The Redundancy and Distribution of the Phrase Lengths of the Fixed-Database Lempel-Ziv Algorithm

by

Abraham J. Wyner
Department. of Statistics
University of California
Berkeley, CA
USA
(ajw@stat.berkeley.edu)

Abstract

The Fixed-Database version of the Lempel Ziv algorithm closely resembles many versions that appear in practice. In this paper, we ascertain several key asymptotic properties of the algorithm as applied to sources with finite memory. First, we determine that for a dictionary of size n, the algorithm achieves a redundancy $\rho_n = H\frac{\log\log n}{\log n} + o(\frac{\log\log n}{\log n})$, where H is the entropy of the process. This is the first, non-trivial, *lower bound* on any Lempel-Ziv type compression scheme. We then find the limiting distribution and all moments of the lengths of the phrases by comparing them to a random-walk like variable with well known behavior.

1 Introduction

The fixed-database Lempel-Ziv algorithm (FDLZ) closely resembles practical versions of the LZ algorithm that are widely in use. In this paper, we establish several theorems concerning the asymptotic properties of this version of the algorithm, concentrating on precise bounds on the redundancy, ρ_n (defined as the difference between the average rate and the source entropy H and when the memory is of size n). Specifically, we demonstrate that the algorithm has a redundancy $\rho_n = H\frac{\log\log n}{\log n} + o(\frac{\log\log n}{\log n})$. Although an upper bound on ρ_n exists for a different version of LZ (see [8]), this is the first lower bound. The calculation of these bounds for the FDLZ is possible by using several theorems that provide the limiting distribution and moments of the length of all the FDLZ phrases. These theorems will be stated in Section 2; we shall save the proofs for Sections 3 and 4.

We will be concerned with a data source which is a random sequence $\{\mathbf{Z}_k\}_{k=-\infty}^{\infty}$, that takes values in an alphabet $\mathcal{A}$, with $|\mathcal{A}| = A < \infty$. We assume that $\{Z_k\}$ is stationary, and has finite memory M. Thus

$$\Pr\left\{\mathbf{Z}_k = z_k \mid \mathbf{Z}_{-\infty}^{k-1} = \mathbf{z}_{-\infty}^{k-1}\right\} = \Pr\left\{\mathbf{Z}_k = z_k \mid \mathbf{Z}_{k-M}^{k-1} = \mathbf{z}_{k-M}^{k-1}\right\} , \tag{1}$$

where $\mathbf{z}_i^j = (z_i, z_{i+1}, \ldots, z_j)$, $-\infty \le i < j \le \infty$. We also assume that $\{Z_k\}$ is aperiodic. [1] Let H be the source entropy. Of course the data source can be losslessly encoded using $(H + \rho_n)$ bits per source symbol, where n is a measure of the complexity of the code, and $\rho_n \to 0$, as $n \to \infty$. The LZ algorithm is a universal procedure (which does not depend on the source statistics) for encoding the source at a rate close to the entropy which has been shown to achieve $\rho_n \to 0$ in a variety of contexts.

2 Description of the Algorithm

We begin with a discussion of the FDLZ algorithm. Let $\mathbf{Z}_1^\infty$ (appreviate simply $\mathbf{Z}$) be the data source, discussed above, which we wish to compress. Assume that the encoder has available a "training sequence" $\mathbf{X}_1^n$ that is statistically independent of, but with the same distribution as $\mathbf{Z}_1^n$. The encoder codes $\mathbf{X}_1^n$ into binary with no compression (this requires $\lceil n \log A \rceil$ bits), and we assume

[1] Define $S_k = \mathbf{Z}_{k-M+1}^k$. The sequence $\{S_k\}$ is a (memory 1) Markov chain. We say that $\{Z_k\}$ is *aperiodic*, if the chain $\{S_k\}$ is aperiodic.

that $\mathbf{X}_1^n$ is available to the decoder. The encoding procedure for $\mathbf{Z}$ is as follows. Let L_1 be the largest integer such that a copy of $\mathbf{Z}_1^{L_1}$ is a sub-string of $\mathbf{X}_1^n$, with

$$\mathbf{Z}_1^{L_1} = \mathbf{X}_{m_1}^{m_1+L_1-1}, \quad m_1 \in [1, n - L_1 + 1] \ . \tag{2}$$

Then $\mathbf{Z}_1^{L_1}$ (called a *phrase*) is encoded by a binary representation of m_1 (which requires $\log n$ bits), and a binary representation of L_1 (which requires about $\log L_1$ bits). With knowledge of (m_1, L_1) and $\mathbf{X}_1^n$, the decoder can reconstruct $\mathbf{Z}_1^{L_1}$ from (2). The entire process is now repeated for $Z_{L_1+1}, Z_{L_1+2}, \ldots$, to obtain $(m_2, L_2), (m_3, L_3)$ etc. Note that we can think of $\mathbf{X}_1^n$ as a "dictionary" which defines the coding scheme, and the parameter n as representative of the "complexity" of the code. Thus described, the FDLZ is a "variable-to-variable" length data compression scheme (which forms variable length phrases and codes them with a variable number of bits).

Suppose that we use this scheme to encode K repetitions of the process. We produce K phrases of length L_k and will require on average $K \log n + E \sum_{k=1}^{K} \log L_k$ bits to encode, on average, $E \sum_{k=1}^{K} L_k$ bits. Thus the average rate is

$$R_n = \lim_{K \to \infty} \frac{K \log n + E \sum_{k=1}^{K} \log L_k}{E \sum_{k=1}^{K} L_k}, \tag{3}$$

where the indicated expectation is with respect to the data-sequence $\mathbf{Z}$ and the training-sequence $\mathbf{X}_1^n$ [2] . Note that since we amortize the bits used for transmitting $\mathbf{X}_1^n$ over the infinite data source $\mathbf{Z}$, these bits do not contribute to the rate. We point out that the compression ratio in (3) is an average over all possible samples of the code and all possible incoming sources.

There is, however, an alternative definition of compression rate which is

$$R_n = E\{ \lim_{K \to \infty} \frac{K \log n + \sum_{k=1}^{K} \log L_k}{\sum_{k=1}^{K} L_k} \} \tag{4}$$

The main difference between definition (3) and definition (4) is that in the former, we define the compression rate as a ratio of expectations. In the latter, the compression rate is defined as an

[2] All logarithms in this paper are base 2, unless stated otherwise.

expectation of a ratio. It has been suggested that the definition (3) lacks a direct operational relationship to data compression. In fact, both definitions are operational. We now endow each with an operational situation so that in each setting there will be an *unequivocal* definition of the compression ratio. Then we display a random code whose compression rate when computed according to (3) is not identical to the compression rate when computed according to (4). Thus we will show that the very definition of compression rate can only be distinguished by the contextual setting in which the code is applied.

Setting A: Generate N independent and identically distributed copies of the (random) code. Distribute this code to N encoders who are each given a single M "megabyte" hard drive to store data. Next, generate N independent identically distributed sequences $\mathbf{Z}$ and ask each of the encoders to store on their M megabyte hard drive as many symbols of their sequence as they can. The achieved compression rate is the total memory which is NM bits, dvided by the the sum of the lengths of the longest prefix of each $\mathbf{Z}$ that can be squeezed onto the encoder's hard drives. We then compute the compression rate by letting N and M go to infinity. We remark that this situation is most akin to a communications context where the space limitation imposed by a hard drive is analogous to a time constraint imposed by a channel open only for a fixed time interval. In the communications setting, the goal would be to trasmit as much information over the channel in the alotted time. We point out that the compression rate that arises in this setting is really definition (3).

Setting B: As before, begin with N independent copies of the code distributed to N encoders. This time each encoder will receive finite sequences $\mathbf{Z}_1^M$ that are each independent and identically distributed. The i^{th} encoder would code his sequence of source symbols using a random B_i bits. The compression achieved by the i^{th} encoder is $\frac{B_i}{M}$. Clearly, in this setting the compression rate is the average value of $\frac{B_i}{M}$ over all N encoders, evaluated in the limit as N and M go to infinity. We point out that the compression rate in the setting is definition (4).

We now display a random code and a stationary ergodic source $\mathbf{Z}$ whose compression rates when computed under definitions (3) and (4) are NOT the same. Let $\mathbf{Z}$ be a binary, i.i.d sequence with entropy 1/2. Define a random code consisting of two equally likely outcomes: a "good" code that compresses every two bits of $\mathbf{Z}$ into 1 bit (with high $\mathbf{Z}$ probability) or a "bad" code that achieves no compression by simply copying the symbols of $\mathbf{Z}$. For such a code can there be a well defined compression rate? If the code is applied in setting B then the correct definition is (4) and the average compression rate is 3/4, since half of all potential encoders will achieve a rate of 1/2 and the other half a rate of 1, which averages to 3/4. However, if the operational context is setting A then the appropriate definition is (3). In that context, half the encoders would store $2M$ bits of $\mathbf{Z}$ and the other half of the encoders would store M bits. The total number of symbols encoded is $2M\frac{N}{2} + M\frac{N}{2} = \frac{3NM}{2}$ bits, stored on a total of NM bits. The compression rate is therefore 2/3 under definition (3).

Which is the correct compression rate? Neither is necessarily correct, although we conjecture that the two definitions are identical if the code itself is somehow stationary and ergodic. For the FDLZ that would certainly be true, since we conjecture that for large n all databases produce nearly equivalent samples of the code. Since we believe that the definition of compression that is most important to information theorists is the one that has a direct relationship to communications, we therefore choose to use definition (3) thoughout this paper.

