Chapter 1

Lempel-Ziv Compression

1.1 Universal Lossless Compression

We first set the benchmark using the performance of an optimal compressor that knows the source statistics, and construct a universal compression scheme that doesn't know the source statistics but is asymptotically optimal.

Consider the problem of compressing a source sequence x^n with some source code. For the sake of brevity, we will consider the most common case that the source code outputs a binary sequence. Our conclusions carry over to non-binary alphabets easily.

Definition 1.1.1. *A source code for an n-block source sequence, $\mathcal{C}_n$, is defined as a mapping from a source sequence x^n to a binary sequence of finite length, i.e.*

$$\mathcal{C}_n : \mathcal{X}^n \to \{0,1\}^*. \tag{1.1}$$

More explicitly,

$$\mathcal{C}_n(x^n) = b_1, b_2, \cdots, b_{l_n}, \tag{1.2}$$

where $l_n = l_n(x^n)$ is a length of the output sequence which depends on the input sequence, and $b_i \in \{0,1\}$, $i = 1, \cdots, l_n$.

Definition 1.1.2. *A source code $\mathcal{C}_n$ for an n-block source sequence is said to be "lossless", or "non-singular" [5], if $\mathcal{C}_n(x^n) \neq \mathcal{C}_n(\tilde{x}^n)$ for all $x^n \neq \tilde{x}^n$. Furthermore, a source code $\mathcal{C}_n$ is said to be uniquely decodable (UD) if all its extensions, i.e. the concatenations of successive blocks, are still non-singular.*

For any random source sequence X^n and any UD source code, we know the following bounds on the minimum achievable average description length

$$H(X^n) \leq \min_{\forall \mathcal{C}_n : \mathrm{UD}} \mathbb{E} l_n(X^n) \leq H(X^n) + 1 \tag{1.3}$$

Thus, when we consider a source process $\mathbf{X}$, and look at the average per-symbol description length, we have

$$\lim_{n\to\infty} \min_{\forall \mathcal{C}_n:\mathrm{UD}} \mathbb{E}\left[\frac{1}{n} l_n(X^n)\right] = \lim_{n\to\infty} \frac{1}{n} H\left(X^n\right) \tag{1.4}$$

$$\triangleq \mathbb{H}\left(\mathbf{X}\right), \qquad \text{(when the limit exists)} \tag{1.5}$$

where $\mathbb{H}\left(\mathbf{X}\right)$ is the *entropy rate* of the random process $\mathbf{X}$.

Exercise 1.1.3. *For a stationary random process* $\mathbf{X}$*, show that*

(a) $\lim_{n\to\infty} \frac{1}{n} H\left(X^n\right)$ *exists.*

(b) $\lim_{n\to\infty} \frac{1}{n} H\left(X^n\right)$ *is equal to* $\lim_{k\to\infty} H\left(X_0|X_{-k}^{-1}\right)$*.*

(c) $\lim_{k\to\infty} H\left(X_0|X_{-k}^{-1}\right)$ *is also equal to* $H(X_0|X_{-\infty}^{-1})$ *(only for those who have taken measure theoretic probability).*

Exercise 1.1.3 implies that the limit exists for any stationary random process $\mathbf{X}$. Furthermore, it is well-known that Huffman code can achieve the minimum in (1.5). However, such a code strongly depends on the distribution of the source sequence. *What if we do not know the distribution of the source sequence?*

Definition 1.1.4. *A (sequence of) scheme(s) is universal if*

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l_n(X^n)\right] = \mathbb{H}\left(X\right) \tag{1.6}$$

for every stationary source $\mathbf{X}$*.*

Clearly, Shannon code and Huffman code do not fall into this category due to their dependence on the source distribution. Also, note that it is not a priori clear that to show any such a scheme exists. However, we will see one celebrated example of such a scheme: the Lempel-Ziv (LZ) compressor. Among various LZ compression schemes, we will focus, for concreteness, on the version known as "LZ78".

