Chapter 4
Zero-Error Data Compression

Department of Information Engineering
The Chinese University of Hong Kong

4.1 The Entropy Bound

Definition 4.1 A D-ary source code $\mathcal{C}$ for a source random variable X is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$, the set of all finite length sequences of symbols taken from a D-ary code alphabet.

Definition 4.2 A code $\mathcal{C}$ is uniquely decodable if for any finite source sequence, the sequence of code symbols corresponding to this source sequence is different from the sequence of code symbols corresponding to any other (finite) source sequence.

Example 4.3 Let $\mathcal{X} = \{A, B, C, D\}$. Consider the code $\mathcal{C}$ defined by

x	$\mathcal{C}(x)$
A	0
B	1
C	01
D	10

$$\begin{array}{lcl} AAD & \rightarrow & 0010 \\ ACA & \rightarrow & 0010 \\ AABA & \rightarrow & 0010 \end{array}$$

Therefore, $\mathcal{C}$ not uniquely decodable.

Theorem 4.4 (Kraft Inequality) Let $\mathcal{C}$ be a D-ary source code, and let $l_1, l_2, \cdots, l_m$ be the lengths of the codewords. If $\mathcal{C}$ is uniquely decodable, then

$$\sum_{k=1}^{m} D^{-l_k} \leq 1.$$

Expected Length

- Source random variable $X \sim \{p_1, p_2, \cdots, p_m\}$
- Expected length of $\mathcal{C}$:

$$L = \sum_i p_i l_i$$

- Intuitively,

$$H_D(X) \leq L$$

because each D-ary symbol can carry at most 1 D-it of information.

Theorem 4.6 (Entropy Bound) Let $\mathcal{C}$ be a D-ary uniquely decodable code for a source random variable X with entropy $H_D(X)$. Then the expected length of $\mathcal{C}$ is lower bounded by $H_D(X)$, i.e.,

$$L \geq H_D(X).$$

This lower bound is tight if and only if $l_i = -\log_D p_i$ for all i.

Corollary 4.7 $H(X) \leq \log |\mathcal{X}|$.

Definition 4.8 The redundancy R of a D-ary uniquely decodable code is the difference between the expected length of the code and the entropy of the source.

By the entropy bound, $R \geq 0$.

4.2 Prefix Codes

Definition 4.9 A code is called a prefix-free code if no codeword is a prefix of any other codeword. For brevity, a prefix-free code will be referred to as a prefix code.

Example 4.10

x	$\mathcal{C}'(x)$
A	0
B	10
C	110
D	1111

Code Tree for Prefix Code

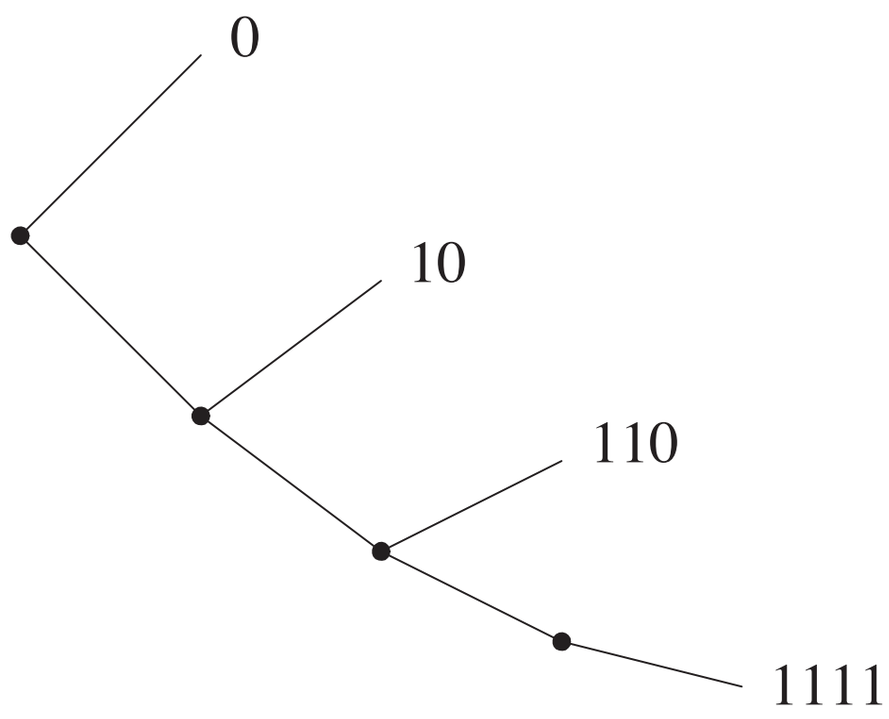
- A D-ary tree is a graphical representation of a collection of finite sequences of D-ary symbols.
- A node is either an *internal node* or a *leaf*.
- The tree representation of a prefix code is called a *code tree*.

