The pumping lemma for contex-free languages

Definition (Unit rules and λ-rules)

A *unit rule* is a rule of the form $X \to Y$ where X and Y are variable symbols.

A *λ-rule* is a rule of the form $X \to \lambda$ where X is a variable symbol.

Lemma (Removing unit rules and λ-rules)

For a given context-free grammar G one can effectively construct a context-free grammar G' such that

$$\mathrm{L}(G') = \mathrm{L}(G) \setminus \{\lambda\}$$

and the set of rules of G' contains neither unit rules nor λ-rules.

Proof. See lecture.

The pumping lemma for contex-free languages

Definition (Useful variable symbols)

Let $G = (N, T, P, S)$ be a context-free grammar.

A variable symbol X is *useful* if there is a word $z \in T^*$ such that X occurs in some derivation of z, i.e., $z = uvw$ and

$$S \overset{G,*}{\Longrightarrow} uXw \overset{G,*}{\Longrightarrow} uvw;$$

otherwise, X is *useless*.

Lemma (Removing useless variable symbols)

If one cancels in a given context-free grammar G all useless variable symbols and all rules in which these variable symbols occur, one obtains a grammar G' that has only useful variable symbols and where $\mathrm{L}(G) = \mathrm{L}(G')$.

Sketch of proof. It suffices to observe that cancelling a useless variable symbol does not create new useless variable symbols. □

The pumping lemma for contex-free languages

Definition (Chomsky normal form for context-free grammars)

A context-free grammar $G = (N, T, P, S)$ is in *Chomsky normal form* if all rules in P are of the form

$$X \to YZ \quad \text{or} \quad X \to a \qquad \text{where } X, Y, Z \in N \text{ and } a \in T,$$

with the possible exception that P may contain the rule $S \to \lambda$, in which case S does not occur on the right-hand side of any rule in P.

Lemma (Transformation into Chomsky normal form)

For a given context-free grammar G one can effectively construct a context-free grammar G' in Chomsky normal form such that $\mathrm{L}(G) = \mathrm{L}(G')$. In addition, the grammar G' can be chosen such that all its variable symbols are useful.

The pumping lemma for contex-free languages

Proof. We construct context-free grammars G_1, G_2, and G_3 where G_3 is in Chomsky normal form and such that

$$\mathrm{L(G)} = \mathrm{L(G_1)} \;\text{ and }\; \mathrm{L(G_1)} \setminus \{\lambda\} = \mathrm{L(G_2)} = \mathrm{L(G_3)}.$$

In case $\lambda \notin \mathrm{L}(G)$, we can simply let $G' = G_3$.

Otherwise, we obtain G' as required by adding to G_3 a new start symbol S' and rules $S' \to \lambda$ and $S' \to S$, where S is the start symbol of G_3.

By the lemma above, we can in addition ensure that all variable symbols of G' fare useful by removing all useless variable symbols.

The pumping lemma for contex-free languages

Proof, cont.: (Replace terminal symbols a by Z_a)

The grammar G_1 is obtained from G as follows.

For every $a \in T$

replace a by Z_a in all rules of P,

add a variable Z_a and a rule $Z_a \to a$,

where the Z_a are mutually distinct new variables.

By construction, we have $\mathrm{L}(G) = \mathrm{L}(G_1)$.

The rules in G_1 have the form $X \to a$ for a terminal symbol a or

$$X \to Y_1 \cdots Y_t \text{ for some } t \geq 0 \text{ and variable symbols } Y_1, \ldots, Y_t.$$

The pumping lemma for contex-free languages

Proof, cont.: (Remove unit rules and λ-rules)

The grammar G_2 is obtained from G_1 as follows.

Let G_2 be a grammar equivalent to G_1 and without unit rules and λ-rules as in the proof of the corresponding lemma above.

By construction, we have $\mathrm{L}(G_1) \setminus \{\lambda\} = \mathrm{L}(G_2)$.

The rules in G_2 have the form $X \to a$ for a terminal symbol a or

$$X \to Y_1 \cdots Y_t \text{ for some } t \geq 2 \text{ and variable symbols } Y_1, \ldots, Y_t.$$

The pumping lemma for contex-free languages

Proof, cont.: (Remove multiple variable symbols)

The grammar G_3 is obtained from G_2 as follows.

