Diplomarbeit

Unäre stochastische Automaten

Februar 2002

Verfasser: Gregor Gramlich

Betreuer: Prof. Dr. Georg Schnitger

Johann Wolfgang Goethe–Universtät

Frankfurt am Main

Fachbereich Biologie und Informatik

Institut für Informatik

Professur für Theoretische Informatik

Erklärung

Hiermit versichere ich, die vorliegende Diplomarbeit *Unäre stochastische Automaten* selbständig verfasst und keine anderen Hilfsmittel als die angegebenen Quellen verwendet zu haben.

Offenbach, den 11. Februar 2002

Gregor Gramlich

Dank

Vielen Menschen ist es gelungen, mein Interesse für die Mathematik und Informatik zu wecken. Stellvertretend sei hier meinen Gymnasiallehrern Herrn Claas und Herrn Schotte gedankt.

Die ursprüngliche Fragestellung von Herrn Schnitger, die meinen ersten Forschungen zu dieser Arbeit zu Grunde lag, habe ich etwas zwangloser behandelt, was schließlich zu einer besseren Antwort geführt hat, als zunächst erhofft. Herrn Schnitger möchte ich insbesondere für die Freiheiten danken, die er mir bei der Bearbeitung des Themas gelassen hat, aber auch für viele Verbesserungsvorschläge und die sehr konkreten Ideen zu Abschnitt 8.

Last but not least gilt der Dank meinen Eltern, die mir die finanzielle Sicherheit für mein Studium geben konnten und meiner freien Entfaltung als Mensch und Informatiker nie im Wege standen.

Zusammenfassung

Wir untersuchen das Verhalten von unären stochastischen endlichen Automaten mit Hilfe von Methoden der Theorie der homogenen Markovketten.

Für unäre stochastische Automaten mit ϵ-isoliertem Cutpoint λ und n Zuständen bestimmen wir eine obere Schranke für die Größe des zyklischen Teils eines optimalen äquivalenten DFAs.

Ein Ergebnis von Milani und Pighizzini zeigt bereits, dass für den zyklischen Teil des äquivalenten DFAs $O(e^{\sqrt{n \ln n}})$ Zustände ausreichen und in unendlich vielen Fällen auch $\Omega(e^{\sqrt{n \ln n}})$ Zustände benötigt werden, wobei die Größe von ϵ keine Rolle spielt.

Wir zeigen die obere Schranke $n^{\frac{1}{2\epsilon}}$ für die Größe des zyklischen Teils und weisen nach, dass der optimale DFA für jedes $c < 1$ in unendlich vielen Fällen mehr als $n^{c\frac{1}{2\epsilon}}$ viele Zustände im zyklischen Teil benötigt.

Wir weisen auch nach, dass es eine unendliche Familie endlicher unärer Sprachen gibt, für die es jeweils einen PFA mit n Zuständen und $\frac{1}{4}$-isoliertem Cutpoint gibt, während der optimale DFA $e^{\Omega(\sqrt{n \ln n})}$ Zustände im Anfangspfad benötigt.

Inhaltsverzeichnis

1 Einführung **1**

2 Notationen **3**

3 Formale Sprachen und Automaten **4**

3.1 Formale Sprachen 4

3.2 Deterministische endliche Automaten und reguläre Sprachen 5

3.2.1 Die Nerode-Relation 6

3.3 Nichdeterministische endliche Automaten 9

3.4 Stochastische endliche Automaten 10

3.4.1 Automaten mit isoliertem Cutpoint 13

3.5 Unäre reguläre Sprachen 15

4 Bekannte Ergebnisse für reguläre Sprachen **17**

4.1 Satz von Rabin 17

4.1.1 Eine untere Schranke von Ambainis 22

4.2 Der unäre Fall 23

4.2.1 Tradeoff zwischen NFA und DFA für unäre Sprachen 23

4.2.2 Tradeoff zwischen PFA mit isoliertem Cutpoint und DFA für unäre Sprachen 23

4.2.3 Weitere Literatur zu unären Sprachen 24

5 Bekannte Ergebnisse für homogene Markovketten **25**

5.1 Grundlagen der Markovkettentheorie 25

5.2 Absorbierende Markovketten 34

5.3 Ergodische Markovketten 37

5.4 Allgemeine Markovketten 42

6 Markovketten und unäre stochastische Automaten **47**

6.1 Die Arbeit von Milani und Pighizzini 47

6.2 Die Arbeit von Mereghetti, Palano und Pighizzini 49

7 Eine obere Schranke für die Periodizität, die von ϵ abhängt **50**

7.1 Die Bedeutung der ϵ-Lücke 50

7.1.1 Eine stärkere Vermutung 57

7.2 Obere Schranke für den äquivalenten DFA 57

7.3 Untere Schranke für den äquivalenten DFA 57

8 Eine untere Schranke für den Anfangspfad **60**

8.1 Ein PFA für die Sprache $\{1^m\}$ 60
8.2 Exponentieller Blowup bei konstanter Isolation 62

9 Folgerungen und offene Fragen **63**
9.1 Fazit 63
9.2 Eine obere Schranke für die Größe des Anfangspfads des äquivalenten DFA . . 63
9.3 Sequentiell unäre Sprachen 63

1 Einführung

Wir untersuchen in dieser Diplomarbeit das Verhalten formaler Sprachen, die über einem Alphabet mit nur einem Buchstaben definiert sind. Wir nennen diese Sprachen unär. Unser Augenmerk liegt bei diesen Untersuchungen auf den endlichen Automaten, die unäre reguläre Sprachen akzeptieren. Besonders interessieren wir uns für die Anzahl der notwendigen Zustände bei stochastischen endlichen Automaten und deterministischen endlichen Automaten, die unäre Sprachen akzeptieren.

Die grundlegenden Arbeiten über unäre reguläre Sprachen stammen von Chrobak [Chr86] und von Milani und Pighizzini [MP00].

Bei unären regulären Sprachen bestimmt nur die Länge eines Wortes, ob dieses Wort Element der Sprache ist. Jede unendliche unäre reguläre Sprache hat eine ultimative Periode χ, was bedeutet, dass bis auf eine endliche Zahl von Ausnahmen ein Wort der Länge m genau dann in der Sprache ist, wenn auch das Wort der Länge $m+\chi$ in der Sprache ist. Ein deterministischer endlicher Automat, der eine unendliche unäre reguläre Sprache akzeptiert, besteht aus einem Anfangspfad und einem anschließenden Zyklus.

Die Arbeit von Milani und Pighizzini nutzt ein Ergebnis, das aus der Markovketten-Forschung bekannt ist [Gan86] und besagt, dass auch die Potenzen von stochastischen Matrizen einer gewissen Periodizität unterworfen sind. Diese Aussage übertragen Milani und Pighizzini auf unäre stochastische endliche Automaten mit isoliertem Cutpoint und erhalten somit eine obere Schranke für die Größe der ultimativen Periode der Sprache, die nur von der Anzahl der Zustände des stochastischen Automaten abhängt.

Aus der für stochastische endliche Automaten fundamentalen Arbeit von Rabin [Rab63] ist bereits eine obere Schranke für die Zahl der Zustände eines äquivalenten deterministischen Automaten bekannt, die sowohl von der Zustandszahl des stochastischen Automaten, als auch von der Größe seiner Isolation abhängt. Dies legt die Frage nahe, ob eine solche Schranke für die ultimative Periode unärer Sprachen nicht auch von diesen beiden Parametern abhängig gemacht werden kann.

Die Periode der Markovketten und somit auch die ultimative Periode der regulären Sprache setzt sich als kleinstes gemeinsames Vielfaches aus verschieden Teilperioden zusammen. Anhand der Beispiele in [MP00] und nach intensiver Studie der Markovketten-Theorie ergab sich die Vermutung, dass die Anzahl der Teilperioden durch den Kehrwert der Größe der Isolation im stochastischen Automaten beschränkt wird. Diese Vermutung bestätigte sich und brachte uns zu einer neuen oberen Schranke für die ultimative Periode der Sprache, die von der Anzahl der Zustände des stochastischen Automaten und von der Größe seiner Isolation abhängt. Für eine feste Größe der Isolation ist diese Schranke polynomiell in der Zahl der Zustände und außerdem auch scharf, da jedes asymptotisch kleinere Polynom in unendlich vielen Fällen übertroffen werden muss.

Wir führen formale Sprachen und die verschiedenen Modelle endlicher Automaten in Abschnitt 3 ein. Mit Ausnahme der stochastischen Automaten sind diese Themen aus dem Grundstudium bekannt. Der Unterabschnitt über stochastische endliche Automaten geht auch ausführlich auf die Klasse der Cutpoint-Sprachen ein.

Abschnitt 4 beschreibt den Satz von Rabin. Dieser stellt die bekannteste obere Schranke für die Größe des minimalen deterministischen endlichen Automaten, der zu einem stochastischen Automaten äquivalent ist, dar. In diesem Abschnitt wird ebenfalls eine untere Schranke von Ambainis zitiert, die jedoch weit unter der oberen Schranke von Rabin liegt. Außerdem geben wir einen Ausblick auf die entsprechenden Ergebnisse, die sich auf unäre Sprachen beschränken und zitieren zwei weitere Arbeiten über unäre Sprachen und endliche Automaten.

In Abschnitt 5 werden sehr ausführlich die Grundlagen zum Verständnis der Markovketten und ihrer Periodizität gelegt. Besonders wurde hier auf die Anschaulichkeit Wert gelegt, die es auch erlaubt, die Ergebnisse ohne tiefes Studium der Beweise nachzuvollziehen. Einige Methoden wurden zwar in der Literatur schon angedeutet, mussten aber für unsere Ziele noch weiter ausgebaut werden.

Die Arbeiten von Milani und Pighizzini, sowie von Mereghetti, Palano und Pighizzini werden in Abschnitt 6 zitiert und mit den Ergebnissen aus Abschnitt 5 in Verbindung gebracht.

Nach diesen Vorarbeiten kommen wir in Abschnitt 7 zum Hauptergebnis der Arbeit. Die obere Schranke für die Periodizität der unären Sprache, die von einem stochastischen endlichen Automaten mit isoliertem Cutpoint akzeptiert wird, wird hergeleitet. Anschließend zeigen wir auch, dass diese Schranke scharf ist.

Uns ist es auch gelungen, eine untere Schranke für die Vergrößerung der Zustandszahl im Anfangspfad eines DFA machzuweisen, der einen unären stochastischen endlichen Automaten mit $\frac{1}{4}$-isoliertem Cutpoint simuliert. Dieses Ergebnis wird in Abschnitt 8 behandelt.

Abschnitt 9 geht auf Fragen ein, die die Arbeit zwar offen lässt, für deren Beantwortung aber einige Grundlagen aus dieser Arbeit nützlich sein könnten.

2 Notationen

Zunächst einige Notationen, die im Folgenden ohne weitere Erklärung benutzt werden.

ln	Natürlicher Logarithmus
log	Logarithmus zur Basis 2
$\subset$	Echte Teilmenge
$\subseteq$	Teilmenge
2^A	Potenzmenge der Menge A: $B \in 2^A \Leftrightarrow B \subseteq A$
$\lvert A\rvert$	Mächtigkeit einer Menge A = Anzahl ihrer Elemente
I_n	Einheitsmatrix der Größe n
$\times$	Matrixmultiplikation
$\forall$	Allquantor
$\exists$	Existenzquantor
$\wedge$	Logisches UND
$\vee$	Logisches ODER
$\neg$	Logisches NICHT
$\mathbb{N}$	Die natürlichen Zahlen $\mathbb{N} = \{1, 2, \ldots\}$
$\mathbb{N}_0$	$\mathbb{N}_0 := \mathbb{N} \cup \{0\}$
$O(g(n))$	$f(n) = O(g(n)) \Leftrightarrow \exists n_0, c > 0 : \forall n \geq n_0 : 0 \leq f(n) \leq cg(n)$
$\Omega(g(n))$	$f(n) = \Omega(g(n)) \Leftrightarrow \exists n_0, c > 0 : \forall n \geq n_0 : 0 \leq cg(n) \leq f(n)$
$\Theta(g(n))$	$f(n) = \Theta(g(n)) \Leftrightarrow f(n) = O(g(n)) \wedge f(n) = \Omega(g(n))$

3 Formale Sprachen und Automaten

Wir beschäftigen uns in dieser Arbeit mit regulären Sprachen und den Automatenmodellen mit denen diese Sprachen akzeptiert werden können. Wir können eine formale Sprache als eine Menge von Wörtern über einem bestimmten Alphabet auffassen. Eine formale Sprache kann auch unendlich viele Wörter enthalten. Oft werden formale Sprachen im Zusammenhang mit Entscheidungsproblemen verwandt. Bekommt ein Programm eine Eingabe, dann muss es entscheiden, ob es diese Eingabe akzeptiert oder verwirft. Die Menge aller akzeptierten Eingaben ist dann eine formale Sprache. In diesem Sinne können wir eine formale Sprache auch als Funktion betrachten, die ein Wort auf die Ausgabe 1 oder 0 abbildet, je nachdem, ob das Wort Element der Sprache ist, oder nicht. Die Sprache selbst ist dann wiederum das volle Urbild der Menge $\{1\}$.
Automaten sind Formalisierungen der Programme, die entscheiden, ob ein Wort Element der Sprache ist. Es gibt verschieden mächtige Automatenmodelle, die verschiedene Sprachklassen akzeptieren. Wir beschäftigen uns mit verschiedenen Sorten von endlichen Automaten, die in diesem Abschnitt definiert werden. Bis auf die Ausnahme allgemeiner stochastischer endlicher Automaten akzeptieren alle die gleiche Sprachklasse, nämlich die regulären Sprachen.

3.1 Formale Sprachen

Wir benötigen zunächst die kleinsten Bausteine eines Wortes, die Buchstaben. Reiht man diese aneinander, ergibt sich ein Wort. Alle zur Verfügung stehenden Buchstaben fassen wir zu einer Menge zusammen, dem Alphabet.

Definition 3.1 *Alphabet und Wort*

Ein Alphabet Σ ist eine endliche Menge von Buchstaben.
Ein Wort w über einem Alphabet Σ ist die endliche Konkatenation von Buchstaben aus Σ.
$w = a_1 a_2 \dots a_n$ mit $a_i \in \Sigma$ für $1 \leq i \leq n$.

Wir können angeben, wie lang ein Wort ist, indem wir die Anzahl seiner Buchstaben zählen und definieren eine Bezeichnung für die Menge aller Wörter derselben Länge.

Definition 3.2 *Die Länge eines Wortes, Σ^n*

Falls $w = a_1 a_2 \dots a_n$ mit $a_i \in \Sigma$, dann hat w die Länge $|w| = n$.
Für $n \in I\!N$ ist $\Sigma^n := \{a_1 a_2 \dots a_n | a_i \in \Sigma\}$ die Menge aller Wörter über Σ, der Länge n.

In einigen Fällen benötigen wir auch das Wort, das die Länge 0 hat.

Definition 3.3 *Das leere Wort ε*

Für jedes Alphabet Σ bezeichnen wir das leere Wort mit ε. Das leere Wort hat die Länge 0.
Wir definieren $\Sigma^0 := \{\varepsilon\}$.

Um eine formale Sprache L als Teilmenge der Menge aller Wörter über einem Alphabet Σ bezeichnen zu können, ist folgende Definition nützlich.

Definition 3.4 Σ^+ *und* Σ^*

$$\Sigma^+ := \bigcup_{n \in \mathbb{N}} \Sigma^n$$

ist die Menge aller Wörter über Σ, die mindestens Länge 1 haben.

$$\Sigma^* := \bigcup_{n \in \mathbb{N}_0} \Sigma^n = \Sigma^+ \cup \{\varepsilon\}$$

ist die Menge aller Wörter über Σ.

Jetzt haben wir alle Grundlagen für die Definition der formalen Sprache geschaffen.

Definition 3.5 *Formale Sprache L*

Wir nennen eine Teilmenge $L \subseteq \Sigma^*$ eine formale Sprache über dem Alphabet Σ.

3.2 Deterministische endliche Automaten und reguläre Sprachen

Das Modell der deterministischen endlichen Automaten ist eines der einfachsten und eingeschränktesten Automatenmodelle, die in der Informatik betrachtet werden. Gerade diese Einfachheit machen die endlichen Automaten zu einem gut erforschbaren und erforschten Modell, auf dem auch andere Automaten aufbauen.

Definition 3.6 *Deterministischer endlicher Automat*

Ein deterministischer endlicher Automat (**DFA**) ist ein 5-Tupel $M = (Q, \Sigma, \delta, q_0, F)$.
Q ist die endliche Zustandsmenge, Σ das Alphabet, δ die Übergangsfunktion, $q_0 \in Q$ der Startzustand und $F \subseteq Q$ die Menge der akzeptierenden Zustände.
Die Übergangsfunktion δ bildet einen Zustand $q \in Q$ und einen Buchstaben $a \in \Sigma$ auf einen Zustand $\delta(q, a) = p \in Q$ ab. δ ist auf allen $(q, a) \in Q \times \Sigma$ definiert.
Die Übergangsfunktion wird rekursiv erweitert, so dass auch Wörter $w \in \Sigma^+$ verarbeitet werden können. Für $u \in \Sigma^*$ und $a \in \Sigma$ definieren wir $\delta(q, ua) := \delta(\delta(q, u), a)$. Außerdem gelte $\delta(q, \varepsilon) := q$.
Ein Wort $w \in \Sigma^*$ wird vom Automaten M genau dann akzeptiert, wenn $\delta(q_0, w) \in F$.
Die vom Automaten M akzeptierte Sprache ist $L(M) := \{w | w \in \Sigma^* \wedge \delta(q_0, w) \in F\}$.

Definition 3.7 *Äquivalente Automaten*
Wir nennen zwei Automaten M und M' äquivalent, wenn $L(M) = L(M')$ gilt.

Wir wollen in der Lage sein, die Größe verschiedener Automaten miteinander zu vergleichen.

Definition 3.8 *Größe eines endlichen Automaten*
Die Größe eines endlichen Automaten, wird durch die Anzahl der Zustände $|Q|$ gemessen.

Definition 3.9 *Reguläre Sprachen*
Die Klasse der von deterministischen endlichen Automaten akzeptierten Sprachen wird als Klasse der regulären Sprachen (**REG**) bezeichnet.

Definition 3.10 *Zustandsdiagramm eines DFA*
Das Zustandsdiagramm eines DFA ist ein gerichteter Graph (Q, E) mit Kantenbeschriftung. Die Knoten in Q entsprechen den Zuständen des Automaten. Eine Kante (p, q) existiert genau dann, wenn es ein $\sigma \in \Sigma$ gibt, so dass $\delta(p, \sigma) = q$. Eine Kante $(p, q) \in E$ wird mit allen Buchstaben der Menge $\{\sigma \in \Sigma | \delta(p, \sigma) = q\}$ beschriftet.

Wenn wir das Zustandsdiagramm eines DFA zeichnen, dann markieren wir die akzeptierenden Zustände durch einen doppelten Kreis. Den Startzustand markieren wir durch eine eingehende Kante ohne Ursprung.

Im Zustandsdiagramm eines DFA gibt es für einen Zustand $q \in Q$ und einen Buchstaben $\sigma \in \Sigma$ genau eine von q ausgehende Kante mit Beschriftung σ, da $\delta(q, \sigma)$ für alle $(q, \sigma) \in Q \times \Sigma$ definiert ist.

Beispiel: Sei M der DFA, der die Wörter $w \in \{0, 1, 2\}^*$ mit einer geraden Anzahl von 0en akzeptiert. Dann ist $M = (\{q_0, q_1\}, \{0, 1, 2\}, \delta, q_0, \{q_0\})$ mit

$$\delta(q_i, \sigma) = \begin{cases} q_i & \sigma \in \{1, 2\} \\ q_{1-i} & \sigma = 0 \end{cases} .$$

Das Zustandsdiagramm von M ist in Abbildung 1 dargestellt.

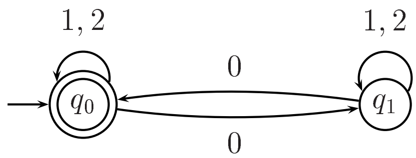

Abbildung 1: Das Zustandsdiagramm eines DFA

3.2.1 Die Nerode-Relation

Wir lösen uns zunächst ein wenig vom Modell der endlichen Automaten und betrachten Äquivalenzrelationen, die auf Wörtern definiert sind. Über die Nerode-Relation finden wir dann jedoch sehr schnell wieder den Weg zurück zu den regulären Sprachen und den DFAs.

Definition 3.11 *Rechtsinvariante Äquivalenzrelation*
Ein Äquivalenzrelation R auf Σ^* heißt rechtsinvariant genau dann, wenn

$$x\,R\,y \quad \Rightarrow \quad \forall z \in \Sigma^* : xz\,R\,yz.$$

Falls $x\,R\,y$ gilt, dann sagen wir, dass x und y in Relation bezüglich R stehen.

Definition 3.12 *Äquivalenzklasse*
Sei R eine Äquivalenzrelation auf Σ^*. Für $x \in \Sigma^*$ ist $[x]_R = \{w | w \in \Sigma^* \wedge w\, R\, x\}$ die Äquivalenzklasse bezüglich R mit Repräsentant x.

Definition 3.13 *Index einer Äquivalenzrelation*
Die Anzahl der Äquivalenzklassen bezüglich einer Äquivalenzrelation R bezeichnen wir als den Index $\text{ind}(R)$ von R.

Definition 3.14 *Nerode-Relation*
Für eine Sprache L wird die Nerode-Relation R_L folgendermaßen definiert:

$$x\, R_L\, y \quad :\Longleftrightarrow \quad (\forall z \in \Sigma^* : xz \in L \Leftrightarrow yz \in L).$$

Bemerkung: Die Nerode-Relation R_L ist rechtsinvariant.

Der Satz von Nerode stellt den Bezug zwischen regulären Sprachen und der Endlichkeit der Anzahl der Äquivalenzklassen her.

Satz 3.15 *Satz von Nerode*
Die folgenden Aussagen sind äquivalent.

(1) $L \subseteq \Sigma^*$ ist regulär.

(2) L ist die Vereinigung von einigen Äquivalenzklassen einer rechtsinvarianten Äquivalenzrelation von endlichem Index.

(3) Die Nerode-Relation R_L hat endlichen Index.

Beweis:

- (1) $\Rightarrow$ (2): Sei L regulär. Dann gibt es einen DFA $M = (Q, \Sigma, \delta, q_0, F)$. Sei R_M die Relation, für die gilt:

  $$x\, R_M\, y \quad :\Longleftrightarrow \quad \delta(q_0, x) = \delta(q_0, y).$$

  Dann ist R_M eine Äquivalenzrelation. R_M ist sogar rechtsinvariant:
  $x\, R_M\, y \;\Rightarrow\; \delta(q_0, x) = \delta(q_0, y) \;\Rightarrow\; \forall z \in \Sigma^* : \delta(q_0, xz) = \delta(q_0, yz) \;\Rightarrow\; \forall z \in \Sigma^* : xz\, R\, yz$

  Der Index von R_M ist endlich, da die Zahl der Äquivalenzklassen durch die Zahl der Zustände des Automaten M beschränkt ist.

  Behauptung:

  $$L = \bigcup_{x:\delta(q_0,x)\in F} [x]_{R_M}$$

  Beweis der Behauptung: Sei $w \in L$, dann gilt $\delta(q_0, w) \in F$ und somit gilt auch $w \in [w]_{R_M}$. Sei andererseits $w \in \bigcup_{x:\delta(q_0,x)\in F} [x]_{R_M}$, dann gibt es ein $x \in \Sigma^*$ mit $\delta(q_0, x) \in F$ und $w\, R_M\, x \;\Rightarrow\; \delta(q_0, w) \in F$, also gilt auch $w \in L$. ■

- (2) $\Rightarrow$ (3): Sei R eine Äquivalenzrelation mit den Eigenschaften aus (2). Behauptung:

$$x\, R\, y \;\Rightarrow\; x\, R_L\, y$$

Beweis der Behauptung: Sei $z \in \Sigma^*$, dann ist

$$\begin{aligned} x\, R\, y \;&\Rightarrow\; \forall z \in \Sigma^* : xz\, R\, yz \\ &\Rightarrow\; \forall z \in \Sigma^* : (xz \in L \;\Leftrightarrow\; yz \in L) \\ &\Rightarrow\; x\, R_L\, y. \end{aligned}$$

Somit ist R_L eine Vergröberung von R.

$$\begin{aligned} &\Rightarrow\; \forall x \in \Sigma^* : [x]_R \subseteq [x]_{R_L} \\ &\Rightarrow\; \operatorname{ind}(R_L) \leq \operatorname{ind}(R) \end{aligned}$$

Da $\operatorname{ind}(R)$ nach Voraussetzung endlich ist, hat auch die Nerode-Relation endlichen Index.

■

- (3) $\Rightarrow$ (1): Sei R_L die Nerode-Relation mit endlichem Index. Wir entwerfen einen DFA M_N, den Nerode-Automaten, der L akzeptiert. Im Rest des Beweises sei $[x] = [x]_{R_L}$.

Wir definieren $M_N = (Q_N, \Sigma, \delta_N, [\varepsilon], F_N)$, wobei Q_N die Menge der Äquivalenzklassen bezüglich R_L ist und die akzeptierenden Zustände durch $F_N := \{[x] | x \in L\}$, sowie die Übergangsfunktion durch $\delta_N([x], a) := [xa]$ definiert sind. Für die Wohldefiniertheit von F_N und δ_N müssen wir zeigen, dass F_N und δ_N vom gewählten Repräsentanten unabhängig sind.

Sei $[x] = [y] \;\Rightarrow\; (x \in L \Leftrightarrow y \in L)$ und F_N ist wohldefiniert.

Sei $[x] = [y] \;\Rightarrow\; x\, R_L\, y \;\Rightarrow\; xa\, R_L\, ya \;\Rightarrow\; [xa] = [ya] \;\Rightarrow\; \delta_N([x], a) = \delta_N([y], a)$ und auch δ_N ist wohldefiniert.

