References

[1] Bender, P. E. and J. Wolf, "New asymptotic bounds and improvements on the Lempel-Ziv data compression algorithm", *IEEE Trans. on I.T.*, **37**, pp. 721–729, 1991.

[2] Feller, *An Introduction to Probability Theory and Its Applications*, 2^{ND} ed., New York: John Wiley, 1957.

[3] 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.

[4] 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).

where $P(a) = \Pr\{Z_0 = a\}$, and $P(b|a) = \Pr\{Z_1 = b \mid Z_0 = a\}$, for $a, b \in \mathcal{A}$. For $a \in \mathcal{A}$, let "E_a" denote expectations conditioned on $Z_1 = a$. Thus, in particular,

$$E\phi_1(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) = \sum_{a \in \mathcal{A}} P(a)\ E_a\phi_1(\mathbf{Z}, \mathbf{X}, \mathbf{Y})\ . \tag{B.8}$$

Now return to $E\phi_k$. Let

$$\Lambda_k = \sum_{i=1}^{k-1} L_i + 1\ . \tag{B.9}$$

Then

$$\phi_k(\mathbf{Z}_1^\infty, \mathbf{X}, \mathbf{Y}) = \phi_1(\mathbf{Z}_{\Lambda_k}^\infty, \mathbf{X}, \mathbf{Y})\ . \tag{B.10}$$

We note that we must take our expectations with care, since $\mathbf{Z}_{\Lambda_k}^\infty$ is *not* independent of $\mathbf{X}_1^n$, $\mathbf{Y}_1^n$, even though $(\mathbf{Z}_1^\infty, \mathbf{X}_1^n, \mathbf{Y}_1^n)$ is an independent triple.

Now define the event

$$\mathcal{E}_{\mathbf{xy}} = \{\mathbf{X}_1^n = \mathbf{x} \ \text{ and } \ \mathbf{Y}_1^n = \mathbf{y}\}\,, \quad \mathbf{x}, \mathbf{y} \in \mathcal{A}^n\ . \tag{B.11}$$

We write

$$\begin{aligned}
E\phi_k(\mathbf{Z}_1^\infty, \mathbf{X}, \mathbf{Y}) &= \sum_{\mathbf{x},\mathbf{y}} \Pr(\mathcal{E}_{\mathbf{xy}}) E[\phi_k \mid \mathcal{E}_{\mathbf{xy}}] \\
&\overset{\text{(a)}}{=} \sum_{\mathbf{x},\mathbf{y}} P(\mathcal{E}_{\mathbf{xy}}) E[\phi_1(\mathbf{Z}_{\Lambda_k}^\infty, \mathbf{x}, \mathbf{y}) \mid \mathcal{E}_{\mathbf{x},\mathbf{y}}] \\
&\overset{\text{(b)}}{=} \sum_{\mathbf{x},\mathbf{y}} P(\mathcal{E}_{\mathbf{xy}}) \sum_{a \in \mathcal{A}} \Pr\{Z_{\Lambda_k} = a \mid \mathcal{E}_{\mathbf{xy}}\} E[\phi_1(\mathbf{Z}_{\Lambda_k}^\infty, \mathbf{x}, \mathbf{y}) \mid Z_{\Lambda_k} = a] \\
&\overset{\text{(c)}}{\leq} \sum_{a \in \mathcal{A}} E_a\phi_1(\mathbf{Z}_1^\infty, \mathbf{X}, \mathbf{Y})\ .
\end{aligned} \tag{B.12}$$

Step (a) in (B.12) follows from (B.10); step (b) from the fact that $(\mathbf{X}, \mathbf{Y}), Z_{\Lambda_k}, \mathbf{Z}_{\Lambda_k}^\infty$ is a Markov chain (i.e. $\mathbf{Z}_{\Lambda_k}^\infty$ depends on $(\mathbf{X}, \mathbf{Y})$ and Z_{Λ_k} only through Z_{Λ_k}) so that

$$E[\phi_1(\mathbf{Z}_{\Lambda_k}^\infty, \mathbf{x}, \mathbf{y}) \mid Z_{\Lambda_k} = a,\ \mathcal{E}_{\mathbf{xy}}] = E[\phi_1(\mathbf{Z}_{\Lambda_k}^\infty, \mathbf{x}, \mathbf{y}) \mid Z_{\Lambda_k} = a]\ ;$$

and step (c) follows on bounding the conditional probability by 1, interchanging the order summation, and using the definition of E_a.

Now let $\sigma \triangleq \min\{P(a) \in \mathcal{A} : P(a) > 0\}$, and write (B.8) as

$$E\phi_1(Z_1^\infty, \mathbf{X}, \mathbf{Y}) = \sum_{a \in \mathcal{A}} P(a) E_a\phi_1(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) \geq \sigma \sum_{a \in \mathcal{A}} E_a\phi_1\ . \tag{B.13}$$

Combining (B.12) and (B.13) we have (B.6) with $c = 1/\sigma$ and the proof is complete.

Appendix B: Average Rate

Suppose that we use the scheme described in Section II to encode K repetitions of the process with $(L_k, \hat{L}_k)$, $1 \le k \le K$. This will require on average about $K \log n + E \sum_{k=1}^{K} \log\ |L_k - \hat{L}_k|$ bits to encode on average $E \sum_{k=1}^{K} L_k$ bits. Thus the average rate[†] is

$$R_n = \lim_{K \to \infty} \frac{K \log n + \sum\limits_{k=1}^{K} E \log |L_k - \hat{L}_k|}{\sum\limits_{k=1}^{K} E L_k} . \tag{B.1}$$

Now it is shown in [4], that as $n \to \infty$,

$$E L_k = E L_1 + O(1) , \tag{B.2}$$

where $O(1)$ does not depend on n or k. We show below that there is a constant c (that does not depend on n or k) such that

$$E \log |L_k - \hat{L}_k| \le c\, E \log |L_1 - \hat{L}_1| , \tag{B.3}$$

so that (B.1) becomes

$$R_n \le \frac{\log n + c\, E \log |L_1 - \hat{L}_1|}{E L_1 + O(1)} . \tag{B.4}$$

Inequality (B.4) is (2.7). We now give a proof of (B.3).

Let us define, for $k = 1, 2, 3, \ldots$,

$$\phi_k(\mathbf{Z}_1^{\infty}, \mathbf{X}_1^n, \mathbf{Y}_1^n) = \log |L_k - \hat{L}_k| , \tag{B.5}$$

where the dependence of ϕ_k on $\mathbf{Z}$, $\mathbf{X}$, $\mathbf{Y}$ is made explicit. (Actually ϕ_k depends on $\mathbf{Z}_1^{\infty}$ only through $\mathbf{Z}_1^{kn}$.) Inequality (B.3) is

$$E \phi_k(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) \le c\, E \phi_1(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) . \tag{B.6}$$

These expectations are taken with $\mathbf{Z}$ having distribution specified by

$$\Pr \left\{ \mathbf{Z}_1^j = \mathbf{z}_1^j \right\} = P(z_1) \prod_{i=2}^{j} P(z_i \mid z_{i-1}) , \tag{B.7}$$

[†] There are other reasonable definitions of the average rate. The definition in (B.1) turns out to be convenient.

where $|\epsilon_j| \leq c|\lambda_2|^j \to 0$, as $j \to \infty$.

