Zur Komplexität des CYK-Algorithmus

Es werden $\frac{n^2+n}{2}$ Mengen V_{ij} berechnet. Für jede dieser Mengen werden $|P|$ Produktionen und höchstens n Werte für k betrachtet. Der Test der Bedingung $(A \to BC) \in P \wedge B \in V_{ik} \wedge C \in V_{k+1,j}$ erfordert bei geeigneter Repräsentation der Mengen V_{ij} konstanten Aufwand. Der Gesamtaufwand ist also $O(|P|n^3)$.

Mit der gleichen Methode und dem gleichen Rechenaufwand kann man zu dem getesteten Wort, falls es in der Sprache ist, auch gleich einen Ableitungsbaum konstruieren, indem man sich bei der Konstruktion der V_{ij} nicht nur merkt, welche Nichtterminale sie enthalten, sondern auch gleich, warum sie sie enthalten, d.h. aufgrund welcher Produktionen sie in die Menge aufgenommen wurden.

7.3 Das Pumping-Lemma für kontextfreie Sprachen

Zur Erinnerung: Das Pumping-Lemma für reguläre Sprachen: Für jede reguläre Sprache L gibt es eine Konstante $n \in \mathbb{N}$, so dass sich jedes Wort $z \in L$ mit $|z| \geq n$ zerlegen lässt in $z = uvw$ mit $|uv| \leq n$, $|v| \geq 1$ und $uv^*w \subseteq L$.

Zum Beweis haben wir $n = |Q|$ gewählt, wobei Q die Zustandsmenge eines L erkennenden DFA war. Das Argument war dann, dass beim Erkennen von z (mindestens) ein Zustand zweimal besucht werden muss und damit der dazwischen liegende Weg im Automaten beliebig oft wiederholt werden kann.

Völlig gleichwertig kann man argumentieren, dass bei der Ableitung von z mittels einer rechtslinearen Grammatik ein Nichtterminalsymbol (mindestens) zweimal auftreten muss und die dazwischen liegende Teilableitung beliebig oft wiederholt werden kann.

Genau dieses Argument kann in ähnlicher Form auch auf kontextfreie Grammatiken (in Chomsky-Normalform) angewendet werden:

Satz 96 (Pumping-Lemma)

Für jede kontextfreie Sprache L gibt es eine Konstante $n \in \mathbb{N}$, so dass sich jedes Wort $z \in L$ mit $|z| \geq n$ zerlegen lässt in

$$z = uvwxy,$$

mit

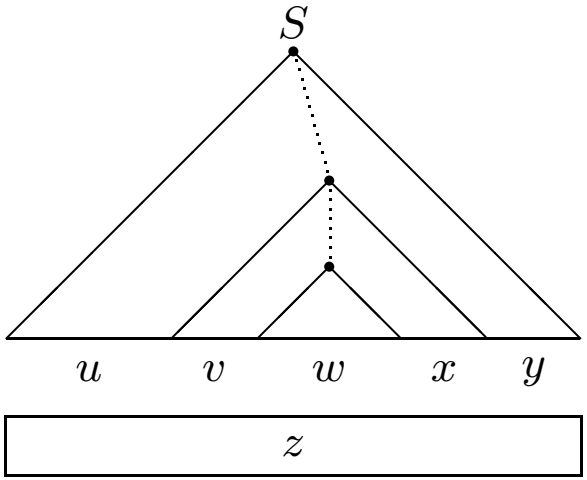
1. $|vx| \geq 1$,
2. $|vwx| \leq n$, *und*
3. $\forall i \in \mathbb{N}_0 : \; uv^i wx^i y \in L$.

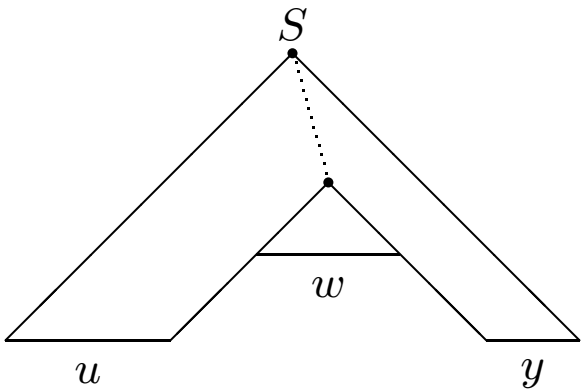
Beweis:

Sei $G = (V, \Sigma, P, S)$ eine Grammatik in Chomsky-Normalform mit $L(G) = L$. Wähle $n = 2^{|V|}$. Sei $z \in L(G)$ mit $|z| \geq n$. Dann hat der Ableitungsbaum für z (ohne die letzte Stufe für die Terminale) mindestens die Tiefe $|V| + 1$, da er wegen der Chomsky-Normalform den Verzweigungsgrad 2 hat.

