1. Source Encoding

The problem of source encoding is to translate a message into a suitable alphabet with no ambiguity and with efficiency.

We suppose that the initial messages are sequences of symbols from a set X, with $|X| = n$ say, and that the coding alphabet is A, with $|A| = q$. We denote as A^* the set of finite, non-empty, words over A; the length of a word w is denoted as $|w|$ or $l(w)$. A code is then a function $c : X \to A^*$. The image of c is the set of **codewords**. The code is naturally extended to X^* by concatenation:

$$c(x_1x_2 \cdots x_r) = c(x_1)c(x_2) \cdots c(x_r).$$

Definition 1. a) *A code is **uniquely decipherable** if the extension of c to X^* is injective (different sequences of symbols are encoded to different words).*

b) *A code is **instantaneously decipherable** if and only if no codeword is a prefix of another codeword.*

Exercise 2. *Let $X = \{N, S, E, W\}$. Classify the following codes with respect to these properties:*

1. $c(N) = 0$, $c(S) = 01$, $c(E) = 11$, $c(W) = 10$;
2. $c(N) = 0$, $c(S) = 01$, $c(E) = 011$, $c(W) = 0111$;
3. $c(N) = 00$, $c(S) = 01$, $c(E) = 11$, $c(W) = 10$.

An instantaneously decipherable code is called a **prefix code**.

1.1. **Prefix Codes and Rooted Trees.** Recall that a **rooted tree** is a tree (a connected simple graph with no cycles) with a vertex identified as the root. Let $l(v)$ denote the distance from the vertex v to the root, also called the length of the vertex; if v' is adjacent to v and $l(v') = l(v)+1$, v' is called a (direct) descendant of v. Vertices with no descendants, which are (with the possible exception of the root) the vertices with degree 1, are called **leaves**. The remaining vertices (including the root) are called **internal** vertices.
We say that T is a q-tree if each internal vertex has at most q descendants, and denote by $T(n, q)$ the set of rooted q-trees with n leaves.

Proposition 3. *There is a bijection between $T(n, q)$ and the set of prefix codes $c : X \to A^*$.*

Proof. (**HW**) □

Remark 4. *Rooted trees may also be identified as decision trees. For example, the problem of determining an integer x from $\{1, \cdots, m\}$ by a sequence of tests of the form - is $x < t$? - may be represented by a tree $T \in T(m, 2)$. The problem of identifying a counterfeit coin (which may be lighter or heavier) from a set of m coins comparing the weights of subsets of coins gives rise to a tree $T \in T(2m, 3)$, as for each coin there are 2 possible positive answers (lighter or heavier) and each decision (weighting) has 3 possible outcomes.*

Exercise 5. *How many tests of the form - is $x < t$? - do we need to find the right answer in the first problem?*

Exercise 6. *Define a strategy to solve the second problem with a minimum number of weightings. How many weightings are needed?*

In both applications of rooted trees, two parameters are of interest: $L(T) = \max\{l(v) : v \text{ a leaf of } T\}$; and $\bar{L}(T) = \frac{1}{n}\sum l(v)$ where the sum is over the set of leaves.
This generalizes, for any probability distribution on the leaves of T, as $\bar{L}(T) = \sum p(v)l(v)$.

Proposition 7. $L(T) \geq \lceil \log_q(n) \rceil$.

Exercise 8. *Prove the proposition by induction on $L = L(T)$.*

Definition 9. *A q-tree is called complete if each internal vertex has exactly q descendants.*

We mention, for future use, a property of complete trees:

Lemma 10. *If $T \in T(n, q)$ is complete, then $(q-1) \mid (n-1)$.*

Exercise 11. *Prove the lemma: start with the case where all leaves have the same length t; in the general case, suppose that T has n_i leaves with length $0 \leq i \leq t$ and append at each leaf a q tree in order to obtain a complete q tree such that all leaves have the same length t.*

The efficiency of a code is determined by the average length of the codewords used in the message. The prefix code defined by

$$c(N) = 00, \qquad c(S) = 01, \qquad c(E) = 11, \qquad c(W) = 10$$

uses 2 bits (binary digits, the length unit) by symbol. If all four elements of X occur with the same probability, this is clearly best possible. But a more realistic assumption is that symbols from the original message occur with different probabilities, as happens for instance in any human language with the symbols of the corresponding alphabet.

Exercise 12. *Suppose X has the probability distribution*

$$p(N) = 0.6, \qquad p(S) = 0.1, \qquad p(E) = 0.2, \qquad p(W) = 0.1;$$

Find a prefix code which is more efficient than the one above.
Is it optimal in terms of efficiency?