Our task is to justify ineq. (4.14). Let

$$\mathcal{E}_3 \triangleq \bigcap_{j=\ell+t}^{n} \{\mathbf{X}_{j-\ell+1}^{j} = \mathbf{z}_1^{\ell}\} ,$$

and as in Section IV, let

$$\mathcal{E}_2 = \{\mathbf{X}_{-\ell}^{0} = \mathbf{z}_1^{\ell}\} .$$

Then

$$\begin{aligned}
\Pr(\mathcal{E}_3 \mid \mathcal{E}_2) &= \Pr(\mathcal{E}_3 \mid X_0 = z_0) \\
&= \sum_{a\in\mathcal{A}} \Pr\{X_t = a \mid X_0 = z_0\} \Pr(\mathcal{E}_3 \mid X_t = a,\ X_0 = z_0) \\
&\overset{\text{(a)}}{=} \sum_{a} \Pr\{X_t = a \mid X_0 = z_0\} \Pr(\mathcal{E}_3 \mid X_t = a) \\
&\overset{\text{(b)}}{\geq} (1 - c|\lambda_2|^t) \sum_{a} \Pr\{X_t = a\} \Pr(\mathcal{E}_3 \mid X_t = a) \\
&= (1 - c|\lambda_2|^t) \Pr(\mathcal{E}_3) . \qquad\qquad \text{(A.5)}
\end{aligned}$$

Step (a) follows from the fact that $\mathcal{E}_3$ is $\mathbf{X}_{t+1}^{\infty}$-measurable, and the Markov property. Step (b) follows from (A.4). Letting t be large enough so that $c|\lambda_2|^t < 1/2$, we have

$$\Pr(\mathcal{E}_3 \mid \mathcal{E}_2) \geq \frac{1}{2} P(\mathcal{E}_3) , \qquad \text{(A.6)}$$

which justifies (4.14).

Appendix A: Comments on the Memory of the Source

Let $\{Z_n\}$ be the source sequence with memory M – so that (1.1) 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 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 $\mathbf{z}_1^\ell \in \mathcal{A}^\ell$,

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

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

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

where $\beta = 1/\alpha$, and $\alpha = \alpha_1^{1/A} < 1$. This allows us to replace the right member of (3.7) by $\beta\alpha^{k-j+1}$, and prove Theorem 2.1 even if (3.6) does not hold.

Our final task is to establish quantitatively the vanishing memory of the source. Let $\mathbb{P}$ be the transition probability matrix defining the Markov chain $\{Z_n\}$. Let λ_i, $1 \leq i \leq A$, be the eigenvalues of $\mathbb{P}$ ordered so that $|\lambda_1| \geq |\lambda_2| \geq \ldots$. Then $\lambda_1 = 1$, and $|\lambda_2| < 1$ (from the aperiodicity of the chain). It is a property of Markov chains (see, e.g. [2, Chapter XVI]), that

$$\Pr\{Z_j = a \mid Z_0 = b\} = (1 + \epsilon_j)\Pr\{Z_j = a\} \ , \tag{A.4}$$

Inequalities (4.6) and (4.10) imply

$$\Pr\left\{L(\mathbf{z}_1^\infty, \mathbf{X}_1^\infty) \le \ell \mid \mathbf{X}_{-\ell+1}^0 = \mathbf{z}_1^\ell\right\} \le \alpha^{T-\ell-1} + \frac{\alpha^{\ell^*/2}}{1-\alpha} \ . \tag{4.11}$$

Now if the source is memoryless, the events

$$\mathcal{E}_1 = \{L(\mathbf{z}_1^\infty, \mathbf{X}_1^\infty) \le \ell\} \quad \text{and} \quad \mathcal{E}_2 = \{\mathbf{X}_{-\ell+1}^0 = \mathbf{z}_1^\ell\}$$

are independent, so that the left member of (4.11) equals $\Pr(\mathcal{E}_1)$, and (4.11) is stronger than Lemma 3.2B (with $t = 0$).

It remains to prove Lemma 3.2B in the Markov case. Inequality (4.11) is an upper bound on $P(\mathcal{E}_1 \mid \mathcal{E}_2)$. Write, for $t > 0$,

$$\mathcal{E}_1 = \bigcap_{j=\ell}^{n} \left\{\mathbf{X}_{j-\ell+1}^j \ne \mathbf{z}_1^\ell\right\} = \bigcap_{j=\ell+t}^{n} \left\{\mathbf{X}_{j-\ell+1}^j \ne z_1^\ell\right\} - \bigcup_{j=\ell}^{\ell+t-1} \left\{\mathbf{X}_{j-\ell+1}^j = z_1^\ell\right\} \ , \tag{4.12}$$

so that

$$\Pr(\mathcal{E}_1 \mid \mathcal{E}_2) \ge \Pr\left[\bigcap_{j=\ell+t-1}^{n} \left\{\mathbf{X}_{j-\ell+1}^j \ne \mathbf{z}_1^\ell\right\} \mid \mathcal{E}_2\right] - t\alpha^\ell \ . \tag{4.13}$$

Now from (A.6), if t is sufficiently large,

$$\begin{aligned} &\Pr\left[\bigcap_{j=\ell+t-1}^{n} \left\{\mathbf{X}_{j-\ell+1}^j \ne \mathbf{z}_1^\ell\right\} \mid \mathcal{E}_2\right] \\ &\ge \frac{1}{2} \Pr\left[\bigcap_{j=\ell+t-1}^{n} \left\{\mathbf{X}_{j-\ell+1}^j \ne \mathbf{z}_1^\ell\right\}\right] \ge \frac{1}{2} \Pr\left\{L(\mathbf{z}_1^\infty, \mathbf{X}_1^n) \le \ell\right\} \ . \end{aligned} \tag{4.14}$$

Inequalities (4.11), (4.13), and (4.14) yield

$$\Pr\left\{L(\mathbf{z}_1^\infty, \mathbf{X}_1^\infty) \le \ell\right\} \le 2\left[\alpha^{T-\ell+1} + \frac{\alpha^{\ell^*/2}}{(1-\alpha)} + t\alpha^{\ell^*}\right] \ , \tag{4.15}$$

which is Lemma 3.2B.

Acknowledgement

We acknowledge with thanks useful conversations with J. Glasgow, A. Orlitsky, J. Wolf, and J. Ziv.

$$\begin{aligned}
&\overset{(b)}{=} \bigcap_{j=\frac{\ell}{2}}^{n} \left\{\mathbf{X}_{j-\ell+1}^{j} \neq \mathbf{z}_1^{\ell}\right\} \overset{(c)}{=} \bigcap_{j=\ell}^{n} \left\{\mathbf{X}_{j-\ell+1}^{j} \neq \mathbf{z}_1^{\ell}\right\} - \bigcup_{j=\ell^*/2}^{\ell-1} \left\{\mathbf{X}_{j-\ell+1}^{j} = \mathbf{z}_1^{\ell}\right\} \\
&\overset{(d)}{=} \left\{L(z_1^{\ell}, \mathbf{X}_1^{n}) \leq \ell\right\} - \bigcup_{j=\ell^*/2}^{\ell-1} \left\{\mathbf{X}_{j-\ell+1}^{j} = \mathbf{z}_1^{\ell}\right\} \ . \qquad (4.5)
\end{aligned}$$

