Algebraische Analyse von approximativem Reinforcement Lernen

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
des Fachbereichs Mathematik/Informatik
der Universität Osnabrück
genehmigte

D i s s e r t a t i o n

devon
Ar tur M erke

Tag der mündlichen Prüfung: 22.07.2005
Erster Gutachter: Prof. Dr. M. Riedmiller
Zweiter Gutachter: Prof. Dr. B. Hammer
Inhaltsverzeichnis

1 Einleitung 1

2 Grundlagen 8
  2.1 Reinforcement Lernen . . . . . . . . . . . . . . . . . . . . . . 8
    2.1.1 Entscheidungen und Strategien . . . . . . . . . . . . . . . . 9
    2.1.2 Strategieauswertung . . . . . . . . . . . . . . . . . . . . . . 10
    2.1.3 Wertiteration, Q-Lernen . . . . . . . . . . . . . . . . . . . . 13
  2.2 Lineare Funktionsapproximation . . . . . . . . . . . . . . . . . . . 15
    2.2.1 Aktualisieren einer Wertfunktion . . . . . . . . . . . . . . . . 16
  2.3 Matrixtheorie . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

3 Matrixiterationen 19
  3.1 Wiederholte Anwendung von $T(x) = (I + ab^T)x + ra$ . . . . . . . 21
  3.2 Wiederholte Anwendung von $T(x) = Ax + b$ . . . . . . . . . . . . . . . 25
  3.3 Stabilität spezieller Iterationsmatrizen . . . . . . . . . . . . . . . . 30
  3.4 Unendliche Produkte von schießen Projektionen . . . . . . . . . . . . . 37
  3.5 Relation von schießen Projektionen und speziellen Iterationsmatrizen . 43
  3.6 Lage von Eigenwerten . . . . . . . . . . . . . . . . . . . . . . . . . . 48
  3.7 Zusammenfassung . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

4 Konvergenz von approximativem Reinforcement Lernen 56
  4.1 TD[0] mit linearer Funktionsapproximation . . . . . . . . . . . . . . 57
  4.2 Divergenz des TD[0] Verfahrens . . . . . . . . . . . . . . . . . . . . . 59
  4.3 Synchrones Reinforcement Lernen . . . . . . . . . . . . . . . . . . . 64
  4.4 Zusammenhang mit linearen Gleichungssystemen . . . . . . . . . . . . 69
  4.5 Residuelles Gradienten Verfahren . . . . . . . . . . . . . . . . . . . . 71
    4.5.1 Asynchrones RG Verfahren . . . . . . . . . . . . . . . . . . . . 73
  4.6 Konvergenzgeschwindigkeit des TD[0] und RG Verfahrens im Vergleich 75
INHALTSVERZEICHNIS

4.7 Approximative Wertiteration ........................................ 75
4.8 Approximatives Q-Lernen ............................................ 77
4.9 Zusammenfassung ...................................................... 77

5 Gitterbasierte Approximatoren ........................................ 80
  5.1 Stückweise konstante Gitterapproximatoren .................... 81
    5.1.1 Eigenschaften beim approximativen RL .................... 82
  5.2 Stückweise lineare Gitterapproximatoren ...................... 84
    5.2.1 Eigenschaften beim approximativen RL .................... 88
  5.3 CMAC Approximator ................................................. 89
  5.4 Zusammenfassung .................................................... 91

6 Zusammenfassung und Ausblick ....................................... 93

Literaturverzeichnis ....................................................... 97
Abbildungsverzeichnis

1.1 Relation zwischen dem approximativen TD[0] Verfahren und Matrixiterationen. .................................. 6

2.1 Schema eines markovschen Entscheidungsprozesses. ................. 8

3.1 Die tangentialen Punkte der Einheitskugel der $|| \cdot ||_1$ Norm im $\mathbb{R}^2$. ... 24

3.2 Bei der Anwendung der optimalen Lernrate wir der ursprüngliche Kreis zum Kreis $K_{\alpha^*}$, der konzentrisch in dem gestrichelten Kreis $K_1$ enthalten ist. .................................................. 49

4.1 Die Hutfunktionen $\phi_1$ und $\phi_2$. ................................. 60

4.2 Identifizierung der Übergänge $g_1 \sim h$ und $g_2 \sim h$ mit Übergängen im Intervall $[0,1]$ .................................................. 60

4.3 Der Spektralradius der Matrix $A = (I+\alpha A_2)(I+\alpha A_3)$ in Abhängigkeit von der Lernrate $\alpha$. ................................................. 62

5.1 Verschiedene gitterbasierte Approximatoren. .......................... 80

5.2 Voronoi Kachelung eines regelmäßigen Gitters in $\mathbb{R}^2$. ........... 81

5.3 Eine Treppenfunktion bzw. ein eindimensionaler stückweise konstanter Gitterapproximator. .............................. 82

5.4 Stückweise lineare Funktion von einer Treppenfunktion. .......... 83

5.5 Transformation stückweise konstanter Funktion zu Hutfunktion. ... 84

5.6 Trägermenge einer Basisfunktion über 5 Simplexen. ................. 86

5.7 Trägermenge einer Basisfunktion in einer Kuhn-Triangulierung. ... 87

5.8 Visualisierung von Beispiel 10 mit Hilfe von 2 Übergängen auf einem stückweise linearen Gitterapproximator mit 2 Basisfunktionen $\phi_1$ und $\phi_2$. ............................................................ 88

5.9 Vergleich zwischen einem stückweise konstanten Gitterapproximator und dem CMAC Approximator. .......................... 90

5.10 Drei Übergänge auf einem CMAC Approximator mit lediglich zwei Basisfunktionen, bei denen das synchrone TD[0] Verfahren für jede konstante Lernrate $\alpha \neq 0$ divergiert. .......................... 91
Tabellenverzeichnis

3.1 Übersicht über die Beschränktheit von Produkten und Potenzen verschiedener Matrixtypen. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44

3.2 Übersicht für dreidimensionale Matrizen $P$ über die Lage der Spektralradii $\tilde{\rho}_{TD}(P)$ und $\rho_{RG}^*(P)$. . . . . . . . . . . . . . . . . . . . . 53

4.1 Eigenschaften von Matrizen der Form $\Phi^\top \hat{D}(I - \gamma \hat{P})\Phi$. . . . . . . . . . 68
Kapitel 1

Einleitung


$^1$Wir werden im Kapitel 2 auf weitere formale Einzelheiten der Repräsentation einer Wertfunktion
Der einzige Ausweg ist meistens eine (parametrische) Approximation der Wertfunktion $v$, um so das Lernen nicht auf dem Raum aller möglichen Wertfunktionen, sondern nur auf einer Teilmenge davon stattzufinden zu lassen. Dies bringt aber neue Schwierigkeiten mit sich, auf die wir noch eingehen werden.


**Konvergenzproblem beim approximativen RL**


**Thema und Umfang der Arbeit**

Ziel der Arbeit ist es das Konvergenzverhalten von RL Verfahren mit Funktionsapproximation zu untersuchen. Wir konzentrieren uns dabei auf das TD[0] Verfahren zur Strategieauswertung, da man anhand dieses Verfahrens die grundlegenden Probleme sieht, und die gewonnenen Erkenntnisse auch auf andere Verfahren übertragen kann. Insbesondere bedeutet die Divergenz des TD[0] Verfahrens, dass auch eine Folge von Aktualisierungen vorgeben kann, so dass das Verfahren der Wertiteration oder das Q-Lernen ebenfalls divergieren werden (vgl. Abschnitte 2.1.3, 4.7 und 4.8)

Die Auswertung einer Strategie durch das TD[0] Verfahren wird durch die Aktualisierungsvorschrift (2.8) erzielt. Diese Aktualisierungsvorschrift lautet

\[ v^{k+1}(s) = (1 - \alpha_k)v^k(s) + \alpha_k(\gamma v^k(j) + r) \]

Dabei sind \( s \) und \( j \) Zustände und \( r \) ein Belohnungssignal. Der Zustand \( j \) ist zudem gemäß der Wahrscheinlichkeit \( p(\cdot | s, a) \) des zu Grunde liegenden markovischen Prozesses gesampelt (vgl. Abschnitt 2.1). Das macht das TD[0] Verfahren zu einem stochastischen Prozess, dessen Konvergenz gegen die gesuchte Wertfunktion nachgewiesen werden kann, falls die Lernrate \( \alpha_k \) im Laufe der Zeit gemäß (2.9) abklingt [BT96]. Die Analyse von solchen stochastischen Prozessen ist meistens recht schwierig, insbesondere wenn zudem ein Funktionsapproximator zum Einsatz kommt. Wie wir im Kapitel 4 zeigen werden, ist die Ursache für die Divergenz eines solchen Prozesses bei Verwendung von Funktionsapproximation aber meist rein algebraischer Natur. Der Ausgangspunkt unserer Untersuchungen ist dabei das TD[0] Verfahren, wobei wir die Bedingung (2.9) der abklingenden Lernrate \( \alpha_k \) durch eine konstante, aber beliebig kleine Lernrate \( \alpha \) ersetzen. Dies entspricht im Übrigen auch dem Vorgehen in der Praxis, wo die Lernrate evtl. anfänglich abgesenkt wird, dann aber bei einem konstanten Wert belassen wird. Dass diese Annahme auch in der Theorie plausibel ist, zeigt sich dadurch, dass das TD[0] Verfahren ohne Funktionsapproximation bei konstanter Lernrate beschränkt bleibt, wie wir im Korollar 3 zeigen werden.

---

2Es handelt sich um fast sichere Konvergenz vgl. [BT96, Bil95].
Diese Beobachtung wird als Forderung auf den approximativen Fall übertragen. Wir interessieren uns daher für Klassen von Approximatoren, für die das approximative TD[0] Verfahren bei konstanter Lernrate $\alpha$ beschränkt bleibt. Als Anschauungsbeispiel dient ein MDP mit lediglich 3 Zuständen und ein Funktionsapproximator, bei dem es Folgen von TD[0] Aktualisierungen gibt, die für jede Lernrate $\alpha \neq 0$, mag sie noch so klein sein, divergieren (vgl. Beispiel 10). Das Beispiel basiert auf Folgen von affinen Abbildungen, und so ist auch das allgemeine Vorgehen in dieser Arbeit. Wir zeigen im Abschnitt 4.1, dass das TD[0] Verfahren der Anwendung einer Folge von verschiedenen affinen Abbildungen der Form

$$T(x) = \left( I + \alpha \varphi(g)(\gamma \varphi(h) - \varphi(g)) \right)^\top x + \alpha r\varphi(g)$$

entspricht. Dabei sind $\varphi(g)$ und $\varphi(h)$ so genannte Merkmalsvektoren von Zuständen $g$ und $h$, die zu einem Übergang $g \rightsquigarrow h$ gehören. Durch Homogenisierung dieser affinen Abbildungen erreichen wir sogar, dass das approximative TD[0] Verfahren als ein unendliches Produkt von Matrizen aufgefasst werden kann (vgl. Abschnitt 3.4). Damit gelingt es in dieser Arbeit eine Brücke zwischen approximativem RL und der relativ neuen Theorie von unendlichen Matrixprodukten zu schlagen [DL92, BW92, BE97]. Dieses Vorgehen ist neu, und erweist sich auch als sehr nützlich.


Im Allgemeinen ist ein solcher Nachweis aber schwierig, so dass es nützlich ist, notwendige Kriterien für die Beschränktheit von solchen unendlichen Produkten zu haben. Damit kann man gezielt bestimmte Funktionsapproximatoren ausschließen, die ein solches notwendiges Kriterium verletzen. Oder man kann sie weiteren Untersuchungen unterziehen, falls sie diesen Test bestehen. Ein einfaches notwendiges Kriterium ist die Beschränktheit von unendlichen Potenzen einer jeden Matrix, die in einem solchen unendlichen Produkt vorkommen kann. Dies entspricht in unserem Fall der wiederholten Anwendung einer affinen Abbildung der Form (1.1). Dieses einfache Kriterium wurde in [MS04] eingeführt. Allerdings gibt es Approximatoren die dieses einfache Kriterium erfüllen, aber beim asynchronen TD[0] Verfahren versagen. Ein Beispiel für einen solchen Approximator ist der CMAC Approximator (vgl. Abschnitt 5.3). In [SM03a] wurde daher ein anderes Kriterium vorgeschlagen, welches sich auf das wiederholte Anwenden von affinen Abbildungen der Form

$$T(x) = \left( I - \alpha \Phi^\top D(I - \gamma P)\Phi \right)^\top x + \alpha \Phi^\top Dr$$

bezieht. Auf diese Form von Matrizen kommt man in natürlicher Weise, wenn man statt nur eines Übergangs $g \rightsquigarrow h$ eine ganze Menge von Übergängen betrachtet, und gleichzeitig über allen solchen Übergängen eine TD[0] Aktualisierung ausführt.

\[
I - \alpha \Phi^T D(I - \gamma P)\Phi
\]

für alle möglichen stochastischen Matrizen \(P\) und nichtnegativen Diagonalmatrizen \(D\) wichtige Indizien, ob der durch die Matrix \(\Phi\) repräsentierte Approximator für das synchrone TD[0] Verfahren geeignet ist. Sollte es nur eine Matrix \(P\) und \(D\) geben, so dass die Eigenwerte *ungeünstig* liegen (vgl. Abschnitte 3.3 und 4.3), dann gibt es Folgen von asynchronen TD[0] Aktualisierung mit diesem Approximator, die sicher divergieren werden. Diesen Zusammenhang zwischen den oben erwähnten Bedingungen haben wir nochmal in Abbildung 1.1 zusammengefasst. Ein Pfeil von Kästchen A zu Kästchen B bedeutet dabei, dass eine mögliche Divergenz im Fall A, auf jeden Fall Divergenz im Fall B hervorrufen wird.


Es wird auch zum ersten Mal eine Untersuchung von möglichen Approximatorenmatrizen \(\Phi\) im Zusammenhang mit dem synchronen TD[0] Verfahren durchgeführt. Für weite Klassen von Approximatorens \(\Phi\) können wir Divergenz auf Grund der Lage von Eigenwerten diagnostizieren. Dazu gehören unter anderem der CMAC Approximator, aber auch \(\Phi\) mit scheinbar guten Eigenschaften, wie zum Beispiel Orthogonalität der Spalten von \(\Phi\). Die einzige Klasse, für die Divergenz ausgeschlossen werden konnte, ist eine Klasse von Approximatorens \(\Phi\), die stückweise konstanten Approximatorens entspricht (vgl. Abschnitte 3.3 und 5.1).

KAPITEL 1. EINLEITUNG

Abbildung 1.1: Relation zwischen dem approximativen TD[0] Verfahren und Matrixiterationen.

Das synchrone und asynchrone RG Verfahren werden als solche in Abschnitt 4.5 untersucht. Insbesondere liefern wir einen im Vergleich zu [SM03a, Sch03] vereinfachten Beweis, dass das synchrone RG Verfahren immer (bedingt) konvergiert. Ferner zeigen wir, dass es im Fall von Funktionsapproximation bei konstanter Lernrate beim asynchronen RG Verfahren durchaus zur Divergenz kommen kann. Die erzeugte Folge von Aktualisierungen ist sehr speziell, zeigt aber, dass man im Allgemeinen aus der Konvergenz im synchronen Fall nicht auf Beschränktheit im asynchronen Fall schließen kann (zumindest nicht wenn es unendlich viele Zustände gibt).

Im Zusammenhang mit dem synchronen TD[0] Verfahren wird im Abschnitt 3.3 (Satz 6) die Tatsache bewiesen, dass für hinreichend kleine Diskontierungsfaktoren $\gamma$ die Stabilität des TD[0] Verfahrens mit beliebigen Funktionapproximatorenp gewährleisten werden kann. Gleichzeitig wird aber ein generisches Beispiel vorgestellt, welches zeigt, dass für $\gamma$ mindestens $\gamma \leq 2/(1 + \sqrt{n})$ gelten muss, wobei $n$ der Anzahl der Zustände entspricht. Das macht es für die Praxis unbrauchbar, ist aber durchaus vom theoretischen Interesse. Auch ein eher theoretisches Ergebnis ist die
Existenz einer Klasse von Systemen $P$, für die der Diskontierungsfaktor $\gamma$ beliebig aus dem Intervall $[0, 1)$ gewählt werden kann, ohne die Stabilität des TD[0] Verfahrens bei allen möglichen Funktionsapproximatoren zu gefährden (vgl. Korollar 2). Diese Eigenschaft war bisher nur von Matrizen $P$ bekannt, die der stationären Verteilung der unterliegenden Markov-Kette entsprechen [BT96].
Kapitel 2
Grundlagen

Dieses Kapitel dient in erster Linie der Einführung von Grundlagen, die für die spätere Behandlung von approximativen Reinforcement Lernen benötigt werden. Insbesondere geben wir eine kurze Einführung in die wichtigsten Verfahren des Reinforcement Lernens, gehen dann auf lineare Funktionsapproximation ein und beenden das Kapitel mit einer kurzen Übersicht über die wichtigsten verwendeten Fakten aus der Matrixtheorie. Gerade die Teile über die Funktionsapproximation und Matrixtheorie dienen vor allem dem Aufsetzen einer einheitlichen Notation, die in weiten Teilen der Arbeit immer wieder benötigt wird. Alle vorgestellten Verfahren und Resultate werden ohne Beweis aber dafür mit entsprechenden Literaturhinweisen versehen.

2.1 Reinforcement Lernen

In diesem Abschnitt werden gängige Begriffe und Verfahren des Reinforcement Lernens (RL) eingeführt. Besonderes Augenmerk gilt dabei Verfahren zur Strategieauswertung, deren Konvergenz im Fall von Funktionsapproximation den Schwerpunkt dieser Arbeit bildet.

Abbildung 2.1: Schema eines markovschen Entscheidungsprozesses.

Ein (stationärer) MDP $M$ wird durch ein Tupel $M = (S, A, r, p)$ dargestellt. Dabei bezeichnen $S$ die Menge aller Zustände und $A$ die Menge aller Aktionen von $M$. Mit einem MDP ist in dieser Arbeit immer ein (zeit-)diskreter MDP gemeint. Das bedeutet, dass Zustandsübergänge nur an diskreten Zeitpunkten erfolgen. Befindet sich der MDP $M$ zum Zeitpunkt $t$ in einem Zustand $s$, so hat die Anwendung einer Aktion $a \in A$ zur Folge, dass

1. der Agent eine Belohnung (bzw. ein Reinforcementsignal) $r(s, a)$ erhält,
2. $M$ mit Wahrscheinlichkeit $p(j|s, a)$ in einen neuen Zustand $j$ übergeht.

Mit $p(\cdot|s, a)$ ist dabei eine Wahrscheinlichkeitsverteilung auf $S$ vorgegeben. Wir konzentrieren uns hier auf den Fall, in dem die Zustandsmenge $S$ endlich ist. Für $p$ muss daher

$$\sum_{j \in S} p(j|s, a) = 1 \quad \forall s \in S, a \in A$$

gelten. Wichtig ist, dass der Übergang zu einem neuen Zustand $j$ nach Anwendung einer Aktion $a$ nur vom gegenwärtigen Zustand $s$ abhängt.

### 2.1.1 Entscheidungen und Strategien

Die Entscheidungen eines Agenten werden formal durch eine Abbildung $\pi : S \to A$ dargestellt. Eine solche Abbildung wird im Rahmen von RL als eine (deterministische) Strategie bezeichnet. Um zwischen verschiedenen Strategien unterscheiden zu können, müssen diese bewertet werden können. Es gibt verschiedene Möglichkeiten dieses zu tun, allen ist gemein, dass einer Strategie $\pi$ für jeden Zustand $s$ eine reelle Zahl $v^{\pi}(s)$ zugeordnet wird. Diese Zahl $v^{\pi}(s)$ wird als Wert (engl. value) der Strategie $\pi$ für den Zustand $s$ bezeichnet.

Ist eine Strategie $\pi$ festgelegt, so ist ausgehend von einem Zustand $s = s_0$ die Abfolge von zukünftigen Zuständen

$$s_0, s_1, s_2, \ldots \; \text{mit} \; s_{t+1} \sim p(\cdot|s_t, \pi(s_t))$$

ein markovscher Prozess. Durch die Belohnungsabbildung $r : S \times A \to \mathbb{R}$ entsteht ein assozierter markovscher Prozess

$$r_t(s, \pi) = r(s_t, \pi(s_t)), \quad t = 0, 1, \ldots$$
Für einen Diskontierungsfaktor $\gamma \in [0, 1)$ ist damit
\[
\sum_{t=0}^{\infty} \gamma^t r_t(s, \pi)
\]
eine (beschränkte) Zufallsvariable. Durch den Diskontierungsfaktor $\gamma$ erhalten weiter in der Zukunft liegende Belohnungen eine kleinere Gewichtung als frühere Belohnungen, was die Beschränktheit der Zufallsvariablen sichert. Der Erwartungswert dieser Zufallsvariablen wird mit $v^\pi$ bezeichnet
\[
v^\pi(s) = E\left( \sum_{t=0}^{\infty} \gamma^t r_t(s, \pi) \right) \quad (2.1)
\]

Die Abbildung, die durch $v^\pi : S \rightarrow \mathbb{R}$ zu Stande kommt, wird als Wertfunktion (engl. value function) bezeichnet. Falls $S$ endlich ist, so wird $v^\pi$ als Vektor im $\mathbb{R}^{|S|}$ aufgefasst.

### 2.1.2 Strategieauswertung

Die durch (2.1) gegebene Definition der Wertfunktion $v^\pi$ einer Strategie $\pi$ beruht auf der Bildung eines Erwartungswertes. Dieser kann durch Sampiln von langen Trajektorien und dem Bilden des empirischen Erwartungswertes approximiert werden (fast sichere Konvergenz). Dieses Vorgehen ist jedoch nicht immer praktisch durchführbar, und das RL befasst sich mit einer Reihe von Techniken und Algorithmen, die das explizite Bilden des empirischen Erwartungswertes vermeiden.

Eine zentraler RL Algorithmus ist das Lernen mit temporalen Differenzen (engl. temporal differences learning, TD). Da wir in dieser Arbeit auf Techniken setzen, die Matrixiterationen verwenden, werden wir uns dem TD Algorithmus aus dieser Richtung nähern.

Der Schlüssel zu den meisten RL Algorithmen ist das so genannte Bellman’sche Optimalitätsprinzip. Für die durch (2.1) definierte Wertfunktion $v^\pi$ einer Strategie $\pi$ besagt es, dass die folgende Gleichung
\[
\forall s \in S : \quad v^\pi(s) = r(s, \pi(s)) + \gamma \sum_{j \in S} p(j|s, \pi(s)) v^\pi(j)
\]
gleichzeitig für alle Zustände $s \in S$ erfüllt sein muss. Man kann dieses System von $|S| = n$ Gleichungen auch in Matrixnotation schreiben
\[
v^\pi = \gamma P^\pi v^\pi + r^\pi \quad (2.2)
\]
mit
\[
v^\pi = (v^\pi_1, \ldots, v^\pi_n)^\top, \quad r^\pi = (r^\pi_1, \ldots, r^\pi_n)^\top, \quad p^\pi = \begin{pmatrix} p(1|1, \pi(1)) & \cdots & p(n|1, \pi(1)) \\ \vdots & \ddots & \vdots \\ p(1|n, \pi(n)) & \cdots & p(n|n, \pi(n)) \end{pmatrix}
\]
\[
P^\pi \in \mathbb{R}^{n \times n}, \quad \gamma \in (0, 1)
\]
Dabei wurden die endliche Zustandsmenge $S$ mit der Menge $\{1, \ldots, n\}$, und der Raum der Abbildungen $\mathbb{R}^S$ von $S$ nach $\mathbb{R}$ mit dem Standard Vektorraum $\mathbb{R}^{|S|} = \mathbb{R}^n$ identifiziert.

Da $P^\pi$ eine stochastische Matrix ist, ist der Spektralradius $\rho(P^\pi)$ von $P^\pi$ gleich 1 (vgl. Abschnitt 2.3). Damit ist die Matrix $I - \gamma P^\pi$ regulär, ansonsten müsste $\gamma P$ den Eigenwert 1 haben, was auf Grund von $\gamma \in [0,1)$ und $\rho(P^\pi) = 1$ unmöglich ist.

Damit ist das lineare Gleichungssystem in (2.2) eindeutig lösbar. Die Lösung kann direkt per Matrix Inversion angegeben werden

$$v^\pi = (I - \gamma P^\pi)^{-1} r^\pi$$  (2.3)

Da große Gleichungssysteme in der Praxis selten direkt per Matrix Inversion gelöst werden (können), ist die durch Gleichung (2.3) vorgeschlagene Methode eher von theoretischer Bedeutung.