3 Results

The exact analysis of the algorithm's performance is centered around a decomposition of each of the phrase lengths into two pieces. To this end we define the function $T_n(\mathbf{z}_1^\infty)$, for a sequence $\{z_k\}_1^\infty$, $z_k \in \mathcal{A}$, by

$$T_n(\mathbf{z}_1^\infty) = \min\left\{k \geq 1 : \ \Pr\left\{\mathbf{Z}_1^k = \mathbf{z}_1^k\right\} \leq 1/n\right\} . \tag{5}$$

Thus $T_n(\mathbf{Z})$ is a random variable (since it is a function of the random sequence $\mathbf{Z}$). The major theorem of this paper finds the relationship of this random variable to the phrase lengths. Before we state the theorem, we try to provide some intuition for how $T_n(\mathbf{Z})$ relates to the parsing algorithm:

Consider a fixed sequence $\mathbf{z}_1^\infty$. The probability (with respect to the probability law of $\mathbf{X}_1^n$) of any prefix of length k of $\mathbf{z}_1^\infty$ decreases as k increases. For a given prefix of length k let $W(\mathbf{z}_1^k)$ be the number of times that the prefix appears in the database. We can express this quantity using indicators

$$W(\mathbf{z}_1^k) = \sum_{i=1}^{n-k} 1(\mathbf{X}_i^{i-k-1} = \mathbf{z}_1^k).$$

Taking expectations and using the stationarity of $\mathbf{X}_1^n$ and we have that

$$E_{\mathbf{X}} W(\mathbf{z}_1^k) = (n-k) P(\mathbf{X}_1^k = \mathbf{z}_1^k).$$

Thus, with $k = T_n(\mathbf{z}_1^\infty)$, we have that the expected number of matches of z_1^k is

$$(n-k)P(\mathbf{X}_1^k = \mathbf{z}_1^k) = (n-k)P(\mathbf{Z}_1^k = \mathbf{z}_1^k) \approx \frac{n-k}{n} \approx 1$$

Furthermore, for prefixes of $\mathbf{z}_1^\infty$ that are very much longer than k, the expected number of matches is far less than 1; similarly for prefixes that are very much shorter than k the expected number of matches is far larger than 1. It is reasonable to expect that the length of the longest prefix that occurs in the database (this is L_1) should be very nearly $T_n(\mathbf{z}_1^\infty)$ since that is the length of a prefix whose expected number of matches is closest to 1. Given this intuition it seems reasonable to decompose the phrase length into two pieces,

$$L_1 = T_n(\mathbf{Z}) + \Delta, \tag{6}$$

with the hope that Δ will be small. This recognizes that the phrase length is nearly determined by the random variable $T_n(\mathbf{Z})$ (which is a function of the incoming data $\mathbf{Z}$) and an adjustment for the difference.

The key mathematical result is an approximation of the distribution of Δ. A proof is given is Section 4.

Theorem 3.1 *Let $\Delta = L_1 - T_n(\mathbf{Z})$. Then for every non-negative integer j, there exists a positive constant $\alpha < 1$, and a positive constant δ such that*

$$\Pr\{\Delta \geq j\} \leq \alpha^j + O(\frac{\log n + j}{n^\delta})\alpha^j. \tag{7}$$

$$\Pr\{\Delta \leq -j\} \leq \exp(-\alpha^{-j}) + O(\frac{(\log n)^2}{n^\delta})\alpha^{-j} \tag{8}$$

This theorem tell us that the right hand tail of Δ is exponential and that the left hand tail is doubly exponential. Furthermore, both tails have distributions that are *independent* of n, and thus Δ is $O(1)$. Thus Theorem 3.1 proves , asymptotically in n, that the phrase length L_1 and $T_n(\mathbf{Z})$

are approximately equal. It is worth pointing out that this result is proved by means of Poisson Approximation and Stein's method, though it is shown in [14] that these powerful tools are not completely necessary. Since the distribution of $T_n(\mathbf{Z})$ (and all its moments) can be determined using well established techniques (renewal theory, random walks), the asymptotic equivalence of L_1 and $T_n(\mathbf{Z})$ leads to the following theorems (we reserve the proofs for section 4).

Theorem 3.2 *Let* $\sigma^2 = Var \log P(X_{m+1}|\mathbf{X}_1^m)$*, where* m *is the memory of the source. Then, if* $H < \log |A|$ *and* $\sigma > 0$*, we have as* $n \to \infty$*,*

$$\frac{L_k - \frac{\log n}{H}}{\sigma\sqrt{\log n / H^3}} \Rightarrow N(0,1).$$

However, if the source is memoryless and uniformly random over the alphabet, with entropy $H = \log |A|$ (and therefore incompressible), we have a different result

Theorem 3.3 *If* $H = \log |A|$*, then the distribution of* L_k*, for any integer* j *and every* n*, is*

$$\Pr\{L_k < j + \frac{\log n}{H}\} = \exp(-|\mathcal{A}|^{-j}) + O(\frac{\log n + j}{n}).$$

We remark that both Theorem 3.2 and Theorem 3.3 provide the distribution of all the phrases in a fixed Lempel-Ziv parse. This is deceptively difficult mathematical task since the phrases are not identically distributed or even independent (not even for memoryless sources); this, due to the construction of every phrase through a common database.

The major mathematical task in calculating the redundancy from the definition in (3), is to find $E \log L_k$ and EL_k.We will prove the following theorem in section 4:

Theorem 3.4 *As* $n \to \infty$*,* $EL_k = \frac{\log n}{H} + O(1)$.

To illustrate the power of Theorem 3.4 we prove a corollary which provides bounds (both upper and lower) on the redundancy, ρ_n.

Corollary *As* $n \to \infty$*,* $\rho_n \approx \frac{H \log\log n}{\log n}$.

Proof: First, upper bound the redundancy using Theorem 3.4 and Jensen's inequality; whereby (3) reduces to

$$R_n \leq \frac{\log n + \log\log n + O(1)}{\frac{\log n}{H} - O(1)}.$$

Hence,

$$\rho_n = R_n - H \leq \frac{H \log\log n}{\log n} + O(\frac{1}{\log n}).$$

To find the lower bound we need an extension of a result from [12] [3], (which is also found in [9]) which holds that for any $\epsilon > 0$ there exists an N such that for all $n > N$

$$\Pr\{L_k > \frac{\log n(1-\epsilon)}{H}\} > 1 - \epsilon. \tag{9}$$

Thus, with high probability the phrase lengths are "close" to their common mean of $\frac{\log n}{H}$. From this result and the Markov inequality we have that

$$E \log L_k \geq \log(\frac{\log n}{H}[1-\epsilon]) \Pr\{L_k > \frac{\log n}{H}[1-\epsilon]\} \geq \log\log n - o(\log n \log n).$$

Now (3) simplifies to

$$\rho_n \geq H\frac{\log\log n}{\log n} + o(\frac{\log\log n}{\log n}).$$

This is the lower bound. •

Finally, we point out that there is an immediate application of Theorem 3.2 towards increasing the algorithm's efficiency. To see this, consider a variation of the FDLZ which encodes not the successive phrase lengths L_i, but the successive difference in phrase lengths: $L_i - L_{i-1}$. Since the standard deviation of the phrase length is (given by Theorem 3.2) $O(\sqrt{\log n})$, we see from (3) and Jensen's inequality, that the redundancy of this version is upper bounded (to first order approximation) by $\frac{H \log\log n}{2\log n}$. To see this directly, observe that the magnitude of the difference in phrase lengths is proportional to the standard deviation of the phrase length distribution, which (by Theorem 3.2) is smaller than the expected value of the phrase length by a square root factor. The translates to an improvement in redundancy of at least a factor of two. In [14] it has been demonstrated that the algorithm can be modified further to reduce the redundancy down to $O(\frac{1}{\log n})$. Yet, this bound is only upper bound; it is still unknown if any more "performance" can be squeezed out of Lempel-Ziv like algorithms. Recent results have shown, however, that a similar version, without the fixed

[3] The inequality in [12] applies only to L_1. The extension to L_k will be discussed in the proof of Theorem 3.4

databse restriction, of the Lempel-Ziv algorithm (the LZ-78 version) performs with an asymptotic redundancy equal to at least $O(1 \log n)$. Thus we conjecture, surprisingly, that the LZ-78 version, which parses sequences into distinct phrases, may perform better than the algorithm considered here (which is the FDLZ-77). Thus the improvements made in [14] are needed to restore the parity in performance.

There have been several recent articles that have considered the asymptotic distributional properties of Lempel-Ziv phrases in various contexts, all quite different from than that which is considered here, but still fundamentally similar in spirit. Our Theorems 3.2,3.3 and 3.4 are generalizations of Proposition 2 in Jacquet and Szpankowski [6]. There it is shown, that for asymmetric memoryless sources, a quantity labeled D_n (which in the context of FDLZ is almost exactly L_1) satisfies a normal limit law (they compute the mean and variance) and in the case of uniform random sources (symmetric memoryless source) that the following holds:

$$\lim_{n \to \infty} \sup_x |\Pr\{D_n \leq x\} - \exp(-n|\mathcal{A}|^{-x})| = 0.$$

We can, by changing variables (replacing j with an appropriate choice of x), recover these limits. In fact, we have proved much more. Theorem's 3.2 and 3.4 extend the aforementioned proposition to Markov sources and provide an approximate bound on the rate of convergence. Theorem 3.3 (the uniform memoryless case) provides an explicit rate of convergence that is the basis for a potentially useful statistical test for pure randomness (see [13]).

There is a more fundamental difference here, however. To apply these results to a direct analysis of the data compression algorithm, we must must generalize the results to *all* phrases. It is insufficient to find the behavior of L_1; the behavior of L_k is needed for all k. How to accomplish this is not immediately obvious even in the memoryless case; the distribution of the length of every phrase depends not only on the database but on every previous phrase. Therefore, only the first phrase L_1, can be handled with relative ease, even though for Markov sources this dependence should be insignificant. Indeed it is; but it requires proof.

Szpankowski and Louchard in [7] take a different attack, presenting an analysis of the so called phrase length "average profile". This they define as the length of a randomly selected phrase. One can not establish redundancy from this quantity. In fairness, they are considering a version of the Lempel-Ziv algorithm that has much more complicated dependencies, the so called LZ-78, which undertakes a different parsing scheme than that presented here (which is the LZ-77 version with a fixed-data base restriction).

Lastly the final difference is in method. While our approach is probabilistic, their's is analytic. The first fruit off of this probabilistic tree is Theorem 3.1. It is our conjecture that this theorem concisely captures the essence of many Lempel-Ziv, string matching based, coding algorithms.

4 Proofs of Theorems

We begin with some definitions, and then proceed with a series of lemmas that will lay the foundation for the proofs of the theorems. First, let $\mathbf{z}_1^k$, $k \leq \infty$, $n < \infty$, be any sequence from the alphabet $\mathcal{A}$. Define $L(\mathbf{z}_1^\infty)$ to be the largest integer L, such that a copy of string $\mathbf{z}_1^L$ is a substring of $\mathbf{X}_1^n$, i.e.

$$\mathbf{z}_1^L = \mathbf{X}_m^{m+L-1}, \quad \text{for some} \quad m \in [1, n-L+1] \ . \tag{10}$$

We remark that the random variable $L(\mathbf{Z})$ is the longest prefix of of $\mathbf{Z}$ that occurs in the database, thus it is immediately identified as the length of the first phrase in the above version of the L-Z algorithm. Later we will consider $L(\mathbf{Z}_k^\infty)$ which is the longest match in the database of the source sequence starting at time k. We point out that $L(\mathbf{Z})$ is a random function of the database *and* the source, while $L(\mathbf{z}_1^\infty)$ is a random function only of the database.

