

# Prefix codes for integers

## Jorcad 2008

Dominique Perrin

September 16, 2008

# Outline

- 1 Some prefix codes for integers
  - ▶ Golomb codes
  - ▶ Golomb-Rice codes
  - ▶ Exponential Golomb codes
  - ▶ Reversible codes
- 2 Infinite Huffman trees
  - ▶ Exponential distributions
  - ▶ Ordered trees
- 3 Length distributions

# Variable-length Codes and Finite Automata

Part of a chapter, by Marie-Pierre Béal, Jean Berstel, Brian Marcus, DP, Christophe Reutenauer and Paul Siegel of **Selected Topics in Information Theory and Coding**, Isaac Woungang ed., World Scientific (to appear).

# Coding integers

The binary expansion of positive integers is not uniquely decipherable: 11010 may represent the number 26, or the sequence 6, 2, or the sequence 1, 2, 2.

The **Elias code** of a positive integer is composed of its binary expansion preceded by a number of zeros equal to the length of this expansion minus one. For instance, the Elias code of 26 is 000011010.

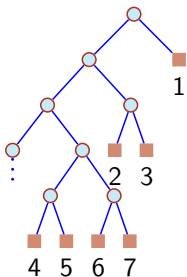


Figure: The Elias code.

# Golomb codes

The **Golomb codes** have the form  $G_m = 1^*0R_m$  where  $R_m$  is the prefix code represented on Figure 2 for  $m \leq 7$ .

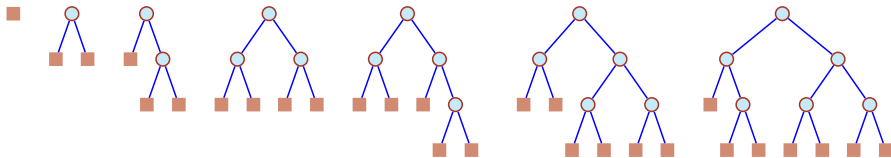


Figure: The sets  $R_1$  to  $R_7$ .

The encoding of the integers is alphabetic. One way to define directly the encoding of an integer is as follows. Set  $r = \lceil \log m \rceil$ . Define the **adjusted** binary representation of an integer  $n < m$  as its representation on  $r - 1$  bits if  $n < 2^r - m$  and on  $r$  bits otherwise (adding 0's on the left if necessary). The encoding of the integer  $n$  in  $G_m$  is formed of  $n/m$  1's followed by 0, followed by the adjusted binary representation of  $n$  modulo  $m$ .

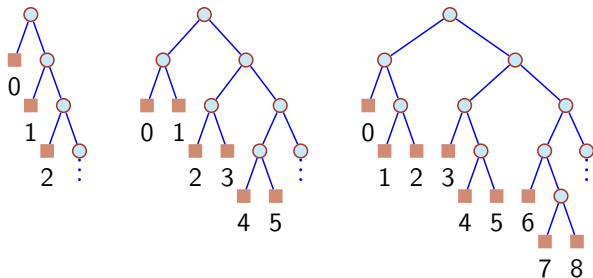


Figure: The Golomb codes of orders 1, 2, 3.

# Golomb-Rice codes

Particular case of the Golomb code for  $m = 2^k$ . Explicit description: The encoding assigns to an integer  $n \geq 0$  the concatenation of two binary words, the **base** and the **offset**. The base is the unary expansion (over the alphabet  $\{1\}$ ) of  $\lfloor n/2^k \rfloor$  followed by a 0. The offset is the remainder of the division written in binary on  $k$  bits. Thus, for  $k = 2$ , the integer  $n = 9$  is coded by 110|01.

