, [6].

In order to use of Cauchy residue theorem [3] we define

$$f(s, w) = \sum_{\ell \geq 0} \mathcal{F}_\ell(s) w^\ell.$$

Let us introduce functional operators **A** and **C** as follows

$$\begin{aligned}\mathbf{C}_{f;s} &= f(s) + f(s-1) + f(s-2) + f(s-3) + \cdots ,\\ \mathbf{A}_{f;s} &= f(s)\mathcal{S}(s) + f(s-1)\mathcal{S}(s-1) + f(s-2)\mathcal{S}(s-2) + \cdots ,\end{aligned}$$

where $\mathcal{S}(s) = p^{-s} + q^{-s}$. Also suppose $g(s,w) = \sum_{\ell\geq 0} \mathbf{A}^{\ell}_{1;s} w^{\ell}$ and

$$\widetilde{f}_{\ell;s} = \mathbf{C}_{f_{\ell-1};s} - \mathbf{C}_{f_{\ell-1};-1}, \qquad \widehat{g}_{\ell;s} = \mathbf{A}^{\ell}_{\widetilde{g}_{\ell-\ell};s} - \mathbf{A}^{\ell}_{\widetilde{g}_{\ell-\ell};-1}.$$

In the following theorem we find an explicit representation of $f(s,w)$ in terms of the operators **A** and **C**.

Theorem 2.1. *The power series $f(s,w)$ satisfies*

$$f(s,w) = H(s,w) + \sum_{\ell\geq 0} \boldsymbol{A}^{\ell}_{N(\cdot,w);s} w^{\ell} - H(s,w)\sum_{\ell\geq 0} \boldsymbol{A}^{\ell}_{N(\cdot,w);-1} w^{\ell},$$

where

$$H(s,w) = \frac{g(s,w)}{g(0,w)}, \qquad N(s,w) = \sum_{\ell\geq 0} \widetilde{H}_{\ell;s} w^{\ell}.$$

Proof. It is obvious. Similar considerations are done in [2] in proof of Theorem 3 where the (somewhat simpler) recurrences appearing there are treated analogously. □

We now show asymptotic behavior of the second moment of the our random variable because by studying the second moment, we can guess the behavior of the t-th moment. First we show that $N(s,w)$ is analytic for $|w| < (\mathcal{S}(\Re(s)) - \nu)^{-1}$ for some $\nu > 0$ and can derive $f(s,w) \approx H(s,w)$. Finally we prove $\mathcal{F}_{\ell}(s)$ behave asymptotically as $\mathcal{S}(s)^{\ell}$ and show that $\mathbb{E}(X^2_{n,\ell}) \approx D_{\ell,2}(n)$. Since for complex s [2], $D^{*}_{\ell,1}(s) \leq C'\Gamma(s)\mathcal{S}(s)^{\ell}$, by the Mellin transform property

$$|D^{'*}_{\ell,1}(s)| = |-(s-1)D^{*}_{\ell,1}(s-1)| \leq C'|s||\Gamma(s-1)|\mathcal{S}(\Re(s)-1)^{\ell},$$

for constant C'. Thus by convolution of Mellin transform:

$$\begin{aligned}|\mathcal{H}_{\ell}(s)| &= \Big|\frac{1}{\Gamma(s)}\int_0^{\infty} 2D^{(1)}_{\ell-1}(px)D^{(1)}_{\ell-1}(qx)x^{s-1}dx\Big|\\ &= \Big|\frac{1}{\Gamma(s)}\Big|\Big|\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} D^{'*(1)}_{\ell}(u)D^{'*(1)}_{\ell}(s-u)du\Big|\\ &\leq \frac{C''}{|\Gamma(s)|}\int_{c-i\infty}^{c+i\infty} \frac{|u||s-u||\Gamma(u)||\Gamma(s-u-1)|}{\Big(\mathcal{S}(\Re(u)-1)\mathcal{S}(\Re(s-u)-1)\Big)^{-\ell}}du\\ &\leq C\mathcal{S}(c-1)^{2\ell} \qquad\qquad \Re(u) = c = \Re(s-u),\\ &= C\mathcal{S}\left(\frac{\Re(s)}{2}-1\right)^{2\ell} \qquad\qquad c = \frac{\Re(s)}{2},\\ &\leq C(\mathcal{S}(\Re(s)) - \nu)^{\ell},\end{aligned}$$

for constant C and for some $\nu > 0$.