In der Regel werden große Gleichungssysteme iterativ gelöst. Unter den iterativen Verfahren gibt es wiederum eine Vielzahl an Varianten. Wir konzentrieren uns hier auf die so genannten Splitting Methoden (vgl. [Mei99]). Dazu wird die Matrix $A = I - \gamma P^\pi = M - N$ als Differenz von einer leicht zu invertierbaren Matrix $M$ und der Differenzmatrix $N = M - A$ geschrieben. Das Iterationsverfahren wird mit einem beliebigen Startwert $v^1$ gestartet, und entspricht der Vorschrift

$$v^{k+1} = M^{-1} N v^k + M^{-1} r^\pi$$  (2.4)

In unserem Fall ist sicherlich die Wahl $M = I$ und $N = \gamma P^\pi$ naheliegend. Dies ergibt die Iteration

$$v^{k+1} = \gamma P^\pi v^k + r^\pi$$

welche gegen die eindeutige Lösung $v^\pi$ konvergiert. Man kann das Verfahren durch so genannte Relaxation (vgl. [Mei99]) modifizieren. Dazu wird $v^{k+1}$ nicht gleich auf $\gamma P^\pi v^k + r^\pi$ gesetzt, sondern es wird ausgehend von $v^k$ entlang des Veränderungsvektors $(\gamma P^\pi v^k + r^\pi) - v^k$ abhängig von einem Schrittweitenparameter bzw. Lernrate $\alpha$ angepasst

$$v^{k+1} = v^k + \alpha(\gamma P^\pi v^k + r^\pi - v^k) = (I + \alpha \gamma P^\pi - \alpha I)v^k + \alpha r^\pi$$

$$= (I - \alpha(I - \gamma P^\pi))v^k + \alpha r^\pi$$  (2.5)

Die Iterationsvorschrift (2.5) ist in der Literatur auch unter dem Namen Richardson Verfahren bekannt ([Mei99]). Wir werden dieser Form von Iteration beim synchronen RL wieder begegnen. Dort wird $P^\pi$ allerdings auf Grund von beobachteten Übergängen durch eine Approximation $\tilde{P}$ ersetzt (vgl. Abschnitt 4.3).

Die Lernrate $\alpha$ in (2.5) sollte optimalerweise so gewählt werden, dass der Spektralradius der Iterationsmatrix $(I - \alpha(I - \gamma P^\pi))$ minimiert wird, oder zumindest kleiner 1 gemacht wird, denn nur dann ist die Konvergenz des Verfahrens gegeben.

Wir kehren nun zur Ausgangssituation $A = I - \gamma P^\pi = M - N$ zurück, und wählen als $M$ die untere Diagonalmatrix von $A$, inklusive der Diagonalen. Mit der
KAPITEL 2. GRUNDLAGEN

Kurzschreibweise \( p_{ij} = p(j|i, \pi(i)) \) hat die Matrix \( M \) die folgende Form

\[
M = \begin{pmatrix}
1 - \gamma p_{1,1} & 0 & 0 & \cdots & 0 \\
\gamma p_{2,1} & 1 - \gamma p_{2,2} & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
\gamma p_{n,1} & \cdots & \cdots & 1 - \gamma p_{n-1,n-1} & 0 \\
\gamma p_{n,1} & \cdots & \cdots & \gamma p_{n-1,n} & 1 - \gamma p_{n,n}
\end{pmatrix}
\]

Diese Wahl von \( M \) ist charakteristisch für das Gauss-Seidel Verfahren. Die Iteration (2.4) kann nun komponentenweise geschrieben werden

\[
v^{k+1}_i = \gamma \left( \sum_{j=1}^{i-1} p_{ij} v^{k+1}_j + \sum_{j=i}^{n} p_{ij} v^k_j \right) + r^\pi_i
\]

wobei die Komponenten \( i = 1, \ldots, n \) nacheinander angepasst und ersetzt werden, ohne die alten Werte zwischenzuspeichern. Aus diesem Grund wird das Gauss-Seidel Verfahren auch oft Einzelschrittverfahren genannt.

Man kann zeigen (vgl. [BT89]), dass die Reihenfolge \( i = 1, \ldots, n \) nicht unbedingt eingehalten werden muss. Die Aktualisierungen der Komponenten von \( v \) können beliebig erfolgen, solange garantiert ist, dass jede Komponenten unendlich oft aktualisiert wird. Es ergibt sich die folgende Aktualisierungsvorschrift

\[
v^{k+1}_i = \gamma \sum_{j=1}^{n} p_{ij} v^{k+1}_j + r^\pi_i
\]

(2.6)

Auch bei der Aktualisierung (2.6) kann man Relaxation einführen

\[
v^{k+1}_i = v^k_i + \alpha \left( \gamma \sum_{j=1}^{n} p(j|i, \pi(i)) v^k_j + r^\pi_i - v^k_i \right)
\]

\[
= (1 - \alpha) v^k_i + \alpha \left( \gamma \sum_{j=1}^{n} p(j|i, \pi(i)) v^k_j + r^\pi_i \right)
\]

(2.7)

Bei allen bisherigen Betrachtungen war es notwendig, dass die Wahrscheinlichkeiten \( p(| i, \pi(i) \rangle \) explizit angegeben werden können. Beim RL wird nun statt des Erwartungswertes \( \sum_{j=1}^{n} p(j|i, \pi(i)) v(j) \) in (2.7) ein einzelner Sample genommen, der entsprechend der Verteilung \( p(|i, \pi(i) \rangle \) zu Stande kommt. Es ergibt sich die Aktualisierung

\[
v^{k+1}(s) = (1 - \alpha_k) v^k(s) + \alpha_k (\gamma v^k(j) + r(s, \pi(s)))
\]

(2.8)

mit \( j \sim p(|s, \pi(s) \rangle \). Bei der Vorschrift (2.8) handelt es sich um das TD[0] Lernen ([Sut88, SB98]). Aus Gründen der Lesbarkeit wurde wieder die Funktions- statt
2.1. REINFORCEMENT LERNEN

Vektorschreibweise verwendet. Die Konvergenz\(^1\) des TD[0] Verfahrens gegen die
gesuchte Wertfunktion \(v^*\) ist garantiert, falls die Lernrate gemäß

\[
\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{und} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty
\]  

abfällt ([BT96]). Die erste Bedingung in (2.9) ist dafür verantwortlich, dass von jedem
Startpunkt \(v\) auch die Lösung \(v^*\) erreicht wird. Die zweite Bedingung bewirkt, dass
die Varianz der Zufallsvariablen \(v(s_t)\) in (2.8) im Laufe des Verfahrens abnimmt.

In der Praxis ist es aber oft so, dass die Lernrate \(\alpha_k\) anfänglich abgesenkt wird, aber
ab einer bestimmten unteren Schranke \(\alpha\) nicht weiter reduziert wird. Man verliert
dadurch die Sicherheit der Konvergenz, dennoch sind die auf diesem Weg erhaltenen
Lösungen oftmals brauchbarer. Wichtig ist aber, dass das Verfahren stabil bleibt.
Eine mögliche Instabilität tritt z.B. in Erscheinung, wenn die Werte \(v\) während der
Iteration (2.8) mit konstanten \(\alpha_k = \alpha\) unbeschränkt wachsen. Dies kann für zu große
Lernraten \(\alpha\) geschehen, allerdings sollte es für genügend kleine Werte von \(\alpha\) nicht
passieren.

Es ist leicht zu sehen, dass bei deterministischen Systemen für geeignete konstante
\(\alpha\) sogar die Konvergenz bewahrt bleibt. Dies liegt daran, dass sich das TD[0] Lernen
im deterministischen Fall auf das asynchrone Gauss Seidel Verfahren reduziert.
Bezi nicht deterministischen Systemen kann man für ein festes \(\alpha\) keine Konvergenz
erwarten. Dass solche Systeme für hinreichend kleine \(\alpha\) nicht divergieren, werden
wir in Kapitel 4 sehen. Dieses Ergebnis ist ein Spezialfall einer Analyse, die RL mit
Funktionsapproximation einschließt. Es dient aber gleichzeitig als Motivation, solche
Systeme überhaupt zu untersuchen, die für geeignete konstante Lernraten \(\alpha\) nicht
divergieren.

\[2.1.3 \text{ Wertiteration, Q-Lernen}\]

Im vorigen Abschnitt haben wir sehr eingehend das TD[0] Verfahren vorgestellt.
Wir haben dabei den besonderen Zusammenhang zu linearen Gleichungssystemen
herausgestellt. Sobald man aber nicht nur eine Strategie auswertet, sondern auch
eine verbesserte Strategie sucht, ist eine Linearisierung nicht mehr möglich. Die
Probleme, die beim RL im Zusammenhang mit Funktionsapproximation auftreten,
sind allerdings sowohl für den linearen als auch den nicht linearen Fall verwandt.

Diese Arbeit hat als Ziel die Probleme im linearen Fall zu studieren, und so mögliche
Ausweise für Divergenz von RL Verfahren mit Funktionsapproximation vorzuschlagen.
Manche der Resultate sind aber auch auf den nicht linearen Fall anwendbar, so dass
wir hier insbesondere auf das Verfahren der Wertiteration (engl. value iteration)
und das Q-Lernen eingehen werden. Streng genommen ist die Wertiteration ein
Verfahren des Dynamischen Programmierrers [Ber87, Put94], da stets ein Modell des
markovschen Prozesses \(p(\cdot|s,a)\) benötigt wird. Den Übergang zum modellfreien Fall
macht aber das Q-Lernen möglich [Wat89, BT96].

\(^1\)Da es sich um eine Zufallsvariable handelt, ist es die fast sichere Konvergenz vgl. [BT96, Bil95].
Um fortzufahren benötigen wir den Begriff einer optimalen Strategie \( \pi^* : S \to A \). Da wir bereits wissen, wie Strategien ausgewertet werden, ist eine optimale Strategie dadurch definiert, dass sie für jeden Zustand die bestmögliche Auswertung erreicht. Es muss also für jede andere Strategie \( \pi : S \to A \) folgende Ungleichung
\[
v^\pi(s) \geq v^\pi(s), \quad \forall s \in S
\]
gelten. Die Existenz einer solchen optimalen Strategie ist durch das Bellman’sche Optimalitätsprinzip gesichert (vgl. [Put94]). Einen algorithmischen Zugang zur Berechnung der optimalen Strategie liefert die Bellman’sche Gleichung
\[
v^*(s) = \max_{a \in A} \left( r(s, a) + \gamma \sum_{j \in S} p(j|s, a)v^*(j) \right) \tag{2.10}
\]
wobei wir \( v^\pi \) mit \( v^* \) abgekürzt haben. In der Tat ist die Bellman’sche Gleichung auch gleichzeitig Grundlage für das oben erwähnte asynchrone Verfahren der Wertiteration
\[
v^{k+1}(s) = \max_{a \in A} \left( r(s, a) + \gamma \sum_{j \in S} p(j|s, a)v^k(j) \right) \tag{2.11}
\]
Diese Wertiteration konvergiert gegen die optimale Wertfunktion \( v^* \). Eine optimale Strategie lässt sich anhand des Modells \( p(|s, a) \) ablesen
\[
\pi^*(s) = \arg\max_{a \in A} \left( r(s, a) + \gamma \sum_{j \in S} p(j|s, a)v^*(j) \right)
\]
Es sind mehrere optimale Strategien möglich, jedoch ist die optimale Wertfunktion \( v^* \) als Fixpunkt einer Kontraktionsabbildung immer eindeutig.

Ein Nachteil der Wertiteration ist die Tatsache, dass immer das Modell \( p(|s, a) \) erforderlich ist, um aus der optimalen Wertfunktion \( v^* \) eine optimale Strategie \( \pi^* \) abzuleiten. Diesen Nachteil behebt das Q-Lernen. Dazu wird die optimale Q-Funktion \( q^*: S \times A \to \mathbb{R} \) eingeführt
\[
q^*(s, a) = r(s, a) + \gamma \sum_{j \in S} p(j|s, a)v^*(j)
\]
Aus der Bellman’schen Gleichung (2.10) folgt direkt
\[
q^*(s, a) = r(s, a) + \gamma \sum_{j \in S} p(j|s, a) \max_{b \in A} q^*(j, b) \tag{2.12}
\]
Die Identität (2.12) wird manchmal auch als Bellman’sche Gleichung bezeichnet. Auch hier kann man analog zum Verfahren der Wertiteration (2.11) einen iterativen Algorithmus zur Berechnung von \( q^* \) ableiten
\[
q^{k+1}(s, a) = r(s, a) + \gamma \sum_{j \in S} p(j|s, a) \max_{b \in A} q^k(j, b) \tag{2.13}
\]
Die Iteration (2.13) konvergiert gegen die optimale Q-Funktion \( q^* \) und eine optimale Strategie lässt sich direkt aus \( q^* \) ohne Kenntnis des Modells ableisen
\[
\pi^*(s) = \arg\max_{a \in A} q^*(s, a)
\]
2.2. LINEARE FUNKTIONSAPPROXIMATION

Allerdings ist bei der Iteration (2.13) noch ein Modell erforderlich. Ähnlich wie beim Übergang von (2.6) über (2.7) zu (2.8) kann bei (2.13) erst Relaxation eingeführt werden, und dann kann der Erwartungswert durch ein Sample ersetzt werden. Insgesamt ergibt sich dann das Q-Lernen

\[ q^{k+1}(s, a) = (1 - \alpha_k)q^k(s, a) + \alpha_k \left( r(s, a) + \gamma \max_{b \in A} q^k(j, b) \right) \]   (2.14)

wobei \( j \sim p(\cdot|s, a) \) einen gemäß dem Modell \( p(\cdot|s, a) \) gesampelten Folgezustand von \( s \) darstellt.

Die Vorschrift (2.14) ist eine Verallgemeinerung der TD[0] Vorschrift (2.8). Wenn man nämlich die Menge der möglichen Aktionen auf die einzige Aktion \( a = \pi(s) \) einschränkt, dann erhält man aus (2.14) mit \( v^k(a) := q^k(s, \pi(s)) \) genau die Vorschrift (2.8). Damit ist das TD[0] Verfahren ein echter Spezialfall des Q-Lernens. Insbesondere sind alle Probleme mit dem TD[0] Verfahren, die sich in Bezug auf Divergenz mit Funktionsapproximation im Kapitel 4 auftun werden, automatisch auch Probleme die im Zusammenhang mit dem Q-Lernen auftreten können. Andererseits können Ergebnisse, die das TD[0] Verfahren im Zusammenhang mit Funktionsapproximation sicherer machen, als Motivation für Untersuchungen beim Q-Lernen dienen.

2.2 Lineare Funktionsapproximation

In diesem Abschnitt betrachten wir eine spezielle endliche Repräsentation einer beliebigen Funktion \( v : S \rightarrow \mathbb{R} \). Diese Funktion kann man als Wertfunktion auffassen, die Elementen aus \( S \) reellwertige Bewertungen zuordnet. Ist die Kardinalität von \( S \) unendlich, so lässt sich \( v \) im Allgemeinen nicht durch endlich viele Werte darstellen. Auch im Fall wenn \( S \) endlich, aber sehr groß ist, ist man oft an einer kompakteren Repräsentation interessiert.

Es gibt verschiedene Möglichkeiten eine Funktion durch endlich viele Parameter zu repräsentieren. Wir beschränken uns hier auf lineare Funktionsapproximation. Wir wählen dazu eine Menge \( \{\phi_1, \ldots, \phi_m\} \) von Basisfunktionen \( \phi_i : S \rightarrow \mathbb{R} \). Die Wertfunktion \( v \) kann durch eine lineare Kombination der Basisfunktionen \( \phi_i \) dargestellt werden

\[ v(s) = v(s; w) = \sum_{i=1}^{m} w_i \phi_i(s) \]

Damit beschränken wir uns auf einen endlich dimensionalen Unterraum von Funktionen, deren Kardinalität im Allgemeinen kleiner als die von \( \mathbb{R}^S \), dem Raum aller möglichen Abbildungen zwischen \( S \) und \( \mathbb{R} \), ist.

Wir führen nun eine kompaktere Schreibweise ein. Die Parameter \( w_i \) können zu einem Vektor

\[ w = (w_1, \ldots, w_m)^\top \in \mathbb{R}^m \]

zusammengefasst werden. Ähnliches ist auch mit den Auswertungen \( \phi_i(s) \) möglich. Wir definieren dazu den Merkmalsvektor

\[ \varphi(s) = (\phi_1(s), \ldots, \phi_m(s))^\top \in \mathbb{R}^m, \]
welcher alle Auswertungen der Basisfunktionen auf einem Zustand $s$ zu einem Vektor zusammenfasst. Damit lässt sich der Wert $v(s)$ kompakt schreiben als

$$v(s) = v(s; w) = w^\top \varphi(s)$$  \hspace{1cm} (2.15)

Der Zustandsraum $S$ kann prinzipiell kontinuierlich sein, man erhält aber eine kompaktere Repräsentation der Wertfunktion wenn $S$ endlich ist. Dabei ist auch der Fall interessant, in dem der Zustandsraum $S$ zwar unendlich groß ist, aber nur endlich viele Zustände $\{s_1, \ldots, s_n\} \subset S$ beobachtet werden, und es praktischer ist, mit $S \cap \{s_1, \ldots, s_n\}$ zu arbeiten. Dies wird z.B. im Fall des synchronen RL zutreffen.

Wir wählen die kanonische Darstellung $S = \{1, \ldots, n\}$. Der Raum aller Wertfunktionen $R^S$ ist demnach gleich dem $n$ dimensional Vektorraum $\mathbb{R}^n = \mathbb{R}^{\{1, \ldots, n\}}$. Eine Wertfunktion $v$ kann damit durch einen Vektor $v = (v_1, \ldots, v_n)^\top$ in $\mathbb{R}^n$ repräsentiert werden. Mit der Notation $v_s = v(s)$ wobei $s \in \{1, \ldots, n\}$ erhält die Formel (2.15) ihre Gültigkeit. Man kann die Formel (2.15) für alle Stellen von $v$ schreiben, und erhält die Darstellung

$$v = \begin{pmatrix}
  w_1^\top \varphi(1) \\
  \vdots \\
  w_1^\top \varphi(n)
\end{pmatrix} = \begin{pmatrix}
  \varphi(1)^\top \\
  \vdots \\
  \varphi(n)^\top
\end{pmatrix} w = \Phi w \in \mathbb{R}^n$$

$\Phi$ ist dabei eine $n \times m$ Matrix, deren Zeilen aus den transponierten Merkmalsvektoren $\varphi(i) \in \mathbb{R}^m$ bestehen.

### 2.2.1 Aktualisieren einer Wertfunktion

Im Verlauf eines Lernverfahrens verändert sich eine Wertfunktion $v$. Wir gehen hier auf das Anpassen des Parameters $w \in \mathbb{R}^m$ einer parametrisierten Wertfunktion $v(\cdot; w)$ an neue Werte ein. Dazu verwenden wir das Gradientenabstiegs-Verfahren, welches wir im Folgenden kurz herleiten. Insbesondere wird man dabei das Grundschema sehen, wonach eine Aktualisierung der Wertfunktion $v(\cdot; w)$ mittels des Gradientenabstiegs der Anwendung einer speziellen affinen Abbildung entspricht.

Die Wertfunktion $v(\cdot; w)$ soll im Zustand $s$ aktualisiert werden. Der neue Wert, den $v$ in diesem Zustand einnehmen soll wird mit $\tau$ bezeichnet. Beim Gradientenabstiegs-Verfahren wird der Fehler

$$E(w, \tau) = \frac{1}{2} (v(s, w) - \tau)^2$$  \hspace{1cm} (2.16)

iterativ minimiert. Dazu wird der Gradient

$$\frac{\partial E(w, \tau)}{\partial w} = (v(s; w) - \tau)(\varphi_1(s), \ldots, \varphi_m(s))^\top = (\varphi(s)^\top w - \tau)\varphi(s)$$  \hspace{1cm} (2.17)

gebildet und der Parametervektor $w$ anhand von

$$w = w - \alpha \frac{\partial E(w, \tau)}{\partial w} = \left( I - \alpha \varphi(s)^\top \right) w + \alpha \tau \varphi(s)$$  \hspace{1cm} (2.18)
2.3. MATRIXTHEORIE

gegen die Gradientenrichtung angepasst. Dazu verwenden wir einen Schrittweitenparameter bzw. eine Lernrate $\alpha$, der die Stärke der Aktualisierung bestimmt. Die Aktualisierung von $w$ in (2.18) kann auch kurz als Anwendung einer affinen Abbildung auffassen

$$w = Aw + b,$$

wobei die Form der Matrix $A$ von den Merkmalssvektoren $\varphi(s)$ und $\varphi(j)$ abhängt, falls $\tau = \gamma \varphi(j)^T w - r$ ist. In jedem Fall hat $A$ die Form $I + cd^T$ wobei $cd^T$ eine Matrix vom Rang 1 ist.

Werden mehrere solcher Aktualisierungen hintereinander gemacht, so erhalten wir eine Folge von Aktualisierungen (2.19) mit jeweils unterschiedlichen Matrizen $A$ bzw. Vektoren $b$, $c$ und $d$.

2.3 Matrixtheorie

In diesem Abschnitt stellen wir einige Definitionen der Matrixtheorie vor, die im weiteren Verlauf der Arbeit immer wieder gebraucht werden. Für eine detaillierte Darstellung verweisen wir auf [HJ85, HJ91]. Wir betrachten hier Matrizen über dem Körper der komplexen Zahlen $\mathbb{C}$, wegen der Einbettung $\mathbb{R} \subset \mathbb{C}$ sind diese Definitionen auch für reellwertige Matrizen gültig.

Eine Matrix $A \in \mathbb{C}^{m \times n}$ kann als lineare Abbildung $A : \mathbb{C}^n \rightarrow \mathbb{C}^m$ aufgefasst werden. Den Bildraum der Matrix $A$ bezeichnen wir mit

$$\text{Im}(A) = A(\mathbb{C}^n) = \{Ax | x \in \mathbb{C}^n\} \subset \mathbb{C}^m$$

und ihren Kern mit

$$\text{Ker}(A) = \{x \in \mathbb{C}^n | Ax = 0\} \subset \mathbb{C}^n.$$ 

Sowohl $\text{Im}(A)$ als auch $\text{Ker}(A)$ sind Untervektorräume von $\mathbb{C}^m$ bzw. $\mathbb{C}^n$. Für quadratische Matrizen $A \in \mathbb{C}^{n \times n}$ sind Kerne der Abbildungen $(\lambda I - A)^k$, $k = 1,2,\ldots$ von Interesse. Ist

$$\mathcal{E}_\lambda^A := \text{Ker}(\lambda I - A) \neq \{0\}$$

so nennt man $\mathcal{E}_\lambda^A$ den Eigenraum von $A$ zum Eigenwert $\lambda \in \mathbb{C}$. Für alle $x \in \mathcal{E}_\lambda^A$ gilt $Ax = \lambda x$. Eine Verallgemeinerung eines Eigenraums $\mathcal{E}_\lambda^A$ stellt ein Hauptaum $\mathcal{H}_\lambda^A$ zum Eigenwert $\lambda$ dar

$$\mathcal{H}_\lambda^A := \text{Ker}((\lambda I - A)^n)$$

Offenbar gilt $\mathcal{E}_\lambda^A \subset \mathcal{H}_\lambda^A$, wobei ein Hauptaum die praktische Eigenschaft hat, dass er in Bezug auf $A$ invariant ist

$$A(\mathcal{H}_\lambda^A) = \{Ax | x \in \mathcal{H}_\lambda^A\} \subset \mathcal{H}_\lambda^A$$

Die Menge aller Eigenwerte einer Matrix $A$ bezeichnet man als Spektrum $\sigma(A)$ von $A$. Das Spektrum lässt sich auch als Nullstellenmenge des charakteristischen Polynoms

$$p(\lambda) := \text{det}(\lambda I - A).$$
von $A$ auffassen. Den Radius des kleinsten Kreises in $\mathbb{C}$ mit Mittelpunkt 0, in den $\sigma(A)$ noch hineinpasst, bezeichnet man als Spektralradius

$$\rho(A) := \max \{ |\lambda| \mid \lambda \in \sigma(A) \}$$

von $A$. Der Spektralradius $\rho(A)$ spielt eine wichtige bei unseren Untersuchungen. Er ist eng verknüpft mit dem Begriff einer Matrixnorm. Eine Abbildung $\| \cdot \| : \mathbb{C}^{n \times n} \rightarrow \mathbb{R}$ wir als Matrixnorm bezeichnet, falls sie die üblichen Normaxiome erfüllt, und zusätzlich submultiplikativ ist

$$\| AB \| \leq \| A \| \| B \|$$

Es gibt Normen auf $\mathbb{C}^{n \times n}$ die keine Matrixnormen sind. Bedingung (2.22) ist aber enorm wichtig, und um dies zu unterstreichen werden Matrixnormen in dieser Arbeit immer mit drei Strichen $\| \cdot \|$ geschrieben. Für alle Matrixnormen $\| \cdot \|$ und Matrizen $A \in \mathbb{C}^{n \times n}$ gilt die Abschätzung

$$\rho(A) \leq \| A \|$$

durch den Spektralradius $\rho(A)$. Da der Spektralradius als Funktion $\rho(\cdot) : \mathbb{C}^{n \times n} \rightarrow \mathbb{R}$ im Gegensatz zu Matrixnormen nicht submultiplikativ ist, ist man trotz dieser Abschätzung auf die Matrixnormen angewiesen. Ein wichtiger Vertreter unter Matrixnormen ist die Spektralnorm. Sie wird von der euklidischen Norm $\| \cdot \|_2$ auf $\mathbb{C}^n$ induziert:

$$\| A \|_2 := \max_{\| x \|_2 = 1} \| Ax \|_2$$

Eine weitere Matrixnorm ist die Zeilensummennorm

$$\| A \|_\infty := \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$ 

Auch die Zeilensummennorm ist durch eine Vektornorm auf $\mathbb{C}^n$ induziert, es handelt sich dabei um die Maximumsnorm $\| \cdot \|_{\infty}$. Dies erklärt auch die Bezeichnung $\| \cdot \|_\infty$ für die Zeilensummennorm.
Kapitel 3

Matrixiterationen


Wir definieren zuerst eine Menge von affinen Abbildungen

\[ \mathcal{T} = \{ T \mid T : \mathbb{C}^n \to \mathbb{C}^n \}, \]

so dass für ein \( T \in \mathcal{T} \)

\[ T(x) = Ax + b, \quad \text{mit } b \in \mathbb{C}^n, A \in \mathbb{C}^{n \times n} \]

gilt. Wir schreiben dafür auch kurz \( T = (A, b) \) bzw. \( T = A \) falls der translatorische Anteil \( b \) gleich dem Nullvektor ist. Durch die Einbettung von \( \mathbb{R} \subset \mathbb{C} \) sind die erhaltenen Resultate auch für reellwertige Vektoren und Matrizen anwendbar.