1.2 Lempel-Ziv Compression

The main idea of LZ compression is:

- parse the source sequence into phrases such that each phrase is the shortest phrase not seen earlier (incremental parsing)
- describe (encode) each new phrase by describing the index of the phrase from the past that forms its prefix and the new symbol at the end of the phrase.

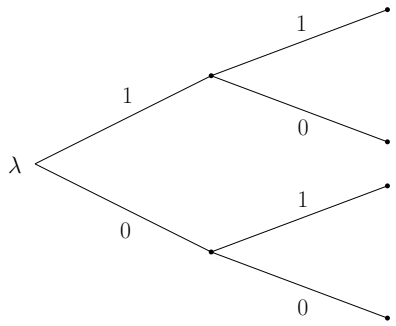

Figure 1.1: LZ tree associated with x^n

Example 1.2.1. *Suppose we want to compress the bit stream*

$$x^n : 0\ 1\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ 0\ \cdots\ x_n$$

Then, we parse x^n into phrases as

$$x^n : 0,1,1\ 0,1\ 1,0\ 1,1\ 1\ 0,0\ 1\ 0,1\ 1\ 0\ 0,0\ \cdots\ x_n$$

Since this operation is basically adding one new symbol to a previously encountered phrase, it naturally induces the tree in Figure 1.1.

Given the complete parsed phrases, LZ encoding is simply to index the previously encountered phrase with the additional symbol comprising the new phrase. Let the index of the empty string be zero. Then, the LZ code for x^n is given by

$$(0,0),(0,1),(2,0),(2,1),(1,1),(4,0),(5,0),(6,0),\cdots .$$

While there are many tweaks that can be applied to boost performance in practice, we will stay with this basic LZ compression scheme for concreteness and ease of analysis.

1.3 The Universality of the LZ Compression

Let $N_{LZ} = N_{LZ}(x^n)$ be the number of phrases in the LZ parsing of x^n. We can describe the length of the source sequence n, and the total number of phrases N_{LZ} with no more than $\log n$ and $\log N_{LZ}$ bits, respectively. Since the LZ compressor encodes each phrase with the index of the phrase from the past that forms its prefix and the new symbol at the end of the phrase, we can describe x^n with a number of bits per source symbol no larger than

$$\begin{aligned}&\frac{1}{n}\left[\log(n) + N_{LZ}(\log N_{LZ} + 1)\right]\\ &\qquad = \frac{1}{n} N_{LZ} \log N_{LZ} + o(1),\end{aligned} \tag{1.7}$$

where we use the fact that $N_{LZ} \leq \frac{n}{(1-\epsilon_n)\log n}$ where $\epsilon_n \to 0$ as $n \to \infty$ [5].

Definition 1.3.1. *A "Markov Kernel" of order k is a mapping*

$$Q: \quad \mathcal{X}^k \to \mathcal{M}(\mathcal{X}),$$

where $\mathcal{M}(\mathcal{X})$ is the simplex of probabilities on $\mathcal{X}$.

This Q corresponds to a Markov source of order k. We write

$$Q(x^n|x_{-(k-1)}^0) = \Pi_{i=1}^n Q(x_i|x_{i-k}^{i-1}). \tag{1.8}$$

Let $\mathcal{M}_k$ denote the class of all Markov kernels of order k.

Theorem 1.3.2. *For $\forall n, k$, and any individual sequence $x_{-(k-1)}^n$*

$$\frac{1}{n} N_{LZ}(x^n) \log N_{LZ}(x^n) \leq \min_{Q \in \mathcal{M}_k} \frac{1}{n} \log \left(\frac{1}{Q(x^n|x_{-(k-1)}^0)} \right) + \epsilon_n^{(k)}, \tag{1.9}$$

where $\epsilon_n^{(k)}$ is independent of the underlyng sequence x^n, and satisfies $\epsilon_n^{(k)} \overset{n\to\infty}{\longrightarrow} 0$.