Bleibt noch zu zeigen, dass $L = L(M_N)$.

$$\begin{aligned} w \in L(M_N) \;&\Leftrightarrow\; \delta_N([\varepsilon], w) \in F_N \\ &\Leftrightarrow\; [\varepsilon w] = [w] \in F_N \\ &\Leftrightarrow\; w \in L \end{aligned}$$

Es gibt also einen DFA, der L akzeptiert. Somit ist L regulär. ■

Bemerkung: Der Nerode-Automat für L ist der minimale DFA, der L akzeptiert. Dies folgt aus dem Beweis von Satz 3.15, da wir aus jedem DFA M', der L akzeptiert, mit Hilfe des Beweisschritts (1) $\Rightarrow$ (2) eine Äquivalenzrelation $R_{M'}$ mit $\operatorname{ind}(R_{M'}) \leq Q'$ konstruieren können. Im nächsten Beweisschritt (2) $\Rightarrow$ (3) weisen wir nach, dass die Nerode-Relation R_L gröber ist als $R_{M'}$ und für die Zustandszahl des Nerode-Automaten M_N gilt schließlich $|Q_N| = \operatorname{ind}(R_L) \leq \operatorname{ind}(R_{M'}) \leq Q'$.

3.3 Nichdeterministische endliche Automaten

Neben den deterministischen Automaten spielen auch die nichtdeterministischen eine wichtige Rolle. Befindet sich der nichtdeterministische Automat in einem Zustand und liest einen Buchstaben des Wortes, dann hat er mehrere mögliche Folgeszustände. Ein Wort wird akzeptiert, wenn es die Möglichkeit gibt, nach Lesen des Wortes in einem akzeptierenden Zustand zu landen.

Definition 3.16 *Nichtdeterministischer endlicher Automat*

Ein nichtdeterministischer endlicher Automat (**NFA**) ist ein 5-Tupel $M = (Q, \Sigma, \delta, q_0, F)$.

Q, Σ, q_0 und F haben die gleiche Bedeutung wie in der Definition von DFAs. Allerdings ist das Akzeptanzverhalten anders definiert und δ wird jetzt als Abbildung $Q \times \Sigma \to 2^Q$ verstanden.

Auch hier setzen wir die Übergangsfunktion für Wörter $w \in \Sigma^+$ fort: Es gelte $\delta(q, \varepsilon) := \{q\}$ und für $u \in \Sigma^*$ in Verbindung mit einem Buchstaben $a \in \Sigma$ sei

$$\delta(q, ua) := \bigcup_{p \in \delta(q,u)} \delta(p, a).$$

Ein Wort w wird vom Automaten akzeptiert, wenn es mindestens einen akzeptierenden Zustand gibt, der nach Lesen von w erreicht werden kann. Also ist $w \in L(M) \Leftrightarrow \delta(q_0, w) \cap F \neq \emptyset$.

Das Zustandsdiagramm eines NFA ist ähnlich definiert wie beim DFA. Hier werden jedoch Kanten (p, q) gesetzt, falls es ein $\sigma \in \Sigma$ gibt, so dass $q \in \delta(p, \sigma)$. Eine Kante $(p, q) \in E$ wird mit allen Buchstaben aus $\{\sigma \in \Sigma | q \in \delta(p, \sigma)\}$ beschriftet.

Im Zustandsdiagramm eines NFA kann es somit für einen Zustand $q \in Q$ und einen Buchstaben $\sigma \in \Sigma$ mehrere von q ausgehende Kanten mit Beschriftung σ geben, oder auch gar keine, falls $\delta(q, \sigma) = \emptyset$.

Bemerkung: Die Klasse der von nichtdeterministischen endlichen Automaten akzeptierten Sprachen stimmt mit der Klasse der regulären Sprachen überein.

Bemerkung: Für jede Sprache L, die von einem NFA mit n Zuständen akzeptiert wird, gibt es einen DFA mit $\leq 2^n$ Zuständen. Der DFA kann mit Hilfe der Potenzmengenkonstruktion konstruiert werden. Der Beweis ist nachzulesen in [Weg93].

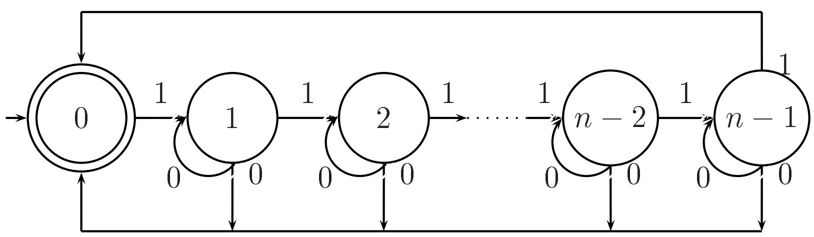

Abbildung 2: Das Zustandsdiagramm des Modulo-Automaten

Bemerkung: Es gibt eine unendliche Folge von regulären Sprachen $(L_n)_{n \in \mathbb{N}}$, für die es jeweils einen NFA mit n vielen Zuständen gibt, während der minimale äquivalente DFA 2^n viele

Zustände benötigt. Um dies zu beweisen, konstruieren wir zunächst den Modulo-Automaten (siehe Abbildung 2), einen NFA mit n Zuständen, und zeigen dann, dass die Potenzmengenkonstruktion bereits den minimalen DFA liefert. Der Beweis wird in [Wot00] dargestellt.

3.4 Stochastische endliche Automaten

Eine weitere Ausprägung endlicher Automaten sind die stochastischen endlichen Automaten. Für den Übergang von einem Zustand zum nächsten bei gelesenem Buchstaben, wird hier eine Wahrscheinlichkeitsverteilung festgelegt. Ein Wort wird akzeptiert, wenn seine Akzeptanzwahrscheinlichkeit einen zusätzlich definierten Schwellwert, den Cutpoint, übersteigt.

Definition 3.17 *Stochastischer endlicher Automat*

Ein stochastischer endlicher Automat oder probabilistischer endlicher Automat (**PFA**) ist ein 5-Tupel $M = (Q, \Sigma, \{A(\sigma)|\sigma \in \Sigma\}, \pi, \eta)$.
Auch hier ist Q die endliche Zustandsmenge mit $|Q| = n$ und Σ das endliche Alphabet. Für jeden Buchstaben $\sigma \in \Sigma$ ist eine stochastische $n \times n$ Übergangsmatrix $A(\sigma)$ definiert, dessen Zeilen und Spalten mit den Zuständen indiziert sind. Es gilt also

$$\begin{aligned} \forall p, q \in Q, \sigma \in \Sigma : A_{p,q}(\sigma) \geq 0 \\ \forall p \in Q, \sigma \in \Sigma : \sum_{q \in Q} A_{p,q}(\sigma) = 1. \end{aligned}$$

π ist ein stochastischer Zeilenvektor, der die Startverteilung repräsentiert und $\eta \in \{0,1\}^n$ ist ein n-dimensionaler Spaltenvektor, der als Inzidenzvektor der Menge der akzeptierenden Zustände angesehen werden kann. Die Komponenten der eben genannten Vektoren sind wieder mit den Zuständen indiziert.

Definition 3.18 *Zustandsdiagramm eines PFA*

Das Zustandsdiagramm eines PFA ist ähnlich wie das eines DFA oder NFA definiert, Hier werden jedoch Kanten (p, q) gesetzt, wenn es einen Buchstaben $\sigma \in \Sigma$ gibt, so dass $A_{p,q}(\sigma) > 0$ ist. Eine Kante $(p, q) \in E$ wird mit allen Buchstaben aus $\{\sigma \in \Sigma | A_{p,q}(\sigma) > 0\}$ und in Klammern mit der entsprechenden Übergangswahrscheinlichkeit $A_{p,q}(\sigma)$ beschriftet.

Beispiele für Zustandsdiagramme von PFAs sind in Abbildungen 3 und 4 zu sehen.

Definition 3.19 *$A(w)$ für $w \in \Sigma^*$*

Sei $M = (Q, \Sigma, \{A(\sigma)|\sigma \in \Sigma\}, \pi, \eta)$ ein PFA mit n vielen Zuständen. Für das leere Wort setzen wir $A(\varepsilon) := I_n$ und für $\sigma_1, \ldots, \sigma_k \in \Sigma$ setzen wir $A(\sigma_1 \cdots \sigma_k) := A(\sigma_1 \cdots \sigma_{k-1}) \times A(\sigma_k)$ mit Hilfe der Matrixmultiplikation.

$$A_{p,q}(\sigma_1 \cdots \sigma_k) = \sum_{r \in Q} A_{p,r}(\sigma_1 \cdots \sigma_{k-1}) \cdot A_{r,q}(\sigma_k)$$

Später werden wir an manchen Stellen zur besseren Lesbarkeit das $\times$-Zeichen bei der Matrixmultiplikation weglassen.

Definition 3.20 *Akzeptanzwahrscheinlichkeit $p_M(w)$ für $w \in \Sigma^*$*

Für einen PFA M mit Übergangswahrscheinlichkeitsmatrix A definieren wir die Akzeptanzwahrscheinlichkeit eines Wortes $w \in \Sigma^*$ durch $p_M(w) := \pi \times A(w) \times \eta$.

Definition 3.21 *Die von einem PFA akzeptierte Sprache*

Sei M ein PFA und $\lambda \in \mathbb{R}$ mit $0 \leq \lambda < 1$. Dann ist

$$T(M, \lambda) := \{w | w \in \Sigma^*, p_M(w) > \lambda\}$$

die Sprache, die vom PFA M mit Cutpoint λ akzeptiert wird.

Die von PFAs akzeptierten Sprachen bilden eine eigene Sprachklasse.

Definition 3.22 *Cutpoint-Sprachen*
Die Klasse der Sprachen L, für die es einen PFA M und einen Cutpoint λ gibt, so dass $L = T(M, \lambda)$, heißt die Klasse der Cutpoint-Sprachen (**CL**).

Die Klasse der Cutpoint-Sprachen umfasst die Klasse der regulären Sprachen.

Satz 3.23 *REG$\subseteq$CL*

Für jede reguläre Sprache L gibt es auch einen PFA M und einen Cutpoint λ, so dass $T(M, \lambda) = L$.

Beweis: Sei $L \in$ REG, dann gibt es einen DFA $M = (Q, \Sigma, \delta, q_0, F)$, der die Sprache L akzeptiert. Wir konstruieren einen PFA $M' = (Q, \Sigma, \{A(\sigma) | \sigma \in \Sigma\}, \pi, \eta)$, der mit einem Cutpoint $\lambda = \frac{1}{2}$ ebenfalls L akzeptiert. Wir wählen

$$\begin{aligned}
\pi_q &:= \begin{cases} 1 & q = q_0 \\ 0 & q \neq q_0 \end{cases} \\
\eta_q &:= \begin{cases} 1 & q \in F \\ 0 & q \notin F \end{cases} \\
A_{p,q}(\sigma) &:= \begin{cases} 1 & \delta(p, \sigma) = q \\ 0 & \text{sonst.} \end{cases}
\end{aligned}$$

Dann gilt für alle $w \in \Sigma^*$:

$$\begin{aligned}
p_{M'}(w) = 1 &\Leftrightarrow w \in L(M) \\
p_{M'}(w) = 0 &\Leftrightarrow w \notin L(M)
\end{aligned}$$

Und so ist $L = T(M', \frac{1}{2})$ und $L \in$ CL. ∎

Andererseits ist die Klasse CL größer als die Klasse REG.

Satz 3.24 *REG≠CL*
Es gibt Cutpoint-Sprachen, die nicht regulär sind.

Für den Beweis werden einige Lemmata benötigt.
Wir untersuchen die Eigenschaften des PFA $M = (Q, \Sigma, \{A(\sigma)|\sigma \in \Sigma\}, \pi, \eta)$, der auch in Abbildung 3 dargestellt ist. Seine Komponenten sind definiert durch

$$
\begin{array}{lll}
Q = \{q_1, q_2\} & \Sigma = \{0, 1\} & \pi = \begin{pmatrix} 1 & 0 \end{pmatrix} \\
A(0) = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} & A(1) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix} & \eta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.
\end{array}
$$

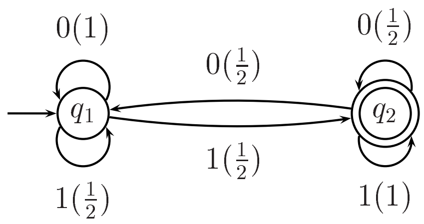

Abbildung 3: Ein PFA für den Beweis, dass REG≠CL

Lemma 3.25 *Akzeptanzwahrscheinlichkeit eines Wortes $\sigma_1\sigma_2\ldots\sigma_k$*
Im oben beschriebenen PFA M ist für jedes $k \geq 1$ und alle $\sigma_1, \ldots, \sigma_k \in \Sigma$ die Akzeptanzwahrscheinlichkeit $p_M(\sigma_1\sigma_2\ldots\sigma_k) = 0.\sigma_k\ldots\sigma_2\sigma_1$ in Binärnotation.

Lemma 3.26 *$T(M, \lambda)$ für verschiedene λ*
Sei M der PFA aus dem obigen Lemma, dann gilt für alle $\lambda_1, \lambda_2 \in [0, 1)$:

$$T(M, \lambda_1) \neq T(M, \lambda_2) \Leftrightarrow \lambda_1 \neq \lambda_2$$

Lemma 3.27 *Überabzählbarkeit der Klasse $T(M, \lambda)$*
Sei M der PFA aus dem obigen Lemma. Dann ist die Sprachklasse $\{T(M, \lambda)|\lambda \in [0, 1)\}$ überabzählbar.

Beweis von Lemma 3.25: Wir führen den Beweis per Induktion. Für $k = 1$:

$$
\begin{array}{lcl}
\sigma_1 = 0 & : & p_M(0) = 0.0 \\
\sigma_1 = 1 & : & p_M(1) = 0.1
\end{array}
$$

Für ein $k \geq 1$ gelte nun $p_M(\sigma_1\sigma_2\ldots\sigma_k) = 0.\sigma_k\ldots\sigma_2\sigma_1$. Dann gilt für $\sigma_{k+1} = 0$:

$$
\begin{aligned}
p_M(\sigma_1\ldots\sigma_k\sigma_{k+1}) &= p_M(\sigma_1\ldots\sigma_k)\cdot\frac{1}{2} + (1 - p_M(\sigma_1\ldots\sigma_k))\cdot 0 \\
&= p_M(\sigma_1\ldots\sigma_k)\cdot\frac{1}{2} \\
&= 0.0\sigma_k\ldots\sigma_1
\end{aligned}
$$

Andererseits gilt für $\sigma_{k+1} = 1$:

$$\begin{aligned} p_M(\sigma_1 \dots \sigma_k \sigma_{k+1}) &= p_M(\sigma_1 \dots \sigma_k) \cdot 1 + (1 - p_M(\sigma_1 \dots \sigma_k)) \cdot \frac{1}{2} \\ &= \frac{1}{2} + p_M(\sigma_1 \dots \sigma_k) \cdot \frac{1}{2} \\ &= 0.1\sigma_k \dots \sigma_1 \end{aligned}$$

Und in beiden Fällen gilt $p_M(\sigma_1 \dots \sigma_k \sigma_{k+1}) = 0.\sigma_{k+1}\sigma_k \dots \sigma_1$. ■

Beweis von Lemma 3.26: Seien $\lambda_1 \neq \lambda_2$ im geforderten Intervall und o.B.d.A. gelte $\lambda_1 < \lambda_2$. Dann gibt es ein $s \in I\!N$, so dass λ_1 und λ_2 sich in (nicht notwendig abbrechender) Binärnotation so darstellen lassen:

$$\begin{aligned} \lambda_1 &= 0.\sigma_1\sigma_2 \dots \sigma_s 0 \dots \\ \lambda_2 &= 0.\sigma_1\sigma_2 \dots \sigma_s 1 \dots \end{aligned}$$

Und für die Zahl $x := 0.\sigma_1\sigma_2 \dots \sigma_s 1$ gilt: $\lambda_1 < x \leq \lambda_2$. Also gilt für das Wort $w := 1\sigma_s \dots \sigma_2\sigma_1$: $w \in T(M, \lambda_1)$, aber $w \notin T(M, \lambda_2)$ und $T(M, \lambda_1) \neq T(M, \lambda_2)$ folgt. ■

Beweis von Lemma 3.27: Es gibt überabzählbar viele $\lambda \in I\!R$ mit $\lambda \in [0, 1)$ und somit gibt es auch überabzählbar viele Sprachen $T(M, \lambda)$. Da die Klasse der Cutpoint-Sprachen CL alle eben genannten Sprachen umfasst, ist auch CL überabzählbar. ■

Beweis von Satz 3.24: Es gibt nur abzählbar viele reguläre Sprachen, da die Menge der DFAs abzählbar ist. Es gibt aber überabzählbar viele Cutpoint-Sprachen und somit auch überabzählbar viele Cutpoint-Sprachen, die nicht regulär sind. Somit ist REG$\neq$CL und zusammen mit Satz 3.23 ergibt sich REG$\subset$CL. ■

3.4.1 Automaten mit isoliertem Cutpoint

Definition 3.28 *Isolierter Cutpoint*
Sei $M = (Q, \Sigma, \{A(\sigma)|\sigma \in \Sigma\}, \pi, \eta)$ ein PFA und sei $\epsilon > 0$. Ein Cutpoint λ heißt ϵ-isoliert bezüglich M genau dann, wenn für jedes Wort $w \in \Sigma^*$ gilt: $|p_M(w) - \lambda| \geq \epsilon$.
Wir nennen ein Cutpoint λ isoliert, wenn es ein $\epsilon > 0$ gibt, so dass λ ein ϵ-isolierter Cutpoint ist.

Rabin [Rab63] hat gezeigt, dass die Klasse der von einem stochastischen endlichen Automaten mit isoliertem Cutpoint erkannten Sprachen mit der Klasse der regulären Sprachen übereinstimmt.
Der Satz von Rabin (Satz 4.1) wird ausführlich im Abschnitt 4.1 behandelt und sei hier schon einmal zitiert.

> Sei $M = (Q, \Sigma, \{A(\sigma)|\sigma \in \Sigma\}, \pi, \eta)$ ein PFA mit $n := |Q|$. λ sei ein ϵ-isolierter Cutpoint bezüglich M. Dann ist $T(M, \lambda)$ regulär und es gibt einen DFA mit höchstens $(1 + \frac{1}{2\epsilon})^{n-1}$ Zuständen, der diese Sprache erkennt.

Offensichtlich spielt die Größe der Isolation ϵ eine entscheidende Rolle für die Größe des minimalen äquivalenten DFA.

Ein einfaches Beispiel zeigt bereits, dass man für allgemeine PFA mit isoliertem Cutpoint und n Zuständen keine obere Schranke für die Größe eines äquivalenten DFA angeben kann, die nur von n abhängt.

Definition 3.29 *Die Sprachfamilie C_m*
Für $m \in I\!N$ sei $C_m \subseteq \{a\}^*$ die Sprache, die alle Wörter der Länge höchstens m enthält.

$$C_m := \{w \in \{a\}^* \mid |w| \leq m\}$$

Satz 3.30 *Ein PFA für C_m*
Für jedes $m \in I\!N$ gibt es einen PFA mit 2 Zuständen und isoliertem Cutpoint $\frac{1}{2}$, der die Sprache C_m akzeptiert.

Beweis: Sei

$$M = \left(\{q_1, q_2\}, \{a\}, \left\{\begin{pmatrix} x & 1-x \\ 0 & 1 \end{pmatrix}\right\}, \begin{pmatrix} 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right)$$

der PFA mit dem abgebildeten Zustandsdiagramm (Abbildung 4).

x sei so gewählt, dass $2^{-\frac{1}{m}} < x < 2^{-\frac{1}{m+1}}$.

Dann gilt für ein Wort $w = a^i$ mit $i \in I\!N_0$: $p_M(w) = x^i$, offensichtlich ist $x^i > \frac{1}{2} \Leftrightarrow 2^{-\frac{1}{i}} < x$. Für den Cutpoint $\lambda = \frac{1}{2}$ heißt das, dass ein Wort mit $k \leq m$ Buchstaben akzeptiert wird, da $2^{-\frac{1}{k}} \leq 2^{-\frac{1}{m}} < x \Rightarrow x^k > \frac{1}{2}$. Andererseits wird ein Wort mit $l > m$ Buchstaben verworfen, da $2^{-\frac{1}{l}} \geq 2^{-\frac{1}{m+1}} > x \Rightarrow x^l < \frac{1}{2}$.

Es gilt $T(M, \frac{1}{2}) = C_m$.

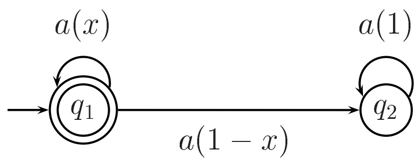

Abbildung 4: Ein PFA für die Sprache C_m mit 2 Zuständen und isoliertem Cutpoint

Außerdem ist der Cutpoint $\frac{1}{2}$ isoliert. Sei $\epsilon_m := \min\{x^m - \frac{1}{2}, \frac{1}{2} - x^{m+1}\} > 0$.

Dann gilt für alle $i \in I\!N_0$: $|p_M(a^i) - \frac{1}{2}| \geq \epsilon_m$. ■

Auf der anderen Seite können wir folgende Aussage über einen äquivalenten DFA treffen.

Satz 3.31 *Mindestgröße eines DFA, der die Sprache C_m akzeptiert.*
Jeder DFA, der die Sprache C_m akzeptiert, benötigt mindestens $m + 2$ viele Zustände.

Beweis: Angenommen, es gäbe einen DFA $M = (Q, \{a\}, \delta, q_0, F)$ mit $|Q| < m + 2$, dann gäbe es zwei $i, j \in I\!N_0$ mit $0 \leq i < j \leq m + 1$, für die gilt: $\delta(q_0, a^i) = \delta(q_0, a^j) = q'$. Da $a^{m+1} \notin C_m$, muss einerseits gelten: $\delta(q', a^{m+1-j}) = \delta(\delta(q_0, a^j), a^{m+1-j}) = \delta(q_0, a^{m+1}) \notin F$. Andererseits akzeptiert der Automat dann für $l = m + 1 - j + i \leq m$ das Wort $a^l \in C_m$ nicht, da $\delta(q_0, a^l) = \delta(\delta(q_0, a^i), a^{m+1-j}) = \delta(q', a^{m+1-j}) \notin F$. Wir haben einen Widerspruch zur Annahme erhalten.

Somit muss jeder DFA, der die Sprache C_m akzeptiert, mindestens $m+2$ Zustände haben. Ein DFA für C_m mit $m+2$ Zuständen ist leicht zu konstruieren. ■

Es gibt also keine obere Schranke für die Größe eines äquivalenten DFA, die nur von der Zustandszahl des zu simulierenden PFA und nicht von der Größe der Isolation ϵ abhängt.

3.5 Unäre reguläre Sprachen

Die Ergebnisse dieser Arbeit beschränken sich auf reguläre Sprachen, die über einem Alphabet mit nur einem Buchstaben definiert sind.

Definition 3.32 *Unäre Sprache*

Eine Sprache $L \subseteq \Sigma^*$ heißt unär, wenn $|\Sigma| = 1$.

Im Folgenden nehmen wir für unäre Sprachen stets $\Sigma = \{1\}$ an.
Wenn wir das Zustandsdiagramm eines DFA $M = (Q, \{1\}, \delta, q_0, F)$ betrachten, der eine unendliche unäre Sprache akzeptiert, dann bemerken wir, dass dieser Automat so aufgebaut sein muss, dass es einen Pfad vom Startzustand aus gibt, der von einem Zyklus gefolgt wird.

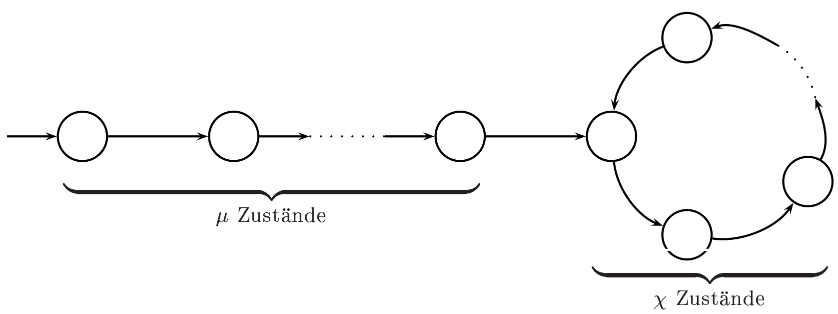

Abbildung 5: Ein unärer DFA

Die Größe eines unären DFA wird durch das Paar (χ, μ) beschrieben, wobei $\chi \geq 1$ die Anzahl der Zustände im Zyklus und $\mu \geq 0$ die Anzahl der Zustände im Anfangspfad repräsentiert.

Satz 3.33 *Struktur einer unären regulären Sprache*

Sei L eine unendliche unäre reguläre Sprache und $\chi \in I\!N$, $\mu \in I\!N_0$, dann sind die folgenden Aussagen äquivalent:

(1) L wird von einem DFA der Größe (χ, μ) akzeptiert.

(2) Für jedes $m \in I\!N_0$ mit $m \geq \mu$ gilt: $1^m \in L \Leftrightarrow 1^{m+\chi} \in L$.

Beweis:

- (1) $\Rightarrow$ (2): Sei $M = (Q, \{1\}, \delta, q_0, F)$ mit Größe (χ, μ) und $L(M) = L$ und es sei $m \geq \mu$. Dann gilt nach Struktur des DFA $\delta(q_0, 1^m) = q$ mit einem Zustand q im Zyklus und deshalb $\delta(q, 1^\chi) = q$.

$$\begin{aligned} 1^m \in L &\Leftrightarrow \delta(q_0, 1^m) = q \in F \\ &\Leftrightarrow \delta(\delta(q_0, 1^m), 1^\chi) = q \in F \\ &\Leftrightarrow 1^{m+\chi} \in L \end{aligned}$$

- (2) $\Rightarrow$ (1): χ und μ seien wie gefordert und L habe die in (2) angegebene Eigenschaft. Wir konstruieren einen DFA $M = (Q, \{1\}, \delta, q_0, F)$ mit $Q = \{q_0, q_1, \ldots, q_{\mu-1}, q_\mu, \ldots, q_{\mu+\chi-1}\}$. Die Übergangsfunktion sei so, dass $\delta(q_{i-1}, 1) = q_i$ für $1 \leq i \leq \mu + \chi - 1$ gelte und zusätzlich der Übergang $\delta(q_{\mu+\chi-1}, 1) = q_\mu$ definiert ist. Der Automat hat offensichtlich die geforderte Größe. Die akzeptierenden Zustände seien für $0 \leq i \leq \mu + \chi - 1$ so definiert: $q_i \in F \Leftrightarrow 1^i \in L$. $L = L(M)$ ergibt sich durch die Definition der akzeptierenden Zustände, den zyklischen Übergang im DFA und die Überlegungen im ersten Teil des Beweises.

