

FORMALE SYSTEME

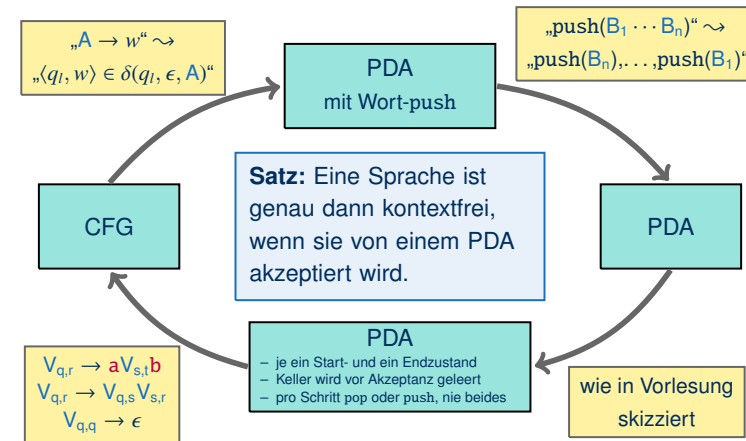
17. Vorlesung: Deterministische Typ-2-Sprachen / Entscheidungsprobleme

Hannes Straß

Folien: © Markus Krötzsch, <https://iccl.inf.tu-dresden.de/web/FS2020>, CC BY 3.0 DE

TU Dresden, 6. Dezember 2021

PDA $\hat{=}$ CFG $\hat{=}$ Typ 2



Deterministische Kellerautomaten

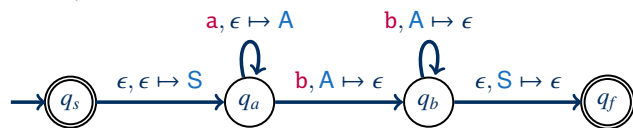
Ein **deterministischer Kellerautomat** (international: „DPDA“) M ist ein Tupel $M = \langle Q, \Sigma, \Gamma, \delta, q_0, F \rangle$ bestehend aus Zustandsmenge Q , Eingabealphabet Σ , Kelleralphabet Γ , Startzustand $q_0 \in Q$, Endzustände $F \subseteq Q$ und partieller Übergangsfunktion

$$\delta : Q \times \Sigma_\epsilon \times \Gamma_\epsilon \rightarrow Q \times \Gamma_\epsilon,$$

so dass für alle $q \in Q$, $a \in \Sigma$ und $A \in \Gamma$ jeweils nur eines der folgenden definiert ist:

$\delta(q, a, A)$ $\delta(q, a, \epsilon)$ $\delta(q, \epsilon, A)$ $\delta(q, \epsilon, \epsilon)$

Beispiel: $\{a^i b^j \mid i \geq 1\}$



Deterministisch vs. nichtdeterministisch

Typ-2-Sprachen

- Erkannt durch PDAs
- Nicht unter Komplement abgeschlossen
- Echte Obermenge der det. Typ-2-Sprachen, z.B. $\{a^i b^j c^k \mid i \neq j \text{ oder } j \neq k\}$
- Wortproblem in $O(|w|^3)$ (CYK-Algorithmus)
- Generiert durch CFGs

Deterministische Typ-2-Sprachen

- Erkannt durch DPDAs
- Unter Komplement abgeschlossen
- Echte Obermenge der regulären Sprachen, z.B. $\{a^i b^j \mid i \geq 0\}$
- Wortproblem in $O(|w|)$ (DPDA-Abarbeitung)
- Generiert durch deterministische CFGs (kein Vorlesungsstoff)

Deterministische Typ-2-Sprachen

Mehrdeutige Grammatiken

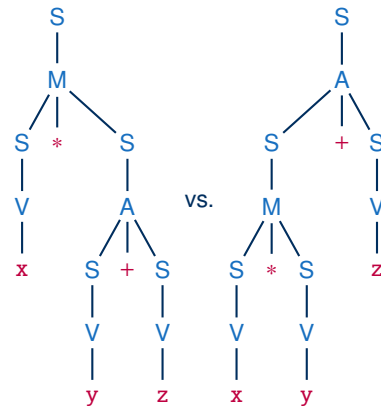
Wir wissen: Ein Ableitungsbaum kann zumeist durch verschiedene Ableitungen erzeugt werden.

