Lossless Source Coding

Geometric distributions and Golomb codes – Part 1

Distributions on the nonnegative integers

- ❑ $\mathbb{N} = \{0, 1, 2, \ldots\}$: the nonnegative integers (natural numbers).
- ❑ Probability mass function $P\colon \mathbb{N} \to [0,1]$, $\sum_{k\geq 0} P(k) = 1$.
- ❑ $X \sim P$ may have finite or infinite entropy

$$H(X) = -\sum_{k=0}^{\infty} P(k) \log P(k)$$

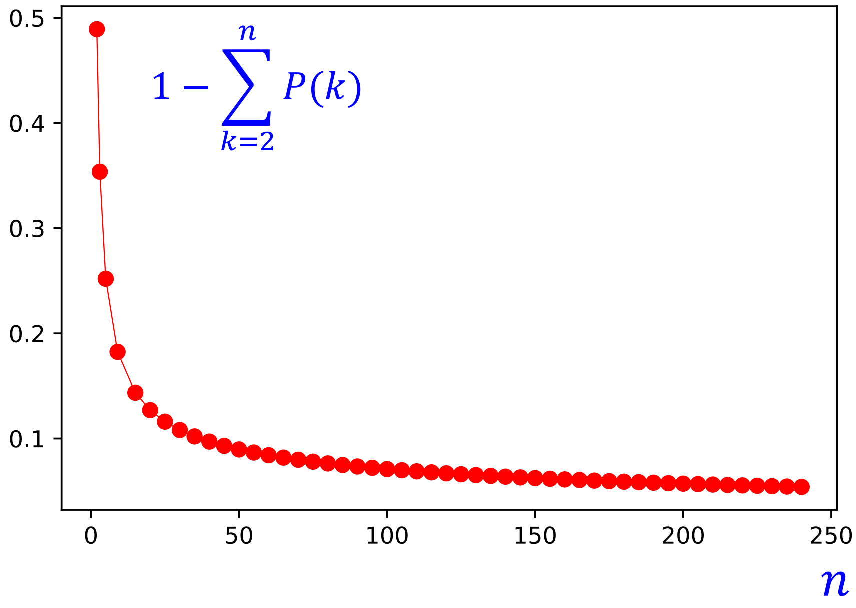
- ❑ Clearly, $\mathbb{N}$ here can be used as proxy for any *countable* alphabet underlying P. We refer to P as a *countable distribution* (or *countable PMF*).

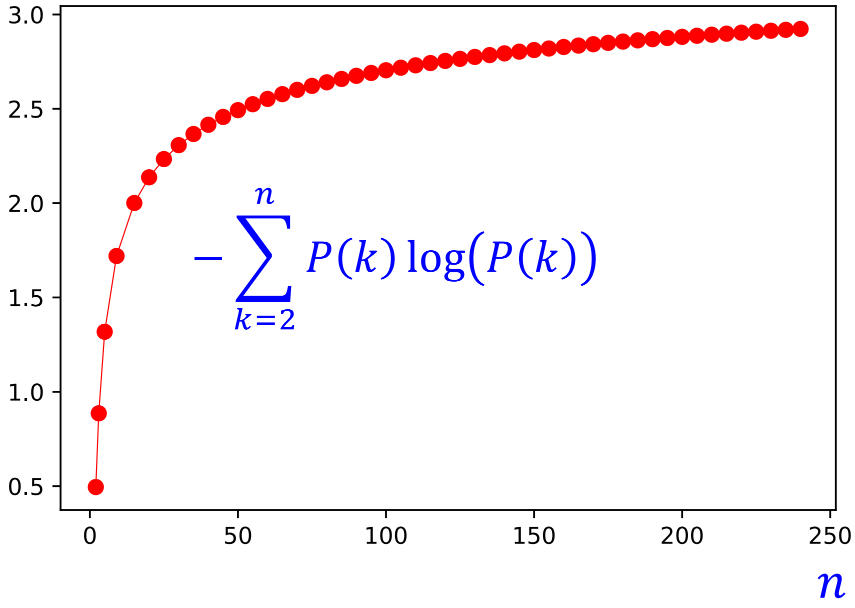
Example: PMF with infinite entropy

$$P(k) = \frac{c}{k \log^2 k}\ , \qquad k \geq 2, \qquad c = \left(\sum_{k=2}^{\infty} \frac{1}{k \log^2 k} \right)^{-1}$$

convergent series

- ❑ We have $H(P) = \infty$
  - ● why: $\sum_{k=2}^{\infty} \frac{1}{k \log k}$ is *divergent*.





Example: PMF with finite entropy (1)

- *Zeta distribution*:

$$P(k) = \frac{1}{\zeta(s)}\frac{1}{k^s}, \quad s > 1,\ k \geq 1, \quad \zeta(s) = \textstyle\sum_{k=1}^{\infty} \frac{1}{k^s}$$

Riemann zeta function

(we'll omit the argument s)

- Writing $s - 1 = 2\epsilon \quad (\epsilon > 0)$,

$$H(x) = -\frac{1}{\zeta}\sum_{k=1}^{\infty}\frac{\log k^{-s} - \log\zeta}{k^s} = \frac{s}{\zeta}\sum_{k=1}^{\infty}\frac{\log k}{k^s} + \log\zeta$$

$$\leq \frac{s}{\zeta}\sum_{k=1}^{K_0-1}\frac{\log k}{k^s} + \frac{s}{\zeta}\sum_{k=K_0}^{\infty}\frac{1}{k^{s-\epsilon}} + \log\zeta < \infty\,.$$

K_0 such that $\log k \leq k^{\epsilon}\ \ \forall k \geq K_0$

finite sum

$s - \epsilon = 1 + \epsilon > 1$

Example: PMF with finite entropy (2)

- *The geometric distribution* $\mathrm{GD}(\gamma)$:

$$P(k) = (1-\gamma)\gamma^k, \qquad \gamma \in (0,1), \qquad k \geq 0$$

- We have $\sum_{k\geq 0} P(k) = 1$ (prove!), and

$$H(x) = -\sum_{k\geq 0}(1-\gamma)\gamma^k\left[\log(1-\gamma) + k\log\gamma\right]$$

$$= -(1-\gamma)\log(1-\gamma)\sum_{k\geq 0}\gamma^k - (1-\gamma)\log\gamma\sum_{k\geq 0}k\gamma^k$$

$$= \frac{-(1-\gamma)\log(1-\gamma) - \gamma\log\gamma}{1-\gamma} = \frac{h_2(\gamma)}{1-\gamma} < \infty.$$