Step (a) in (4.5) follows from the definition of $N(\cdot,\cdot)$; step (b) from $\mathbf{z} \in \mathcal{S}_\ell$; step (c) from $\mathcal{E}_1 \cap \mathcal{E}_2 = \mathcal{E}_1 - \mathcal{E}_2^c$, for any two events $\mathcal{E}_1$ and $\mathcal{E}_2$; and step (d) from (4.3). It follows, using $\Pr(\mathcal{E}_1 - \mathcal{E}_2^c) \geq \Pr(\mathcal{E}_1) - \Pr(\mathcal{E}_2^c)$ and the union bound, that for any $\mathbf{z}_1^\infty$ such that $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}$,

$$\begin{aligned}
&\Pr\left\{N(\mathbf{z}_1^{\ell}, \mathbf{X}_1^{\infty}) > n \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\} \\
&\quad\geq \Pr\left\{L(\mathbf{z}_1^{\infty}, \mathbf{X}_1^{\infty}) \leq \ell \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\} \\
&\qquad - \sum_{j=\ell^*/2}^{\ell-1} \Pr\left\{\mathbf{X}_{j-\ell+1}^{j} = \mathbf{z}_1^{\ell} \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\} \\
&\quad\geq \Pr\left\{L(\mathbf{z}_1^{\infty}, \mathbf{X}_1^{\infty}) \leq \ell \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\} - \frac{\alpha^{\ell^*/2}}{1-\alpha} \ . \qquad (4.6)
\end{aligned}$$

The final step in (4.6) follows from the fact that for $\frac{\ell^*}{2} \leq j \leq \ell - 1$, the vector $\mathbf{X}_{j-\ell+1}^{j} = (X_{j-\ell+1}, \ldots, X_j)$ has j components X_k, $k \geq 1$. Thus

$$\Pr\left\{\mathbf{X}_{j-\ell+1}^{j} = \mathbf{z}_1^{\ell} \mid \mathbf{X}_{j-\ell+1}^{0} = z_1^{\ell}\right\} \leq \alpha^j \ . \qquad (4.7)$$

Now from Kac' Theorem for average recurrence time (see for example [3], Appendix),

$$E\left\{N(\mathbf{z}_1^{\ell}, \mathbf{X}_1^{\infty}) \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\} = \frac{1}{\Pr\{\mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\}} \ .$$

Thus from the Markov inequality,

$$\Pr\left\{N(\mathbf{z}_1^{\ell}, \mathbf{X}_1^{\infty}) > n \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\} \leq \frac{1}{n \Pr\left\{\mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\right\}} \ . \qquad (4.8)$$

Now let $\mathbf{z}_1^\infty$ be such that $\mathbf{z}^{\ell^*} \in \mathcal{S}_{\ell^*}$, and let ℓ satisfy $\ell^* \leq \ell \leq T_n(\mathbf{z}_1^\infty)$. We have, from the definition of T_n,

$$\begin{aligned}
\frac{1}{n} &\leq \Pr\left\{\mathbf{X}_1^{T-1} = \mathbf{z}_1^{T-1}\right\} = \Pr\left\{\mathbf{X}_1^{\ell} = \mathbf{z}_1^{\ell}\right\} \Pr\left\{\mathbf{X}_{\ell+1}^{T-1} = \mathbf{z}_{\ell+1}^{T-1} \mid \mathbf{X}_1^{\ell} = \mathbf{z}_1^{\ell}\right\} \\
&\leq \Pr\left\{\mathbf{X}_1^{\ell} = \mathbf{z}_1^{\ell}\right\} \alpha^{T-\ell-1} \ . \qquad (4.9)
\end{aligned}$$

Thus (4.8) becomes

$$\Pr\{N(\mathbf{z}_1^{\ell}, \mathbf{X}_1^{\infty}) > n \mid \mathbf{X}_{-\ell+1}^{0} = \mathbf{z}_1^{\ell}\} \leq \alpha^{T-\ell-1} \ . \qquad (4.10)$$

$$\begin{aligned}
&\overset{(b)}{\leq} \quad n \ \Pr\left\{\mathbf{X}_1^\ell = \mathbf{z}_1^\ell\right\} \\
&= \quad n \ \Pr\left\{\mathbf{X}_1^{T_n(\mathbf{z})} = \mathbf{z}_1^T\right\} \Pr\left\{\mathbf{X}_{T+1}^\ell = \mathbf{z}_{T+1}^\ell \mid \mathbf{X}_1^T = \mathbf{z}_1^T\right\} \\
&\overset{(c)}{\leq} \quad 1 \cdot \alpha^{\ell - T(\mathbf{z})} \ . \qquad (4.1)
\end{aligned}$$

The steps in (4.1) are justified as follows: (a) from the definition of L (3.1); (b) from the union bound and the stationarity of $\{X_j\}$; (c) from the definition of T_n (2.10), which implies

$$\Pr\left\{\mathbf{X}_1^{T_n(\mathbf{z})} = \mathbf{z}_1^{T_n(\mathbf{z})}\right\} \leq 1/n \ , \qquad (4.2)$$

and (3.7). Inequality (4.1) is Lemma 3.2A.

We now turn to the proof of Lemma 3.2B. We will need some new definitions. First, as a mathematical convenience, imagine that the two data-bases $\mathbf{X}_1^n$ and $\mathbf{Y}_1^n$ are substrings of an infinite realization $\mathbf{X}_{-\infty}^\infty$, $\mathbf{Y}_{-\infty}^\infty$. For $\ell \leq n$, and arbitrary $\mathbf{z}_1^\infty$ we can write the event

$$\{L(\mathbf{z}_1^\infty, \mathbf{X}_1^n) \leq \ell\} = \bigcap_{j=\ell}^n \left\{\mathbf{X}_{j-\ell+1}^j \neq \mathbf{z}_1^\ell\right\} \ . \qquad (4.3)$$

The last quantity that we will need is the following "recurrence time". Let $\mathbf{x}_{-\ell+1}^\infty$ be arbitrary, $\ell \geq 1$. Let $N_1(\mathbf{x}_{-\ell+1}^\infty)$ be the minimum $N > 0$ such that

$$\mathbf{x}_{-\ell+1}^0 = \mathbf{x}_{N-\ell+1}^N \ .$$

Think of $\mathbf{x}_{-\ell+1}^0$ as a template, and slide it into "positive" time until we get a perfect match. Then N_1 is the number of positions we will need to slide. For example, when $\ell = 4$, and $\mathbf{x}_{-\ell+1}^\infty = \mathbf{x}_{-3}^\infty = (a\,b\,c\,a \mid b\,c\,a\,b\,b\,c\,d\ldots)$, then $N_1 = 3$, since $\mathbf{x}_0^3 = \mathbf{x}_{-3}^0 = (a\,b\,c\,a)$. To make the dependence of $N_1(\mathbf{x}_{-\ell+1}^0)$ on $\mathbf{x}_{-\ell+1}^0$ explicit, define

$$N(\mathbf{x}_{-\ell+1}^0, \mathbf{x}_1^\infty) = N_1(\mathbf{x}_{-\ell+1}^\infty) \ . \qquad (4.4)$$