Successively, each rule of the form $X \to Y_1 \cdots Y_t$ where $t \geq 3$ is replaced by the rules

$$\begin{aligned} X &\to Y_1 Z_1, \\ Z_1 &\to Y_2 Z_2, \\ &\vdots \\ Z_{t-1} &\to Y_{t-1} Y_t, \end{aligned}$$

where for each replacement the Z_i are chosen as mutually distinct new variable symbols.

By construction, we have $\mathrm{L}(G_2) = \mathrm{L}(G_3)$.

The rules in G_3 have the form $X \to a$ for a terminal symbol a or

$$X \to Y_1 Y_2 \text{ for variable symbols } Y_1 \text{ and } Y_2. \ \square$$

The pumping lemma for contex-free languages

In what follows, we derive a pumping lemma for contex-free languages, as well as a variant for the subclass of linear languages.

Similar to the case of regular languages, these pumping lemmas are the standard tools for showing that a certain language is not context-free or is not linear.

Theorem (Pumping lemma for context-free languages)

Let L be a context-free language. Then there is a constant k such that every word $z \in L$ of length at least k can be written in the form

$$z = uvwxy$$

where the words u, v, w, x, and y have the following properties

(i) $|vwx| \leq k$,

(ii) $vx \neq \lambda$,

(iii) $uv^iwx^iy \in L$ *for all* $i \geq 0$.

The pumping lemma for contex-free languages

Proof. Fix some context-free grammar $G = (N, T, P, S)$ in Chomsky normal form that generates L and let $k = 2^{|N|+1}$.

For any given word $z \in L$ such that $|z| \geq k$, consider a left derivation α of z in G and the corresponding parse tree $\mathrm{T}(\alpha)$.

The parse tree $\mathrm{T}(\alpha)$ has depth of at least $|N| + 2$ because a parse tree of depth at most $|N| + 1$ has at most $2^{|N|} < |z|$ leave nodes.

Fix a path between the root and a leave node of $\mathrm{T}(\alpha)$ of maximum length among all such paths.

The length of the path is at least $|N| + 1$, hence there is a node K on the path at distance $|N| + 1$ from the leave node of the path.

The subtree of $\mathrm{T}(\alpha)$ with root K has depth $|N| + 1$ because

- the leave node of the path is at this distance from K while
- there cannot be any node at larger distance, otherwise the path would not have maximum length.

The pumping lemma for contex-free languages

Proof, cont.: The part of the path between K and the leave node of the path contains $|N| + 2$ nodes, i.e., contains $|N| + 1$ nodes that are not leave nodes.

Each of these $|N| + 1$ nodes is marked with one of the $|N|$ variable symbols, i.e., there are two distinct nodes K' and K'' that are marked with the same variable symbol, say, with X.

So there are words u, v, w, x, and y over T such that

$$S \overset{G,*}{\Longrightarrow} uXy \overset{G,t}{\Longrightarrow} uvXxy \overset{G,*}{\Longrightarrow} uvwxy = z \qquad \text{for some } t > 0.$$

From this derivation we obtain

(i) $|vwx| \leq 2^{|N|+1} = k$ by choice of the node K,

(ii) $vx \neq \lambda$ by $t > 0$ and because P contains neither unit rules nor λ-rules,

(iii) $uv^iwx^iy \in L$ for all $i \geq 0$ because $X \overset{G,*}{\Longrightarrow} v^iwy^i$ for all $i \geq 0$.

□

The pumping lemma for contex-free languages

Example (A language that is not context-free)

The language $L = \{0^n1^n0^n : n > 0\}$ is not context-free.

For a proof by contradiction, assume that L is context-free.

Choose both, a constant k and a partition $uvwxy$ of $0^k1^k0^k$, as in the pumping lemma for context-free languages, i.e.,

$$0^k1^k0^k = uvwxy, \quad |vwx| \leq k, \quad vx \neq \lambda, \quad \text{and } uwy \in L.$$

Then at least one of the words u and y must have length at least k.

So the word uwy has a prefix 0^k or a suffix 1^k but has length strictly less than $3k$.

Thus $uwy \notin L$, contradicting the choice of u, w, and y.

The pumping lemma for contex-free languages

Example (Words of prime length)

The language $L = \{w \in \{1\}^* : |w| \text{ is prime}\}$ of all words of prime length over the unary alphabet is not context-free.

For a proof by contradiction, assume that L_1 is regular and accordingly choose a constant k as in the pumping lemma for context-free languages and a prime number $p \geq k + 2$.

Let $z = uvwxy$ be a partition of $z = 1^p$ as in the pumping lemma.