Before proving Theorem 1.3.2, we state and prove the main result about the LZ's universality in the stochastic setting.

Theorem 1.3.3. *The LZ scheme is universal, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E} l_{\mathrm{LZ}}(X^n) = \mathbb{H}(\mathbf{X}) \quad \textit{for every stationary process } \mathbf{X}. \tag{1.10}$$

We first prove Theorem 1.3.3 based on Theorem 1.3.2. In order to use Theorem 1.3.2, we have to define the conditional empirical entropy.

Definition 1.3.4. *The conditional empirical entropy of order k associated with $x_{-(k-1)}^n$ is defined as:*

$$H_k(x_{-(k-1)}^n) = H(U_{k+1}|U^k), \tag{1.11}$$

where $H(U_{k+1}|U^k)$ is the conditional entropy of random variable U_{k+1} given U^k with joint distribution $P_{U^{k+1}}(u^{k+1}) = \frac{1}{n}|\{1 \leq i \leq n : x_{i-k}^i = u^{k+1}\}|$.

The following exercise shows two important properties of H_k:

Exercise 1.3.5. *Show That*

(a) $\min_{Q\in\mathcal{M}_k} \frac{1}{n} \log \frac{1}{Q(x^n|x_{-(k-1)}^0)} = H_k(x_{-(k-1)}^n)$
Hint: *The proof follows closely the way to prove* $\min_{Q\in\mathcal{M}_0} \frac{1}{n} \log \frac{1}{Q(x^n)} = H_0(x^n)$.

(b) For any stationary process $\mathbf{X}$, $\mathbb{E}H_k(X_{-(k-1)}^n) \leq H(X_0|X_{-k}^{-1})$.
Hint: *Jensen's inequality.*

Now, let us prove Theorem 1.3.3.

Proof

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\mathbb{E}l_{\mathrm{LZ}}(X^n) &= \limsup_{n\to\infty}\frac{1}{n}\mathbb{E}N_{\mathrm{LZ}}(X^n)\log N_{\mathrm{LZ}}(X^n)\\ &\overset{(i)}{\leq} \limsup_{n\to\infty}\mathbb{E}\left[\min_{Q\in\mathcal{M}_k}\frac{1}{n}\log\frac{1}{Q(X^n|X^0_{-(k-1)})}\right]\\ &\overset{(ii)}{=} \limsup_{n\to\infty}\mathbb{E}\left[H_k(X^n_{-(k-1)})\right]\\ &\overset{(iii)}{\leq} H(X_0|X_{-k}^{-1}),\end{aligned} \tag{1.12}$$

where (i) comes from Theorem 1.3.2, (ii) is because of Exercise 1.3.5 Part (a), and (iii) is due to Exercise 1.3.5 Part (b).

Let $k\to\infty$ on the right hand side. It implies that

$$\limsup_{n\to\infty}\frac{1}{n}\mathbb{E}l_{\mathrm{LZ}}(X^n)\leq\mathbb{H}(\mathbf{X}) \tag{1.13}$$

Obviously, the LZ code is a lossless so

$$\liminf_{n\to\infty}\frac{1}{n}\mathbb{E}l_{\mathrm{LZ}}(X^n)\geq\mathbb{H}(\mathbf{X}), \tag{1.14}$$

which completes the proof.

□

Let us introduce a few more notations before proving Theorem 1.3.2. Denote y_i as the ith phrase in the LZ parsing of x^n, e.g.

$$x^n=\underbrace{x_1}_{y_1},\underbrace{x_2x_3}_{y_2},\underbrace{x_4\ldots}_{y_3},\ldots,\underbrace{\ldots x_{n-1}x_n}_{y_{N_{\mathrm{LZ}}}}$$

Let v_i denote the index of the start of the ith phrase, and therefore $y_i=x_{v_i}^{v_{i+1}-1}$. We append an arbitrary k-tuple $x^0_{-(k-1)}$ to x^n in order to prevent an edge effect. Denote c_{l,u^k} as the number of phrases with length l and left context u^k, i.e.,

$$c_{l,u^k}=|\left\{1\leq i\leq N_{\mathrm{LZ}}:|y_i|=l,x_{v_i-k}^{v_i-1}=u^k\right\}|,$$

where $|y_i|$ denotes the length of phrase y_i.