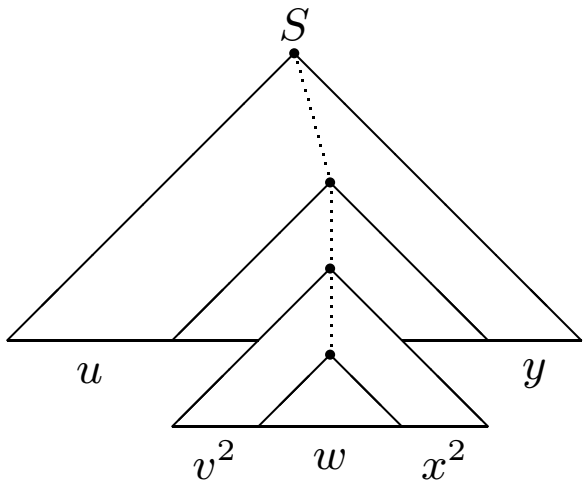
Auf einem Pfadabschnitt der Länge $\geq |V| + 1$ kommt nun mindestens ein Nichtterminal wiederholt vor. Die zwischen diesen beiden Vorkommen liegende Teilableitung kann nun beliebig oft wiederholt werden.

Beweis:





Dieser Ableitungsbaum zeigt
$uwy \in L$



Dieser Ableitungsbaum zeigt
$uv^2wx^2y \in L$

Beweis:

Sei $G = (V, \Sigma, P, S)$ eine Grammatik in Chomsky-Normalform mit $L(G) = L$. Wähle $n = 2^{|V|}$. Sei $z \in L(G)$ mit $|z| \geq n$. Dann hat der Ableitungsbaum für z (ohne die letzte Stufe für die Terminale) mindestens die Tiefe $|V| + 1$, da er wegen der Chomsky-Normalform den Verzweigungsgrad 2 hat.

Auf einem Pfadabschnitt der Länge $\geq |V| + 1$ kommt nun mindestens ein Nichtterminal wiederholt vor. Die zwischen diesen beiden Vorkommen liegende Teilableitung kann nun beliebig oft wiederholt werden.

Um $|vwx| \leq n$ zu erreichen, muss man das am weitesten unten liegende Doppelvorkommen eines solchen Nichtterminals wählen.

□

Beispiel 97

Wir wollen sehen, dass die Sprache

$$\{a^i b^i c^i;\ i \in \mathbb{N}_0\}$$

nicht kontextfrei ist.

Wäre sie kontextfrei, so könnten wir das Wort $a^n b^n c^n$ (n die Konstante aus dem Pumping-Lemma) aufpumpen, ohne aus der Sprache herauszufallen. Wir sehen aber leicht, dass dann für die Zerlegung $z = uvwxy$

$$\#_a(vx) = \#_b(vx) = \#_c(vx) > 0 \text{ und } v, x \in a^* + b^* + c^*$$

gelten muss, letzteres, damit die a's, b's und c's beim Pumpen nicht in der falschen Reihenfolge auftreten. Damit ergibt sich aber Widerspruch!

Zur Vereinfachung von Beweisen wie in dem gerade gesehenen Beispiel führen wir die folgende Verschärfung des Pumping-Lemmas ein:

Satz 98 (Ogdens Lemma)

Für jede kontextfreie Sprache L gibt es eine Konstante $n \in \mathbb{N}$, so dass für jedes Wort $z \in L$ mit $|z| \geq n$ die folgende Aussage gilt: Werden in z mindestens n (beliebige) Buchstaben markiert, so lässt sich z zerlegen in

$$z = uvwxy,$$

so dass

1. *in vx mindestens ein Buchstabe und*
2. *in vwx höchstens n Buchstaben markiert sind und*
3. $(\forall i \in \mathbb{N}_0)[uv^i wx^i y \in L]$.

Bemerkung: Das Pumping-Lemma ist eine triviale Folgerung aus Ogdens Lemma (markiere alle Buchstaben in z).

Beweis:

Sei $G = (V, \Sigma, P, S)$ eine Grammatik in Chomsky-Normalform mit $L(G) = L$. Wähle $n = 2^{|V|+1}$. Sei $z \in L$ und seien in z mindestens n Buchstaben markiert. In einem Ableitungsbaum für z markieren wir alle (inneren) Knoten, deren linker *und* rechter Teilbaum *jeweils* mindestens ein markiertes Blatt enthalten. Es ist nun offensichtlich, dass es einen Pfad von der Wurzel zu einem Blatt gibt, auf dem mindestens $|V| + 1$ markierte innere Knoten liegen.