Lemma 2.2. *There exists $\nu > 0$ such that $N(s,w)$ is analytic for $|w| < (\mathcal{S}(\Re(s)) - \nu)^{-1}$.*

Proof. It is obvious $|\mathcal{S}(s-j)| \leq |\mathcal{S}(s)| \max(pq)^j$ for $j \geq 0$. By definition of $\widetilde{H}_{\ell;s}$, if $|w| < (\mathcal{S}(\Re(s) - \nu))^{-1}$, then $N(s,w) = \sum_{\ell \geq 0} \widetilde{H}_{\ell;s} w^\ell$ converges absolutely and represents an analytic function. □

For a real number θ with $(\log \frac{1}{p})^{-1} < \theta < (\log \frac{1}{q})^{-1}$, let

$$\lambda = \lambda(\theta) = \frac{1}{\log(p/q)} \log \Big(\frac{1 - \theta \log(1/p)}{\theta \log(1/q) - 1} \Big).$$

Equivalently,

$$\theta = \frac{p^{-\lambda} + q^{-\lambda}}{p^{-\lambda} \log \frac{1}{p} + q^{-\lambda} \log \frac{1}{q}}.$$

Theorem 2.3. *$\mathcal{F}_\ell(s)$ behave asymptotically as $\mathcal{S}(s)^\ell$ and $\mathbb{E}(X_{n,\ell}^2) \approx D_{\ell,2}(n)$.*

Proof. For some $\nu > 0$, $N(s,w)$ is analytic for $|w| < (\mathcal{S}(\Re(s)) - \nu)^{-1}$. Then $\mathcal{F}_\ell(s)$ behave asymptotically as $\mathcal{S}(s)^\ell$ as was the case of the first moment of the internal profile in [2]. By applying the saddle point method for inverse Mellin transform (in case $x = n$) for

$$D_{\ell,2}(n) = \frac{1}{2\pi i} \int_{\lambda - i\infty}^{\lambda + i\infty} D^*_{\ell,2}(n)(s) n^{-s} ds,$$

we can obtain a similar Theorem 2.1 in [2] for $\mathbb{E}(X_{n,\ell}^2)$ (see [2] for details and calculations). □

Let $\mu_{n,\ell,t} = \mathbb{E}(X_{n,\ell}^t)$ be the t-th moment of the $X_{n,\ell}$. By the similar manner, for

$$E_\ell^{(t)}(x) = \sum_{n \geq 0} \mu_{n,\ell,t} \frac{x^n}{n!},$$

we obtain

$$\begin{aligned} E_\ell^{'(t)}(x) &= e^{qx} E_{\ell-1}^{(t)}(px) + e^{px} E_{\ell-1}^{(t)}(qx) \\ &\quad + \sum_{m=1}^{t-1} \sum_{n=1}^{t-1} \beta(m,n) E_{\ell-1}^{(m)}(px) E_{\ell-1}^{(n)}(qx), \end{aligned} \tag{5}$$

where $\beta(m,n) \in \mathbb{Z}$ and $E_0^{(t)}(x) = e^x - 1$ [2, 11].