We show in Appendix A that with no loss of generality we can take the memory-size $M = 1$, i.e. (1) is

$$\Pr\left\{Z_k = z_k \mid \mathbf{Z}_{-\infty}^{k-1} = \mathbf{z}_{-\infty}^{k-1}\right\} = \Pr\left\{Z_k = z_k \mid Z_{k-1} = z_{k-1}\right\} \ . \tag{11}$$

We assume that

$$\alpha \triangleq \max_{a,b \in \mathcal{A}} \Pr\left\{Z_k = b \mid Z_{k-1} = a\right\} < 1 \ . \tag{12}$$

and that

$$\beta \triangleq \max_{a,b \in \mathcal{A}} \Pr\left\{Z_k = b \mid Z_{k+1} = a\right\} < 1 \ . \tag{13}$$

We also show in Appendix A how to prove all of the theorems when $\alpha = 1$ and $\beta = 1$. We will make use of the fact that for $i < j < k$, and arbitrary $\mathbf{z}$,

$$\Pr\left\{\mathbf{Z}_j^k = \mathbf{z}_j^k \mid \mathbf{Z}_i^{j-1} = \mathbf{z}_i^{j-1}\right\} = \Pr\left\{\mathbf{Z}_j^k = \mathbf{z}_j^k \mid Z_{j-1} = z_{j-1}\right\}$$
$$= \prod_{i=j}^{k} \Pr\left\{Z_i = z_i \mid Z_{i-1} = z_{i-1}\right\} \leq \alpha^{k-j+1} . \tag{14}$$

We will also use the fact that for $j < k < i$, and arbitrary $\mathbf{z}$,

$$\Pr\left\{\mathbf{Z}_j^k = \mathbf{z}_j^k \mid \mathbf{Z}_{k+1}^i = \mathbf{z}_{k+1}^i\right\} = \Pr\left\{\mathbf{Z}_j^k = \mathbf{z}_j^k \mid Z_{k+1} = z_{k+1}\right\}$$
$$= \prod_{i=j}^{k} \Pr\left\{Z_i = z_i \mid Z_{i+1} = z_{i+1}\right\} \leq \beta^{k-j+1} . \tag{15}$$

Finally, we also define

$$\alpha_{\mathrm{MIN}} \triangleq \min_{a,b \in \mathcal{A}} \left\{\Pr\left\{Z_1 = b \mid Z_0 = a\right\} > 0\right\} \tag{16}$$

$$\beta_{\mathrm{MIN}} \triangleq \min_{a,b \in \mathcal{A}} \left\{\Pr\left\{Z_0 = b \mid Z_1 = a\right\} > 0\right\}. \tag{17}$$

4.1 Outline

Let $\ell(\mathbf{z}_1^\infty)$ be a positive integer valued function of infinite sequences in $\mathcal{A}$ and let $\mathbf{X}_{-\infty}^\infty$ be the doubly infinite extension of the training sequence (abbreviate as $\mathbf{X}$). Thus the distribution of the *random variable* $\ell(\mathbf{Z}_1^\infty)$ is a function on the incoming data $\mathbf{Z}$ and is therefore independent of $\mathbf{X}$. For ease of presentation we will write ℓ instead of the cumbersome $\ell(\mathbf{Z}_1^\infty)$, but the reader should keep the dependence on $\mathbf{Z}$ and the independence on $\mathbf{X}$ in mind.

Define $R(\mathbf{z}_1^\ell)$ to be the smallest integer N such that $\mathbf{X}_N^{N+\ell-1}$ is a copy of $\mathbf{z}_1^\ell$. In plain language, $R(\mathbf{z}_1^\ell)$ is the first time the sequence $\mathbf{z}_1^\ell$ appears in $\mathbf{X}_1^\infty$. The connection of the waiting time $R(\mathbf{z}_1^\ell)$, to the longest match $L(\mathbf{Z})$ is through the following equivalence: $\{R(\mathbf{Z}_1^\ell\} > n\} = \{L(\mathbf{Z}) < \ell\}$.
Let us recall the random variable

$$W(\mathbf{z}_1^\ell) = \sum_{i=0}^{n-\ell} 1(\mathbf{X}_i^{i+\ell-1} = \mathbf{z}_1^\ell). \tag{18}$$

Notice that

$$\{R(\mathbf{z}_1^\ell\} > n\} = \{W(\mathbf{z}_1^\ell) = 0\},$$

and hence, for a random sequence $\mathbf{Z}_1^\ell$ (of random length) that

$$\Pr\{L_1 < \ell\} = \Pr\{L(\mathbf{Z}) < \ell\} = P\{W(\mathbf{Z}_1^\ell) = 0\}. \tag{19}$$

Consequently, if we can establish that $W(\mathbf{z}_1^\ell)$ is Poisson (for most choices of $\mathbf{z}_1^\ell$) then we could show that $\Pr\{L_1 < \ell\} \approx \exp(-\lambda(\mathbf{z}_1^\ell))$, where $\lambda(\mathbf{z}_1^\ell) = E_{\mathbf{X}} W(\mathbf{z}_1^\ell)$ is the expected number of matches of $\mathbf{z}_1^\ell$ in the database. If we average again over the source $\mathbf{Z}$ with $\lambda \triangleq E_{\mathbf{Z}} E_{\mathbf{X}} W(\mathbf{Z}_1^\ell)$ we would prove that

$$\Pr\{L_1 < \ell\} \approx \exp(-\lambda).$$

In light of this discussion we can outline the basic steps of the proof of Theorem 3.1. First, we bound the difference between $\Pr\{W(\mathbf{z}_1^\ell) = 0\}$ and $\exp(-\lambda(\mathbf{z}_1^\ell))$ (the Poisson approximation of that probability). This will be Lemma 4.1. It turns out that this bound is good for most choices of $\mathbf{z}_1^\ell$, but we will only need to show that the bound is good for a random choice of the sequence $\mathbf{Z}_1^\ell$ and for a particular choice of ℓ (namely $\ell = T_n(\mathbf{Z}) + j$). This will be Lemma 4.2. Then with $\Delta = \ell - T_n(\mathbf{Z})$ we can use (19) to approximate $\Pr\{\Delta < j\}$.

Theorem 3.2 will follow from Theorem 3.1 in three steps. First we use renewal theory to show that the distribution of $T_n(\mathbf{Z})$ is Normal. Then we apply Theorem 3.1 to show that L_1 is normal. Finally, for every k we relate L_k to L_1 and prove that L_k is also Normal.

Observe that (18) implies that $W(\mathbf{z}_1^\ell)$ is the sum of a sequence of dependent indicator random variables with $Y_i = 1(\mathbf{X}_i^{i+\ell-1} = \mathbf{z}_1^\ell)$ and $W(\mathbf{z}_1^\ell) = \sum_{i=1}^{n-\ell} Y_i$. With fixed $\mathbf{z}_1^\infty$ and fixed [4] ℓ the random variables Y_i are only weakly dependent. If we define

$$\Phi = \sum_{i=-2\ell}^{2\ell} Y_i.$$

then, Φ is the number of matches of $\mathbf{z}_1^\ell$ that occur in the database somewhere in a 2ℓ-neighborhood of X_0. The first Lemma of the next section shows that the quality of the Poisson approximation will depend on the behavior of the quantity:

$$B(\mathbf{z}_1^\ell) \triangleq E_{\mathbf{X}}\Phi + E_{\mathbf{X}}(\Phi | Y_0 = 1) - 1. \tag{20}$$

[4] ℓ is necessarily fixed given $\mathbf{z}_1^\infty$.

4.2 Key Lemmas

Lemma 4.1 *Given a sequence* $\mathbf{z}_1^\infty$ *and an integer* $\ell = \ell(\mathbf{z}_1^\infty)$ *with* $W(\mathbf{z}_1^\ell)$ *is defined as in (18) and* $\lambda(\mathbf{z}_1^\ell) = E_{\mathbf{X}} W(\mathbf{z}_1^\ell)$*, then there exists a constant* γ*, with* $0 \leq \gamma < 1$ *such that the following holds*

$$|E_{\mathbf{X}} 1(W(\mathbf{z}_1^\ell) = 0) - \exp(-\lambda(\mathbf{z}_1^\ell))| \leq (1 \wedge \lambda(\mathbf{z}_1^\ell))[B(\mathbf{z}_1^\ell) + O(\gamma^\ell)]. \tag{21}$$

Furthermore, if the distribution of $\mathbf{X}$ *is memoryless, then* $\gamma = 0$.

Proof: See appendix B.

Thus Lemma 4.1 is a Poisson approximation of a probability whose accuracy we can evaluate be probing the terms in the right hand side of (21) First we consider $\lambda(\mathbf{z}_1^\ell)$. This we do by taking expectations in (18) with respect to $\mathbf{X}$. Using the stationarity of $\mathbf{X}$ and the linearity of expectations we have

$$\lambda(\mathbf{z}_1^\ell) = (n - \ell) P(\mathbf{z}_1^\ell). \tag{22}$$

Next we evaluate the expected value of Φ which by definition is the number of times that indicator random variables Y_i are equal to 1 in a 2ℓ-neighborhood of zero. The expectation of Φ, by linearity is

$$E_{\mathbf{X}} \Phi = E \sum_{i=-2\ell}^{2\ell} Y_i = (4\ell + 1) P(\mathbf{z}_1^\ell). \tag{23}$$

Finally, we evaluate the conditional expectation of Φ given that $Y_0 = 1$: we write

$$\begin{aligned} E_{\mathbf{X}}[\Phi | Y_0 = 1] &= E_{\mathbf{X}}[\sum_{i=-2\ell}^{2\ell} Y_i | Y_0 = 1] \\ &= E_{\mathbf{X}}[\sum_{i=-2\ell}^{2\ell} 1(\mathbf{X}_i^{i+\ell} = \mathbf{z}_1^\ell) | \mathbf{X}_1^\ell = \mathbf{z}_1^\ell] \\ &= \sum_{i=-2\ell}^{2\ell} E_{\mathbf{X}}[1(\mathbf{X}_i^{i+\ell} = \mathbf{X}_1^\ell) | \mathbf{X}_1^\ell = \mathbf{z}_1^\ell] \\ &= \sum_{i=-2\ell}^{2\ell} \Pr(\mathbf{X}_i^{i+\ell} = \mathbf{X}_1^\ell | \mathbf{X}_1^\ell = \mathbf{z}_1^\ell) \end{aligned} \tag{24}$$

Thus

$$B(\mathbf{z}_1^\ell) = (4\ell + 1) P(z_1^\ell) + \sum_{i=-2\ell}^{2\ell} \Pr(\mathbf{X}_i^{i+\ell} = \mathbf{X}_1^\ell | \mathbf{X}_1^\ell = \mathbf{z}_1^\ell) - 1. \tag{25}$$

It is intuitively clear that for most $\mathbf{z}_1^\infty$ that $B(\mathbf{z}_1^\ell)$ tends to 0 as $\ell \to \infty$. The only exceptions are sequences $\mathbf{z}_1^\ell$ that display a high degree of self-symmetry; for example 0_1^ℓ. Then the bound in (25) is at least $\Pr\{X_2^{1+\ell} = 0_1^\ell | \mathbf{X}_1^\ell = 0_1^\ell\} = \Pr\{X_{1+\ell} = 0 | X_\ell = 0\}$, which does not tend to zero.