Mit \( \mathcal{T}^\mathbb{N} \) bezeichnen wir den Raum aller Folgen in \( \mathcal{T} \). Sei \( \tau = (\tau_1, \tau_2, \ldots) \in \mathcal{T}^\mathbb{N} \) dann bezeichnet

\[ \tau^k(x) = \tau_k \circ \ldots \circ \tau_1(x) \]

die hintereinander Ausführung der ersten \( k \) Abbildungen \( \tau_i : \mathbb{C}^n \to \mathbb{C}^n \). Besteht \( \mathcal{T} \) aus nur einem Element \( T \), so enthält auch \( \mathcal{T}^\mathbb{N} \) nur eine Folge \( \tau = (T, T, \ldots) \) von
Abbildungen. In diesem Fall gilt offenbar
\[ \tau^k(x) = T^k(x) \]
Für eine Folge \( \tau \) von affinen Abbildungen können verschiedene Grenzverhalten auftreten, die wir nun formal beschreiben werden
Konvergenz
\[ \text{Conv}(\tau) : \iff \forall x \in \mathbb{C}^n : \lim_{k \to \infty} \tau^k(x) = a < \infty \]
Bedingte Konvergenz (der Grenzwert hängt vom Startwert \( x \) ab)
\[ \text{CondConv}(\tau) : \iff \forall x \in \mathbb{C}^n : \lim_{k \to \infty} \tau^k(x) = a(x) < \infty \]
Beschränktheit (auch bekannt als Lagrange Stabilität)
\[ \text{Bounded}(\tau) : \iff \forall x \in \mathbb{C}^n : \limsup_{k \to \infty} \|\tau^k(x)\| < \infty \]
Divergenz
\[ \text{Diverge}(\tau) : \iff \exists x \in \mathbb{C}^n : \limsup_{k \to \infty} \|\tau^k(x)\| = \infty. \]
Offenbar gelten folgende Implikationen zwischen den verschiedenen Grenzverhalten:
\[ \text{Conv}(\tau) \Rightarrow \text{CondConv}(\tau) \Rightarrow \text{Bounded}(\tau) \text{ und } \text{Bounded}(\tau) \Leftrightarrow \neg \text{Diverge}(\tau). \]
Oft ist man nicht nur am Verhalten einer einzigen Folge \( \tau \) von Abbildungen interessiert. Man möchte vielmehr obige Eigenschaften für alle möglichen Folgen \( \tau \in T^N \) nachweisen, die Abbildungen aus einer bestimmten Menge \( T \) von Abbildungen enthalten. Dies wird allerdings nicht immer möglich sein. So zeigt z.B. ein Resultat aus [BT00], dass bereits für den Fall, dass \( T \) aus 2 Abbildungen besteht, die Frage der Beschränktheit für alle möglichen Folgen \( \tau \in T^N \) unentscheidbar ist.
\[ T = \{(I + ab^T, ra)\} \]
untersucht. Hierbei enthält \( T \) nur eine affine Abbildung, deren Matrix aus der Summe der Identitätsmatrix \( I \) und einer Rang 1 Matrix \( ab^T \) mit \( a, b \in \mathbb{R}^n \) besteht. Zum synchronen RL korrespondieren Systeme
\[ T = \{(A, b)\} \]
mit einer beliebigen affinen Abbildung \( T(x) = Ax + b \) mit beliebigem \( A \in \mathbb{C}^{n \times n}, b \in \mathbb{C}^n \). Wir werden das Konvergenzverhalten von solchen Iterationen untersuchen, und dabei hinreichende und notwendige Bedingungen für alle möglichen Grenzverhalten angeben. Anschließend werden wir Systeme
\[ T = \{(I + ab^T, ra) \mid a \in A \subset \mathbb{R}^n, b \in B \subset \mathbb{R}^n \} \quad (3.1) \]
3.1 WIEDERHOLTE ANWENDUNG VON $T(x) = (I + ab^\top)x + ra$

In diesem Abschnitt betrachten wir Systeme von der Form

$$x^{k+1} = (I + ab^\top)x^k + ra$$

(3.2)

mit $a, b \in \mathbb{R}^n$ und $r \in \mathbb{R}$. Die Matrix $ab^\top$ ist eine reelle Rang 1 Matrix, und hat damit immer die 0 als Eigenwert (jeder zu $b$ orthogonale Vektor ist ein zugehöriger
KAPITEL 3. MATRIXITERATIONEN

Eigenvector). Daraus folgt, dass der Spektralradius der Matrix $I + ab^\top$ größer oder gleich 1 ist. Standard Methoden aus der Numerik betrachten meistens aber den Fall eines Spektralradiuses der kleiner 1 ist. Die Betrachtung des relativ einfachen Systems (3.2) dient uns als Einstieg in die Frage, was passieren kann, wenn der Spektralradius einer Iterationsmatrix gleich 1 ist. Ferner wird demonstriert, wie durch Homogenisierung einer Iterationsvorschrift, das Grenzverhalten auf eine einfachere Weise bestimmt werden kann. Die Technik der Homogenisierung wird uns in diesem Kapitel mehrmals begegnen.

Im folgenden betrachten wir das homogene System

$$\tilde{x}^{k+1} = \tilde{A}\tilde{x}^k = (I + \tilde{a}b^\top)\tilde{x}^k$$

(3.3)

mit $\tilde{A} \in \mathbb{R}^{(n+1)\times(n+1)}$ und $\tilde{x}^k \in \mathbb{R}^{n+1}$ definiert durch

$$\tilde{A} = \begin{pmatrix} I + ab^\top & ra \\ 0 & \vdots & 0 \\ & & 1 \end{pmatrix}, \quad \tilde{a} = \begin{pmatrix} a \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \tilde{b} = \begin{pmatrix} b \\ r \end{pmatrix}, \quad \tilde{x}^1 = \begin{pmatrix} x^1 \\ 1 \end{pmatrix}$$

Direktes Nachrechnen liefert die Gültigkeit der Identität

$$\tilde{x}^k = \begin{pmatrix} x^k \\ 1 \end{pmatrix}$$

auch für $k > 1$. Damit weisen die Systeme (3.2) und (3.3) gleiches Konvergenzverhalten auf.

Setzen wir nun $B = \tilde{a}\tilde{b}^\top$ so gilt

$$B^2 = (\tilde{a}^\top \tilde{b})B = (a^\top b)B$$

und folglich auch $B^{k+1} = (a^\top b)kB$. Betrachten wir nun zuerst den Fall $a^\top b = 0$, dann ist $B^2 = 0$. Für $(I + B)^2$ gilt damit

$$(I + B)^2 = I + 2B + B^2 \overset{=0}{=} I + 2B$$

und per Induktion zeigt man auch die allgemeine Identität

$$(I + B)^k = I + kB$$

Die Iteration (3.3) wird also divergieren, falls $B \neq 0$ ist. $B = \tilde{a}\tilde{b}^\top$ kann dabei nur dann annulliert werden, falls $a = 0$, oder $b = 0 \wedge r = 0$ gesetzt werden.

Ist dagegen $a^\top b \neq 0$ so erhalten wir mit Hilfe der binomischen Formel

$$(I + B)^k = \sum_{i=0}^{k} \binom{k}{i} B^i$$

$$= I + \frac{1}{a^\top b}B\left(-1 + \sum_{i=0}^{k} \binom{k}{i} (a^\top b)^i\right)$$

$$= I + \frac{1}{a^\top b}B\left(-1 + \binom{1 + a^\top b}{\Delta}\right)$$

(3.4)
3.1. WIEDERHOLTE ANWENDUNG VON $T(x) = (I + ab^\top)x + ra$

Das Verhalten von $(I + B)^k$ hängt damit im Grenzwert von $\Delta = 1 + a^\top b$ ab. Ist $|\Delta| > 1$ so divergiert $(I + B)^k$ für große Werte von $k$, ansonsten konvergiert bzw. oszilliert (für $\Delta = -1$) es. Die fassen wir im folgenden Satz zusammen.

**Satz 1** Die Iteration (3.2) konvergiert für $-2 < a^\top b < 0$ und oszilliert zwischen zwei Werten für $a^\top b = -2$. Für $a^\top b = 0$ divergiert die Iteration, außer in den trivialen Fällen $a = 0$ oder $b = 0$ und $r = 0$. Für $|1 + a^\top b| > 1$ divergiert die Iteration.

Beweis: siehe obige Herleitung. ◊

Bis auf die trivialen Bedingungen $a = 0$ bzw. $b = 0$ und $r = 0$ lassen sich alle Eigenschaften der Iteration (3.2) mit Hilfe des Skalarproduktes $a^\top b$ ausdrücken. Da $a^\top b$ und 0 Eigenwerte von $ab^\top$ bzw. $\tilde{a}\tilde{b}^\top$ sind, beruht die obige Aussage im wesentlichen auf den Eigenschaften des Spektrums der involvierten Matrix.

Die Ergebnisse lassen sich leicht auf den Fall des komplexen Körpers $\mathbb{C}$ anstatt der reellen Zahlen $\mathbb{R}$ erweitern, was im nächsten Abschnitt mit der allgemeinen Iteration (3.6) durchgeführt wird, und daher hier der Übersichtlichkeit wegen weggelassen wurde.

Es ist interessant zu untersuchen, ob es generische Formen von Vektoren $a$ und $b$ in (3.2) gibt, so dass die Iteration (3.2) konvergiert. Dies ist offensichtlich für $a = -b$ der Fall, da für alle $a \in \mathbb{R}$ immer $-a^\top a \leq 0$ gilt. Im Hinblick auf die Anwendungen im approximativen RL, ist der Fall $b = (\gamma c - a)$ und einer vorangestellten Lernrate $\alpha$ interessant, also die Iteration

$$x^{k+1} = (I + \alpha a(\gamma c - a)^\top)x^k + ra$$

(3.5)

Aus Satz 1 ist ersichtlich, dass diese Iteration für ein geeignetes $\alpha$ beschränkt bleibt, falls $a^\top a \geq \gamma a^\top c$ gilt. Geometrisch gesehen bedeutet es, dass der Vektor $c$ in dem von Vektor $a$ aufgespannten Halbraum

$$H_a = \{x \mid x^\top a \leq a^\top a\}$$

liegen muss. Insbesondere wird die Iteration (3.5) für beliebige Vektoren $a$ und $c \in H_a$ beschränkt bleiben.

In der Praxis ist aber eher eine Menge $\mathcal{X}$ vorgegeben, und man interessiert sich für die $a \in \mathcal{X}$, so dass die Iteration (3.5) für beliebige $c \in \mathcal{X}$ beschränkt bleibt. Laut obiger Herleitung sind es genau die $a \in \mathcal{X}$, für die $\mathcal{X} \subseteq H_a$ gilt. Diese Eigenschaft fassen wir in der folgenden Definition zusammen

**Definition 1** Gegeben sei eine Menge $\mathcal{X} \subset \mathbb{R}^n$ und ein $a \in \mathbb{R}^n$, wobei $a$ nicht unbedingt in $\mathcal{X}$ liegen muss. Gilt

$$\mathcal{X} \subseteq H_a$$

dann heißt $a$ ein tangentialer Punkt (von $\mathcal{X}$). ◊

Insgesamt ergibt sich folgende Formulierung

**Korollar 1** Die Iteration (3.5) ist genau dann für beliebige $c \in \mathcal{X}$ beschränkt, wenn $a$ ein tangentialer Punkt von $\mathcal{X}$ ist. ◊
Abbildung 3.1: Die tangentialen Punkte der Einheitskugel der $|| \cdot ||_1$ Norm im $\mathbb{R}^2$.

Für bestimmte Mengen lassen sich die tangentialen Punkte direkt angeben.

**Beispiel: 1** Betrachte die Menge

$$X = \{ x \mid ||x||_1 = |x_1| + |x_2| \leq 1 \} \subset \mathbb{R}^2$$

Dann sind die Punkte

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$$

die einzigen tangentialen Punkte, die selbst in $X$ liegen. Dies ist in Abbildung 3.1 nochmal veranschaulicht.

Das Ergebnis aus Beispiel 1 lässt sich für die Einheitskugel der $|| \cdot ||_1$ Norm in $\mathbb{R}^n$ verallgemeinern.

**Lemma 1** Betrachte die Einheitskugel $B_1$ der $|| \cdot ||_1$ Norm in $\mathbb{R}^n$

$$B_1 = \{ x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n \mid \sum_{i=1}^n |x_i| \leq 1 \}$$

dann ist $a = (a_1, \ldots, a_n)^T \in B_1$ genau dann ein tangentialer Punkt von $B_1$, wenn $a_i \in \{0, \kappa, -\kappa\}, |\kappa| \leq 1$ und $\sum_{i=1}^n |a_i| = 1$ gilt.

Beweis: "$\Rightarrow$" folgt aus

$$a^T a = \sum_{i=0}^n a_i^2 = \kappa \sum_{i:a_i \neq 0} |a_i| \geq \kappa \sum_{i:a_i \neq 0} |c_i| = \sum_{i:a_i \neq 0} |a_i||c_i| \geq |a^T c| \geq a^T c$$

"$\Rightarrow$" Sei O.B.d.A. $|a_1| > |a_2| > 0$, $|a_1| \geq |a_i|$ und setze $c = (1, 0, \ldots, 0)^T$ falls $a_1 > 0$, sonst $c = (-1, 0, \ldots, 0)^T$. Nun gilt

$$a^T a = \sum_{i=1}^n a_i^2 < |a_1| \sum_{i=1}^n |a_i| = |a_1| \sum_{i=1}^n |c_i| = |a_1||c_1| = a_1 c_1 = a^T c$$

also ist $a^T a < a^T c$ und $a$ kann damit kein tangentialer Punkt von $B_1$ sein.
Für die Einheitskugel der Spektralnorm sind alle Randpunkte auch tangentielle Punkte.

**Lemma 2** Betrachte die Einheitskugel $B_2$ der Spektralnorm $\| \cdot \|_2$ Norm in $\mathbb{R}^n$

$$B_2 = \left\{ x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i^2 \leq 1 \right\}$$

dann sind alle $a \in B_2$ für die $a^\top a = 1$ gilt, tangentielle Punkte von $B_2$.

Beweis: die Behauptung folgt unmittelbar aus der Cauchy Schwarz'schen Ungleichung: $(a^\top c)^2 \leq (a^\top a)(c^\top c) \leq (a^\top a)^2$

Als letzten Fall betrachten wir die Einheitskugel $B_1$ der Maximumsnorm

**Lemma 3** Betrachte die Einheitskugel $B_1$ der $\| \cdot \|_\infty$ Norm in $\mathbb{R}^n$

$$B_1 = \left\{ x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n \mid \max_{i} |x_i| \leq 1 \right\}$$

dann ist $a = (a_1, \ldots, a_n)^\top \in B_1$ genau dann ein tangentialer Punkt von $B_1$, wenn $a_i \in \{0, 1, -1\}$.

Beweis: "$\Leftarrow$" folgt aus

$$a^\top a = \sum_{i=0}^{n} a_i^2 = \sum_{i:a_i \neq 0} |a_i| \geq \sum_{i:a_i \neq 0} |c_i| = \sum_{i:a_i \neq 0} |a_i||c_i| \geq |a^\top c| \geq a^\top c$$

"$\Rightarrow$" Sei O.B.d.A $1 > |a_1| > 0$ und setze $c = (1, a_2, \ldots, a_n)^\top$ falls $a_i > 0$, sonst $c = (-1, a_2, \ldots, a_n)^\top$. Nun gilt

$$a^\top a = \sum_{i=1}^{n} a_i^2 < |a_1| + \sum_{i=2}^{n} a_i^2 = \sum_{i=1}^{n} a_i c_i = a^\top c$$

also ist $a^\top a < a^\top c$ und $a$ kann damit kein tangentialer Punkt von $B_1$ sein.

Wegen $B_1 \subset B_2 \subset B_\infty$ ist natürlich jeder tangentielle Punkt von $B_\infty$ oder $B_2$ auch ein tangentialer Punkt der entsprechenden Untermengen.

### 3.2 Wiederholte Anwendung von $T(x) = Ax + b$

In diesem Abschnitt betrachten wir die allgemeine inhomogene Matrixiteration

$$x^{k+1} = Ax^k + b$$

mit $A \in \mathbb{C}^{n \times n}$ und $x^k, b \in \mathbb{C}^n$. Wir haben also $T = \{(A, b)\}$ und betrachten das Konvergenzverhalten der konstanten Folge $\tau$ mit $\tau_i = (A, b)$. Wir führen die verkürzte Schreibweise $\text{Conv}(A, b)$ für $\text{Conv}(\tau)$ etc. ein, wobei $\tau$ die oben erwähnte konstante Folge von affinen Abbildungen darstellt.
Das Ziel dieses Abschnitts besteht darin, notwendige und hinreichende Bedingungen für die verschiedenen Grenzwertverhalten \( \text{Conv}(A, b) \), \( \text{CondConv}(A, b) \), \( \text{Bounded}(A, b) \) bzw. \( \text{Diverge}(A, b) \) der Iteration (3.6) zu erhalten.

Für die Eigenschaft \( \text{Conv}(A, b) \) existieren Resultate aus der Numerik, die solche notwendigen und hinreichenden Bedingungen aufstellen [Gre97]. Für den homogenen Fall \( b = 0 \) und die Eigenschaften \( \text{CondConv}(A, 0) \) und \( \text{Bounded}(A, 0) \) findet man ebenfalls entsprechende Bedingungen in [Lud85] (siehe Satz 2 weiter unten). Für die inhomogene Iteration (3.6) existierte bislang keine vollständige Aufstellung solcher Bedingungen. Diese Charakterisierung wird aber im Kapitel 4 die Analyse von synchronem RL von Bedeutung sein, wo Fragen über Konvergenz oder Beschränktheit des synchronen TD[0] Verfahrens in Verbindung mit Funktionsapproximation in natürlicher Weise aufkommen.

Unsere Analyse wird sich auf die Eigenschaften des Spektrums der Matrix \( A \) stützen. Wie üblich bezeichnen wir mit
\[
\rho(A) := \max \{|\lambda| \mid \lambda \in \sigma(A)\}
\]
den Spectralradius von \( A \). Der zum Eigenwert \( \lambda \) korrespondierende Eigenraum wird mit \( E^A_\lambda = \{ x \mid Ax = \lambda x \} \) bezeichnet. Der korrespondierende Hauptraum wird durch \( H^A_\lambda = \{ x \mid \exists k : (\lambda I - A)^k x = 0 \} \) gekennzeichnet. Aus der Definition folgt direkt, dass \( E^A_\lambda \subseteq H^A_\lambda \) gilt, im Allgemeinen ist aber \( E^A_\lambda \neq H^A_\lambda \). Der Fall \( E^A_\lambda = H^A_\lambda \) bedeutet, dass ein Vektor, der durch Matrix \( (\lambda I - A)^k \) für ein \( k \geq 1 \) annulliert wird, bereits durch die Matrix \( I - A \) annulliert wird (cf. [GLR86]).

Im folgenden definieren wir zwei Eigenschaften einer Matrix \( A \) mit Eigenwerten \( \lambda \in \sigma(A) \) welche nützlich sein werden, um die Resultate zu präsentieren:

\[
\text{Cond1}(A) :\iff \text{aus } |\lambda| = 1 \text{ folgt } \lambda = 1 \quad (3.7)
\]
\[
\text{CondE}(A) :\iff \text{aus } |\lambda| = 1 \text{ folgt } E^A_\lambda = H^A_\lambda \quad (3.8)
\]

Die erste Bedingung sichert zu, dass die 1 der einzige Eigenwert mit Betrag eins ist. Die zweite besagt, dass Eigenräume von Eingenwerten mit Betrag eins, identisch mit ihren Haupräomen sind. Der folgende Satz fasst existierende Aussagen für homogene Iterationen zusammen

\textbf{Satz 2} Für eine homogene Iteration
\[
x^{k+1} = Ax^k
\]
mit \( A \in \mathbb{C}^{n \times n} \) und \( x^k \in \mathbb{C}^n \) gelten folgende Aussagen

\[
\text{Conv}(A, 0) \iff \rho(A) < 1
\]
\[
\text{CondConv}(A, 0) \iff \rho(A) \leq 1 \land \text{CondE}(A) \land \text{Cond1}(A)
\]
\[
\text{Bounded}(A, 0) \iff \rho(A) \leq 1 \land \text{CondE}(A).
\]

Beweis: es wird die Jordansche Normalform von \( A \) benutzt, ein entsprechender Beweis findet sich in [Lud85, Ela99]. \( \diamondsuit \)

An den obigen Bedingungen fällt auf, dass aus der Beschränktheit von (3.9) die bedingte Konvergenz folgt, wenn 1 der einzige Eigenwert mit Betrag eins ist. Im Fall der Konvergenz von (3.9) sind überhaupt keine Eigenwerte mit Betrag eins zugelassen.
3.2. WIEDERHOLTE ANWENDUNG VON $T(x) = Ax + b$


**Satz 3** Für eine homogene Iteration (3.9) gelten folgende Aussagen

\[
\begin{align*}
\text{Conv}(A, 0) & \iff \exists \text{ Matrixnorm } \| \cdot \| \text{ mit } \| A \| < 1 \\
\text{CondConv}(A, 0) & \iff \exists \text{ Vektornorm } \| \cdot \| \text{ mit } \| Ax \| < \| x \| \text{ gdw. } Ax \neq x \\
\text{Bounded}(A, 0) & \iff \exists \text{ Matrixnorm } \| \cdot \| \text{ mit } \| A \| \leq 1
\end{align*}
\]

Beweis: Die erste Bedingung gilt sogar im Fall \(\text{Conv}(A, b)\) mit \(b \neq 0\), siehe [Gre97]. Für die zweite und dritte Äquivalenz siehe [BE97]. Wir gehen hier kurz auf die dritte Äquivalenz ein. Die Bedingung \(\| A \| \leq 1\) ist wegen der Submultiplikativität von Matrixnormen hinreichend für \(\text{Bounded}(A, 0)\). Die Notwendigkeit dieser Bedingung ist Äquivalent zu der Existenz einer Vektornorm \(\| \cdot \|\) mit der Eigenschaft \(\| Ax \| \leq \| x \|\) für alle \(x \in \mathbb{C}^n\). Die entsprechende Matrixnorm ist dann die durch diese Vektornorm induzierte Matrixnorm \(\| A \| = \max_{\| x \| = 1} \| Ax \|\). Die passende Vektornorm erhalten wir mit Hilfe der Definition \(\| x \| = \sup_{n \geq 0} \nu(A^n x)\) wobei \(\nu\) eine beliebige Vektornorm darstellt (alle Normen sind äquivalent). Diese Vektornorm ist wegen der Voraussetzung \(\text{Bounded}(A, 0)\) für \(A\) wohldefiniert. Die Behauptung folgt dann aus

\[
\| Ax \| = \sup_{n \geq 0} \nu(AA^n x) = \sup_{n \geq 1} \nu(A^n x) \leq \sup_{n \geq 0} \nu(A^n x) = \| x \| \quad \Diamond
\]

Nun sind wir für die Vorstellung unseres Hauptsatzes für inhomogene Iterationen gerüstet. Es wird vollständig geklärt, wann sich welches Verhalten für die Iteration (3.6) einstellt. Die Grundidee ist eine Reduktion auf den homogenen Fall, ähnlich wie dies im Abschnitt 3.1 vorgeführt wurde. Durch Homogenisierung reduzieren wir die inhomogene Iteration (3.6) in eine homogene Iteration \(\tilde{x}^{k+1} = \tilde{A}\tilde{x}^k\). Das versetzt uns in die Lage, Satz 2 anzuwenden, um die verschiedenen Verhalten der inhomogenen Iteration zu studieren.

