SOURCE CODING based on WAITING

Frans M.J. Willems[1]

IEEE EURASIP Spain Seminar on Signal Processing, Communication and Information Theory,
Universidad Carlos III de Madrid,
December 11, 2014

[1]Eindhoven University of Technology

WinZip and LZ77

WinZip is based on LZ77, a lossless compression method proposed by Lempel & Ziv [1977].
Compression is achieved by replacing repeated segments in the data with pointers. To avoid deadlock an uncoded symbol is added to each pointer.

Example LZ77:

search buffer	look-ahead buffer		output
	a b r a c a d a b	r a _	(0,-,a)
a	b r a c a d a b r	a _	(0,-,b)
a b	r a c a d a b r a	_	(0,-,r)
a b r	a c a d a b r a _		(3,1,c)
a b r a c	a d a b r a _		(2,1,d)
a b r a c a d	a b r a _		(7,4,_)
a b r a c a d a b r a _			

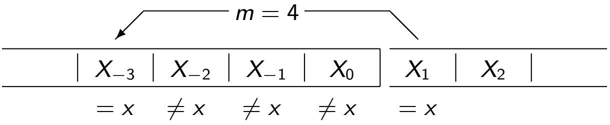
Question: Why does this method work? Note that the statistics of the data are unknown!

Waiting Times

Consider the discrete stationary and ergodic process

$$\cdots, X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, \cdots .$$

Suppose that $X_1 = x$ for symbol-value $x \in \mathcal{X}$ with $\Pr\{X_1 = x\} > 0$. We say that the *waiting time* of the x that occurred at time $t = 1$ is m if $X_{1-m} = x$ and $X_t \neq x$ for $t = 2 - m, \cdots, 0$.



Let $Q_x(m)$ be the conditional probability that the waiting time of the x occurring at $t = 1$ is m. Hence

$$Q_x(m) \stackrel{\Delta}{=} \Pr\{X_{1-m} = x, X_{2-m} \neq x, \cdots, X_0 \neq x | X_1 = x\}.$$

The *average* waiting time for symbol-value x with $\Pr\{X_1 = x\} > 0$ is defined as

$$T(x) \stackrel{\Delta}{=} \sum_{m=1,2,\cdots} m Q_x(m).$$

Kac's Result

Example: Consider an i.i.d. (binary) process and assume that $\Pr\{X_1 = 0\} = p > 0$. Then

$$\begin{aligned} Q_0(m) &= p(1-p)^{m-1} \text{ and} \\ T(0) &= \sum_{m=1,2,\cdots} mp(1-p)^{m-1} = \frac{1}{p}. \end{aligned}$$

Theorem (Kac,1947)

For stationary and ergodic processes

$$T(x) = \frac{1}{\Pr\{X_1 = x\}}, \tag{1}$$

for any x with $\Pr\{X_1 = x\} > 0$.

Kac's Result for Sliding Blocks

Let L be a positive integer.
When $\cdots, X_{-1}, X_0, X_1, X_2, \cdots$ is stationary and ergodic, then also

$$\cdots, \begin{pmatrix} X_{-1} \\ X_0 \\ \vdots \\ X_{L-2} \end{pmatrix}, \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{L-1} \end{pmatrix}, \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_L \end{pmatrix}, \begin{pmatrix} X_2 \\ X_3 \\ \vdots \\ X_{L+1} \end{pmatrix}, \cdots$$

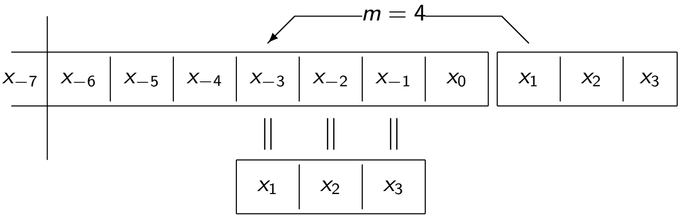
is stationary and ergodic.
Therefore Kac's result holds also for "sliding" L-blocks, hence

$$T((x_1, x_2, \cdots, x_L)) = \frac{1}{\Pr\{(X_1, X_2, \cdots, X_L) = (x_1, x_2, \cdots, x_L)\}},$$

if $\Pr\{(X_1, X_2, \cdots, X_L) = (x_1, x_2, \cdots, x_L)\} > 0$.
Now a waiting time equal to m implies that m is the smallest positive integer such that $(x_{1-m}, x_{2-m}, \cdots, x_{L-m}) = (x_1, x_2, \cdots, x_L)$.

Universal Source Coding Based on Waiting Times

Suppose that our source is *binary* i.e. $X_t \in \{0,1\}$ for all integer t.



An encoder wants to transmit a source block $x_1^L \triangleq (x_1, x_2, \cdots, x_L)$ to a decoder. Both encoder and decoder have access to buffers containing all previous source symbols $\cdots, x_{-2}, x_{-1}, x_0$.
Using these previous source symbols the encoder can determine the waiting time m of x_1^L. It is the smallest integer m that satisfies

$$x_{1-m}^{L-m} = x_1^L,$$

where $x_{1-m}^{L-m} \triangleq (x_{1-m}, x_{2-m}, \cdots, x_{L-m})$.

