Codes I

Prof. Dan A. Simovici

UMB

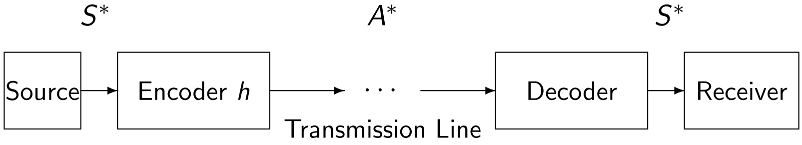
Definition

An *information source* (in short, a *source*) is a pair $\mathcal{S} = (S, D)$, where $S = \{s_0, s_1, \ldots\}$ is a nonempty, countable set referred to as the *source set*, and D is is a probability distribution

$$D = \begin{pmatrix} s_0 & s_1 & \cdots \\ p_0 & p_1 & \cdots \end{pmatrix}$$

where $\sum_{i \in \mathbb{N}} p_i = 1$.
If S is a finite set, then we refer to $\mathcal{S} = (S, D)$ as a *finite source*.

The symbols generated by the source are encoded as words over an alphabet A, which is, of course, finite, using a morphism $h : S^* \longrightarrow A^*$ referred to as the *encoding morphism*. The encoding of a word $s_0 \cdots s_{m-1}$ generated by the source, $h(s_0) \cdots h(s_{m-1}) \in A^*$, is sent through a communication line to a decoder that converts the word $h(s_0) \cdots h(s_{m-1})$ back to a word over the set S.



Different words produced by the source must yield distinct coded messages. This amounts to requiring that h be an injective morphism between S^* and A^*.

Definition

Let A be an alphabet and let $\mathcal{S} = (S, D)$ be a source. A *code* on an alphabet A is a triple $C = (\mathcal{S}, A, h)$, where $h : S^* \longrightarrow A^*$ is an injective morphism.

The *code set* of C is the set of images of symbols of S under the morphism h,

$$h(S) = \{h(s) \mid s \in S\}.$$

Often, when the source and the alphabet are clear from context we will use the term *code* to refer to either h or the code set $h(S)$.

Example

Let S be a finite source set, A be an alphabet such that $|A| \geq 2$, and $k \in \mathbb{N}$ be a number such that $|S| \leq |A|^k$. Any injective mapping $h : S \longrightarrow A^*$ such that $h(s)$ is a word of length k can be extended to an injective morphism from S^* to A^*. Codes constructed in this manner are known as *block codes of length k*.
For instance, let $S = \{s_0, s_1, s_2\}$ and let $A = \{0, 1\}$. By choosing $k = 2$, we can define a block code of length 2 by $h(s_0) = 00$, $h(s_1) = 01$, and $h(s_2) = 10$.

If we do not require that $|h(s)| = k$ for each $s \in S$, then even if $h : S \longrightarrow A^*$ is an injective mapping, its extension $h : S^* \longrightarrow A^*$ is not necessarily an injective morphism as shown in the next example.

Example

Let $S = \{s_0, s_1, s_2\}$, $A = \{0, 1\}$, and let $h : S \longrightarrow A^*$ be the injective mapping $h(s_0) = 0$, $h(s_1) = 01$, and $h(s_2) = 10$. Observe that the extension $h : S^* \longrightarrow A^*$ is not injective because $h(s_1 s_0) = h(s_0 s_2) = 010$.

Definition

Let A be an alphabet, and let $L = \{x_0, x_1, \ldots\}$ be a language on A, $L \neq \emptyset$. L is *uniquely decipherable* if the equality

$$x_{i_0} \cdots x_{i_{m-1}} = x_{j_0} \cdots x_{j_{n-1}}$$

implies $m = n$ and $x_{i_\ell} = x_{j_\ell}$, for $0 \leq \ell \leq n-1$.

If L is a code set, then $\lambda \notin L$. Indeed, if $\lambda \in L$, then we would have $x = \lambda x$ for every $x \in A^*$, which contradicts the uniquely decipherability property.

Theorem

A language $L \subseteq A^*$ is uniquely decipherable if and only if it is code set.

Proof

Suppose that $L = \{x_0, \ldots, x_{k-1}, \ldots\}$ is a uniquely decipherable language. Let S be a source set such that S has the same cardinality as L. There exists a bijection $h : S \longrightarrow L$ such that $h(s_i) = x_i$ for every $x_i \in L$. Suppose that $h(s_{i_0} \ldots s_{i_{m-1}}) = h(s_{j_0} \ldots s_{j_{n-1}})$. This is equivalent to $x_{i_0} \cdots x_{i_{m-1}} = x_{j_0} \cdots x_{j_{n-1}}$, so $m = n$ and $x_{i_\ell} = x_{j_\ell}$ for $0 \leq \ell \leq n-1$ by the unique decipherability condition, which, in turn, implies $h(s_{i_\ell}) = h(s_{j_\ell})$ for $0 \leq \ell \leq m-1$. Since $h : S \longrightarrow L$ is a bijection, $s_{i_\ell} = s_{j_\ell}$ for $0 \leq \ell \leq m-1$, which means that $s_{i_0} \ldots s_{i_{m-1}} = s_{j_0} \ldots s_{j_{n-1}}$. This shows that the morphism $h : S^* \longrightarrow A^*$ is injective, so $L = h(S)$ is a code set.