We now state and prove a Lemma which considers that average value of $B(\mathbf{Z}_1^\ell)$ with respect to a random source sequence $\mathbf{Z}$ and with ℓ defined (as outlined above) for any integer j to be

$$\ell = T_n(\mathbf{Z}) + j. \tag{26}$$

Here we define $Y_i = 1\{\mathbf{X}_i^{i+\ell-1} = \mathbf{Z}_1^\ell\}$ and $\Phi = \sum_{i=-2\ell}^{2\ell} Y_i$. This reflects the fact that we are now considering not only a random database $\mathbf{X}$ but also a random pattern $\mathbf{Z}_1^\ell$.

Lemma 4.2 *Let $\ell = T_n(\mathbf{Z})+j$, and let $E_{\mathbf{Z}}B(\mathbf{Z}_1^\ell) = E_{\mathbf{Z}}E_{\mathbf{X}}\Phi + E_{\mathbf{Z}}E_{\mathbf{X}}(\Phi|Y_0 = 1) - 1$. Then, if $j \geq 0$, we have*

$$E_{\mathbf{Z}}B(\mathbf{Z}_1^\ell) \leq 4(\eta_2 + j + 1)(\frac{1}{n} + \frac{1}{n^{\frac{\mu_{\mathrm{MIN}}}{\mu}}})\alpha^j. \tag{27}$$

If $j < 0$ we have

$$E_{\mathbf{Z}}B(\mathbf{Z}_1^\ell) \leq 4(\eta_2 + j + 1)(\frac{1}{n} + \frac{1}{n^{\frac{\mu_{\mathrm{MIN}}}{\mu}}})\alpha_{\mathrm{MIN}}^j. \tag{28}$$

Remark: If more precise knowledge of the phase length is desired, and the distribution of the source is known, then much better bounds can be obtained by evaluating these expectations explicitly. We do this type of calculation in the proof of Theorem 3.3.

Proof: The first step in the proof is to bound the size of ℓ for all source sequences $\mathbf{Z}$. Define

$$\mu = \log\frac{1}{\alpha}, \quad \mu_{\mathrm{MIN}} = \log\frac{1}{\alpha_{\mathrm{MIN}}}. \tag{29}$$

Thus, with $T_n(\mathbf{z})$ defined by (5), for all[5] $\mathbf{z} = \mathbf{z}_1^\infty$

$$\eta_1 \triangleq \frac{\log n}{\mu_{\mathrm{MIN}}} \leq T_n(\mathbf{z}) \leq \frac{\log n}{\mu} + 1 \triangleq \eta_2. \tag{30}$$

With $\ell = T_n(z_1^\infty) + j$ and all sequences $\mathbf{z}_1^\infty$, we have

$$\eta_1 + j \leq \ell \leq \eta_2 + j. \tag{31}$$

Observe that we may write, for $j > 0$,

$$P(\mathbf{z}_1^\ell) = \Pr(\mathbf{Z}_1^{T_n} = \mathbf{z}_1^{T_n}) \prod_{i=T_n+1}^{T_n+j} \Pr\{Z_i = z_i | Z_i = z_i\}. \tag{32}$$

[5]We restrict ourselves to those sequences $\mathbf{z}_1^\infty$ such that for $j = 1, 2, \ldots$, $P(\mathbf{z}_1^j) > 0$. The probability of these $\mathbf{z}$'s is 1.

From (5) we know that $P(\mathbf{z}_1^{T_n}) \leq \frac{1}{n}$. So (32) in light of (14) proves that

$$P(\mathbf{z}_1^{\ell}) \leq \frac{1}{n}\alpha^j \tag{33}$$

and thus the bound

$$E_{\mathbf{Z}}E_{\mathbf{X}}\Phi = \leq 4(\eta_2+j+1)\max_{\mathbf{z}_1^{\ell}} P(\mathbf{z}_1^{\ell}) \leq 4(\eta_2+j+1)\frac{1}{n}\alpha^j, \tag{34}$$

follows from (23) and (31). We have a reverse inequality for $j<0$ using the same argument as well as (15) to show that

$$P(\mathbf{z}_1^{\ell}) \leq \frac{1}{n}\alpha_{\mathrm{MIN}}^j. \tag{35}$$

Thus, from (23) and (31) imply that

$$E_{\mathbf{Z}}E_{\mathbf{X}}\Phi = \leq 4(\eta_2+j+1)\max_{\mathbf{z}_1^{\ell}} P(\mathbf{z}_1^{\ell}) \leq 4(\eta_2+j+1)\frac{1}{n}\alpha_{\mathrm{MIN}}^j. \tag{36}$$

An upper bound on $E_{\mathbf{Z}}E_{\mathbf{X}}(\Phi|Y_0=1)$ is obtained beginning with (24) (with $j \geq 0$):

$$\begin{aligned}
E_{\mathbf{Z}}E_{\mathbf{X}}(\Phi|Y_0=1)-1 &\overset{(a)}{=} \sum_{i=-2\ell}^{2\ell}\Pr\{\mathbf{X}_{i+1}^{i+\ell}=\mathbf{X}_1^{\ell}|\mathbf{X}_1^{\ell}=\mathbf{Z}_1^{\ell}\}-1\\
&\overset{(b)}{=} \sum_{i=-2\ell}^{2\ell}\Pr\{\mathbf{X}_{i+1}^{i+\ell}=\mathbf{X}_1^{\ell}\}-1\\
&\overset{(c)}{=} \sum_{i=-2\ell}^{-1}\Pr\{\mathbf{X}_{i+1}^{i+\ell}=\mathbf{X}_1^{\ell}\}+\sum_{i=1}^{2\ell}\Pr\{\mathbf{X}_{i+1}^{i+\ell}=\mathbf{X}_1^{\ell}\}\\
&\overset{(d)}{\leq} \sum_{i=-2(\eta_2+j)}^{-1}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}+\sum_{i=1}^{2(\eta_2+j)}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}\\
&\overset{(e)}{=} \sum_{i=-2(\eta_2+j)}^{-(\eta_1+j)}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}+\sum_{i=-(\eta_1+j)+1}^{-1}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}\\
&\quad+\sum_{i=1}^{\eta_1+j-1}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}+\sum_{i=\eta_1+j}^{2(\eta_2+j)}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}\\
&\overset{(f)}{\leq} 2(\eta_2-\eta_1)\beta^{\eta_1+j}+2(\eta_2-\eta_1)\alpha^{\eta_1+j}\\
&\quad+\sum_{i=-(\eta_1+j)+1}^{-1}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}+\sum_{i=1}^{\eta_1+j-1}\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j}=\mathbf{X}_1^{\eta_1+j}\}\\
&\overset{(g)}{\leq} 2(\eta_2-\eta_1)\beta^{\eta_1+j}+2(\eta_2-\eta_1)\alpha^{\eta_1+j}+(\eta_1+j)\alpha^{\eta_1+j}+(\eta_1+j)\beta^{\eta_1+j}\\
&\overset{(h)}{\leq} 2\Big(\frac{\eta_2+j}{n^{\frac{\mu_{\mathrm{MIN}}}{\mu}}}\Big)\alpha^j+2\Big(\frac{\eta_2+j}{n^{\frac{\mu_{\mathrm{MIN}}}{\mu}}}\Big)\beta^j
\end{aligned} \tag{37}$$

Step (a) follows by taking the expectation over $\mathbf{Z}_1^\ell$ of Φ_2 in (24). Step (b) follows from the definition of conditional probability and the fact that $\mathbf{X}$ and $\mathbf{Z}$ are independent and identically distributed. Step (c) considers the sum into two pieces; matches that occur before time 1 and those that occur after time 1. Step (d) follows since $\eta_1+j \leq \ell \leq \eta_2+j$ and the fact that matches of shorter sequences are more likely to occur than longer ones. Step (e) further splits the summation into those matches that overlap $\mathbf{X}_1^{\eta_1+j}$ and those that don't. Step (f) bounds the probability of a non-overlapping match using (14). Step (g) is a technical issue and is justified by lemma 4.9 (see Appendix B), and step (h) totals the terms.

Equation (27) follows by totaling (36) and (37). Without loss of generality assume that $\alpha > \beta$. Otherwise, replace α with β in (27). Finally, if $j < 0$ we repeat the step with α_{MIN} instead of α and β_{MIN} instead of β. $\bullet$.

We complete our series of lemmas with the following:

Lemma 4.3

$$ET_n(\mathbf{Z}) = \frac{\log n}{H} + O(1). \tag{38}$$

Proof: Observe that $T_n(\mathbf{Z})$ is a stopping time for the Markov process $\mathbf{Z}$ with respect to $\sigma(\mathbf{Z}_1^n)$. We would like to apply the Wald's theorem to a random walk, so let $S_0 = -\log P(Z_0)$ and define

$$S_k \triangleq S_{k-1} - \log P(Z_k|Z_{k-1}).$$

Recognize that

$$T = T_n(\mathbf{Z}) = \inf\{k : S_k \geq \log n\}.$$

But $\mathbf{Z}$ is not i.i.d. so we cannot apply Wald's theorem directly. Instead, form a set of conditional stopping times indexed by every possible opening state $a \in \mathcal{A}$ that may precede Z_1;

$$T_a = \inf\{k : S_k \geq \log n; Z_k = a | Z_0 = a\}.$$

The idea here is to construct a new process formed by parsing the original sequence into blocks of excursions from each occurrence of state a to the next. To this end, we define $M_0(a) = 0$

and let $M_i(a) = \inf\{k : Z_k = a \text{ and } k > M_{i-1}(a)\}$. The excursion blocks are the sequences $V_i(a) = \mathbf{Z}_{M_{i-1}(a)+1}^{M_i(a)}$. The excursion random walk starting in state a is defined to be

$$Q_k(a) = \sum_{i=1}^{k} -\log P(V_i(a)).$$

This is simple random walk since these excursion blocks are independent and identically distributed. If we define $N(T_a)$ to equal the number of excursions that have occurred by time T_a, we have that $S_{T_a} = Q_{N(T_a)}$. We know from [4] that for the Markov chain that starts in a state a, the "entropy of the trajectory" is given by

$$E[-\log P(\mathbf{Z}_1^{M(a)})] = E[M(a)]\,H$$

where $M(a) = M_1(a)$ (this is lemma 4.10 and is proved in Appendix B).