We now make the important observation that if $N(\mathbf{x}_{-\ell+1}^0, \mathbf{x}_1^\infty) = \nu < \ell$, then the ℓ-sequence $\mathbf{x}_{-\ell+1}^0$ is periodic with period ν. In the above example $\nu = 3$, and $x_{-\ell+1}^0 = \mathbf{x}_{-3}^0 = (a\,b\,c\,a)$ is periodic with period 3. It follows that if $\mathbf{x}_{-\ell+1}^0 \in \mathcal{S}_\ell$, then for all $\mathbf{x}_1^\infty$, $N(\mathbf{x}_{\ell+1}^0, x_1^\infty) > \frac{\ell}{2}$. Let $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}$. Then for $\ell \geq \ell^*$, *given that* $\mathbf{X}_{-\ell+1}^0 = \mathbf{z}_1^\ell$, the event

$$\left\{N(\mathbf{z}_1^\ell, \mathbf{X}_1^\infty) > n\right\} \overset{(a)}{=} \bigcap_{j=1}^n \left\{\mathbf{X}_{j-\ell+1}^j \neq \mathbf{z}_1^\ell\right\}$$

Combining (3.22) and (3.29) we have

$$E|T - L_1| = O(1) \ .$$

It remains to examine $E[T^* - L^*]^+$. As in (3.28), write

$$E[T^* - L_1^*]^+ = E[T^* - L_1^*]^+ I_{\{\mathbf{Z}_1^{*^{\ell^*}} \in \mathcal{S}_{\ell^*}\}} + E[T^* - L_1^*]^+ I_{\{\mathbf{Z}_1^{*^{\ell^*}} \notin \mathcal{S}_{\ell^*}\}} \ . \tag{3.30}$$

¿From (3.27), the first term is $O(1)$, and the second term $\leq \left(\frac{1}{\mu} \log n + 1\right) \Pr \{\mathbf{Z}_1^{*^{\ell^*}} \notin \mathcal{S}_{\ell^*}\}$. We complete the proof (that $E|T^* - L_1^*| = O(1)$) by showing that

$$\Pr \{\mathbf{Z}_1^{*^{\ell^*}} \notin \mathcal{S}_{\ell^*}\} = O\left(\frac{1}{n^{b_2}}\right) \ , \tag{3.31}$$

for $b_2 > 0$. Define the events

$$\begin{aligned}
\mathcal{E}_1 &= \{\mathbf{Z}_1^{*^{\ell^*}} \notin \mathcal{S}_{\ell^*}\} \ , \\
\mathcal{E}_2 &= \{\mathbf{Z}_1^{\ell^*} \notin \mathcal{S}_{\ell^*}\} \ , \\
\mathcal{E}_3 &= \{L_1 > \ell^*\} \ .
\end{aligned}$$

Then

$$\Pr \{\mathbf{Z}_1^{*^{\ell^*}} \notin \mathcal{S}_{\ell^*}\} = \Pr(\mathcal{E}_1) = P(\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3) + P(\mathcal{E}_1\mathcal{E}_2^c\mathcal{E}_3) + P(\mathcal{E}_1\mathcal{E}_3^c) \ . \tag{3.32}$$

Since $\mathcal{E}_1\mathcal{E}_2^c\mathcal{E}_3 =$ the nullset, (3.32) becomes

$$P\{Z^* \notin \mathcal{S}_{\ell^*}\} \leq P(\mathcal{E}_2) + P(\mathcal{E}_3^c) = O\left(\frac{\log n}{n^{1/2\mu_{\mathrm{MIN}}}}\right) + O\left(\frac{1}{n^{b_1}}\right) = O\left(\frac{1}{n^{b_2}}\right) \ , \tag{3.33}$$

where we have used Lemma 3.1 with $\ell = \ell^*$ to bound $P(\mathcal{E}_2)$ and Lemma 3.3A to bound $P(\mathcal{E}_3^c)$. Inequality (3.33) is (3.31), and the proof is complete.

IV Proof of Lemma 3.2

In this section we give the proof of our principal lemma, Lemma 3.2. Let n and $\mathbf{z}_1^\infty$, such that $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell_n^*}$, be given and fixed for this entire section. We begin with part (A) of the lemma.

Write, for $\ell \geq T_n(\mathbf{z})$ $(= T)$,

$$\Pr \{L(\mathbf{z}_1^\infty, \mathbf{X}_1^n) \geq \ell\} \overset{(a)}{=} \Pr \bigcup_{j=\ell}^{n} \left\{\mathbf{X}_{j-\ell+1}^j = \mathbf{z}_1^\ell\right\}$$

Let $\mathbf{z}_1^n$ be such that $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}$ be given, and held fixed (for a while). Write

$$
\begin{aligned}
& E[T_n(\mathbf{z}) - L(\mathbf{z}, \mathbf{X}_1^n)]^+ \\
& \overset{(a)}{=} \sum_{k=1}^{\infty} \Pr\{T_n(\mathbf{z}) - L(\mathbf{z}, \mathbf{X}_1^n) \geq k\} \\
& = \sum_{k=1}^{\infty} \Pr\{L(\mathbf{z}, \mathbf{X}_1^n) \leq T_n(\mathbf{z}) - k\} \\
& \overset{(b)}{=} \sum_{k=1}^{T_n(\mathbf{z})} \Pr\{L(\mathbf{z}, \mathbf{X}_1^n) \leq T_n(\mathbf{z}) - k\} \\
& = \sum_{k=1}^{T_n(\mathbf{z})-\ell^*-1} \Pr\{L(\mathbf{z}, \mathbf{X}_1^n) \leq T_n(\mathbf{z}) - k\} \\
& \quad + \sum_{k=T_n(\mathbf{z})-\ell^*}^{T_n(\mathbf{z})} \Pr\{L(\mathbf{z}, \mathbf{X}_1^n) \leq T_n(\mathbf{z}) - k\} \\
& \overset{(c)}{\leq} \sum_k \left(\alpha^k + t\alpha^{\ell^*} + \frac{\alpha^{\ell^*/2}}{(1-\alpha)}\right) + \ell^* \Pr\{L(\mathbf{z}, \mathbf{X}_1^n) \leq \ell^*\} \\
& \overset{(d)}{=} O(1)\ .
\end{aligned}
\tag{3.26}
$$

The steps in (3.26) are justified as follows: (a) from the tail-sum identity; (b) from $L(\mathbf{z}, \mathbf{X}_1^n) \geq 0$; (c) by applying Lemma 3.2B for $\ell = T_n(\mathbf{z}) - k$ on the first summation, and the fact that $\Pr\{L(\mathbf{z}, \mathbf{X}_1^n) \leq T_n(\mathbf{z}) - k\} \leq \Pr\{L \leq \ell^*\}$, for $T_n(\mathbf{z}) - k \leq \ell^*$; (d) by applying Lemma 3.3A to the second term.

Again by replacing $\mathbf{X}_1^n$ by $\mathbf{Y}_1^n$ and using Lemma 3.3B, we obtain for $\mathbf{z} \in \mathcal{S}_{\ell^*}$,

$$
E[T_n(\mathbf{z}) - L(\mathbf{z}, Y_1^n)]^+ = O(1)\ . \tag{3.27}
$$

Note that the $O(1)$ term in (3.26) and (3.27) does not depend on $\mathbf{z}$.