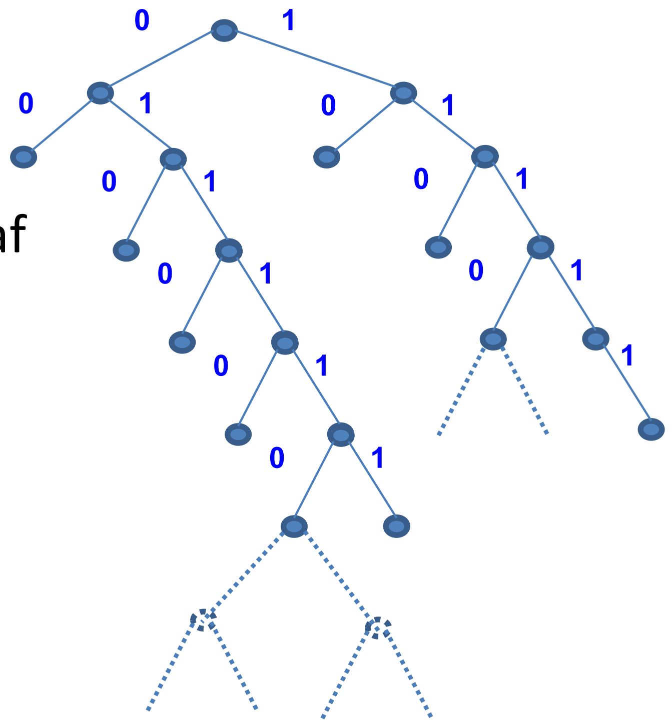
$$h_2(x) = -x\log x - (1-x)\log(1-x)$$

binary entropy

Binary prefix codes for countable distributions

- $\mathcal{C}: \mathbb{N} \rightarrow \{0,1\}^*$, such that $\mathcal{C}(i)$ is not a prefix of $\mathcal{C}(j)$ for any $i \neq j$.
- As in the finite case, a prefix code must satisfy *Kraft's condition:*

$$\sum_{k \geq 0} 2^{-\mathrm{length}(\mathcal{C}(k))} \leq 1.$$



- $\mathcal{C}$ can be represented by an infinite *binary tree*.
- The tree is *complete* if every node that is not a leaf has exactly two children.
  - Differently from the finite case, a complete infinite tree may have a Kraft sum < 1.
- Given a PMF P, the average code length of $\mathcal{C}$ is

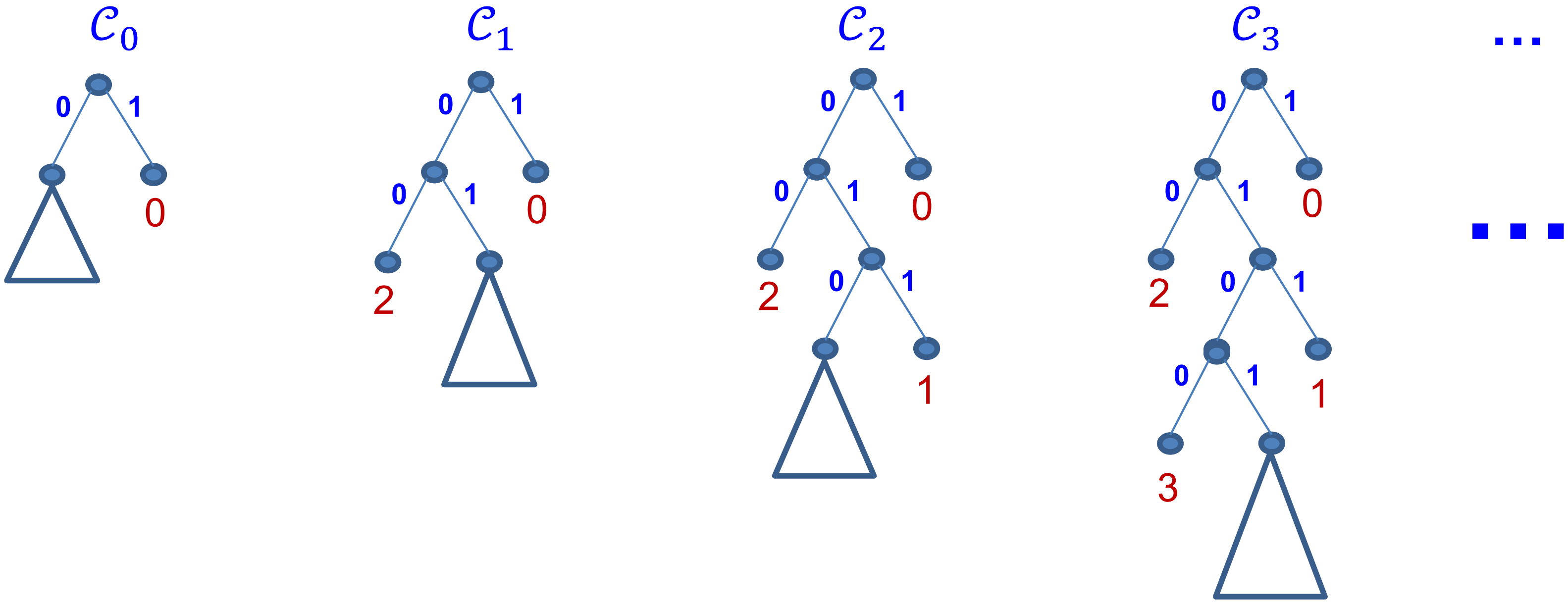
$$L(\mathcal{C}) = \textstyle\sum_{k \geq 0} P(k) \cdot \mathrm{length}(\mathcal{C}(k))$$

  which, again, may be finite or infinite.
- $\mathcal{C}$ is optimal for P if $L(\mathcal{C}) \leq L(\mathcal{C}')$ for any code $\mathcal{C}'$;

  $\Rightarrow$ makes sense only when $L(\mathcal{C}) < \infty$.