If we apply this result in conjunction with Wald's theorem to the random walk $Q_{N(T_a)}$ we have that

$$EQ_{N(T_a)} = EN(T_a)EM(a)\,H.$$

Now, since $S_{T_a} = Q_{N(T_a)}$ it must be that $EQ_{N(T_a)} \geq \log n$. This, and the fact that there is, on average, a multiplicative factor of $EM(a)$ more steps in the original process for every step in the excursion process (that is $ET_a = EN(T_a)EM(a)$), leads to

$$\log n \leq EQ_{N(T_a)} = EM(a)H\frac{ET_a}{EM(a)} = HET_a. \tag{39}$$

We can easily upper bound the expected value of the random walk at the stopping time since

$$EQ_{N(T_a)} \leq \log n + \alpha \max_{a\in\mathcal{A}} EM(a) = \log n + O(1),$$

which implies that for all $a \in \mathcal{A}$,

$$\log n \leq HET_a \leq \log n + O(1) \tag{40}$$

To complete the argument we define the unconditional stopping time,

$$T_Z = \inf\{k : S_k \geq \log n;\; Z_k = Z_0\}$$

where the opening state is the random variable Z_0. It follows from (40) that $ET_Z = \frac{\log n}{H} + O(1)$. To finish the lemma we bound the difference between ET to ET_Z by the the maximum expected excursion length,

$$|ET - ET_Z| \leq \max_a EM(a) = O(1).$$

This implies that $ET_n(\mathbf{Z}) = \frac{\log n}{H} + O(1)$, which completes the proof of the lemma. •

4.3 Proof of Theorem 3.1

We begin by defining, for a fixed sequence $\mathbf{z}_1^\ell$ the quantity

$$\Delta(\mathbf{z}_1^\infty) \triangleq L(\mathbf{z}_1^\infty) - T_n(\mathbf{z}_1^\infty)$$

and for a random sequence $\mathbf{Z}_1^\ell$ we define $\Delta \triangleq \Delta(\mathbf{Z})$. For a fixed string $\mathbf{z}_1^\ell$, with $\ell = \ell(\mathbf{z}_1^\infty) = T_n(\mathbf{Z})+j$ and $j > 0$ we have from (19) that

$$\Pr\{\Delta(\mathbf{z}_1^\ell) < j\} = \Pr\{W(\mathbf{z}_1^\ell) = 0\},$$

and therefore, we can use Lemma 4.1 to show that

$$|E_{\mathbf{X}} 1(\Delta(\mathbf{z}_1^\infty) < j) - \exp(-\lambda(\mathbf{z}_1^\ell))| \le (1 \wedge \lambda(\mathbf{z}_1^\ell))[B(\mathbf{z}_1^\ell) + O(\gamma^\ell)]. \tag{41}$$

From (33) and (22) we see that $\lambda(\mathbf{z}_1^\ell) \le \alpha^j$. Thus (41) reduces to

$$E_{\mathbf{X}} 1(\Delta(\mathbf{z}_1^\infty) < j) \ge \exp(-\alpha^j) - \alpha^j [B(\mathbf{z}_1^\ell) + O(\gamma^\ell)].$$

Averaging over all $\mathbf{Z}$ and hence over all ℓ, yields

$$P(\Delta < j) \ge \exp(-\alpha^j) - \alpha^j [E_{\mathbf{Z}} B(\mathbf{z}_1^\ell) + O(\gamma^\ell)] \tag{42}$$

Applying Lemma 4.2 to (42), with $\delta = \min(1, \frac{\mu_{\text{MIN}}}{\mu}, \gamma)$, and using the fact that $\exp(-\alpha^j) \le 1 - \alpha^j$,

$$P(\Delta \ge j) \le 1 - \exp(-\alpha^j) + \alpha^j O(\frac{\log n + j}{n^\delta}) \le \alpha^j + \alpha^j O(\frac{\log n + j}{n^\delta}). \tag{43}$$

This is (7).

The other direction follows the same steps, with $\ell = \ell(\mathbf{z}_1^\infty) = T_n(\mathbf{z}_1^\infty) - j$. Thus,

$$\eta_1 - j \le \ell \le \eta_2 - j \tag{44}$$

and for all $\mathbf{z}_1^\ell$

$$\alpha^{-j} \le \lambda(\mathbf{z}_1^\ell) \le \alpha_{\text{MIN}}^{-j}. \tag{45}$$

Following the steps which led to (42), using (45), Lemma 4.1, and the fact that $1 \wedge \lambda(\mathbf{z}_1^\ell) = 1$:

$$\begin{aligned}
|E_{\mathbf{X}} 1(\Delta(\mathbf{z}_1^\infty) < -j) - \exp(-\lambda(\mathbf{z}_1^\ell))| &\le (1 \wedge \lambda(\mathbf{z}_1^\ell))[B(\mathbf{z}_1^\ell) + O(\gamma^\ell)] && (46)\\
E_{\mathbf{X}} 1(\Delta(\mathbf{z}_1^\infty) < -j) &\le \exp(-\alpha^{-j}) + [B(\mathbf{z}_1^\ell) + O(\gamma^\ell)] \\
P(\Delta < -j) &\le \exp(-\alpha^{-j}) + [E_{\mathbf{Z}} B(\mathbf{z}_1^\ell) + O(\gamma^\ell)] && (47)
\end{aligned}$$

Now if we apply Lemma 4.2 to (47) we see have that

$$P(\Delta < -j) \leq \exp(-\alpha^{-j}) + 4(\eta_2 - j)(\frac{1}{n} + \frac{1}{n^{\frac{\mu}{\mu_{\text{MIN}}}}})\alpha^{-j} + O(\frac{\gamma^j}{n^\gamma}). \tag{48}$$

Let us take a closer look at the right hand side of (48). Let

$$f(j) = \exp(-\alpha^{-j})$$

and let

$$g(j,n) = 4(\eta_2 - j)(\frac{1}{n} + \frac{1}{n^{\frac{\mu_{\text{MIN}}}{\mu}}})\alpha^{-j}.$$

Now $f(j)$ is clearly monotonically decreasing with increasing j, while $g(j,n)$ is monotonically increasing in j for fixed n. For small values of j, $f(j) > g(j,n)$; but as j gets large there may exist a j_0 such that $f(j_0) \leq g(j_0, n)$. With this in mind we construct a new function:

$$h(j,n) = \begin{cases} f(j) + g(n, j_0) & \text{if } j < j_0 \\ 2g(n, j_0) & \text{if } j \geq j_0 \end{cases}$$

Since $P(\Delta < -j)$ is non-increasing in j, we would then prefer the bound

$$P(\Delta < -j) \leq h(j,n) + O(\frac{\gamma^j}{n^\gamma}). \tag{49}$$

A good candidate for j_0, is $j_0 = -\log_\alpha \log_e n^{\frac{\mu_{\text{MIN}}}{\mu}}$. Since

$$f(j_0) = \frac{1}{n^{\frac{\mu_{\text{MIN}}}{\mu}}},$$

and

$$g(n, j_0) = O(\frac{(\log n)^2}{n^{\frac{\mu_{\text{MIN}}}{\mu}}}).$$

This choice of j_0 satisfies the condition that $f(j_0) \leq g(n, j_0)$. Therefore, (49) becomes

$$P(\Delta < -j) \leq \exp(-\alpha^{-j}) + O(\frac{(\log n)^2}{n^\delta}).$$

This is (8) and the theorem is proved. •

4.4 Proof of Theorem 3.4

We are finally in a a position to easily prove Theorem 3.4. First consider the special case of the length of the first phrase L_1, This special case is important, and distinct from the other cases, since Z_1 is in the stationary distribution. For this phrase we use the simple inequality,

$$L_1 \leq T + [L_1 - T]1(L_1 \geq T). \tag{50}$$

Taking expectations of both sides over $\mathbf{X}$ and $\mathbf{Z}$, we have that

$$\begin{aligned} EL_1 &\leq ET + E[L_1 - T; L_1 \geq T] \\ &\overset{(a)}{\leq} \frac{\log n}{H} + \sum_{j=0}^{\infty} P(L_1 - T \geq j) + O(1) \\ &\overset{(b)}{\leq} \frac{\log n}{H} + O(1). \end{aligned} \tag{51}$$

where step (a) follows from the "tail-sum" formula for expectations together with Lemma 4.3, and step (b) follows from Theorem 3.1 (7) and the fact that the sum of a geometric series is finite. This is the upper bound on EL_1.

To find the lower bound, begin with the identity,

$$L_1 \geq T - [T - L_1]1(L_1 \leq T).$$

Taking expectations we have

$$\begin{aligned} EL_1 &\geq ET - E[T - L_1; L_1 \leq T] \\ &\overset{(a)}{\geq} \frac{\log n}{H} + O(1) - \sum_{j=0}^{\infty} P(L_1 - T \geq -j) \\ &\overset{(b)}{\geq} \frac{\log n}{H} + O(1) - \sum_{j=0}^{\eta_2} P(L_1 - T \geq -j) \end{aligned} \tag{52}$$

Where step (a) follows from Lemma 4.3 and the tail-sum formula, and step (b) follows from the fact that $L_1 > 0$, so that for all $j \geq \eta_2$, we know that $P(T - L_1 \geq j) = 0$.