Allerdings gilt im Allgemeinen nicht die Äquivalenz \(\text{CondE}(A) \leftrightarrow \text{CondE}(\tilde{A})\). Deswegen werden wir eine zusätzliche Bedingung \(b \in \text{Im}(I - A)\) brauchen, wobei \(\text{Im}(A) = \{ Ax \mid x \in \mathbb{C}^n \}\) den Bildraum der linearen Abbildung \(A\) bezeichnet (vgl. Abschnitt 2.3).

**Satz 4** Für die inhomogene Matrix Iteration (3.6) gelten folgende Äquivalenzen

\[
\begin{align*}
\text{Conv}(A, b) & \iff \rho(A) < 1 \quad (3.10) \\
\text{CondConv}(A, b) & \iff \rho(A) \leq 1 \land \text{CondConv}(A) \land \text{Cond1}(A) \quad (3.11) \\
\text{Bounded}(A, b) & \iff \rho(A) \leq 1 \land \text{CondE}(A) \quad (3.12)
\end{align*}
\]

Fall 1: \(b \notin \sigma(A)\)
KAPITEL 3. MATRIXITERATIONEN

Fall 2: $1 \in \sigma(A)$

\[
\text{CondConv}(A, b) \iff \rho(A) \leq 1 \land \\
\text{CondE}(A) \land \text{Cond1}(A) \land \\
b \in \text{Im}(I - A)
\]

\[
\text{Bounded}(A, b) \iff \rho(A) \leq 1 \land \\
\text{CondE}(A) \land b \in \text{Im}(I - A).
\]

Beweis: Äquivalenz (3.10) stellt ein Standard Resultat aus der Numerik dar (siehe [Gre97]), und wurde der Vollstandigkeit wegen aufgenommen. Die Iteration (3.6) kann auch in homogener Form geschrieben werden

\[
x^{k+1} = \tilde{A}x^k
\]

mit $\tilde{A} \in \mathbb{C}^{(n+1) \times (n+1)}$ und $\tilde{x}^k \in \mathbb{C}^{n+1}$:

\[
\tilde{A} = \begin{pmatrix} A & b \\ 0 & \ddots & 0 \\ & & 1 \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} x \\ \vdots \\ 1 \end{pmatrix}
\]

Wir bezeichnen mit $p(\lambda)$ und $\tilde{p}(\lambda)$ die charakteristischen Polynome von $A$ bzw. $\tilde{A}$. Auf Grund der Form von $\tilde{A}$ gilt $\tilde{p}(\lambda) = p(\lambda)(1 - \lambda)$ und wir erhalten

\[
\sigma(\tilde{A}) = \sigma(A) \cup \{1\}.
\]

Aus der Blockform von $\tilde{A}$ folgt auch, dass ein Vektor $(x_1, \ldots, x_n, 0)^T$ genau dann einen Eigen- bzw. Hauptvektor von $A$ zum Eigenwert $\lambda$ darstellt, wenn $(x_1, \ldots, x_n, 0)^T$ ein Eigen- bzw. Hauptvektor von $\tilde{A}$ zum gleichen Eigenwert ist. Ist ferner ein Vektor $(x_1, \ldots, x_n, x_{n+1})^T$ mit $x_{n+1} \neq 0$ ein Eigen- bzw. Hauptvektor von $\tilde{A}$, dann kann es nur ein Eigen- oder Hauptvektor von $A$ zum Eigenwert 1 sein. Daher gilt für $\lambda \in \sigma(A)$

\[
E_{\lambda}^A = H_{\lambda}^A \iff E_{\lambda}^A = H_{\lambda}^A \quad \text{und} \\
E_{\lambda}^A = H_{\lambda}^A \iff E_{\lambda}^A = H_{\lambda}^A \text{ für } \lambda \neq 1.
\]

Da die Iteration (3.15) eine homogenisierte Form der Iteration (3.6) ist, gelten die Relationen CondConv$(A, b) \iff$ CondConv$(\tilde{A}, 0)$ und Bounded$(A, b) \iff$ Bounded$(\tilde{A}, 0)$. Wir betrachten zuerst den Fall $1 \notin \sigma(A)$. Aus (3.16) folgt, dass $\lambda = 1$ ein einfacher Eigenwert von $\tilde{A}$ ist. Der korrespondierende Eigenraum ist nämlich mindestens eindimensional, andererseits kann der umschließende Hauptaum maximal eindimensional sein. Daher gilt $E_{\lambda}^A = H_{\lambda}^A$. Zusammen mit (3.17) erhalten wir die Äquivalenz CondE$(A) \iff$ CondE$(\tilde{A})$. Aus (3.16) ergibt sich direkt $\rho(A) \leq 1 \iff \rho(\tilde{A}) \leq 1$. Daher gilt auch die Äquivalenz $\rho(A) \leq 1 \land \text{CondE}(A) \iff \rho(\tilde{A}) \leq 1 \land \text{CondE}(\tilde{A})$. Aus Satz 2 folgt die Äquivalenz zu Bounded$(\tilde{A}, 0)$, welches wiederum äquivalent zu Bounded$(A, b)$ ist. Damit ist die Gültigkeit von (3.12) gezeigt. Wegen (3.16) gilt auch Cond1$(A) \iff \text{Cond1}(\tilde{A})$. Zusammen mit (3.12) ergibt sich die Äquivalenz (3.11).
Wir betrachten nun den interessanteren Fall $1 \in \sigma(A)$. Behauptungen (3.13) und (3.14) werden direkt aus Satz 2 und (3.16) folgen, falls wir die Gültigkeit der Äquivalenz
\[
\text{CondE}(\hat{A}) \iff \text{CondE}(A) \land b \in \text{Im}(I - A)
\]
geben können. Wir nehmen zuerst an, dass $\text{CondE}(\hat{A})$ gilt. Mit (3.17) folgt die Gültigkeit von $\text{CondE}(A)$. Ebenfalls aus $\text{CondE}(\hat{A})$ und den Überlegungen die zu (3.17) geführt haben folgt, dass $\hat{A}$ einen Eigenvektor $\hat{x} = (x, x_{n+1})^\top$ zum Eigenwert 1 hat, für den $x_{n+1} \neq 0$ gilt. ODbA können wir $x_{n+1} = 1$ annehmen. Es gilt
\[
\hat{A}\hat{x} = \hat{x} \iff Ax + b = x
\]
und wir erhalten $\text{CondE}(\hat{A}) \Rightarrow \text{CondE}(A) \land b \in \text{Im}(I - A)$. Umgekehrt sei $\text{CondE}(A) \land b \in \text{Im}(I - A)$. Die Bedingung $b \in \text{Im}(I - A)$ und die Äquivalenz in (3.18) implizieren, dass die Matrix $\hat{A}$ einen Eigenvektor der Form $\hat{x} = (x, x_{n+1})^\top$ mit $x_{n+1} = 1$ besitzt. Daher gilt $\dim \mathcal{E}^A_1 = \dim \mathcal{E}^A_1 + 1$. Aus $\text{CondE}(A)$ folgt, dass $\mathcal{E}^A_1 = \mathcal{H}^A_1$. Und zusammen mit $\mathcal{E}^A_1 \subset \mathcal{H}^A_1$ erhalten wir $\mathcal{E}^A_1 = \mathcal{H}^A_1$. Kombiniert mit (3.17) ergibt sich $\text{CondE}(\hat{A}) \iff \text{CondE}(A) \land b \in \text{Im}(I - A)$. Wir haben damit die Äquivalenz $\text{CondE}(A) \iff \text{CondE}(A) \land b \in \text{Im}(I - A)$ gezeigt. Zusammen mit Satz 2 erhalten wir (3.14). Analog zu oben folgt die Äquivalenz (3.13) aus den Identitäten (3.16) und (3.14).

Um die Aussagen von Satz 4 zu veranschaulichen betrachten wir ein einfaches Beispiel. Die Translation $x^{k+1} =Ix^k + b$ mit $b \neq 0$ divergiert offensichtlich wegen $x^k = kb$. Das kann man aber auch aus $\sigma(I) = \{1\}$ und $b \notin \text{Im}(I - I) = \{0\}$ ablesen, da dies eine Verletzung der Äquivalenz (3.14) in Satz 4 darstellt.

Die Bedingung $b \in \text{Im}(I - A)$ gleich der Existenz eines $c \in \mathbb{C}^n$, so dass $b = (I - A)c$ gilt. Nun ist
\[
\hat{A}^k = \begin{pmatrix} A^k & a_k \\ 0 & \ldots & 0 & 1 \end{pmatrix}
\]
mit
\[
a_k = \sum_{i=0}^{k-1} A^k b
\]
Die Summe in (3.19) wird mit der Bedingung $b = (I - A)c$ zu einer Teleskopsumme
\[
a_k = \sum_{i=0}^{k-1} A^k (I - A)c = \sum_{i=0}^{k-1} A^k c - \sum_{i=1}^{k} A^k c = c - A^k c = (I - A^k)c
\]
Dies beweist erneut, dass die Zusatzbedingung $b \in \text{Im}(I - A)$ hinreichend für die Beschränktheit der Iteration (3.6) ist. Die Notwendigkeit dieser Bedingung im Fall $1 \in \sigma(A)$ ist allerdings nicht so einfach zu sehen. Die Bedingung $b \in \text{Im}(I - A)$ ist für manche Matrizen trivialerweise erfüllt, so ist z.B. für die Matrix $A = (I + ba^\top)$ immer $(I - (I + ba^\top))(-a^\top a)^{-1}a = b$, also $b \in \text{Im}(I - A)$. Diese Tatsache ermöglicht es die Ergebnisse von Satz 1 aus dem allgemeineren Satz 4 auch für komplexwertige Matrizen abzuleiten.
3.3 Stabilität spezieller Iterationsmatrizen

Im Abschnitt 3.2 haben wir allgemeine Matrixiterationen betrachtet. Diese Betrachtungen wollen wir hier für eine bestimmte Klasse von Matrizen spezialisieren. Wir werden insbesondere Matrizen der Form

\[ A = I - \alpha \Phi^T D(I - \gamma P)\Phi \]  

(3.20)

untersuchen, mit \( \Phi \in \mathbb{R}^{n \times m} \) und \( D, P \in \mathbb{R}^{n \times n} \). Dabei werden \( D \) als eine nichtnegative Diagonalmatrix und \( P \) als eine stochastische Matrix angenommen. Wir werden im Kapitel 4 sehen, dass Matrizen aus Gleichung (3.20) in natürlicher Weise beim synchronen TD(0) Verfahren auftreten werden. In diesem Abschnitt wollen wir dies noch außer Acht lassen, und die Konvergenzeigenschaften der Matrixiteration (3.6) mit der spezialisierten Matrix \( A \) aus (3.20) betrachten. Insbesondere werden uns Eigenschaften der Matrix \( \Phi \) interessieren, die \( \rho(A) \leq 1 \) oder sogar \( \rho(A) < 1 \) implizieren. Dies wird uns in natürlicher Weise zum Begriff der positiven Stabilität einer Matrix, und zu der Klasse der so genannten \( M \)-Matrizen führen. Wir beginnen mit einigen Definitionen und Fakten, die aus [HJ91] entnommen wurden.

**Definition 2** Eine Matrix \( A \in \mathbb{C}^{n \times n} \) heißt positiv stabil, falls für den Realteil \( \text{Re}\lambda \) eines jeden Eigenwertes \( \lambda \) von \( A \)

\[ \text{Re}\lambda > 0 \]

gilt. Gilt lediglich \( \text{Re}\lambda \geq 0 \), so heißt \( A \) semi positiv stabil.

**Definition 3** Sei \( A = (a_{ij}) \in \mathbb{R}^{n \times n} \) eine Matrix für die alle nicht diagonalen Einträge kleiner oder gleich 0 sind

\[ a_{ij} \leq 0 \text{ für } i \neq j \]

Dann heißt die Matrix \( A \) eine \( M \)-Matrix, falls \( A \) positiv stabil ist. Ist \( A \) lediglich semi positiv stabil, so spricht man von einer singulären \( M \)-Matrix.  

\( M \)-Matrizen haben viele interessante Eigenschaften, auf einige davon werden wir noch eingehen. Eine umfassende Übersicht findet sich z.B. in [HJ91, BP79]. Die wichtigsten davon fassen wir im folgenden Lemma zusammen

**Lemma 4** Sei \( A \) eine \( M \)-Matrix, und \( D \) eine nichtnegative Diagonalmatrix. Dann gilt

1. Es existieren eine nichtnegative Matrix \( N \) und ein \( \theta > \rho(N) \), so dass man \( A \) als die Differenz \( \theta I - N \) schreiben kann. Andererseits ist jede Matrix der Form \( \theta I - N \) mit \( N \) nichtnegativ und \( \theta > \rho(N) \) automatisch eine \( M \)-Matrix.

2. Die Matrizen \( DA \) und \( AD \) sind \( M \)-Matrizen falls \( D \) regulär ist. Ansonsten sind \( DA \) und \( AD \) singuläre \( M \)-Matrizen.

3. Jede Hauptuntermatrix von \( A \) ist eine \( M \)-Matrix.

Beweis: siehe [HJ91]
3.3. STABILITÄT SPEZIELLER ITERATIONSMATRIZEN

Nun können wir folgendes Lemma formulieren

**Lemma 5** Seien $D$ eine nichtnegative Diagonalmatrix, $P$ eine stochastische Matrix und $\gamma \in [0, 1]$, dann ist $D(I - \gamma P)$ eine $M$-Matrix bzw. ein singuläre $M$-Matrix, falls $D$ singulär ist.

Beweis: Da $P$ eine stochastische Matrix ist, gilt $\rho(P) \leq 1$. Wegen $0 \leq \gamma < 1$ haben alle Eigenwerte von $I - \gamma P$ einen positiven Realteil. Da $P$ ferner nichtnegativ ist, sind die nicht diagonalen Einträge von $I - \gamma P$ nicht positiv, und $I - \gamma P$ ist damit eine $M$-Matrix. Die Behauptungen folgen nun mit Lemma 4. ◊

Wir interessieren uns nun für Matrizen $P \in \mathbb{R}^{n \times m}$ mit $m \leq n$, für die das Produkt $\Phi^T D(I - \gamma P) \Phi$ (semi) positiv stabil ist. Dabei soll die positive Stabilität unabhängig von der Wahl der stochastischen Matrix $P$ und der nichtnegativen Diagonalmatrix $D$ gelten.

Eine andere Möglichkeit besteht darin, die Klasse der Matrizen $P$ und $D$ so einzuschränken, dass $\Phi^T D(I - \gamma P) \Phi$ für beliebige (semi) positiv stabil ist. Eine solche Möglichkeit ist die Wahl von stochastischen Matrizen $P$, die einen stationären Verteilungsvektor $d$ mit $d^T P = d^T$ besitzen\(^1\). Für ein solches $P$ kann man zeigen, dass

$$x^T \Phi^T D(I - \gamma P) \Phi x \geq 0, \quad x \in \mathbb{R}^n$$

(3.21) gilt (bzw. > 0, falls die Spalten von $\Phi$ linear unabhängig sind), falls $D$ die spezielle Form $D = \text{diag}(d)$ hat (siehe [BT96], Lemma 6.6). Da die Ungleichung (3.21) nur für reelle $x$ gilt, kann man nicht auf die Symmetrie von $A$ schließen, insbesondere kann $A$ komplexwertige Eigenwerte haben. Das folgende Lemma zeigt aber, dass der Realteil eines solchen Eigenwertes nichtnegativ sein muss

**Lemma 6** Sei $A \in \mathbb{R}^{n \times n}$, und es gelte $x^T A x > 0$ (bzw. $\geq 0$) für alle $x \in \mathbb{R}^n$. Dann ist $A$ (semi) positiv stabil.

Beweis: sei $\lambda \in \mathbb{C}$ ein Eigenwert von $A$ mit dem Eigenvektor $0 \neq z = x + iy \in \mathbb{C}^n$ ($x, y \in \mathbb{R}^n$). Aus $A z = \lambda z$ und der Voraussetzung folgt

$$\lambda \frac{z^* z}{>0} = (x^T - iy^T) A (x^T + iy^T) = x^T A x + y^T A y + i(x^T A y - y^T A x) \geq 0$$

$$\in \mathbb{R}$$

Da $z^* z = |z|^2 > 0$ reell ist, sieht man, dass der Realteil von $\lambda$ nichtnegativ sein muss. Er ist gar positiv, falls die verschärfte Voraussetzung $x^T A x > 0$ gilt. ◊

Lemma 6 und Ungleichung (3.21) implizieren zusammen, dass für die spezielle Wahl von $P$ und $D = \text{diag}(d)$ mit $d^T P = d^T$ die Matrix $\Phi^T D(I - \gamma P) \Phi$ (semi) positiv stabil ist, und zwar unabhängig von der Matrix $P$. Im weiteren Verlauf, wollen wir dies umkehren, und nach Möglichkeit eine große Klasse von Matrizen $P$ bestimmen, für die $\Phi^T D(I - \gamma P) \Phi$ unabhängig von $P$ und $D$ (semi) positiv stabil sind.

Wir betrachten zuerst den quadratischen Fall $\Phi \in \mathbb{R}^{n \times n}$. Die einfachste Wahl stellen dabei orthogonale $\Phi$ dar, da in diesem Fall $\Phi^{-1} = \Phi^T$ gilt, und die

\(^1\)Dazu gehören insbesondere Grenzwerte $\lim_{k \to \infty} P^k$ von primitiven stochastischen Matrizen, d.h. Matrizen für die ein $k \in \mathbb{N}$ existiert, so dass $P^k > 0$ gilt, vgl. [HJ85].
Ähnlichkeitstransformation bei Matrizen ihre Eigenwerte erhält. Wir fassen dies in folgendes

**Lemma 7** Für eine $M$-Matrix $A$ und eine orthogonale Matrix $\Phi \in \mathbb{R}^{n \times n}$ ist das Produkt $B = \Phi^T A \Phi$ eine positiv stabile Matrix. $B$ braucht aber keine $M$-Matrix zu sein, auch nicht wenn die Matrix $A$ die spezielle Form $A = I - \gamma P$ hat, wobei $P$ eine stochastische Matrix ist.

Beweis: Der erste Teil der Behauptung folgt aus $\Phi^T = \Phi^{-1}$ und der Tatsache, dass die Matrizen $A$ und $\Phi^{-1} A \Phi$ die gleichen Eigenwerte haben. Wir begründen jetzt, warum die resultierende Matrix nicht unbedingt eine $M$-Matrix ist. Wäre $B = \Phi^T A \Phi$ eine $M$-Matrix, so müssten laut Lemma 4 alle Hauptuntermatrizen von $B$ ebenfalls $M$-Matrizen sein, wir werden aber im Beispiel 2 sehen, dass dies im Allgemeinen nicht der Fall ist.

Nun wollen wir uns dem Fall $m < n$ widmen. Um den nächsten Satz zu formulieren, benötigen wir die Blocknotation $\Psi = [\Phi|\Theta]$. Diese drückt aus, dass die ersten Spalten der Matrix $\Psi$ mit der Matrix $\Phi$ übereinstimmen, $\Phi$ also eine Teilmatrix von $\Psi$ darstellt. Manchmal kann man ausgehend von der Teilmatrix $\Phi$ auch Eigenschaften der umgebenden Matrix $\Psi$ fordern. Hat z.B. $\Phi$ orthogonale Spalten, so existiert nach dem Schmidt’schen Orthogonalisierungsverfahren eine Matrix $\Theta$, so dass $\Psi = [\Psi|\Theta]$ eine orthogonale Matrix ist. Diese Tatsache wird auch im folgenden Satz benutzt.

**Satz 5** Sei $A$ eine $M$-Matrix und $\Phi \in \mathbb{R}^{n \times m}$ eine Matrix, die Teilmatrix einer nichtnegativen orthogonalen Matrix $\Psi = [\Phi|\Theta] \in \mathbb{R}^{n \times n}$ ist. Dann ist $\Phi^T A \Phi$ eine $M$-Matrix, also insbesondere positiv stabil.

Beweis: laut Lemma 4 kann man $A$ schreiben als $A = \theta I - N$ mit einer nichtnegativen Matrix $N$ und $\theta > \rho(N)$. Nun ist

$$\Psi^T A \Psi = \Psi^T (\theta I - N) = \theta \Psi^T \Psi - \Psi^T N \Psi = \theta I - \Psi^T N \Psi$$

Da $\Psi$ und $N$ nichtnegativ sind, ist $\Psi^T N \Psi$ ebenfalls nichtnegativ. Wegen der Orthogonalität von $\Psi$ ist ferner $\rho(\Psi^T N \Psi) = \rho(N)$. Damit ist nach Lemma 4 die Matrix $\Psi^T A \Psi$ eine $M$-Matrix. Man kann $\Psi^T A \Psi$ unter Ausnutzung der Blockstruktur schreiben als

$$\Psi^T A \Psi = [\Phi, \Theta]^T A [\Phi, \Theta] = \begin{pmatrix} \Phi^T A \Phi \\ \Phi^T A \Theta \\ \Theta^T A \Phi \\ \Theta^T A \Theta \end{pmatrix}.$$ 

Daraus folgt, dass $\Phi^T A \Phi$ eine Hauptuntermatrix einer $M$-Matrix ist, und damit nach Lemma 4 selbst eine $M$-Matrix ist.

Die Aussage des obigen Satzes ist relativ beschränkt, da man leicht nachweisen kann, dass die Klasse der orthogonalen und nichtnegativen Matrizen mit der Klasse der Permutationsmatrizen übereinstimmt. Die Matrix $\Phi$ ist damit Teil einer Permutationsmatrix, und falls $m < n$ gilt, muss $\Phi$ mindestens $n - m$ Zeilen enthalten, die nur 0 Einträge haben.

Man könnte versuchen die Forderungen an die Matrix $\Phi$ im Satz 5 abzuschwächen, und lediglich fordern, dass $\Phi$ orthogonale Spalten und nichtnegative Einträge hat. Diese Forderung ist aber zu schwach, wie das folgende Beispiel zeigt.

$\diamondsuit$
3.3. STABILITÄT SPEZIELLER ITERATIONSMATRizen

Beispiel: 2 Wir betrachten die Matrizen $\Phi$ und $I - \gamma P$ mit

$$\Phi = \begin{pmatrix} 0.9 & 0 \\ \sqrt{1 - 0.9^2} & 0 \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$\Phi$ ist nichtnegativ und hat orthogonale Spalten, ferner ist $P$ eine stochastische Matrix. Nun gilt

$$B = \Phi^T (I - \gamma P) \Phi = I - \gamma \begin{pmatrix} d & 0 \\ 0 & 1 \end{pmatrix}$$

mit $d = 0.9^2 + 0.9\sqrt{1 - 0.9^2} \approx 1.2023$. Damit ist $B$ für $\gamma > \frac{1}{d} \approx 0.83$ nicht positiv stabil. Andererseits ist $B$ für $\gamma < \frac{1}{d}$ positiv stabil.

Die Probleme im Beispiel 2 stammen davon, dass eine Hauptuntermatrix einer positiv stabilen Matrix nicht positiv stabil zu sein braucht. Im Beispiel 2 gab es aber eine Schranke $d$, so dass für $\gamma \in [0, d^{-1})$ diese Eigenschaft sicher gestellt werden konnte. Wir werden im folgenden Satz sehen, dass sich so ein Schranke immer finden lässt, und zwar sogar für beliebige Matrizen $A \in \mathbb{R}^{n \times m}$, die vollen Spaltenrang besitzen.

Satz 6 Sei $A \in \mathbb{R}^{n \times m}$ eine beliebige $m \times n$ Matrix mit $m \leq n$ und mit vollem Spaltenrang. Dann existiert eine Schranke $r = \rho(H(P))^{-1} > 0$, so dass für $\gamma \in [0, r)$ die Matrix $D(I - \gamma P) \Phi$ (semi) positiv stabil ist.