Es gilt aber auch: Ein Wort kann mehrere verschiedene Ableitungsbäume haben.

Beispiel:

$$S \rightarrow A \mid M \mid V \quad A \rightarrow S+S$$

$$M \rightarrow S*S \quad V \rightarrow x \mid y \mid z$$



Rückblick: Ableitungsbäume

Rückblick:

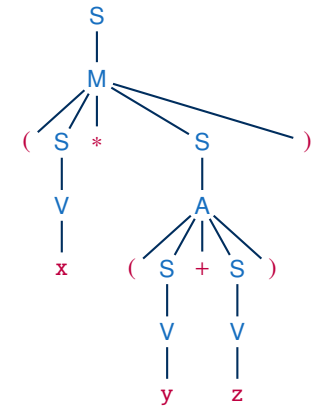
Die Interpretation von Wörtern kontextfreier Sprachen basiert zumeist auf dem Syntaxbaum.

Beispiel:

$$S \rightarrow A \mid M \mid V \quad A \rightarrow (S+S)$$

$$M \rightarrow (S*S) \quad V \rightarrow x \mid y \mid z$$

Wort: $(x*(y+z))$



Mehrdeutige Grammatiken (2)

Eine Grammatik G ist genau dann **mehrdeutig**, wenn es ein Wort $w \in L(G)$ gibt, das mehrere Syntaxbäume zulässt (äquivalent: wenn es für w mehrere unterschiedliche Linksableitungen gibt).

Mehrdeutige Grammatiken in der Praxis:

- unerwünscht in technischen Sprachdefinitionen (Programmiersprachen, Datenformate)
- relevant in der Sprachverarbeitung (um die Mehrdeutigkeit natürlicher Sprache abzubilden)

Für viele mehrdeutige Grammatiken kann man eindeutige Grammatiken finden, welche die gleiche Sprache generieren, z.B.:

$$S \rightarrow A \mid M \quad A \rightarrow A+A \mid M \quad M \rightarrow M*M \mid V \quad V \rightarrow x \mid y \mid z$$

Aber: Es gibt inhärent mehrdeutige Sprachen, die nur mehrdeutige Grammatiken haben, z.B. $\{a^i b^j c^k \mid i, k \geq 0\} \cup \{a^i b^k c^k \mid i, k \geq 0\}$.

Mehrdeutig vs. deterministisch

Fakten:

- Deterministische Sprachen haben immer auch eindeutige Grammatiken. (Allerdings womöglich nicht ausschließlich solche.)
- Aber: Eindeutige Grammatiken können nichtdeterministische Typ-2-Sprachen beschreiben.

deterministisch \subseteq eindeutig \subseteq Typ 2

Unterschiedliche Motivationen:

- eindeutige Sprachdefinitionen: eindeutige Syntaxbäume, eindeutige Interpretation
- deterministische Sprachen: effizienteres Parsing möglich

Viele Programmiersprachen verwenden deterministische (und eindeutige) Grammatiken.

Nach vorne blicken

Deterministische Typ-2-Sprachen sind praktisch sehr relevant:

- Programmiersprachen sind meist im Kern deterministisch
- Man kann sie sehr effizient parsen (kein CYK)



Donald Knuth, 2005
CC-BY-SA 2.5
(C) J. Appelbaum

In der Praxis hilft eine Verallgemeinerung von DCFGs: **Grammatiken mit endlicher Vorschau** (Lookahead).

Idee:

- Wort wird von links nach rechts gelesen
- Grammatikregeln werden rückwärts angewendet, um Teile des gelesenen Worts zu reduzieren
- Die Wahl der Grammatikregel hängt nur vom schon gelesenen Wort und von bis zu k weiteren Symbolen ab (Vorschau)

Grammatiken, die das erlauben, sind vom Typ **LR**(k), wobei **LR**(0) (keine Vorschau) DCFGs sind.

Deterministische CFGs

Man kann die Klasse der deterministischen Typ-2-Sprachen auch durch Grammatiken beschreiben:

- Dies führt zu **deterministischen kontextfreien Grammatiken** (DCFGs).
- Die eigentliche Definition ist relativ technisch.
- DPDAs erkennen **im Prinzip** dieselben Sprachen wie DCFGs.