Code convergence

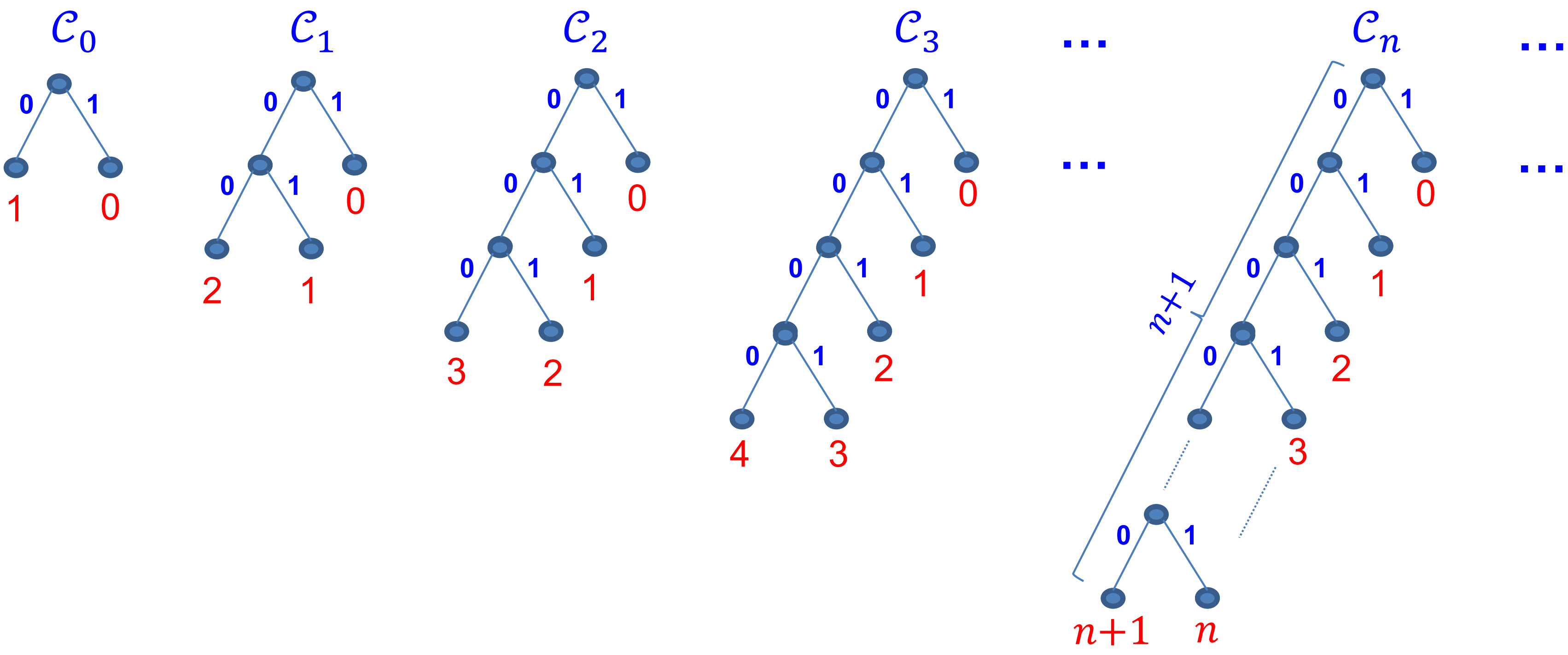
- A sequence of *finite* binary prefix codes $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2, \ldots$ for subsets of $\mathbb{N}$ *converges* to an *infinite* code $\mathcal{C}$ for $\mathbb{N}$ iff
  - for every integer $i \in \mathbb{N}$ there is an index $J_i \geq 0$ such that $\mathcal{C}_j$ assigns a codeword to i for all $j \geq J_i$,
  - for every integer $i \in \mathbb{N}$ there is an index $J'_i \geq J_i$ such $\mathcal{C}_j(i)$ remains *constant*, and equal to $\mathcal{C}(i)$, for all $j \geq J'_i$.



Code convergence: Example

- The *unary* code $C(k) = \overbrace{00\ldots0}^{k}1$ is the limit of the sequence of codes

$$\mathcal{C}_n = \{1, 01, 001, \ldots, 0^n 1,\ 0^n 0\},\ n \geq 0\ .$$



- Say $P(k) = 2^{-(k+1)}$ (geometric distribution $\gamma = \frac{1}{2}$)

  Then, $L(\mathcal{C}) = \sum_{k\geq0}(k+1)2^{-(k+1)} = 2$, and

$$H(X) = -\sum_{k\geq0} P(k)\log P(k) = \sum_{k\geq0} 2^{-(k+1)}\,(k+1) = 2.$$

Questions of interest

- ❑ How does the average code length $L(\mathcal{C})$ relate to the entropy $H(X)$?
- ❑ Are there optimal codes for countable distributions?
- ❑ If so, for what distributions?
- ❑ Can we construct them?
- ❑ Can we describe them compactly?

- ❑ Some answers:
  - Shannon's lower bound applies also to countable distributions, i.e.,
  $$L(\mathcal{C}) \geq H(X).$$
  - Therefore, the code in the previous example is optimal. Clearly, it can be described compactly.
  - How about more general cases? *We cannot use Huffman's procedure!*

Existence of optimal codes

- ❑ $X \sim P$, where P is a countable distribution. The *truncated* random variable $X_n \sim P_n$ has *finite* support $\{0,1,\ldots,n\}$, with $P_n(k) = P(k)/\sum_{j=0}^{n} P(j)$.
- ❑ A *truncated Huffman code* $\mathcal{C}_n^{\mathrm{Huf}}$ for X is a Huffman code for X_n.

Theorem [Linder, Tarokh, Zeger '97], [Kato, Han, Nagoka '96]
Let X be a random variable with countable support, and with finite entropy. Then,

- ● there exists a sequence of binary truncated Huffman codes for X which converges to an optimal code for X,
- ● the sequence of average code lengths of the truncated Huffman codes converges to the minimum possible average code length for X,
- ● any optimal prefix code for X must satisfy the Kraft condition with equality.

- ❑ The proof is not constructive: it does not tell us how to choose or construct the sequence of truncated Huffman codes.
- ❑ *In fact, there are very few classes of countable distributions for which an optimal prefix code can be constructed and described compactly.*
- ❑ We will study such a construction for arbitrary *geometric distributions*.

Why geometric distributions?