We complete the proof of the Theorem by writing

$$
E[T - L_1]^+ = E[T - L_1]^+ I_{\{\mathbf{Z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}\}} + E[T - L_1]^+ I_{\{\mathbf{Z}_1^{\ell^*} \notin \mathcal{S}_{\ell^*}\}}\ . \tag{3.28}
$$

¿From (3.26), the first term is $O(1)$. Using (3.10) and Lemma 3.1, the second term is $\leq \left(\frac{1}{\mu}\log n + 1\right)\frac{\ell^*}{2}\alpha^{\ell^*/2} = o(1)$. Thus

$$
E[T - L_1]^+ = O(1)\ . \tag{3.29}
$$

Also (3.19b) implies

$$\mu_{\mathrm{MIN}}(\Delta^* + L_1^* - L_1) \geq \mu\Delta - \mu_{\mathrm{MIN}}$$

or

$$L_1^* - L_1 \geq \frac{\mu\Delta - \mu_{\mathrm{MIN}}\Delta^*}{\mu_{\mathrm{MIN}}} - 1 \ . \tag{3.21b}$$

Inequalities (3.21) imply that for $L_1 < T$,

$$|L_1^* - L_1| \leq c_1|\Delta| + c_2|\Delta^*| + c_3 \ ,$$

which, together with (3.16), implies Lemma 3.5.

We are finally ready to give the proof of Theorem 2.1. In the light of Lemma 3.5, it will suffice to show that

$$E|L_1 - T|, \quad E|L_1^* - T^*| = O(1) \ .$$

We remark that the fact that $E|L_1 - T| = O(1)$ is established in [4], but since we get this fact essentially for free, we will show it too.

For $-\infty < u < \infty$, let $[u]^+ = \max(u, 0)$, so that $|u| = [u]^+ + [-u]^+$. We show first that

$$E[L_1 - T]^+ = O(1) \ . \tag{3.22}$$

We use the "tail-sum" identity (for a r.v. $U \geq 0$, $EU = \sum_{k=1}^{\infty} \Pr\{U \geq k\}$), which implies

$$\begin{aligned} & E\{[L_1 - T]^+ \mid \mathbf{Z} = \mathbf{z}\} \\ & \overset{(a)}{=} \sum_{k=1}^{\infty} \Pr\{L(\mathbf{z}, \mathbf{X}_1^n) - T_n(\mathbf{z}) \geq k\} \\ & = \sum_{k=1}^{\infty} \Pr\ \{L(\mathbf{z}, \mathbf{X}_1^n) \geq T_n(\mathbf{z}) + k\} \overset{(b)}{\leq} \sum_{k=1}^{\infty} \alpha^k = \frac{\alpha}{1-\alpha} < \infty \ . \end{aligned} \tag{3.23}$$

Step (a) follows from the independence of $\mathbf{Z}$ and $L(\mathbf{z}, X_1^n)$ for fixed $\mathbf{z}$, and step (b) from Lemma 3.2A. Averaging (3.23) over $\mathbf{z}$ yields (3.22). On replacing $\mathbf{X}_1^n$ by $\mathbf{Y}_1^n$ in (3.23), we get

$$E[L_1^* - T^*]^+ = O(1) \ . \tag{3.24}$$

We now show that

$$E[T - L_1]^+ = O(1) \ . \tag{3.25}$$

Proof: Let $T = T_n(\mathbf{Z})$ and $T^* = T_n(\mathbf{Z}^*)$. Observe that if $L_1 \geq T$, then $\mathbf{Z}_1^T = \mathbf{Z}_1^{*T}$, and $T = T^*$. Thus, from the triangle inequality,

$$|L_1 - L_1^*| \leq |L_1 - T| + |L_1^* - T^*| \ , \tag{3.16}$$

for $L_1 \geq T$.

Now consider $L_1 < T$. Then, from (2.10), the probability of $\mathbf{Z}_1^{L_1} > 1/n$. Since $\mathbf{Z}^*$ agrees with $\mathbf{Z}$ in the first L_1 symbols, $L_1 \leq T^*$. Write

$$\Pr\left\{\mathbf{Z}_1^k = \mathbf{z}_1^k\right\} = P(z_1) \prod_{j=2}^{k} P(z_k \mid z_{k-1}) \ . \tag{3.17}$$

We have from (2.10) and (3.9) and (3.17) with $k = T$,

$$\log n \leq -\log P(Z_1) - \sum_{j=2}^{T} \log P(Z_j \mid Z_{j-1}) \leq \log n + \mu_{\text{MIN}} \tag{3.17a}$$

and

$$\log n \leq -\log P(Z_1^*) - \sum_{j=2}^{T^*} \log P(Z_j^* \mid Z_{j-1}^*) \leq \log n + \mu_{\text{MIN}} \ . \tag{3.17b}$$

Subtracting (3.17a) from (3.17b), and using $\mathbf{Z}_1^{L_1} = \mathbf{Z}_1^{*L_1}$ we obtain

$$-\mu_{\text{MIN}} \leq -\sum_{j=L_1+1}^{T} \log P(Z_j \mid Z_{j-1}) + \sum_{j=L_1+1}^{T^*} \log P(Z_j^* \mid Z_{j-1}^*) \leq \mu_{\text{MIN}} \ . \tag{3.18}$$

The first inequality in (3.18) and (3.6) and (3.9) yield

$$-\mu_{\text{MIN}} \leq \mu_{\text{MIN}}(T - L_1) - \mu(T^* - L_1) \ , \tag{3.19a}$$

and the second inequality in (3.18) yields

$$\mu(T - L_1) - \mu_{\text{MIN}}(T^* - L_1) \leq \mu_{\text{MIN}} \ . \tag{3.19b}$$

Letting

$$\Delta = T - L_1, \quad \Delta^* = T^* - L_1^* \ , \tag{3.20}$$

(3.19a) implies

$$\mu(\Delta^* + L_1^* - L_1) \leq \mu_{\text{MIN}}\Delta + \mu_{\text{MIN}} \ ,$$

or

$$L_1^* - L_1 \leq \frac{\mu_{\text{MIN}}\Delta - \mu\Delta^*}{\mu} + \frac{\mu_{\text{MIN}}}{\mu} \ . \tag{3.21a}$$

Thus

$$\begin{aligned}\Pr(\mathcal{E}_2) &= \Pr(\mathcal{E}_1^c\mathcal{E}_2) + \Pr(\mathcal{E}_1\mathcal{E}_2) \\ &\le O\left(\frac{1}{n^{b_1}}\right) + \Pr(\mathcal{E}_1) = O\left(\frac{1}{n^{b_1}}\right) ,\end{aligned}$$

where the last step follows on another application of part (A). This is part (B).

Let us look at the "pseudo-source" sequence $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}(\mathbf{Z}, \mathbf{X}_1^n)$, given by (2.5). $\hat{\mathbf{Z}}$ will, of course, be a finite length sequence, which means that $T_n(\hat{\mathbf{Z}})$ may not be well defined. To get around this awkward possibility we extend $\hat{\mathbf{Z}}$ to make an infinite sequence in the following obvious way. We determine (L_1, m_1) exactly as before so that $\hat{\mathbf{Z}}$ is given by (2.5), i.e.

$$\hat{\mathbf{Z}} = (X_{m_1}, X_{m_2}, \ldots, X_n) .$$

For purposes of extending $\hat{\mathbf{Z}}$ only, imagine that $\mathbf{X}$ is an infinite sequence $\mathbf{X}_1^\infty$ (instead of $\mathbf{X}_1^n$), and define the extended pseudo-source

$$\mathbf{Z}^* = \mathbf{Z}_1^{*\infty} = (X_{m_1}, X_{m_2}, \ldots) . \tag{3.14}$$

Then define

$$L_1^* = L(\mathbf{Z}^*, \mathbf{Y}_1^n) . \tag{3.15}$$

Of course $L_1^* \ge \hat{L}_1$. The following lemma allows us to prove Theorem 2.1 using L_1^* instead of L_1.