Theorem 1.3.6 (Ziv's Inequality)**.** *For any $x^n_{-(k-1)}$, and $Q\in\mathcal{M}_k$, the following inequality holds:*

$$\sum_{l,u^k}c_{l,u^k}\log c_{l,u^k}\leq\log\frac{1}{Q(x^n|x^0_{-(k-1)})}. \tag{1.15}$$

Proof

$$\begin{aligned}
\log Q(x^n|x^0_{-(k-1)}) &= \log Q(y_1, y_2, ..., y_{N_{\mathrm{LZ}}}|x^0_{-(k-1)}) \\
&= \sum_{i=1}^{N_{\mathrm{LZ}}} \log Q(y_i|y^{i-1}, x^0_{-(k-1)}) \\
&= \sum_{l,u^k} c_{l,u^k} \sum_{i:|y_i|=l, x_{v_i-k}^{v_i-1}=u^k} \frac{1}{c_{l,u^k}} \log Q(y_i|x_{v_i-k}^{v_i-1}) \\
&\overset{(i)}{\leq} \sum_{l,u^k} c_{l,u^k} \log \left(\frac{1}{c_{l,u^k}} \sum_{i:|y_i|=l, x_{v_i-k}^{v_i-1}=u^k} Q(y_i|x_{v_i-k}^{v_i-1}) \right) \\
&\overset{(ii)}{\leq} \sum_{l,u^k} c_{l,u^k} \log \left(\frac{1}{c_{l,u^k}} \right),
\end{aligned} \tag{1.16}$$

where (i) comes from Jensen's inequality and the concavity of log, and (ii) comes from the fact that y_i are distinct. □

Remark

- Since $\sum_{l,u^k} c_{l,u^k} = N_{\mathrm{LZ}}$, we have $\sum_{l,u^k} \frac{c_{l,u^k}}{N_{\mathrm{LZ}}} = 1$. Thus $\frac{c_{l,u^k}}{N_{\mathrm{LZ}}}$ can be interpreted as a probability mass function on a pair (L, U^k).
- Since $\sum_{l,u^k} l c_{l,u^k} = n$, we have $\sum_{l,u^k} \frac{c_{l,u^k}}{N_{\mathrm{LZ}}} l = \frac{n}{N_{\mathrm{LZ}}}$. Thus $\mathbb{E}L = \frac{n}{N_{\mathrm{LZ}}}$.

Exercise 1.3.7. *(a) Let L be a nonnegative integer-valued random variable with $\mathbb{E}L \leq \mu$, then*

$$H(L) \leq (\mu + 1) \log \mu - \mu \log \mu$$

Hint: *Prove that equality is attained when L has a geometric distribution.*

(b) Show that

$$N_{\mathrm{LZ}} \leq K \frac{n}{\log n},$$

where K is a constant independent of n and x^n.
Hint: *the lengths of the phrases are growing so one cannot pack too many of them in a sequence of length n.*

(c) Show that

$$\frac{N_{\mathrm{LZ}}}{n} \sum_{l,u^k} \frac{c_{l,u^k}}{N_{\mathrm{LZ}}} \log \frac{N_{\mathrm{LZ}}}{c_{l,u^k}} \leq \epsilon_n^{(k)},$$

where $\epsilon_n^{(k)}$ *is independent of* x^n *and* $\lim_{n\to\infty}\epsilon_n^{(k)}=0$.
Hint: $\sum_{l,u^k}\frac{c_{l,u^k}}{N_{\rm LZ}}\log\frac{N_{\rm LZ}}{c_{l,u^k}}=H(L,U^k)\le H(L)+H(U^k)\le H(L)+k\log|\mathcal{X}|$, *and now use the previous two parts.*

Now we are ready to prove Theorem 1.3.2.