Geometric distributions are useful in practice

- Consider random variable $B \sim \text{Bernoulli}(\gamma)$ (i.e., $P(0) = \gamma$). We are interested in describing long sequences of independent realizations of B.
  - We could use an arithmetic coder, but we are interested in a simpler solution.
  - Let b_1^n be the sequence of interest, emitted by B^n. Parse b_1^n as

$$b_1^n = \overbrace{00\ldots0}^{k_1}1\,\overbrace{00\ldots0}^{k_2}1\,\overbrace{00\ldots0}^{k_3}1\ldots\ldots\overbrace{00\ldots0}^{k_N}1$$

We have

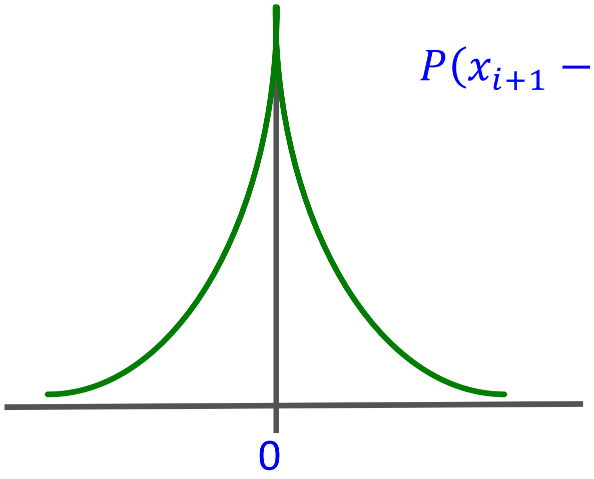
$$P(\overbrace{00\ldots0}^{k}1) = \gamma^k(1-\gamma)$$

⇒ B_1^n can be represented by a sequence of independent random variables distributed as $\text{GD}(\gamma)$.

Why geometric distributions?

Geometric distributions are useful in practice

- ❑ In natural, continuous tone images, differences between contiguous pixels are well modeled by a *two-sided geometric distribution* (discrete Laplacian)



$$P(x_{i+1} - x_i = \Delta) = \frac{1-\gamma}{1+\gamma}\gamma^{|\Delta|}$$

+ we will see that optimal codes for geometric distributions are very easy to implement!

Golomb codes

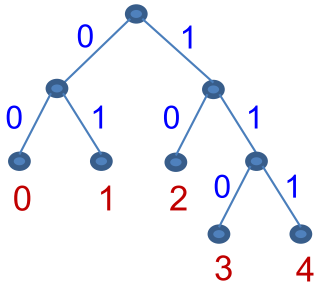
- In 1966, Golomb described a family of prefix-free codes for $\mathbb{N}$ (motivated by sequences of Bernoulli trials).
- Consider an integer $m \geq 1$. The mth order *Golomb code* G_m encodes an integer $i \geq 0$ in two parts, as follows:

concatenation

$$G_m(i) = \text{binary}_m(i \bmod m) \mid \text{unary}(i \operatorname{div} m)$$

C/C++:
`i%m`
`i/m`

- Here,
  - $i \bmod m,\ i \operatorname{div} m =$ remainder and quotient in integer division $\frac{i}{m}$ (resp.)
  - $\text{binary}_m(j) =$ *binary* encoding of j in an optimal code for $\{0, 1, \ldots, m-1\}$ under a uniform distribution ($\lfloor \log m \rfloor$ or $\lceil \log m \rceil$ bits, shorter codes for smaller numbers)
    - Example: $m = 5$, lengths 2 and 3: 0: 00 1: 01 2: 10 3: 110 4: 111



  - $\text{unary}(j) = \overbrace{00\ldots0}^{j}1$ *unary* representation of j.
- Given m and $G_m(i)$, a decoder uniquely reconstructs

$$i = (i \operatorname{div} m) \cdot m + (i \bmod m)$$

Golomb codes – Examples

$m = 5$

i	$G_m(i)$	$\ell(i)$	
0	00 1	3	3
1	01 1	3	
2	10 1	3	
3	110 1	4	5
4	111 1	4	
5	00 01	4	
6	01 01	4	
7	10 01	4	
8	110 01	5	5
9	111 01	5	
10	00 001	5	
11	01 001	5	
12	10 001	5	
13	110 001	6	
14	111 001	6	⋮
⋮	⋮	⋮	

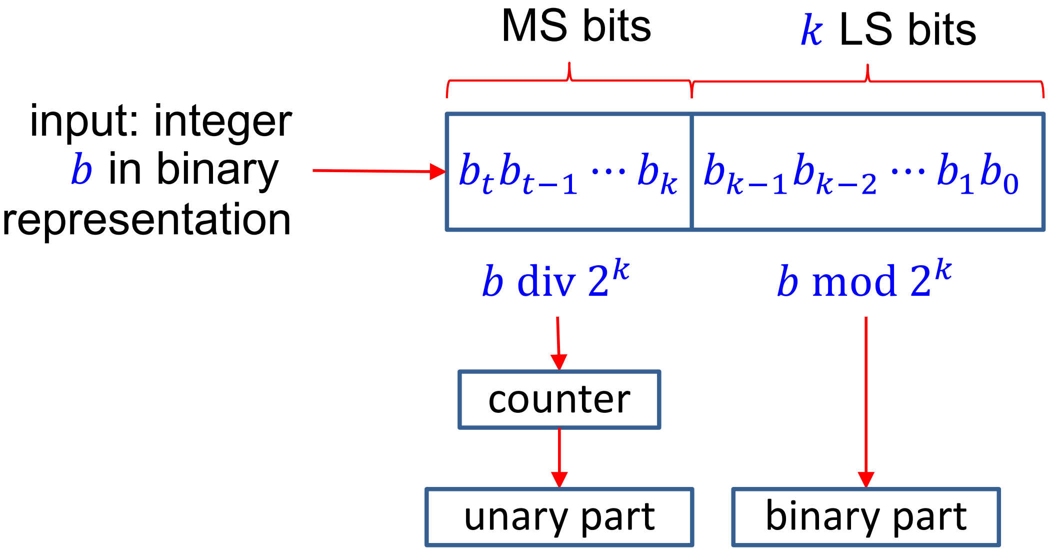
$m = 2^k = 4, \qquad k = 2$

i	i (*binary*)	$G_m(i)$	$\ell(i)$	
0	00	00 1	3	4
1	01	01 1	3	
2	10	10 1	3	
3	11	11 1	3	
4	1 00	00 01	4	4
5	1 01	01 01	4	
6	1 10	10 01	4	
7	1 11	11 01	4	
8	10 00	00 001	5	4
9	10 01	01 001	5	
10	10 10	10 001	5	
11	10 11	11 001	5	
12	11 00	00 0001	6	
13	11 01	01 0001	6	⋮
14	11 10	10 0001	6	
⋮		⋮	⋮	

Golomb PO2 codes

- When $m = 2^k$, we call G_m a *Golomb power of two* (PO2) code and use k as the defining parameter: $G_k^* \triangleq G_{2^k}$.
- PO2 codes are especially simple to implement!