Beweis: Laut [HJ91] Seite 95 ist die positive Definitheit von $H(A) := \frac{1}{2}(A + A^*)$, eine hinreichende Bedingung sowohl für die positive Stabilität von $A$ als auch für die positive Stabilität sämtlicher Hauptuntermatrizen von $A$. Betrachten wir nun den Fall für reguläre $\Phi \in \mathbb{R}^{n \times n}$. Es gilt

$$H(\Phi^T D(I - \gamma P) \Phi) = \Phi^T H(D(I - \gamma P)) \Phi = \Phi^T D(I - \frac{\gamma}{2}(P + P^T)) D\Phi$$

und damit ist für

$$\gamma < \frac{1}{\rho(H(P))} = \frac{2}{\rho(P + P^T)}$$

(3.22) die Matrix $G = I - \frac{\gamma}{2}(P + P^T)$ positiv definit. Da aus $x^T G x > 0$ ebenfalls $x^T \Phi^T D G D \Phi x = (D \Phi x)^T A(D \Phi x) > 0$ folgt, ist insgesamt auch $H(\Phi^T D(I - \gamma P) \Phi)$ positiv definit. Für $\Phi \in \mathbb{R}^{n \times m}$ mit $m < n$ und vollem Spaltenrang ist $\Phi$ eine Untermatrix einer regulären Matrix $\Psi = [\Phi, \star]$. Damit ist aber $\Phi^T D(I - \gamma P) \Phi$ eine Hauptuntermatrix von $\Psi^T D(I - \gamma P) \Psi$ und somit eine positiv stabile Matrix.

Ob die Schranke $r = \rho(H(P))^{-1}$ in Satz 6 weiter nach oben verbessert werden kann, ist zur Zeit nicht bekannt. Für symmetrische stochastische Matrizen gilt aber immer $r = \rho(H(P)) = \rho(P) = 1$. Man kann dies auf stochastische Matrizen erweitern, die mittels einer orthogonalen Transformation symmetrisiert werden können. Dies führt zum folgenden Korollar
Korollar 2 Seien $\Phi$ und $P$ wie im Satz 6. Gilt $\rho(H(P)) = 1$ so ist $\Phi^\top D(I - \gamma P)\Phi$ für $\gamma \in [0,1)$ immer positiv stabil. Der Fall $\rho(H(P)) = \rho(P) = 1$ ist insbesondere immer dann gegeben, wenn $P$ symmetrisch ist. Es gilt aber auch für alle $P$, für die die eine orthogonale Matrix $U$ gibt, so dass $UPU^\top$ symmetrisch ist. Zu dieser Klasse gehören unter anderem die normalen Matrizen (Matrizen mit der Eigenschaft $PP^\top = P^\top P$).

Beweis: es folgt direkt aus dem Beweis von Satz 6, dass die Bedingung $\rho(H(P)) = 1$ hinreichend für die positive Stabilität der Matrix $\Phi^\top D(I - \gamma P)\Phi$ ist. Ist $P$ symmetrisch, so gilt trivialerweise $\rho(H(P)) = \rho(P) = 1$. Ist die Matrix $UPU^\top$ für eine orthogonale Matrix $U$ symmetrisch, so gilt wegen der Ähnlichkeitsinvarianz des Spektralradius $\rho(H(U PU^\top)) = 1$ für $P$ normal, dann existiert eine orthogonale Matrix $U$, so dass $UPU^\top$ diagonal, also insbesondere symmetrisch ist (vgl. [HJ85] Satz 2.5.4).

Satz 6 bzw. Korollar 2 lassen sich ferner für $\Phi$ mit nicht vollem Spaltenrang erweitern, allerdings muss man dann die positive Stabilität durch semi positive Stabilität ersetzen.

Leider liefern die Aussagen von Satz 6 keine guten Schranken für den Diskontierungsfaktor $\gamma$, wenn die Matrix $P$ nicht symmetrisch ist (bzw. nicht den anderen Bedingungen aus Korollar 2 entspricht). Wir werden im Beispiel 3 eine stochastische Matrix $P \in \mathbb{R}^{n \times n}$ konstruieren, für die $\rho(H(P)) = \frac{1}{2}(1 + \sqrt{n})$ gilt. Damit die Forderungen aus Satz 6 genügt, müsste es aus dem Intervall $[0, 2/(1 + \sqrt{n})]$ stammen, dessen Länge mit steigendem $n$ gegen $0$ geht. Allerdings sind die Forderungen in Satz 6 hinreichend aber nicht notwendig. Zur Zeit sind keine hinreichenden und notwendigen Bedingungen bekannt, die das Vererben der positiven Stabilität einer Matrix an ihre Hauptuntermatrizen charakterisieren würden.

Beispiel: 3 Betrachte die stochastische Matrix $P \in \mathbb{R}^{n \times n}$ mit

$$P = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix} \quad \text{und} \quad \frac{1}{2}(P + P^\top) = \frac{1}{2} \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

Ferner benötigen wir die Transformationsmatrix $S = I - T$ mit

$$T = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$
3.3. STABILITÄT SPEZIELLER ITERATIONSMATRIZEN

Man rechnet leicht nach, dass \( S^{-1} = I + T \) gilt, und dass die Matrix \( S^{-1} = \frac{1}{2}(P + P^\top)S^{-1} \) fast eine obere Dreiecksmatrix ist

\[
S^{-1} = \frac{1}{2} \begin{pmatrix}
2 & (n-1) & * & \cdots & * \\
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & 0 \\
0 & 0 & 0 & \cdots & 0 
\end{pmatrix}
\]

Damit erhalten wir

\[
\det \left( \lambda I - S^{-1} \right) = \lambda^{n-2} \left( (\lambda - 1)\lambda - \frac{1}{4}(n-1) \right)
\]

woraus unmittelbar

\[
\rho(H(P)) = \rho \left( \frac{1}{2}(P + P^\top) \right) = \rho \left( S^{-1} \right) = \frac{1 + \sqrt{n}}{2}
\]

folgt. Der Spektralradius von \( H(P) \) ist also proportional zu \( \sqrt{n} \). Man könnte vermuten, dass dies nur für irreduzible Matrizen der Fall sein kann. Dem ist nicht so, man kann die Matrix \( P \) durch eine positive und damit irreduzible Matrix \( \Phi \) ersetzen

\[
P_\epsilon = \begin{pmatrix}
1 - \epsilon(n-1) & \epsilon & \cdots & \epsilon \\
\vdots & \vdots & \ddots & \vdots \\
1 - \epsilon(n-1) & \epsilon & \cdots & \epsilon
\end{pmatrix}
\]

mit \( \epsilon < (n-1)^{-1} \). Es gilt nämlich, dass \( \det(\lambda I - H(P_\epsilon)) = \det(\lambda I - H(P)) \) für \( \epsilon \to 0 \), und damit strebt auch der Spektralradius \( \rho(P_\epsilon) \) von \( P_\epsilon \) gegen \( \frac{1}{2}(1 + \sqrt{n}) \). 

Wir wollen nun auf eine Klasse von Matrizen \( \Phi \in \mathbb{R}^{n \times m} \) eingehen, für die \( \Phi^\top D(I - \gamma P) \Phi \) immer positiv stabil bleibt.

**Satz 7** Sei \( \Phi \in \mathbb{R}^{n \times m} \) eine Matrix mit orthogonalen Spalten und Einträgen aus \( \{0, r\} \) mit \( r > 0 \). Ferner seien \( P \) eine beliebige stochastische und \( D \) eine nichtnegative Diagonalmatrix. Dann ist die Matrix \( \Phi^\top D(I - \gamma P) \Phi \) semi positiv stabil, und sogar positiv stabil, falls \( \Phi \) vollen Spaltenrang hat und \( D \) regulär ist.

Beweis: O.B.d.A. setzen wir \( r = 1 \), und betrachten zuerst den Fall, dass \( \Phi \) vollen Spaltenrang hat. Da die Spalten von \( \Phi \) zueinander orthogonal sind, ist \( B = \Phi^\top D \Phi \) eine positive Diagonalmatrix. Ferner gilt \( \Phi e = e \), mit \( e = (1, \ldots, 1)^\top \), da \( \Phi \) in jeder Zeile genau einen positiven Eintrag \( r = 1 \) hat. Der Vektor \( e \) ist als generischer Vektor beliebiger Dimension mit lauter Einträgen zu verstehen. Die Dimension dieses Vektors ist immer aus dem Kontext ersichtlich, und wird daher nicht besonders gekennzeichnet. Die Matrix \( P \) ist eine stochastische Matrix ist, und es gilt folglich \( P e = e = \Phi e \). Damit ergibt sich

\[
\Phi^\top D P \Phi e = \Phi^\top D P e = \Phi^\top D P e = B e
\]

und wir erhalten für \( Q = B^{-1} \Phi^\top D P \Phi \)

\[
Q e = B^{-1} \Phi^\top D P \Phi e = e
\]
KAPITEL 3. MATRIXITERATIONEN

also ist \( Q \) eine stochastische Matrix. Nun ist die Matrix

\[
B^{-1}\Phi^\top D(I - \gamma P)\Phi = B^{-1}\Phi^\top D\Phi - \gamma B^{-1}\Phi^\top D\Phi = I - \gamma Q
\]

wegen der Form \( I - \gamma Q \) nach Lemma 5 eine \( M \)-Matrix. Da mit \( B \) auch \( B^{-1} \) eine positive Diagonalmatrix ist, und \( M \)-Matrizen diagonal stabil sind, folgt aus der Tatsache, dass \( B^{-1}\Phi^\top D(I - \gamma P)\Phi \) eine \( M \)-Matrix ist, auch dass \( \Phi^\top D(I - \gamma P) \) eine \( M \)-Matrix ist, also insbesondere positiv stabil ist.

Der Fall, in dem die Matrizen \( B \) und/oder \( D \) nicht den vollen Spaltenrang haben erfolgt analog, man muss lediglich die Inverse durch die Pseudoinverse und stochastisch durch substochastisch ersetzen.

Im Beweis von Satz 7 haben wir gesehen, dass die Matrix \( \Phi \) die Eigenschaft \( \Phi e = e \) hat. Es stellt sich die Frage, ob diese Bedingung alleine ausreicht, um die Aussagen aus Satz 7 zu beweisen. Dies ist leider nicht möglich, wie das folgende Beispiel zeigt.

**Beispiel: 4** Betrachte die Matrizen

\[
\Phi = \frac{1}{3} \begin{pmatrix}
1 & 2 & 0 & 0 \\
1 & 0 & 2 & 0 \\
1 & 0 & 2 & 0 \\
1 & 0 & 0 & 2 \\
2 & 0 & 0 & 1
\end{pmatrix}, \quad P = \begin{pmatrix}
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]

dann gilt \( \Phi(1,1,1,1)^\top = (1,1,1,1,1)^\top \) aber für \( A = \Phi^\top D(I - \gamma P)\Phi \) mit \( \gamma = 0.95 \) ist

\[
\sigma(A) \approx \{ -0.021 \pm 0.076i, 0.793, 0.443 \}
\]

und damit ist die Matrix \( A \) nicht positiv stabil. Für kleinere Werte von \( \gamma \), z.B. für \( \gamma \leq 0.92 \) ist die Matrix \( A \) positiv stabil. Dies ist konform zu der Aussage in Satz 6. Für andere Matrizen \( \Phi \) kann es aber passieren, dass \( \gamma \leq 0.92 \) nicht ausreichend ist. Laut (3.22) im Beweis von Satz 6 ist dies erst für \( \gamma < \rho(H(P))^{-1} = 2/(\rho(P + P^\top)) \) garantiert. Da \( P \) analog zu der stochastischen Matrix im Beispiel 3 aufgebaut ist, gilt \( \rho(H(P))^{-1} = 2/(1 + \sqrt{5}) \approx 0.61803. \) Für \( \gamma < 0.61803 \) ist daher die positive Stabilität von \( A \) für beliebige \( \Phi \) garantiert.

Das Beispiel 4 ähnelt dem bekannten Beispiel von Baird [Bai95] mit

\[
\Phi = \begin{pmatrix}
1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}, \quad P = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]

Auch hier gilt \( \frac{1}{5}\Phi e = e \), es ist allerdings \( m = 8 > 7 = n \). Das Beispiel 4 ist kleiner, ferner ist es mit \( m < n \) realistischer, denn \( \Phi \) wird später als Approximatormatrix \( \Phi \)
3.4 UNENDLICHE PRODUKTE VON SCHIEFEN PROJEKTIONEN

aufgefasst, welche normalerweise die Anzahl der Dimensionen reduziert, anstatt sie zu erhöhen.

In [Sch03] findet sich eine vollständige Analyse des Beispiels von Baird mit den Matrizen \( \Phi \) und \( P \) gegeben durch Formel (3.23). Unter anderem wird auch auf das Phänomen hingewiesen, dass für kleinere \( \gamma \) die Realteile der Eigenwerte von \( \Phi^\top D(I - \gamma P)\Phi \) positiv werden, ohne es aber als eine Notwendigkeit zu charakterisieren, die bei uns allgemein aus Satz 6 folgt.

Neben der Eigenschaft \( \Phi e = e \) im Beweis von Satz 7 war es gefordert, dass \( \Phi \) Einträge in \( \{0, r\} \) für ein \( r > 0 \) haben muss. Auch diese Forderung alleine reicht nicht, um eine zum Satz 7 äquivalente Aussage zu machen. Dieses Beispiel werden wir später im Abschnitt 5.3 wiedersprechen, um die Divergenz des CMAC Approximators zu demonstrieren.

**Beispiel: 5** Betrachte die Matrizen

\[
\Phi = \begin{pmatrix}
1 & 0 \\
0 & 1 \\
1 & 1 \\
\end{pmatrix}, \quad P = \begin{pmatrix}
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{pmatrix}
\]

dann sind die Einträge der Matrix \( \Phi \) in \( \{0, 1\} \) aber die Matrix \( A = \Phi^\top D(I - \gamma P)\Phi \) mit \( \gamma = 0.9 \) hat sowohl positive als auch negative Eigenwerte:

\[\sigma(A) \approx \{1, -0.6\}\]

Damit ist die Matrix \( A \) nicht semi positiv stabil.

Außer der Ergebnisse in den Sätzen 5 und 7 konnte kein weiteres positives Ergebnis einer Matrixklasse \( \Phi \) aufgestellt werden, die die positive Stabilität der \( M \)-Matrix \( I - \gamma P \) aufrechterhalten würde. Wir haben dafür eine Reihe von Gegenbeispielen konstruieren können. Es wäre interessant der Frage nachzugehen, ob es überhaupt Klassen von Approximatrixmatrizen \( \Phi \) gibt, die die positive Stabilität der Matrix \( I - \gamma P \) aufrechterhalten, und andere als die in den Sätzen 5 und 7 geforderten Eigenschaften haben. Eine negative Antwort auf diese Frage würde bedeuten, dass sobald eine Matrix \( \Phi \) die Voraussetzungen von Satz 5 oder Satz 7 erfüllt, dann auch Matrizen \( D \) und \( P \) existieren, für die die Matrix \( \Phi^\top D(I - \gamma P)\Phi \) nicht stabil ist. Eine Klärung dieser Frage würde weitreichende Folgen für die generelle Eignung von Funktionsapproximatorenbem Reinforcement Lernen haben (vgl. Abschnitt 4.3).

### 3.4 Unendliche Produkte von schiefen Projektionen

In diesem Abschnitt betrachten wir den bisher allgemeinsten Fall von Matrixiterationen. Es handelt sich dabei nicht um die wiederholte Anwendung einer einzigen affinen Abbildung, sondern um abwechselnde Ausführung verschiedener affiner Abbildungen. Wir betrachten dabei eine nicht unbedingt endliche Menge \( T \) von speziellen affinen Abbildungen \( T : \mathbb{R}^n \rightarrow \mathbb{R}^n \)

\[T(x) = (I + ab^\top)x + \tau a\]
Wir werden vornehmlich die homogenisierte Form (3.3) verwenden, so dass wir uns auf unendliche Produkte

\[ \ldots A_k A_{k-1} \ldots A_1, \quad A_i \in \mathcal{T} \]

von Matrizen konzentrieren können, wobei in unserem Fall alle Matrizen die spezielle Form \( A = I + ab^\top \) haben. Solche Matrizen werden auch schiefe Projektionen genannt [VEB99]. Die Operation

\[ (I + ab^\top)x = x + (b^\top x)a \]

ist im Fall von \( a = -b/(b^\top b) \) eine Orthogonalprojektion, dabei wird \( x \) auf den zum Vektor \( b \) orthogonalen Unterraum \([b]^\perp\) projiziert. Im Fall einer schieben Projektion wird allerdings in Richtung des Vektors \( a \) verschoben, der sich im Allgemeinen von \( b \) unterscheidet.

In [VEB99] wird behauptet, dass das Konvergenzverhalten von unendlichen Produkten schiefer Projektionen nicht einfacher zu analysieren ist, als das Verhalten von unendlichen Produkten normaler Matrizen, die nicht diese besondere Struktur besitzen. Wir werden auf die Resultate über allgemeine unendliche Matrixprodukte im weiteren Verlauf eingehen. Im Anschluss werden wir allerdings zeigen, dass für bestimmten Arten von solchen schieben Projektionen (\( a \) und \( b \) haben eine spezielle Form), Eigenschaften, wie z.B. Beschränktheit, garantiert werden können. Dies wird später Anwendungen im asynchronen RL haben.

Aussagen über das Verhalten von unendlichen Matrixprodukten, lassen sich nicht mehr nur mit Hilfe von Spektralradien der beteiligten Matrizen charakterisieren. Es ist zwar in allen Fällen notwendig, dass für eine Matrix \( A \in \mathcal{T} \) der Spektralradius \( \rho(A) \leq 1 \) ist. Aber es gibt z.B. Fälle, in denen für zwei Matrizen \( A \) und \( B \), sowohl \( \rho(A) < 1 \) als auch \( \rho(B) < 1 \) gilt, aber der Spektralradius \( \rho(AB) \) der Produktmatrix \( AB \) größer als 1 ist.

**Beispiel: 6** Betrachte die Matrizen

\[ A = \begin{pmatrix} \kappa & 0 \\ \kappa & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \kappa & \kappa \\ 0 & 0 \end{pmatrix}, \quad \text{und} \ AB = \begin{pmatrix} \kappa^2 & \kappa^2 \\ \kappa^2 & \kappa^2 \end{pmatrix} \]

Es gilt \( \rho(A) = \rho(B) = \kappa \) und \( \rho(AB) = 2\kappa^2 \). Ist \( 1 > \kappa > 1/\sqrt{2} \), dann sind einseitig \( \rho(A) < 1 \) und \( \rho(B) < 1 \), während \( \rho(AB) > 1 \) gilt. Damit divergiert das endliche Produkt

\[ (AB)^k \to \infty \]

für \( 1 > \kappa > 1/\sqrt{2} \) und \( k \to \infty \), obwohl sowohl \( A^k \) als auch \( B^k \) für gleiches \( \kappa \) und \( k \to \infty \) gegen die Null-Matrix konvergieren.

Damit ist klar, dass man kein Analogon zum Satz 2 im Fall von mehr als einer Matrix aufstellen kann. Allerdings können die Aussagen von Satz 3 weitgehend gerettet werden.
3.4. UNENDLICHE PRODUKTE VON SCHIEFEN PROJEKTIONEN

Satz 8 Sei $T$ eine (nicht unbedingt endliche) Menge von Matrizen, $T^n$ bezeichnet den Raum aller Folgen in $T$. Ferner bezeichnen wir mit $S \subset T$ eine endliche Teilmenge von $T$. Dann gelten

\[
\forall \tau \in T^n \text{ gilt } \text{Conv}(\tau) \iff \exists \text{ Matrixnorm } \| \cdot \| \text{ und } \kappa < 1 \text{ mit } \|\tau_i\| \leq \kappa < 1 \\
\forall \tau \in S^n \text{ gilt } \text{CondConv}(\tau) \iff \exists \text{ Vektornorm } \| \cdot \| \text{ mit } \|\tau_i(x)\| < \|x\| \text{ gdw. } \tau_i(x) \neq x \\
\forall \tau \in T^n \text{ gilt } \text{Bounded}(\tau) \iff \exists \text{ Matrixnorm } \| \cdot \| \text{ mit } \|\tau_i\| \leq 1
\]

Wichtig ist dabei die Tatsache, dass die geforderte (Matrix-)Norm für alle Matrizen $A$ in $T$ bzw. $S$ die gleiche ist.

Beweis: Die erste Implikation folgt unmittelbar aus der Existenz einer Matrixnorm mit $\|\tau_i\| \leq \kappa < 1$ und der Submultiplikatheit von Matrixnormen. Es ist auch klar, dass man die Forderung nach einer uniformen Schranke $\kappa < 1$, die für alle Matrizen gilt, nicht auf lediglich die Forderung $\|A\| < 1$ abschwächen kann. Ein Gegenbeispiel stellt die Menge der eindimensionalen Matrizen $T = \{0.9^{1/k} | k = 1, 2, \ldots\}$ dar. Es gilt $\|A\| \leq 1$ für alle $A$ in $T$ aber das unendliche Produkt

\[
\prod_{k=1}^{\infty} (0.9^{1/k}) \longrightarrow 0.9^{e^2}
\]

konvergiert nicht gegen 0, da $\sum_{k=1}^{n} 1/n^2 = \pi^2/6$ gleich (vgl. [Heu90]).


\[
\|x\| = \sup_{\nu \geq 0} \sup_{\tau \in T_n} \nu(\tau_n \cdots \tau_1 x)
\]

definiert wird. Dabei bezeichnet $T_n := (T \cup \{I\})^n$ alle endlichen Folgen der Länge $n$, wobei $T$ um um die Identität erweitert wurde. Auf Grund der Normenaquivalenz ist $\nu$ ist eine beliebige Vektornorm auf $\mathbb{C}^n$.

Satz 8 ist praktisch, wenn man eine entsprechende Norm zur Verfügung hat. Bereits anhand der Beweisskizze für die Beschränktheit von beliebigen Matrixprodukten kann man aber erahnen, dass eine entsprechende Norm unter Umständen schwer zu finden sein wird. Praktischer wäre es, wenn die Matrizen in Bezug auf eine geläufige Matrixnorm die Kriterien von Satz 8 erfüllen würden. Manchmal ist dies möglich, wie wir im Satz 10 sehen werden. Wir beginnen allerdings mit einem negativen Ergebnis für die Spektralnorm.

Lemma 8 Sei $A = I + ab^T \in \mathbb{C}^{n \times n}$, $a, b \neq 0$ und $a \neq rb$ für alle $r \in \mathbb{C}$ ($a$ und $b$ sind nicht kolinear), dann gilt für die Spektralnorm $\|A\|_2 > 1$.  

\[
\text{Lemma 8: } \text{Sei } A = I + ab^T \in \mathbb{C}^{n \times n}, a, b \neq 0 \text{ und } a \neq rb \text{ für alle } r \in \mathbb{C} \text{ (}a \text{ und } b \text{ sind nicht kolinear), dann gilt für die Spektralnorm } \|A\|_2 > 1.
\]
Beweis: Die Spektralnorm $\| \cdot \|_2$ ist eine von der euklidischen Norm $\| \cdot \|_2$ induzierte Matrixnorm $\| A \|_2 := \max_{\|x\|_2 = 1} \| Ax \|_2$ mit $\| Ax \|_2^2 = x^\top A^\top Ax$. Es gilt

$$\|(I + ab^\top)x\|_2^2 = x^\top (I + ab^\top)^\top(I + ab^\top)x$$
$$= x^\top (I + (ba^\top + ab^\top) + ba^\top ab^\top )x^\top$$
$$= 1 + 2(x^\top a)(x^\top b) + (x^\top b)^2 a^\top a \geq 0$$

Auf Grund der Voraussetzung $a \neq rb$ existiert ein normalisiertes $x$ für welches $x^\top a = 0$ und $x^\top b = s \neq 0$ gelten (nehmen z.B. $x = \|b\|_2^{-1}(b - (a^\top b)/(a^\top a)a$ und normalisieren es). Damit ist $\|(I + ab^\top)x\|_2 = 1 + s^2/a^\top a > 1$ und insbesondere $\|A\|_2 = \max_{\|b\|_2 = 1} \|Ay\|_2 \geq \|Ax\|_2 > 1$.

Lemma 8 ist insofern von Bedeutung, als in der homogenisierten Form (3.3) für die Matrix $I + \tilde{a}b^\top$ immer $\tilde{a} \neq \tilde{tb}$ für alle $t \in \mathbb{R}$ gegeben ist. Man kann aus Lemma 8 weitere Aussagen für andere Matrixnormen ableiten. Z.B. sieht man wegen der Identität $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$ (vgl. [HJ85], Seite 313), dass mit $\|A\|_2 > 1$ immer mindestens eine der beiden Normen $\|A\|_1$ oder $\|A\|_\infty$ größer 1 ist.