■

Die Größe χ in Satz 3.33 und ihre Bestimmung ist das Hauptthema dieser Arbeit. Sie gibt Anlass zur folgenden Definition.

Definition 3.34 *Ultimative Periodizität einer Sprache*

Eine unäre Sprache L heißt ultimativ χ-periodisch, wenn es ein $\mu \in I\!N_0$ gibt, so dass für jedes $m \in I\!N_0$ mit $m \geq \mu$ gilt: $1^m \in L \Leftrightarrow 1^{m+\chi} \in L$.

Eine unäre Sprache heißt minimal ultimativ χ-periodisch, wenn sie ultimativ χ-periodisch ist, und für alle $\chi' < \chi$ nicht ultimativ χ'-periodisch ist.

Korollar 3.35 *Jede unendliche reguläre unäre Sprache ist ultimativ periodisch*

Für jede unendliche reguläre unäre Sprache L gibt es ein $\chi \in I\!N$, so dass L ultimativ χ-periodisch ist. Andererseits ist jede unäre ultimativ χ-periodische Sprache regulär.

Beweis: Das Korollar folgt direkt aus der Definition der ultimativen Periodizität und Satz 3.33, sowie der Definition regulärer Sprachen. ■

4 Bekannte Ergebnisse für reguläre Sprachen

In diesem Abschnitt zitieren wir einige Sätze über Schranken für die Größe von DFAs, die zu PFAs mit isoliertem Cutpoint äquivalent sind. Wir sprechen auch vom Blowup, der nötig ist, um einen stochastischen Automaten durch einen deterministischen Automaten zu simulieren. Die Beweise sind detailiert in den angegebenen Quellen nachzulesen. Für die unären PFAs führen wir die Beweise in Abschnitt 6 aus.

4.1 Satz von Rabin

Wie bereits erwähnt, sind die Sprachen, die von einem PFA mit isoliertem Cutpoint akzeptiert werden regulär.

Der Beweis ist auch in [Rab63], [Paz71] und [Wot00] nachzulesen.

Satz 4.1 *Satz von Rabin*

Sei $M = (Q, \Sigma, \{A(\sigma) | \sigma \in \Sigma\}, \pi, \eta)$ ein PFA mit $n := |Q|$. λ sei ein ϵ-isolierter Cutpoint bezüglich M. Dann ist $T(M, \lambda)$ regulär und es gibt einen DFA mit höchstens $(1 + \frac{1}{2\epsilon})^{n-1}$ Zuständen, der diese Sprache erkennt.

Beweis: Zunächst benötigen wir ein kombinatorisches Lemma.

Lemma 4.2 *Höchstzahl der stochastischen Vektoren mit Mindestabstand δ*

$\mathcal{P}_n$ sei die Menge der stochastischen Vektoren der Dimension n:

$$\mathcal{P}_n := \left\{ \vec{x} = (x_i) \in I\!R^n \;\middle|\; \sum_{i=1}^{n} x_i = 1, x_i \geq 0 \text{ für } 1 \leq i \leq n \right\}$$

Für ein $\delta > 0$ sei $\mathcal{U}_\delta$ eine Teilmenge von $\mathcal{P}_n$, in der je zwei Vektoren in der 1-Norm den Mindestabstand δ haben. Es gelte also

$$\forall \vec{x}, \vec{y} \in \mathcal{U}_\delta : \quad \vec{x} \neq \vec{y} \Rightarrow \sum_{i=1}^{n} |x_i - y_i| \geq \delta.$$

Dann ist $\mathcal{U}_\delta$ eine Menge mit höchstens $\left(1 + \frac{2}{\delta}\right)^{n-1}$ Elementen.

Beweis des Lemmas: Es sei $\vec{x} = (x_1, x_2, \ldots, x_n) \in U_\delta$. Wir definieren

$$\begin{aligned} v_{\vec{x}} &:= \{\vec{z} | z_i \geq x_i, \sum_{i=1}^{n} (z_i - x_i) = \frac{\delta}{2}\} \\ V_\delta &:= \{\vec{z} | z_i \geq 0, \sum_{i=1}^{n} z_i = 1 + \frac{\delta}{2}\}. \end{aligned}$$

Wir weisen zunächst nach, dass $v_{\vec{x}} \subseteq V_\delta$ für jedes $\vec{x} \in \mathcal{U}_\delta$.

Es gilt nämlich

$$\begin{aligned}
\vec{z} \in v_{\vec{x}} \quad &\Rightarrow \quad \sum_{i=1}^{n} (z_i - x_i) = \frac{\delta}{2} \\
&\Rightarrow \quad \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} x_i + \frac{\delta}{2} = 1 + \frac{\delta}{2} \\
&\Rightarrow \quad \vec{z} \in V_\delta .
\end{aligned}$$

In Abbildung 6 ist die Situation für den Fall $n = 2$ dargestellt.

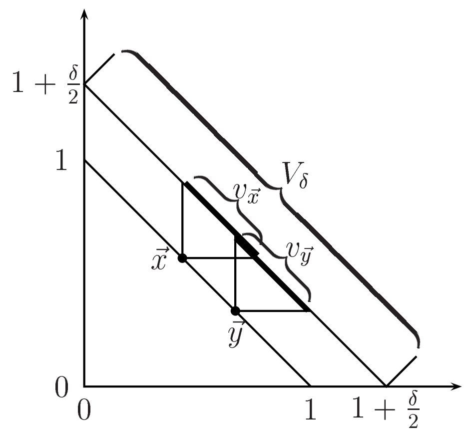

Abbildung 6: Die Mengen $v_{\vec{x}}$ und V_δ

Wir nennen einen Punkt $\vec{z} \in v_{\vec{x}}$ genau dann einen inneren Punkt von $v_{\vec{x}}$, wenn $z_i > x_i$ für jede Komponente gilt.

Behauptung: Für je zwei $\vec{x}, \vec{y} \in \mathcal{U}_\delta$ haben $v_{\vec{x}}$ und $v_{\vec{y}}$ keinen gemeinsamen inneren Punkt.

Beweis der Behauptung: Angenommen $v_{\vec{x}}$ und $v_{\vec{y}}$ hätten einen gemeinsamen inneren Punkt $\vec{z}$. Dann gilt $z_i > x_i$ und $z_i > y_i$ für alle i. Für die Differenz zweier Komponenten erhalten wir

$$\begin{aligned}
x_i - y_i &= x_i - z_i + z_i - y_i \\
&= -|x_i - z_i| + |z_i - y_i| \\
&< |x_i - z_i| + |z_i - y_i| \\
&\text{und} \\
-(x_i - y_i) &= z_i - x_i - (z_i - y_i) \\
&= |z_i - x_i| - |z_i - y_i| \\
&< |z_i - x_i| + |z_i - y_i| ,
\end{aligned}$$

woraus $|x_i - y_i| < |z_i - x_i| + |z_i - y_i|$ folgt.

Für die Summe gilt

$$\begin{aligned}\sum_{i=1}^{n}|x_i - y_i| &< \sum_{i=1}^{n}|z_i - x_i| + \sum_{i=1}^{n}|z_i - y_i| \\ &= \frac{\delta}{2} + \frac{\delta}{2} \\ &= \delta.\end{aligned}$$

Dies widerspricht aber der Tatsache, dass $\vec{x}$ und $\vec{y}$ Elemente von $\mathcal{U}_\delta$ sind. ■

Die Gesamtzahl der stochastischen Vektoren $\vec{x} \in \mathcal{U}_\delta$ kann somit nicht größer sein, als die Zahl der Simplizes $v_{\vec{x}}$, die in den Simplex V_δ passen. Mit $S(v_{\vec{x}})$ bezeichnen wir das Volumen von $v_{\vec{x}}$.

Behauptung: $S(v_{\vec{x}}) = c_n \cdot (\frac{\delta}{2})^{n-1}$, wobei c_n von δ und $v_{\vec{x}}$ unabhängig ist.

Beweis-Skizze für die Behauptung: Für $n = 2$ und $n = 3$ reicht es, die Skizzen in Abbildung 7 zu betrachten.

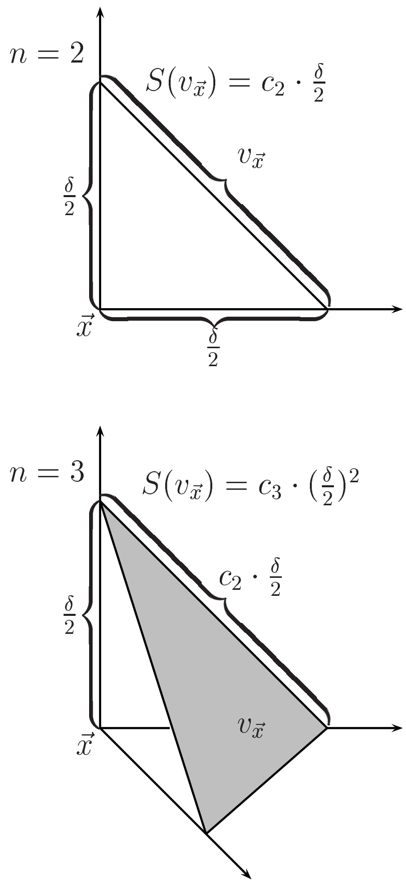

Abbildung 7: Skizzen zur Volumenberechnung eines Simplex im zwei- und drei-dimensionalen Fall

Für $n = 2$ ist die Länge der Geraden $\sqrt{2} \cdot \frac{\delta}{2}$, während wir für $n = 3$ ein gleichseitiges Dreieck der Seitenlänge $l := \sqrt{2} \cdot \frac{\delta}{2}$ und dem Flächeninhalt $\frac{l^2 \cdot \sqrt{\frac{3}{4}}}{2} = \sqrt{\frac{3}{4}} \cdot (\frac{\delta}{2})^2$ erhalten.

Im Allgemeinen müssen wir das $(n-1)$-dimensionale Volumen eines $(n-1)$-dimensionalen Simplex betrachten. Dieses Volumen ist noch von der Wahl der Basisvektoren des $(n-1)$-

dimensionalen Raums abhängig (vergleiche [For99] Seiten 46-51). Bei beliebiger aber fester Wahl der Basisvektoren erhalten wir die gewünschte Konstante c_n. □

V_δ ist ein Simplex von der gleichen Form wie $v_{\vec{x}}$, jedoch mit „Kantenlänge“ $1 + \frac{\delta}{2}$, und hat deshalb das Volumen

$$S(V_\delta) = c_n \cdot \left(1 + \frac{\delta}{2}\right)^{n-1}.$$

k sei die Zahl der Simplizes $v_{\vec{x}}$, die in den Simplex V_δ passen. Dann muss gelten:

$$\begin{array}{rrcl} k \cdot c_n \cdot \left(\frac{\delta}{2}\right)^{n-1} & & \leq & c_n \cdot \left(1 + \frac{\delta}{2}\right)^{n-1} \\ \Rightarrow & k & \leq & \left(\frac{1+\frac{\delta}{2}}{\frac{\delta}{2}}\right)^{n-1} \\ \Rightarrow & k & \leq & \left(1 + \frac{2}{\delta}\right)^{n-1}. \end{array}$$

$\mathcal{U}_\delta$ ist also eine Menge mit höchstens $\left(1 + \frac{2}{\delta}\right)^{n-1}$ Elementen. ■

Nun zeigen wir, dass der Abstand der Wahrscheinlichkeitsverteilungen auf den Zuständen im Automat nach dem Lesen zweier Nerode-inäquivalenter Wörter mindestens 4ϵ ist. Wir können somit die Menge der Wahrscheinlichkeitsverteilungen die jeweils zu einer Äquivalenzklasse der Nerode-Relation gehören als eine Menge $\mathcal{U}_\delta$ aus dem vorherigen Lemma mit $\delta = 4\epsilon$ auffassen.

Definition 4.3 *Positiv- und Negativ-Teil einer reellen Zahl*

Sei a eine reelle Zahl, dann ist $a^+ := \max(a, 0)$ und $a^- := \min(a, 0)$.

Bemerkung: Es gilt $a = a^+ + a^-$ und $|a| = a^+ - a^-$.

Lemma 4.4 *Summe der Positivteile des Differenzvektors zweier stochastischer Vektoren*

Sei $M = (Q, \Sigma, \{A(\sigma)|\sigma \in \Sigma\}, \pi, \eta)$ ein PFA. Dann gilt für alle $x, y \in \Sigma^*$:

$$\sum_{q \in Q} \left((\pi A(x))_q - (\pi A(y))_q\right)^+ = \frac{1}{2} \cdot \sum_{q \in Q} \left|(\pi A(x))_q - (\pi A(y))_q\right|.$$

Beweis: Für alle $x, y \in \Sigma^*$ gilt $\sum_{q \in Q} (\pi A(x))_q = 1$ und $\sum_{q \in Q} (\pi A(y))_q = 1$. Daraus folgt

$$\begin{array}{rrcl} & 0 & = & \sum_{q \in Q} (\pi A(x))_q - \sum_{q \in Q} (\pi A(x))_q \\ \Rightarrow & 0 & = & \sum_{q \in Q} \left((\pi A(x))_q - (\pi A(x))_q\right) \\ \Rightarrow & 0 & = & \sum_{q \in Q} \left((\pi A(x))_q - (\pi A(x))_q\right)^+ + \sum_{q \in Q} \left((\pi A(x))_q - (\pi A(x))_q\right)^- \end{array}$$

$$\Rightarrow \qquad \sum_{q\in Q}\left((\pi A(x))_q - (\pi A(x))_q\right)^+ \quad = \quad -\sum_{q\in Q}\left((\pi A(x))_q - (\pi A(x))_q\right)^- . \tag{1}$$

Für jeden (nicht notwendig stochastischen) Vektor $p = (p_1, \ldots, p_n)$ gilt

$$\sum_{i=1}^n |p_i| = \sum_{i=1}^n (p_i)^+ - \sum_{i=1}^n (p_i)^- .$$

Mit (1) ergibt sich

$$\begin{aligned}
\sum_{q\in Q}|(\pi A(x))_q - (\pi A(y))_q| \quad &= \quad \sum_{q\in Q}\left((\pi A(x))_q - (\pi A(x))_q\right)^+ - \sum_{q\in Q}\left((\pi A(x))_q - (\pi A(x))_q\right)^- \\
&= \quad 2\cdot\sum_{q\in Q}\left((\pi A(x))_q - (\pi A(x))_q\right)^+ .
\end{aligned}$$

Wir erhalten also das gewünschte Ergebnis

$$\sum_{q\in Q}\left((\pi A(x))_q - (\pi A(y))_q\right)^+ = \frac{1}{2}\cdot\sum_{q\in Q}|(\pi A(x))_q - (\pi A(y))_q| .$$

■

Beweis von Satz 4.1: Wir zeigen, dass $T(M,\lambda)$ für jedes ϵ-isolierte λ regulär ist, indem wir den Satz von Nerode (Satz 3.15) benutzen. Insbesondere zeigen wir, dass $\mathrm{ind}(R_{T(M,\lambda)}) \leq \left(1+\frac{1}{2\epsilon}\right)^{n-1}$ gilt.

Es sei $S \subseteq \Sigma^*$ eine beliebige Wortmenge mit der Eigenschaft, dass für alle $x, y \in S$ mit $x \neq y$ gilt: $x\neg R_{T(M,\lambda)}y$. Wir werden zeigen, dass $|S| \leq \left(1+\frac{1}{2\epsilon}\right)^{n-1}$ gilt.

Falls $x\neg R_{T(M,\lambda)}y$ gilt, dann gibt es ein $z \in \Sigma^*$, so dass

$$\begin{array}{ccc}
xz \in T(M,\lambda) & \Rightarrow & p_M(xz) = (\pi A(x)) \times (A(z)\eta) \geq \lambda + \epsilon \\
& \text{und} & \\
yz \notin T(M,\lambda) & \Rightarrow & p_M(yz) = (\pi A(y)) \times (A(z)\eta) \leq \lambda - \epsilon
\end{array}$$

gilt. Somit muss auch $(\pi A(x)) \times (A(z)\eta) - (\pi A(y)) \times (A(z)\eta) \geq 2\epsilon$ gelten und es folgt

$$\sum_{q\in Q}(\pi A(x) - \pi A(y))_q \cdot (A(z)\eta)_q \geq 2\epsilon.$$

Auf der anderen Seite gilt

$$\begin{aligned}
& \sum_{q\in Q}(\pi A(x) - \pi A(y))_q \cdot (A(z)\eta)_q \\
\leq \quad & \sum_{q\in Q}(\pi A(x) - \pi A(y))_q^+ \cdot \max_{p\in Q}(A(z)\eta)_p \quad + \quad \sum_{q\in Q}(\pi A(x) - \pi A(y))_q^- \cdot \min_{p\in Q}(A(z)\eta)_p \\
\leq \quad & \sum_{q\in Q}(\pi A(x) - \pi A(y))_q^+ \cdot \max_{p\in Q}(A(z)\eta)_p \\
\leq \quad & \sum_{q\in Q}(\pi A(x) - \pi A(y))_q^+ \\
\underbrace{=}_{\textit{Lemma 4.4}} \quad & \frac{1}{2}\cdot\sum_{q\in Q}|(\pi A(x))_q - (\pi A(y))_q| .
\end{aligned}$$

Wir erhalten die folgenden Ungleichungen:

$$\begin{aligned} \frac{1}{2} \cdot \sum_{q \in Q} \left|(\pi A(x))_q - (\pi A(y))_q\right| &\geq 2\epsilon \\ \Rightarrow \quad \sum_{q \in Q} \left|(\pi A(x))_q - (\pi A(y))_q\right| &\geq 4\epsilon \end{aligned}$$

Also ist $\{\pi A(x) | x \in S\}$ eine Teilmenge von $\mathcal{P}_n$ mit den Eigenschaften, die wir von $\mathcal{U}_{4\epsilon}$ in Lemma 4.2 fordern. Außerdem gilt für alle $x, y \in S : x \neq y \Rightarrow \pi A(x) \neq \pi A(y)$, da wir ja wissen, dass $x \neg R_{T(M,\lambda)} y$ gilt. Zusammen mit Lemma 4.2 ergibt sich

$$|S| = |\{\pi A(x) | x \in S\}| \leq |\mathcal{U}_{4\epsilon}| \leq \left(1 + \frac{2}{4\epsilon}\right)^{n-1} \leq \left(1 + \frac{1}{2\epsilon}\right)^{n-1}.$$

Es gibt also höchstens $\left(1 + \frac{1}{2\epsilon}\right)^{n-1}$ viele $x \in \Sigma^*$, die nicht äquivalent bezüglich der Nerode-Relation sind. Somit gilt auch $\mathrm{ind}(R_{T(M,\lambda)}) \leq \left(1 + \frac{1}{2\epsilon}\right)^{n-1}$ und wir können den Nerode-Automaten mit höchstens $\left(1 + \frac{1}{2\epsilon}\right)^{n-1}$ Zuständen konstruieren. ■

4.1.1 Eine untere Schranke von Ambainis

Die obere Schranke von Rabin scheint nicht scharf zu sein. Die beste bisher bekannte untere Schranke wird im folgenden Satz beschrieben [Amb96]:

Satz 4.5 *Satz von Ambainis*

Es gibt eine unendliche Folge von PFAs mit isoliertem Cutpoint und n Zuständen, so dass der jeweils minimale äquivalente DFA mindestens $\Omega\left(2^{n \frac{\log \log n}{\log n}}\right)$ Zustände benötigt.

Beweis-Skizze: Ambainis beweist diesen Satz mit Hilfe der Sprachfamilie $L_m \subseteq \Sigma_m^*$ mit dem Alphabet $\Sigma_m = \{a_1, a_2, \ldots, a_m\}$. Jedes Wort $w \in L_m$ hat die Länge m^2 und enthält jeden der Buchstaben a_i jeweils genau m mal.

Jeder DFA, der L_m akzeptiert, benötigt mindestens $(m+1)^m$ viele Zustände, da er sich die Anzahl der Vorkommnisse jedes einzelnen Buchstaben merken muss.

Außerdem gibt es einen PFA mit ϵ-isoliertem Cutpoint, der mit $O\left(m \frac{\log^2 m}{\log \log m}\right)$ Zuständen auskommt und L_m akzeptiert.

Setzen wir n als Zustandszahl des minimalen PFA mit isoliertem Cutpoint, der L_m akzeptiert, dann gilt $n = O\left(m \frac{\log^2 m}{\log \log m}\right)$. Jeder äquivalente DFA hat $k \geq (m+1)^m \geq 2^{m \log m}$ Zustände. Asymptotisch ergibt sich

$$k = \Omega\left(2^{\frac{n \log \log n}{\log n}}\right).$$

□

4.2 Der unäre Fall

Wenn wir unsere Untersuchungen auf unäre Sprachen beschränken, erkennen wir, dass PFAs mit isoliertem Cutpoint und NFAs in Bezug auf den zyklischen Teil des DFA keine mächtigen Einsparungen liefern können.

4.2.1 Tradeoff zwischen NFA und DFA für unäre Sprachen

Chrobak [Chr86] hat eine obere Schranke für die Umwandlung eines NFA in einen DFA im unären Fall gezeigt, die niedriger als die bekannte 2^n Schranke für den allgemeinen Fall ist.

Satz 4.6 *Obere Schranke beim Übergang von NFA auf DFA im unären Fall*
Zu jedem unären NFA mit n Zuständen gibt es einen äquivalenten DFA der Größe

$$(O(e^{\sqrt{n \ln n}}), O(n^2)).$$

Außerdem ist diese Schranke bezogen auf den zyklischen Teil scharf:

Satz 4.7 *Untere Schranke beim Übergang von NFA auf DFA im unären Fall*
Für jedes n gibt es einen unären NFA mit n Zuständen, so dass jeder äquivalente DFA $\Omega(e^{\sqrt{n \ln n}})$ viele Zustände im Zyklus benötigt.

4.2.2 Tradeoff zwischen PFA mit isoliertem Cutpoint und DFA für unäre Sprachen

Ein ähnliches Ergebnis, haben Milani und Pighizzini [MP00] für den PFA mit isoliertem Cutpoint gezeigt:

Satz 4.8 *Obere Schranke beim Übergang von PFA auf DFA im unären Fall*
Zu jedem unären PFA, der eine Sprache L mit isoliertem Cutpoint akzeptiert, gibt es einen äquivalenten DFA mit $O(e^{\sqrt{n \ln n}})$ vielen Zuständen im zyklischen Teil.

Dieser Satz wird in Abschnitt 6.1 bewiesen.

Wir erhalten also eine obere Schranke für die ultimative Periodizität der von einem PFA mit ϵ-isoliertem Cutpoint akzeptierten Sprache. Diese Schranke ist unabhängig von ϵ.

Milani und Pighizzini zeigen auch, dass die Schranke scharf ist, denn für jedes n gibt es einen PFA mit n Zuständen, für den jeder äquivalente DFA mindestens $\Omega(e^{\sqrt{n \ln n}})$ Zustände benötigt.

Wir zeigen in unserer Arbeit, dass es sinnvoll ist, die Größe von ϵ wieder in die Betrachtung miteinzubeziehen. Für große ϵ, erhalten wir eine viel kleinere obere Schranke. Insbesondere ergibt sich für konstantes ϵ eine polynomielle obere Schranke. Außerdem können wir zeigen, dass unsere obere Schranke in gewissem Sinne scharf ist.

Auf diesem Weg kommen wir auch zu einem Ergebnis, das Mereghetti, Palano und Pighizzini in ihrer Arbeit [MPP01] bewiesen haben.

Satz 4.9 *Größe eines unären PFA für L_k*

Für jedes $k = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_m^{\alpha_m}$ mit p_i prim und $\alpha_i \in I\!N$ ist die notwendige und ausreichende Zahl der Zustände eines PFA, der die Sprache $L_k = \{1^{lk} | l \in I\!N_0\}$ mit isoliertem Cutpoint akzeptiert, $p_1^{\alpha_1} + p_2^{\alpha_2} + \ldots + p_m^{\alpha_m}$.

Wir gehen auf diesen Satz noch einmal in Abschnitt 6.2 ein.

4.2.3 Weitere Literatur zu unären Sprachen

Hier seien zwei Arbeiten noch kurz erwähnt, ohne näher auf die Ergebnisse einzugehen.

In [GMP01] beschäftigen sich die Autoren mit unären Zwei-Weg NFAs (2NFAs), also mit nichtdeterministischen endlichen Automaten, die beim Lesen des Eingabewortes die Richtung wechseln können. Sie zeigen, dass auf Eingaben der Länge mindestens $5n^2$ jeder unäre 2NFA mit n Zuständen durch einen 2NFA mit $(2n + 2)$ Zuständen simuliert werden kann, der nur am Anfang und Ende des Worts die Richtung wechselt. Aus diesem Resultat ergibt sich neben der Entwicklung einer Normalform für 2NFAs auch die Möglichkeit einer subexponentiellen Simulation unärer 2NFAs durch Zwei-Weg DFAs (2DFAs). Es wird sogar gezeigt, dass jeder unäre 2NFA mit n Zuständen von einem 2DFA mit $O(n^{\log n+4})$ Zuständen simuliert werden kann.

Auch die Arbeit [MP01] beschäftigt sich mit der Umwandlung der verschiedenen Arten unärer endlicher Automaten. Das Hauptergebnis ist eine *scharfe* Schranke beim Übergang von einem unären 2NFA mit n Zuständen auf den äquivalenten DFA. Dieser benötigt $O(e^{\sqrt{n \ln n}})$ Zustände. Zu einem 2NFA mit n Zuständen kann außerdem ein äquivalenter 2NFA mit $O(n^2)$ Zuständen konstruiert werden, der sowohl die Richtungswechsel, als auch die nichtdeterministischen Wahlen nur am Anfang oder Ende des zu verarbeitenden Worts macht.

5 Bekannte Ergebnisse für homogene Markovketten

Wir benötigen für unsere Ergebnisse sehr detailierte Kenntnisse über die Markovketten und ihre Periodizität. Wir werden in erster Linie den Darstellungen in [Rob76], [TK98] und [Gan86] folgen.