  Example: *Golomb PO2 encoder*



C/C++:

b mod 2^k : `b & ((1<<k)-1)`

b div 2^k : `b >> k`

Optimality of Golomb codes

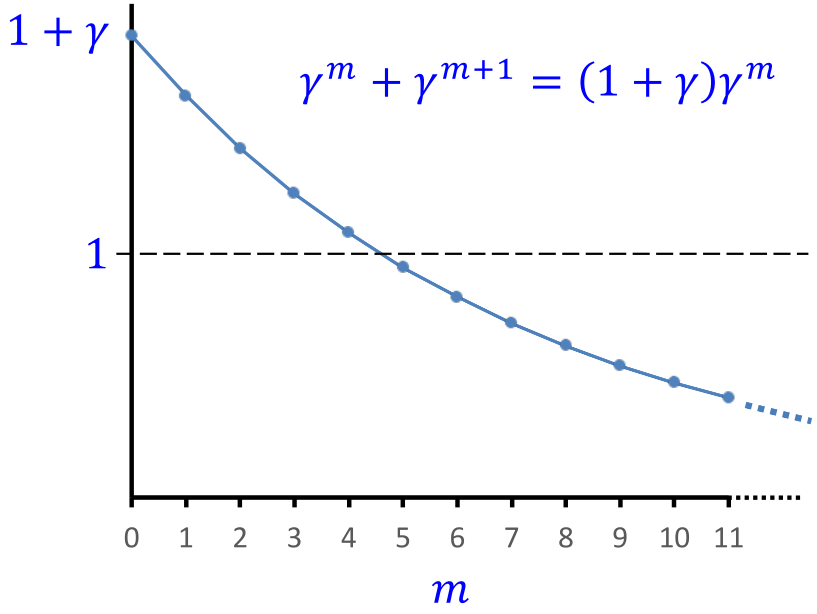
Theorem [Gallager, Van Voorhis 1975]

Let $X \sim \mathrm{GD}(\gamma)$ and let m be the *unique* integer satisfying

$$\gamma^m + \gamma^{m+1} \leq 1 < \gamma^m + \gamma^{m-1}.$$

Then, G_m is an optimal prefix-free code for X.

Why is there a *unique* such value of m ?



Given γ, we have
$m = \min\{\, m' \mid \gamma^{m'} + \gamma^{m'+1} \leq 1 \,\}$.

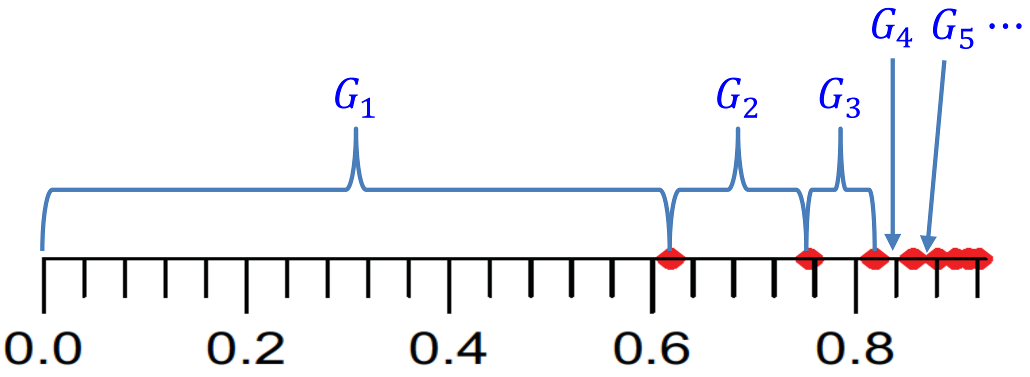
Golomb (1966) had proved optimality for $\gamma = 2^{-\frac{1}{m}}$, i.e., $\gamma^m = \frac{1}{2}$.

Optimality of Golomb codes

What range of γ is G_m optimal for?

Solution of $\gamma^m + \gamma^{m+1} = 1$

m	γ_m
1	0.6180339887
2	0.7548776662
3	0.8191725134
4	0.8566748839
5	0.8812714616
6	0.8986537126
7	0.9115923535
8	0.9215993196



Proof of optimality

Consider γ fixed and m as determined above. Define an *r-reduced source S_r*, for any $r \geq 0$, as a source with $r+1+m$ symbols, with the following probabilities:

$$P_r(i) = \begin{cases} (1-\gamma)\gamma^i, & 0 \leq i \leq r, \\ \dfrac{(1-\gamma)\gamma^i}{1-\gamma^m}, & r+1 \leq i \leq r+m. \end{cases}$$

We have $\sum_{i=0}^{r+m} P_r(i) = 1$. In fact, S_r can be interpreted as defined over an alphabet of regular symbols and "super-symbols",

$$S_r = \left\{ 0, 1, 2, \ldots, r,\ A_1, A_2, \ldots, A_m \right\},$$

where

$$A_j = \left\{ r + j + t \cdot m \;\middle|\; t = 0, 1, 2, \ldots \right\}, \qquad 1 \leq j \leq m\,.$$

Indeed, we have

$$P_r(A_j) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^{r+j+t\cdot m} = \frac{(1-\gamma)\gamma^{r+j}}{1-\gamma^m}, 1 \leq j \leq m.$$

Proof of optimality (cont.)

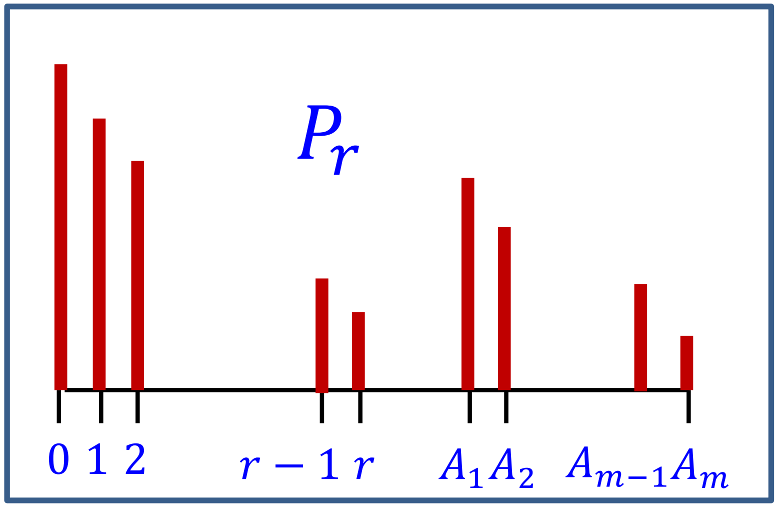
Recall: $\gamma^m + \gamma^{m+1} \leq 1 < \gamma^m + \gamma^{m-1}$ *definition of m* (**)

$S_r = \{ 0,1,2, \dots , r, A_1 , A_2, \dots , A_m\},$

$P_r(i) = (1-\gamma)\gamma^i,\ 0 \leq i \leq r,$

$P_r(A_j) = \frac{(1-\gamma)\gamma^{r+j}}{1-\gamma^m}, 1 \leq j \leq m.$

Consider Huffman coding of S_r.