Ab jetzt betrachten wir Matrizen mit der speziellen Form $A = I + a(\gamma b - a)^\top$. Wie man leicht aus Lemma 8 sieht, ist auch hier $\|A\|_2 > 1$ wenn $a$ und $b$ nicht kollinear sind.

Wir suchen insbesondere nach Klassen $\mathcal{A}$ und $\mathcal{B}$ von Vektoren, so dass die Menge von Matrizen $\mathcal{T} = \{I + a(\gamma b - a)^\top \mid a \in \mathcal{A}, b \in \mathcal{B}\}$ produktbeschränkt ist. Dies ist laut Satz 8 äquivalent zu der Existenz einer Matrixnorm $\| \cdot \|$, so dass $\|A\| \leq 1$ für alle $A \in \mathcal{T}$. Diese Norm wird nur in bestimmten Fällen explizit angegeben werden können. Im folgenden Satz ist das der Fall.

**Satz 9** Seien $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ und $b = (b_1, \ldots, b_n)^\top \in \mathbb{R}^n$ mit $a_i \in \{0, 1\}$, $\sum_{i=1}^n a_i \leq 1$, $b_i \geq 0$ und $\sum_{i=1}^n b_i \leq 1$. Dann ist $A = I + a(\gamma b - a)^\top$ für $\alpha \in [0, 1]$ eine substochastisch Matrix. Insbesondere gilt $\|A\|_\infty \leq 1$.

Beweis: betrachten wir die $i$-te Zeile von $A$. Falls $a_i = 0$ ist, dann entspricht diese Zeile der $i$-ten Zeile der Identiitätsmatrix, die Zeilensumme ist also gleich 1. Falls $a_i = 1$ ist, dann hat die $i$-te Zeile die Form $\alpha (\gamma b + (1 - \alpha)a)^\top$. Für $\alpha \in [0, 1)$ gilt für die $i$-te Zeilensumme $s$ die Abschätzung

$$0 \leq s = \alpha \sum_{i=1}^n a_i b_i + (1-\alpha) a_i \leq \alpha \gamma + (1-\alpha) = 1 - \alpha (1-\gamma) \leq 1$$

Es gilt sogar $s < 1$ falls $\alpha \neq 0$. Da die anderen Zeilen von $A$ denen der Identiitätsmatrix entsprechen, ist die Matrix $A$ substochastisch.

Mit Satz 9 erhalten wir eine ganze Klasse von Matrizen der Form $A = I + \alpha a(\gamma b - a)^\top$, deren Zeilensummennorm $\leq 1$ ist. Dies haben sie mit substochastischen Matrizen gemeinsam, allerdings haben die betrachteten Matrizen noch weitere Eigenschaften,
die auch bei inhomogenen Iterationen beschränkt lassen. Wir betrachten dazu die affine Abbildung

\[ T(x) = (I + \alpha a(\gamma b - a)) x + ra \]

und schreiben sie homogenisiert in Form einer Matrix

\[
\begin{pmatrix}
I + \alpha a(\gamma b - a) & ra \\
0 & 1
\end{pmatrix}
\]

Der nächste Satz wird zeigen, dass die Menge solcher Matrizen produktbeschränkt ist.

**Satz 10** Sei \( A \subset \mathbb{R}^n \) die Menge aller Vektoren \( a = (a_1, \ldots, a_n)^\top \), die die Eigenschaft \( a_i \in \{0, 1\}, \sum_{i=1}^n a_i \leq 1 \) haben. Ferner sei \( B \subset \mathbb{R}^n \) die Menge aller Vektoren \( b = (b_1, \ldots, b_n)^\top \) mit \( b_i \geq 0 \) und \( \sum_{i=1}^n b_i \leq 1 \). \( R > 0 \) bezeichne eine beliebige aber feste Schranke. Dann ist die Menge

\[
T = \left\{ \begin{pmatrix} I + \alpha a(\gamma b - a)^\top & ra \\ 0 & 1 \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} | a \in A, b \in B, \alpha \in [0, 1), |r| < R \right\}
\]

eine produktbeschränkte Menge von Matrizen.

Beweis: wir zeigen die Behauptung induktiv. Sei \( \tilde{C} \) ein Produkt von \( n \) Matrizen in \( T \), und \( \tilde{A} \) eine beliebige Matrix in \( T \)

\[
\tilde{C} = \begin{pmatrix} C & c \\ 0^\top & 1 \end{pmatrix}, \quad \tilde{A} = \begin{pmatrix} A & a \\ 0^\top & 1 \end{pmatrix}, \quad \text{und} \quad \tilde{A}\tilde{C} = \begin{pmatrix} AC & Ac + a \\ 0^\top & 1 \end{pmatrix}
\]

Wir setzen \( S = R \sum_{i=0}^\infty \theta^i \) mit später noch zu definierendem \( 0 \leq \theta < 1 \). Die Induktionsvoraussetzung lautet nun: \( \|C\|_\infty \leq 1 \) und \( c \|c\|_\infty \leq S \). Diese ist im Fall \( n = 1 \) nach Satz 9 erfüllt. Ebenfalls nach Satz 9 gilt \( \|A\|_\infty \leq 1 \) und wegen der Submultiplikativität von Matrixnormen ist \( \|AC\|_\infty \leq \|A\|_\infty \|C\|_\infty \leq 1 \).

Es bleibt zu zeigen, dass \( \|d\|_\infty \leq S \) mit \( d = (d_1, \ldots, d_n)^\top = Ac + a \) gilt. Aus dem Beweis von Satz 9 wissen wir, dass \( A \) nur in der \( i \)-ten Zeile, die wir mit \( s = (s_1, \ldots, s_n)^\top \) bezeichnen, von der Identitätsmatrix \( I \) abweicht. Ferner ist die \( i \)-te Zeilensumme von \( A \) gleich \( \theta \) := \( \sum_{j=1}^n s_j = 1 - \alpha(1 - \gamma) < 1 \). Damit gilt \( d_j = c_j \) für \( j \neq i \). Für \( d_i \) gilt

\[
d_i = s^\top c + a_i \leq \sum_{j=1}^n s_j = 1 - \alpha(1 - \gamma) + \theta \sum_{j=1}^n A_j = S + R = S
\]

Insgesamt erhalten wir \( \|d\|_\infty \leq S \).

\[ \Box \]

In den Sätzen 9 und 10 haben die Vektoren in der Menge \( A \) nur eine Stelle, die sich von 0 unterscheidet. Eine Verallgemeinerung davon wäre es, wenn man für \( \mathcal{A} \ni a = (a_1, \ldots, a_n)^\top \) auch \( a_i \in \{0, \kappa\} \), mit \( \kappa > 0 \) und \( \sum_{i=1}^n a_i = 1 \), zulässt. Dass solche
Matrizen einen Spektralradius \( \leq 1 \) haben ist eine Folgerung von Lemma 1. Es zeigt
nämlich, dass solche Vektoren \( a \) tangentiale Punkte der \( \| \cdot \|_1 \)-Einheitskugel sind, die
eine Obermenge der Menge \( \mathcal{B} \) ist. Folgendes Beispiel zeigt allerdings, dass das Produkt
von zwei solchen Matrizen einen Spektralradius haben kann, der größer als 1 ist. Es
existiert daher auch keine Matrixnorm, so dass alle solche Matrizen in Bezug auf
diese Norm \( \| \cdot \|_1 \) waren.

**Beispiel: 7** Betrachte

\[
a = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad c = \frac{1}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad b = \frac{1}{10} \begin{pmatrix} 1 \\ 9 \\ 0 \end{pmatrix}
\]

Setze \( A = (I + \alpha a(\gamma b - a)^\top) \) und \( C = (I + \alpha c(\gamma b - c)^\top) \) mit \( \alpha = 0.5 \) und \( \gamma = 0.99 \).
Man weist leicht nach, oder folgert es allgemein aus Lemma 1, dass \( \rho(A) \leq 1 \) und
\( \rho(C) \leq 1 \) gilt. Für das Produkt \( AC \) kann man es nicht behaupten, denn es gilt

\[
\sigma(AC) = \sigma \begin{pmatrix} 0.902169 & 0.107305 & -0.0122188 \\ -0.0730807 & 1.205055 & -0.137219 \\ 0.02475 & 0.09775 & 0.875 \end{pmatrix} = \{0.881635, 1, 1.100589\}
\]

wobei \( x = (0.416924, 0.816435, 0.399511)^\top \) der zu \( \rho(AC) \approx 1.100589 \)
korrespondierende Eigenvektor ist. Damit ist \( D_i \) für \( D = AC \) und \( i \to \infty \)
unbeschränkt. Es existiert daher auch keine Matrixnorm \( \| \cdot \| \) für die \( \| A \|, \| C \| \leq 1 \)
gelten würde.

Eine andere mögliche Verallgemeinerung von Satz 10 wäre es, wenn man für die
Menge \( \mathcal{B} \) die Menge aller Vektoren \( b = (b_1, \ldots, b_n)^\top \in \mathbb{R}^n \) nimmt, für die \( b_i \geq 0 \) und
\( b^\top b = \sum_{i=1}^n b_i^2 \leq 1 \) gilt. Damit wäre \( \mathcal{B} \) eine Teilmenge der \( \| \cdot \|_2 \)
Einheitskugel, und aus Lemma 2 folgt damit, dass der Spektralradius der beteiligten Matrizen trotz der
Erweiterung von \( \mathcal{B} \) immer noch kleiner oder gleich 1 ist. Folgendes Beispiel zeigt, dass
es für Produkte von solchen Matrizen nicht mehr der Fall zu sein braucht.

**Beispiel: 8** Betrachte

\[
a = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c = \begin{pmatrix} \sqrt{0.7} \\ \sqrt{0.3} \end{pmatrix}
\]

Setze \( A = (I + \alpha a(\gamma b - a)^\top) \) und \( C = (I + \alpha c(\gamma c - b)^\top) \) mit \( \alpha = 0.5 \) und \( \gamma = 0.99 \).
Aus Lemma 2 folgt, dass sowohl \( \rho(A) \leq 1 \) als auch \( \rho(C) \leq 1 \) gilt. Das Produkt \( AC \)
hat aber einen Spektralradius \( \rho(AC) \), der größer als 1 ist, denn

\[
\sigma(AC) = \sigma \begin{pmatrix} 1.04835 & 0.135561 \\ 0.495 & 0.5 \end{pmatrix} = \{1.15137, 0.396982\}
\]

und \( x = (0.796187, 0.605051)^\top \) der zu \( \rho(AC) \approx 1.15137 \) korrespondierende
Eigenvektor ist. Damit ist \( D_i \) für \( D = AC \) und \( i \to \infty \) unbeschränkt. Es existiert
daher auch keine Matrixnorm \( \| \cdot \| \) für die \( \| A \|, \| C \| \leq 1 \) gelten würde. \( \diamond \)
Eine dritte Möglichkeit der Verallgemeinerung besteht darin, in der Menge $B$ nur tangentielle Punkte der $\| \cdot \|_\infty$ Einheitskugel zulassen. Beliebige Punkte der $\| \cdot \|_\infty$ Einheitskugel kann man nicht zulassen, da die $\| \cdot \|_\infty$ Einheitskugel eine Obermenge der $\| \cdot \|_2$ Einheitskugel darstellt, und man damit auch schon mit Beispiel 8 ein Gegenbeispiel in der Hand hat. Das folgende Beispiel zeigt aber, dass auch diese potenzielle Verallgemeinerung von Satz 10 nicht funktioniert.

Beispiel: 9 Betrachte

\[
a = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}
\]

Setze $A = (I + a(a(b(c-a))^\top)$ und $C = (I + ab(\gamma c - b)^\top) \text{ mit } \alpha = 0.5$ und $\gamma = 0.99$. Erneut sind die Spektralradii $\rho(A) \leq 1$ und $\rho(C) \leq 1$ (vgl. Lemma 3), aber das Spektrum ihres Produkts ist gleich

\[
\sigma(AC) = \sigma\left( \begin{pmatrix} 1.24002 & 0.492525 \\ 0.495 & 0.995 \end{pmatrix} \right) = \{1.62625, 0.60878\}
\]

und damit ist auch der Spektralradius $\rho(AC) \approx 1.62625 > 1$, wobei $x = (0.786911, 0.617067)^\top$ ein korrespondierender Eigenvektor ist. Damit ist $D^i$ für $D = AC$ und $i \to \infty$ unbeschränkt. Es existiert daher auch keine Matrixnorm $\| \cdot \|$ für die $\| A \|, \| C \| \leq 1$ gelten würde.

Anhand der Beispiele haben wir gezeigt, dass sich Satz 10 nicht auf eine größere Menge von Matrizen erweitern lässt. Es wäre aber interessant zu untersuchen, ob es gänzlich andere Typen von schießen Projektionen gibt, für die eine zu Satz 10 äquivalente Aussage möglich wäre. Dies erscheint allerdings unwahrscheinlich, da wir mit den Lemmata 1, 2 und 3 bereits Klassen von möglichen Matrizen eingeschränkt haben, deren Spektralradius $\leq 1$ ist. Dies ist aber eine notwendige Bedingung, damit beliebige Produkte solcher Matrizen beschränkt bleiben. Satz 10 ist quasi ein Schnitmenge dieser in den Lemmata 1, 2 und 3 vorgestellten Matrizen, für die unendliche Produkte nicht divergieren. In Tabelle 3.4 haben diese Ergebnisse nochmals zusammengefasst.

3.5 Relation von schiefen Projektionen und speziellen Iterationsmatrizen

Im Abschnitt 3.4 haben wir unendliche Produkte von Matrizen

\[
\ldots A_k A_{k-1} \ldots A_1, \quad A_i \in \mathcal{T}
\]

betrachtet. Speziell wurden Matrizen von der Form $A = I + ab^\top$ behandelt, und in den Sätzen 9 und 10 hinreichende Kriterien an die Form der Vektoren $a$ und $b$ gestellt, damit unendliche Produkte dieser Matrizen beschränkt bleiben. In diesem Abschnitt werden wir ein notwendiges Kriterium vorstellen, ohne die Form der einzelnen Vektoren $a$ und $b$ weiter einzuschränken.
Tabelle 3.1: Übersicht über die Beschränktheit von Produkten und Potenzen verschiedener Matrixtypen.

<table>
<thead>
<tr>
<th>Eigenschaften von Matrizen der Form $\begin{pmatrix} I + \alpha a(\gamma b - a)^\top &amp; ra \ 0 &amp; \ldots &amp; 0 \ 1 \end{pmatrix}$</th>
<th>Beschränktheit von Potenzen</th>
<th>Beschränktheit von beliebigen Produkten</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_i \in {0, 1}, \sum_i a_i = 1$; $b_i \in [0, 1], \sum_i b_i \leq 1$</td>
<td>ja: Korollar 1, Lemma 1</td>
<td>ja: Satz 10</td>
</tr>
<tr>
<td>$a_i \in {0, \kappa}, \kappa &gt; 0, \sum_i a_i = 1$; $b_i \in [0, 1], \sum_i b_i \leq 1$</td>
<td>ja: Korollar 1, Lemma 1</td>
<td>nein: Beispiel 7</td>
</tr>
<tr>
<td>$\sum_i a_i^2 = 1$; $\sum_i b_i^2 \leq 1$</td>
<td>ja: Korollar 1, Lemma 2</td>
<td>nein: Beispiel 8</td>
</tr>
<tr>
<td>$a_i \in {0, 1}, \sum_i a_i = 1$; $\sum_i b_i^2 \leq 1$</td>
<td>ja: Korollar 1, Lemma 2</td>
<td>nein: Beispiel 8</td>
</tr>
<tr>
<td>$a_i \in {0, 1}$; $b_i \in {0, 1}$</td>
<td>ja: Korollar 1, Lemma 3</td>
<td>nein: Beispiel 9</td>
</tr>
</tbody>
</table>

Die Menge der Matrizen $\mathcal{T}$ wird im Folgenden als endlich angenommen. Der Begriff des Joint Spectral Radius $\rho(\mathcal{T})$ wird sich als nützlich erweisen. Dabei handelt sich um

$$
\rho(\mathcal{T}) = \limsup_{k \to \infty} \max_{\tau \in \mathcal{T}_k} \|\tau_k \tau_{k-1} \cdots \tau_1\|^{\frac{1}{k}}
$$

(3.24)

wobei $\mathcal{T}_k := (\mathcal{T} \cup \{I\})^k$ wieder alle endlichen Folgen der Länge $k$ von Matrizen in $\mathcal{T} \cup \{I\}$ bezeichnet [BN05]. Der Wert von $\rho(\mathcal{T})$ ist von der verwendeten Matrixnorm $\| \cdot \|$ unabhängig.

Die Definition von $\rho(\mathcal{T})$ erinnert an die Identität

$$
\rho(A) = \limsup_{k \to \infty} \|A^k\|^{\frac{1}{k}}
$$

(3.25)

für den Spektralradius einer Matrix (vgl. [HJ85]). Insbesondere ist $\rho(\mathcal{T}) \leq 1$ notwendig für die Beschränktheit von unendlichen Produkten von Matrizen in $\mathcal{T}$. Im Übrigen, es ist ein unentscheidbares Problem, ob $\rho(\mathcal{T}) \leq 1$ gilt [BT00]. Uns interessiert hier, ob es eine untere Schranke für $\rho(\mathcal{T})$ gibt, und wie diese aussieht. In [BN05, BN03] ist eine solche Schranke angegeben, die wir im folgenden Lemma wiedergeben.

\[\text{Aber nicht hinreichend, ähnlich wie } \rho(A) \leq 1 \text{ lediglich notwendig aber nicht hinreichend für die Beschränktheit von } A^k, k \to \infty \text{ ist.}\]
Lemma 9 Für eine endliche Menge von Matrizen $T = \{A_1, \ldots, A_m\}$ gilt

$$\rho(T) \geq \max_{\alpha_i \geq 0, \sum_i \alpha_i = 1} \rho \left( \sum_{i=1}^m \alpha_i A_i \right).$$

Beweis: Ein Beweis ist in [BN03] zu finden. Da diese Quelle nicht immer verfügbar zu sein scheint, wird hier der Beweis wiedergegeben. Für ein festes $k$ und $\alpha_i \geq 0$, $\sum_{i=1}^m \alpha_i = 1$ gilt

$$1 = \left( \sum_{i=1}^m \alpha_i \right)^k = \sum_{\sigma \in \{1, \ldots, m\}^k} \frac{\alpha_{\sigma_1} \alpha_{\sigma_2} \cdots \alpha_{\sigma_k}}{\alpha_0 \geq 0}$$

wobei $\sigma$ alle möglichen $k$ Tupel mit Einträgen in der Indexmenge $\{1, \ldots, m\}$ durchläuft. Die $\alpha_{\sigma}$ bilden damit erneut eine konvexe Kombination, und man erhält

$$\| (A_1 + \ldots + A_m)^k \| = \left\| \sum_{\sigma \in \{1, \ldots, m\}^k} \alpha_{\sigma} A_{\sigma_1} A_{\sigma_2} \cdots A_{\sigma_k} \right\|$$

$$\leq \sum_{\sigma \in \{1, \ldots, m\}^k} \alpha_{\sigma} \| A_{\sigma_1} A_{\sigma_2} \cdots A_{\sigma_k} \|$$

$$\leq \max_{\sigma \in \{1, \ldots, m\}^k} \| A_{\sigma_1} A_{\sigma_2} \cdots A_{\sigma_k} \|$$

Wenn man nun auf beiden Seiten den lim sup anwendet, erhält man zusammen mit (3.24) und (3.25) die Behauptung. \hfill \Box

Lemma 9 zeigt, dass der Joint Spectral Radius $\rho(T)$ größer oder gleich dem Maximum ist, welches der normale Spektralradius auf der konvexen Hülle der Matrizen $\{A_1, \ldots, A_m\}$ erreichen kann.

Wir werden nun das Ergebnis von Lemma 9 auf den Fall anwenden, in dem die Matrizen $A \in T$ die Form $I + \alpha \varphi_i (\varphi_j - \varphi_i)^\top$ mit $\varphi_1, \ldots, \varphi_n \subset \mathbb{R}^m$ haben. Wir schreiben die Vektoren $\varphi_i$ als Zeilen einer Matrix $\Phi$

$$\Phi = \begin{pmatrix} \varphi_1^\top \\ \vdots \\ \varphi_n^\top \end{pmatrix} = (\varphi_1 | \ldots | \varphi_n)^\top$$

(3.26)

und betrachten die Matrix $I - \alpha \Phi^\top D (I - \gamma P) \Phi$, die wir bereits im Abschnitt 3.3 verwendet haben. Dabei sind $D = \text{diag}(d_1, \ldots, d_n)$, $\sum_i d_i = 1$ eine nichtnegative Diagonalmatrix und $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ eine stochastische Matrix. Es gilt

$$DP \Phi = (\Phi^\top (DP)^\top)^\top = \begin{pmatrix} d_1 \sum_{k=1}^n p_{1k} \varphi_1 \\ \vdots \\ d_n \sum_{k=1}^n p_{nk} \varphi_n \end{pmatrix}^\top$$

Zusammen mit den allgemeinen Identitäten für Vektoren $\varphi_i, \psi_i, \psi \in \mathbb{R}^m$

$$(\varphi_1 | \ldots | \varphi_n)(\psi_1 | \ldots | \psi_n)^\top = \varphi_1 \psi_1^\top + \ldots + \varphi_n \psi_n^\top$$
KAPITEL 3. MATRIXITERATIONEN

und

\[ \psi \left( \sum_i \lambda_i \varphi_i \right)^\top = \left( \psi^\top \right)^\top \left( \sum_i \lambda_i \varphi_i \right)^\top = \left( \sum_i \lambda_i \varphi_i \psi_i^\top \right)^\top \]

erhält man

\[ \Phi^\top DP\Phi = \varphi_1 \left( d_1 \sum_{j=1}^n p_{1j} \varphi_j \right)^\top + \ldots + \varphi_n \left( d_n \sum_{j=1}^n p_{nj} \varphi_j \right)^\top \]

\[ = d_1 \sum_{j=1}^n p_{1j} \varphi_j^\top + \ldots + d_n \sum_{j=1}^n p_{nj} \varphi_n \varphi_j^\top \]

Insgesamt folgt

\[ I - \alpha \Phi^\top D(I - \gamma P)\Phi = I - \alpha \Phi^\top D\Phi + \alpha \gamma \Phi^\top DP\Phi \]

\[ = \left( \sum_{i=1}^n d_i \right) I - \alpha \sum_{i=1}^n d_i \varphi_i^\top \varphi_i^\top + \alpha \gamma \sum_{i=1}^n d_i \sum_{j=1}^n p_{ij} \varphi_i^\top \varphi_j^\top \]

\[ = \sum_{i=1}^n d_i \left( \sum_{j=1}^n p_{ij} I - \alpha \sum_{j=1}^n p_{ij} \varphi_i^\top \varphi_i^\top + \alpha \gamma \sum_{j=1}^n p_{ij} \varphi_i^\top \varphi_j^\top \right) \]

\[ = \sum_{i=1}^n d_i \sum_{j=1}^n p_{ij} \left( I + \alpha \varphi_i (\gamma \varphi_j - \varphi_i)^\top \right) \]

(3.27)

Da nach Voraussetzung \( d_i \geq 0 \), \( \sum_i d_i = 1 \) gilt, und \( P = (p_{ij}) \) eine stochastische Matrix ist, ist \( I - \alpha \Phi^\top D(I - \gamma P)\Phi \) eine konvexe Kombination der einzelnen Matrizen

\[ I + \alpha \varphi_i (\gamma \varphi_j - \varphi_i)^\top, \quad i, j = 1, \ldots, n \]

Man zeigt leicht anhand der Definition (3.26), dass ebenfalls für \( r := (r_1, \ldots, r_n)^\top \)

\[ \Phi^\top Dr = \sum_{i=1}^n d_i r_i \varphi_i \]

gilt. Damit ist auch die Matrix

\[ H_{\Phi, P, D} := \begin{pmatrix} I - \alpha \Phi^\top D(I - \gamma P)\Phi & \alpha \Phi^\top Dr \\ 0 & \ldots & 0 & 1 \end{pmatrix} \]

(3.28)

als konvexe Kombination von Matrizen der Form

\[ \begin{pmatrix} I + \alpha \varphi_i (\gamma \varphi_j - \varphi_i)^\top & \alpha r_i \varphi_i \\ 0 & \ldots & 0 & 1 \end{pmatrix} \]

(3.29)

darstellbar.