Lemma 3.4 *With probability* 1,

$$|\hat{L}_1 - L_1| \le |L_1^* - L_1| .$$

Proof: If $\hat{L}_1 \ge L_1$, then (since $L_1^* \ge \hat{L}_1$) $|\hat{L}_1 - L_1| \le |L_1^* - L_1|$. Thus we need consider only the case $\hat{L}_1 < L_1$. If $\hat{L}_1 < L_1$, then $\hat{\mathbf{Z}}_1^{\hat{L}_1}$ is a substring of $\mathbf{X}_1^n$ and $\mathbf{Y}_1^n$, and $\hat{\mathbf{Z}}_1^{\hat{L}_1+1}$ is a substring of $\mathbf{X}_1^n$ but *not* of $\mathbf{Y}_1^n$. Thus $\hat{L}_1 = \hat{L}_1^*$, and $|\hat{L}_1 - L_1| = |L_1^* - L_1|$, completing the proof.

The final lemma that we need is

Lemma 3.5 *There exist constants* $c_1, c_2, c_3 > 0$ *which do not depend on* n *such that, with probability* 1,

$$|L_1 - L_1^*| \le c_1|L_1 - T_n(\mathbf{Z})| + c_2|L_1^* - T_n(\mathbf{Z}^*)| + c_3 .$$

Remarks: (A) The probabilities bounded in (3.12) and (3.13) are computed for fixed $\mathbf{z}$. Of course these bounds will hold if $\mathbf{X}_1^n$ is replaced by $\mathbf{Y}_1^n$.

(B) Lemma 3.2 tells us roughly that, with high probability, $L(\mathbf{z}, \mathbf{X}_1^n)$ is about $T_n(\mathbf{z})$.

We use Lemma 3.2 to prove

Lemma 3.3 *(A) For* $\mathbf{z} = \mathbf{z}_1^n$ *such that* $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}$*,*

$$\Pr\{L(\mathbf{z}, X_1^n) \le \ell^*\} = O\left(\frac{1}{n^{b_1}}\right) ,$$

where $b_1 > 0$.

(B) For $\mathbf{z} = \mathbf{z}_1^n$ *such that* $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}$*, and* $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}(\mathbf{z}_1^n, \mathbf{X}_1^n)$*,*

$$\Pr\{L(\hat{\mathbf{Z}}, Y_1^n) \le \ell^*\} = O\left(\frac{1}{n^{b_1}}\right) ,$$

where $b_1 > 0$.

Proof: To prove (A), apply Lemma 3.2B with $\ell = \ell^*$. Since, from (3.10)

$$T_n(\mathbf{z}) - \ell^* \ge \frac{1}{2\mu_{\mathrm{MIN}}} \log n = \ell^* ,$$

we have from (3.13)

$$\Pr\ \{L(\mathbf{z}, \mathbf{X}_1^n) \le \ell^*\} \le 2\left[\alpha^{\ell^*} + t\alpha^{\ell^*} + \frac{\alpha^{\ell^*/2}}{(1-\alpha)}\right] ,$$

which implies part (A) with $b_1 = \frac{\mu}{4\mu_{\mathrm{MIN}}}$.

To prove (B), for fixed $\mathbf{z}$, define the events

$$\begin{aligned} \mathcal{E}_1 &= \{L(\mathbf{z}, \mathbf{X}_1^n) \le \ell^*\} , \\ \mathcal{E}_2 &= \left\{L(\hat{\mathbf{Z}}(\mathbf{z}, \mathbf{X}_1^n), \mathbf{Y}_1^n) \le \ell^*\right\} . \end{aligned}$$

If $\mathcal{E}_1^c$ occurs, then the first ℓ^* symbols in $\hat{\mathbf{Z}}(\mathbf{z}, X_1^n)$ are $\mathbf{z}_1^{\ell^*}$. Thus if $\mathbf{z}_1^{\ell^*} \in \mathcal{S}_{\ell^*}$, then so will $\hat{\mathbf{Z}}_1^{\ell^*}$. Thus part (A) will apply yielding

$$\Pr(\mathcal{E}_1^c \mathcal{E}_2) = O\left(\frac{1}{n^{b_1}}\right) .$$

Lemma 3.1 $\Pr\left\{\mathbf{Z}_1^\ell \notin \mathcal{S}_\ell\right\} \leq \frac{\ell}{2}\alpha^{\ell/2}$.

Proof: Write

$$\Pr\{\mathbf{Z}_1^\ell \notin \mathcal{S}_\ell\} = \sum_{\mathbf{z}\in\mathcal{A}^{\ell/2}} \Pr\left\{\mathbf{Z}_1^{\ell/2} = \mathbf{z}\right\} \Pr\left\{\mathbf{Z}_1^\ell \notin \mathcal{S}_\ell \mid \mathbf{Z}_1^{\ell/2} = \mathbf{z}\right\} . \tag{3.8}$$

Now if $\mathbf{z}_1^\ell \notin \mathcal{S}_\ell$, it must be periodic with period $\tau \leq \ell/2$. Thus $\mathbf{z}_{\frac{\ell}{2}+1}^\ell$ is determined by $\mathbf{z}_1^{\ell/2}$ and τ, and for a given $\mathbf{z}_1^{\ell/2}$ there are at most $(\ell/2)$ vectors $\mathbf{z}_{\frac{\ell}{2}+1}^\ell$ such that $\mathbf{z}_1^\ell \notin \mathcal{S}_8$ (corresponding to the $\ell/2$ possible values of τ). Thus, using (3.7),

$$\Pr\left\{\mathbf{Z}_1^\ell \notin \mathcal{S}_\ell \mid \mathbf{Z}_1^{\ell/2} = \mathbf{z}\right\} \leq \left(\frac{\ell}{2}\right)\alpha^{\ell/2} ,$$

which, with (3.8), yields Lemma 3.1.

To state the remaining lemmas we need to define

$$\alpha_{\text{MIN}} \triangleq \min_{a,b\in\mathcal{A}} \{\Pr\{Z_1 = b \mid Z_0 = a\} > 0\} , \tag{3.9a}$$

and

$$\mu = \log\frac{1}{\alpha}, \quad \mu_{\text{MIN}} = \log\frac{1}{\alpha_{\text{MIN}}} . \tag{3.9b}$$

Thus, with $T_n(\mathbf{z})$ defined by (2.10), for all[*] $\mathbf{z} = \mathbf{z}_1^\infty$

$$\frac{\log n}{\mu_{\text{MIN}}} \leq T(\mathbf{z}) \leq \frac{\log n}{\mu} + 1 . \tag{3.10}$$

Let us also define

$$\ell^* = \ell_n^* = \frac{1}{2\mu_{\text{MIN}}}\log n . \tag{3.11}$$

We now state what turns out to be the key lemma, leaving a proof for Section IV.