For $m = |vx|$, it holds that $1 \leq m \leq |vwx| \leq k \leq p - 2$ $(*)$.

Furthermore, the word $uv^{p-m}wx^{p-m}y$ is in L, however its length cannot be prime because of

$$|uv^{p-m}wx^{p-m}y| = \underbrace{|uwy|}_{=p-m} + m(p-m) = \underbrace{(m+1)}_{\geq 2 \text{ by } (*)}\ \underbrace{(p-m)}_{\geq 2 \text{ by } (*)},$$

which contradicts the definition of L. □

The pumping lemma for contex-free languages

Definition (Linear grammars and languages)

A *grammar* $G = (N, T, P, S)$ is *linear* if all its rules are of the form

$$X \to uYv \quad \text{where } X, Y \in N \text{ and } u, v \in T^*.$$

A language is *linear* if it is generated by a linear grammar.

Example (A linear and a nonlinear languages)

The language $L = \{0^n1^n \colon n \geq 0\}$ is generated by the linear grammar $(\{S\}, \{0,1\}, \{S \to 0S1|\lambda\}, S)$, hence L is linear.

The language $L = \{0^m1^m0^n1^n \colon m, n \geq 0\}$ is not linear, see below.

The pumping lemma for contex-free languages

Definition (Chomsky normal form for linear grammars)

A linear grammar $G = (N, T, P, S)$ is in *Chomsky normal form* if all rules in P are of one of the forms

$$X \to aY, \quad \text{or} \quad X \to Ya, \quad \text{or} \quad X \to a \qquad \text{where } X, Y \in N \text{ and } a \in T,$$

with the possible exception that P may contain the rule $S \to \lambda$, in which case S does not occur on the right-hand side of any rule in P.

Lemma (Transformation into Chomsky normal form)

For a given linear grammar G one can effectively construct a linear grammar G' in Chomsky normal form such that $\mathrm{L}(G) = \mathrm{L}(G')$.

In addition, the grammar G' can be chosen such that all its variable symbols are useful.

Proof. Similar to the context-free case, see lecture.

The pumping lemma for contex-free languages

The following pumping lemma for linear languages is the standard tool for showing that a language is not linear.

Theorem (Pumping lemma for linear languages)

For every linear language L there is a constant k such that every word $z \in L$ of length at least k can be written in the form

$$z = uvwxy$$

where the words u, v, w, x, and y have the following properties

(i) $|uv| \leq k$ *and* $|xy| \leq k$,

(ii) $vx \neq \lambda$,

(iii) $uv^iwx^iz \in L$ *for all* $i \geq 0$.

The pumping lemma for contex-free languages

Proof. Fix some linear grammar $G = (N, T, P, S)$ in Chomsky normal form that generates L and let $k = |N| + 1$.

For any given word $z \in L$ such that $|z| \geq k$, consider a derivation α of z in G, which then must have length of at least k.

Each of the first k sentential forms that occur in α contain a single variable symbol, hence some variable symbol X occurs twice, i.e., there are words u, v, w, x, and y in T^* such that

$$S \overset{G,t_1}{\Longrightarrow} uXy \overset{G,t_2}{\Longrightarrow} uvXxy \overset{G,*}{\Longrightarrow} uvwxy = z,$$

$0 \leq t_1$, $0 < t_2$ and $t_1 + t_2 \leq k - 1$.

Since G is a linear grammar in Chomsky normal form we obtain

(i) $|uv| + |xy| \leq t_1 + t_2 \leq k$,

(ii) $xy \neq \lambda$ by $t_2 > 0$,

(iii) $uv^iwx^iy \in L$ for all $i \geq 0$ because $X \overset{G,*}{\Longrightarrow} v^iwy^i$ for all $i \geq 0$.

□

The pumping lemma for contex-free languages

Example (Linear and context-free languages)

The language $L = \{0^m 1^m 0^n 1^n : m, n \geq 0\}$ is context-free but not linear.

Obviously L can be generated by a context-free grammar.

In order to prove that L is not linear, assume otherwise and let k be a constant as in the pumping lemma for linear languages.

Fix a partition $uvwxy$ of $z = 0^k 1^k 0^k 1^k$ according to the pumping lemma, i.e.,

$$0^k 1^k 0^k 1^k = uvwxy.$$

Then we have by (i) $|uv| \leq k$ and $|xy| \leq k$ and by (ii) $vx \neq \lambda$, hence $v = 0^s$ and $x = 1^t$ where $s + t > 0$.