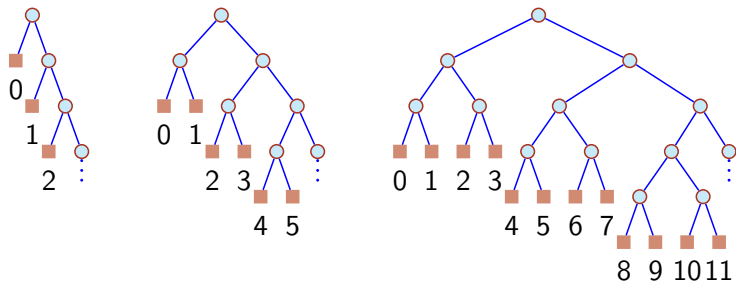


Figure: The codes  $GR_k$  for  $k = 0, 1, 2$ .

# Generating series

Regular expression

$$GR_k = 1^*0\{0, 1\}^k. \quad (1)$$

Generating series of the Golomb–Rice code of order  $k$

$$f_{GR_k}(z) = \frac{2^k z^{k+1}}{1 - z} = \sum_{i \geq k+1} 2^k z^i.$$

For a uniform Bernoulli distribution on the channel symbols, the weighted generating series for the resulting probabilities of the Golomb–Rice codes  $GR_k$  and the corresponding average length  $\lambda_{GR_k}$  are

$$\begin{aligned} p_{GR_k}(z) &= f_{GR_k}(z/2) = \frac{z^{k+1}}{2 - z}, \\ \lambda_{GR_k} &= p'_{GR_k}(1) = k + 2. \end{aligned} \quad (2)$$

For  $p^m = 1/2$ , the series  $p_{GR_k}(z)$ , and thus also the average length  $\lambda_{GR_k}$  happens to be the same for the probability distribution on the code induced by the geometric distribution on the source.

# Exponential Golomb codes

Let  $x$  be the binary expansion of  $1 + \lfloor n/2^k \rfloor$  and let  $i$  be its length. The **base** is made of the unary expansion of  $i - 1$  followed by  $x$  with its initial 1 replaced by a 0. The **offset** is, as before, the binary expansion of the rest of the division of  $n$  by  $2^k$ , written on  $k$  bits. Thus, for  $k = 1$ , the codeword for 9 is 11001|1.

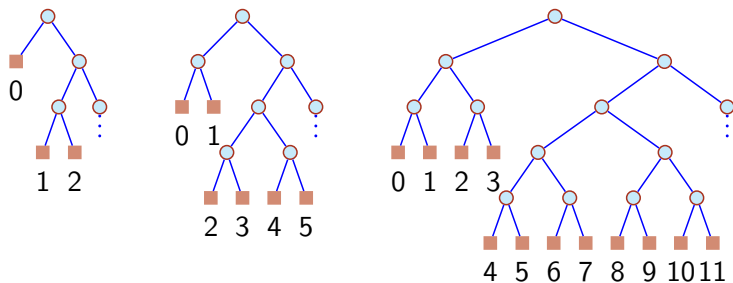


Figure: The exponential Golomb codes of orders 0, 1, 2.

# Generating series

An expression describing the exponential Golomb code of order  $k$  is

$$EG_k = \bigcup_{i \geq 0} 1^i 0 \{0, 1\}^{i+k},$$

and we have the simple relation

$$EG_k = EG_0 \{0, 1\}^k.$$

The generating series of  $EG_k$  is

$$f_{EG_k}(z) = \frac{2^k z^{k+1}}{1 - 2z^2}.$$

The weighted generating series for the probabilities of codewords corresponding to a uniform Bernoulli distribution and the average length are

$$p_{EG_k}(z) = \frac{z^{k+1}}{2 - z^2}$$

$$\lambda_{EG_k} = k + 3.$$

# Bifix codes

Possibility of limiting the consequences of errors occurring in the transmission using a bidirectional decoding scheme as follows.