Lemma 3.2 *(A) for all* $\mathbf{z} = \mathbf{z}_1^\infty$*, and all* $\ell \geq T_n(\mathbf{z})$*,*

$$\Pr\left\{L(\mathbf{z}, \mathbf{X}_1^n) \geq \ell\right\} \leq \alpha^{(\ell - T_n(\mathbf{z}))} . \tag{3.12}$$

(B) For all $\mathbf{z}_1^\infty$ *such that* $\mathbf{z}_1^\ell \in \mathcal{S}_{\ell^*}$*, and all* $\ell \in [\ell^*, T_n(\mathbf{z})]$*, there exists a* $t > 0$ *(which does not depend on* n*) such that*

$$\Pr\left\{L(\mathbf{z}, X_1^n) \leq \ell\right\} \leq 2\left[\alpha^{(T_n(\mathbf{z})-\ell)} + t\alpha^{\ell^*} + \frac{\alpha^{\ell^*/2}}{(1-\alpha)}\right] . \tag{3.13}$$

---

[*] 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.

For $\mathbf{z}_1^n, \mathbf{x}_1^n$ sequences from $\mathcal{A}$, let $\hat{\mathbf{Z}}(\mathbf{z}_1^n, \mathbf{x}_1^n)$ be the pseudo-source sequence defined in Section II. Thus if $L(\mathbf{z}_1^n, \mathbf{x}_1^n) = L$, and

$$\mathbf{z}_1^L = \mathbf{x}_m^{m+L-1} \ , \tag{3.3}$$

then $\hat{\mathbf{Z}}(\mathbf{z}_1^n, \mathbf{x}_1^n) = (x_m, x_{m+1}, x_{m+2}, \ldots, x_n)$. If there are more than one m that satisfy (3.3), use the smallest m satisfying (3.3) to define $\hat{\mathbf{Z}}$. Note that $L(\mathbf{Z}_1^n, \mathbf{X}_1^n)$ and $\hat{\mathbf{Z}}$ are random quantities that depend on $\mathbf{Z}_1^n$ and $\mathbf{X}_1^n$. Of course, L_1 (defined by (2.1)) is $L(\mathbf{Z}_1^n, \mathbf{X}_1^n)$, and $\hat{L}_1$ (defined by (2.6)) is $L(\hat{\mathbf{Z}}, \mathbf{Y}_1^n)$.

Let $\mathbf{z}_1^k$ be a k-vector, where $z_i \in \mathcal{A}$, $1 \le i \le k$, and $1 \le k \le \infty$. We say that $\mathbf{z}_1^k$ is *periodic with period* τ (τ an integer) if $z_i = z_{i+j\tau}$, for all integers i, j such that $1 \le i$, $i + j\tau \le k$. For $\ell = 1, 2, \ldots$, define the subset $\mathcal{S}_\ell \subseteq \mathcal{A}^\ell$ by

$$\mathcal{S}_\ell = \{\mathbf{z}_1^\ell \in \mathcal{A}^\ell : \ \mathbf{z}_1^\ell \text{ is } \textit{not} \text{ periodic with period } \tau, \text{ for some } \tau \le \ell/2\} \ . \tag{3.4}$$

Thus, for example, with $\ell = 8$,

$$(a\,b\,c\,a\,b\,c\,a\,b) \notin \mathcal{S}_8 \ ,$$

since it is periodic with period $3 \le 8/2$. But

$$(a\,b\,c\,d\,e\,a\,b\,c) \in \mathcal{S}_8 \ ,$$

since its period is $5 > 8/2$.

Now we show in Appendix A that with no loss of generality we can take the memory-size $M = 1$, i.e. (1.1) is

$$\Pr\left\{Z_k = z_k \mid \mathbf{Z}_{-\infty}^{k-1} = \mathbf{z}_{-\infty}^{k-1}\right\} = \Pr\left\{Z_k = z_k \mid Z_{k-1} = z_{k-1}\right\} \ . \tag{3.5}$$

We assume that

$$\alpha \triangleq \max_{a,b \in \mathcal{A}} \Pr\{Z_k = b \mid Z_{k-1} = a\} < 1 \ . \tag{3.6}$$

We also show in Appendix A how to prove Theorem 2.1 when $\alpha = 1$. We will make use of the fact that for $i < j < k$, and arbitrary $\mathbf{z}$,

$$\begin{aligned} &\Pr\left\{\mathbf{Z}_j^k = \mathbf{z}_j^k \mid \mathbf{Z}_i^{j-1} = \mathbf{z}_i^{j-1}\right\} = \Pr\left\{\mathbf{Z}_j^k = \mathbf{z}_j^k \mid Z_{j-1} = z_{j-1}\right\} \\ &= \prod_{i=j}^{k} \Pr\left\{Z_i = z_i \mid Z_{i-1} = z_{i-1}\right\} \le \alpha^{k-j+1} \ . \end{aligned} \tag{3.7}$$

We can now give the first of a series of lemmas that will lead us to a proof of Theorem 2.1.

Here is some intuition as to why Theorem 2.1 holds. We depend on some facts established in [4]. For a sequence $\{z_k\}_1^\infty$, $z_k \in \mathcal{A}$, define the function $T_n(\mathbf{z})$ by

$$T_n(\mathbf{z}) = \min \{k \geq 1 : \ \Pr\{\mathbf{Z}_1^n = \mathbf{z}_1^n\} \leq 1/n\} \ . \tag{2.10}$$

Thus $T_n(\mathbf{Z}_1^\infty)$ is a random variable. It is shown in [4] that, as $n \to \infty$,

$$ET_n(\mathbf{Z}) = \frac{\log n}{H} + O(1) \ , \tag{2.11}$$

and

$$E|L_1 - T_n(\mathbf{Z})| = O(1) \ . \tag{2.11a}$$

¿From this, it follows that $EL_1 = \frac{\log n}{H} + O(1)$. In our modified algorithm, the source sequence $\mathbf{Z}$ and the pseudo-source $\hat{\mathbf{Z}}$ are "approximately" the same (use (2.1) and (2.5)). Thus $T_n(\mathbf{Z})$ and $T_n(\hat{\mathbf{Z}})$ are about the same, and

$$\begin{aligned} E|\hat{L}_1 - L_1| \ &\leq \ E|T_n(\mathbf{Z}) - L_1| + E|T_n(\hat{\mathbf{Z}}) - \hat{L}_1| \\ &= \ O(1) \end{aligned} \tag{2.12}$$

by (2.11a).

This is an imprecise sketch of the main idea of the proof. A careful argument is required to show that $T_n(\mathbf{Z})$ is close to $T_n(\hat{\mathbf{Z}})$, and that $E|T_n(\hat{\mathbf{Z}}) - \hat{L}_1| = O(1)$. In fact, as we shall see, we will take a more indirect route.

III Proof of Theorem 2.1

We begin with some definitions. Let $\mathbf{z}_1^k, \mathbf{x}_1^n$, $k \leq \infty$, $n < \infty$, be two sequences from the alphabet $\mathcal{A}$. Define $L(\mathbf{z}_1^k, \mathbf{x}_1^n)$ to be the largest integer $L \leq k$, such that a copy of string $\mathbf{z}_1^L$ is a substring of $\mathbf{x}_1^n$, i.e.

$$\mathbf{z}_1^L = \mathbf{x}_m^{m+L-1}, \quad \text{for some} \ \ m \in [1, n-L+1] \ . \tag{3.1}$$

Of course $L \leq k, n$, and for $k \geq n$,

$$L(\mathbf{z}_1^k, \mathbf{x}_1^n) = L(\mathbf{z}_1^n, \mathbf{x}_1^n) \ . \tag{3.2}$$

We can think of $\{\hat{Z}_k\}$ as being a pseudo-data-source that is a substring of $\mathbf{X}_1^n$. We match this pseudo-source in $\mathbf{Y}_1^n$. Thus let $\hat{L}_1$ be the largest integer such that $\hat{\mathbf{Z}}_1^{\hat{L}_1}$ is a substring of $\mathbf{Y}_1^n$, i.e.

$$\hat{\mathbf{Z}}_1^{\hat{L}_1} = \mathbf{Y}_{\hat{m}_2}^{\hat{m}_2+\hat{L}_1-1}, \quad \hat{m}_2 \in [1, n - \hat{L}_1 + 1] \ . \tag{2.6}$$

The crucial point is that $\hat{L}_1$ can be found from $\mathbf{X}_1^n$, $\mathbf{Y}_1^n$ and m_1.