Beweis:

...

Wir betrachten die letzten $|V|+1$ markierten inneren Knoten eines Pfades mit maximaler Anzahl markierter Knoten; nach dem Schubfachprinzip sind zwei mit demselben Nichtterminal, z.B. A, markiert. Wir nennen diese Knoten v_1 und v_2. Seien die Blätter des Teilbaumes mit der Wurzel v_2 insgesamt mit w und die Blätter des Teilbaumes mit der Wurzel v_1 insgesamt mit vwx beschriftet. Es ist dann klar, dass die folgende Ableitung möglich ist:

$$S \to^* uAy \to^* uvAxy \to^* uvwxy\,.$$

Es ist auch klar, dass der Mittelteil dieser Ableitung weggelassen oder beliebig oft wiederholt werden kann.

Beweis:

...

Es bleibt noch zu sehen, dass vx mindestens einen und vwx höchstens n markierte Buchstaben enthält. Ersteres ist klar, da auch der Unterbaum von v_1, der v_2 nicht enthält, ein markiertes Blatt haben muss.

Letzteres ist klar, da der gewählte Pfad eine maximale Anzahl von markierten inneren Knoten hatte und unterhalb von v_1 nur noch höchstens $|V|$ markierte Knoten auf diesem Pfad sein können. Der Teilbaum mit Wurzel v_1 kann also maximal $2^{|V|+1} = n$ markierte Blätter haben. Formal kann man z.B. zeigen, dass ein Unterbaum, der auf jedem Ast maximal k markierte (innere) Knoten enthält, höchstens 2^k markierte Blätter enthält. □

Beispiel 99

$$L = \{a^i b^j c^k d^l;\ i = 0 \text{ oder } j = k = l\}\,.$$

Hier funktioniert das normale Pumping-Lemma nicht, da für z mit $|z| \geq n$ entweder z mit a beginnt und dann z.B. $v \in \{a\}^+$ sein kann oder aber z nicht mit a beginnt und dann eine zulässige Zerlegung $z = uvwxy$ sehr einfach gewählt werden kann.

Sei n die Konstante aus Ogdens Lemma. Betrachte das Wort $ab^n c^n d^n$ und markiere darin $bc^n d$. Nun gibt es eine Zerlegung $ab^n c^n d^n = uvwxy$, so dass vx mindestens ein markiertes Symbol enthält und $uv^2wx^2y \in L$.

Es ist jedoch leicht zu sehen, dass dies einen Widerspruch liefert, da vx höchstens zwei verschiedene der Symbole b, c, d enthalten kann, damit beim Pumpen nicht die Reihenfolge durcheinander kommt.

Bemerkung:
Wie wir gerade gesehen haben, gilt die Umkehrung des Pumping-Lemmas nicht allgemein (d.h., aus dem Abschluss einer Sprache unter der Pumpoperation des Pumping-Lemmas folgt i.A. nicht, dass die Sprache kontext-frei ist).
Es gibt jedoch stärkere Versionen des Pumping-Lemmas, für die auch die Umkehrung gilt. Siehe dazu etwa

David S. Wise:
A strong pumping lemma for context-free languages.
Theoretical Computer Science **3**, pp. 359–369, 1976

Richard Johnsonbaugh, David P. Miller:
Converses of pumping lemmas.
ACM SIGCSE Bull. **22**(1), pp. 27–30, 1990

Satz 100

Sei $G = (V, \Sigma, P, S)$ kontextfrei. Dann kann die Menge V' der Variablen $A \in V$, für die gilt:

$$(\exists w \in \Sigma^*)[A \to^* w]$$

in Zeit $O(|V| \cdot s(G))$ berechnet werden.

Beweis:

Betrachte folgenden Algorithmus:

$\Delta := \{A \in V;\ (\exists (A \to w) \in P \text{ mit } w \in \Sigma^*\};\ V' := \emptyset;$
while $\Delta \neq \emptyset$ **do**
  $V' := V' \cup \Delta$
  $\Delta := \{A \in V \setminus V';\ (\exists A \to \alpha) \in P \text{ mit } \alpha \in (V' \cup \Sigma)^*\}$
od

Induktion über die Länge der Ableitung. □

Definition 101

$A \in V$ heißt nutzlos, falls es keine Ableitung

$$S \to^* w, \quad w \in \Sigma^*$$

gibt, in der A vorkommt.

Satz 102

Die Menge der nutzlosen Variablen kann in Zeit $O(|V| \cdot s(G))$ bestimmt werden.