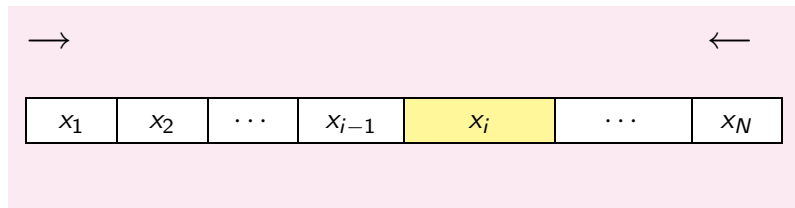


Figure: The transmission of a block of codewords.

In a block of  $N$  encoded source symbols, at most one codeword will be read incorrectly.

# Reversible Golomb-Rice codes

These codes are used for the compression of moving pictures. Indeed, there are reversible codes with the same length distribution as the Golomb-Rice codes. The Advanced Video Coding (AVC) standard recommends the use of these codes instead of the ordinary Golomb-Rice codes to obtain an error resilient coding. The difference with the ordinary codes is that, in the base, the word  $1^i0$  is replaced for  $i \geq 1$  by  $10^{i-1}1$ .

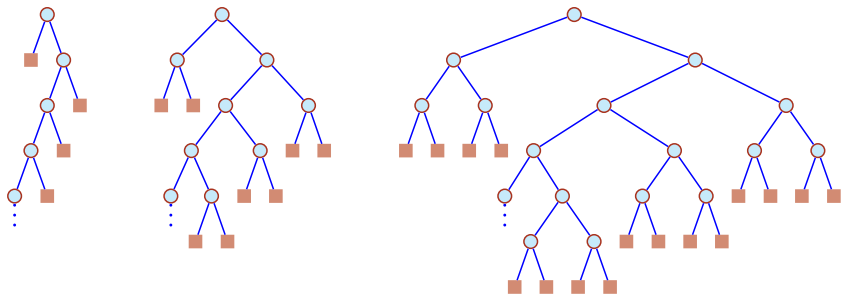


Figure: The reversible Golomb-Rice codes of orders 0, 1, 2.

# Reversible exponential Golomb codes

There is also a reversible version of the exponential Golomb codes, denoted by  $REG_k$ , which are bifix codes with the same length distribution. The code  $REG_0$  is given by

$$REG_0 = 0 + 1\{00, 10\}^*\{0, 1\}1.$$

It is a bifix code because it is equal to its reversal. This comes from the fact that indeed, the set  $\{00, 10\}^*\{0, 1\}$  is equal to its reversal because it is the set of words of odd length which have a 0 at each even position, starting at position 1.

The code of order  $k$  is

$$REG_k = REG_0(0 + 1)^k.$$

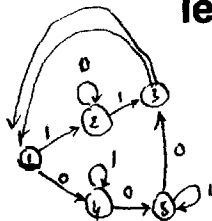


DCT data: The DCT encoded macroblock data consists of four luminance

### Reversible VLCs

When this mode is enabled, a Reversible VLC (RVLC) table shown in

## 8 Appendix B: Transform coefficient length)



$$0 = (1453) (2)$$

$$1 = (123) (4)(5)$$

Table 103: RVLC table for

INDEX	intra			inter		
	LAST	RUN	LEVEL	LAST	RUN	LEVEL
0	0	0	1	0	0	1
1	0	0	2	0	1	1
2	0	1	1	0	0	2
3	0	0	2	0	2	1

# Infinite Huffman trees

Given a probability distribution on the integers, find an optimal prefix code  $X = \{x_0, x_1, \dots\}$ . (i.e. with minimal average length)

$$\lambda(X) = \sum_{i \geq 0} P(i) |x_i|.$$

Theorem (Linder et al. 1997)

*There is a prefix code with finite average length if and only if*

$$H = - \sum_{i \geq 0} P(i) \log P(i) < \infty.$$

# Exponential distribution

Probability distribution on the integers (origin in run length coding). For  $0 < p < 1$  and  $q = 1 - p$ ,

$$\pi(n) = p^n q \quad n \geq 0.$$

Theorem (Gallager, van Vooris, 1975)

*The Golomb code  $G_m$  is optimal for*

$$p^{m-1} + p^m > 1 \geq p^m + p^{m+1}.$$

$p$	$p^2$	$p^3$	$p^4$	$p^5$	$p^6$	$p^7$	$p^8$	$p^9$	$m$
.6	.36								1
.7	.49	.343							2
.8	.64	.51	.41						3
.9	.81	.73	.66	.59	.53	.47	.43	.38	8

# Proof

Set  $G'_m = R_m 1^* 0$ .