Milani und Pighizzini beschränkten sich darauf, ein Ergebnis (Satz 5.41) von Gantmacher [Gan86] zu zitieren, das dort mit Hilfe der Eigenwerte einer stochastischen Matrix hergeleitet wurde. Wir sind der Meinung, dass dieser wenig intuitive, aber ähnlich aufwändig einzuführende Weg, die Periodizität zu berechnen, über den Rahmen dieser Arbeit hinausgeht. Wir wählen den Weg der anschaulichen graphischen Repräsentation der Markovkette in ihrem Zustandsdiagramm, um die Intuition zu stärken, ohne jedoch die mathematische Exaktheit außer Acht zu lassen.

Einige Details zur Berechnung der Periodizität auf diesem Weg haben wir in der Literatur noch nicht komplett ausgeführt gesehen und leiten diese selbst her. Auch diese Zwischenergebnisse sind Grundlagen für unsere Überlegungen zu den endlichen stochastischen unären Automaten.

5.1 Grundlagen der Markovkettentheorie

Definition 5.1 *Markovprozess*

Ein Markovprozess $\{X_i\}$ ist ein stochastischer Prozess mit der Eigenschaft, dass bei gegebenem Wert von X_t, die Werte von X_s für $s > t$ nicht beeinflusst werden von den Werten von X_u für $u < t$.

Bemerkung: Kennen wir also den Zustand genau, in dem sich der Markovprozess zu einem bestimmten Zeitpunkt befindet, so können wir die Wahrscheinlichkeiten des zukünftigen Verhaltens vorhersagen. Eine zusätzliche Information über die Vergangenheit verändert daran nichts.

Definition 5.2 *Markovkette in diskreter Zeit*

Eine Markovkette in diskreter Zeit (oder kurz Markovkette) ist ein Markovprozess, dessen Zustandsraum eine endliche oder abzählbare Menge ist und dessen (Zeit-)Indexmenge $I\!N_0$ ist.

Bemerkung: In diesem Fall bedeutet die Markov-Eigenschaft:

$$\text{Ws}\{X_{n+1} = q | X_0 = p_0, \ldots, X_{n-1} = p_{n-1}, X_n = p\} = \text{Ws}\{X_{n+1} = q | X_n = p\}$$

für alle Zeitpunkte n und alle Zustände $p_0, \ldots, p_{n-1}, p, q$.

Für unsere Zwecke reicht es, endliche Markovketten zu betrachten, das heißt solche mit endlichem Zustandsraum.

Außerdem benötigen wir nur homogene Markovketten.

Definition 5.3 *Homogene Markovkette*

Eine Markovkette $\{X_i\}$ heißt homogen, wenn für alle $m, n \in I\!N_0$ und alle Zustände p, q gilt:

$$\text{Ws}\{X_{n+1} = q | X_n = p\} = \text{Ws}\{X_{m+1} = q | X_m = p\}.$$

Bemerkung: Bei einer homogenen Markovkette spielt es also keine Rolle, zu welchem Zeitpunkt ein Übergang beobachtet wird.

Definition 5.4 *Übergangswahrscheinlichkeitsmatrix einer homogenen Markovkette*

Einer endlichen homogenen Markovkette $\{X_i\}$ ist ihre Übergangswahrscheinlichkeitsmatrix A zugeordnet. A ist eine stochastische Matrix, deren Zeilen und Spalten mit den Zuständen aus dem Zustandsraum der Markovkette indiziert sind, und für die $A_{p,q} = \mathrm{Ws}\{X_{m+1} = q | X_m = p\}$ gilt.

$A_{p,q}$ ist also die Wahrscheinlichkeit, dass die Markovkette vom Zustand p ausgehend in einem Schritt den Zustand q erreicht.

Die Wahrscheinlichkeit, dass die Markovkette vom Zustand p ausgehend in zwei Schritten den Zustand q erreicht, ist offensichtlich die Summe über die Wahrscheinlichkeiten, von p zu einem möglichen Zwischenzustand r und dann zu q zu gehen.

$$\begin{aligned}
\mathrm{Ws}\{X_{n+2} = q | X_n = p\} &= \sum_r \mathrm{Ws}\{X_{n+1} = r | X_n = p\} \cdot \mathrm{Ws}\{X_{n+2} = q | X_{n+1} = r\} \\
&= \sum_r A_{p,r} \cdot A_{r,q} \\
&= (A^2)_{p,q}
\end{aligned}$$

Wenden wir diese Überlegung rekursiv an, dann ergibt sich, dass die Wahrscheinlichkeit, von Zustand p ausgehend den Zustand q, in $t \in \mathbb{N}_0$ Schritten zu erreichen, gerade

$$\mathrm{Ws}\{X_{n+t} = q | X_n = p\} = (A^t)_{p,q}$$

ist. Wir setzen $A^0 = I$.

Zur Verbesserung der Lesbarkeit werden wir $A^{(t)}_{p,q}$ anstelle von $(A^t)_{p,q}$ schreiben.

Definition 5.5 *Zustandsdiagramm einer Markovkette*

$\{X_i\}$ sei eine homogene Markovkette mit endlichem Zustandsraum Q und der Übergangswahrscheinlichkeitsmatrix A. Das Zustandsdiagramm D dieser Markovkette ist ein gerichteter Graph (V, E) mit $V = Q$ und $E = \{(p, q) | p, q \in V \wedge A_{p,q} > 0\}$. Oft wird auch eine Kantenbeschriftung benutzt, die eine Kante $(p, q) \in E$ mit der Übergangswahrscheinlichkeit $A_{p,q}$ beschriftet.

Definition 5.6 *Erreichbarkeit*

Ein Zustand $q \in Q$ ist von einem Zustand $p \in Q$ aus erreichbar (beziehungsweise p kann q erreichen), wenn es ein $t \in \mathbb{N}_0$ gibt, so dass $A^{(t)}_{p,q} > 0$. Wir schreiben $p \to q$.

Bemerkung: Wegen $A^{(0)}_{q,q} = 1$, gilt für alle Zustände $q \to q$.

Satz 5.7 *Erreichbarkeit im Zustandsdiagramm*

D sei das Zustandsdiagramm einer homogenen Markovkette mit Übergangswahrscheinlichkeitsmatrix A und Zustandsmenge Q. Für zwei Zustände $p, q \in Q$ gilt $p \to q$ genau dann, wenn in D ein Pfad von p nach q existiert.

Beweis: „$\Rightarrow$": Es gelte $p \to q$. Zu zeigen ist, dass es eine Folge von Zuständen $p_0, \ldots, p_t$ gibt, so dass für $0 \le i < t$ gilt: $(p_i, p_{i+1}) \in E$, wobei $p_0 = p$ und $p_t = q$.
Nach Voraussetzung gibt es ein $t \in I\!N_0$, so dass $A^{(t)}_{p,q} > 0$. Wir werden einen Pfad der Länge t, also mit t Kanten, finden.
Falls $t = 0$, dann muss $A^{(0)}_{p,q} > 0$ und somit $p = q$ gelten und wir betrachten den Pfad mit nur einem Knoten als den trivialen Pfad.
Falls $t = 1$, dann gilt $A_{p,q} > 0$ und dies entspricht genau der Definition einer Kante (p, q) in D.
Sei also $t > 1$. Dann gilt $A^{(t)}_{p,q} = \sum_{r \in Q} A^{(t-1)}_{p,r} \cdot A_{r,q} > 0$. Wegen der Nichtnegativität aller Einträge einer stochastischen Matrix muss es also mindestens ein $r \in Q$ geben, so dass sowohl $A^{(t-1)}_{p,r} > 0$, als auch $A_{r,q} > 0$ gelten muss. Wir haben somit die letzte Kante des Pfades $(r, q) \in E$ gefunden und können das Vorgehen mit Voraussetzung $A^{(t-1)}_{p,r} > 0$ rekursiv wiederholen, um die letzte Kante eines Pfades der Länge der $t-1$ von p nach r zu konstruieren. Dies wiederholen wir, bis wir nur noch einen Pfad der Länge 1 haben, für den wir ja schon wissen, dass diese eine Kante in D existiert.
„$\Leftarrow$": Angenommen, es gibt eine Folge von Zuständen $p_0, \ldots, p_t$, so dass für $0 \le i < t$ gilt: $(p_i, p_{i+1}) \in E$, wobei $p_0 = p$ und $p_t = q$. Das heißt, dass die entsprechenden Einträge in der Übergangswahrscheinlichkeitsmatrix positiv sind: $A_{p_i,p_{i+1}} > 0$ für $0 \le i < t$.
Wir wollen zeigen, dass $A^{(t)}_{p,q} > 0$ gilt:

$$\begin{aligned}
A^{(t)}_{p,q} = A^{(t)}_{p_0,p_t} &= \sum_{r \in Q} A^{(t-1)}_{p_0,r} A_{r,p_t} \\
&\ge A^{(t-1)}_{p_0,p_{t-1}} A_{p_{t-1},p_t} \\
&\ge \Big(\sum_{r \in Q} A^{(t-2)}_{p_0,r} A_{r,p_{t-1}}\Big) A_{p_{t-1},p_t} \\
&\ge \ldots \\
&\ge A_{p_0,p_1} A_{p_1,p_2} \cdots A_{p_{t-1},p_t} \\
&> 0
\end{aligned}$$

Denn jeder einzelne Faktor ist > 0. ■

Definition 5.8 *Kommunizierende Zustände*
Zwei Zustände p und q kommunizieren miteinander, wenn sowohl $p \to q$, als auch $q \to p$ gilt. Wir schreiben $p \leftrightarrow q$.

Satz 5.9 $\leftrightarrow$ *als Äquivalenzrelation*
Die Kommunikation zwischen je zwei Zuständen $\leftrightarrow$ definiert eine Äquivalenzrelation:

Beweis:

(1) $q \leftrightarrow q$ (Reflexivität) folgt aus $A^0 = I$ und $I_{q,q} = 1$.

(2) Falls $q \leftrightarrow p$, dann gilt auch $p \leftrightarrow q$ (Symmetrie). Dies folgt aus der Definition der kommunizierenden Zustände.

(3) Falls $p \leftrightarrow r$ und $r \leftrightarrow q$, dann gilt auch $p \leftrightarrow q$ (Transitivität). Die Transitivität folgt so: Wenn $p \leftrightarrow r$ und $r \leftrightarrow q$, dann gibt es $n, m \in I\!N_0$, so dass $A^{(n)}_{p,r} > 0$ und $A^{(m)}_{r,q} > 0$. Wir folgern daraus, zusammen mit der Nichtnegativität jedes Eintrags der Matrix A und jeder ihrer Potenzen A^t:
$$A^{(n+m)}_{p,q} = \sum_{s \in Q} A^{(n)}_{p,s} \cdot A^{(m)}_{s,q} \geq A^{(n)}_{p,r} \cdot A^{(m)}_{r,q} > 0$$

■

Die Zustandsmenge lässt sich somit in Äquivalenzklassen, die kommunizierenden Klassen, partitionieren. Nach unseren Überlegungen zur Erreichbarkeit entsprechen die Äquivalenzklassen im Zustandsdiagramm den starken Zusammenhangskomponenten. Jeder Knoten gehört zu genau einer starken Zusammenhangskomponente.

Definition 5.10 *Kondensationsgraph D^**
Sei D ein gerichteter Graph mit den starken Zusammenhangskomponenten $\{K_1, K_2, \ldots, K_m\}$. Dann definieren wir den Kondensationsgraph $D^* = (V, E)$ als gerichteten Graphen mit der Knotenmenge $V = \{K_1, K_2, \ldots, K_m\}$. Eine Kante (K_i, K_j) existiert in D^* genau dann, wenn es in D zwei Knoten $v \in K_i$ und $u \in K_j$ gibt, für die eine Kante (v, u) in D existiert.

Wir kontrahieren also die Äquivalenzklassen bezüglich der kommunizierenden Zustände zu einem einzigen Zustand. Ein Beispiel für den Kondensationsgraphen eines gerichteten Graphen D aus Abbildung 8 ist in Abbildung 9 zu sehen.

Satz 5.11 *Azyklizität des Kondensationsgraphen*

Sei D ein gerichteter Graph und D^* der zugehörige Kondensationsgraph, dann ist D^* azyklisch.

Beweis: Angenommen, es gibt einen Zyklus $K_{i_1}, K_{i_2}, \ldots, K_{i_t}, K_{i_1}$ in D^*. Die K_i in dieser Aufzählung sind paarweise verschieden, denn sonst könnten wir den Zyklus kürzer beschreiben. u sei ein Knoten in K_{i_1} und v sei ein Knoten in K_{i_2}. Über den Zyklus $K_{i_1}, K_{i_2}, \ldots, K_{i_t}, K_{i_1}$ ist offensichtlich u von v aus erreichbar, sowie v von u aus erreichbar. Also müssten u und v in derselben starken Zusammenhangskomponente liegen. Dies ist ein Widerspruch zur Annahme, dass die K_i paarweise verschieden sind. ■

Für die folgenden Definitionen sei wieder A die Übergangswahrscheinlichkeitsmatrix einer endlichen homogenen Markovkette mit Zustandsmenge Q.

Definition 5.12 *Abgeschlossene Zustandsmenge*
Eine Zustandsmenge $C \subseteq Q$ heißt abgeschlossen, wenn für alle $p \in C$ und für alle $q \notin C$ gilt $A_{p,q} = 0$.

Es gibt also keine Möglichkeit, eine abgeschlossene Menge zu verlassen. Zum Beispiel ist in Abbildung 8 die Menge $\{7, 8, 9, 10, 11\}$ abgeschlossen, ebenso wie $\{9, 10, 11\}$.

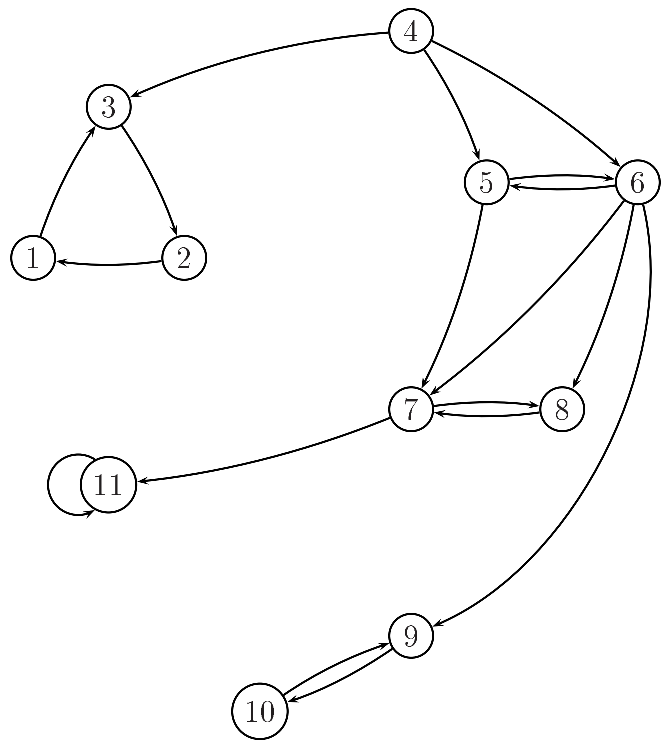

Abbildung 8: Beispiel aus [Rob76] zur Klassifizierung der Zustände

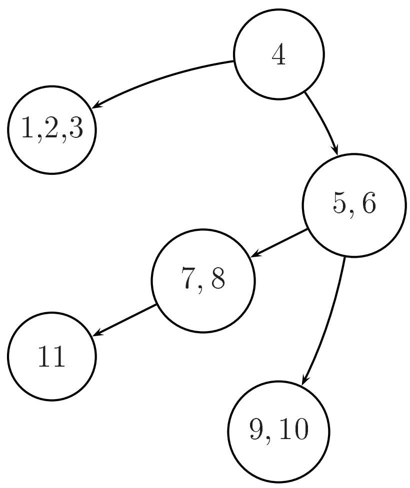

Abbildung 9: Beispiel aus [Rob76]: Der Kondensationsgraph D^*

Definition 5.13 *Ergodische Zustandsmenge*
Eine Zustandsmenge $C \subseteq Q$ heißt ergodisch (oder irreduzibel), wenn sie abgeschlossen ist und keine echte Teilmenge $B \subset C$ ebenfalls abgeschlossen ist.[1]

In unserem Beispiel in Abbildung 8 sind die ergodischen Mengen die $\{9, 10\}$, $\{11\}$ und $\{1, 2, 3\}$.

[1] Einige Autoren fordern bereits für die Ergodizität auch die Aperiodizität. Wir folgen hier der Bezeichnung von [Rob76] und [KS60] und nennen aperiodische ergodische Markovketten regelmäßige Ketten. Siehe auch Definition 5.35.

Satz 5.14 *Ergodische Menge im Kondensationsgraphen*

Eine Zustandsmenge $C \subseteq Q$ ist genau dann ergodisch, wenn C im Zustandsdiagramm (und entsprechend im Kondensationsgraphen) eine starke Zusammenhangskomponente ist, die keine ausgehenden Kanten hat.

Beweis:

„$\Leftarrow$“: Sei C im Zustandsdiagramm eine starke Zusammenhangskomponente ohne ausgehende Kanten. Dann ist C auch eine Äquivalenzklasse kommunizierender Zustände, bei der für alle $p \in C$ und für alle $q \notin C$ gilt: $p \not\to q$, also insbesondere $A_{p,q} = 0$. Somit ist C abgeschlossen.
Außerdem gilt für ein $B \subset C$, dass B nicht abgeschlossen sein kann, denn in C kann jeder Zustand jeden anderen erreichen. Es gibt also immer ein Paar $p \in B$ und $q \in C \setminus B$, so dass $A_{p,q} > 0$ gilt.

„$\neg \Leftarrow \neg$“: Sei C eine Knotenmenge, die aus C ausgehende Kanten hat. Es gibt dann $p \in C$ und $q \notin C$, so dass eine Kante von p nach q führt. Für diesen Eintrag in der Wahrscheinlichkeitsübergangsmatrix gilt $A_{p,q} > 0$. C kann nicht abgeschlossen sein.
Angenommen C ist eine Knotenmenge, die keine ausgehenden Kanten hat, aber C ist selbst keine starke Zusammenhangskomponente. Wir können dann den Kondensationsgraphen eingeschränkt auf C betrachten. Es muss eine starke Zusammenhangskomponente $B \subset C$ geben, die keine ausgehende Kante hat (vergleiche den Beweis zu Satz 5.11). B ist somit eine echte Teilmenge von C, die abgeschlossen ist. C kann nicht ergodisch sein. ∎

Definition 5.15 *Transiente Zustandsmenge*

Ein Zustandsmenge $C \subseteq Q$, die eine Äquivalenzklasse bezüglich der Kommunikation ist und die nicht ergodisch ist, heißt transiente Zustandsmenge.

Bemerkung: Im Zustandsdiagramm ist eine transiente Zustandsmenge eine starke Zusammenhangskomponente mit ausgehenden Kanten.

Definition 5.16 *Ergodische, absorbierende und transiente Zustände*

Wir nennen einen Zustand $q \in Q$ ergodisch, wenn die Äquivalenzklasse aller mit ihm kommunizierender Zustände $[q]_\leftrightarrow$ ergodisch ist.
Wir nennen einen Zustand $q \in Q$ transient, wenn $[q]_\leftrightarrow$ transient ist.
Wir nennen einen Zustand $q \in Q$ absorbierend, wenn $[q]_\leftrightarrow = \{q\}$ gilt und $\{q\}$ abgeschlossen ist.

Bemerkung: Ein Zustand q ist genau dann absorbierend, wenn $A_{q,q} = 1$ gilt. Ein absorbierender Zustand q ist ergodisch.
Diese Definition des transienten Zustands ist im Falle der endlichen Zustandsmenge äquivalent zu der bekannten Definition in der Stochastik, die sich auf die Wahrscheinlichkeit der Rückkehr zu einem Zustand q in endlicher Zeit bezieht.

Satz 5.17 *Struktur einer Markovkette*

Eine endliche homogene Markovkette besteht aus transienten und ergodischen Zustandsmengen (Komponenten). Die folgenden Eigenschaften gelten:

(1) Hat die Markovkette einmal eine transiente Komponente erreicht und diese dann wieder verlassen, dann kann sie nicht mehr zu dieser transienten Komponente zurückkehren.

(2) Von einem transienten Zustand aus kann man immer mindestens eine ergodische Komponente erreichen.

(3) Hat die Markovkette einmal eine ergodische Komponente erreicht, dann kann sie diese nicht mehr verlassen.

(4) In einer endlichen homogenen Markovkette gibt es immer mindestens eine ergodische Komponente.

Beweis: Die Eigenschaften sind eine direkte Konsequenz aus unseren bisherigen Überlegungen zur Erreichbarkeit und ihrer Übertragbarkeit auf das Zustandsdiagramm (Satz 5.7). Wir wissen, dass der Kondensationsgraph azyklisch ist, und für gerichtete azyklische Graphen existiert bekanntlich eine topologische Sortierung. Wir betrachten also eine beliebige aber feste topologische Sortierung auf dem Kondensationsgraphen.

(1) Es ist nicht möglich, von einem Knoten zu einem anderen Knoten zu gelangen, der früher in der topologischen Sortierung erscheint.

(2) Hat ein Knoten noch mindestens eine ausgehende Kante, dann kann man dieser folgen und wieder zu einem Knoten gelangen, der in der topologischen Sortierung später erscheint. Dies kann man solange fortführen, bis ein Knoten keine ausgehende Kante mehr hat. Die zu diesem Knoten gehörige Komponente ist ergodisch.

(3) Es gibt keine ausgehende Kante aus einer Knotenmenge, die einer ergodischen Komponente entspricht.

(4) In der topologischen Sortierung des Kondensationsgraphens gibt es einen letzten Knoten, dieser kann keine ausgehende Kante haben und die zugehörige Komponente ist somit ergodisch.

■

Definition 5.18 *Ergodische Markovkette*
Eine Markovkette heißt ergodisch (oder irreduzibel), wenn jeder Zustand mit jedem anderen kommunizieren kann.

Bemerkung: Eine Markovkette ist genau dann ergodisch, wenn ihr Zustandsdiagramm stark zusammenhängend ist. Eine ergodische Markovkette besteht nur aus ergodischen Zuständen. Der Kondensationsgraph einer ergodischen Markovkette besteht aus einem einzigen Zustand.

Definition 5.19 *Absorbierende Markovkette*
Eine Markovkette heißt absorbierend, wenn jede ergodische Menge nur aus einem Zustand besteht.

Bemerkung: Eine Markovkette ist genau dann absorbierend, wenn sie wenigstens einen absorbierenden Zustand hat und jeder nichtabsorbierende Zustand kann irgendeinen absorbierenden Zustand erreichen.
Wir können die Untersuchung einer beliebigen Markovkette in zwei Stufen aufteilen. Das Verhalten bevor wir in einen ergodischen Zustand wechseln und das Verhalten danach. Beim Verhalten vor dem Eintritt in einen ergodischen Zustand können wir die Struktur der ergodischen Mengen vernachlässigen und diese jeweils als einen absorbierenden Zustand auffassen.

Definition 5.20 *Absorptionsgraph*

Sei D das Zustandsdiagramm einer Markovkette mit Zustandsmenge Q und Wahrscheinlichkeitsübergangsmatrix A. Die Markovkette habe die ergodischen Zustandskomponenten $E_1, E_2, \ldots, E_m$. B sei die Menge aller ergodischen Zustände $B := \bigcup_{i=1}^m E_i$. Dann ist $D^\diamond = (V, E)$ der Absorptionsgraph mit der Knotenmenge $V = (Q \setminus B) \cup \bigcup_{i=1}^m \{E_i\} =: Q^\diamond$. Das heißt: Jeder transiente Zustand ist ein Knoten. Außerdem wird jede ergodische Komponente zu einem einzelnen Knoten kontrahiert. Eine Kante (p, q) existiert in $D^\diamond$ für zwei transiente Zustände $p, q \in Q$ genau dann, wenn $A_{p,q} > 0$. Von einem transienten Zustand p zum Knoten einer ergodischen Komponente E_i existiert die Kante (p, E_i) genau dann, wenn $\sum_{q \in E_i} A_{p,q} > 0$. Außerdem gibt es für jedes E_i eine Schleife $(E_i, E_i) \in E$. Darüber hinaus gibt es keine Kanten.

In Abbildung 10 ist der Absorptionsgraph $D^\diamond$ dargestellt. Die Abbildung setzt die Reihe der Beispiele in Abbildungen 8 und 9 fort.

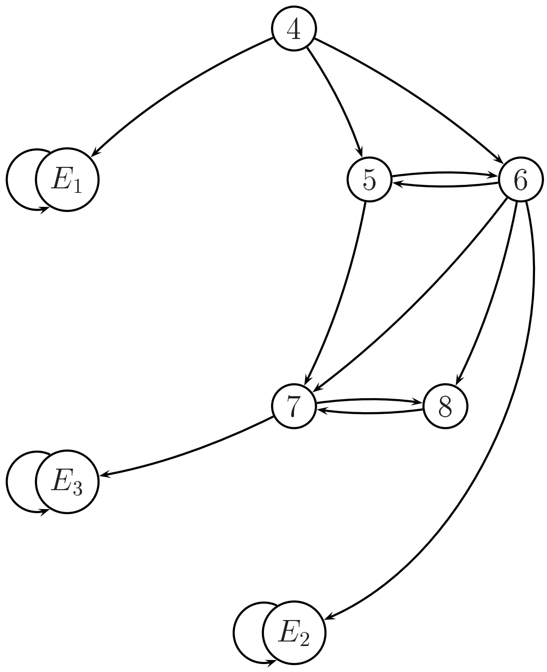

Abbildung 10: Absorptionsgraph $D^\diamond$ zur Analyse der Markovkette vor dem Eintritt in einen ergodischen Zustand

Wir können den Absorptionsgraphen wieder als Zustandsdiagramm einer absorbierenden Markovkette mit Zustandsmenge $Q^\diamond$ betrachten und uns so zunächst auf die Analyse einer absorbierenden Markovkette beschränken. Auf diese Tatsache wird in [Rob76] zwar hingewiesen, sie wird allerdings nicht weiter mathematisch ausgeführt. Wir benötigen jedoch eine exakte Theorie der Zerlegung der Analyse des Verhaltens vor und nach dem Eintreten in eine ergodische Komponente und leiten deshalb die benötigten Ergebnisse selbst her.