Auf diese Tatsache kann Lemma 9 angewendet werden, und man erhält
Satz 11  Gegeben sei eine Menge von Matrizen

\[ T = \left\{ \begin{pmatrix} I + \alpha \varphi_i (\gamma \varphi_j - \varphi_i)^\top & \alpha r_i \varphi_i \\ 0 & 0 \end{pmatrix} \left| \begin{array}{c} \varphi_i \in \mathbb{R}^m, r_i \in \mathbb{R} \end{array} \right. \right\} \quad (3.30) \]

Ferner sei \( \Phi \) gegeben durch (3.26). Es gilt dann

\[ \rho(T) \geq \rho(H_{\Phi,P,D}) \]

wobei die Matrix \( H_{\Phi,P,D} \) durch (3.28) definiert ist. Insbesondere ist \( \rho(H_{\Phi,P,D}) \leq 1 \) notwendig für die Beschränktheit der Matrizen in \( T \). Existiert andererseits eine nichtnegative Diagonalmatrix \( D = \text{diag}(d_1,\ldots,d_n) \), mit \( \sum d_i = 1 \) und eine stochastische Matrix \( P \), so dass \( \rho(H_{\Phi,P,D}) > 1 \) gilt, so gibt es eine Sequenz von Matrizen in \( T \), deren Produkt divergiert.

Beweis: Die Ungleichung (3.31) folgt direkt aus Lemma 9 und Gleichung (3.27), also aus der Tatsache, dass sich die Matrix \( H_{\Phi,P,D} \) als eine konvexe Kombination der Matrizen in \( T \) schreiben lässt. Da \( \rho(T) \leq 1 \) notwendig für die Beschränktheit von unendlichen Produkten von Matrizen in \( T \) ist, ist nach (3.31) auch \( \rho(H_{\Phi,P,D}) \leq 1 \) dafür notwendig.

Satz 11 zeigt, dass die Analyse der für Beschränktheit geeigneter Matrizen \( \Phi \), die in Abschnitt 3.3 untersucht wurden, verwendet werden kann, um potentielle Kandidaten für Klassen von \( \varphi_i \) zu bestimmen, so dass unendliche Produkte von Matrizen der Form (3.29) nicht divergieren. Für die Bedingung \( \rho(H_{\Phi,P,D}) \leq 1 \) ist nämlich insbesondere notwendig, dass die Matrix \( \Phi^\top D (I - \gamma P) \Phi \) semi positiv stabil ist.

Es bleibt noch die Frage zu klären, was passieren kann, falls zwar \( \rho(H_{\Phi,P,D}) = 1 \) gilt, aber die Potenzen \( H_{\Phi,P,D}^k \) von \( H_{\Phi,P,D} \) divergieren. Nehmen wir an, dass dann alle Folgen von Matrizen aus \( T \) in (3.30) beschränkt bleiben würden. Laut Satz 8 müsste dann eine Matrixnorm \( \| \cdot \| \) existieren, so dass für alle \( A_i \in T \) die Ungleichung \( \| A_i \| \leq 1 \) gelten würde. Nun ist aber \( H_{\Phi,P,D} \) eine konvexe Kombination der Matrizen \( A_i \). Damit würde folgen

\[ \| H_{\Phi,P,D} \| = \left\| \sum_{i=1}^n \alpha_i A_i \right\| \leq \sum_{i=1}^n \alpha_i \| A_i \| \leq \sum_{i=1}^n \alpha_i = 1 \]

Die Abschätzung \( \| H_{\Phi,P,D} \| \leq 1 \) ist aber laut Satz 3 ein Widerspruch zur Divergenz der Potenzen \( H_{\Phi,P,D}^k \) von \( H_{\Phi,P,D} \). Dieser Widerspruch zeigt, dass es eine Folge von Matrizen aus \( T \) in (3.30) gibt, deren unendliches Produkt divergieren muss. Diese wichtige Erkenntnis fassen wir im folgenden Satz zusammen.
KAPITEL 3. MATRIXITERATIONEN

Satz 12 Gegeben seien $T$ und $H_{\Phi, P, D}$ wie im Satz 11. Dann folgt aus der Divergenz von Potenzen $H_{\Phi, P, D}^k$ der Matrix $H_{\Phi, P, D}$ die Divergenz mindestens eines unendlichen Produktes von Matrizen in $T$. Formal kann man dies durch

\[ \neg \text{Bounded}(H_{\Phi, P, D}, 0) \iff \neg \text{Bounded}(I - \alpha \Phi^\top D(I - \gamma P)\Phi, \alpha \Phi^\top Dr) \]

ausdrücken.

Beweis: siehe obige Herleitung.


3.6 Lage von Eigenwerten

In diesem Abschnitt beschäftigen wir uns mit der Lage von Eigenwerten der Matrizen

\[ A_{TD}(\alpha, P) := I - \alpha(I - \gamma P) \]
\[ A_{RG}(\alpha, P) := I - \alpha(I - \gamma P)^\top (I - \gamma P) \]

wobei $P$ eine stochastische Matrix ist. Die Bezeichnungen $A_{TD}$ und $A_{RG}$ deuten auf eine Verbindung zum synchronen TD[0] bzw. zum synchronen RG Verfahren hin, die wir im Abschnitt 4.6 untersuchen werden. Uns interessieren insbesondere die minimalen Spektralradii

\[ \rho_{TD}^*(P) := \min_{\alpha \geq 0} \rho(A_{TD}(\alpha, P)) = \min_{\alpha \geq 0} \rho(I - \alpha(I - \gamma P)) \]
\[ \rho_{RG}^*(P) := \min_{\alpha \geq 0} \rho(A_{RG}(\alpha, P)) = \min_{\alpha \geq 0} \rho(I - \alpha(I - \gamma P)^\top (I - \gamma P)) \]

Eine entscheidende Frage ist dabei, ob $\rho_{TD}^*(P) \leq \rho_{RG}^*(P)$ gilt, und wenn ja, dann für welche Matrizen $P$. Der Parameter $\alpha$ wird in Anlehnung an die spätere Anwendung im Reinforcement Lernen als Lernrate bezeichnet. Eine Lernrate $\alpha$ für die das Minimum angenommen wird, wird als optimale Lernrate bezeichnet.

Nach Lemma 5 ist die Matrix $I - \gamma P$ eine $M$-Matrix, ihre Eigenwerte haben also einen positiven Realteil. Wir betrachten daher zuerst Matrizen von der Form $C = I - \alpha M$, wobei $M$ eine $M$-Matrix ist. Folgendes Lemma, das aus der Literatur bekannt ist, beschränkt $M$ sogar weiter, und erlaubt nur reelle Eigenwerte.
3.6. LAGE VON EIGENWERTEN

Abbildung 3.2: Bei der Anwendung der optimalen Lernrate wir der ursprüngliche Kreis zum Kreis $K_{\alpha^*}$, der konzentrisch in dem gestrichelten Kreis $K_1$ enthalten ist.

**Lemma 10** Die Matrix $M \in \mathbb{R}^{n \times n}$ habe nur positive Eigenwerte, wobei $\lambda_1$ den größten und $\lambda_n$ den kleinsten Eigenwert von $M$ bezeichnet. Dann gilt

$$\inf_{\alpha \geq 0} \rho(I - \alpha M) = \rho \left( I - \frac{2}{\lambda_1 + \lambda_n} M \right) = 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n}$$

und $\alpha^* = 2/(\lambda_1 + \lambda_n)$ ist die optimale Lernrate für $I - \alpha M$.

Beweis: siehe [Mei99, Sch03]

Die in Lemma 10 angegebene optimale Lernrate kann man sich anschaulich vorstellen. Die Eigenwerte $\lambda_1$ und $\lambda_n$ definieren eine Kreis in der komplexen Ebene, dessen Mittelpunkt in der Mitte zwischen $\lambda_1$ und $\lambda_n$ auf der positiven reellen Achse liegt. Dieser Mittelpunkt ist $\frac{1}{2}(\lambda_1 + \lambda_n)$ und wird per Multiplikation mit einem $\alpha$ verschoben, dass er auf $1 = (1,0) \in \mathbb{C}$ zeigt. Also ist $\alpha^*$ gerade das Reziproke des Mittelpunktes.

Der so entstandene Kreis $K_{\alpha^*}$ liegt konzentrisch im umschließenden Kreis $K_1$ mit Radius 1 und Mittelpunkt $(1,0) \in \mathbb{R}^2$, der in Abbildung 3.2 gestrichelt dargestellt ist. Eine Lernrate $\alpha \neq \alpha^*$ würde den Kreis $K_{\alpha}$ nicht konzentrisch zum Kreis $K_1$ ausrichten, und damit letztendlich den Spektralradius von $I - \alpha M$ vergrößern.

Obige anschauliche Herleitung zeigt aber auch, dass es nicht unbedingt erforderlich ist, dass alle Eigenwerte von $M$ reell sind. Es reicht, dass der betragsmäßig größte und kleinste Eigenwert reell sind, und die restlichen Eigenwerte innerhalb des von diesen beiden Eigenwerten gebildetem Kreis liegen. Dies wird im folgenden Lemma zusammengefasst

**Lemma 11** Sei $M \in \mathbb{R}^{n \times n}$ eine positiv stabile Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$. Der betragsmäßig größte und kleinste Eigenwert sind reell, d.h. es gilt

$$\lambda_1 = |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}| \geq |\lambda_n| = \lambda_n > 0$$
Falls für die übrigen Eigenwerte \( \lambda_i \in \{ \lambda_2 \ldots \lambda_{n-1} \} \)
\[
\left| \lambda_i - \frac{1}{2}(\lambda_1 + \lambda_n) \right| \leq \frac{1}{2}(\lambda_1 - \lambda_n)
\]
gilt, dann ist
\[
\inf_{\alpha \geq 0} \rho(I - \alpha M) = \rho \left( I - \frac{2}{\lambda_1 + \lambda_n} M \right) = 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n} \tag{3.33}
\]
und \( \alpha^* = 2/(\lambda_1 + \lambda_n) \) ist die optimale Lernrate für \( I - \alpha M \).

Beweis: laut Lemma 10 gilt für \( i \in \{1, n\} \) die Ungleichung \( |1 - \alpha^* \lambda_i| \leq 1 - 2\lambda_n/(\lambda_1 + \lambda_n) \). Es bleibt dies für \( i \in \{2, \ldots, n-1\} \) zeigen. Laut (3.33) gilt aber
\[
|1 - \alpha^* \lambda_i| = \frac{\alpha^*}{\alpha^*} |1 - \alpha^* \lambda_i| = \alpha^* |\lambda_i - \alpha^*| = \alpha^* \left| \lambda_i - \frac{1}{2}(\lambda_1 + \lambda_n) \right|
\]
\[
\leq \alpha^* \frac{1}{2}(\lambda_1 - \lambda_n) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n}
\]
wobei bei der Ungleichung die Voraussetzung benutzt wurde. \( \Diamond \)

Hat \( M \) die spezielle Form \( I - \gamma P \), wobei \( P \) eine stochastische Matrix ist, so ist der betragsmäßig kleinste Eigenwert immer reell. Es liegt daran, dass jede stochastische Matrix \( P \) den Eigenwert 1 hat, und damit ist \( 1 - \gamma \) der betragsmäßig kleinste Eigenwert von \( I - \gamma P \). Hat ferner die Matrix \( P \) noch \(-1\) als Eigenwert, dann ist \( 1 + \gamma \) der betragsmäßig größte Eigenwert, und auf Grund von \( \rho(P) = 1 \) liegen alle anderen (u.U. auch komplexwertigen) Eigenwerte innerhalb des von \( 1 - \gamma \) und \( 1 + \gamma \) gebildeten Kreises. Ein Beispiel für eine solche Matrix ist
\[
\begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0
\end{pmatrix}
\]
mit den Eigenwerten \( \{1, i, -1, -i\} \). Damit sehen wir, dass die in Lemma 10 eingeführte Klasse von Matrizen echt größer als die von Matrizen mit reellen Eigenwerten ist. Für diese Klasse von Matrizen ist es möglich Aussagen über die optimale Lernrate zu machen. Bevor wir dazu kommen benötigen wir noch das folgende Lemma.

**Lemma 12** Betrachte eine reguläre Matrix \( A \in \mathbb{C}^{n \times n} \). A habe die Singulärwerte \( \sigma_1 \geq \ldots \geq \sigma_n \) und die Eigenwerte \( \{\lambda_1, \ldots, \lambda_n\} \) mit der Anordnung \( |\lambda_1| \geq \ldots \geq |\lambda_n| > 0 \). Dann gilt
\[
\frac{|\lambda_1|}{|\lambda_n|} \leq \frac{\sigma_1}{\sigma_n}.
\]

Beweis: Der Betrag des größten Eigenwertes einer Matrix \( A \) gleicht ihrem Spektralradius \( \rho(A) = \lambda_1 \), und man sieht leicht, dass auch \( \rho(A^{-1}) = \frac{1}{|\lambda_1|} \) gilt. Da
\[
\rho(A) \leq \|A\| \tag{3.34}
\]
3.6. LAGE VON EIGENWERTEN

Für jede Matrixnorm \( \| \cdot \| \) gilt (vgl. [HJ85]), ist dies insbesondere für die Spektralnorm \( \| A \|_2 = \max_{\| x \|_2 = 1} \| Ax \|_2 \) der Fall. Die Spektralnorm ist auch gleich der Wurzel des größten Singularwertes \( \sigma_1 \) von \( A \), welcher wiederum der Wurzel des größten Eigenwertes von \( A^*A \) bzw. von \( AA^* \) gleich (vgl. [HJ85])

\[
\| A \|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \sigma_1
\]

(3.35)

Das alles ergibt insgesamt

\[
\frac{\lambda_1(A)}{\lambda_n(A)} = \rho(A)\rho(A^{-1})
\]

(3.34)

\[
\leq \| A \|_2 \| A^{-1} \|_2
\]

(3.35)

\[
= \sqrt{\rho(A^*A)}\sqrt{\rho((A^{-1})^*A^{-1})}
\]

(3.35)

\[
= \sqrt{\rho(A^*A)}\sqrt{\rho((A^*A)^{-1})} = \frac{\sigma_1(A)}{\sigma_n(A)}.
\]

\[\diamond\]

Nun sind wir bereit für den angekündigten Satz über die optimalen Lernraten und Spektralradii der Matrizen \( A_{TD}(\alpha, P) \) und \( A_{RG}(\alpha, P) \).

**Satz 13** Gegeben seien die Matrizen \( A_{TD}(\alpha, P) \) und \( A_{RG}(\alpha, P) \) wie in (3.32). Ferner erfülle \( M = I - \gamma P \) die Voraussetzungen von Lemma 11, dann gilt

\[
\rho_{TD}^*(P) \leq \rho_{RG}^*(P)
\]

und sogar

\[
\rho_{TD}^*(P) < \rho_{RG}^*(P)
\]

falls \( A_{RG}(\alpha, P) \) mindestens 2 unterschiedliche Eigenwerte hat.

**Beweis:** Mit der Notation \( M = I - \gamma P \) erhalten wir

\[
A_{TD}(\alpha, P) = I - \alpha M, \quad (3.36)
\]

\[
A_{RG}(\alpha, P) = -I - \alpha M^\top M. \quad (3.37)
\]

Wir bezeichnen die Eigenwerte von \( M \) mit dem größten bzw. kleinsten Betrag mit \( \lambda_1 := \lambda_1(M) \) bzw. \( \lambda_n := \lambda_n(M) \). Aus Lemma 5 ist bekannt, dass \( M \) eine \( M \)-Matrix ist, und daher nur Eigenwerte mit positiven Realteil hat. Insbesondere ist \( \lambda_n \neq 0 \). Da \( M \) die Voraussetzungen von Lemma 11 erfüllt, gilt für den minimalen Spektralradius \( \rho_{TD}^*(P) \)

\[
\rho_{TD}^*(P) = \inf_{\alpha \geq 0} \rho(I - \alpha M) = 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n} = 1 - 2 \left( 1 + \frac{\lambda_1}{\lambda_n} \right)^{-1}
\]

Da \( M \) als \( M \)-Matrix regulär ist, ist \( M^\top M \) eine positiv definite Matrix, insbesondere sind die Eigenwerte von \( M^\top M \) positiv. Damit erfüllt \( M^\top M \) die Voraussetzungen von Lemma 10 und Lemma 11. Daher gilt

\[
\rho_{RG}^*(P) = \inf_{\alpha \geq 0} \rho(I - \alpha M^\top M) = 1 - \frac{2\mu_n}{\mu_1 + \mu_n} = 1 - 2 \left( 1 + \frac{\mu_1}{\mu_n} \right)^{-1}
\]
wobei $\mu_1$ bzw. $\mu_n$ den größten bzw. kleinsten Eigenwert von $M^\top M$ bezeichnen. Um $\rho^*_TD(P) \leq \rho^*_RG(P)$ zu zeigen, genügt es daher die Ungleichung
\[
1 - 2 \left( 1 + \frac{\lambda_1}{\lambda_n} \right)^{-1} \leq 1 - 2 \left( 1 + \frac{\mu_1}{\mu_N} \right)^{-1}
\]
zu beweisen. Dies ist aber äquivalent zu der Ungleichung $\lambda_1/\lambda_n \leq \mu_1/\mu_n$. Die Eigenwerte $\mu_1$ und $\mu_n$ von $M^\top M$ entsprechen den quadrierten Singularwerten von $M$ (vgl. [HJ85]), daher muss nur noch $\lambda_1/\lambda_n \leq \sigma_1^2/\sigma_n^2$ gezeigt werden. Lemma 12 garantiert, dass $\lambda_1/\lambda_n = |\lambda_1|/|\lambda_n| \leq \sigma_1/\sigma_n$ gilt. Per Definition gilt $\sigma_1/\sigma_n \geq 1$, und es folgt nun
\[
\lambda_1/\lambda_n \leq \sigma_1/\sigma_n \leq (\sigma_1/\sigma_n)^2
\]
Damit ist $\rho^*_TD(P) \leq \rho^*_RG(P)$ gezeigt, falls $M$ die Voraussetzungen von Lemma 11 erfüllt. Hat $M^\top M$ mindestens 2 unterschiedliche Eigenwerte, dann ist $\sigma_1 > \sigma_n$ und folglich auch $\sigma_1/\sigma_n > 1$. Hieraus folgt aber $\sigma_1/\sigma_n < (\sigma_1/\sigma_n)^2$, so dass man in (*) das $\leq$ Zeichen durch $<$ Zeichen ersetzen kann. Es gilt daher $\rho^*_TD(P) < \rho^*_RG(P)$, falls $M^\top M$ (und damit auch $A_{RG}(\alpha, P)$ für $\alpha \neq 0$) mindestens 2 unterschiedliche Eigenwerte hat.

In Kapitel 4 wird Satz 13 verwendet, um zu zeigen, dass das synchrone TD[0]-Verfahren in manchen Fällen nachweislich schneller als das synchrone residuelle Verfahren konvergiert. Die Klasse der Matrizen für die dieses gilt ist mindestens so groß wie die Klasse der $M$-Matrizen in Satz 13.

Es bleibt eine offene Frage, ob die Klasse von stochastischen Matrizen aus Satz 13 die größt mögliche Klasse darstellt, für die $\rho^*_TD(P) \leq \rho^*_RG(P)$ gilt. Wir werden jetzt empirische Untersuchungen vorstellen, die dafür sprechen, dass dies sogar für alle stochastischen Matrizen gelten könnte.

Als erstes kann man versuchen, die Aussage von Satz 13 so zu verallgemeinern, dass auch dann die Lernrate $2/(|\lambda_1| + |\lambda_n|)$ verwendet wird, wenn die Voraussetzungen für $P$ aus Satz 13 nicht gegeben sind. Dabei bezeichnen $\lambda_1$ bzw. $\lambda_n$ die Eigenwerte von $I - \gamma P$ mit dem größten bzw. kleinsten Betrag. Es soll also hierbei keine Rolle spielen, dass $\lambda_1$ nicht unbedingt reell sein wird, oder dass die anderen Eigenwerte die Bedingungen von Satz 13 bzw. Lemma 11 nicht unbedingt erfüllen.

Den Spektralradius der korrespondierenden Matrix bezeichnen wir mit $\hat{\rho}_{TD}(P)$
\[
\hat{\rho}_{TD}(P) = \rho \left( I - \frac{2}{|\lambda_1| + |\lambda_n|} (I - \gamma P) \right)
\]
Nun kann es aber passieren, dass $\hat{\rho}_{TD}(P) > 1$ ist. Dies kann man anhand der folgenden Matrix
\[
P = \begin{pmatrix}
0.2 & 0.4 & 0.4 \\
0.3 & 0.6 & 0.1 \\
0.1 & 0.7 & 0.2
\end{pmatrix}, \quad M = I - 0.98 * P = \begin{pmatrix}
0.804 & -0.392 & -0.392 \\
-0.294 & 0.412 & -0.098 \\
-0.098 & -0.686 & 0.804
\end{pmatrix}
\]

Der betragsmäßig kleinste Eigenwert $\lambda_n$ von $I - \gamma P$ ist immer reell, da 1 immer ein Eigenwert von $P$ ist.
3.6. LAGE VON EIGENWERTEN

Tabelle 3.2: Übersicht für dreidimensionale Matrizen $P$ über die Lage der Spektralradii $\tilde{\rho}_{TD}(P)$ und $\rho^*_{RG}(P)$.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>Anzahl von $P$ mit $\rho^*_{RG}(P) \leq \gamma$</th>
<th>Anzahl von $P$ mit $\rho^*<em>{RG}(P) &lt; \tilde{\rho}</em>{TD}(P)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>absolut</td>
<td>prozentual</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0 %</td>
</tr>
<tr>
<td>0.95</td>
<td>27</td>
<td>0 %</td>
</tr>
<tr>
<td>0.9</td>
<td>123</td>
<td>0.04 %</td>
</tr>
<tr>
<td>0.8</td>
<td>553</td>
<td>0.2 %</td>
</tr>
<tr>
<td>0.7</td>
<td>1306</td>
<td>0.4 %</td>
</tr>
<tr>
<td>0.6</td>
<td>2460</td>
<td>0.8 %</td>
</tr>
<tr>
<td>0.5</td>
<td>4197</td>
<td>1.5 %</td>
</tr>
<tr>
<td>0.4</td>
<td>6427</td>
<td>2.2 %</td>
</tr>
<tr>
<td>0.3</td>
<td>9534</td>
<td>3.3 %</td>
</tr>
<tr>
<td>0.2</td>
<td>13335</td>
<td>4.6 %</td>
</tr>
<tr>
<td>0.1</td>
<td>18343</td>
<td>6.4 %</td>
</tr>
</tbody>
</table>

sehen. Es gilt $\sigma(M) = \{0.02, 1 \pm 0.219135i\}$, und damit ist $\tilde{\rho}_{TD}(P) \approx 1.00785 > 1$. Damit ist auch $\tilde{\rho}_{TD}(P) > \rho^*_{RG}(P)$, denn $\rho^*_{RG}(P)$ ist für eine geeignete Lernrate $\alpha$ immer kleiner als 1. Hier speziell hat die Matrix $M^TM$ die Eigenwerte $\{0.000321,105641,24019\}$, woraus sich für den optimalen Spektralradius $\rho^*_{RG}(P) = 0.999483 > 0.98 = \gamma$ ergibt. Da mit der Lernrate $\alpha = 1$ immer $\rho^*_{TD}(P) \leq \gamma$ erreicht werden kann, ist durch dieses Beispiel die Eigenschaft (3.32) nicht verletzt.

Aus dem obigen Beispiel resultiert eine Überlegung, entweder die Lernrate $\alpha = 2/(|\lambda_1| + |\lambda_n|)$ oder $\alpha = 1$ zu wählen. Man erhält damit eine bessere Abschätzung von $\rho^*_{TD}(P)$ nach oben

$$\tilde{\rho}_{TD}(P) := \min(\tilde{\rho}_{TD}(P), \gamma) \geq \rho^*_{TD}(P) \quad (3.38)$$

Wir werden nun empirisch zeigen, dass in allen untersuchten Fällen

$$\tilde{\rho}_{TD}(P) \leq \rho^*_{RG}(P)$$

gilt. Als erstes werden alle dreidimensionalen stochastischen Matrizen mit Einträgen aus $\{i/10 \mid i = 0, \ldots, 10\}$ untersucht. Es existieren genau $66^3 = 287496$ dieser Matrizen.

In Tabelle 3.6 wurden die Ergebnisse zusammengefasst. Es zeigt sich, dass für übliche Diskontierungsfaktoren $\gamma \geq 0.6$ der Anteil der Matrizen $P$ für die $\rho^*_{RG}(P) < \gamma$ ausfällt, unter einem Prozent liegen. Wir werden später sehen, dass mit steigender Dimension dieser Anteil sinkt. Aber selbst für die Matrizen, für die $\rho^*_{RG}(P) < \gamma$
gilt, erweist sich $\hat{\rho}_{TD}(P)$ kleiner als $\rho^*_{RG}(P)$. Es wurde somit keine einzige Matrix gefunden, für die $\rho^*_{RG}(P) < \hat{\rho}_{TD}(P)$ gelten würde.