(... sofern man sich auf Sprachen beschränkt, bei denen jedes Wort als letztes Zeichen ein spezielles Schlussymbol verwendet, um das Ende zu markieren.)

Gute Nachricht: Man kann effektiv entscheiden, ob eine gegebene CFG deterministisch ist.

Schlechte Nachricht: Praktische Grammatiken erfüllen die strengen Bedingungen oft nicht, auch wenn sie eine deterministisch kontextfreie Sprache beschreiben.

Abschlusseigenschaften (1)

Wir wissen bereits, dass deterministische Typ-2-Sprachen unter Komplement abgeschlossen sind.

Für Schnitt gilt das allerdings nicht:

Satz: Deterministische Typ-2-Sprachen sind nicht unter Schnitt abgeschlossen.

Beweis: Der Beweis für Typ-2-Sprachen funktioniert auch hier. Die Sprachen

$$L_1 = \{a^i b^i c^k \mid i \geq 0, k \geq 0\}$$

$$L_2 = \{a^i b^k c^k \mid i \geq 0, k \geq 0\}.$$

sind deterministisch kontextfrei (Übung: Geben Sie entsprechende DPDAs an). Ihr Schnitt $L_1 \cap L_2 = \{a^i b^i c^i \mid i \geq 0\}$ ist dagegen nicht einmal kontextfrei. \square

Abschlusseigenschaften (2)

Der Nichtabschluss unter Schnitt hat weitere Konsequenzen:

Satz: Deterministische Typ-2-Sprachen sind nicht unter Vereinigung abgeschlossen.

Beweis: Angenommen, sie wären unter Vereinigung abgeschlossen, dann wären sie (mittels der Regeln von De Morgan) auch unter Schnitt abgeschlossen, da sie bereits unter Komplement abgeschlossen sind. Widerspruch. \square

Es gilt aber folgende schwächere Aussage:

Lemma: Ist L_1 deterministisch kontextfrei und L_2 regulär, so ist $L_1 \cap L_2$ deterministisch kontextfrei.

Beweisidee: Wir betrachten einen DPDA \mathcal{P} für L_1 und einen DFA \mathcal{M} für L_2 und konstruieren den Produktautomaten $\mathcal{P} \otimes \mathcal{M}$. \square

Quiz: Abschlusseigenschaften DCF

Quiz: Positionieren Sie sich zu folgendem Beweisversuch.

...

Abschlusseigenschaften (3)

Satz: Deterministische Typ-2-Sprachen sind nicht unter Konkatenation abgeschlossen.

Beweis: Wir betrachten die Sprachen $L_0 = \emptyset^*$, $L_1 = \{a^i b^j c^k \mid i, j, k \geq 0, i \neq j\}$, $L_2 = \{a^i b^j c^k \mid i, j, k \geq 0, j \neq k\}$, $L_3 = \emptyset L_1 \cup L_2$. Die Sprachen L_0 und L_3 sind deterministisch kontextfrei. Angenommen, die Konkatenation $L_0 \circ L_3 = \emptyset^*(\emptyset L_1 \cup L_2)$ wäre ebenfalls deterministisch kontextfrei. Dann wäre auch der Schnitt dieser Konkatenation mit einer regulären Sprache wiederum deterministisch kontextfrei. Dementsprechend wäre auch $\emptyset^*(\emptyset L_1 \cup L_2) \cap \emptyset a^* b^* c^* = \emptyset L_1 \cup \emptyset L_2 = \emptyset(L_1 \cup L_2)$ und damit auch $L_1 \cup L_2$ deterministisch kontextfrei. Nun ist aber die Sprache $L_1 \cup L_2 = \overline{L_1} \cap \overline{L_2} = \{a^i b^j c^i \mid i \geq 0\}$ nicht deterministisch kontextfrei. Widerspruch. \square

Mit ähnlichen Beweisideen ergibt sich: (Ohne Beweis.)

Satz: Deterministische Typ-2-Sprachen sind nicht unter Kleene-Stern abgeschlossen.