Definition 5.21 *Absorptionsmatrix P der absorbierenden Markovkette zu $D^\diamond$*

Wir übernehmen die Voraussetzungen aus der vorigen Definition. Dann sei P die Übergangswahrscheinlichkeitsmatrix der zum Zustandsdiagramm $D^\diamond$ gehörigen Markovkette. P sei für alle transienten Zustände p und q und für alle ergodischen Komponenten E_i, E_j so definiert:

$$\begin{aligned}
P_{p,q} &:= A_{p,q} \\
P_{p,E_i} &:= \sum_{r \in E_i} A_{p,r} \\
P_{E_i,q} &:= 0 \\
P_{E_i,E_j} &:= \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}
\end{aligned}$$

Wir nennen P die zu A gehörige Absorptionsmatrix.

Satz 5.22 *Potenzen der Absorptionsmatrix P*

P sei die zu A gehörige Absorptionsmatrix. Dann gilt für $t \in \mathbb{N}_0$, für alle transienten Zustände p, q und alle ergodischen Komponenten E_i, E_j:

$$\begin{aligned}
P^{(t)}_{p,q} &= A^{(t)}_{p,q} \\
P^{(t)}_{p,E_i} &= \sum_{r \in E_i} A^{(t)}_{p,r} \\
P^{(t)}_{E_i,q} &= 0 \\
P^{(t)}_{E_i,E_j} &= \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}
\end{aligned}$$

Beweis: Die letzten beiden Gleichungen sind offensichtlich, da es sich um eine absorbierende Markovkette handelt.

Wir zeigen $P^{(t)}_{p,q} = A^{(t)}_{p,q}$ für zwei transiente Zustände p, q durch Induktion: $P^{(0)}_{p,q} = A^{(0)}_{p,q}$ gilt, weil der Eintrag 1 ist, falls $p = q$ und 0 sonst. Für $t = 1$ greift die Definition von P. Gelte jetzt also $P^{(t)}_{p,q} = A^{(t)}_{p,q}$ für alle transienten Zustände p, q. Dann folgt

$$\begin{aligned}
P^{(t+1)}_{p,q} &= \sum_{r \in Q^\diamond} P^{(t)}_{p,r} P_{r,q} \\
&= \sum_{\substack{r \in Q^\diamond, \\ r \text{ transient}}} P^{(t)}_{p,r} P_{r,q} + \sum_{i=1}^{m} P^{(t)}_{p,E_i} \underbrace{P_{E_i,q}}_{0} \\
&= \sum_{\substack{r \in Q, \\ r \text{ transient}}} A^{(t)}_{p,r} A_{r,q} + \sum_{\substack{r \in Q, \\ r \text{ ergodisch}}} A^{(t)}_{p,r} \underbrace{A_{r,q}}_{0} \\
&= \sum_{r \in Q} A^{(t)}_{p,r} A_{r,q} \\
&= A^{(t+1)}_{p,q}.
\end{aligned}$$

Nun bleibt noch zu zeigen, dass auch $P^{(t)}_{p,E_i} = \sum_{r \in E_i} A^{(t)}_{p,r}$ für transiente p und ergodische E_i gilt. Wir gehen auch hier wieder induktiv vor. $P^{(0)}_{p,E_i} = 0 = \sum_{r \in E_i} A^{(0)}_{p,r}$ gilt, ebenso wie

$P^{(1)}_{p,E_i} = \sum_{r\in E_i} A^{(1)}_{p,r}$ nach Definition von P. Es gelte also $P^{(t)}_{p,E_i} = \sum_{r\in E_i} A^{(t)}_{p,r}$ für alle transienten Zustände p und alle ergodischen Komponenten E_i. Hier folgt

$$
\begin{aligned}
P^{(t+1)}_{p,E_i} &= \sum_{s\in Q^\diamond} P^{(t)}_{p,s} P_{s,E_i} \\
&= \sum_{\substack{s\,\in\, Q^\diamond,\\ s \text{ transient}}} P^{(t)}_{p,s} P_{s,E_i} + \sum_{j=1}^{m} P^{(t)}_{p,E_j} \underbrace{P_{E_j,E_i}}_{\substack{1 \quad i=j\\ 0 \quad i\neq j}} \\
&= \sum_{\substack{s\,\in\, Q,\\ s \text{ transient}}} A^{(t)}_{p,s} \sum_{r\in E_i} A_{s,r} + \sum_{s\in E_i} A^{(t)}_{p,s} \underbrace{\sum_{r\in E_i} A_{s,r}}_{1} + \sum_{\substack{s\,\in\, Q\setminus E_i,\\ s \text{ ergodisch}}} A^{(t)}_{p,s} \underbrace{\sum_{r\in E_i} A_{s,r}}_{0} \\
&= \sum_{r\in E_i} \sum_{\substack{s\,\in\, Q,\\ s \text{ transient}}} A^{(t)}_{p,s} A_{s,r} + \sum_{r\in E_i}\sum_{s\in E_i} A^{(t)}_{p,s} A_{s,r} + \sum_{r\in E_i} \sum_{\substack{s\,\in\, Q\setminus E_i,\\ s \text{ ergodisch}}} A^{(t)}_{p,s} A_{s,r} \\
&= \sum_{r\in E_i}\sum_{s\in Q} A^{(t)}_{p,s} A_{s,r} \\
&= \sum_{r\in E_i} A^{(t+1)}_{p,r}.
\end{aligned}
$$

■

Wir können die Analyse einer allgemeinen Markovekette also vorerst wie für eine absorbierende Kette bis zur Absorption durchführen. Sind wir einmal in einer ergodischen Menge absorbiert worden, können wir unsere Analyse auf diese beschränken und sie so vornehmen, als befänden wir uns in einer ergodischen Kette.

5.2 Absorbierende Markovketten

Wir beschäftigen uns zunächst mit der Frage, wie groß die Wahrscheinlichkeit ist, in einer endlichen Markovkette von einem Zustand q ausgehend irgendwann einmal einen ergodischen (beziehungsweise absorbierenden) Zustand zu erreichen.

Die Ergebnisse dieses Unterabschnitts haben wir wieder [Rob76] und [TK98] entnommen.

Satz 5.23 *Erreichen eines ergodischen Zustands*

In einer endlichen Markovkette strebt für jeden Startzustand die Wahrscheinlichkeit, dass der Prozess nach t Schritten einen ergodischen Zustand erreicht, für $t \to \infty$, gegen 1.

Vor dem Beweis zunächst ein Korollar für absorbierende Markovketten.

Korollar 5.24 *Absorptionswahrscheinlichkeit*

Die Wahrscheinlichkeit der Absorption in einer absorbierenden Markovkette ist 1.

Beweis von Satz 5.23: Da es nach Satz 5.17 für jeden Zustand p mindestens einen erreichbaren ergodischen Zustand q gibt, und da nur endlich viele Zustände existieren, gibt es ein $t \in \mathbb{N}_0$, so dass für jeden transienten Zustand $p \in Q$ ein ergodischer Zustand $q_p \in Q$ und ein

$t_p \leq t$ existiert, mit der Eigenschaft $A^{(t_p)}_{p,q_p} > 0$. Somit ist die Wahrscheinlichkeit, unabhängig vom Startzustand, in höchstens t Schritten einen ergodischen Zustand zu erreichen eine positive Zahl w. Die Wahrscheinlichkeit, nach höchstens t Schritten keinen ergodischen Zustand erreicht zu haben, ist damit $1 - w$. Wegen der Markov-Eigenschaft ist die Wahrscheinlichkeit, innerhalb der nächsten höchstens t Schritten, keinen ergodischen Zustand zu erreichen, wieder $1 - w$. Somit können wir schließen, dass die Wahrscheinlichkeit, nach höchstens kt Schritten keinen ergodischen Zustand zu erreichen $(1 - w)^k$ ist. Da $0 < w \leq 1$ gilt, nähert sich diese Wahrscheinlichkeit an 0 an. ∎

Definition 5.25 *Kanonische Form für eine absorbierende Markovkette*

A sei die Übergangswahrscheinlichkeitsmatrix einer endlichen homogenen absorbierenden Markovkette. Wenn wir die Zustände der Markovkette so sortieren, dass alle absorbierenden Zustände zuerst aufgezählt werden, dann ist A in kanonischer Form und hat folgendes Aussehen.

$$A = \left(\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{R} & \mathbf{S} \end{array}\right)$$

Die Einheitsmatrix I steht für die Übergänge der absorbierenden Zustände zu sich selbst, die Nullmatrix besagt, dass keine Übergänge von einem absorbierenden zu einem transienten Zustand existieren. R stellt die Übergänge von transienten zu absorbierenden Zuständen dar. In der quadratischen Matrix S sind die Übergänge der transienten Zustände untereinander angegeben.
Es gilt nach Satz 5.23: $\mathbf{S}^t \to \mathbf{0}$ komponentenweise für $t \to \infty$.

Lemma 5.26 *Potenzen von A*

A sei eine Übergangswahrscheinlichkeitsmatrix einer absorbierenden Markovkette in kanonischer Form. Dann hat A^t die Form

$$A^t = \left(\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{R}_{(t)} & \mathbf{S}^t \end{array}\right),$$

wobei $\mathbf{R}_{(t)}$ eine Matrix ist, für die $\mathbf{R}_{(t)} = (\mathbf{I} + \mathbf{S} + \mathbf{S}^2 + \ldots + \mathbf{S}^{t-1}) \times \mathbf{R}$ gilt.

Beweis: $B \subseteq Q$ sei die Menge der absorbierenden Zustände und $C := Q \setminus B$ sei die Menge der transienten Zustände. Dann gilt für $q \in B$:

$$A^{(t)}_{q,q} = \sum_{r \in Q} A^{(t-1)}_{q,r} \cdot \underbrace{A_{r,q}}_{1 \Leftrightarrow r=q} = A^{(t-1)}_{q,q} \cdot 1 = A^{(t-2)}_{q,q} = \ldots = A_{q,q} = 1$$

Und da A^t wieder eine stochastische Matrix ist, gilt falls $p \in B$ für alle $q \neq p$: $A^{(t)}_{p,q} = 0$
Wenn wir die Einträge von A^t für zwei transiente (nichtnotwendig verschiedene) $p, q \in C$ betrachten, also solche mit Einträgen in der $\mathbf{S}$-Teilmatrix, dann erhalten wir

$$A^{(t)}_{p,q} = \sum_{r \in Q} A^{(t-1)}_{p,r} \cdot A_{r,q} = \sum_{r \in B} A^{(t-1)}_{p,r} \cdot \underbrace{A_{r,q}}_{0} + \sum_{r \in C} A^{(t-1)}_{p,r} \cdot A_{r,q} = 0 + \sum_{r \in C} \mathbf{S}^{(t-1)}_{p,r} \cdot \mathbf{S}_{r,q} = \mathbf{S}^{(t)}_{p,q}.$$

Für Zustände $p \in C$ und $q \in B$, also Einträge in $\mathbf{R}_{(t)}$, gilt:

$$\begin{aligned}
\mathbf{R}_{(t)p,q} &= \sum_{r\in B} A^{(t-1)}_{p,r} \cdot \underbrace{A_{r,q}}_{1\Leftrightarrow r=q} + \sum_{r\in C} A^{(t-1)}_{p,r} \cdot A_{r,q} \\
&= A^{(t-1)}_{p,q} + \sum_{r\in C} \mathbf{S}^{(t-1)}_{p,r} \cdot \mathbf{R}_{r,q} \\
&= \mathbf{R}_{(t-1)p,q} + (\mathbf{S}^{t-1} \times \mathbf{R})_{p,q} \\
&= \mathbf{R}_{(t-2)p,q} + (\mathbf{S}^{t-2} \times \mathbf{R})_{p,q} + (\mathbf{S}^{t-1} \times \mathbf{R})_{p,q} \\
&= \ldots \\
&= \mathbf{R}_{(1)p,q} + (\mathbf{S}^{1} \times \mathbf{R})_{p,q} + \ldots + (\mathbf{S}^{t-1} \times \mathbf{R})_{p,q} \\
&= ((I + \mathbf{S} + \mathbf{S}^2 + \ldots + \mathbf{S}^{(t-1)}) \times \mathbf{R})_{p,q}.
\end{aligned}$$

■

Wir erhalten sofort den folgenden Satz:

Satz 5.27 *Grenzmatrix einer absorbierenden Markovkette*
A sei eine Übergangswahrscheinlichkeitsmatrix einer absorbierenden Markovkette in kanonischer Form. Dann hat der komponentenweise Grenzwert von A^t die Form

$$\lim_{t\to\infty} A^t = \left(\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{R}_\infty & \mathbf{0} \end{array}\right),$$

wobei $\mathbf{R}_\infty$ eine nichtnegative Matrix mit Zeilensumme 1 ist.

Beweis: Zu zeigen ist nur noch, dass der komponentenweise Grenzwert

$$\mathbf{R}_\infty = \lim_{t\to\infty} \mathbf{R}_{(t)} = \lim_{t\to\infty} (\sum_{i=0}^{t} \mathbf{S}^i) \times \mathbf{R}$$

existiert.

Wir wissen aus Satz 5.23 bereits, dass $\lim_{t\to\infty} \mathbf{S}^t \to \mathbf{0}$ komponentenweise gilt. Wir benötigen außerdem den Beweis zu den zwei folgenden Behauptungen:

- Die Matrix $(\mathbf{I} - \mathbf{S})$ hat eine Inverse.
- Es gilt $(\mathbf{I} - \mathbf{S})^{-1} = \mathbf{I} + \mathbf{S} + \mathbf{S}^2 + \ldots = \sum_{i=0}^{\infty} \mathbf{S}^i$.

Zunächst erinnern wir uns an die Eigenschaft der Determinanten: $\det(A\times B) = \det(A)\cdot\det(B)$, $\det(\mathbf{I}) = 1$, A hat genau dann eine Inverse, wenn $\det(A) \neq 0$ gilt. Außerdem gilt:

$$\begin{aligned}
(\mathbf{I} - \mathbf{S}) \times (\mathbf{I} + \mathbf{S} + \mathbf{S}^2 + \ldots + \mathbf{S}^{t-1}) &= (\mathbf{I} - \mathbf{S}) + (\mathbf{S} - \mathbf{S}^2) + \ldots + (\mathbf{S}^{t-1} - \mathbf{S}^t) \\
\Rightarrow (\mathbf{I} - \mathbf{S}) \times (\mathbf{I} + \mathbf{S} + \mathbf{S}^2 + \ldots + \mathbf{S}^{t-1}) &= \mathbf{I} - \mathbf{S}^t. \qquad (2)
\end{aligned}$$

Also gilt $\mathbf{I} - \mathbf{S}^t \to \mathbf{I}$. Da $\det(\mathbf{I}) = 1$ und die Determinante eine stetige Funktion ist, gilt für genügend große t: $\det(\mathbf{I} - \mathbf{S}^t) \neq 0$. Weil $\det(A \times B) \neq 0 \Leftrightarrow \det(A) \neq 0 \wedge \det(B) \neq 0$, muss auch $\det(\mathbf{I}-\mathbf{S}) \neq 0$ gelten. Das heißt die Inverse existiert. Multiplizieren wir beide Seiten von (2) von links mit $(\mathbf{I}-\mathbf{S})^{-1}$ dann erhalten wir

$$\mathbf{I}+\mathbf{S}+\mathbf{S}^2+\ldots+\mathbf{S}^{t-1} = (\mathbf{I}-\mathbf{S})^{-1} \times (\mathbf{I}-\mathbf{S}^t).$$

Und für $t \to \infty$ erhalten wir $(\mathbf{I}-\mathbf{S})^{-1} = \sum_{i=0}^{\infty} \mathbf{S}^i$.
Die beiden Behauptungen stimmen also und wir folgern, dass $\mathbf{R}_\infty = (\mathbf{I}-\mathbf{S})^{-1} \times \mathbf{R}$ existiert.

■

Bemerkung: Für einen ausgezeichneten Startzustand $q_0 \in C$, können wir in der Zeile q_0 der Matrix $\mathbf{R}_\infty$ die Wahrscheinlichkeit ablesen, von q_0 ausgehend in einem Zustand $p \in B$ absorbiert zu werden.

5.3 Ergodische Markovketten

Wie bereits gesehen, können wir die ergodischen Komponenten einer allgemeinen Markovkette gut getrennt betrachten und für die asymptotische Analyse der Markovkette später die Wahrscheinlichkeiten der ergodischen Komponente mit der asymptotischen Absorptionswahrscheinlichkeit dieser ergodischen Komponente vom Startzustand aus multiplizieren.
Gehen wir also im Folgenden von ergodischen Markovketten (siehe Definition 5.18) aus. Das Zustandsdiagramm ist somit stark zusammenhängend.

Definition 5.28 *Periode eines Zustands*

Für eine Markovkette mit Übergangswahrscheinlichkeitsmatrix A bilden wir für Zustände p, q die Menge $N_{p,q} := \{t \in I\!N | A^{(t)}_{p,q} > 0\} = \{t \in I\!N | p \text{ kann } q \text{ in } t \text{ Schritten erreichen}\}$. Die Periode des Zustands q ist dann $d(q) := \mathrm{ggt}(N_{q,q})$. Falls $A^{(t)}_{q,q} = 0$ für alle $t \in I\!N$, dann definieren wir $d(q) := 0$.

Wir haben jetzt die Periode eines Zustands einer ergodischen Markovkette eingeführt und wollen auch die Periode einer ergodischen Markovkette definieren. Um zu zeigen, dass diese Definition möglich ist und Sinn macht, benötigen wir eine Folge von Lemmata. (Vergleiche [KS60], Seite 5ff.)

Lemma 5.29 *$N_{q,q}$ ist unter Addition abgeschlossen*

$N_{q,q}$ sei für einen Zustand q wie oben definiert. Dann gilt für $a, b \in N_{q,q}$, dass auch $a+b \in N_{q,q}$ ist.

Beweis: Falls $a \in N_{q,q}$, dann gilt $A^{(a)}_{q,q} > 0$ und für b gilt Entsprechendes. Es folgt wegen der Nichtnegativität von A:

$$A^{(a+b)}_{q,q} = \sum_{r \in q} A^{(a)}_{q,r} \cdot A^{(b)}_{r,q} \geq A^{(a)}_{q,q} \cdot A^{(b)}_{q,q} > 0$$

und $a+b$ ist ein Element von $N_{q,q}$. ■

Lemma 5.30 *Eigenschaften einer unter Addition abgeschlossenen Menge*

Sei $N \subseteq I\!N$ eine unter Addition abgeschlossene Menge, dann enthält N, bis auf endlich viele Ausnahmen, alle Vielfachen von $\text{ggt}(N)$.

Beweis: Wir definieren $d := \text{ggt}(N)$. Falls $d > 1$, teilen wir alle Elemnte von N durch d, so dass wir o.B.d.A. annehmen können, dass $d = 1$ gilt. Wir haben jetzt eine Menge ganzer Zahlen, deren größter gemeinsamer Teiler 1 ist und es gibt auch eine endliche Teilmenge $\{n_1, \ldots, n_k\}$ mit dieser Eigenschaft. Dann gibt es also eine Linearkombination $a_1n_1 + a_2n_2 + \ldots + a_kn_k = 1$ mit ganzzahligen a_i. Sammeln wir alle positiven Summanden zu m einerseits und alle negativen Summanden zu $-n$, andererseits, dann erhalten wir zwei Zahlen $m \in I\!N$ und $n \in I\!N$ mit $m - n = 1$, die wegen der Abgeschlossenheit unter Addition auch Elemente von N sein müssen. Sei $t \geq n(n-1)$. Dann können wir t immer zusammensetzen aus $t = an + b$, wobei $a \geq n - 1$ und $0 \leq b \leq n - 1$ ist. Dann gilt aber auch $t = an + (m - n)b = (a - b)n + bm$ und wegen $m, n \in N$ und der Abgeschlossenheit unter Addition ist $t \in N$. Wenn jedes genügend große t in N liegt, gibt es nur endlich viele natürliche Zahlen, die nicht in N liegen. ∎

Wir wissen jetzt, dass für einen Zustand q mit Periode $d(q)$ in der Menge $N_{q,q}$ nur Vielfache von $d(q)$ liegen und sogar, dass jedes genügend große Vielfache von $d(q)$ in $N_{q,q}$ liegt.
Falls zwei Zustände in derselben kommunizierenden Klasse liegen, also $p \leftrightarrow q$, dann ist $N_{p,q}$ nicht leer. Außerdem haben alle Zustände in dieser kommunizierenden Klasse dieselbe Periode.

Lemma 5.31 *Periode zweier Zustände in derselben kommunizierenden Klasse*

A sei die Übergangswahrscheinlichkeitsmatrix einer homogenen Markovkette mit Zustandsmenge Q. Dann gilt für $p, q \in Q$ falls $p \leftrightarrow q$, dass $d(p) = d(q) =: d$. Außerdem sind alle Elemente aus $N_{p,q}$ kongruent zueinander modulo d.

Beweis: Es gelte $p \leftrightarrow q$. Angenommen $a \in N_{p,q}$, $b \in N_{p,q}$ und $c \in N_{q,p}$.
Zunächst einmal gilt $a + c \in N_{p,p}$, weil $A_{p,p}^{(a+c)} \geq A_{p,q}^{(a)} \cdot A_{q,p}^{(c)} > 0$. Wenn wir dies auf das Zustandsdiagramm übersetzen heißt das, dass es einen Pfad der Länge a von p nach q und einen Pfad der Länge c wieder zurück nach p gibt. Mit der gleichen Überlegung lässt sich zeigen, dass auch $a + kd(q) + c \in N_{p,p}$ für jedes genügend große $k \in I\!N$ gilt, denn es gibt den Pfad von p nach q, man kann mehrfach einen Pfad der Länge $k \cdot d(q) \in N_{q,q}$ von q nach q gehen und dann wieder mit c Schritten in p landen.
Da $a + c$ Vielfaches von $d(p)$ sein muss und für genügend große k auch $a + c + kd(q)$ Vielfaches von $d(p)$ sein muss, muss $d(q)$ Vielfaches von $d(p)$ sein.
All diese Überlegungen führen andererseits auch dazu, dass $d(p)$ ein Vielfaches von $d(q)$ sein muss, woraus $d(q) = d(p)$ folgt.
Nun zeigen wir noch, dass für $a, b \in N_{p,q}$ gilt: $a \equiv b (\text{mod } d)$. Für jedes $c \in N_{q,p}$ gilt $a + c \in N_{q,q}$ und $b + c \in N_{q,q}$. Also sind sowohl $a + c$ als auch $b + c$ durch d teilbar und die Behauptung muss gelten. ∎

Nach der Definition der Periode eines Zustands und den vorangegangenen Lemmata können wir die Periode einer ergodische Markovkette definieren:

Definition 5.32 *Periode einer ergodischen Markovkette*

Für eine ergodische Markovkette und einen beliebigen Zustand q sei $d = d(q)$ die Periode der Markovkette.

Wir nennen eine Markovkette periodisch, wenn sie eine Periode $d > 1$ hat. Falls $d = 1$ ist, dann wird die Markovkette aperiodisch genannt.

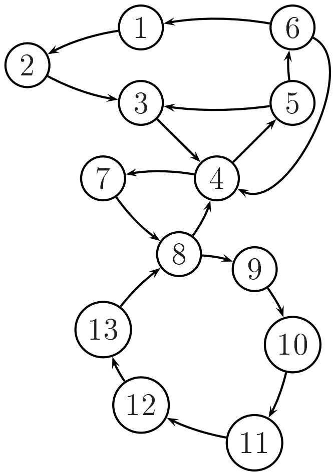

Abbildung 11: Beispiel zur Analyse der Periodizität einer ergodischen Markovkette

Im Beispiel aus Abbildung 11 gilt $N_{1,1} = \{6, 9, 12, 15, \ldots\}$ und wir erhalten somit die Periode $d = \mathrm{ggT}(N_{1,1}) = 3$.

Definition 5.33 *Periodischer Abstand zweier Zustände*
Für zwei Zustände p und q in einer ergodischen Markovkette mit Periode d sei $t_{p,q}$ eine ganze Zahl mit der Eigenschaft $0 \leq t_{p,q} < d$, so dass für jedes Element $a \in N_{p,q}$ gilt $a \equiv t_{p,q} (\mathrm{mod}\ d)$.

Man sieht leicht ein, dass $N_{p,q}$ alle außer endlich vielen Werten $t_{p,q} + kd$ enthält.
Insbesondere gilt $t_{q,q} = 0$ und $t_{p,q} + t_{q,p} \equiv 0 (\mathrm{mod}\ d)$, sowie $t_{p,r} + t_{r,q} \equiv t_{p,q} (\mathrm{mod}\ d)$.
Diese Überlegung führt uns zu einer neuen Äquivalenzrelation $\doteq$ für Zustände, die definiert ist durch $p \doteq q :\Leftrightarrow t_{p,q} = 0$. Reflexivität, Symmetrie und Transitivität sind wegen der Definition von $t_{p,q}$ und den Eigenschaften von $N_{p,q}$ offensichtlich. Wir nennen die Äquivalenzklassen bezüglich $\doteq$ periodische Klassen.

Wir drücken diese Tatsache noch einmal im folgenden Satz aus.

Satz 5.34 *Periodische Klassen*
Die Zustandsmenge Q einer ergodischen Markovkette mit Periode d kann in d periodische Klassen $C_0, \ldots, C_{d-1} \subseteq Q$ partitioniert werden, so dass ausgehend von einem Zustand $p \in C_i$ nach genau t Schritten nur ein Zustand $q \in C_{i+t \bmod d}$ erreicht werden kann.

Bemerkung: Ist $C_0, \ldots, C_{d-1}$ eine solche Partitionierung der Zustandsmenge, dann gilt:

$$\forall p \in C_i, t \in \mathbb{N}_0, q \in Q : \quad A^{(t)}_{p,q} > 0 \Rightarrow q \in C_{i+t \bmod d}.$$

Definition 5.35 *Regelmäßige Markovkette*
Eine Markovkette heißt regelmäßig[2], wenn sie eine ergodische Markovkette ist und Periode $d = 1$ hat, also aperiodisch ist.

[2]Der englische Begriff regular wird hier wie in [Gan86] mit regelmäßig und nicht mit regulär übersetzt.

Bemerkung: Falls eine Markovkette regelmäßig ist, dann fallen alle Zustände in eine einzige periodische Klasse. Aus der Definition der Periode d und der Definition und den Eigenschaften von $N_{p,q}$ ergibt sich: Eine Markovkette ist genau dann regelmäßig, wenn für ihre Übergangswahrscheinlichkeitsmatrix A gilt: $\exists t \forall p, q \in Q : A^{(t)}_{p,q} > 0$.
Jede ergodische Markovkette, in der es einen Zustand q gibt, für den $A_{q,q} > 0$ gilt, ist regelmäßig. Dies folgt, weil $1 \in N_{q,q}$ und damit $d = d(q) = 1$. Die Umkehrung gilt im Allgemeinen nicht.