Theorem 2.4. *The asymptotics of $\mu_{n,\ell,t}$ is of the same order of magnitude as for the average value.*

Proof. Let $\Delta_\ell^{(t)}(x) = e^{-x}E_\ell^{(t)}(x)$ be the Poisson transform $E_\ell^{(t)}(x)$. Then

$$(\Delta_\ell^{(t)}(z))' = e^{-z}E_\ell^{'(t)}(z) - \Delta_\ell^{(t)}(z).$$

Thus recurrence (5) translates into

$$\begin{aligned}\Delta_\ell^{'(t)}(x) + \Delta_\ell^{(t)}(x) &= \Delta_{\ell-1}^{(t)}(px) + \Delta_{\ell-1}^{(t)}(qx) \\ &\quad + \sum_{m=1}^{t-1}\sum_{n=1}^{t-1}\beta(m,n)\Delta_{\ell-1}^{(m)}(px)\Delta_{\ell-1}^{(n)}(qx), \quad \ell \geq 1, \qquad (6)\end{aligned}$$

with initial conditions $\Delta_\ell^{(t)}(0) = 0$ ($\ell \geq 1$) and $\Delta_0^{(t)}(x) = 1 - e^{-x}$, since $p + q = 1$.

It is easy to show that

$$\Delta_\ell^{(t)}(x) = \sum_{\ell_1}\sum_{\ell_2}\theta(\ell_1,\ell_2)\exp\Big\{-\sum_i\sum_j \xi(i,j)p^{\ell_i}q^{\ell_j}x\Big\},$$

with $\ell_i, \ell_j \geq 0$ and $\theta(\ell_1,\ell_2), \xi(i,j) \in \mathbb{Z}$. We express $\Delta_\ell^{*(t)}(x)$ as

$$\Delta_\ell^{*(t)}(x) = \int_0^\infty \Delta_\ell^{(t)}(x)x^{s-1}dx = -\Gamma(s)\mathcal{F}_\ell^{(t)}(s),$$

where

$$\mathcal{F}_\ell^{(t)}(s) = \sum_{\ell_1}\sum_{\ell_2}\theta(\ell_1,\ell_2)\Big\{\sum_i\sum_j \xi(i,j)^{-s}p^{-\ell_i s}q^{-\ell_j s}\Big\}.$$

Thus, $\mathcal{F}_\ell^{(t)}(s)$ can be assumed an entire function and (6) translates into

$$\mathcal{F}_{\ell+1}^{(t)}(s) - \mathcal{F}_{\ell+1}^{(t)}(s-1) = \mathcal{S}(s)\mathcal{F}_\ell^{(t)}(s) + \mathcal{H}_\ell^{(t)}(s), \quad \ell \geq 0, \ \ \mathcal{F}_0^{(t)}(s) = 1, \qquad (7)$$

where for $p < q$,

$$\begin{aligned}\mathcal{H}_\ell^{(t)}(s) &= \sum_{m=1}^{t-1}\sum_{n=1}^{t-1}\frac{\beta(m,n)}{\Gamma(s)}\int_0^\infty \Delta_\ell^{(m)}(px)\Delta_\ell^{(n)}(qx)x^{s-1}dx \\ &\leq p^{-s}\sum_{m=1}^{t-1}\sum_{n=1}^{t-1}\frac{\beta(m,n)}{\Gamma(s)}\int_0^\infty \Delta_\ell^{(m)}(x)\Delta_\ell^{(n)}(x)x^{s-1}dx \\ &= p^{-s}\sum_{m=1}^{t-1}\sum_{n=1}^{t-1}\frac{\beta(m,n)}{2\pi i\,\Gamma(s)}\int_{c-i\infty}^{c+i\infty}\Delta_\ell^{*(m)}(x)\Delta_\ell^{*(n)}(x)dy.\end{aligned}$$

With the same consideration of [2], $\Delta_\ell^{*(t)}(x) \leq K\Gamma(s)\mathcal{S}(s)^\ell$ for some constant K.

