Electrical Engineering 229A Lecture 13 Notes

Daniel Raban

October 7, 2021

1 Optimality of Lempel-Ziv Coding, The Burrows-Wheeler Transform, and Optimal Compression of IID Sequences

1.1 Asymptotic optimality of Lempel-Ziv coding

Last time, we were in the discussing LZ'77 for a general ergodic process $(X_n, n \in \mathbb{Z})$ with $X_n \in \mathscr{X}$ (finite). For any $L \geq 1$, we defined

$$R_L(X_0^{L-1}) := \min\{j \geq 1 : X_{-j}^{-j+L-1} = X_0^{L-1}\}.$$

The compressor conveys $R_L(X_0^{L-1})$ to the decompressor. The compressor knows $(X_n, n \leq -1)$ and X_0^{L-1}; the decompressor only knows the past, $(X_n, n \leq -1)$. This suffices for the decompressor to determine X_0^{L-1}. By comma-free encoding, it suffices to send

$$\log R_L(X_0^{L-1}) + \log\log R_L(X_0^{L-1}) + 5$$

many bits. Then

$$\begin{aligned}\mathbb{E}[R_L(X_0^{L-1})] &\leq \log \mathbb{E}[R_L(X_0^{L-1})] \\ &= \log \frac{1}{p(X_0^{L-1})}, \\ &= H(X_0, \dots, X_{L-1})\end{aligned}$$

by Kac's lemma. So for fixed L,

$$\frac{1}{L}\mathbb{E}[\log R_L(X_0^{L-1})] \leq \frac{1}{L} H(X_0^{L-1}).$$

So

$$\limsup_{L \to \infty} \frac{1}{L}\mathbb{E}[\log R_L(X_0^{L-1})] \leq H,$$

the entropy rate. Also,

$$\frac{1}{L}\mathbb{E}[\log\log R_L(X_0^{L-1})] \leq \frac{1}{L}\log \mathbb{E}[\log R_L(X_0^{L-1})]$$

$$\begin{aligned} &\le \frac{1}{L}\log H(X_0^{L-1}) \\ &\xrightarrow{L\to\infty} 0. \end{aligned}$$

Finally, $5/L \to 0$ as $L \to \infty$, as well. So in total, we get

$$\frac{1}{L}\,\mathbb{E}[\log R_L(X_0^{L-1}) + \log\log R_L(X_0^{L-1}) + 5] \xrightarrow{L\to\infty} H.$$

1.2 The Burrows-Wheeler transform

Here is an algorithm that some people claim works better than the Lempel-Ziv coding scheme.

Example 1.1. To compress the string SHANNON, a string from the English alphabet, we'll consider all the cyclic permutations and lexicographically order them:

$$\begin{array}{lcl} \text{SHANNON} & & \text{ANNONSH} \\ \text{HANNONS} & & \text{HANNONS} \\ \text{ANNONSH} & & \text{NNONSHA} \\ \text{NNONSHA} & \mapsto & \text{NONSHAN} \\ \text{NONSHAN} & & \text{NSHANNO} \\ \text{ONSHANN} & & \text{ONSHANN} \\ \text{NSHANNO} & & \text{SHANNON} \end{array}$$

Transmit the last column (in compressed form) and the number of the row that has the empirical string. The decompressor (after decompression) gets HSANONN and the number 7.

The decompressor can now recover the first column by lexicographically ordering the symbols (because each symbol in the last column shows up the same number of times it does in the original string). Then, the decompressor knows a list of pairs of symbols (the first and last symbol of each row). Using this, the decompressor can now figure out the second column by cyclically permuting these pairs and lexicographically ordering them, and so on. In this way, the decompressor can recover the original string.

Why does this compress the message? Compressing the last column can be done by e.g. arithmetic coding and works to compress down to the entropy rate for sequences from an ergodic process because (as the length of the sequence goes to infinity, and for each fixed L), the last column becomes piecewise iid with $|\mathscr{X}|^L$ pieces. The piece for x_0^{L-1} appears (asymptotically in n) $np(x_0^{L-1})$ times and has marginal with law $p(x \mid x_0^{L-1})$. Here,

$$H(X_L \mid X_0^{L-1}) = \sum_{x_0^{L-1}} p(x_0^{L-1}) H(p(x \mid x_0^{L-1}), x \in \mathscr{X})$$

is just the L-Markov approximation to the entropy rate. So we can compress to the entropy rate as $L \to \infty$.