(Proof cont'd)

Conversely, suppose that L is a code set, that is, $L = h(S)$, where $h : S \longrightarrow A^*$ is an injective mapping whose extension to S^* is an injective morphism, and that $h(s_i) = x_i$ for every $x_i \in L$. If $x_{i_0}, \ldots, x_{i_{m-1}}, x_{j_0}, \ldots, x_{j_{n-1}}$ are words in L such that $x_{i_0} \cdots x_{i_{m-1}} = x_{j_0} \cdots x_{j_{n-1}}$, then $s_{i_0} \cdots s_{i_{m-1}} = s_{j_0} \cdots s_{j_{n-1}}$, because of the injectivity of the morphism $h : S^* \longrightarrow A^*$. Consequently, $m = n$, $s_{i_\ell} = s_{j_\ell}$ for $0 \leq \ell \leq n - 1$, so h is a code, and L is a code set.

Corollary

A language $L \subseteq A^+$ is not a code set if and only if there exist words $x_{i_0}, \ldots, x_{i_{m-1}}, x_{j_0}, \ldots, x_{j_{n-1}}$ in L such that $x_{i_0} \cdots x_{i_{m-1}} = x_{j_0} \cdots x_{j_{n-1}}$ and x_{i_0} is a proper prefix of x_{j_0}.

Suppose that L is not a code set. Then there exist words

$$x_{i_0}, \ldots, x_{i_{m-1}}, x_{j_0}, \ldots, x_{j_{n-1}} \in L$$

such that $x_{i_0} \cdots x_{i_{m-1}} = x_{j_0} \cdots x_{j_{n-1}}$. Suppose that we choose these words such that $\ell = m + n$ is minimal. Then, $x_{i_0} \neq x_{j_0}$ since otherwise, we would have $x_{i_1} \cdots x_{i_{m-1}} = x_{j_1} \cdots x_{j_{n-1}}$ and this would contradict the minimality of ℓ. Therefore, one of the words x_{i_0}, x_{j_0} is a proper prefix of the other.

Conversely, if $x_{i_0} \cdots x_{i_{m-1}} = x_{j_0} \cdots x_{j_{n-1}}$ and x_{i_0} is a proper prefix of x_{j_0} for some words $x_{i_0}, \ldots, x_{i_{m-1}}, x_{j_0}, \ldots, x_{j_{n-1}}$ in L, then L is not uniquely decipherable, so it is not a code set.

Example

Let A be an alphabet and $L \subseteq A^*$ be a language such that for every $x, y \in L$ with $x \neq y$ we have $x \notin \mathrm{PREF}(y)$. By the previous Corollary L is a code set.

Definition

Let A be an alphabet. A *prefix code* on A is a language $L \subseteq A^*$ such that for every $x, y \in L$ with $x \neq y$ we have $x \notin \mathrm{PREF}(y)$.

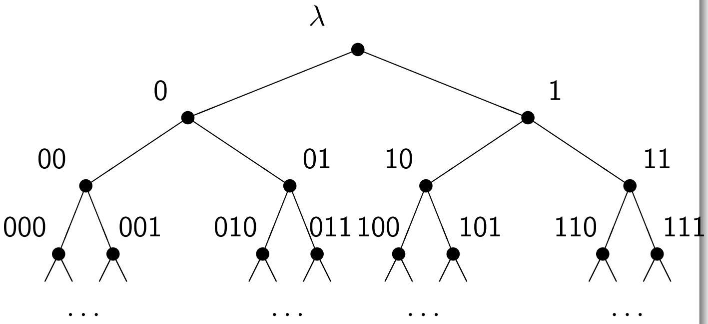
Example

Let $k \in \mathbb{N}$, and let $L_k \subseteq \{a, b\}^*$ be defined by $L_k = \{a^n b \mid 0 \leq n \leq k\}$. Then, L_k is a prefix code, since each code word has exactly one symbol b, which marks its end.

Prefix codes can be obtained using a labeled ordered tree $\mathtt{T}_A$ as a representation of the set of words over an alphabet A. The root of $\mathtt{T}_A$ is labeled by λ; if $A = \{a_0, \ldots, a_{k-1}\}$, then every node labeled by a word $x \in A^*$ has k successors labeled (from left to right) by the words $xa_0, xa_1, \ldots, xa_{k-1}$.