Satz 5.36 *Grenzverteilung einer regelmäßigen Markovkette*
Für die Übergangswahrscheinlichkeitsmatrix einer regelmäßigen Markovkette existiert der (komponentenweise) Grenzwert

$$W = \lim_{t\to\infty} A^t = \begin{pmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \vdots & \vdots & & \vdots \\ w_1 & w_2 & \dots & w_n \end{pmatrix},$$

wobei alle Einträge der Matrix W positiv sind.

Um den Beweis zu führen, benötigen wir zunächst folgendes Lemma.

Lemma 5.37 *Kontraktion bei Matrix-Vektormultiplikation*
Angenommen P ist eine stochastische $n \times n$ Matrix, dessen kleinster Eintrag ϵ ist,[3] und es sei

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

ein n-dimensionaler Spaltenvektor. Nehmen wir an, dass m_0 der kleinste und M_0 der größte Eintrag von x ist. Gleichsam sei m_1 der kleinste und M_1 der größte Eintrag des Spaltenvektors Px.
Dann gilt $M_1 \leq M_0$, sowie $m_1 \geq m_0$ und $M_1 - m_1 \leq (1 - 2\epsilon)(M_0 - m_0)$.

Beweis: Wir definieren x' von x ausgehend, indem wir alle Einträge außer einem m_0 durch M_0 ersetzen. Dann gilt komponentenweise $x' \geq x$. Da jede Zeile in P Zeilensumme 1 hat, ist jede Komponente von Px' von der Form $am_0 + (1 - a)M_0 = M_0 - a(M_0 - m_0)$, wobei $a \geq \epsilon$. Somit gilt für jede Komponente von Px', dass sie $\leq M_0 - \epsilon(M_0 - m_0)$ ist. Da $x \leq x'$, gilt auch $Px \leq Px'$ komponentenweise und wir schließen, dass

$$M_1 \leq M_0 - \epsilon(M_0 - m_0). \tag{3}$$

Woraus wir wegen $M_0 - m_0 \geq 0$ und $\epsilon \geq 0$ wiederum schließen können, dass $M_1 \leq M_0$ gilt. Wenden wir dieses Vorgehen auf den Vektor $-x$ an, dann erhalten wir

$$-m_1 \leq -m_0 - \epsilon(-m_0 + M_0). \tag{4}$$

[3] ϵ kann den Wert 0 annehmen.

Und somit $m_1 \geq m_0$. Wenn wir die Gleichungen (3) und (4) addieren, dann erhalten wir

$$M_1 - m_1 \leq M_0 - m_0 - 2\epsilon(M_0 - m_0) = (1 - 2\epsilon)(M_0 - m_0).$$

■

Beweis von Satz 5.36: A sei die Übergangswahrscheinlichkeitsmatrix der regelmäßigen Markovkette mit Zustandsmenge $Q = \{1, 2, \ldots, n\}$. Falls wir nur einen Zustand haben, sind die Ergebnisse trivial. Nehmen wir also $n \geq 2$ an. Ausserdem nehmen wir an, dass $P = A^k$ für ein genügend großes k eine stochastische Matrix mit positiven Einträgen ist, was wegen der Regelmäßigkeit der Markovkette möglich ist. ϵ sei der kleinste Eintrag in P. Da wir mindestens zwei Einträge in jeder Zeile von P haben, muss $0 < \epsilon \leq \frac{1}{2}$ gelten. e_q sei der Einheitsvektor, der nur für q den Eintrag 1 hat.

$$e_q = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$P^t e_q$ habe den maximalen Eintrag M_t und den minimalen Eintrag m_t. Da $P^t e_q = PP^{t-1}e_q$, folgt aus dem Lemma, dass

$$\begin{aligned} M_1 \geq M_2 \geq M_3 \geq \ldots \\ m_1 \leq m_2 \leq m_3 \leq \ldots \end{aligned}$$

und

$$M_t - m_t \leq (1 - 2\epsilon)(M_{t-1} - m_{t-1}).$$

Als Kürzel verwenden wir $d_t := M_t - m_t$ und es ergibt sich

$$d_t \leq (1 - 2\epsilon)^{t-1} d_1.$$

Aus $0 < \epsilon \leq \frac{1}{2}$ folgt, dass $d_t \to 0$ für $t \to \infty$. Also auch $M_t - m_t \to 0$ und somit strebt $P^t e_q$ gegen einen Spaltenvektor, dessen Einträge alle gleich sind, wir nennen den Wert w_q. Weil $P^t e_q$ die q-te Spalte von P^t ist, schließen wir, dass P^t gegen ein Matrix W strebt mit

$$W = \begin{pmatrix} w_1 & w_2 & \ldots & w_n \\ w_1 & w_2 & \ldots & w_n \\ \vdots & \vdots & & \vdots \\ w_1 & w_2 & \ldots & w_n \end{pmatrix}.$$

Auch W ist eine stochastische Matrix. Die Einträge von W sind alle positiv, da für jedes t und jedes q gilt: $0 < m_1 \leq m_t \leq w_q$.
Die gleichen Überlegungen gelten für A^t, wobei wir jedoch nur für jeden k-ten Schritt die $(1-2\epsilon)$-Kontraktion garantieren können, gleichzeitig ist aber $M_t - m_t$ auch für $A^t e_q$ zumindest eine nicht aufsteigende Folge. Im Grenzwert erhalten wir

$$\lim_{t\to\infty} A^t = \lim_{t\to\infty} P^t = W.$$

■

Satz 5.38 *Grenzverteilungen periodischer ergodischer Markovketten*
Für eine ergodische Markovkette mit Periode d und Übergangswahrscheinlichkeitsmatrix A existieren für $0 \leq j < d$ die Grenzmatrizen

$$\overset{(j)}{A} = A^j A^\infty = \lim_{t\to\infty} A^{j+dt}$$

wobei $A^\infty := \lim_{t\to\infty}(A^d)^t$.

Beweis: Für jeden Zustand q der ergodischen Markovkette gilt $d = d(q)$. Für $0 \leq i < d$ sei C_i eine periodische Klasse. Dann gilt für jedes feste C_i für alle $p, q \in C_i$, dass in $N_{p,q}$ jedes genügend große Vielfache von d vorkommt, somit gilt für jedes genügend große Vielfache kd von d für alle $p, q \in C_i$: $A^{(kd)}_{p,q} > 0$, während für p, q mit $p \in C_i$ und $q \in C_j$ mit $i \neq j$ $A^{(kd)}_{p,q} = 0$ gilt.
Wir können A^{kd} jeweils eingeschränkt auf eine periodische Klasse auch als Übergangswahrscheinlichkeitsmatrix einer regelmäßigen Kette betrachten. Bei entsprechender Sortierung nach periodischen Klassen hat die Matrix A^{kd} folgendes Aussehen:

$$A^{kd} = \begin{pmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & P_d \end{pmatrix}.$$

Mit $P_i > 0$ für alle Einträge. Nach diesen Überlegungen existiert die Grenzmatrix für Potenzen von A^d. Denn wie man leicht nachrechnen kann, haben die Potenzen von A^{kd} die Form

$$(A^{kd})^t = \begin{pmatrix} P_1^t & 0 & \cdots & 0 \\ 0 & P_2^t & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & P_d^t \end{pmatrix}.$$

Wir definieren die Grenzmatrix als

$$A^\infty := \lim_{t\to\infty}(A^d)^t.$$

Für $0 \leq j < d$ muss die Matrixmultiplikation nur entsprechend geklammert werden. ∎

5.4 Allgemeine Markovketten

Allgemeine Markovketten bestehen aus ergodischen und transienten Zuständen, die sich noch durch die Äquivalenzrelation $\leftrightarrow$ in kommunizierende Klassen unterteilen lassen. Wir sprechen im Folgenden von den transienten Komponenten und den ergodischen Komponenten.
Die Übergangswahrscheinlichkeitsmatrix einer allgemeinen Markovkette lässt sich durch Umsortierung der Zustände in eine kanonische Form bringen. Zuerst werden die m vielen ergodischen Komponenten aufgeführt, danach die transienten Komponenten in umgekehrter Reihenfolge ihrer topologischen Sortierung im Kondensationsgraphen. Die Matrix hat dann die folgende Form:

$$A = \begin{pmatrix} A_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A_m & 0 & \cdots & 0 \\ A_{m+1,1} & A_{m+1,2} & \cdots & A_{m+1,m} & A_{m+1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,m} & A_{n,m+1} & \cdots & A_n \end{pmatrix}.$$

Für $1 \leq i \leq m$ ist A_i die Wahrscheinlichkeitsübergangsmatrix der i-ten ergodischen Komponente. Eine Matrix A_i für $i > m$ beschreibt die Übergänge innerhalb einer transienten Komponente, hier ist die Zeilensumme < 1. Eine doppelt indizierte Matrix $A_{i,j}$ mit $i > j$ steht für die Übergange von der i-ten Komponente zur j-ten Komponente. Hier ist die Zeilensumme ≤ 1. Für eine Potenz A^t gilt

$$A^t = \begin{pmatrix} A_1^t & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & A_2^t & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A_m^t & 0 & \cdots & 0 \\ * & * & \cdots & * & * & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ * & * & \cdots & * & * & \cdots & * \end{pmatrix}.$$

Dies folgt wie bei der Betrachtung in Lemma 5.26. Ebenso folgt, dass die Wahrscheinlichkeit eines Übergangs von einem transienten Zustand p zu einem anderen transienten Zustand q für $t \to \infty$ gegen 0 geht, also $A_{p,q}^{(t)} \to 0$.

Die Ergebnisse im Rest dieses Abschnitts haben wir zum Teil [TK98] und [Gan86] entnommen, sie aber mit unseren Methoden selbst hergeleitet.

Satz 5.39 *Absorptionswahrscheinlichkeit ergodischer Komponenten*
Entsprechend der Sortierung in der kanonischen Form seien $E_1, E_2, \ldots, E_m$ die ergodischen Komponenten. q_0 sei ein ausgezeichneter Startzustand. Dann existieren für $1 \leq i \leq m$ die Grenzwerte

$$r_i := \lim_{t\to\infty} \sum_{p \in E_i} A_{q_0,p}^{(t)}.$$

Beweis: Wenn wir wie in Definition 5.21 die zu A gehörige Absorptionsmatrix P betrachten, die mit den transienten Zuständen und den ergodischen Komponenten indiziert ist, dann erhalten wir nach Satz 5.27, dass $\lim_{t\to\infty} P^t$ existiert. Nach Satz 5.22 gilt dann $\sum_{r\in E_i} A_{p,r}^{(t)} = P_{p,E_i}^{(t)}$ und somit auch

$$\lim_{t\to\infty} \sum_{r\in E_i} A_{q_0,r}^{(t)} = \lim_{t\to\infty} P_{q_0,E_i}^{(t)}.$$

■

Wir bezeichnen r_i auch als Absorptionswahrscheinlichkeit der ergodischen Komponente E_i von q_0 aus. Der Startzustand q_0 kann transient oder ergodisch sein. Im ergodischen Fall ist r_i dann 0 oder 1, je nachdem, ob $q_0 \in E_i$.

Lemma 5.40 *Grenzmatrix einer Markovkette mit regelmäßigen Komponenten*
Sei A die Übergangswahrscheinlichkeitsmatrix einer Markovkette, deren ergodische Komponenten $E_1, E_2, \ldots, E_m$ alle die Periode $d_1 = d_2 = \ldots = d_m = 1$ haben. Dann existiert die Grenzmatrix

$$A^\infty := \lim_{t\to\infty} A^t.$$

Beweis: A sei in kanonischer Form. Dann ist jedes A_i für $1 \leq i \leq m$ Übergangswahrscheinlichkeitsmatrix einer regelmäßigen Markovkette. Das heißt nach Satz 5.36 $\lim_{t\to\infty}(A_i^t)_{p,q} = w_q$ für $p, q \in E_i$. Wir betrachten Potenzen A^t von A in kanonischer Form und deren komponentenweisen Grenzwert für $t \to \infty$

- Für p, q transient folgt
$$A_{p,q}^{(t)} \to 0$$
aus Satz 5.23.

- Für p ergodisch und q transient:
$$A_{p,q}^{(t)} = 0.$$
Dies läßt sich schon an der kanonischen Form der Matrix ablesen.

- Für p, q ergodisch, $p \in E_i$, $q \in E_j$
$$A_{p,q}^{(t)} \begin{cases} = 0 & i \neq j \\ \to w_q & i = j \end{cases}.$$
Dies ergibt sich aus den Überlegungen am Anfang des Beweises.

- Für p transient und $q \in E_i$ ergodisch:
$$\begin{aligned} A_{p,q}^{(2t)} &= \sum_{s\in Q} A_{p,s}^{(t)} A_{s,q}^{(t)} \\ &= \sum_{s \text{ transient}} \underbrace{A_{p,s}^{(t)}}_{\to 0} \underbrace{A_{s,q}^{(t)}}_{\leq 1} + \sum_{\substack{s \in Q\setminus E_i,\\ s \text{ ergodisch}}} A_{p,s}^{(t)} \underbrace{A_{s,q}^{(t)}}_{0} + \sum_{s\in E_i} A_{p,s}^{(t)} \underbrace{A_{s,q}^{(t)}}_{\to w_q} \\ &\to w_q \lim_{t\to\infty} \sum_{s\in E_i} A_{p,s}^{(t)} \\ &= w_q r_i. \end{aligned}$$
Hier ist r_i die Absorptionswahrscheinlichkeit der ergodischen Menge E_i von p aus.

■

Satz 5.41 *„Periodizität" allgemeiner Markovketten*
Für eine Markovkette sei die Übergangswahrscheinlichkeitsmatrix A gegeben. Dann gibt es eine ganze Zahl d, so dass der Grenzwert

$$\lim_{t\to\infty} (A^d)^t$$

existiert.

Beweis: A sei in kanonischer Form. $E_1, E_2, \ldots, E_m$ seien die ergodischen Komponenten der Markovkette. $d_1, d_2, \ldots, d_m$ seien die Perioden der ergodischen Komponenten, also $d_i = d(q)$ für ein $q \in E_i$. Die Zahl d definieren wir als $d := \mathrm{kgV}\{d_1, d_2, \ldots, d_m\}$.
Die Matrix A^d ist dann die Übergangswahrscheinlichkeitsmatrix einer Markovkette, die nur ergodische Komponenten mit Periode 1 hat.

$$A^d = \begin{pmatrix} A_1^d & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & A_2^d & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & A_m^d & 0 & \cdots & 0 \\ * & * & \cdots & * & * & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ * & * & \cdots & * & * & \cdots & * \end{pmatrix}$$

So wie im Beweis zu Satz 5.38 werden im ergodischen Teil die periodischen Klassen der ursprünglichen Markovkette zu ergodischen Komponenten der d-ten Potenz der Markovkette. Die Übergänge von transienten zu ergodischen Zuständen, sowie von transienten zu transienten Zuständen in der d-ten Potenz sind nur von qualitativer Bedeutung, ihre tatsächliche Wahrscheinlichkeit spielt für die Grenzwertbetrachtung keine Rolle.
Wir können also Lemma 5.40 auf A^d anwenden und der Grenzwert muss existieren.

$$A^\infty := \lim_{t\to\infty} A^{td}$$

■

Korollar 5.42 *Grenzverteilungen allgemeiner Markovketten*
Für eine Markovkette sei A die Übergangswahrscheinlichkeitsmatrix. Dann gibt es eine ganze Zahl d, so dass für $0 \leq j < d$ die Grenzmatrizen

$$\overset{(j)}{A} = A^j A^\infty = \lim_{t\to\infty} A^{j+dt}$$

existieren, wobei $A^\infty = \lim_{t\to\infty}(A^d)^t$.

Beweis: Das Korollar ergibt sich direkt durch entsprechende Klammerung des Matrixprodukts und Anwendung des Satzes 5.41. ■

Definition 5.43 *Landau-Funktion $F(n)$*
$F(n)$ sei die Landau-Funktion $F(n) := \max\{\mathrm{kgV}\{x_1, \ldots, x_m\} | m \in I\!N \wedge x_1, \ldots, x_m \in I\!N \wedge x_1 + \ldots + x_m = n\}$.

Bemerkung: Es gilt auch $F(n)) = \max\{\mathrm{kgV}\{x_1, \ldots, x_m\} | m \in I\!N \wedge x_1, \ldots, x_m \in I\!N \wedge x_1 + \ldots + x_m \leq n\}$, da $x_i = 1$ in der Definition auch zulässig ist.
Landau hat die Funktion $F(n)$ untersucht [Lan03] und gezeigt, dass die Abschätzung $F(n) = \Theta(e^{\sqrt{n \ln n}})$ gilt.

Satz 5.44 *Größe der Periodizität einer allgemeinen Markovkette*
Für eine Markovkette mit n Zuständen und d aus Satz 5.41 gilt $d \leq F(n)$.

Beweis: d kann als das kleinste gemeinsame Vielfache der Perioden aller ergodischen Komponenten der Markovkette gewählt werden. Für die Periode d_i einer ergodischen Komponente E_i gilt $d_i \leq |E_i|$, da in jeder der d_i periodischen Klassen mindestens ein Zustand sein muss. Ausserdem gilt wegen $\bigcup_{i=1}^{m} E_i \subseteq Q$ auch $\sum_{i=1}^{m} d_i \leq \sum_{i=1}^{m} |E_i| \leq |Q| = n$. Und somit folgt $d \leq F(n)$. ∎

6 Markovketten und unäre stochastische Automaten

Den Zusammenhang zwischen unären PFAs und homogenen Markovketten herzustellen, ist nicht schwer. In der Definition eines PFA gibt es im unären Fall nur eine stochastische Übergangsmatrix. Betrachten wir diese als Übergangswahrscheinlichkeitsmatrix einer homogenen Markovkette, dann können wir sehr einfach unsere Ergebnisse zur Periodizität homogener Markovketten auf stochastische Automaten übertragen. Im Folgenden habe ein unärer PFA M die vereinfachte Beschreibung $M = (Q, A, \pi, \eta)$.

6.1 Die Arbeit von Milani und Pighizzini

Milani und Pighizzini ist es in ihrer Arbeit *Tight bounds on the simulation of unary probabilistic automata by deterministic automata* [MP00] gelungen, eine obere Schranke für die Größe der ultimativen Periodizität einer unären Sprache, die von einem PFA mit isoliertem Cutpoint akzeptiert wird, zu zeigen.

Satz 6.1 *Ultimative Periodizität von $T(M, \lambda)$*

Gegeben sei ein unärer PFA M mit n Zuständen. $T(M, \lambda)$ sei die Sprache, die mit isoliertem Cutpoint λ von M akzeptiert wird. Dann gibt es eine ganze Zahl $d = O(e^{\sqrt{n \ln n}})$, so dass für jedes m, das genügend groß ist, $1^m \in T(M, \lambda)$ genau dann gilt, wenn $1^{m+d} \in T(M, \lambda)$ gilt. Das heißt $T(M, \lambda)$ ist ultimativ d-periodisch.

Beweis: Der PFA sei $M = (Q, A, \pi, \eta)$. Es existieren nach Korollar 5.42 für ein $d \leq F(|Q|)$ und $j = 0, 1, \ldots, d-1$ die Grenzmatrizen

$$\overset{(j)}{A} = A^j A^\infty = \lim_{t\to\infty} A^{j+dt}$$

mit $A^\infty = \lim_{t\to\infty}(A^d)^t$.
Also gibt es für jedes $\epsilon > 0$ eine ganze Zahl μ, so dass für jedes $m \geq \mu$ gilt, dass

$$|\pi A^m \eta - \pi \overset{(m \bmod d)}{A} \eta| < \epsilon.$$

Wir nehmen an, dass der Cutpoint ϵ-isoliert ist, also für jedes $m \in I\!N_0$: $|\pi A^m \eta - \lambda| \geq \epsilon$ gilt. Daraus folgt für jedes $m \geq \mu$:

$$\pi A^m \eta > \lambda \quad \Leftrightarrow \quad \pi \overset{(m \bmod d)}{A} \eta > \lambda.$$

Und somit folgt, da $(m + d) \bmod d = m \bmod d$, dass für jedes $m \geq \mu$

$$\pi A^m \eta > \lambda \quad \Leftrightarrow \quad \pi A^{m+d} \eta > \lambda,$$

oder anders ausgedrückt:

$$1^m \in T(M, \lambda) \quad \Leftrightarrow \quad 1^{m+d} \in T(M, \lambda).$$

Somit ist die Sprache ultimativ d-periodisch. ■

Korollar 6.2 *Obere Schranke für den zyklischen Teil eines DFA*
Zu jedem PFA der eine unäre Sprache mit isoliertem Cutpoint akzeptiert, existiert ein äquivalenter DFA, der im zyklischen Teil höchstens $F(n)$ Zustände hat.

Beweis: Dies ist eine direkte Konsequenz aus Satz 6.1 und Satz 3.33. ■

Außerdem ist diese Schranke scharf.

Satz 6.3 *Untere Schranke für die ultimative Periodizität*
Für jedes $n \in I\!N$ gibt es einen unären PFA mit isoliertem Cutpoint und n Zuständen, der eine Sprache mit minimaler ultimativer Periodizität $F(n)$ akzeptiert.

Beweis: Für gegebenes n sei die zu akzeptierende Sprache $L_n := (1^{F(n)})^*$. Die Sprache hat die minimale ultimative Periodizität $F(n)$. Wäre die ultimative Periodizität $\chi < F(n)$, dann müsste es auch ein $k \in I\!N_0$ mit $1^{k \cdot F(n)} \in L_n$ und $1^{k \cdot F(n)+\chi} \in L_n$ geben, dies widerspricht aber der Definition von L_n.

Nach der Definition von $F(n)$ (Definition 5.43) muss es ein Tupel natürlicher Zahlen $(d_1, \ldots, d_s)$ geben, für die $F(n) = \text{kgv}\{d_1, \ldots, d_s\}$ und $d_1 + \ldots + d_s = n$ gilt.
Wir beschreiben den PFA, der die Sprache $(1^{F(n)})^*$ akzeptiert, informell. Abbildung 12 zeigt das Zustandsdiagramm des PFA.
Für jedes d_i gibt es im PFA einen Zyklus E_i der Länge d_i. Der Zyklus E_i enthalte die Zustände $q_{i,0}, q_{i,1}, \ldots, q_{i,d_i-1}$, so dass vom einen zum nächsten ein deterministischer Übergang definiert ist: $A_{q_{i,j}, q_{i,(j+1) \bmod d_i}} = 1$ für $0 \leq j < d_i$. Die akzeptierenden Zustände seien für $1 \leq i \leq k$ die $q_{i,0}$. Die Startverteilung sei so, dass $\pi_q = \frac{1}{k}$ für alle $q = q_{i,0}$ und 0 sonst. Betrachten wir ein

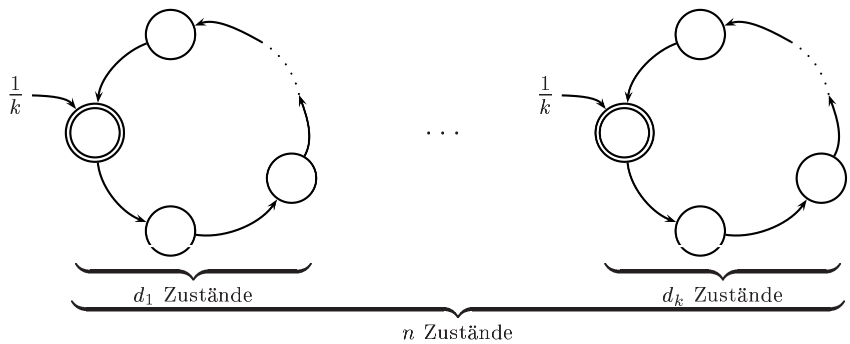

Abbildung 12: Der unäre PFA für den Beweis der unteren Schranke der ultimativen Periodizität

Wort 1^m der Länge m. Falls m ein Vielfaches von $F(n)$ ist, dann ist es auch Vielfaches von jedem d_i und das Wort wird mit Wahrscheinlichkeit 1 akzeptiert. Ist m kein Vielfaches von d, dann gibt es mindestens einen ergodischen Zyklus, der das Wort 1^m nicht akzeptiert. Also ist die Akzeptanzwahrscheinlichkeit $\leq 1 - \frac{1}{k}$. Wählen wir nun den Cutpoint $\lambda = 1 - \frac{1}{2k}$, dann ist dieser offensichtlich $\frac{1}{2k}$ isoliert und die akzeptierte Sprache ist wie gefordert $(1^{F(n)})^*$. ■

Korollar 6.4 *Untere Schranke für den zyklischen Teil eines DFA*
Für jedes $n \in I\!N$ gibt es einen unären PFA mit isoliertem Cutpoint und n Zuständen, so dass jeder äquivalente DFA mindestens $F(n)$ Zustände in seinem zyklischen Teil haben muss.

Beweis: Der PFA sei wie im Beweis zu Satz 6.3 definiert. Dann ist $L_n = (1^{F(n)})^*$ die akzeptierte Sprache. Die Sprache ist minimal ultimativ $F(n)$ periodisch und nach Satz 3.33 muss auch der zyklische Teil des minimalen DFA, der L_n akzeptiert, $F(n)$ Zustände haben. ■

6.2 Die Arbeit von Mereghetti, Palano und Pighizzini

Mereghetti, Palano und Pighizzini haben in ihrer Arbeit *On the succinctness of deterministic, nondeterministic, probabilistic and quantum finite automata* [MPP01] unter anderem den folgenden Satz bewiesen.

Satz 6.5 *Größe eines unären PFA für* L_k
Für jedes $k = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_m^{\alpha_m}$ mit p_i prim und $\alpha_i \in \mathbb{N}$ ist die notwendige und ausreichende Zahl der Zustände eines PFA, der die Sprache $L_k = \{1^{lk} | l \in \mathbb{N}_0\}$ mit isoliertem Cutpoint akzeptiert, $p_1^{\alpha_1} + p_2^{\alpha_2} + \ldots + p_m^{\alpha_m}$.