Claim: The 2 symbols with lowest probability in S_r are r, A_m.

Proof: It suffices to prove

$P_r(r) < P_r(A_{m-1}), \quad P_r(A_m) \leq P_r(r-1).$

$(1-\gamma)\gamma^r < \frac{(1-\gamma)\gamma^{r+m-1}}{1-\gamma^m} \Leftrightarrow 1 < \frac{\gamma^{m-1}}{1-\gamma^m} \Leftrightarrow 1-\gamma^m < \gamma^{m-1}$ RHS of (**).

Similarly, $P_r(A_m) \leq P_r(r-1)$ is implied by the LHS of (**).

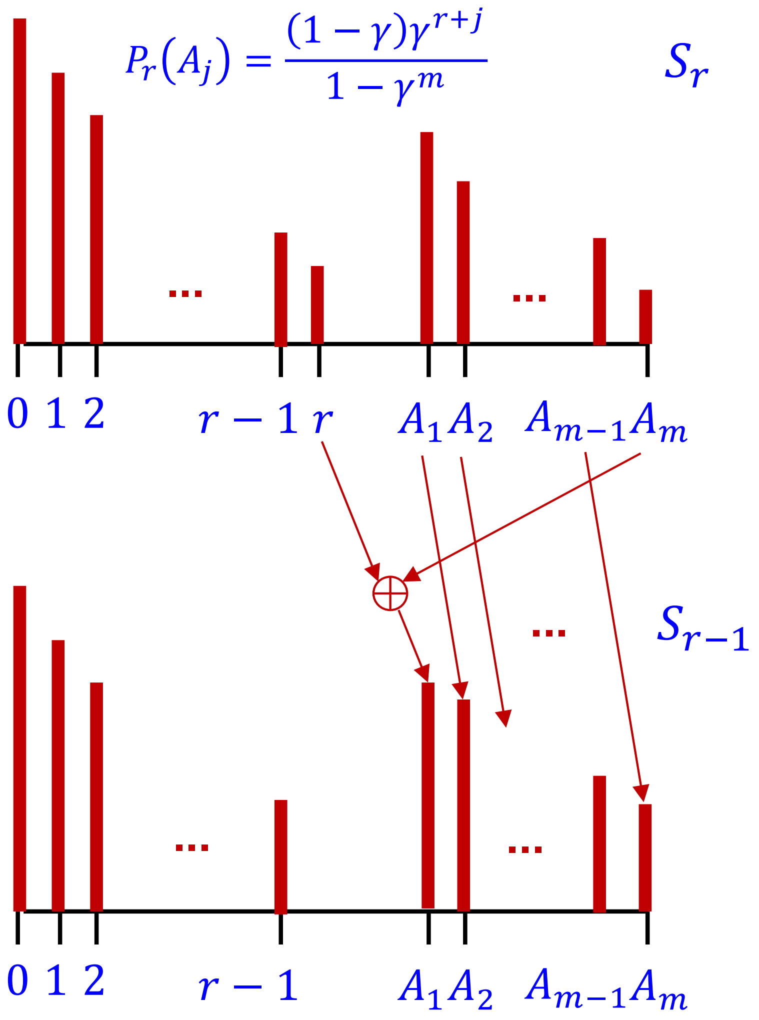
Proof of optimality (cont.)

- The 2 symbols with lowest probability are r, A_m

  ⇒ first step of Huffman procedure merges r, A_m, resulting in a probability

$$(1-\gamma)\gamma^r + \frac{(1-\gamma)\gamma^{r+m}}{1-\gamma^m} = \frac{(1-\gamma)\gamma^r}{1-\gamma^m}$$

= prob. of symbol A_1 in S_{r-1} !



$$P_r(A_j) = \frac{(1-\gamma)\gamma^{r+j}}{1-\gamma^m}$$

- Also, A_1 in S_r is A_2 in S_{r-1},

  A_2 in S_r is A_3 in S_{r-1}, … , etc.

- *⇒ Huffman step transforms S_r into S_{r-1}.*

  Continue until we obtain S_{-1} with

$$P_{-1}(A_i) = \frac{(1-\gamma)\gamma^{i-1}}{1-\gamma^m},\ 1 \le i \le m.$$

Proof of optimality (cont.)

We obtain $S_{-1} = \{ A_1, A_2, \dots, A_m\}$ with

$$P_{-1}(A_i) = \frac{(1-\gamma)\gamma^{i-1}}{1-\gamma^m}, \qquad 1 \le i \le m.$$