To conclude, we recall that from (8), for $j < \eta_2$ that

$$P(L_1 - T \geq -j) \leq \exp(-\alpha^{-j}) + O(\frac{(\log n)^2}{n^\delta}).$$

From (30) we know that $\eta_2 = O(\log n)$, which implies that for sufficiently large n,

$$\begin{aligned} EL_1 &\geq \frac{\log n}{H} + O(1) - \sum_{j=0}^{\eta_2} \exp(-\alpha^{-j}) \\ &\quad - \sum_{j=0}^{\eta_2} O(\frac{(\log n)^2}{n^\delta}) \\ &\geq \frac{\log n}{H} + O(1) + O(\frac{(\log n)^3}{n^\delta}) \\ &= \frac{\log n}{H} + O(1). \end{aligned}$$

This is the lower bound on EL_1.

The final step in the proof is to extend to all phrases L_k. We cannot simply repeat the procedure for L_1 since the first phrase is unique in that its opening state is independent of $\mathbf{X}_1^n$. Consider the k^{th} phrase, for arbitrary k. We hope to prove that the difference between the moments of L_k and the moments of L_1 are $O(1)$ (the same analysis will be used to demonstrate that their normalized distributions are asymptotically identical). Let c_k be the index of the first symbol of the k^{th} phrase. That is $c_k = 1 + \sum_{i=1}^{k-1} L_i$, where the k^{th} phrase is the sequence $\mathbf{Z}_{c_k}^{c_k+L_k-1}$. Define,

$$M = \inf\{j \geq 0 : Z_{c_k+j} = Z_1\},$$

also define,

$$R = \inf\{j \geq 0 : Z_{c_k+M+j} = Z_{c_k}\}.$$

We can define the match length $\hat{L}_k = L(\mathbf{Z}_{c_k+M}^{\infty})$, with the reminder that this is not the length of an actual phrase, although the distributions of L_1 and $\hat{L}_k$ are identical by the Markovity,

$$\mathcal{L}(L_1) = \mathcal{L}(\hat{L}_k). \tag{53}$$

Now define the match length $\tilde{L} = L(\mathbf{Z}_{c_k+M+R}^{\infty})$. Since $Z_{c_k+M+R} = Z_{c_k}$ the distribution of $\tilde{L}$ is the same as L_k. That is

$$\mathcal{L}(L_k) = \mathcal{L}(\tilde{L}). \tag{54}$$

Now consider the set of sequences $\mathbf{Z}$ with $L_k > \hat{L}_k + M$. For any such sequence the longest prefix of $\mathbf{Z}_{c_k}^{\infty}$ that is found in the database is *longer*, by M, than the longest prefix of the sequence $\mathbf{Z}_{c_k+M}^{\infty}$. Since the sequence $\mathbf{Z}_{c_k}^{c_k+L_k}$ is contained in the database, the fact that $L_k > \hat{L} + M$ implies that the sequence $\mathbf{Z}_{c_k+M}^{c_k+M+\hat{L}+1}$ must be contained in the database. This is a contridiction, since $\hat{L}$ is defined to be the longest prefix of $\mathbf{Z}_{c_k+M}^{\infty}$ contained in the database; therefore its one symbol extension cannot be contained in the database. Hence,

$$L_k \leq \hat{L} + M. \tag{55}$$

Now we cannot get a reverse inequality for L_k, but for $\tilde{L}$ we have (again using the fact that match length are maximal) that

$$\hat{L} \leq \tilde{L} + R. \tag{56}$$

From (53) and (54)we know that $EL_1 = E\hat{L}$ and $E\tilde{L} = EL_k$. It is easy to see that EM and ER are bounded, and therefore, by taking expectations in (55) and (56), we have

$$|EL_k - EL_1| = O(1) \tag{57}$$

Applying equation (57) to Lemma 4.3 proves Theorem (3.4). $\bullet$

4.5 Proof of Theorem 3.2

We will prove the theorem using a sequence of lemmas. First we show that the distribution of T_n when properly normalized is $N(0,1)$. Then we will show that this normalized limit is the same as the normalized limit of L_1. Finally, we will demonstrate that the normalized limit of L_1 is the same as the normalized limit of L_k.

Lemma 4.4 *For $n \to \infty$,*

$$\frac{T_n - \frac{\log n}{H}}{\sigma\sqrt{\log n / H^3}} \Rightarrow N(0,1)$$

Proof: Recall that

$$T_n = T_n(\mathbf{Z}) = \max\{m : S_m = \sum_{i=1}^{m} -\log P(Z_i|Z_{i-1}) \leq \log n\}.$$

Observe that this implies that

$$\Pr\{T_n > m\} = \Pr\{S_m \leq \log n\}. \tag{58}$$

The key property is the following lemma which is an application of Theorem 17.3 in Billingsly [2].

Lemma 4.5 (Billingsly) *Since S_n and T_n satisfy the relationship 58, then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to N(0,1) \tag{59}$$

implies

$$\frac{T_n - \frac{\log n}{\mu}}{\sigma(\mu)^{-3/2}\sqrt{\log n}} \to N(0,1)$$

where μ and σ are positive constants.

Lemma 4.5 follows immediately from Lemma 4.6 using any Markov Central Limit Theorem (for example, [3]) to show that $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ is asymptotically Normal(0,1). $\bullet$

Now we state the lemma which provides the necessary connection of T_n with L_1.

Lemma 4.6 *For $n \to \infty$,*

$$\frac{L_1 - \frac{\log n}{H}}{\sigma\sqrt{\log n / H^3}} \Rightarrow N(0,1)$$

Proof: From Theorem 3.1 we have that

$$\frac{|T_n - L_1|}{\sigma\sqrt{\log n / H^3}} = \frac{|\Delta|}{\sigma\sqrt{\log n / H^3}} \to 0$$

as $n \to \infty$. Since the normalized limit of T_n is Normal, the normalized limit of L_1 is Normal. •

Finally we put the pieces together to prove Theorem 3.2. It is easy to see from (55) and the fact that M is $O(1)$ that the distibution of L_k, when normalized is stochastically bounded by the distribution of $\hat{L}$. But the distribution of $\hat{L}$ is identical to that of L_1 from (53), so we have that the normalized distribution of L_k is stochastically upper bounded by a $N(0,1)$ (as $n \to \infty$). From (56) we have that the asymptotic distribution of $\tilde{L}$ is stochastically lower bounded by the distribution of $\hat{L}$. Since the distribution of $\tilde{L}$ is identical to that of L_k from (54), so we have that the asymptotic normalized distribution of L_k is stochastically lower bounded by the normalized distribution of $\hat{L}$ which is a $N(0,1)$. Combining the upper and lower bounds proves the theorem. We also point out, that (9) can be established using the identical analysis (applied to the result cited in [12]. •

4.6 Proof of Theorem 3.3

For this special case, and only this case, $P(\mathbf{z}_1^\ell) = |\mathcal{A}|^{-\ell}$. Hence $T_n(\mathbf{z}_1^\infty) = \frac{\log n}{\log|\mathcal{A}|} = \frac{\log n}{H}$ and for all $\mathbf{z}_1^\infty$, $\ell = T_n(\mathbf{Z}) + j$ is constant. We prove the theorem by refining the error estimate of Lemma 4.2 which earlier was evaluated only crudely. In this special case, we know from (22) that $\lambda = \lambda(\mathbf{z}_1^\ell) = \frac{n-\ell}{|\mathcal{A}|^\ell}$ which is also a constant for all sequences $\mathbf{z}_1^\ell$. Therefore, applying Lemma 4.1 we have that for all sequences $\mathbf{z}_1^\ell$

$$|E_{\mathbf{X}} 1(W(\mathbf{z}_1^\ell) = 0) - \exp(-\lambda)| \le (1 \wedge \lambda) B(\mathbf{z}_1^\ell).$$

Now $B(\mathbf{z}_1^\ell) = E_{\mathbf{X}}\Phi + E_{\mathbf{X}}(\Phi | Y_0 = 1) - 1$. From (23) we have that

$$E_{\mathbf{X}}\Phi = (4\ell + 1) P(\mathbf{z}_1^\ell) = (4\ell + 1)|\mathcal{A}|^{-\ell}.$$

Since this quantity is indepedent of $\mathbf{z}_1^\ell$ we have

$$E_{\mathbf{Z}} E_{\mathbf{X}} \Phi = (4\ell + 1)|\mathcal{A}|^{-\ell}.$$

According to (24) we have that

$$E_{\mathbf{X}}(\Phi | Y_0 = 1) - 1 = \sum_{i=-\ell}^{\ell} \Pr\{\mathbf{X}_{i+1}^{i+\ell} = \mathbf{X}_1^\ell | \mathbf{X}_1^\ell = \mathbf{z}_1^\ell\} - 1$$

Taking expectations with respect to $\mathbf{Z}$:

$$\begin{aligned} E_{\mathbf{Z}} E_{\mathbf{X}}(\Phi | Y_0 = 1) - 1 &= \sum_{i=-\ell}^{\ell} \Pr\{\mathbf{X}_{i+1}^{i+\ell} = \mathbf{X}_1^\ell | \mathbf{X}_1^\ell = \mathbf{Z}_1^\ell\} - 1 \\ &= 2 \sum_{i=1}^{\ell} \Pr\{\mathbf{X}_{i+1}^{i+\ell} = \mathbf{X}_1^\ell\} \\ &= 2 \sum_{i=1}^{\ell} E_{\mathbf{X}_{\ell+1}^{2\ell}} E_{\mathbf{X}_1^\ell} [1\{\mathbf{X}_1^\ell = \mathbf{X}_{i+1}^{i+\ell}\} | \mathbf{X}_{\ell+1}^{2\ell}]. \end{aligned} \tag{60}$$

We claim that for every $1 \le i \le \ell$, and ever outcome of the "future" sequence $x_{1+\ell}^{2\ell}$ there exists a unique sequence $U_i(x_{\ell+1}^{2\ell})$ of length ℓ such that

$$\Pr\{\mathbf{X}_1^\ell = \mathbf{X}_{i+1}^{i+\ell} | \mathbf{X}_{1+\ell}^{2\ell} = x_{1+\ell}^{2\ell}\} = \Pr\{\mathbf{X}_1^\ell = U_i(x_{1+\ell}^{2\ell}) | \mathbf{X}_{1+\ell}^{2\ell} = x_{1+\ell}^{2\ell}\}.$$

In other words, *given* the outcome x of the process from time $1+\ell$ to 2ℓ there exists a single string $U_i(x)$ which will admit a match at position i. We prove by construction: write $U_i(x) = \{u_1, \ldots, u_\ell\}$. Clearly, the first i positions of $U_i(x)$ must be the first i positions of x: $U_{\ell-j+1} = x_{i+1-j}$ for $j = 1, 2, \ldots i$. The rest of the sequence is defined recursively so that the matching condition $X_1^\ell = X_{i+1}^{i+\ell}$ is satisfied, that is we let $u_{\ell-j+1} = u_{\ell-j+i+1}$ for $j = i+1, \ldots, \ell$.