Beweis:

Sei V'' die Menge der nicht nutzlosen Variablen.
Offensichtlich gilt: $V'' \subseteq V'$ (V' aus dem vorigen Satz).
Falls $S \notin V'$, dann sind alle Variablen nutzlos.

Ansonsten:

$\Delta := \{S\}$; $V'' := \emptyset$;
while $\Delta \neq \emptyset$ **do**
$V'' := V'' \cup \Delta$
$\Delta := \{B \in V' \setminus V''$; $(\exists A \to \alpha B \beta) \in P$ mit $A \in V''$,
$\alpha, \beta \in (V' \cup \Sigma)^*\}$
od

Induktion über Länge der Ableitung: Am Ende des Algorithmus ist V'' gleich der Menge der nicht nutzlosen Variablen. □

Bemerkung: Alle nutzlosen Variablen und alle Produktionen, die nutzlose Variablen enthalten, können aus der Grammatik entfernt werden, ohne die erzeugte Sprache zu ändern.

Korollar 103

Für eine kontextfreie Grammatik G kann in Zeit $O(|V| \cdot s(G))$ entschieden werden, ob $L(G) = \emptyset$.

Beweis:

$L(G) = \emptyset \Longleftrightarrow S \notin V''$ (bzw. $S \notin V'$) □

Satz 104

Für eine kontextfreie Grammatik $G = (V, \Sigma, P, S)$ ohne nutzlose Variablen und in Chomsky-Normalform kann in linearer Zeit entschieden werden, ob

$$|L(G)| < \infty\,.$$

Beweis:

Definiere gerichteten Hilfsgraphen mit Knotenmenge V und

$$\text{Kante } A \to B \iff (A \to BC) \text{ oder } (A \to CB) \in P\,.$$

$L(G)$ ist endlich $\iff$ dieser Digraph enthält keinen Zyklus.

Verwende Depth-First-Search (DFS), um in linearer Zeit festzustellen, ob der Digraph Zyklen enthält. □

Satz 105

Seien kontextfreie Grammatiken $G_1 = (V_1, \Sigma_1, P_1, S_1)$ *und* $G_2 = (V_2, \Sigma_2, P_2, S_2)$ *gegeben. Dann können in linearer Zeit kontextfreie Grammatiken für*

1. $L(G_1) \cup L(G_2)$,
2. $L(G_1)L(G_2)$,
3. $(L(G_1))^*$

konstruiert werden. Die Klasse der kontextfreien Sprachen ist also unter Vereinigung, Konkatenation und Kleene'scher Hülle abgeschlossen.

Beweis:

Ohne Beschränkung der Allgemeinheit nehmen wir an, dass $V_1 \cap V_2 = \emptyset$.

1. $V = V_1 \cup V_2 \cup \{S\}$; S neu
   $P = P_1 \cup P_2 \cup \{S \to S_1 | S_2\}$
2. $V = V_1 \cup V_2 \cup \{S\}$; S neu
   $P = P_1 \cup P_2 \cup \{S \to S_1 S_2\}$
3. $V = V_1 \cup \{S, S'\}$; S, S' neu
   $P = P_1 \cup \{S \to S' | \epsilon, S' \to S_1 S' | S_1\}$

Falls $\epsilon \in L(G_1)$ oder $\epsilon \in L(G_2)$, sind noch Korrekturen vorzunehmen, die hier als Übungsaufgabe überlassen bleiben. □

Satz 106

Die Klasse der kontextfreien Sprachen ist nicht abgeschlossen unter Durchschnitt oder Komplement.

Beweis:

Es genügt zu zeigen (wegen de Morgan (1806–1871)): nicht abgeschlossen unter Durchschnitt.

$$L_1 := \{a^i b^i c^j;\ i,j \geq 0\} \text{ ist kontextfrei}$$

$$L_2 := \{a^i b^j c^j;\ i,j \geq 0\} \text{ ist kontextfrei}$$

$$L_1 \cap L_2 = \{a^i b^i c^i;\ i \geq 0\} \text{ ist nicht kontextfrei}$$

□

Satz 107

Die Klasse der kontextfreien Sprachen ist abgeschlossen gegenüber Substitution (mit kontextfreien Mengen).

Beweis:

Ersetze jedes Terminal a durch ein neues Nichtterminal S_a und füge zu den Produktionen P für jedes Terminal a die Produktionen einer kontextfreien Grammatik $G_a = (V_a, \Sigma, P_a, S_a)$ hinzu.
Forme die so erhaltene Grammatik in eine äquivalente Chomsky-2-Grammatik um. □