Zusammenfassung: Deterministische Typ-2-Sprachen sind abgeschlossen unter Komplement, aber nicht unter Vereinigung, Schnitt, Konkatenation oder Kleene-Stern.

Entscheidungsprobleme auf kontextfreien Sprachen

Rückblick Entscheidungsprobleme

Für reguläre Sprachen haben wir eine Reihe von Problemstellungen kennengelernt:

- **Leerheitsproblem:** Ist die beschriebene Sprache gleich der leeren Menge \emptyset ?
- **Inklusionsproblem:** Ist eine beschriebene Sprache Teilmenge einer anderen?
- **Äquivalenzproblem:** Wird durch zwei Beschreibungen die selbe Sprache gegeben?
- **Endlichkeitsproblem:** Ist die beschriebene Sprache endlich?
- **Universalitätsproblem:** Ist die beschriebene Sprache Σ^* ?

Dabei könnten Sprachen durch verschiedene Beschreibungen gegeben sein (Automaten, Grammatiken, ...)

Zudem gibt es freilich noch das Wortproblem (für [D]CFGs bereits besprochen).

Leerheit entscheiden

Satz: Die Leerheit einer CFG ist entscheidbar.

Beweis: Man markiert Variablen mit folgender Prozedur:

- Markiere alle Variablen, welche direkt in ein Wort aus Terminalzeichen umgeschrieben werden können.
- Markiere iterativ alle Variablen, welche in ein Wort aus Terminalzeichen und markierten Variablen umgeschrieben werden können.

Die Sprache ist genau dann nicht leer, wenn bei diesem Verfahren das Startsymbol markiert wird. \square

Meistens unentscheidbar

Viele interessante Fragen sind leider im Allgemeinen nicht durch Algorithmen lösbar:

Satz: Inklusion, Äquivalenz und Universalität von CFGs ist unentscheidbar.

(Ohne Beweis, da wir noch gar nicht über Entscheidbarkeit gesprochen haben ...)

Einiges ist aber doch machbar:

Satz: Leerheit und Endlichkeit einer CFG sind entscheidbar.

Diese Ergebnisse gelten ebenso, wenn PDAs statt CFGs gegeben sind, da wir diese ja in CFGs umwandeln können.

Endlichkeit entscheiden

Satz: Endlichkeit der Sprache $L(G)$ einer CFG G ist entscheidbar.

Beweis: Sei n die Zahl aus dem Pumpinglemma (also $2^{|V|}$ nach Umwandlung in Chomsky-Normalform).

- Wenn es ein Wort $z \in L(G)$ mit $n \leq |z| \leq 2n$ gibt, dann ist $L(G)$ unendlich (da man das Pumping-Lemma auf z anwenden kann).
- Wenn $L(G)$ unendlich ist, dann gibt es ein Wort $z \in L(G)$ mit $n \leq |z| \leq 2n$ (Beweis: Es muss Wörter mit mehr als n Zeichen geben. Sei z ein kürzestes Wort dieser Art. Laut Pumping-Lemma gibt es eine Zerlegung $z = uvwxy$ mit $1 \leq |vx| \leq |vwx| \leq n$ und $uv^0wx^0y = uwy \in L(G)$. Da uwy kürzer ist als z (und z ein Wort kürzester Länge größer als n ist), muss $|uwy| \leq n$ gelten. Mit $|vx| \leq n$ folgt daraus $|z| = |uwy| + |vx| \leq 2n$.)

Das heißt, wir müssen nur testen, ob es so ein Wort $z \in L(G)$ mit $n \leq |z| \leq 2n$ gibt. Das kann man (Brute Force) für alle Wörter dieser Länge tun (da das Wortproblem lösbar ist). \square

(Es gibt effizientere Verfahren, aber dieses ist das einfachste für den Beweis.)

Alles unentscheidbar

Viele weitere interessante Fragen sind leider ebenfalls unentscheidbar:

- **Regularität:** Ist die durch eine CFG gegebene Sprache regulär?
- **Mehrdeutigkeit:** Ist eine gegebene CFG mehrdeutig?
- **Determinisierbarkeit:** Ist die durch eine CFG gegebene Sprache deterministisch?¹
- **Schnittproblem:** Haben zwei gegebene Sprachen gemeinsame Wörter?