Proof The idea is to show that $\frac{1}{n}N_{\rm LZ}\log N_{\rm LZ}$ is close to $\frac{1}{n}\sum_{l,u^k}c_{l,u^k}\log c_{l,u^k}$ so that we can use Ziv's Inequality to finish the proof.

$$\begin{aligned}\sum_{l,u^k}c_{l,u^k}\log c_{l,u^k}=&N_{\rm LZ}\sum_{l,u^k}\frac{c_{l,u^k}}{N_{\rm LZ}}\log\frac{c_{l,u^k}}{N_{\rm LZ}}+N_{\rm LZ}\sum_{l,u^k}\frac{c_{l,u^k}}{N_{\rm LZ}}\log N_{\rm LZ}\\ \ge&-n\epsilon_n^{(k)}+N_{\rm LZ}\log N_{\rm LZ},\end{aligned}\tag{1.17}$$

where the inequality comes from Exercise 1.3.7 Part 3 and the fact $\sum_{l,u^k}\frac{c_{l,u^k}}{N_{\rm LZ}}=1$. Thus

$$\begin{aligned}\frac{1}{n}N_{\rm LZ}\log N_{\rm LZ}&\le\frac{1}{n}\sum_{l,u^k}c_{l,u^k}\log c_{l,u^k}+\epsilon_n^{(k)}\\ &\le\log\frac{1}{Q(x^n|x^0_{-(k-1)})}+\epsilon_n^{(k)},\end{aligned}\tag{1.18}$$

where the second inequality comes from Ziv's Inequality. The proof is complete by the arbitrariness of $Q\in\mathcal{M}_k$. □

1.4 LZ78 and Individual Sequences

We have considered the performance of the LZ in expectation sense and we proved that it is universal in stochastic setting. In this section, we will show that the LZ is as good as any finite-state encoder for any individual sequence.

A finite-state encoder E is characterized by a triplet $(\mathcal{S},g,f)$.

- $\mathcal{S}$ is a finite set of states.
- $g:\mathcal{S}\times\mathcal{X}\to\mathcal{S}$ is a state-update function.
- $f:\mathcal{S}\times\mathcal{X}\to\{0,1\}^*$ is an encoding function.

Let $E_n(s)$ be the set of all lossless (for n-blocks) finite-state encoders for which $|\mathcal{S}|\le s$. The compression rate of a sequence x^n by an encoder E be defined as $\rho_E(x^n)=\frac{l(x^n)}{n}$.

Define $\rho_s(x^n)=\min_{E\in E_n(s)}\rho_E(x^n)$, which is the best compression rate for seqeunce x^n among all s-state compressors. We further define the performance for an infinite sequence x^∞ in a limit supremum sense: $\rho_s(x^\infty)=\limsup_{n\to\infty}\rho_s(x^n)$. Finally, by allowing the number of states to grow we define the *finite-state compressibility* of a sequence: $\rho(x^\infty)=\lim_{s\to\infty}\rho_s(x^\infty)$. Observe that this limit exists because $\rho_s(x^\infty)$ is both nonincreasing in s and bounded.

One can characterize $\rho(x^\infty)$ as the best asymptotic compression possible within the set of finite-state encoding schemes — even when a scheme is designed *specifically for the given sequence* x^n. The following theorem is then somewhat surprising.

Theorem 1.4.1. *Let $\ell_{LZ}(x^n)$ be the description length of a sequence x^n by the LZ78 encoder. Then for any infinite sequence x^∞,*

$$\limsup_{n\to\infty} \frac{1}{n}\ell_{LZ}(x^n) \leq \rho(x^\infty).$$

Proof See the proof of Theorem 2 in [6]. □

1.5 A bit about Sliding Window Compression (LZ77)

This alternative to LZ78 was actually introduced earlier, in [7]. The structure of the algorithm is quite similar, but the method of parsing (and referring to previous encodings) is slightly different.