1.3 Compression of iid sequences at rate R bits/symbol

Leading up to distributed data compression, we will first discuss the fixed length to fixed length (fixed-to-fixed) formulation of point to point data compression. To recognize the relevance of entropy, we need to allow for a probability of error in decompression (which becomes vanishingly small as the block length increases).

Definition 1.1. We'll say that compression can be done **at rate R bits/symbol** if there is a sequence of pairs of maps $((e_n, d_n) : n \geq 1)$ where $e_n : \mathscr{X}^n \to [M_n] := \{1, \dots, M_n\}$ and $d_n : [M_n] \to \mathscr{X}^n$ such that

$$\limsup_{n\to\infty} \frac{1}{n} M_n \leq R$$

and

$$\lim_{n\to\infty} \mathbb{P}(d_n(e_n(X_1^n)) \neq X_1^n) = 0.$$

Theorem 1.1. *Let $X_1, X_2, \dots$ be iid $\mathscr{X}$-valued with entropy rate H. Then compression can be done at rate H and cannot be done at any rate $< H$.*

Remark 1.1. This theorem is also true for arbitrary stationary sequences, but we will not prove that here.

Proof. Achievability: First, observe that it's enough to show that for all $\varepsilon > 0$, compression can be done at rate $H + \varepsilon$; this is because we can take $\varepsilon = 1/m$ for large enough n (depending on m). Recall that $A_\delta^{(n)}$ denotes the set of weakly ε-typical sequences. We know that $|A_\delta^{(n)}| \leq 2^{n(H+\delta)}$ and

$$\lim_{n\to\infty} \mathbb{P}(X_1^n \in A_\delta^{(n)}) = 1.$$

So if $e_n : \mathscr{X}^n \to [\lceil 2^{n(H+\delta)} \rceil + 1]$ gives a unique image to each element of $A_\delta^{(n)}$ and maps $(A_\delta^{(n)})^c$ to a single image, then

$$\limsup_{n\to\infty} \frac{1}{n} \log(\lceil 2^{n(H+\delta)} \rceil + 1) = H + \delta,$$

and

$$\mathbb{P}(d_n(e_n(X_1^n)) \neq X_1^n) \leq \mathbb{P}(X_1^n \in (A_\delta^{(n)})^c) \xrightarrow{n\to\infty} 0.$$

So take $\delta = \varepsilon$.

Converse: Given any $((e_n, d_n), n \geq 1)$, denote $W_n = e_n(X_1^n)$ and $\widehat{X}_1^n = d_n(W_n)$. Then we have the Markov chain $X_1^n - W_n - \widehat{X}_1^n$. We get from Fano's inequality that

$$H(X_1^n \mid W_n) \leq \underbrace{\mathbb{P}(d_n(e_n(X_1^n)) \neq \widehat{X}_1^n)}_{p_{\text{error}}^{(n)}} \log |\mathscr{X}| + \underbrace{h(p_{\text{error}}^{(n)})}_{\leq 1}$$

So if $p_{\text{error}}^{(n)} \to 0$ as $n \to \infty$, then

$$\frac{1}{n} H(X_1^n \mid W_n) \to 0$$

as $n \to \infty$. But

$$\frac{1}{n} H(X_1^n, W_n) = \frac{1}{n} H(W_n) + \frac{1}{n} H(X_1^n \mid W_n) \to 0,$$

and

$$\frac{1}{n} H(X_1^n, W_n) = \frac{1}{n} H(X_1^n) + \frac{1}{n} H(W_n \mid X_1^n).$$

So

$$\liminf_{n\to\infty} \frac{1}{n} H(W_n) \geq H(X_1),$$

which is the entropy rate of the iid sequence. Hence,

$$\liminf_{n} \frac{1}{n} \log M_n \geq H(X_1). \qquad \square$$