Aus dem Beweis zu Satz 6.3 wissen wir bereits, dass $p_1^{\alpha_1} + p_2^{\alpha_2} + \ldots + p_m^{\alpha_m}$ Zustände für den PFA ausreichen. Die untere Schranke erhalten wir in dieser Arbeit als Satz 7.6.

7 Eine obere Schranke für die Periodizität, die von ϵ abhängt

Die Schranke $\Theta(e^{\sqrt{n \ln n}})$ für die Größe des zyklischen Teils eines äquivalenten DFA von Milani und Pighizzini berücksichtigt nicht die Größe der ϵ-Isolation, wie dies zum Beispiel die obere Schranke von Rabin tut.
Wir werden zeigen, dass es auch im unären Fall eine Schranke für die Größe des zyklischen Teils des DFA gibt, die wesentlich von der Größe von ϵ abhängt.

7.1 Die Bedeutung der ϵ-Lücke

An dem Beweis für die untere Schranke der ultimativen Periodizität (Satz 6.3) sieht man bereits, dass für große Zustandszahlen die Isolation deutlich kleiner wird. Jeder ergodische Zyklus im konstruierten PFA hat eine Absorptionswahrscheinlichkeit von mindestens dem Doppelten der Isolationslücke.
Die Vermutung, dass dies auch nötig ist, wird sich bestätigen. Das heißt, dass jede ergodische Komponente, die etwas zur Gesamtperiode beiträgt, mindestens Absorptionswahrscheinlichkeit 2ϵ hat und es somit andererseits höchstens $\frac{1}{2\epsilon}$ viele ergodischen Komponenten geben kann, deren Periode etwas zur Gesamtperiode beitragen.
Intuitiv läßt sich dies dadurch erklären, dass eine ergodische Komponente, die eine kleinere Absorptionswahrscheinlichkeit als 2ϵ hat, und deren Periode teilerfremd zu allen anderen Perioden ist, entweder gar keinen Einfluss auf die Akzeptanz des Wortes hat, oder aber bei dem Versuch, die Akzeptanzwahrscheinlichkeit zum Beispiel von einem Wert unter $\lambda - \epsilon$, der sich schon durch die anderen ergodischen Komponenten ergeben hat, auf einen Wert oberhalb des Cutpoints zu heben, in den isolierten (verbotenen) Bereich kommt.

Zunächst benötigen wir noch zwei Definitionen

Definition 7.1 *Absorptionswahrscheinlichkeit einer ergodischen Komponente im PFA*
$M = (Q, A, \pi, \eta)$ sei ein unärer PFA. Wir betrachten die zur stochastischen Matrix A gehörige Markovkette. Die Markovkette habe m viele ergodische Komponenten $E_i \subseteq Q$. Die Absorptionswahrscheinlichkeit einer ergodischen Komponente E_i im PFA sei

$$r_i := \lim_{t\to\infty} \sum_{q\in Q, p\in E_i} \pi_q \cdot A^{(t)}_{q,p} = \lim_{t\to\infty} \sum_{p\in E_i} (\pi A^t)_p.$$

Bemerkung: Der Grenzwert existiert nach denselben Überlegungen wie in Satz 5.39. An dieser Stelle gehen wir jedoch nicht von einem ausgezeichneten Startzustand aus, sondern von der Anfangsverteilung π.

Satz 7.2 *Redundanz der Perioden unwahrscheinlicher ergodischer Komponenten*
Gegeben sei ein unärer PFA $M = (Q, A, \pi, \eta)$. Die zugehörige Markovkette habe m ergodische Komponenten $E_1, \ldots, E_m$ mit jeweiliger Periodizität d_i und Absorptionswahrscheinlichkeit r_i. Für eine Indexmenge $I \subseteq \{1, \ldots, m\}$ möge die folgende Eigenschaft gelten: Hat die Vereinigung ergodischer Komponenten $\bigcup_{i\in I} E_i$ in der Markovkette eines stochastischen ϵ-isolierten Automaten eine Absorptionswahrscheinlichkeit $\sum_{i\in I} r_i < 2\epsilon$, dann ist die Sprache ultimativ D-periodisch mit

$$D := \text{kgV}(\{d_i | i \in \{1, \ldots, m\} \setminus I\}).$$

Beweis: Wir wissen, dass für $d = \mathrm{kgV}(\{d_1, \ldots, d_m\})$, der Grenzwert $A^\infty = \lim_{t\to\infty}(A^d)^t$ existiert und für $j \in \{0, 1, \ldots, d-1\}$ ist

$$\overset{(j)}{A} := A^j A^\infty = \lim_{t\to\infty} A^{j+dt},$$

wobei der Grenzwert komponentenweise zu verstehen ist.
Daraus folgt, dass zu jedem $\delta > 0$ ein $\mu_1 \in I\!N$ existiert, so dass

$$\forall x \geq \mu_1 : \sum_{q\in Q} |(\pi A^x)_q - (\pi \overset{(x \bmod d)}{A})_q| < \delta$$

gilt und wegen Satz 6.1 existiert auch ein $\mu_2 \in I\!N$, so dass

$$\begin{array}{lll} & \forall x \geq \mu_2 : & 1^x \in L \Leftrightarrow 1^{x+d} \in L \\ \Rightarrow & \forall x \geq \mu_2 : & 1^x R_{T(M,\lambda)} 1^{x+d} \end{array}$$

gilt, wobei $R_{T(M,\lambda)}$ die Nerode-Relation ist. Wir wählen

$$\mu := \max\{\mu_1, \mu_2\}. \tag{5}$$

Für eine Indexmenge $I \subseteq \{1, \ldots, m\}$ gelte $\sum_{i\in I} r_i < 2\epsilon$.

E bezeichne die Vereinigung $E := \bigcup_{i\in I} E_i$.

Wir behaupten, dass bereits $D = \mathrm{kgV}(\{d_i | i \in \{1, \ldots, m\} \setminus I\})$ ultimative Periode der Sprache sei.

Wenn $\sum_{i\in I} r_i < 2\epsilon$, dann gibt es auch ein $\delta > 0$, so dass $\sum_{i\in I} r_i < 2\epsilon - 2\delta$ und somit auch $\sum_{q\in E}(\pi A^n)_q < 2\epsilon - 2\delta$ für jedes $n \in I\!N$. Denn die Absorptionswahrscheinlichkeit der ergodischen Komponenten ergibt sich als aufsteigender Limes.

Zu diesem δ wählen wir wie in (5) ein μ.

Da D ein Vielfaches jedes d_i für $i \notin I$ ist, konvergieren die Einträge der Matrizen A^{tD} für die Zustandspaare $(p, q) \in (Q \setminus E) \times (Q \setminus E)$ nach den gleichen Überlegungen wie im Beweis zu Satz 5.41, diesmal jedoch eingeschränkt auf die Zustände, die nicht in E liegen. Außerdem gilt $A^{(tD)}_{p,q} = 0$ für $p \in E$ und $q \in Q \setminus E$. Und es folgt

$$\forall x \geq \mu : \sum_{q\notin E} |(\pi A^x)_q - (\pi \overset{(x \bmod D)}{A})_q| < \delta,$$

wobei $q \notin E$ natürlich $q \in Q \setminus E$ bedeutet und im Folgenden Komplemente von Zustandsmengen immer bezüglich Q gemeint sind.

Hieraus ergibt sich mit $(x+D) \bmod D = x \bmod D$ und der Dreiecksungleichung, dass für alle $x \geq \mu$ gilt:

$$2\delta > \sum_{q\notin E} |(\pi A^x)_q - (\pi \overset{(x \bmod D)}{A})_q| + \sum_{q\notin E} |(\pi \overset{((x+D) \bmod D)}{A})_q - (\pi A^{x+D})_q| \geq \sum_{q\notin E} |(\pi A^x)_q - (\pi A^{x+D})_q|.$$

Und da $a^+ \leq |a|$ gilt, folgt

$$\forall x \geq \mu : \sum_{q\notin E} ((\pi A^x)_q - (\pi A^{x+D})_q)^+ < 2\delta. \tag{6}$$

Nehmen wir nun an, dass $d > D$ (und damit d Vielfaches von D) ist, und dass D nicht bereits ultimative Periode der Sprache $T(M, \lambda)$ ist. Dann gibt es ein $y \in I\!N$, so dass

$$y > \mu \quad \wedge \quad 1^y \in T(M, \lambda) \quad \wedge \quad 1^{y+D} \notin T(M, \lambda)$$

gilt. Für ein solches y muss wegen der ϵ-Isoliertheit des Cutpoints λ gelten:

$$\begin{aligned} & \pi A^y \eta \geq \lambda + \epsilon \quad \wedge \quad \pi A^{y+D} \eta \leq \lambda - \epsilon \\ \Rightarrow \quad & \pi A^y \eta - \pi A^{y+D} \eta \geq 2\epsilon. \end{aligned} \tag{7}$$

Gleichzeitig gilt offensichtlich

$$\pi A^y \eta - \pi A^{y+D} \eta = \sum_{f \in Q \wedge \eta_f = 1} (\pi A^y - \pi A^{y+D})_f \leq \sum_{f \in Q \wedge \eta_f = 1} (\pi A^y - \pi A^{y+D})_f^+ \leq \sum_{q \in Q} (\pi A^y - \pi A^{y+D})_q^+.$$

Zusammen mit (7) folgt

$$\begin{aligned} & \sum_{q \in Q} (\pi A^y - \pi A^{y+D})_q^+ \geq 2\epsilon \\ \Rightarrow \quad & \sum_{q \in E} (\pi A^y - \pi A^{y+D})_q^+ + \underbrace{\sum_{q \notin E} (\pi A^y - \pi A^{y+D})_q^+}_{<2\delta \text{ mit } (6)} \geq 2\epsilon \\ \Rightarrow \quad & \sum_{q \in E} (\pi A^y - \pi A^{y+D})_q^+ > 2\epsilon - 2\delta. \end{aligned} \tag{8}$$

Andererseits waren die Absorptionswahrscheinlichkeiten sowie δ gerade so gewählt, dass

$$\sum_{q \in E} (\pi A^y - \pi A^{y+D})_q^+ \leq \sum_{q \in E} (\pi A^y)_q \leq \sum_{i \in I} r_i < 2\epsilon - 2\delta$$

gilt, denn πA^y und πA^{y+D} sind nichtnegative Zeilenvektoren. Das Ergebnis steht im direkten Widerspruch zu (8). Es kann also kein solches y geben und die Sprache muss bereits ultimativ D-periodisch sein. ■

Wir können somit die Perioden d_i, die zu den ergodischen Komponenten mit zu geringer Wahrscheinlichkeit gehören, aus der ultimativen Periode der Sprache herausnehmen. Wir konkretisieren dieses Vorgehen in den folgenden Korollaren.

Korollar 7.3 *Größe der tatsächlich benötigten ultimativen Periode*

Sind E_i $(i = 1, \ldots, m)$ die ergodischen Komponenten in der Markovkette eines unären stochastischen ϵ-isolierten endlichen Automaten und sind r_i die entsprechenden Absorptionswahrscheinlichkeiten von der Startverteilung ausgehend, sowie d_i die Perioden, dann ist die Sprache ultimativ D-periodisch mit

$$D := \mathrm{ggT}(\{\mathrm{kgV}(\{d_i | i \in I\}) | I \subseteq \{1, \ldots, m\} \wedge \sum_{i \in I} r_i > 1 - 2\epsilon\}).$$

Beweis: Zunächst benötigen wir die Überlegung, dass wenn in einer Menge von Zahlen $\{D_j | j \in J\}$ jedes D_j ultimative Periode der Sprache ist, dann ist jedes D_j auch Vielfaches der minimalen ultimativen Periode. Also muss auch $\max(\{n \in I\!N | \text{jedes } D_j \text{ ist Vielfaches von } n\}) = \mathrm{ggT}(\{D_j | j \in J\})$ Periode der Sprache, beziehungsweise Vielfaches der minimalen Periode der Sprache sein.
Nun zählen wir als D_j einfach die $\mathrm{kgV}(\{d_i | i \in I\})$ für alle $I \subseteq \{1, \ldots, m\}$ mit $\sum_{i \in I} r_i > 1 - 2\epsilon$ auf. Nach Satz 7.2 wissen wir, dass sie alle ultimative Perioden der Sprache sind. (Hier betrachten wir Indexmengen I, die den Komplementen der Indexmengen in Satz 7.2 entsprechen.) Das Korollar ist bewiesen. ■

Wir wissen, dass eine Zahl $n \in I\!N$ eindeutig in ihre Primfaktoren zerlegt werden kann. Dies gibt Anlass zur folgenden Definition.

Definition 7.4 *Primpotenzen*

$I\!P$ sei die Menge der Primzahlen, $\alpha_p(n)$ für $p \in I\!P$ und $n \in I\!N$ sei der Exponent der Primzahl p in der Primfaktorzerlegung der Zahl n ($\alpha_p(n) = 0$, falls p kein Primfaktor von n ist). Es gilt also

$$n = \prod_{p \in I\!P} p^{\alpha_p(n)}.$$

Bemerkung: Dann gilt für eine Primpotenz p^α mit $p \in I\!P$ und $\alpha \in I\!N_0$ und eine natürliche Zahl D: $p^\alpha | D \Leftrightarrow \alpha_p(D) \geq \alpha$. Für eine endliche Menge $M \subset I\!N$ gilt

$$\mathrm{ggT}(M) = \prod_{p \in I\!P} p^{\min\{\alpha_p(n) | n \in M\}}$$

$$\mathrm{kgV}(M) = \prod_{p \in I\!P} p^{\max\{\alpha_p(n) | n \in M\}}.$$

Satz 7.5 *Mindestwahrscheinlichkeit für Primpotenzen in D*

Sei D wie in Korollar 7.3 definiert, ferner sei $p \in I\!P$ und $\alpha \in I\!N_0$. Dann gibt es für jede Primpotenz p^α, die D teilt ein $I \subseteq \{1, \ldots, m\}$, so dass $\sum_{i \in I} r_i \geq 2\epsilon$ und p^α teilt jedes d_i für $i \in I$. Außerdem gilt die Umkehrung: Wenn es für eine Primpotenz p^α ein $I \subseteq \{1, \ldots, m\}$ gibt, so dass $\sum_{i \in I} r_i \geq 2\epsilon$ gilt und p^α teilt jedes d_i für $i \in I$, dann teilt p^α auch D.

$$\forall p^\alpha : \left(p^\alpha | D \quad \Longleftrightarrow \quad \exists I \subseteq \{1, \ldots, m\} : \left(\sum_{i \in I} r_i \geq 2\epsilon \wedge \forall i \in I : p^\alpha | d_i \right) \right)$$

Beweis: Nach den Überlegungen zur Primpotenzzerlegung in Bezug auf den ggT und das kgV, ergibt sich für D aus Korollar 7.3:

$$\alpha_p(D) = \min\{\max\{\alpha_p(d_i) | i \in I\} | I \subseteq \{1, \ldots, m\} \wedge \sum_{i \in I} r_i > 1 - 2\epsilon\} \tag{9}$$

Im Folgenden seien $I, J \subseteq \{1, \ldots, m\}$. Außerdem sei $J^c = \{1, \ldots, m\} \setminus J$ das Komplement von J in der gesamten Indexmenge.

Wir müssen somit zeigen, dass eine Primpotenz p^α die Zahl D genau dann teilt, wenn es ein J gibt, so dass p^α jedes d_i für alle $i \in J$ teilt und $\sum_{i \in J} r_i \geq 2\epsilon$ gilt. Oder anders ausgedrückt: Es ist zu zeigen, dass

$$\forall p \in I\!P, \alpha \in I\!N_0 : \quad \alpha_p(D) \geq \alpha \quad \iff \quad \exists J : \ (\sum_{i \in J} r_i \geq 2\epsilon \ \wedge \ \forall i \in J : \alpha_p(d_i) \geq \alpha). \qquad (10)$$

„$\Rightarrow$“: Wir folgern zunächst eine Aussage (11) aus der linken Seite von (10), nehmen dann an, dass die rechte Seite von (10) nicht gilt und führen dies zum Widerspruch zu Aussage (11).

Es gelte die linke Seite von (10) zusammen mit Gleichung (9).

$$\alpha_p(D) = \min\{\max\{\alpha_p(d_i) | i \in I\} | I \subseteq \{1, \ldots, m\} \wedge \sum_{i \in I} r_i > 1 - 2\epsilon\} \geq \alpha$$

Damit das Minimum $\geq \alpha$ wird, muss jedes der Maxima $\geq \alpha$ sein. Damit ein Maximum $\geq \alpha$ wird, muss es ein Element geben, dass $\geq \alpha$ ist. Wir erhalten so

$$\begin{aligned} &\Rightarrow \ \forall I : \ (\sum_{i \in I} r_i > 1 - 2\epsilon \ \Rightarrow \max\{\alpha_p(d_i) | i \in I\} \geq \alpha) \\ &\Rightarrow \ \forall I : \ (\sum_{i \in I} r_i > 1 - 2\epsilon \ \Rightarrow \exists i \in I : \alpha_p(d_i) \geq \alpha). \end{aligned} \qquad (11)$$

Nehmen wir nun an, dass die rechte Seite von (10) nicht stimmt, dann hieße das

$$\begin{aligned} &\neg(\exists J : \ (\sum_{i \in J} r_i \geq 2\epsilon \ \wedge \ \forall i : (i \in J \Rightarrow \alpha_p(d_i) \geq \alpha))) \\ \Rightarrow \ &\forall J : \ (\sum_{i \in J} r_i < 2\epsilon \ \vee \ \exists i : (i \in J \wedge \alpha_p(d_i) < \alpha)) \\ \Rightarrow \ &\forall J : \ (\neg(\exists i : (i \in J \wedge \alpha_p(d_i) < \alpha)) \ \Rightarrow \ \sum_{i \in J} r_i < 2\epsilon) \\ \Rightarrow \ &\forall J : \ ((\forall i \in J : \alpha_p(d_i) \geq \alpha) \ \Rightarrow \ \sum_{i \in J} r_i < 2\epsilon). \end{aligned} \qquad (12)$$

Sei nun J eine maximale Indexmenge mit der Eigenschaft, dass $\forall i \in J : \alpha_p(d_i) \geq \alpha$, dann gilt $\forall i \in J^c : \alpha_p(d_i) < \alpha$. Es müsste nach Annahme, dass die rechte Seite von (10) nicht stimmt, und somit (12) gilt, $\sum_{i \in J} r_i < 2\epsilon$ folgen. Da wegen Satz 5.23 $\sum_{i=1}^m r_i = 1$ gilt, folgt außerdem $\sum_{i \in J^c} r_i > 1 - 2\epsilon$. Was im Widerspruch zu (11) steht, wenn wir J^c als eines der I betrachten.

„$\Leftarrow$“:

Es gelte die rechte Seite von (10):

$$\exists J \ (\sum_{i \in J} r_i \geq 2\epsilon \ \wedge \ \forall i \in J : \alpha_p(d_i) \geq \alpha)$$

Nehmen wir nun ein solches J, dann gilt für jedes I mit $\sum_{i \in I} r_i > 1-2\epsilon$, dass $I \cap J \neq \emptyset$, denn sonst wäre ja die Gesamtwahrscheinlichkeit > 1. Es gibt also ein $i \in I \cap J$. Wir erhalten:

$$\Rightarrow \quad \forall I : \; (\sum_{i \in I} r_i > 1 - 2\epsilon \; \Rightarrow \; \exists i \in I : \alpha_p(d_i) \geq \alpha).$$

Da wir in (9) nur über solche I minimieren, ist bei der Maximierung immer mindestens ein $\alpha_p(d_i) \geq \alpha$ zu berücksichtigen.

$$\Rightarrow \quad \alpha_p(D) = \min\{\max\{\alpha_p(d_i) | i \in I\} | I \subseteq \{1, \ldots, m\} \wedge \sum_{i \in I} r_i > 1 - 2\epsilon\} \geq \alpha$$

■

Bemerkung: Salopp ausgedrückt bedeutet Satz 7.5, dass eine Primpotenz nur dann in der Zerlegung von D auftreten kann, wenn sie in so vielen d_i auftritt, dass deren Gesamtwahrscheinlichkeit $\geq 2\epsilon$ ist.

Wir wissen jetzt bereits genug, um zu zeigen, dass die Primpotenzen der Periode wesentlich die Größe des PFA bestimmen und können selbst eines der Ergebnisse aus [MPP01] herleiten.

Satz 7.6 *Untere Schranke der Zustände eines PFA*

L sei eine unäre reguläre Sprache mit minimaler ultimativer Periodizität D. D habe die Primpotenzzerlegung $D = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_s^{\alpha_s}$. Dann benötigt jeder PFA, der L mit isoliertem Cutpoint akzeptiert, mindestens $p_1^{\alpha_1} + p_2^{\alpha_2} + \ldots + p_s^{\alpha_s}$ Zustände.

Beweis: $M = (Q, A, \pi, \eta)$ sei der minimale PFA, der L mit isoliertem Cutpoint akzeptiert. Die Perioden der ergodischen Komponenten seien d_i für $i = 1, \ldots, m$. Jedes $p_j^{\alpha_j}$ muss mindestens ein d_i teilen, denn andernfalls würde diese Primpotenz nicht die Periode der Markovkette teilen und D wäre nicht minimale Periode der Sprache. Deshalb gilt

$$\sum_{i=1}^{m} d_i \;\; \geq \;\; \sum_{j=1}^{s} p_j^{\alpha_j}.$$

Die Zahl der Zustände einer ergodischen Komponente ist mindestens so groß wie ihre Periode. Die Gesamtzahl der Zustände ist mindestens so groß wie die aller ergodischen Zustände. Wir erhalten das gewünschte Ergebnis $|Q| \geq p_1^{\alpha_1} + p_2^{\alpha_2} + \ldots + p_s^{\alpha_s}$. ■

Bevor wir zur Aussage über den Zusammenhang der Größe der ultimativen Periodizität der Sprache und der Anzahl der Zustände im ϵ-isolierten PFA kommen, benötigen wir noch ein Lemma.

Lemma 7.7 *Ungleichung zwischen arithmetischem und geometrischem Mittel*

Für eine Folge reeller Zahlen $(r_1, \ldots, r_m)$ mit $r_i \geq 0$ und $\sum_{i=1}^{m} r_i = 1$ und für beliebige Zahlen $x_i \geq 0$ gilt:

$$\sum_{i=1}^{m} r_i x_i \geq \prod_{i=1}^{m} x_i^{r_i}.$$

Beweis: Wenn die r_i und x_i so gewählt sind, dann gilt für jede konkave Funktion f:

$$f(\sum_{i=1}^{m} r_i x_i) \geq \sum_{i=1}^{m} r_i f(x_i).$$

Da der Logarithmus eine konkave Funktion ist, gilt.

$$\begin{aligned}
& \log(\sum_{i=1}^{m} r_i x_i) \geq \sum_{i=1}^{m} r_i \log(x_i) \\
\Rightarrow\ & \log(\sum_{i=1}^{m} r_i x_i) \geq \sum_{i=1}^{m} \log(x_i^{r_i}) \\
\Rightarrow\ & 2^{\log(\sum_{i=1}^{m} r_i x_i)} \geq \prod_{i=1}^{m} 2^{\log(x_i^{r_i})} \\
\Rightarrow\ & \sum_{i=1}^{m} r_i x_i \geq \prod_{i=1}^{m} x_i^{r_i}
\end{aligned}$$

■

Satz 7.8 *Maximale Größe der ultimativen Periodizität*

Die von einem unären PFA mit n Zuständen und ϵ-isoliertem Cutpoint λ akzeptierte Sprache $T(M, \lambda)$ hat ultimative Periodizität $D \leq n^{\frac{1}{2\epsilon}}$.

Beweis: Wir betrachten die Primpotenzzerlegung von D.

$$D = \prod_{j=1}^{k} p_j$$

$(p_j)_j$ ist die Folge aller Primpotenzen, die D teilen, also $\{p_j | j = 1, \ldots, k\} = \{p^{\alpha_p(D)} | p \in \mathbb{P} \wedge p | D\}$. Durch Gruppieren der Primpotenzen in allen d_i und da ja für jedes p_j wegen Satz 7.5 gelten muss: $\sum_{i:p_j|d_i} r_i \geq 2\epsilon$, erhalten wir

$$\prod_{i=1}^{m} d_i^{r_i} \geq \prod_{j=1}^{k} p_j^{2\epsilon}.$$

Insgesamt folgt

$$n \geq \sum_{i=1}^{m} d_i \geq \sum_{i=1}^{m} r_i d_i \underbrace{\geq}_{\text{Lemma 7.7}} \prod_{i=1}^{m} d_i^{r_i} \geq \prod_{j=1}^{k} p_j^{2\epsilon} \geq D^{2\epsilon}.$$

Wir können die Aussage zunächst so lesen, dass jeder unäre PFA mit ϵ-isoliertem Cutpoint mindestens $D^{2\epsilon}$ Zustände benötigt, wenn die zu akzeptierende Sprache minimal ultimativ D-periodisch ist. Andererseits folgt aus $n \geq D^{2\epsilon}$ auch das gewünschte Ergebnis.

$$D \leq n^{\frac{1}{2\epsilon}}$$

■

7.1.1 Eine stärkere Vermutung

Wir haben sogar die Vermutung, dass für ein beliebiges, aber festes ϵ und genügend großes n gilt:

$$D \leq (n2\epsilon)^{\frac{1}{2\epsilon}}$$

Die Vermutung gründet darauf, dass wir eine möglichst große Gesamtperiode D erhalten, die sich ja als kleinstes gemeinsames Vielfaches der einzelnen Perioden der ergodischen Komponenten d_i ergibt, wenn wir $\lfloor\frac{1}{2\epsilon}\rfloor$ ergodische Komponenten haben und alle Perioden d_i etwa gleich groß sind, also $d_i \approx n2\epsilon$, wobei alle d_i paarweise teilerfremd sind.