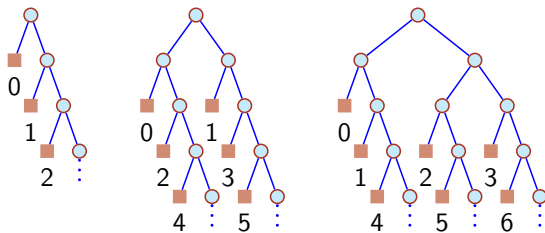


Figure: The modified Golomb codes of orders 1, 2, 3.

Set  $Q = 1 - p^m$ . By the choice of  $m$ , one has  $p^{m-1} > Q > p^{m+1}$ . We consider, for  $k \geq 0$ , the **truncated source**

$$B_k = \{0, \dots, k-1, k, \dots, k+m-1\}.$$

In particular,  $B_0 = \{0, \dots, m-1\}$ . We consider on  $B_k$  the distribution

$$\pi(i) = \begin{cases} p^i q & \text{for } 0 \leq i < k, \\ p^i q / Q & \text{for } k \leq i < k+m-1. \end{cases}$$

One has  $\sum_{i \in B_k} \pi(i) = 1$ .

We have

$$\pi(0) > \pi(1) > \dots > \pi(k-2) > \pi(k-1)$$

and

$$\pi(k) > \pi(k+1) > \dots > \pi(k+m-2) > \pi(k+m-1).$$

Since  $p^{m-1} > Q > p^{m+1}$ ,

$$\begin{aligned}\pi(k+m-2) &= p^{k+m-2}q/Q > p^{k-1}q = \pi(k-1) \\ \pi(k-2) &= p^{k-2}q > p^{k+m-1}q/Q = \pi(k+m-1)\end{aligned}$$

Thus  $k-1$  and  $k+m-1$  form the **pair of minimal weight**. Huffman's algorithm replaces them with a new symbol  $k'$  with weight is

$$\pi(k') = \pi(k-1) + \pi(k+m-1) = p^{k-1}q(1 + p^m/Q) = p^{k-1}q/Q.$$

We obtain  $B_{k-1}$  and we may iterate.

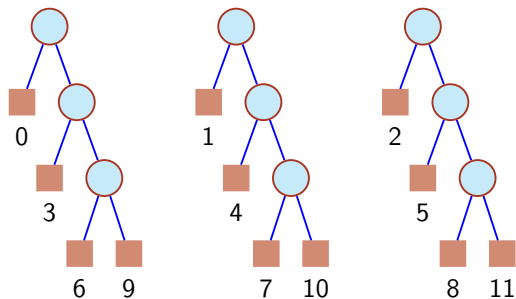


Figure: The result after 9 steps ( $m = 3, k = 9$ ).

# Quasi-uniform sequences

The sequence

$$w_0 \geq w_1 \geq \dots \geq w_{m-2} \geq w_{m-1}$$

is **quasi-uniform** if  $w_{m-2} + w_{m-1} \geq w_0$ . At the end, we are left with  $B_0 = \{0, 1, \dots, m-1\}$  on which Huffman algorithm produces  $R_m$ . Indeed,

## Lemma (Gallager, Van Voochris)

*For any quasi-uniform sequence, Huffman algorithm produces a prefix code with at most two different heights for the leaves.*

Proof: Set  $s = w_{m-2} + w_{m-1}$ . Then  $w_{m-4} + w_{m-3} \geq w_{m-2} + w_{m-1} = s$  and thus  $s, w_0, \dots, w_{m_3}$  is flat.