Applying this construction to (60) shows that

$$\begin{aligned} E_{\mathbf{X}}(\Phi(\mathbf{Z}_1^\ell) | Y_0 = 1) - 1 &= 2 \sum_{i=1}^{\ell} E_{\mathbf{X}_{\ell+1}^{2\ell}} E_{\mathbf{X}_1^\ell} 1\{\mathbf{X}_1^\ell = \mathbf{X}_{i+1}^{i+\ell}\} | \mathbf{X}_{\ell+1}^{2\ell}]\} \\ &= 2 \sum_{i=1}^{\ell} E_{\mathbf{X}_1^\ell} E_{\mathbf{X}_{1+\ell}^{2\ell}} 1\{\mathbf{X}_1^\ell = U_i(\mathbf{X}_{i+1}^{i+\ell}) | \mathbf{X}_{\ell+1}^{2\ell}\} \end{aligned} \tag{61}$$

Now since $\mathbf{X}$ is memoryless we can drop the conditioning and simply write

$$E_{\mathbf{X}}(\Phi(\mathbf{Z}_1^\ell) | Y_0 = 1) - 1 = 2 \sum_{i=1}^{\ell} E_{\mathbf{X}} 1\{\mathbf{X}_1^\ell = U_i(\mathbf{X}_{i+1}^{i+\ell})\} = \frac{2\ell}{|\mathcal{A}|^\ell}.$$

This brings us to a refinement of lemma 4.2, with $\ell = T_n(\mathbf{Z}) + j = \frac{\log n}{\log |\mathcal{A}|} + j$ then

$$E_{\mathbf{Z}} B(\mathbf{Z}_1^\ell) \leq (1 \wedge \frac{1}{|\mathcal{A}|^{\log n + j}}) \frac{4(\log n + j)}{|\mathcal{A}|^\ell}.$$

Now, for our choice in ℓ, $\lambda = \frac{n-\ell}{|\mathcal{A}|^\ell} = \frac{1}{|\mathcal{A}|^j} + O(\frac{\log n}{n})$, which implies, together with lemma 4.1, that

$$|\Pr\{L_1 < j + \log n\} - \exp(-|\mathcal{A}|^{-j} - O(\frac{\log n}{n}))| \leq (1 \wedge \frac{1}{|\mathcal{A}|^{\log n + j}}) \frac{4(\log n + j)}{|\mathcal{A}|^\ell}.$$

Which reduces to

$$\Pr\{L_1 < j + \frac{\log n}{H}\} = \exp(-|\mathcal{A}|^{-j}) + O(\frac{\log n + j}{n}).$$

Appendix A: Comments on the Memory of the Source

Let $\{Z_n\}$ be the source sequence with memory M – so that (1) is satisfied. For $-\infty < n < \infty$, define $S_n = \mathbf{Z}_{n-M+1}^n$. The transformation $\{Z_n\} \to \{S_n\}$ is $1-1$ and the entropy of the sequence $\{S_n\}$ is also H. Further, $\{S_n\}$ is the state sequence of a Markov chain (with state space $\mathcal{A}^M$). It follows from the stationarity and ergodicity of $\{Z_n\}$ that the Markov chain is irreducible. We also assumed that the Markov chain is also aperiodic.

Now if the source memory $M > 1$, we can replace the source $\{Z_n\}$ by the Markov chain $\{S_n\}$ with memory 1 and an enlarged alphabet. Thus with no loss of generality we will take the source memory $M = 1$, and finite alphabet $\mathcal{A}$ with $|\mathcal{A}| = A$.

Now write $\Pr\{Z_1 = a \mid Z_0 = b\} = P(a|b)$, $a, b \in \mathcal{A}$. Then, for $\mathbf{z}_1^\ell \in \mathcal{A}^\ell$,

$$\Pr\{\mathbf{Z}_1^\ell = \mathbf{z}_1^\ell \mid Z_0 = z_0\} = \prod_{i=1}^{\ell} P(z_c \mid z_{i-1}) \ . \tag{A.1}$$

Now if this probability equals 1 for $\ell \geq A$, the Markov chain transition diagram defining $\{Z_n\}$ must have a loop of probability 1 transitions. Then either $\{Z_n\}$ is degenerate (with $H = 0$) or it is not irreducible. Hence for $\ell \geq A$, the probability in (A.1) < 1. Define

$$\alpha_1 = \max_{\mathbf{z}_0^A} \ \Pr\{\mathbf{Z}_1^A = \mathbf{z}_1^A \mid Z_0 = z_0\} < 1 \ . \tag{A.2}$$

Then for $\ell \geq 1$, write $\ell = kA + j$, $0 \leq j < A$. Then

$$\begin{aligned} \Pr\{\mathbf{Z}_1^\ell = \mathbf{z}_1^\ell \mid \mathbf{Z}_0 = \mathbf{z}_0\} \ &\leq \ \alpha_1^k = \alpha_1^{(\ell-j)/A} \leq \alpha_1^{\ell/A-1} \\ &= \ \left(\frac{1}{\alpha_1}\right)(\alpha_1^{1/A})^\ell = \beta_0\alpha^\ell \ , \end{aligned} \tag{A.3}$$

where $\beta_0 = 1/\alpha$, and $\alpha = \alpha_1^{1/A} < 1$. This allows us to replace the right member of (14) by $\beta_0\alpha^{k-j+1}$, and prove the theorems even if (12) does not hold.

Appendix B: Stein's Method and proof of lemma 4.1

Theorem 4.7 (Stein) *Let Y_1^n be a sequence of random variables taking only the values 0 or 1. Also let*

$$W = \sum_{1=1}^{n} Y_i$$

and

$$\lambda = E[W]$$

Let the random variable U be uniformly distributed over the integers from 1 to n, independent of Y_1^n. Now, let $Y_1^*, \ldots, Y_n^*$ be any sequence of random variables having the distribution of Y_1^n conditioned on $Y_U = 1$. Then let

$$W^* = \sum_{i=1}^{n} Y_i^*$$

Then for all subsets A of the positive integers,

$$|Pr\{W \in A\} - e^{-\lambda} \sum_{w \in A} \frac{\lambda^w}{w!}| \leq (1 \wedge \lambda) E|W + 1 - W^*| \tag{A.4}$$

For a proof of Theorem 4.7 See [10].

One might ask what Y^* might look like. One simple way to find Y^*, (a random variable with the right conditional distribution) is to pick U and pick Y. If $Y_U = 1$ then let $Y^* = Y$. If not, repeat the process by picking a new U and a new Y. The trick to the application of the theorem is to choose a Y^* in a way that W and W^* will be as closely 'coupled' as possible. A proper coupling can force the $|W + 1 - W^*|$ term to be small, and thereby provide a good bound.

To help find Y^* we make use of the following lemma:

Lemma 4.8 (Stein) *Let U be distributed $Uniform[1, 2, \ldots, n]$. Let U' be a random variable on $[1, n]$ with*

$$Pr\{U' = i\} = Pr\{U = i | Y_i = 1\}$$

and let $Y_1', Y_2', \ldots, Y_n'$ be a sequence of random variables, conditionally distributed given U' according to

$$Pr\{Y' = y | U' = i\} = Pr\{Y = y | U = i \;\&\; Y_i = 1\}$$

for all $i \in [1,n]$ and $y \in \{0,1\}^n$. Then Y' has the distribution specified for Y^* in Theorem(4.7) and consequently W', defined by

$$W' = \sum_{i=1}^{n} Y_i'$$

has the distribution specified for W^*.

To prove Lemma 4.1 let

$$Y_i = 1\{\mathbf{X}_{i+1}^{i+\ell} = \mathbf{z}_1^\ell\}$$

for $i = 0, 1, 2 \ldots, n-\ell$. Observe that

$$W(\mathbf{z}_1^\ell) = W = \sum_{i=0}^{n-\ell} Y_i.$$

We now define W^* by introducing some auxilliary randomness in the form of a doubly infinite sequence $\hat{\mathbf{X}}_{-\infty}^{\infty}$, independent of $\mathbf{X}_{-\infty}^{\infty}$, that has as its unconditional distribution the conditional distribution of $\mathbf{X}_{-\infty}^{\infty}$ given that $\mathbf{X}_1^\ell = \mathbf{z}_1^\ell$. That is

$$\Pr\{\hat{\mathbf{X}} = \mathbf{x}_{-\infty}^{\infty}\} = \Pr\{\mathbf{X} = \mathbf{x}_{-\infty}^{\infty} | \mathbf{X}_1^\ell = \mathbf{z}_1^\ell\}.$$

We then construct the indicator variable out of $\hat{\mathbf{X}}$,

$$\hat{\mathbf{Y}}_i = 1\{\hat{\mathbf{X}}_{i+1}^{i+\ell} = \mathbf{z}_1^\ell\}.$$

We set up a coupling [6] by introducing, for U uniformly distributed on $[1, n-\ell]$, the integer random variables

$$T^+ = \inf\{n - U \geq k > \ell : X_{U+k} = \hat{X}_k\} \tag{A.5}$$

and

$$T^- = \inf\{U - \ell \geq k > \ell : X_{U-k} = \hat{X}_{-k}\}. \tag{A.6}$$

After the coupling times T^+ and T^- the two sequences are probabilistically equivalent. Thus, for sources with no memory, we may let $T^+ = T^- \triangleq \ell$. We form a new sequence by splicing the sequence $\hat{\mathbf{Y}}$ at times T^+ and T^- and patching it into $\mathbf{Y}$ at the random time U. Accordingly define,

$$Y' = \begin{cases} Y_i & \text{if } i \notin [U - T^-, U + T^+] \\ \hat{Y}_i & \text{if } i \in [U - T^-, U + T^+] \end{cases}$$

[6] A coupling time is defined to be the first time i, when two independent Markov chains are in the same state. That is, $X_i = \hat{X}_i$, but $X_j \neq \hat{X}_j$ for all $j < i$. After two independent, identically distributed chains couple they are henceforth probabilistically equivalent.