Now the sequence $\mathbf{Z}_1^{L_1}$ is encoded by a binary representation of m_1 (as above, requiring $\log n$ bits), and a binary representation of $\hat{L}_1 - L_1$ (requiring about $\log|\hat{L}_1 - L_1|$ bits). The decoder, which knows $\mathbf{X}_1^n, \mathbf{Y}_1^n$, m_1, $(\hat{L}_1 - L_1)$, computes $\hat{L}_1$ (from $\mathbf{X}_1^n, \mathbf{Y}_1^n, m_1$), and then L_1 from $\hat{L}_1$ and $\hat{L}_1 - L_1$. Thus the decoder can recover the data $\mathbf{Z}_1^{L_1}$ using (2.1).

Again the process is repeated with $\mathbf{Z}_{L_1+1}^{\infty}$ instead of $\mathbf{Z}_1^{\infty}$. The average code rate will be shown in Appendix B to be

$$R_n \approx \frac{\log n + c\, E \log|\hat{L}_1 - L_1|}{E L_1} \ , \tag{2.7}$$

where $c > 0$ is a constant. Since, as above, $EL_1 = \frac{\log n}{H} + O(1)$, we have that the redundancy is

$$\rho_n \sim \frac{H E \log|\hat{L}_1 - L_1|}{\log n} \ . \tag{2.8}$$

The main mathematical task is to estimate $E \log|\hat{L}_1 - L_1|$. We prove the following in Section III.

Theorem 2.1 *As* $n \to \infty$, $E|\hat{L}_1 - L_1| = O(1)$.

Thus from (2.8) and Jensen's inequality,

$$\rho_n = O\left(\frac{1}{\log n}\right), \quad \text{as} \quad n \to \infty \ . \tag{2.9}$$

This is an improvement over (2.4). Although it is of theoretical interest, it is not clear that the improvement in (2.9) is significant from a practical point of view. After all, we replaced $H \log\log n$, which is not very large, by a constant of unknown size. But simulations reported in [1] on a similar scheme indicate that this modification might in fact be practically significant.

(When n is large, L_1 is well defined with high probability.) 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 (2.1). 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.

The average code rate in bits per source symbol can be shown to be [4] (for large n)

$$R = R_n \approx \frac{\log n + E \log L_1}{E L_1}, \tag{2.2}$$

where the indicated expectation is with respect to the data-sequence $\mathbf{Z}$ and the training-sequence $\mathbf{X}$.[†] The numerator in (2.2) is the average number of bits required to transmit a phrase, and the denominator is the average phrase length. Note that since we amortize the bits used for transmitting $\mathbf{X}_1^n$ over the infinite data source $\mathbf{Z}_1^\infty$, these bits do not contribute to the rate.

In Reference [4] it is shown that, as $n \to \infty$,

$$E L_1 = \frac{\log n}{H} + O(1), \tag{2.3a}$$

and

$$E \log L_1 \sim \log \log n \ . \tag{2.3b}$$

Thus the redundancy

$$\rho_n \triangleq R_n - H \sim H \frac{\log \log n}{\log n} \ , \tag{2.4}$$

as $n \to \infty$.

We now describe the modified algorithm, which was motivated by Bender and Wolf [1]. Assume that the encoder has *two* training sequences $\mathbf{X}_1^n, \mathbf{Y}_1^n$, which are independent of each other, and independent of the data sequence $\mathbf{Z}_1^\infty$. Also $\mathbf{X}_1^n, \mathbf{Y}_1^n$ have the same distribution as $\mathbf{Z}_1^n$. The encoder codes both $\mathbf{X}_1^n, \mathbf{Y}_1^n$ with no compression, so that they will be available to the decoder. We now begin to encode $\mathbf{Z}_1^\infty$. Let (m_1, L_1) be exactly as above, so that (2.1) is satisfied. Let $\{\hat{Z}_k\}$ be defined by

$$\left(\hat{Z}_1, \hat{Z}_2, \ldots\right) \triangleq (X_{m_1}, X_{m_1+1}, \ldots, X_n) \ . \tag{2.5}$$

[†] All logarithms in this paper are taken to the base 2.

I Introduction

The fixed-database Lempel-Ziv algorithm (FDLZ) closely resembles practical versions of the LZ algorithm that are widely in use. In Reference [4], it was shown that for finite-memory sources, the *redundancy* ρ_n (defined as the difference between the average code rate and the source entropy, H, when the memory size for the algorithm is n) is $\sim H\frac{\log\log n}{\log n}$, as $n \to \infty$. In Reference [1], Bender and Wolf suggested a variant of LZ which empirically appears to perform well. In this paper we suggest a finite memory version of their scheme, and show that it has redundancy $\rho_n = O\left(\frac{1}{\log n}\right)$, where n is the memory size.

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

$$\Pr\left\{Z_k = z_k \mid \mathbf{Z}_{-\infty}^{k-1} = \mathbf{z}_{-\infty}^{k-1}\right\} = \Pr\left\{Z_k = z_k \mid \mathbf{Z}_{k-M}^{k-1} = \mathbf{z}_{k-M}^{k-1}\right\} , \tag{1.1}$$

where $\mathbf{z}_i^j = (z_i, z_{i+1}, \ldots, z_j)$, $-\infty \leq i < j \leq \infty$. We also assume that $\{Z_k\}$ is aperiodic.[†] 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.

II Description of the Algorithm

We begin with a discussion of the FDLZ algorithm. Let $\mathbf{Z}_1^{\infty}$ be the data source, discussed above, which we want 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 that $\mathbf{X}_1^n$ is available to the decoder. The encoding procedure for $\mathbf{Z}_1^{\infty}$ 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 $\mathbf{X}_1^n$, that is

$$\mathbf{Z}_1^{L_1} = \mathbf{X}_{m_1}^{m_1+L_1-1}, \quad m_1 \in [1, n - L_1 + 1] \ . \tag{2.1}$$

[†] 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.

Improved Redundancy of a Version of the Lempel-Ziv Algorithm

by

Aaron D. Wyner
AT&T Bell Laboratories
Murray Hill, NJ 07974
USA
(adw@research.att.com)

and

Abraham J. Wyner
Dept. of Statistics
Stanford University
Stanford, CA 94305
USA
(ajw@playfair.stanford.edu)

Abstract

The Lempel-Ziv data compression algorithm has the property that for finite-memory sources the redundancy ρ_n (defined as the difference between the average code rate and the entropy when the memory size is n) is $O\left(\frac{\log\log n}{\log n}\right)$. We suggest a new version of the algorithm with redundancy $\rho_n = O\left(\frac{1}{\log n}\right)$.