Minimal linear grammars and codes

Arturo Carpi

Dipartimento di Matematica e Informatica
Università di Perugia

Giornate di Informatica Teorica
in ricordo di Aldo de Luca

Rome, July 11–12, 2019

Encoding and decoding

Encoding requires an injective map $f : B^* \to A^*$, where B and A are, respectively, the source alphabet and the code alphabet.

A morphism $f : A^* \to B^*$ is injective if and only if

- f is injective on letters,
- $f(A)$ is a variable length code.

Minimal linear grammars

Definition

A linear grammar is minimal if it has a unique non-terminal.

Example

Let G be the minimal linear grammar with the productions

$$\begin{aligned} p_a: &\quad X \to aXb, \\ p_b: &\quad X \to bXa, \\ p_T: &\quad X \to \epsilon. \end{aligned}$$

G generates the language of binary antipalindromes.

Unambiguous minimal linear grammars

Let $G = (V, T, P, X)$ be a minimal linear grammar with a unique terminating production p_T, A an alphabet, and $f : A \to P \setminus \{p_T\}$ a bijection.
Let $c_f : A^* \to L(G)$ be the function mapping any word $w = a_1 \cdots a_n$, $(a_1, \ldots, a_n \in A)$ to the word generated in G by the sequence of productions $f(a_1), \ldots, f(a_n), p_T$.

Proposition

The map c_f is a bijection if and only if G is unambiguous.

Thus, one may encode by means of unambiguous minimal linear grammars (in the sequel, briefly, UMLG)

Our goal

Study of some classical problem of Theory of Codes in the more general framework of UMLG.

Minimal linear grammars

Remark

In the sequel, all grammars have a unique terminating production, denoted p_T.

Remark

Let G be a minimal right-linear grammar with productions

$$X \to y_1 X, \quad \cdots, \quad X \to y_n X, \quad X \to z.$$

The grammar G is unambiguous if and only if $Y = \{y_1, \ldots, y_n\}$ is a code.

Proposition (Greibach)

Ambiguity of minimal linear grammars is undecidable.

(no Sardinas-Patterson algorithm for UMLG)

Bernoulli distributions

Definition

A Bernoulli distribution on the alphabet T is a map $\pi\colon T \to \mathbb{R}_+$ such that $\sum_{a\in T} \pi(a) = 1$.
For any word $w = a_1 \cdots a_n$ $(a_1, \ldots, a_n \in T)$, we set

$$\pi(w) = \pi(a_1) \cdots \pi(a_n)$$

and for any subset R of T^*,

$$\pi(R) = \sum_{w\in R} \pi(w).$$

Definition

A subset L of T^* is dense if for all $w \in T^*$ one has $T^*wT^* \cap L \neq \emptyset$.

Maximal codes

Proposition (Schützenberger)

Let X be a regular code on the alphabet T and π a positive Bernoulli distribution on the alphabet T. The following propositions are equivalent:

- X is a maximal code,
- X^* is dense,
- $\pi(X) = 1$.

Bernoulli distributions and UMLG

Definition

Let G be a minimal linear grammar with productions

$$X \to u_1 X v_1, \quad \cdots, \quad X \to u_n X v_n, \quad X \to w_T.$$

We set $\pi(G) = \sum_{i=1}^{n} \pi(u_i v_i)$.

Remark

$\pi(G) = \pi(L_1)/\pi(w_T)$, where $L_1 = \{w \mid X \Rightarrow^2 w\}$.

Proposition

Let G be an UMLG. For all Bernoulli distribution π on the terminal alphabet, one has $\pi(G) \leq 1$. Moreover,

$$\pi(L(G)) = \frac{\pi(w_T)}{1 - \pi(G)}.$$

Maximal UMLG

Definition

An UMLG G is maximal if there does not exists another UMLG on the same terminal alphabet whose set of non-terminating productions properly contains the one of G.

Definition

A subset L of T^* is *very dense* (in T^*) if there exists a finite set $F \subseteq T^*$ such that, for all w in T^*, $FwF \cap L \neq \emptyset$.

Maximality and density

Proposition

Let G be an UMLG. If $L(G)$ is very dense, then $\pi(G) = 1$ for all Bernoulli distribution π on the terminal alphabet.

Proposition

Let G be an UMLG. If one has $\pi(G) = 1$ for some positive Bernoulli distribution π on the terminal alphabet, then G is maximal.

Proposition

Let G be a maximal UMLG. Then $L(G)$ is dense.