Universal Source Coding Based on Waiting Times (cont.)

Waiting time m is encoded and sent to the decoder. The code for m consists of a preamble $p(m)$ and an index $i(m)$ and has length $l(m)$. Code table for the waiting time m for $L = 3$:

m	$p(m)$	$i(m)$	$l(m)$
1	00	-	2+0=2
2	01	0	2+1=3
3	01	1	2+1=3
4	10	00	2+2=4
5	10	01	2+2=4
6	10	10	2+2=4
7	10	11	2+2=4
≥ 8	11	copy of $x_1 x_2 x_3$	2+3=5

After decoding m the decoder can reconstruct x_1^L using the previous source symbols

For arbitrary L we get index lengths $0, 1, \cdots, L-1$ and a "copy"-code with length L. We use a preamble $p(m)$ of $\lceil \log_2(L+1) \rceil$ bits to specify one of these $L+1$ alternatives.

Universal Source Coding Based on Waiting Times (cont.)

For arbitrary L we get for the code-block length $l(m)$

$$l(m) = \begin{cases} \lceil \log_2(L+1) \rceil + \lfloor \log_2 m \rfloor & \text{if } m < 2^L, \\ \lceil \log_2(L+1) \rceil + L & \text{if } m \geq 2^L. \end{cases}$$

This results in the upper bound

$$l(m) \leq \lceil \log_2(L+1) \rceil + \log_2 m. \tag{2}$$

After processing the block x_1^L both the encoder and decoder can update their buffers. Then the next block

$$x_{L+1}^{2L} \triangleq x_{L+1}, x_{L+2}, \cdots, x_{2L}$$

is processed in a similar way, etc.

Note: Buffers need only contain the previous $2^L - 1$ source symbols!

Analysis of the Waiting-Time Algorithm

Assume that a certain x_1^L occurred as first block. What is then the average codeword length $L(x_1^L)$ for x_1^L?

$$\begin{aligned}
L(x_1^L) &= \sum_{m=1,2,\cdots} Q_{x_1^L}(m) l(m) \\
&\overset{(a)}{\leq} \sum_{m=1,2,\cdots} Q_{x_1^L}(m) \lceil \log_2(L+1) \rceil + \sum_{m=1,2,\cdots} Q_{x_1^L}(m) \log_2 m \\
&\overset{(b)}{\leq} \lceil \log_2(L+1) \rceil + \log_2 \left(\sum_{m=1,2,\cdots} m Q_{x_1^L}(m) \right) \\
&\overset{(c)}{=} \lceil \log_2(L+1) \rceil + \log_2 \frac{1}{\Pr\{X_1^L = x_1^L\}}.
\end{aligned}$$

Here (a) follows from the bound (2) on $l(m)$, (b) from Jensen's inequality $E[\log_2 M] \leq \log_2 E[M]$ since the log is a convex-$\cap$ function. Furthermore (c) follows from Kac's theorem (1).

Analysis of the Waiting-Time Algorithm (cont.)

The probability that x_1^L occurred as first block is $\Pr\{X_1^L = x_1^L\}$. For the average codeword length $L(X_1^L)$ we therefore get

$$\begin{aligned} L(X_1^L) &= \sum_{x_1^L} \Pr\{X_1^L = x_1^L\} L(x_1^L) \\ &\leq \sum_{x_1^L} \Pr\{X_1^L = x_1^L\} \left(\lceil \log_2 (L+1) \rceil + \log_2 \frac{1}{\Pr\{X_1^L = x_1^L\}} \right) \\ &= \lceil \log_2 (L+1) \rceil + H(X_1^L). \end{aligned}$$

For the rate R_L we now obtain

$$R_L = \frac{L(X_1^L)}{L} \leq \frac{H(X_1^L)}{L} + \frac{\lceil \log_2 (L+1) \rceil}{L}.$$

Achieving Entropy

First note that

$$\lim_{L\to\infty} \frac{H(X_1^L)}{L} \triangleq H_\infty(X)$$

and

$$\lim_{L\to\infty} \frac{\lceil \log_2(L+1) \rceil}{L} = 0.$$

Theorem (W., 1986, 1989)

The Waiting-Time Algorithm achieves entropy since

$$\lim_{L\to\infty} R_L = \lim_{L\to\infty} \left(\frac{H(X_1^L)}{L} + \frac{\lceil \log_2(L+1) \rceil}{L} \right) = H_\infty(X).$$

Note: This algorithm is **universal**. Although the statistics of the source are unknown, entropy is achieved.

Relation Waiting Times and Entropy

Again assume that $\cdots, X_{-1}, X_0, X_1, X_2, \cdots$ is stationary and ergodic with entropy $H_\infty(X)$. Let the random variable M be the waiting time of the source block X_1^L.

Theorem (Wyner & Ziv, 1989)

Fix an $\epsilon > 0$*. Then*

$$\lim_{L \to \infty} \Pr\left\{ M \geq 2^{L(H_\infty(X)+\epsilon)} \right\} = 0.$$

This result was crucial in proving that the LZ77 algorithm achieves entropy (Wyner & Ziv [1994]).