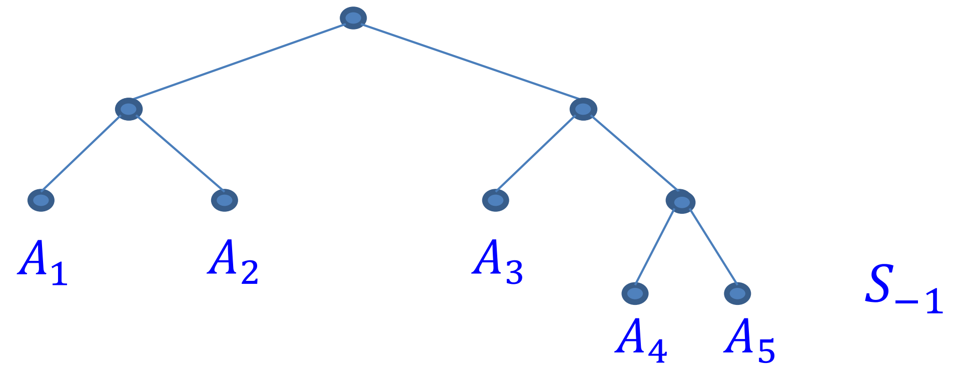
We have

$$P_{-1}(A_1) < P_{-1}(A_{m-1}) + P_{-1}(A_m) \quad \text{from } (**)$$

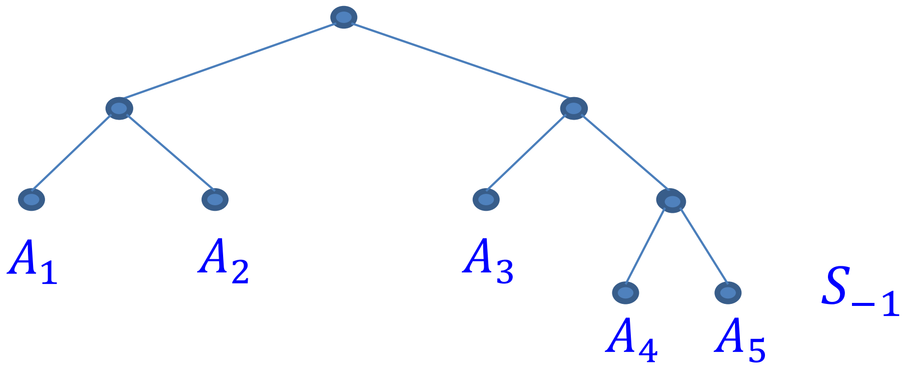
$\Rightarrow$ S_{-1} is a *quasi-uniform* source with m symbols. An optimal code for such a source has $2^{\lceil \log m \rceil} - m$ words of length $\lfloor \log m \rfloor$ and $2m - 2^{\lceil \log m \rceil}$ words of length $\lceil \log m \rceil$

(shortest codewords assigned to highest probability symbols).