# Ordered trees

The weight of the father is the sum of the weights of its sons.  
The weights are nonincreasing in military order (Gallager's **sibling property**). For  $p^m < .5$ , the tree  $R_m 1^* 0$  is ordered, otherwise exchange 0, 1.

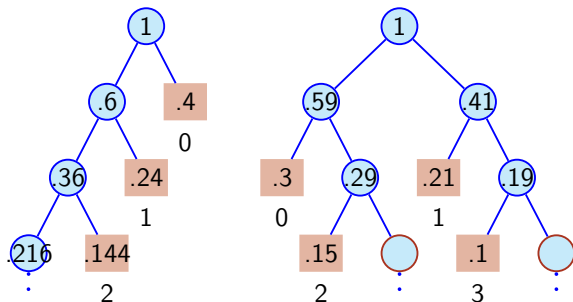


Figure: The cases  $m = 1$  ( $p = .6$ ),  $m = 2$  ( $p = .7$ ).

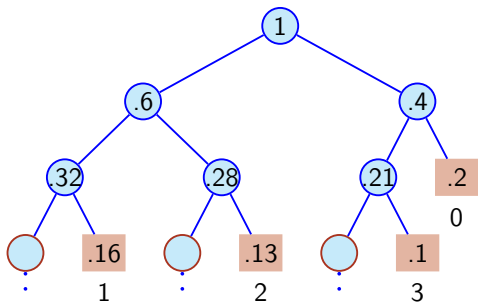


Figure: The case  $m = 2$  ( $p = .7$ ).

For finite trees, any Huffman tree is ordered (up to the order of the siblings). False for infinite trees.

### Theorem

*For infinite trees, one can always find an ordered tree among the optimal ones.*

## Two sided exponential distribution

$$P_{\theta,d}(x) = c_{\theta,d}\theta^{|x+d|} \quad (x \in \mathbb{Z})$$

with  $0 < \theta < 1$  and  $0 \leq d \leq 1/2$  and

$$c_{\theta,d} = \frac{1 - \theta}{\theta^{1-d} + \theta^d}$$

$d = 0$  centered at 0.

$d = 1/4$  OSGD with  $p = \sqrt{\theta}$ .

Characterization of optimal prefix codes in four types indexed by  $\ell \geq 1$  (Merhav, Seroussi, Weinberger, 2000).

# Length distributions

## Theorem (Schutzenberger, 1961)

*The average length*

$$\lambda_X = \sum_{x \in X} |x| 2^{-|x|}$$

*of a thin maximal bifix code  $X$  is an integer.*

Golomb Rice  $\lambda_{GR_k} = k + 2$ . Exponential Golomb  $\lambda_{EG_k} = k + 3$ . There is no reversible version of the other Golomb codes since  $\lambda_{G_m} = \lambda_{R_m} + 2$  and  $\lfloor \log m \rfloor < \lambda_{R_m} < \lceil \log m \rceil$ .

## Problem (Alswhehde et al.)

*Does there always exist a binary bifix code with generating series  $f(z)$  when  $f(1/2) \leq \frac{3}{4}$ ?*

## Length distributions 2

### Theorem (Bassino, Béal, DP)

*There exists a rational binary prefix code with generating series  $f(z)$  if and only if  $f(z)$  is  $\mathbb{N}$ -rational and  $f(1/2) \leq 1$ .*

Golomb codes:  $f_{G_m}(z) = \frac{zf_{R_m}(z)}{1-z}$ .

Exponential Golomb codes:  $f_{EG_k}(z) = \frac{2^k z^{k+1}}{1-2z^2}$ .

### Problem

*When does there exist a rational bifix code with generating series  $f(z)$ ?*