Instantaneous Decoding

$$BCDAC\cdots \rightarrow 10110111110110\cdots$$

Decode by tracing the code tree from the root:

$$10110111110110\cdots \rightarrow 10, 110, 1111, 0, 110, \cdots$$



Theorem 4.11 There exists a D-ary prefix code with codeword lengths l_1, $l_2, \cdots, l_m$ if and only if the Kraft inequality

$$\sum_{k=1}^{m} D^{-l_k} \leq 1$$

is satisfied.

Proof Direct part follows because a prefix code is uniquely decodable and hence satisfies Kraft's inequality.

D-adic Distributions

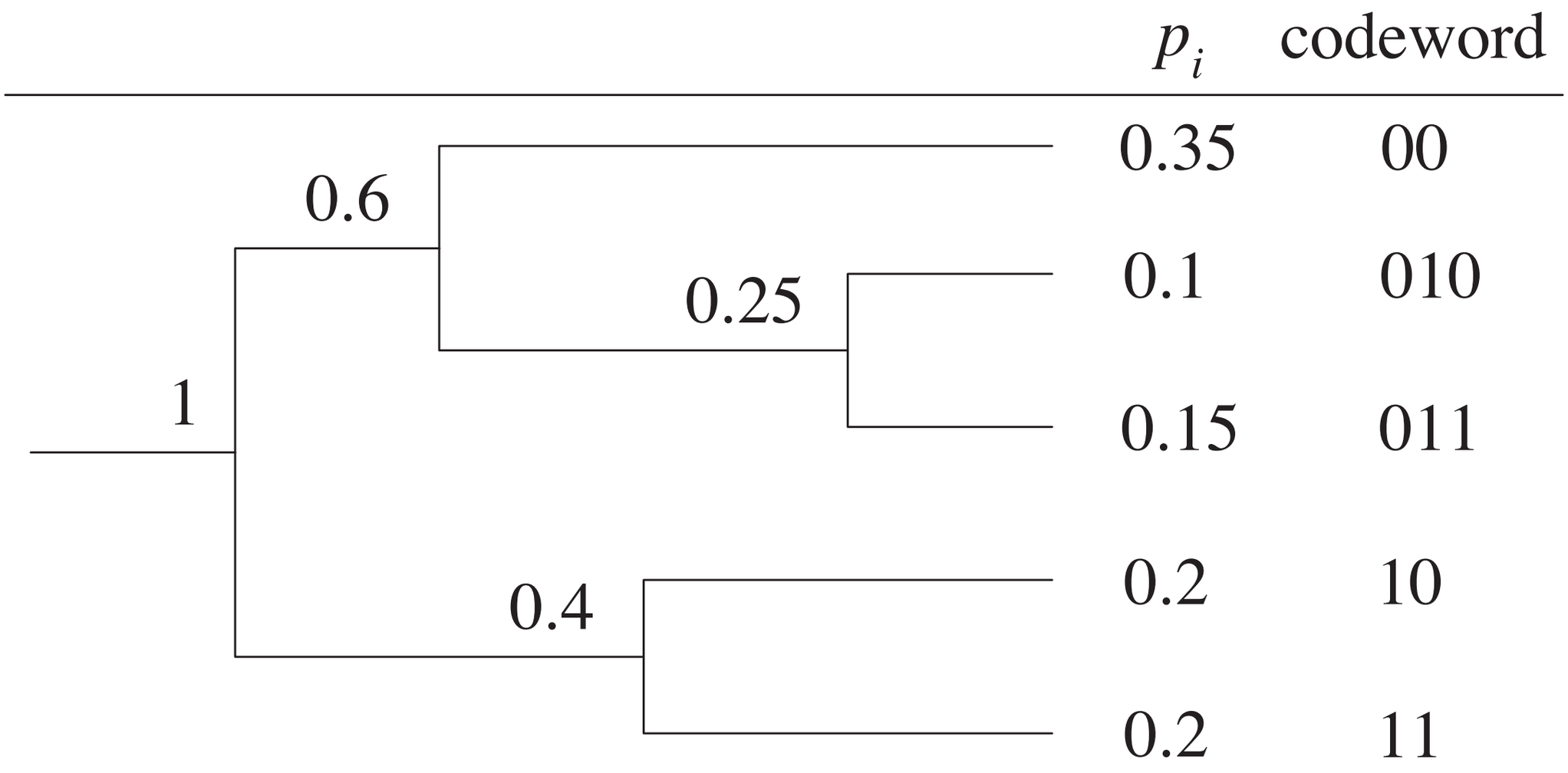
- $p_i = D^{-t_i}$ for all i, where t_i is integer
- *dyadic* when $D = 2$