It is easily shown that this sequence, by Lemma 4.8 and the stationarity of $\mathbf{X}$, has the required distribution of Y^*. We now have the proper candidate for W^*, namely

$$W^* = \sum_{i=1}^{n-\ell} Y_i^{'}. \tag{A.7}$$

The mathematical task will be to evaluate $E|W+1-W^*|$. By construction, most of the terms cancel, so that

$$W+1-W^* = 1 + \sum_{i=U-T^-}^{U+T^+} (Y_i - \hat{Y}_i^{'}).$$

Before taking expectations, truncate the sum at 2ℓ (so that if ℓ is sufficiently large, then with very high probability the coupling times T^+ and T^- will both be less than 2ℓ). Therefore,

$$W+1-W^* \leq \sum_{i=U-2\ell}^{U+2\ell} (Y_i - \hat{Y}_i) + 1 + [T^+ - 2\ell]1\{T^+ > 2\ell\} + [T^- - 2\ell]1\{T^- > 2\ell\} \tag{A.8}$$

Hence,

$$E_{\mathbf{X},\hat{\mathbf{X}}}|W+1-W^*| \leq \sum_{i=U-2\ell}^{U+2\ell} E_{\mathbf{X}} Y_i + \sum_{i=U-2\ell}^{U+2\ell} E_{\hat{\mathbf{X}}} \hat{Y}_i - 1 + E_{\mathbf{X},\hat{\mathbf{X}}}[T^+; T^+ > 2\ell] + E[T^-; T^- > 2\ell]$$

Now let

$$\begin{aligned} \Phi &\triangleq \sum_{i=-2\ell}^{2\ell} Y_i \\ C(\mathbf{z}_1^\ell) &\triangleq E_{\mathbf{X},\hat{\mathbf{X}}}[T^+; T^+ > 2\ell] + E_{\mathbf{X},\hat{\mathbf{X}}}[T^-; T^- > 2\ell]. \end{aligned}$$

Therefore the stationarity of $\mathbf{X}$ and the definition of $\hat{Y}_i$ imply that

$$E_{\mathbf{X},\hat{\mathbf{X}}}|W+1-W^*| \leq E_{\mathbf{X}}\Phi + E_{\mathbf{X}}(\Phi | Y_0 = 1) - 1 + C(\mathbf{z}_1^\ell)$$

Since it is well known (see [5]) that for any two independent identically distributed, aperiodic, irreducible Markov chains $\mathbf{X}$ and $\hat{\mathbf{X}}$, there exists an integer k and a constant $0 < p < 1$ such that $\inf_{a,b \in \mathcal{A}} \Pr\{X_k = \hat{X}_k | X_0 = a, \hat{X}_0 = b\} > p$. It follows that there exists a constant $\gamma > 0$ such that for all $\mathbf{z}_1^\ell$

$$C(\mathbf{z}_1^\ell) = O(\gamma^\ell). \tag{A.9}$$

We also note that for memoryless sources the coupling times are defined to be equal to ℓ and thus $C(\mathbf{z}_1^\ell) = 0$. Finally, we are able to establish Lemma 4.1 by evaluating (A.4) with $A = \{0\}$, and $\lambda(\mathbf{z}_1^\ell) = EW(\mathbf{z}_1^\ell)$ •

Lemma 4.9 *For $\eta_1 \geq i > 0$*

$$\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j} = \mathbf{X}_1^{\eta_1+j}\} \leq \beta^{\eta_1+j},$$

and for $\eta_1 \leq i < 0$

$$\Pr\{\mathbf{X}_{i+1}^{i+\eta_1+j} = \mathbf{X}_1^{\eta_1+j}\} \leq \alpha^{\eta_1+j}.$$

Proof: We will do only the case where $i > 0$. The case where $i < 0$ is the same. Observe that

$$\Pr\{\mathbf{X}_{i+1}^{i+\eta_1} = \mathbf{X}_1^{\eta_1}\} = E[1\{\mathbf{X}_1^{\eta_1} = \mathbf{X}_{i+1}^{i+\eta_1}\}|\mathbf{X}_{\eta_1+1}^{2\eta_1}].$$

Now we will show that for every $1 \leq i \leq \eta_1$, and every outcome of the "future" sequence $x_{1+\eta_1}^{2\eta_1}$ there exists a unique sequence $U_i(x_{\eta_1}^{2\eta_1})$ of length η_1 such that

$$\Pr\{\mathbf{X}_1^{\eta_1} = \mathbf{X}_{i+1}^{i+\eta_1}|\mathbf{X}_{1+\eta_1}^{2\eta_1} = x_{1+\eta_1}^{2\eta_1}\} = \Pr\{\mathbf{X}_1^{\eta_1} = U_i(x_{1+\eta_1}^{2\eta_1})|\mathbf{X}_{1+\eta_1}^{2\eta_1} = x_{1+\eta_1}^{2\eta_1}\}.$$

In other words, *given* the outcome x of the process from time $1+\eta_1$ to $2\eta_1$ there exists a single string $U_i(x)$ which will admit a match at position i. We prove by construction: write $U_i(x) = \{u_1, \ldots, u_{\eta_1}\}$. Clearly, the first i positions of $U_i(x)$ must be the first i positions of x: $U_{\eta_1-j+1} = x_{i+1-j}$ for $j = 1, 2, \ldots i$. The rest of the sequence is defined recursively so that the matching condition $X_1^{\eta_1} = X_{i+1}^{i+\eta_1}$ is satisfied, that is we let $u_{\eta_1-j+1} = u_{\eta_1-j+i+1}$ for $j = i+1, \ldots, \eta_1$. Therefore,

$$\begin{aligned}
\Pr\{\mathbf{X}_{i+1}^{i+\eta_1} = \mathbf{X}_1^{\eta_1}\} &= E[1\{\mathbf{X}_1^{\eta_1} = \mathbf{X}_{i+1}^{i+\eta_1}\}|\mathbf{X}_{\eta_1+1}^{2\eta_1}] \\
&= E[1\{\mathbf{X}_1^{\eta_1} = U_i(\mathbf{X}_{i+1}^{i+\eta_1})\}|\mathbf{X}_{\eta_1+1}^{2\eta_1}] \\
&\leq \beta^{\eta_1}
\end{aligned}$$

Lemma 4.10 *Let T_a be the length of the path from state a to state a: $T_a = \inf\{k > 0 : X_k = a|X_1 = a\}$. The entropy of this "trajectory" from state a to a, with length T_a, is defined to be $H_a = -E\log P(X_1^{T_a}|X_1 = a)$, and is equal to $ET_a H$.*

Proof: Let a be fixed with stationary probability $p_a = \Pr\{X_1 = a\}$. Let For each return of the process to state a we can define a random variable, $T_a(j)$ equal to the recurrence time of the j^{th} return to state a. It is clear that $T_a(j)$ are *i.i.d.* with $ET_a(j) = \frac{1}{p_a}$ (by Kac's theorem). Observe that the entropy of P can be written as an almost sure limit:

$$H = \lim_{n\to\infty} \frac{-\log P(X_1^n|X_1 = a)}{n}$$

which implies that

$$H = \lim_{n\to\infty} \sum_{j=1}^{n} \frac{-\log P(X_{C_{r-1}}^{C_j}|X_1 = a)}{\sum_{j=1}^{n} T_a(j)}$$

(where $C_0 = 1$ and $C_i = \sum_{j=1}^{i} T_a(j)$). If we divide the numerator and denominator by k and apply the strong law of large numbers we have that $H = \frac{H_a}{ET_a}$, which proves the theorem. QED.

Corollary 4.1 *The variance, σ_a^2, in the entropy of the Markov trajectory from state a to a is equal to $\frac{\sigma^2}{p_a}$.*

Proof:

$$\begin{aligned}
\sigma^2 &= \lim_{n\to\infty} E\frac{\log^2 P(X_1^n|X_1 = a) - n^2H^2}{n} \\
&= \lim_{n\to\infty} \frac{\sum_{j=1}^{n}[\log^2 P(X_{C_{j-1}}^{C_j}|X_1 = a) - T_a^2 H_a^2]}{\sum_{j=1}^{n} T_a(j)} \\
&= \frac{Var(\log P(X_1^{T_a}|X_1 = a))}{ET_a}
\end{aligned}$$

Therefore $\sigma_i^2 = \frac{\sigma^2}{p_a}$, as desired. QED.

References

[1] Barbour, Holst, and Janson, *Poisson Approximation*, Oxford Press, 1992.

[2] P. Billingsley, *Convergence of Probability Measures*, Wiley and Sons, New York, 1968.

[3] Kai Lai Chung, *Markov Chains With Stationary Transition Probabilities*, Springer-Verlag, New York, 1967.

[4] L. Ekroot and Thomas M. Cover, "The Entropy of a randomly stopped sequence", *IEEE Transactions on Information Theory*, Vol. 37, pp.1641-1644, 1991.

[5] Hoel, Port and Stone. *An Introduction to Stochastic Process* Houghton Mifflin, Boston, 1972.

[6] P. Jacquet and W. Szpankowski, " Autocorrelation on Words and Its Applications: Analysis of Suffix trees by String-Ruler Approach" *Journal of Combinatorial Theory* Vol. 66, NO. 2, May 1994.

[7] G. Louchard and W. Szpankowski, " Average Profile and Limiting Distribution of a Phrase Size in the Lempel-Ziv Parsing Algorithm",

[8] E. Plotnick, M.J. Weinberger, and Jacob Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the Redundancy of the Lempel-Ziv Algorithm", IEEE Trans. Inform. Theory, Vol 38, No. 1 pp 66-72, January 1992

[9] W. Szpankowski, "A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors" SIAM J. COmpt., 22. 1176-1198, 1993. *IEEE Transactions on Information Theory*, Vol. 41, March 1995, pp 478-488.

[10] C. Stein. *Approximate Computation of Expectations. Lecture Notes 7*, Institute of Mathematical Statistics, Hayward, CA, 1986.

[11] C. Stein. Lecture notes: Stanford University, Dept. of Statistics, 1992.

[12] Wyner, A. D. and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic source with applications to data compression," *IEEE Trans. on Information Theory*, **35**, pp. 1250-1258, Nov. 1989.

[13] Wyner, A. J., *String Matching Theorems and Applications to Data Compression and Statistics*, Ph.D. Thesis, Stanford University, 1993. Also "The redundancy and distribution of phrase lengths of the FDLZ algorithm", paper submitted to 1994 I.S.I.T. (Trondheim).

[14] Wyner, A. D. and Wyner, A.J. "Improved Redundancy of a Version of the Lempel-Ziv algorithm" *IEEE Trans. on Information Theory*, Volume 41, pp. 723-731, May, 1995.