¹Aber, wie zuvor angemerkt: Man kann entscheiden, ob eine gegebene CFG bereits deterministisch ist. (Wenn sie es nicht ist, dann bedeutet das aber nicht, dass es keine äquivalente DCFG geben könnte.)

Entscheidungsprobleme für DPDAs

Die Situation ist etwas besser bei DPDAs:

- Leerheit** **entscheidbar** (wie bei CFGs)
- Endlichkeit** **entscheidbar** (wie bei CFGs)
- Universalität** **entscheidbar** (entspricht Leerheit des Komplements)
- Regularität** **entscheidbar**
(Stearns: A Regularity Test for Pushdown Machines, 1967)
- Inklusion** **unentscheidbar**
(Ginsburg & Greibach: Deterministic context-free languages, 1966)
- Schnitt** **unentscheidbar**
(Wie Inklusion, da wir Komplemente haben: $X \subseteq Y$ gdw. $X \cap \bar{Y} = \emptyset$)
- Äquivalenz** **entscheidbar!**
(Séizergues: $L(A)=L(B)$? decidability results from complete formal systems, 2001; komplexes Verfahren ohne Komplexitätsschranken; Ergebnis bekannt seit 1997)

(Mehrdeutigkeit und Determinisierbarkeit sind bei DPDAs trivial.)

Übersicht

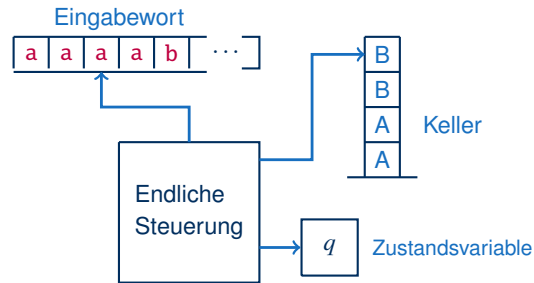
	CFG	DPDA
Wortproblem	in $O(w ^3)$	in $O(w)$
Leerheit	entscheidbar	entscheidbar
Endlichkeit	entscheidbar	entscheidbar
Universalität	unentscheidbar	entscheidbar
Inklusion	unentscheidbar	unentscheidbar
Schnitt	unentscheidbar	unentscheidbar
Äquivalenz	unentscheidbar	entscheidbar
Regularität	unentscheidbar	entscheidbar
Mehrdeutigkeit	unentscheidbar	trivial
Determinisierbarkeit	unentscheidbar	trivial

Rechnen mit Typ 1 und Typ 0

Wortprobleme jenseits von Typ 2

Wir haben gesehen:

- Endliche Automaten erkennen Typ-3-Sprachen.
- Endliche Automaten erweitert um Kellerspeicher erkennen Typ-2-Sprachen.



Für Typ 1 und Typ 0 benötigen wir mehr als das – aber was?

Berechnungsmodelle nach Kellerautomaten?

Beobachtung: Auch jenseits von Typ 2 kann man Wortprobleme algorithmisch lösen.

Beispiel: Die Sprache $\{a^i b^i c^i \mid i \geq 0\}$ ist nicht kontextfrei, wird also von keinem PDA erkannt. Dennoch wäre es nicht sehr schwer, ein Programm in einer beliebigen Programmiersprache zu schreiben, welches feststellt, ob eine Eingabe diese Form hat.

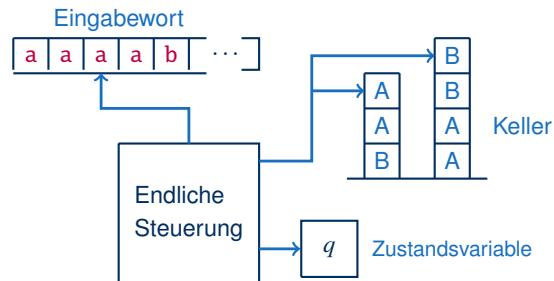
Aber: Praktische Programmiersprachen eignen sich schlecht als allgemeine Berechnungsmodelle, da sie viel zu kompliziert sind.

~ Wir wollen lieber unser Automatenmodell erweitern.