Corollary 4.12 There exists a D-ary prefix code which achieves the entropy bound for a distribution $\{p_i\}$ if and only if $\{p_i\}$ is D-adic.

Huffman Codes

- A simple construction of optimal prefix codes.
- (Binary Case) Keep merging the two smallest probability masses until one probability mass (i.e., 1) is left.
- (D-ary Case) Insert zero probability masses until there are $D + k(D-1)$ masses.
- In general there can be more than one Huffman code.

Huffman Procedure



Optimality of Huffman Codes

- Assume $p_1 \geq p_2 \geq \cdots \geq p_m$.
- Denote the codeword assigned to p_i by c_i, and denote its length by l_i.

Lemma 4.15 In an optimal code, shorter codewords are assigned to larger probabilities.

Lemma 4.16 There exists an optimal code in which the codewords assigned to the two smallest probabilities are siblings, i.e., the two codewords have the same length and they differ only in the last symbol.

Lemma 4.17 The Huffman procedure produces an optimal prefix code.

Proof Reduce the problem until its size is equal to 2, which we know how to solve.

Upper Bound on L_{Huff}

Theorem 4.18 The expected length of a Huffman code, denoted by L_{Huff}, satisfies

$$L_{\text{Huff}} < H_D(X) + 1.$$

This bound is the tightest among all the upper bounds on L_{Huff} which depend only on the source entropy.

Proof

- Construct a code with codeword lengths $l_i = \lceil -\log_D p_i \rceil$ by showing that the Kraft inequality is satisfied.
- Show that $L = \sum_i p_i l_i < H(X) + 1$.
- Then $L_{\text{Huff}} \leq L < H(X) + 1$.
- For tightness, consider $P_k = \left\{1 - \frac{D-1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right\}$ and let $k \to \infty$.

Asymptotic Achievability of H(X)

- $$H(X) \leq L_{\text{Huff}} < H(X) + 1.$$
- Use a Huffman code to encode $X_1, X_2, \cdots, X_n$, n i.i.d. copies of X. Then

$$nH(X) \leq L^n_{\text{Huff}} < nH(X) + 1.$$

- Divide by n to obtain

$$H(X) \leq \frac{1}{n} L^n_{\text{Huff}} < H(X) + \frac{1}{n} \to H(X) \text{ as } n \to \infty$$