Suppose x_1 through x_n have already been parsed. In LZ78, the next phrase $x_n^{n+\ell}$ would be chosen so as to be the shortest phrase that has not yet been selected as a phrase. In LZ77, it is instead the shortest phrase that has not occurred as a subsequence *anywhere* in x^n. This procedure can be formally defined as follows.

Define

$$L(n, x^\infty) = \min\left\{\ell \neq 0 : x_{n+1}^{n+\ell} \neq x_{i+1}^{i+\ell} \text{ for any } i \in \{0, \ldots, n-\ell-1\}\right\}.$$

The kth phrase in the LZ77-parsing of a sequence x^∞ is then given by $x_{N_k}^{N_k+L(N_k,x^\infty)}$ where $N_1 = 1$ and $N_{k+1} = N_k + L(N_k, x^\infty) + 1$. In the remaining discussion, we will refer to the kth phrase length $L(N_k, x^\infty)$ as L_k.

The encoder encodes $x_{N_k}^{N_k+L_k}$ by specifying the following:

(a) "Where" in the past ($x_1^{N_k-1}$) the unoriginal component $x_{N_k}^{N_k+L_k-1}$ of the new phrase $x_{N_k}^{N_k+L_k}$ appeared.

(b) The length of the phrase L_k.

(c) The "novel" component $x_{N_k+L_k}$ of the new phrase $x_{N_k}^{N_k+L_k}$.

We first quantify the number of bits expended for encoding the kth phrase. The first component costs no more than $\log N_k$ bits (since this is an integer between 1 and N_k). The second component costs no more than $\log L_k + O(\log\log L_k)$ bits (since, as one can show, any integer i can be losslessly represented with length no greater than $\log i + O(\log\log i)$). The third component requires only $\log(|\mathcal{X}| - 1)$ bits.

The number of bits expended per source symbol is then given by (in the limit of large L_k and N_k

$$\frac{\text{bits expended}}{\text{source symbol}} = \frac{\log N_k}{L(N_k, x^\infty)}.$$

In a perfect world, whenever the source is drawn from a stationary/ergodic process $\mathbf{X}$, we would like this quantity to approach the entropy rate $H(\mathbf{X})$ as k grows. Lempel and Ziv demonstrated the following fact, which isn't quite as powerful a statement, but close.

Theorem 1.5.1. *For any stationary and ergodic process* $\mathbf{X}$*,*

$$\frac{\log n}{L(n, \mathbf{X})} \overset{n\to\infty}{\to} \mathbb{H}(\mathbf{X}).$$

Proof See Theorem 1 in [8]. □

Bibliography

[1] R. Durrett, Probability: Theory and Examples, Duxbury Press, 1996.

[2] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo Weinberger,"Universal Discrete Denoising: Known Channel," *IEEE Trans. Inf. Theory,* vol. IT-51, no. 1, pp. 5–28, 2005.

[3] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger, "Inequalities for the L1 Deviation of the Empirical Distribution", *Tech. Report, Information Theory Research Group, HP Laboratories*, 2003.

[4] K. Viswanathan and E. Ordentlich, "Lower Limits of Discrete Universal Denoising," *in Proc. of IEEE International Symposium on Information Theory 2006,* July 2006, pp. 2363–2367.

[5] T. M. Cover and J. A. Thomas, Elements of Information Theory, Wiley, 1991.

[6] J. Ziv, A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Info. Theory*, vol. IT-24, pp. 530 – 536, Sept 1978.

[7] J. Ziv, A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Info. Theory*, vol. IT-23, pp. 337 - 343, May 1977.

[8] Wyner, A.D., Ziv, J. , "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Info. Theory*, vol.35, no.6, pp.1250-1258, Nov 1989.