Example

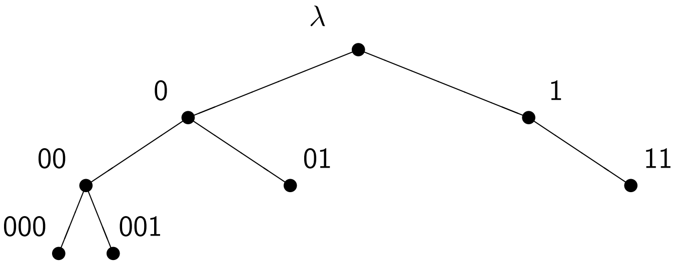
Let $A = \{0, 1\}$ be an alphabet. The labeled ordered tree $\mathtt{T}_A$ is shown here:



(Example cont'd)

Note that a word u is a prefix of another word v if and only if u is the label of a node that occurs on the path that joins the root with v. Therefore, to obtain a prefix code we need to consider a subtree $\mathtt{T}$ of $\mathtt{T}_A$. The prefix code that corresponds to $\mathtt{T}$ comprises the labels of the leaves of $\mathtt{T}$.

For instance, the prefix code that corresponds to the subtree shown below is $\{000, 001, 01, 11\}$.



Definition

A language $L \subseteq A^*$ is *catenatively independent* if $L \cap L^n = \emptyset$ for every $n \geq 2$.

In other words, L is catenatively independent if no word $w \in L$ can be written as $w = w_0 \cdots w_{n-1}$ where $n \geq 2$ and $w_i \in L$ for $0 \leq i \leq n-1$.

Example

The language $L = \{a, aba, baba, bb, bbba\}$ over the alphabet $\{a, b\}$ is catenatively independent.
Also, the language $\{x \in A^* \mid |x| = n\}$ is catenatively independent for any n.

No catenatively independent language may contain λ.

Theorem

(Schützenberger Theorem) *A language L over the alphabet A is a code if and only if L is catenatively independent and $L^*w \cap L^* \neq \emptyset$, $wL^* \cap L^* \neq \emptyset$ for a word $w \in A^*$ imply $w \in L^*$.*

Proof

The conditions of the theorem are sufficient:
Let $L \subseteq A^*$ be a language that satisfies these conditions. Note that $\lambda \notin L$ because of the catenative independence of L.
If L were not a code, we would have words $x_{i_0}, \ldots, x_{i_{n-1}}, x_{j_0}, \ldots, x_{j_{m-1}}$ from L such that

$$x_{i_0} \cdots x_{i_{n-1}} = x_{j_0} \cdots x_{j_{m-1}}$$

and $x_{j_0} = x_{i_0} z$ for some $z \neq \lambda$. Thus, $Lz \cap L \neq \emptyset$, which implies $L^* z \cap L^* \neq \emptyset$. This also gives, by the cancellation property,

$$x_{i_1} \cdots x_{i_{n-1}} = z x_{j_1} \cdots x_{j_{m-1}},$$

so $zL^* \cap L^* \neq \emptyset$. Hence, $z \in L^*$ and $z \neq \lambda$. Since $x_{j_0} = x_{i_0} z$, this contradicts the catenative independence of L.

To prove that the conditions are necessary, assume that L is a code. The catenative independence of L is immediate.
Suppose that $L^*w \cap L^* \neq \emptyset$ and $wL^* \cap L^* \neq \emptyset$ for a word $w \in A^*$. This means that we have words $x_{i_0}, \ldots, x_{i_{m-1}}, x_{j_0}, \ldots, x_{j_{n-1}}$ and $x_{k_0}, \ldots, x_{k_{p-1}}, x_{l_0}, \ldots, x_{l_{q-1}}$ in L such that

$$\begin{aligned} x_{i_0} \cdots x_{i_{m-1}} w &= x_{j_0} \cdots x_{j_{n-1}}, \\ w x_{k_0} \cdots x_{k_{p-1}} &= x_{l_0} \cdots x_{l_{q-1}}. \end{aligned}$$

Combining the above equalities, we obtain

$$x_{i_0} \cdots x_{i_{m-1}} x_{l_0} \cdots x_{l_{q-1}} = x_{j_0} \cdots x_{j_{n-1}} x_{k_0} \cdots x_{k_{p-1}}.$$

The fact that L is a code implies $m + q = n + p$, and in addition, $x_{i_0} = x_{j_0}, \ldots, x_{l_{q-1}} = x_{k_{p-1}}$.
We must have $m \leq n$, because if $m > n$, then $x_{i_n} \ldots x_{i_{m-1}} w = \lambda$, and this would imply $x_{i_n} = \cdots = x_{i_{m-1}} = w = \lambda$, which contradicts the catenative independence of the language L.
If $m = n$, then $w = \lambda \in L^*$; otherwise, $m < n$, and this implies $w = x_{j_m} \cdots x_{j_{n-1}}$, which gives $w \in L^*$.