Example

Binary antipalindromes are generated by the UMLG with productions

$$X \to aXb, \quad X \to bXa, \quad X \to \epsilon.$$

The language is dense, but the UMLG is not maximal. Moreover,

$$\pi(G) = 2\pi(a)(1 - \pi(a)) \leq 1/2,$$

for all Bernoulli distribution.

Example

Let G be the grammar with the productions

$$X \to uXv, \quad u \in P, v \in S, \qquad X \to \epsilon,$$

where P and S are, respectively, a maximal prefix code and a maximal suffix code on the terminal alphabet.

- G is unambiguous,
- $\pi(G) = 1$ for all Bernoulli distribution π,
- G is maximal,
- $L(G)$ is dense,
- In general, G is not very dense
  (e.g., taking $P = \{a, ba, bb\}$, $S = \{a, ab, bb\}$).

Proportional UMLG

Problem

Does there exists a maximal UMLG G such that $\pi(G) < 1$ for some positive Bernoulli distribution π ?

Definition

An UMLG is proportional if there exists a rational number q such that for all non-terminating production $X \to uXv$ one has $|v| = q|u|$.

Proposition

Let G be a proportional UMLG and π be the uniform Bernoulli distribution. The following conditions are equivalent:

- G is a maximal UMLG,
- G is a maximal proportional UMLG,
- $\pi(G) = 1$.

Optimality

A transmitter computes an injective encoding function $h\colon A^* \to T^*$ on messages $w \in A^*$ generated by a source with probability p_w.
Any letter $a \in T$ has a transmission cost $c(a) \in \mathbb{R}_+$.
The map c is extended to a morphism of T^* into the additive semigroup $\mathbb{R}_+$.
The average cost per letter of the transmission is given by

$$C = \lim_{n\to\infty} \sum_{w\in A^n} p_w\, c(h(w)),$$

provide the limit exists.
In the simplest case, $p_w = \pi(w)$ for some Bernoulli distribution π.

Optimality - 2

If h is a monomorphism, then $Y = h(A)$ is a code and

$$C = \sum_{y \in Y} p_y c(y), \qquad \text{where } p_y = \pi(h^{-1}(y))$$

If, moreover, all letters have the same cost (say 1), then $C = \sum_{y \in Y} p_y |y|$.

If h is obtained from an UMLG G with productions

$$X \to u_1 X v_1, \quad \cdots, \quad X \to u_n X v_n, \quad X \to w_T,$$

then

$$C = \sum_{i=1}^{n} p_i c(u_i v_i), \qquad \text{where } p_i = \pi(h^{-1}(u_i w_T v_i))$$

If all letters have the same cost (say 1), then $C = \sum_{y \in Y} p_y |y|$.

Optimality - 3

Proposition

Let G be an UMLG with productions

$$X \to u_1 X v_1, \quad \cdots, \quad X \to u_n X v_n, \quad X \to w_T.$$

There exists a prefix code $Y = \{y_1, \ldots, y_n\}$ such that $|y_i| = |u_i v_i|$, $i = 1, \ldots, n$.

Proof.

By inequality $\pi(G) \leq 1$ and Kraft-McMillan Theorem. □

Thus, in the case of constant letter cost, coding by UMLG cannot accelerate the transmission rate obtained by optimal prefix codes.

Commutative equivalence

Definition

Two productions $X \to \alpha$ and $X \to \beta$ are commutatively equivalent if β is a permutation of α. Two grammars $G_1 = (V, T, P_1, X)$ and $G_2 = (V, T, P_2, X)$ are commutatively equivalent if there is a bijection $f : P_1 \to P_2$ such that, for all $p \in P_1$, p and $f(p)$ are commutatively equivalent.

Extending an argument of Carter and Gill, one has:

Proposition

The following conjectures are equivalent:

- every UMLG is commutatively equivalent to an unambiguous regular grammar,
- every UMLG can be replaced by a code with the same average cost per letter for every assignment of symbol costs and every probability distribution on source symbols.

Completability

Example

Let G be the UMLG with productions

$$X \to a^5X \mid bX \mid abX \mid ba^2X \mid \epsilon.$$

Since $\{a^5, b, ab, ba^2\}$ is an uncompletable code, the productions of G cannot occur in a regular maximal UMLG.

Problem

Is it possible to obtain a (non-regular) maximal UMLG adding productions to G?
More generally, does there exist 'uncompletable' UMLG ?

Thank you