Für einen Spezialfall können wir die Vermutung beweisen:
Wir nehmen an, dass die Absorptionswahrscheinlichkeiten der m ergodischen Komponenten gleichverteilt sind, also $r_i = \frac{1}{m}$ sind. Es gelte $m \geq \frac{1}{2\epsilon}$. Dann müsste jede Primpotenz p_j, die wie im Beweis zu Satz 7.8 D teilt mindestens auch $m \cdot 2\epsilon$ viele der d_i teilen, da andernfalls die Gesamtwahrscheinlichkeit $\sum_{i:p_j|d_i} r_i < 2\epsilon$ wäre und es gäbe eine kleinere ultimative Periodizität als D. Offensichtlich muss also

$$\prod_{i=1}^{m} d_i \geq \prod_{j=1}^{k} p_j^{m2\epsilon} \geq D^{m2\epsilon}$$

gelten. Dies ergibt:

$$\frac{1}{m}n \geq \frac{1}{m}\sum_{i=1}^{m} d_i \geq (\prod_{i=1}^{m} d_i)^{\frac{1}{m}} \geq (D^{m2\epsilon})^{\frac{1}{m}} \geq D^{2\epsilon}.$$

Woraus mit $m \geq \frac{1}{2\epsilon}$

$$n \geq m \cdot D^{2\epsilon} \geq \frac{1}{2\epsilon}D^{2\epsilon}$$

folgt. Und damit ist $D \leq (n2\epsilon)^{\frac{1}{2\epsilon}}$ gezeigt.

7.2 Obere Schranke für den äquivalenten DFA

Korollar 7.9 *Obere Schranke für den äquivalenten DFA*

Akzeptiert ein unärer PFA mit n Zuständen eine Sprache L mit ϵ-isoliertem Cutpoint, dann braucht der minimale äquivalente DFA höchstens $n^{\frac{1}{2\epsilon}}$ Zustände in seinem zyklischen Teil.

Beweis: Das Korollar ergibt sich aus Satz 7.8 und Satz 3.33. ∎

7.3 Untere Schranke für den äquivalenten DFA

Wir zeigen in diesem Abschnitt, dass die obere Schranke $n^{\frac{1}{2\epsilon}}$ in dem Sinne scharf ist, dass man den Exponenten für beliebige n nicht verkleinern kann. Es ist nicht möglich, die Schranke mit einem konstanten Faktor im Exponenten, zum Beispiel auf $n^{c\cdot\frac{1}{2\epsilon}}$ mit $c < 1$, zu verkleinern.

Satz 7.10 *Untere Schranke für die ultimative Periodizität*
Es gibt eine unendliche Folge von regulären Sprachen $L_{k,m}$, die jeweils von einem unären PFA mit n_k Zuständen und ϵ-isoliertem Cutpoint λ akzeptiert werden, wobei $m := \lfloor \frac{1}{2\epsilon} \rfloor \geq 2$ gilt. Für jedes $c < 1$ gibt es eine Sprache $L_{k,m}$, die die minimale ultimative Periodizität $D > n_k^{c \cdot \frac{1}{2\epsilon}}$ hat.

Zunächst benötigen wir ein Lemma.

Lemma 7.11 *Verhältnis zweier Primzahlen p_k und p_{k+m}*
Für $2 < k \in I\!N$ und $2 \leq m \in I\!N$ gilt

$$p_{k+m} \leq 4m^2 p_k.$$

Beweis: Es sei $(p_i)_{i \in I\!N}$ die aufsteigende Folge aller Primzahlen. Aus der Primzahlforschung ist bekannt (vergleiche [Cal]):

$$k \ln k \leq p_k \leq 2k \ln k \qquad \text{für } k > 2.$$

Seien m und k wie im Lemma angegeben.

$$4m^2 \geq 4m \cdot \ln(2m) \geq \frac{2}{2\ln 2} 2m \ln(2m) \underbrace{\geq}_{2m \geq 2+m} \frac{2(2+m)\ln(2+m)}{2\ln 2} \geq \frac{2(k+m)\ln(k+m)}{k \ln k}$$

Die letzte Ungleichung gilt, da der letzte Term als Funktion von k für jedes $m \geq 2$ monoton fällt.

$$\Rightarrow p_{k+m} \leq 2(k+m)\ln(k+m) \leq k \ln k \underbrace{\frac{2(k+m)\ln(k+m)}{k \ln k}}_{\leq 4m^2} \leq 4m^2 p_k$$

■

Beweis von Satz 7.10: Sei $m := \lfloor \frac{1}{2\epsilon} \rfloor \geq 2$.
Für ein $k \in I\!N$ sei

$$x_k := \prod_{i=k}^{k+m-1} p_i$$

und

$$n_k := \sum_{i=k}^{k+m-1} p_i.$$

Wir zeigen nun, dass es für jedes $m \geq 2$ und jedes $0 < c = 1 - \delta < 1$ ein k_0 gibt, so dass für alle $k > k_0$ gilt: $x_k > n_k^{cm}$.
Sei k_0 so gewählt, dass $p_{k_0} \geq (4m^3)^{\frac{c}{\delta}}$. Es gilt dann für alle $k > k_0$:

$$p_k^\delta > (4m^3)^c \Rightarrow p_k^{1-\delta} \cdot p_k^\delta > (4m^3)^c \cdot p_k^c \Rightarrow p_k > (m \cdot 4m^2 p_k)^c \Rightarrow p_k > (m \cdot p_{k+m})^c.$$

Sei also $k > k_0$, dann gilt:

$$\prod_{i=k}^{k+m-1} p_i > p_k^m > (m \cdot p_{k+m})^{cm} > (\sum_{i=k}^{k+m-1} p_i)^{cm}.$$

Und wir haben das gewünschte Ergebnis.

$$x_k > n_k^{cm}$$

Die Sprache

$$L_{k,m} := \{1^{z \cdot x_k} | x_k = \prod_{i=k}^{k+m-1} p_i \ \wedge \ z \in I\!N_0\}$$

wird von einem PFA mit m ergodischen Komponenten, deren Perioden die Primzahlen von p_k bis p_{k+m-1} sind, akzeptiert. Der PFA wird so wie in Abbildung 12 konstruiert. Die Startverteilung gewichtet den akzeptierenden Zustand jedes ergodischen Zyklus mit Wahrscheinlichkeit $\frac{1}{m}$. Der Cutpoint sei $1 - \frac{1}{2m}$. Die Isolation ist dann $\frac{1}{2m} \geq \epsilon$,

$$m = \left\lfloor \frac{1}{2\epsilon} \right\rfloor \Rightarrow m \leq \frac{1}{2\epsilon} \Rightarrow \epsilon \leq \frac{1}{2m}.$$

Der PFA hat n_k viele Zustände. Andererseits benötigt der DFA mindestens so viele Zustände, wie die minimale (ultimative) Periodizität. Diese ist hier x_k. ■

8 Eine untere Schranke für den Anfangspfad

Bisher haben wir uns nur mit dem ultimativ periodischen Verhalten der unären Sprachen beschäftigt und uns nicht um die Größe des Anfangspfads im DFA gekümmert.

In Abschnitt 3.4.1 haben wir bereits eine unäre Sprachfamilie kennengelernt, die jeweils von einem PFA mit 2 Zuständen akzeptiert werden kann, während der äquivalente DFA $m+2$ Zustände benötigt. Der Preis für den kleinen PFA ist allerdings auch eine kleine Isolationslücke. Denn es gilt $\epsilon(m) = \Theta(1/m)$.

Bisher ist keine bessere obere Schranke für den Anfangspfad eines unären DFA bekannt, als die allgemeine obere Schranke von Rabin.

Wir können allerdings nachweisen, dass in unendlich vielen Fällen ein superpolynomieller Blowup selbst bei einer Isolationslücke von $\frac{1}{4}$ nötig ist.

8.1 Ein PFA für die Sprache $\{1^m\}$

Wir beschreiben eine unendliche Familie endlicher Sprachen, für die ein PFA mit $\frac{1}{4}$-isoliertem Cutpoint ausreicht, dessen Größe polylogarithmisch in der Größe des Anfangspfads des äquivalenten DFA ist.

Satz 8.1 *PFA für die Sprache* $\{1^m\}$

Es gibt einen PFA mit $O\left(\frac{(\ln m)^2}{\ln\ln m}\right)$ Zuständen und $\frac{1}{4}$-isoliertem Cutpoint für die Sprache $L_m = \{1^m\}$.

Beweis: Wir benötigen zunächst das folgende Lemma.

Lemma 8.2 *Höchstzahl übereinstimmender Restklassen*
Gegeben sei eine natürliche Zahl m. Für $P > 12\frac{\ln m}{\ln\ln m}$ seien $p_1, p_2, \ldots p_P$ die ersten P Primzahlen, die größer als $\ln m$ sind. Dann gilt für jede natürliche Zahl $x \neq m$ mit $x \leq m^3$: $x \equiv m \mod p_i$ für höchstens $\frac{P}{4}$ der Primzahlen p_i.

Beweis des Lemmas: Die natürliche Zahl m sei gegeben und $p_1, \ldots, p_P$ seien die Primzahlen. Es gilt

$$\begin{aligned} & P > 12\frac{\ln m}{\ln\ln m} \\ \Rightarrow\ & \frac{P}{4}\ln\ln m > 3\ln m \\ \Rightarrow\ & (\ln m)^{\frac{P}{4}} > m^3. \end{aligned}$$

Für eine Auswahl von Primzahlen $p_{i_1}, \ldots, p_{i_k}$ und zwei beliebige natürliche Zahlen x, y gilt

$$(\forall 1 \leq j \leq k : x \equiv y \mod p_{i_j}) \quad \Leftrightarrow \quad x \equiv y \mod \prod_{j=1}^{k} p_{i_j}.$$

Außerdem gilt für $x, y < N$: $(x \equiv y \mod N) \Leftrightarrow (x = y)$. Nehmen wir an, dass $x \leq m^3$ gilt, dann folgt für $k > \frac{P}{4}$ auch $x, m \leq m^3 < (\ln m)^{\frac{P}{4}} < \prod_{j=1}^{k} p_{i_j}$.

Nach der Wahl der P Primzahlen $p_1, \ldots, p_P$ folgt, dass für jede Auswahl der Primzahlen $p_{i_1}, \ldots, p_{i_k}$ mit $k > \frac{P}{4}$ und für jedes $x \leq m^3$ gilt:

$$\begin{aligned} x \neq m \quad &\Rightarrow \quad x \not\equiv m \mod \prod_{j=1}^{k} p_{i_j} \\ &\Rightarrow \quad \exists j : x \not\equiv m \mod p_{i_j}. \end{aligned}$$

Somit haben wir gezeigt, dass für höchstens $\frac{P}{4}$ der Primzahlen $p_1, \ldots, p_P$ die Äquivalenz $x \equiv m \mod p_i$ gelten kann. ■

Bemerkung: Falls andererseits für mehr als $\frac{P}{4}$ der Primzahlen $x \equiv m \mod p_i$, dann gilt auch $x = m$ und $x \equiv m \mod p_i$ für jedes p_i.

Wir setzen $P := 1 + \lfloor 12 \frac{\ln m}{\ln \ln m} \rfloor$ und $h := \frac{1}{m^2}$. Die Primzahlen $p_1, \ldots, p_P$ seien wie in Lemma 8.2 passend zu m gewählt.

Abbildung 13 zeigt eine Skizze des Zustandsdiagramms des PFA M.
Für jede Primzahl p_i gibt es in M einen Zyklus E_i der Länge p_i. Der Zyklus E_i enthalte die Zustände $q_{i,0}, q_{i,1}, \ldots, q_{i,p_i-1}$, so dass von einem Zustand zum nächsten ein Übergang mit Wahrscheinlichkeit $1-h$ definiert ist: $A_{q_{i,j}, q_{i,(j+1) \bmod p_i}} = 1-h$ für $0 \leq j < p_i$. Für $1 \leq i \leq P$ seien die Zustände $q_{i, m \bmod p_i}$ akzeptierend.

Zusätzlich existiert ein absorbierender nicht akzeptierender Zustand v. Jeder Zustand $q_{i,j}$ hat einen Übergang nach v mit Wahrscheinlichkeit h.

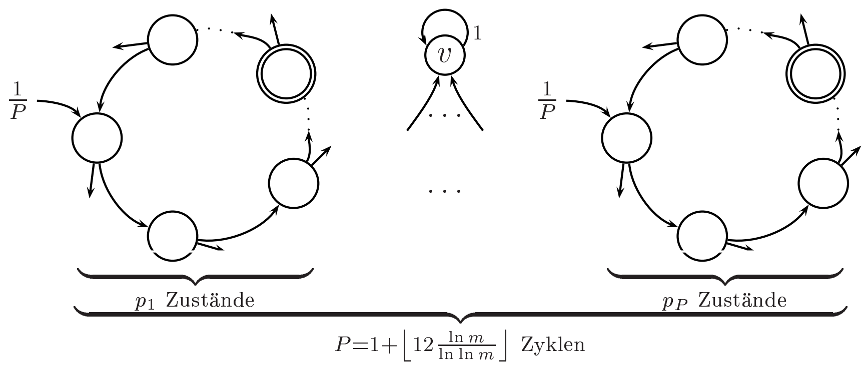

Abbildung 13: Der unäre PFA für den Beweis der unteren Schranke der Länge des Anfangspfads

Die Zahl der Zustände des PFA M kann nach oben als $P \cdot p_P$ abgeschätzt werden.

Nach dem Primzahlsatz (vergleiche [Cal]) gibt es ungefähr

$$\frac{b}{\ln b} - \frac{a}{\ln a}$$

Primzahlen im Intervall $[a, b]$. Wenn wir forden, dass es mindestens P Primzahlen im Intervall $[p_1, p_P]$ gibt, dann folgt daraus in der Näherung:

$$\begin{aligned} P \quad &\leq \quad \frac{p_P}{\ln p_P} - \frac{p_1}{\ln p_1} \\ \Rightarrow \frac{p_P}{\ln p_P} \quad &\geq \quad 12 \frac{\ln m}{\ln \ln m} + \frac{\ln m}{\ln \ln m} \\ \Rightarrow \frac{p_P}{\ln p_P} \quad &\geq \quad 13 \frac{\ln m}{\ln \ln m} \end{aligned}$$

Setzen wir $p_P = c \cdot \ln m$, dann erfüllen wir unsere Forderung zum Beispiel mit $c = 100$ für alle $m \geq 8$. Wir können also $p_P = O(\ln m)$ annehmen. Die Gesamtzahl der Zustände ist somit $n = O\left(\frac{(\ln m)^2}{\ln\ln m}\right)$.

Bleibt noch zu zeigen, dass M nur das Wort 1^m akzeptiert, wenn wir $\lambda = \frac{1}{2}$ setzen.

Die Akzeptanzwahrscheinlichkeit des Worts 1^m ist $(1-h)^m > \frac{3}{4}$ für $m > 3$. Denn wir treffen jeden akzeptierenden Zustand im Automaten, falls wir nicht vom Zustand v absorbiert werden. Die Wahrscheinlichkeit, nicht absorbiert zu werden ist $(1-h)^m$.
Betrachten wir ein Wort 1^x mit $m \neq x \leq m^3$, dann können nach Lemma 8.2 höchstens $\frac{P}{4}$ der akzeptierenden Zustände erreicht werden. Das heißt, dass $p_M(1^x) \leq \frac{P}{4} \cdot \frac{1}{P} \leq \frac{1}{4}$ gilt. Das Wort wird nicht akzeptiert.
Für ein Wort 1^x mit $x > m^3$ spielt die Absorption die Hauptrolle: Selbst wenn alle Restklassen mit denen von m übereinstimmen, gilt $p_M(1^x) \leq (1-h)^x \leq \left(1 - \frac{1}{m^2}\right)^{m^3}$. Der letzte Term ist für jedes $m > 1$ höchstens $\frac{1}{4}$.

Der Cutpoint $\lambda = \frac{1}{2}$ ist also $\frac{1}{4}$-isoliert und Satz 8.1 ist bewiesen. ■

8.2 Exponentieller Blowup bei konstanter Isolation

Wir haben folglich auch eine untere Schranke für den Anfangspfad gefunden:

Satz 8.3 *Untere Schranke für den Anfangspfad*
Es gibt eine unendliche Folge endlicher Sprachen, für die es jeweils einen PFA mit $\frac{1}{4}$-isoliertem Cutpoint und n Zuständen gibt. Der minimale äquivalente DFA benötigt jeweils $e^{\Omega(\sqrt{n \ln n})}$ Zustände im Anfangspfad.

Beweis: Wir benutzen die Sprachfamilie L_m aus Satz 8.1. Offensichtlich benötigt der minimale DFA, der L_m akzeptiert mindestens m Zustände im Anfangspfad. Der PFA kommt mit $n = O\left(\frac{(\ln m)^2}{\ln\ln m}\right)$ Zuständen aus. Formen wir die asymptotische Abschätzung um, dann erhalten wir für jedes genügend große m:

$$\begin{aligned} n &\leq c \cdot \frac{(\ln m)^2}{\ln\ln m} \\ \Rightarrow \frac{n \ln\ln m}{c} &\leq (\ln m)^2 \\ \Rightarrow \frac{\frac{n}{2}\ln\frac{n}{c}}{c} &\leq (\ln m)^2 && (13) \\ \Rightarrow \frac{n \ln n}{4c^2} &\leq (\ln m)^2 && (14) \\ \Rightarrow \sqrt{\frac{n\ln n}{4c^2}} &\leq \ln m \\ \Rightarrow m &\geq e^{\frac{1}{2c}\sqrt{n \ln n}} \end{aligned}$$

Die Ungleichung (13) ergibt sich aus $n \leq c \cdot \frac{(\ln m)^2}{\ln\ln m} \leq c \cdot (\ln m)^2$, während sich (14) ergibt, falls $n \geq c^2$ gilt. Denn dann folgt $\ln n \geq 2\ln c \geq 2\frac{c\ln c}{2c-1}$ und $(2c-1)\ln n \geq 2c\ln c$ und somit auch $\ln n - \ln c \geq \frac{\ln n}{2c}$.

Wir haben das gewünschte Ergebnis von $m = e^{\Omega(\sqrt{n \ln n})}$ bewiesen. ■

9 Folgerungen und offene Fragen

9.1 Fazit

Wir haben nachgewiesen, dass für unendliche unäre Sprachen, also (ultimativ) periodische unäre Sprachen ein superpolynomieller Blowup der Zustandszahl im Zyklus beim Übergang vom PFA auf den DFA nur erzwungen werden kann, wenn die Isolationslücke ϵ klein wird. Bei konstanter Isolationslücke erhalten wir für die Zustandszahl im Zyklus nur einen polynomiellen Blowup.

Andererseits haben wir für endliche unäre Sprachen nachgewiesen, dass hier auch bei einer großen konstanten Isolationslücke ein superpolynomieller Blowup benötigt wird.

Dies wirft die Frage auf, ob allgemein gilt, dass ein großer Blowup nur für endliche oder co-endliche Sprachen zu erzielen ist, wenn wir eine konstante Isolation voraussetzen.

Das folgende Argument zeigt jedoch, dass dies nicht der Fall ist. Ein großer Blowup des Anfangspfads ist auch für Sprachen mit nicht-trivialer ultimativer Periode zu erzielen.

Die Sprache $L_m := \{1^{2m}\} \cup \{1^k | k \text{ ungerade}\}$ wird von einem PFA akzeptiert, der ähnlich dem PFA in Abschnitt 8 konstruiert ist, mit dem Unterschied, dass statt des einen absorbierenden Zustands zwei Zustände g und u, die einen deterministischen Zyklus bilden, eingesetzt werden. Zustand u ist akzeptierend und kann nur von Wörtern ungerader Länge erreicht werden, g ist nicht akzeptierend und kann nur bei gerader Wortlänge erreicht werden. Die transienten „Zyklen“ bestehen jetzt nicht mehr aus p_i Zuständen, sondern aus $2p_i$ Zuständen. Die Übergänge aus den transienten Zuständen führen immer abwechselnd nach g und u mit Wahrscheinlichkeit h. Alle ungeraden Zustände in den transienten „Zyklen“ sind akzeptierend. Zusätzlich ist noch der $2m \bmod 2p_i$ Zustand akzeptierend. Die Analyse der akzeptierten Sprache und der Größe der Isolation erfolgt ähnlich wie in Abschnitt 8.

9.2 Eine obere Schranke für die Größe des Anfangspfads des äquivalenten DFA

Uns ist es gelungen, eine scharfe obere Schranke für die ultimative Periodizität und somit für die Anzahl der Zustände im Zyklus eines unären DFA zu finden. Ebenso haben wir für eine unendliche Folge von Sprachen eine untere Schranke für die Größe des Anfangspfads nachgewiesen.

Uns ist es nicht gelungen, eine scharfe obere Schranke für den Anfangspfad anzugeben. Die einzige bekannte obere Schranke ist die von Rabin. Sie scheint jedoch gerade im Falle unärer Sprachen nicht scharf zu sein.

9.3 Sequentiell unäre Sprachen

Möglicherweise lassen sich die hier gewonnenen Methoden auch auf sequentiell unäre Sprachen anwenden, also auf Sprachen der Form

$$L \subseteq \{a_1^{n_1} a_2^{n_2} \dots a_k^{n_k} | a_i \in \Sigma, n_i \in \mathbb{N}_0\}.$$

Hier gibt es im DFA eine baumartige Struktur: Erst ein Pfad für die a_1, dann ein Zyklus für die a_1. Aus jedem bisher genannten Zustand kann es eine Kante für a_i mit $i > 1$ geben, die der Anfang eines neuen Pfades und Zyklus sein kann und so weiter. Wenn wir nacheinander die Markovketten betrachten, die einem PFA, der L akzeptiert, für die einzelnen a_i zugrunde liegen, können wir eventuell ähnliche Ergebnisse für die Zyklen des DFA erhalten wie in dieser Arbeit.

Abbildungsverzeichnis

1 Beispiel für ein Zustandsdiagramms eines DFA 6
2 Das Zustandsdiagramm des Modulo-Automaten 9
3 PFA zum Beweis REG≠CL 12
4 PFA für die Sprache C_m 14
5 Unärer DFA 15
6 Die Mengen $v_{\vec{x}}$ und V_δ 18
7 Skizzen zur Volumenberechnung eines Simplex 19
8 Beispiel: Zustandsklassifizierung 29
9 Beispiel: Kondensationsgraph 29
10 Beispiel: Absorptionsgraph zur Analyse vor der Absorption 32
11 Beispiel: Periodizität 39
12 Unärer PFA für den Beweis der unteren Schranke $F(n)$ 48
13 Unärer PFA für die Sprache $\{1^m\}$ 61

Literatur

[Amb96] Andris Ambainis. The complexity of probabilistic versus deterministic finite automata. *Proc. 7th International Symposium on Algorithms and Computation (ISAAC), Lecture Notes in Computer Science*, 1178:233–238, 1996.

[Cal] Chris K. Caldwell. How many primes are there? . `http://www.utm.edu/research/primes/howmany.shtml`.

[Chr86] Marek Chrobak. Finite automata and unary languages. *Theoretical Computer Science*, 47:149–158, 1986.

[For99] Otto Forster. *Analysis 3*. Vieweg, 1999.

[Fre83] David Freedman. *Markov Chains*. Springer-Verlag, 1983.

[Gan86] Felix R. Gantmacher. *Matrizentheorie*. Springer-Verlag, 1986.

[GMP01] V. Geffert, C. Mereghetti, and G. Pighizzini. Converting two-way nondeterministic unary automata into simpler automata. *Mathematical Foundations of Computer Science 2001, Lecture Notes in Computer Science 2001*, 2136:398–407, 2001.

[KS60] John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. D. Van Nostrand Company, 1960.

[Lan03] Edmund Landau. Über die Maximalordnung der Permutationen gegebenen Grades. *Archiv der Mathematik und Physik*, 3:92–103, 1903.

[MP00] Massimiliano Milani and Giovanni Pighizzini. Tight bounds on the simulation of unary probabilistic automata by deterministic automata. *Descriptional Complexity of Automata, Grammars and Related Structures (DCAGRS2000)*, 2000.

[MP01] C. Mereghetti and G. Pighizzini. Optimal simulations between unary automata. *SIAM Journal on Computing*, 30:1976–1992, 2001.

[MPP01] Carlo Mereghetti, Beatrice Palano, and Giovanni Pighizzini. On the succinctness of deterministic, nondeterministic, probabilistic and quantum finite automata. *Descriptional Complexity of Automata, Grammars and Related Structures (DCAGRS 2001)*, 2001.

[Paz71] Azaria Paz. *Introduction to probabilistic automata*. Academic Press, 1971.

[Rab63] M.O. Rabin. Probabilistic automata. *Information and Control*, 6(3):230–245, 1963.

[Rob76] Fred S. Roberts. *Discrete mathematical models, with applications to social, biological, and environmental problems*. Prentice Hall, 1976.

[Sen73] Eugene Seneta. *Non-negative Matrices and Markov Chains*. Springer-Verlag, 2nd edition, 1973.

[TK98] Howard M. Taylor and Samuel Karlin. *An introduction to stochastic modeling*. Academic Press, 3rd edition, 1998.

[Weg93] Ingo Wegener. *Theoretische Informatik.* B.G. Teubner, 1993.

[Weg98] Stefan Wegenkittl. *Generalized Phi-Divergence and Frequency Analysis in Markov Chains.* Dissertation, Universität Salzburg, 1998.

[Wot00] Detlef Wotschke. Beschreibungskomplexität I. Skript zur Vorlesung, Johann Wolfgang Goethe-Universität Frankfurt, 2000.

Index

Absorptionsgraph, 32
Absorptionsmatrix, 32
Absorptionswahrscheinlichkeit, 50
Akzeptanzwahrscheinlichkeit, 11
Alphabet, 4
Äquivalenzklasse, 7
Äquivalenzrelation
    Index, 7
    rechtsinvariant, 6
Automat
    äquivalente Automaten, 5
    deterministischer, 5
    Größe, 5, 15
    Modulo-Automat, 9
    nichtdeterministischer, 9
    probabilistischer, 10
    stochastischer, 10

Blowup, 17

CL, 11
Cutpoint, 10, 11
    isolierter, 13

DFA, 5

Klasse
    kommunizierende, 28
    periodische, 39
Kondensationsgraph, 28

Landau-Funktion, 45

Markovkette, 25
    absorbierende, 31
    ergodische, 31
    homogene, 25
    Periode, 38
    regelmäßige, 39
Modulo-Automat, 9

Nerode-Relation, 7
NFA, 9

Periode eines Zustands, 37
Periodizität
    ultimative, 16
PFA, 10

REG, 6

Sprache
    Cutpoint-Sprache, 11
    formale, 5
    reguläre, 6
    unäre, 15

Übergangswahrscheinlichkeitsmatrix, 26

Wort, 4
    leeres Wort ε, 4

Zustandsdiagramm
    DFA, 6
    Markovkette, 26
    NFA, 9
    PFA, 10
Zustandsmenge
    abgeschlossene, 28
    ergodische, 29
    transiente, 30