Thus

$$\begin{aligned}\left|\mathcal{H}_\ell^{(t)}(s)\right| &\leq K^{'}p^{-s}\sum_{m=1}^{t-1}\sum_{n=1}^{t-1}\int_{c-i\infty}^{c+i\infty}\frac{\beta(m,n)}{2\pi i\,\Gamma(s)}|\Gamma(z-1)||\Gamma(s-z-1)| \\ &\times\Big(\mathcal{S}(\Re(z)-1)\mathcal{S}(\Re(s-z)-1)\Big)^{\ell}dz \\ &\leq K(s,p)\mathcal{S}(z-1)^{2\ell}, \qquad \Re(z)=z=\Re(s-z) \\ &= K(s,p)\mathcal{S}\left(\Re(s)/2-1\right)^{2\ell}.\end{aligned}$$

Thus $\mathcal{H}_\ell^{(t)}(s)=\mathcal{O}(\mathcal{S}(\Re(s))/2-1)^{2\ell}$. Similar to [11], one can see the inhomogeneous part in (7) is relatively small and proof is completed. □

4. Conclusion

We obtained an asymptotics for the $\mu_{n,\ell,t}$ built over a string generated by an asymmetric Bernoulli model through the relation between this algorithm and digital search tree. This result was derived by applying analytic combinatorics.

Conflicts of Interest. The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] J. E. Coffman and J. Eve, File structures using hashing functions, *Commun. ACM* **13** (1970) 427–432.

[2] M. Drmota and W. Szpankowski, Un(expected) behavior of digital search tree profile, *SIAM* (2009) 130–138.

[3] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2008.

[4] E. Gilbert and T. Kadota, The Lempel-Ziv algorithm and message complexity, *IEEE Trans. Inform. Theory* **38** (1992) 1839–1842.

[5] P. Jacquet and W. Szpankowski, Asymptotic behavior of the Lempel-Ziv parsing scheme and [in] digital search trees, *Theor. Comput. Sci.* **144** (1-2) (1995) 161–197.

[6] R. Kazemi, On the phrases in Lempel-Ziv parsing algorithm, *Int. J. Aca. Res.* (IJAR) **4** (4) (2012) 112–115.

[7] R. Kazemi and M. Q. Vahidi-Asl, The variance of the profile in digital search trees, *Discrete Math. Theor. Comput. Sci.* **13** (3) (2011) 21–38.

[8] P. Kirschenhofer, H. Prodinger and W. Szpankowski, On the variance of the external path in a aymmetric digital search trie, *Discrete Appl. Math.* **25** (1989) 129–143.

[9] A. Lempel and J. Ziv, On the complexity of finite sequences, *IEEE Inform. Theory* **22** (1) (1976) 75–81.

[10] G. Louchard and W. Szpankowski, Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm, *IEEE Trans. Inform. Theory* **41** (1995) 478–488.

[11] G. Park, H. K. Hwang, P. Nicodeme and W. Szpankowski, Profile of tries, *SIAM J. Comput.* **8** (2009) 1821–1880.

[12] A. Wyner and J. Ziv, Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression, *IEEE Trans. Inform. Theory* **35** (1989) 1250–1258.

[13] J. Ziv, On classification with empirically observed statistics and universal data compression, *IEEE Trans. Inform. Theory* **34** (1988) 278–286.

[14] J. Ziv, Compression, *Test of Randomness, and Estimating the Statistical Model of Individual Sequence–Sequences*, Combinatorics, Compression, Security, and Transmission. Ed. by R. M. Capocelli, Springer-Verlag, New York, 1990, pp. 366–373.

[15] J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Trans. Inform. Theory* **24** (5) (1978) 530–536.

Hojjat Naeini
Department of Statistics,
Science and Research Branch,
Islamic Azad University,
Tehran, I. R. Iran
e-mail: hojjatnaeini223344@gmail.com

Ramin Kazemi
Department of Statistics,
Imam Khomeini International University,
Qazvin, I. R. Iran
e-mail: r.kazemi@sci.ikiu.ac.ir

Mohammad Hasan Behzadi
Department of Statistics,

Science and Research Branch,
Islamic Azad University,
Tehran, I. R. Iran
e-mail:behzadi@srbiau.ac.ir