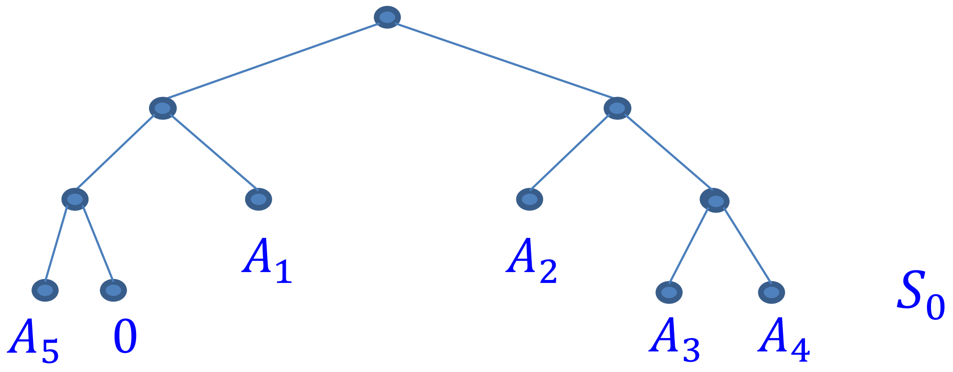
Example: $m = 5$



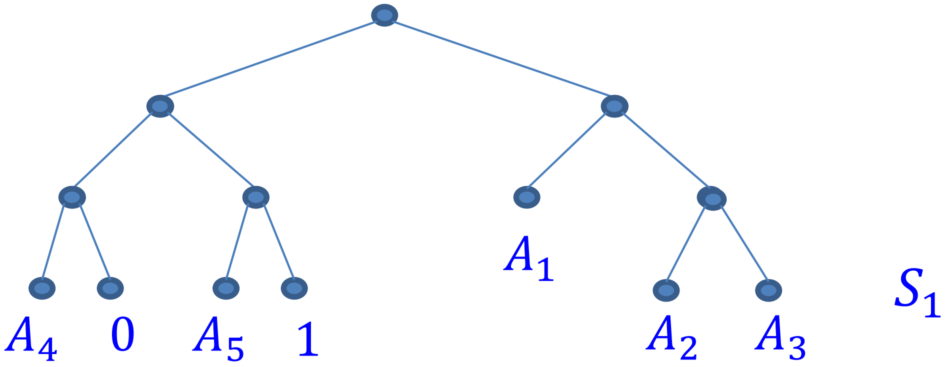
Unfolding reduced sources



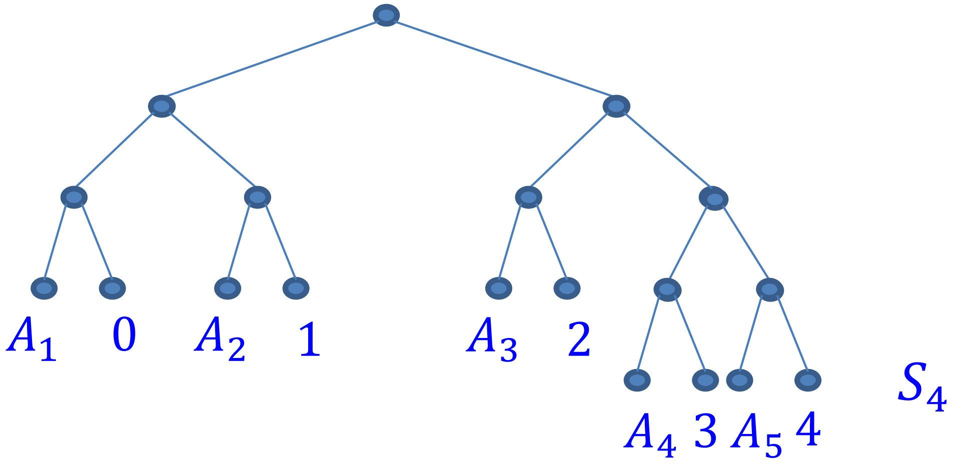
Unfolding reduced sources



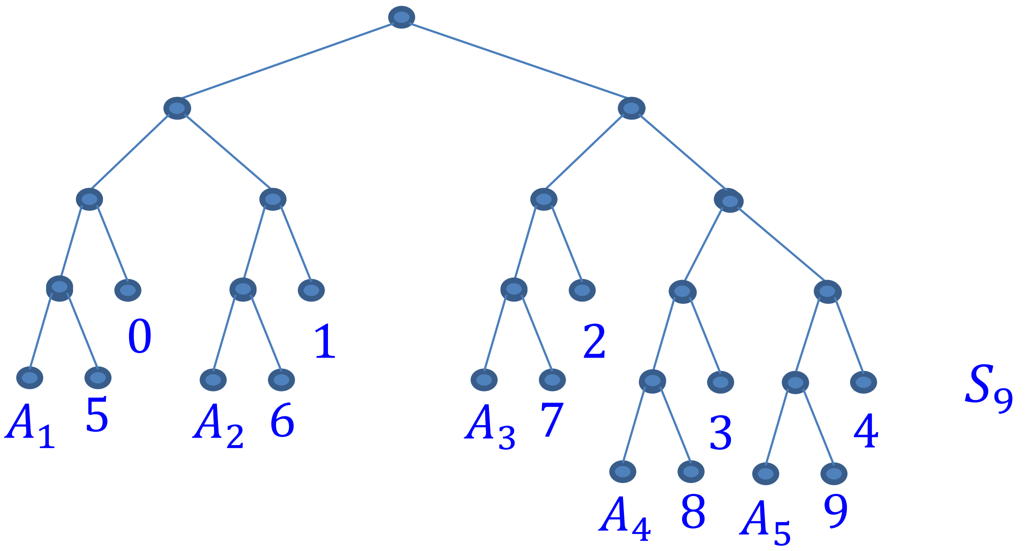
Unfolding reduced sources



Unfolding reduced sources

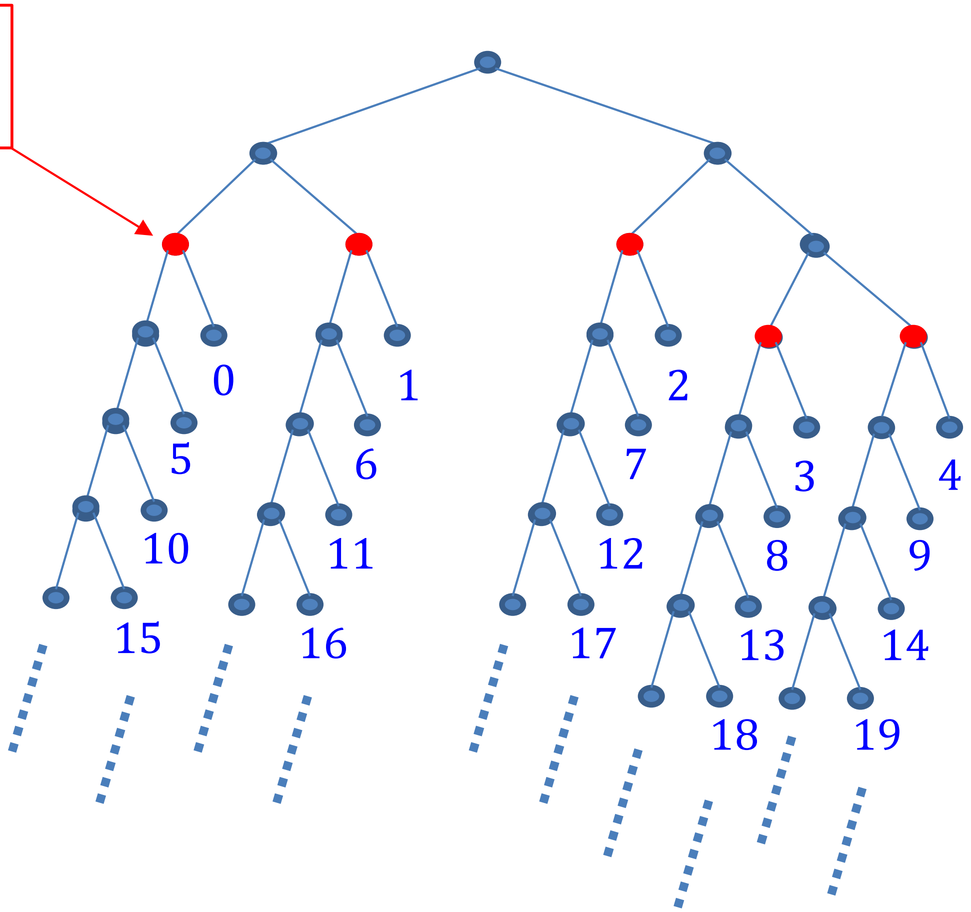


Unfolding reduced sources



Unfolding reduced sources

● leaves of the binary$_m$ tree



$\mathrm{GD}(\gamma)$

From each leaf of the binary$_m$ tree we "hang" a unary tree: equivalent to concatenating the two codes!

Proof of optimality (cont.)

- We have proved that the sequence of optimal codes $\mathcal{C}_{-1}, \mathcal{C}_0, \mathcal{C}_1, \ldots$ for the reduced sources $S_{-1}, S_0, S_1, \ldots$ converges to the Golomb code G_m for m satisfying $(**)$.
- Why is the code *optimal* for $\mathrm{GD}(\gamma)$? (intuition is obvious, but …)

  $\bar{L} = \inf \bar{L}(\mathcal{C})$ over all uniquely decipherable codes $\mathcal{C}$ for $\mathrm{GD}(\gamma)$.

  $\bar{L}_G =$ expected code length for G_m

  $\bar{L}_r =$ expected code length for $\mathcal{C}_r$ on S_r

  - Clearly, we have $\bar{L} \le \bar{L}_G$
  - Also, $\bar{L}_r \le \bar{L}$ because we can use a subset of the codewords of $\mathcal{C}$ for S_r, taking the original codeword from $\mathcal{C}$ for $0, 1, \ldots, r$, and the codeword $\mathcal{C}$ assigns to $r + j$ for A_j .

$$\begin{aligned} \bar{L} &= \textstyle\sum_{i=0}^{r} P(i)|\mathcal{C}(i)| + \sum_{j=1}^{m} \sum_{i\in A_j} P(i)|\mathcal{C}(i)| \\ &> \textstyle\sum_{i=0}^{r} P(i)|\mathcal{C}(i)| + \sum_{j=1}^{m} \sum_{i\in A_j} P(i)|\mathcal{C}(r+j)| \\ &= \textstyle\sum_{i=0}^{r} P(i)|\mathcal{C}(i)| + \sum_{j=1}^{m} P\big(A_j\big)|\mathcal{C}(r+j)| = \bar{L}_r \end{aligned}$$

$r + j$ has shortest codeword in A_j

  - For similar reasons, $\bar{L}_r$ is increasing with r, and it has a limit as $r \to \infty$, so $\lim\limits_{r\to\infty} \bar{L}_r \le \bar{L}$. But $\lim\limits_{r\to\infty} \bar{L}_r = \bar{L}_G$, so $\bar{L}_G \le \bar{L}$.

Expected code length

- Short calculation shows that

$$\bar{L}_G = \lfloor \log m \rfloor + 1 + \frac{\gamma^t}{1-\gamma^m} \qquad (t = 2^{\lfloor \log m \rfloor + 1} - m)$$

(The G-vV paper has an error in this formula—$\lceil\ \rceil$ instead of $\lfloor\ \rfloor$)

  - This holds for any γ and m , not necessarily optimal.