Consequently, $uwx \notin L$, contradicting the choice of $uvwxy$ and the pumping property (iii).

The pumping lemma for contex-free languages

Corollary (Regular, linear, and context-free languages)

The classes of regular, linear, and context-free languages form a strict hierarchy in the sense that

$$\{L\colon L \text{ regular}\} \subsetneqq \{L\colon L \text{ linear}\} \subsetneqq \{L\colon L \text{ context-free}\}.$$

Proof. The inclusion relations hold because every right-linear grammar is linear, and every linear grammar is context-free.

That the inclusions are proper is witnessed by already discussed counterexamples, i.e., by the languages

$$\{0^n 1^n : n \geq 0\}, \text{ which is linear but not regular and}$$
$$\{0^m 1^m 0^n 1^n : m, n \geq 0\}, \text{ which is context-free but not linear.}$$

□

The pumping lemma for contex-free languages

Theorem (Closure properties of the context-free languages)

The class of context-free languages is closed under union, concatenation, and Kleene closure, i.e.,

(i) *if L_1 and L_2 are context-free, then $L_1 \cup L_2$ is context-free,*

(ii) *if L_1 and L_2 are context-free, then $L_1 L_2$ is context-free,*

(iii) *if L is context-free, then L^* is context-free,*

Proof. (i), (ii) For given context-free languages L_1 and L_2, let $G_1 = (N_1, T_1, P_1, S_1)$ and $G_2 = (N_2, T_2, P_2, S_2)$ be context-free grammars where $L_1 = \mathrm{L}(G_1)$ and $L_2 = \mathrm{L}(G_2)$.

The pumping lemma for contex-free languages

Proof, cont.: Then the languages $L_1 \cup L_2$ and $L_1 L_2$ are generated by the context-free grammars

$$(N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, P_1 \cup P_2 \cup \{S \to S_1 | S_2\}, S) \text{ and}$$
$$(N_1 \cup N_2 \cup \{S\}, T_1 \cup T_2, P_1 \cup P_2 \cup \{S \to S_1 S_2\}, S),$$

respectively, where S is a new variable symbol and for the latter grammar one has to assume that N_1 and N_2 are disjoint.

(iii) For a given context-free language L, let $G = (N, T, P, S)$ be a context-free grammar where $L = \mathrm{L}(G)$.

Then the language L^* is generated by the context-free grammar

$$(N \cup \{S'\}, T, P \cup \{S' \to SS' | \lambda\}, S')$$

where S' is a new variable symbol. □

Remark (Closure properties of the context-free languages)

The class of context-free languages is closed under transition to the mirror language, i.e., if a context-free language L is context-free, then the mirror language $L^{\mathrm{R}} = \{w^{\mathrm{R}} \colon w \in L\}$ is context-free, too.

For a proof, let the language L be context-free.

Choose a context-free grammar $G = (N, T, P, S)$ where $L = \mathrm{L}(G)$.

Then the language L^{R} is generated by the context-free grammar $G^{\mathrm{R}} = (N, T, P^{\mathrm{R}}, S)$ where

$$P^{\mathrm{R}} = \{X \to w^{\mathrm{R}} \colon X \to w \in P\}.$$

It can be shown by induction over n for all words $w \in (N \cup T)^*$ and all n that we have $S \overset{G,n}{\Longrightarrow} w$ if and only if $S \overset{G^{\mathrm{R}},n}{\Longrightarrow} w^{\mathrm{R}}$.

Theorem (Closure properties of the context-free languages)

The class of context-free languages is neither closed under intersection nor under complement, i.e.,

(i) *L_1 and L_2 context-free does not imply $L_1 \cap L_2$ context-free,*

(ii) *L context-free does not imply that $\overline{L}$ is context-free.*

Proof. (i) The languages

$$L_1 = \{0^m 1^m 0^n \colon m, n \geq 0\} \quad \text{and} \quad L_2 = \{0^m 1^n 0^n \colon m, n \geq 0\},$$

are both context-free but have the non-context-free intersection

$$L_1 \cap L_2 = \{0^m 1^m 0^m \colon m \geq 0\}$$

(ii) The context-free languages are closed under union, hence closure under complement would imply closure under intersection by de Morgan's law $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$. □

It can be shown by specifying an appropriate grammar that the complement of the language $\{0^m 1^m 0^m \colon m \geq 0\}$ is context-free.