Jenseits PDAs (1): Mehr Stapel

Die Haupteinschränkung von Kellerautomaten war das eingeschränkte Speichermodell. Wie könnte man das erweitern?

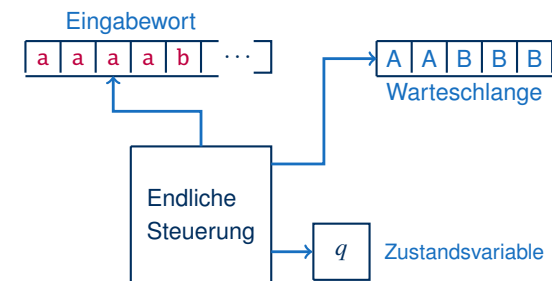
- Man könnte statt eines Stapelspeichers zwei (oder mehr) Stapel verwenden
- Automatenübergänge werden zum Zugriff auf weitere Stapel entsprechend erweitert



Jenseits PDAs (2): „Warteschlangenautomaten“

Die Haupteinschränkung von Kellerautomaten war das eingeschränkte Speichermodell. Wie könnte man das erweitern?

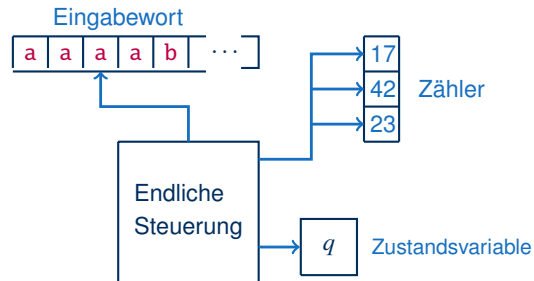
- Man könnte statt eines Stapelspeichers eine Warteschlange (Queue) verwenden
~ first-in/first-out (FIFO)
- Definition fast genau wie bei PDAs, aber mit enqueue/dequeue statt push/pop



Jenseits PDAs (3): Zählerautomaten

Die Haupteinschränkung von Kellerautomaten war das eingeschränkte Speichermodell. Wie könnte man das erweitern?

- Man könnte statt eines Stapelspeichers (endlich viele) Speicherplätze für natürliche Zahlen einführen.
- Automatenübergänge könnten einzelne Variablen inkrementieren, dekrementieren, auf Gleichheit mit 0 testen, ...



Viele mögliche Wege

Bisher gesammelte Ideen:

- PDAs mit zwei Stapeln
- PDAs mit einer noch größeren Zahl an Stapeln
- Warteschlangenautomaten
- Zählerautomaten
- WHILE-Programme
- GOTO-Programme

Man kann jedes dieser Berechnungsmodelle formal definieren ...

Es ergeben sich daher viele Fragen ...

Welche Sprachklasse können diese Modelle jeweils erkennen?

... aber immer wieder die gleiche Antwort:

Genau die Typ-0-Sprachen.

Jenseits PDAs (4): Programme statt Automaten

Eventuell könnte man auch vom Automatenmodell abweichen und stattdessen eine einfache Programmiersprache definieren.

Einfache Ausdrucksmittel:

- **Variablen**, die Zahlen speichern können
- **Wertezuweisungen**, die Variablen das Ergebnis eines Ausdrucks (z.B. aus +, -, Variablen, Zahlen) zuweisen
- **Schleifen** der Form **while** $x \neq 0$ **do**: ...

~> Sogenannte **WHILE-Programme**

Statt **while** könnte man auch **if** und **goto** einführen:

~> sogenannte **GOTO-Programme**.

Zusammenfassung und Ausblick

Deterministische Typ-2-Sprachen sind praktisch wichtig, da effizient parsebar.

Viele **Fragestellungen für Typ-2-Sprachen** sind unentscheidbar, wobei deterministische Sprachen noch etwas mehr erlauben.

Zahlreiche naheliegende Erweiterungen von PDAs führen alle zur gleichen Ausdrucksstärke (Typ 0).

Offene Fragen:

- Welches Berechnungsmodell sollen wir nun verwenden?
- Wenn alle PDA-Verallgemeinerungen Typ 0 liefern, was ist dann mit Typ 1?
- Unterscheiden sich Typ 0 und Typ 1 überhaupt?