Dies bestätigts sich bei weiteren empirischen Untersuchungen. So wurden 100000 fünf bzw. zehn dimensionale stochastische Matrizen $P$ zufällig erzeugt. Für rein stochastische Matrizen findet sich für $\gamma \in [0.2, 0.99]$ keine Matrix, für die $\rho^*_{RG}(P) \leq \gamma$ gelten würde. Erst für $\gamma = 0.1$ wurden bei 5 dimensionalsten Matrizen 8 Beispiele gefunden, für die $\rho^*_{RG}(P) \leq \gamma$ war. Allerdings war für alle diese Matrizen $\tilde{\rho}_{TD}(P) \leq \rho^*_{RG}(P)$ welches unsere Vermutung $\tilde{\rho}_{TD}(P) \leq \rho^*_{RG}(P)$ stützt. Bei stochastischen Matrizen $P$, die Einträge aus $\{0, 1\}$ haben, und daher ein deterministisches System beschreiben, liegt der Anteil der Matrizen mit $\rho^*_{RG}(P) \leq \gamma$ generell höher. So findet man für fünfdimensionale Matrizen $P$ und $\gamma = 0.9$ ca. 41 Matrizen, für die $\rho^*_{RG}(P) \leq \gamma$ gilt. Allerdings wiederholen sich auch hier die Erfahrungen aus dem rein stochastischen Fall, und es gilt für alle Matrizen $\tilde{\rho}_{TD}(P) \leq \rho^*_{RG}(P)$.

Im zehndimensionalen Fall finden sich unter den zufällig erzeugten Matrizen keine Matrizen $P$, für die $\rho^*_{RG}(P) \leq \gamma$ gelten würde.

Diese Untersuchungen untermauern die Vermutung, dass die Ungleichung (3.32) für alle stochastischen Matrizen $P$ gilt. Ferner scheint die Anzahl der Matrizen $P$, für die $\rho^*_{RG}(P) \leq \gamma$ gilt, mit wachsender Dimension zu sinken. Für die meisten Fälle, für die also $\rho^*_{RG}(P) > \gamma$ gilt ist die triviale Lernrate $\alpha = 1$ bereits ausreichend, um $\rho^*_{TD}(P) < \rho^*_{RG}(P)$ zu garantieren.

### 3.7 Zusammenfassung

Dieses Kapitel diente der Vorbereitung von theoretischen Resultaten, die in den Kapiteln 4 und 5 auf Verfahren des approximativen Reinforcement Lernens angewendet werden. In Kapitel 4 werden wir sehen, dass sich das approximative TD\[0\]-Verfahren als Folge von Anwendungen von speziellen affinen Abbildungen studieren lässt. In diesem Kapitel haben wir solche Folge von affinen Abbildungen untersucht. Insgesamt wurden folgende Resultate erzielt:

- Für eine wiederholte Anwendung einer speziellen affinen Abbildung der Form (3.2) wurde das Grenzverhalten durch Satz 1 bestimmt. Diese Ergebnis entspricht dem in [MS02] vorab veröffentlichten Kriterium. Durch eine Homogenisierung der Iteration, wurde allerdings der Beweis vereinfacht.

- In den Lemmata 1, 2 und 3 haben wir Resultate über zulässige Vektoren bei der wiederholten Anwendung einer affinen Abbildung der Form (3.5) untersucht. Diese Ergebnisse sind eine Verallgemeinerung der Untersuchungen in [Sch03] über sichere Startzustände.

- Satz 4 hat zum ersten Mal alle Grenzverhalten einer allgemeinen Iteration (3.6) mit einer beliebigen affinen Abbildung geklärt. Der Beweis stützt sich auf Ergebnisse aus [Lud85, Ela99], die lediglich homogene Systeme behandeln. Da hinreichende und notwendige Kriterien für alle möglichen Grenzverhalten angegeben werden, stellt Satz 4 eine Verallgemeinerung eines der Hauptresultate aus [Sch03] dar, wo lediglich hinreichende Bedingungen für eine spezielle Form
3.7. ZUSAMMENFASSUNG

der Iterationsmatrix angegeben werden. Dieses Ergebnis wurde vorab in [MS04] veröffentlicht.

- Es wurden spezielle Iterationsmatrizen der Form (3.20) untersucht. Insbesondere wurde systematisch der Frage nach der Erhaltung der positiven Stabilität der Matrix \( D(I - \gamma P) \) bei Transformation mit einer Approximatorenmatrix \( \Phi \) nachgegangen. Dabei wurde konsequent die Tatsache ausgenutzt, dass die Matrix \( D(I - \gamma P) \) eine so genannte \( M \)-Matrix ist. Neue Ergebnisse sind hier Satz 6 und Korollar 2, die die positive Stabilität für eine Klasse von stochastischen Matrizen \( P \) zusichern, und zwar für beliebige Approximatorenmatrizen \( \Phi \). Ebenfalls neu sind die Sätze 5 und 7, die Klassen von Approximatorenmatrizen \( \Phi \) vorstellen, die die positive Stabilität der Matrix \( D(I - \gamma P) \) erhalten. Ferner werden viele andere Matrixtypen untersucht, die allerdings die positive Stabilität nicht erhalten. Eine Übersicht dieser Ergebnisse findet sich wegen der Bedeutung für das synchrone TD[0]-Verfahren in Tabelle 4.3 im Kapitel 4.


- Satz 12 ist ein zentrales Ergebnis dieser Arbeit. Es zeigt, dass die Beschränktheit einer bestimmten Matrixiteration zwingend notwendig für die Beschränktheit von Folgen von speziellen affinen Abbildungen ist. Dieser Satz ist ein Bindeglied zwischen dem synchronen und asynchronen TD[0] Verfahren, die in Kapitel 4 untersucht werden.


Kapitel 4

Konvergenz von approximativem Reinforcement Lernen

Es ist ein bekanntes Phänomen, dass Verfahren des Reinforcement Lernens (RL) zusammen mit Funktionsapproximation divergieren können [BM95, Bai95]. In diesem Kapitel untersuchen wir vornehmlich das approximative TD[0] Verfahren, gehen aber am Ende des Kapitels auch auf die sich ergebenden Implikationen für das Verfahren der Wertiteration und für das Q-Lernen ein. Das TD[0] Verfahren studieren wir in zwei Formen, die wir mit asynchron und synchron bezeichnen. Die asynchrone Version entspricht dem normalen Vorgehen in der Praxis, hier werden verschiedene Zustandsübergänge \( g \circ h \) nacheinander betrachtet, und die zu approximierende Wertfunktion wird nach jedem Übergang neu angepasst. Wenn wir im weiteren Verlauf vom TD[0] Verfahren sprechen, so meinen wir damit immer die asynchrone Form des TD[0] Verfahrens. Nur der synchrone Fall wird immer besonders gekennzeichnet. Im Gegensatz zum asynchronen TD[0] Verfahren, werden beim synchronen TD[0] Verfahren alle Übergänge gleichzeitig betrachtet und die Wertfunktion wird unter Berücksichtigung aller dieser Daten aktualisiert. Der großer Vorteil des synchronen TD[0] Verfahrens liegt darin, dass es sich besser als die asynchrone Variante untersuchen lässt. Wir werden auch zeigen, dass die im synchronen Fall gewonnenen Erkenntnisse zum Teil auf den asynchronen Fall übertragbar sind.

Wie bereits in Kapitel 1 angesprochen, interessiert uns vor allem der Fall einer konstanten Lernrate \( \alpha \). Dies steht scheinbar im Widerspruch zu der sonstigen Forderung nach einer abfallenden Lernrate (vgl. Abschnitt 2.1). In der Praxis wird aber meistens eine konstante Lernrate verwendet, so dass die Forderung nach einer abfallenden Lernrate eher theoretischer Natur ist. Unter Verwendung von Satz 9 werden wir zeigen, dass das klassische TD[0] Verfahren (d.h. ohne Funktionsapproximation) auch für konstante Lernraten beschränkt bleibt. Daher erscheint es plausibel, auch im Fall von Funktionsapproximation die Beschränktheit des TD[0] oder anderer RL Verfahren zu fordern, wenn die eingesetzte Lernrate konstant gehalten wird. Wir werden anhand von Beispiel 10 konkret, und später auch allgemein zeigen, dass es Fälle geben kann, bei denen für alle Lernraten \( \alpha \neq 0 \) Divergenz auftreten kann. Allerdings gibt es durchaus kompatible
Funktionsapproximatoren, die sowohl im synchronen als auch im asynchronen Fall die Beschränktheit des TD[0] Verfahrens garantieren.

Wir beginnen zuerst mit der angesprochenen Motivation, dass das TD[0] Verfahren ohne Funktionsapproximation immer beschränkt bleibt. Die Vorschrift für das TD[0] Verfahren für einen Zustandsübergang \( g \rightarrow h \in S \) laut (2.8) definiert durch

\[
v(g) = (1 - \alpha)v(g) + \alpha(\gamma v(h) + r(g))
\]

wobei der Zustand \( h \) gemäß der Verteilung \( p(\cdot|g, \pi(g)) \) gesampelt wurde. Der besseren Lesbarkeit wegen, wurde in der obigen Formel \( r(g) \) statt dem ursprünglichen \( r(g, \pi(g)) \) verwendet, die Abhängigkeit von der verwendeten Strategie ist aber wie beim Übergang \( g \rightarrow h \) implizit vorhanden.

Ist der Zustandsraum \( S \) \( \{1, \ldots, n\} \) endlich, so kann \( v(i) \) als die \( i \)-te Koordinate des Vektors \( v = (v(1), \ldots, v(n))^\top \) aufgefasst werden. Mit

\[
\theta_i := (0, \ldots, 0, \underbrace{1}_{\text{i-te Stelle}}, 0, \ldots, 0)^\top \in \mathbb{R}^n
\]

kann die Aktualisierung (4.1) auf Grund der Identität \( v(i) = \theta_i^\top v \) in Vektorform geschrieben werden

\[
v = (I + \alpha \theta_g(\gamma \theta_h - \theta_g)^\top)v + r\theta_g
\]

Man kann obige Vorschrift auch in homogener Form schreiben. So kommt auf Iterationsmatrizen, die den Voraussetzungen von Satz 10 genügen, und damit die Beschränktheit von \( v \) auch bei beliebigen Anwendungen der TD[0] Vorschrift (4.1) mit festem \( \alpha \) wahren. Dies fassen wir im folgenden Korollar zusammen

**Korollar 3** Das TD[0] Verfahren bleibt für Lernraten \( \alpha \in [0, 1) \) beschränkt. Die Lernrate kann sich dabei durchaus verändern, für die Beschränktheit ist nicht gefordert, dass \( \alpha \) gemäß (2.9) abfällt.

*Beweis:* siehe obige Herleitung und Satz 10.

Damit bleibt das TD[0] Verfahren auch bei konstanten Lernraten beschränkt. Im Fall von Funktionsapproximation wird das nicht immer so sein. Wir werden sehen, dass der Grund für eine mögliche Divergenz rein algebraischer Natur ist. Unsere Untersuchungen werden uns dabei zu verschiedenen Verbindungen mit den im Kapitel 3 vorgestellten Resultaten führen.

### 4.1 TD[0] mit linearer Funktionsapproximation

Die TD[0] Aktualisierungsvorschrift (4.1) muss für eine parametrisierte Darstellung \( v(\cdot; w) \) der Wertfunktion \( v \) angepasst werden. Wir konzentrieren uns hier auf eine Darstellung durch lineare Funktionsapproximation mit einem Parametervektor \( w = (w_1, \ldots, w_m)^\top \in \mathbb{R}^m \)

\[
v(s; w) = \sum_{i=1}^m w_i \phi_i(s) = w^\top \Phi(s) = \sum_{i=1}^m w_i \phi_i(s)
\]

wobei

\[
\Phi(s) := \left[ \phi_1(s), \ldots, \phi_m(s) \right]^\top
\]
KAPITEL 4. KONVERGENZ VON APPROXIMATIVEM RL

wobei \{\phi_1, \ldots, \phi_m\} eine Menge von Basisfunktionen \(\phi_i : S \to \mathbb{R}\) darstellt. Der Vektor

\[ \varphi(s) = (\phi_1(s), \ldots, \phi_m(s))^\top \]  

(4.3)

aller Auswertungen der Basisfunktionen für den Zustand \(s\) wir als Merkmalsvektor von \(s\) bezeichnet.

Die Darstellung (4.2) kann man nicht direkt in (4.1) einsetzen, denn es beinhaltet noch keine Vorschrift, wie der Parametervektor \(w\) angepasst werden muss. Wir deuten dies durch die Verwendung eines Pfeils statt der Gleichheit an

\[ v(g; w) \mapsto (1 - \alpha) v(g; w) + \alpha (\gamma v(h; w) + r(g)) =: \tau \]

Der eigentliche Parameter \(w\) wird so angepasst, dass die Diskrepanz zwischen \(v(g; w)\) und dem neuen Wert \(\tau\) möglichst klein wird. Es geht also darum den Fehler

\[ E(w, \tau) = \frac{1}{2} (v(g; w) - \tau)^2 \]  

(4.4)

zu minimieren. Da \(E(w, \tau)\) eine konvexe Funktion ist, wird das Minimum für den Vektor \(w\) angenommen, für den der Gradient von \(E(w, \tau)\) verschwindet. Laut (2.17) ist

\[ \frac{\partial E(w, \tau)}{\partial w} = (\varphi(g)^\top w - \tau) \varphi(g) \]  

(4.5)

daher wird \(E(w, \tau)\) von \(w = (\tau/\varphi(g)^\top \varphi(g)) \varphi(g)\) minimiert. Insgesamt erhalten wir die Vorschrift

\[ w^{k+1} = \frac{1}{\varphi(g)^\top \varphi(g)} \varphi(g) \left( (1 - \alpha) v(g; w^k) + \alpha (\gamma v(h; w^k) + r(g)) \right) \]

(4.6)

Auf die Vorschrift (4.6) wird erneut Relaxation angewendet, so dass insgesamt die folgende vollständige Aktualisierung

\[ w^{k+1} = (1 - \beta) w^k + \beta \frac{1}{\varphi(g)^\top \varphi(g)} \varphi(g) ((1 - \alpha) \varphi(g) + \alpha \gamma \varphi(h))^\top w^k + \frac{\alpha r(g)}{\varphi(g)^\top \varphi(g)} \varphi(g) \]

zu Stande kommt. Allerdings wird in der Praxis selten die Relaxation doppelt angewendet. Dies bedeutet, dass der erste Schrittweitenparameter \(\alpha\) gleich 1 gesetzt wird, denn der zweite Relaxationsschritt hat mehr Einfluss auf die Aktualisierung von \(w\). Damit erhalten wir die vereinfachte Formel

\[ w^{k+1} = \left( I + \kappa \beta \varphi(g)(\gamma \varphi(h) - \varphi(g))^\top \right) w^k + \kappa \beta r(g) \varphi(g) \]

wobei \(\kappa = (\varphi(g)^\top \varphi(g))^{-1}\) gesetzt wurde. Der Faktor \(\kappa\) ist ein Normalisierungsfaktor, er sichert zu, dass bei der Lernrate \(\beta = 1\) wirklich das Minimum der Fehlerfunktion erreicht wird. Wird dieser Normalisierungsfaktor verwendet, so wird die TD[0]...
4.2. DIVERGENZ DES TD[0] VERFAHRENS

Vorschrift auch manchmal Kaczmarz Aktualisierung genannt [Par97, MS02]. Da es sich aber lediglich um eine Skalierung der Lernrate handelt, ändert die Verwendung des Normalisierungsfaktors nichts an den Konvergenzeigenschaften des TD[0] Verfahrens. Wir werden daher die folgende vereinfachte Aktualisierungsformel

\[ w_{k+1} = \left( I + \alpha \varphi(g) (\gamma \varphi(h) - \varphi(g)) \right) w_k + \alpha r(g) \varphi(g) \]  

verwenden. Formel (4.7) stellt den Ausgangspunkt unserer Untersuchungen dar. Sie wird stellvertretend für alle TD[0] Aktualisierungen mit Funktionsapproximation verwendet. Der Parameter \( \alpha \) wird als Lernrate bezeichnet. Da wir ohne den Normalisierungsfaktor \( \kappa \) arbeiten, muss \( \alpha \) nicht unbedingt aus dem Intervall \([0,1]\) stammen.

4.2 Divergenz des TD[0] Verfahrens

Problematisch am TD[0] Verfahren ist, dass es es unter Anwendung von Funktionsapproximation divergieren kann. In diesem Abschnitt zeigen wir ein elementares Beispiel (vgl. Beispiel 10) bei dem keine noch so kleine aber konstante Lernrate \( \alpha \) die Divergenz der Verfahrens aufhalten kann. Dieses Beispiel wird uns auch als Motivation für das synchrone TD[0] Verfahren dienen, das wir im Abschnitt 4.3 behandeln werden.

Wir beginnen allerdings mit wiederholten Einzelübergängen, um ein einfaches notwendiges Kriterium (vgl. Satz 14) für die Beschränktheit des TD[0] Verfahrens zu motivieren. Wir zeigen dann im Beispiel 10, dass es trotz der Erfüllung dieses Kriteriums zu Divergenz kommen kann.

Wir betrachten zwei Übergänge \( g_1 \sim h \in S \) und \( g_2 \sim h \in S \) mit

\[ \varphi(g_1) = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix}, \quad \varphi(g_2) = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \quad \varphi(h) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \]  

wobei über den Zustandsraum \( S \) nichts weiter vorausgesetzt wird. Da dieses Beispiel aber durch Anwendung von gitterbasierten Approximatorenpaaren entstanden ist, kann man sich den Zustandsraum anschaulich als das Intervall \([0,1] \subset \mathbb{R}\) vorstellen.

Der angewendete Funktionsapproximator ist

\[ \phi(s) = w_1 \max \{ 0, 1 - |s| \} + w_2 \max \{ 0, 1 - |s - 1| \} \]

Die Funktionen \( \phi_1 \) und \( \phi_2 \) stellen so genannte Hutfunktionen dar, \( w_1 \) und \( w_2 \) sind die korrespondierenden Koeffizienten. In Abbildung 4.1 sind \( \phi_1 \) und \( \phi_2 \) dargestellt. Man sieht auch, dass \( \phi_2 \) eine um 1 nach rechts verschobene Kopie von \( \phi_1 \) darstellt.

---

1 Gitterapproximatorenpaare werden ausführlich in Kapitel 5 behandelt.
2 Wenn man bei endlichen Zustandsräumen bleiben möchte, so ist auch jede Diskretisierung von \([0,1]\) denkbar, die die Zustände \( g_1, g_2 \) und \( h \) enthält.
KAPITEL 4. KONVERGENZ VON APPROXIMATIVEM RL

Abbildung 4.1: Die Hutfunktionen $\phi_1$ und $\phi_2$.

Die Funktionen $\phi_1$ und $\phi_2$ sind auf ganz $\mathbb{R}$ definiert. Falls man sie auf das Intervall $[0, 1]$ einschränkt, so ergibt sich eine besonders einfache Darstellung

$$\varphi(s) = (\phi_1(s), \phi_2(s))^\top = (1 - s, s)^\top, \quad \text{für } s \in [0, 1] \quad (4.9)$$

Damit kann man die durch (4.8) charakterisierten Zustände $g_1$, $g_2$ und $h$ mit den reellen Zahlen $g_1 = 0.6$, $g_2 = 0.5$ und $h = 1.0$ identifizieren. Dies wurde in Abbildung 4.2 veranschaulicht.

Die TD[0] Aktualisierungsvorschrift (4.7) ergibt für die Zustände $g_1 = 0.6$ und $h = 1.0$

$$w^{k+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \alpha \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} \begin{pmatrix} \gamma(0) & 0 \\ 0 & \gamma(1) \end{pmatrix} w^{k+1} + \alpha r(0.4) \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} =: A_1$$

und für die Zustände $g_2 = 0.5$ und $h = 1.0$

$$w^{k+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \alpha \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} \gamma(0) & 0 \\ 0 & \gamma(1) \end{pmatrix} w^{k} + \alpha r(0.5) \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} =: b_2$$

Wir betrachten die obigen Aktualisierungsvorschriften mit einem festen Diskontierungsfaktor $\gamma = 0.95$. Wenn man nun ein $\alpha > 0$ festlegt, dann hat die Matrix $I + \alpha A_1$ zwangsläufig einen Spektralradius $\rho(I + \alpha A_1) > 1$. Dies liegt
4.2. DIVERGENZ DES TD[0] VERFAHRENS

daran, dass $A_1$ nur die nichtnegativen Eigenwerte $\{0, 0.05\}$ besitzt. Somit würden wiederholte Iterationen mit dieser Matrix divergieren.

Ist dagegen $\alpha < 0$ so hat die Matrix $I + \alpha A_2$ einen Spektralradius $\rho(I + \alpha A_2) > 1$. Die Matrix $A_2$ hat nämlich nur die nichtpositiven Eigenwerte $\{0, -0.025\}$. Damit würden wiederholte Iterationen mit der Matrix $I + \alpha A_2$ divergieren.

Insgesamt sieht man, dass es keine Lernrate $\alpha \neq 0$ gibt, so dass man beide Übergänge $g_1 \sim h$ und $g_2 \sim h$ ohne Divergenz wiederholen kann. Dies fassen wir im folgenden Satz zusammen

**Satz 14** Die Eigenschaft

$$\forall g, h \in S : \varphi(g)^T (\gamma \varphi(h) - \varphi(g)) \geq 0$$

bzw.

$$\forall g, h \in S : \varphi(g)^T (\gamma \varphi(h) - \varphi(g)) \leq 0$$

ist notwendig, damit beliebige approximative TD[0] Aktualisierungen (4.7) für ein festes $\alpha \neq 0$ nicht divergieren.

Beweis: $\varphi(g)^T (\gamma \varphi(h) - \varphi(g))$ ist ein Eigenwert der Matrix $\varphi(g)(\gamma \varphi(h) - \varphi(g))^T$. Wenn Zustände $g_1, g_2, h_1, h_2 \in S$ mit der Eigenschaft

$$\varphi(g_1)^T (\gamma \varphi(h_1) - \varphi(g_1)) < 0$$

und

$$\varphi(g_2)^T (\gamma \varphi(h_2) - \varphi(g_2)) > 0$$

existieren, dann wird, wie in der obigen Herleitung gezeigt, eine der Aktualisierungen für die Übergänge $g_1 \sim h_1$ bzw. $g_2 \sim h_2$ bei wiederholter Anwendung divergieren.

Das in Satz 14 angegebene Kriterium ist nicht hinreichend, damit keine Divergenz auftritt. Wir werden jetzt ein elementares Beispiel vorstellen, in dem zwar die Bedingungen von Satz 14 eingehalten werden, es aber trotzdem zu Divergenz bei zwei aufeinander folgenden Übergängen kommen kann. Das Beispiel zeigt insbesondere, dass es keine konstante Lernrate $\alpha \neq 0$ gibt, für die eine spezielle Folge von asynchronen TD[0] Aktualisierungen beschränkt bleibt.

**Beispiel: 10** Sei $S = \{0, 0.5, 1\} \subset [0,1]$ eine Einschränkung des Zustandsraums $[0,1]$. Durch direktes Nachrechnen kann man zeigen, dass die Bedingungen von Satz 14 erfüllt sind, d.h. es gilt $\varphi(g)^T (\gamma \varphi(h) - \varphi(g)) < 0$ für $\varphi$ aus (4.9) und für alle $g, h \in S$.

Betrachte die Zustandsübergänge $0 \sim 0$ und $0.5 \sim 1$. Der Übergang $0 \sim 0$ entspricht der folgenden TD[0] Aktualisierung

$$w^{k+1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \end{pmatrix} + \alpha \begin{pmatrix} \varphi(0) \\ \gamma \begin{pmatrix} 1 \\ 0 \\ \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \\ \end{pmatrix} \end{pmatrix} =: A_3$$

$$w^{k+1} = \alpha r(0) \begin{pmatrix} 1 \\ 0 \\ \end{pmatrix} =: b_3$$
KAPITEL 4. KONVERGENZ VON APPROXIMATIVEM RL

Abbildung 4.3: Der Spektralradius der Matrix $A = (I + \alpha A_2)(I + \alpha A_3)$ in Abhängigkeit von der Lernrate $\alpha$.

Die Aktualisierung des zweiten Übergangs $0.5 \sim 1$ wurde bereits in Gleichung (4.10) angegeben. Das hintereinander Ausführen beider Aktualisierungen liefert die Vorschrift

$$w^{k+1} = (I + \alpha A_2)((I + \alpha A_3)w^k + \alpha b_3) + \alpha b_2$$
$$= (I + \alpha A_2)(I + \alpha A_3)w^k + \alpha^2 A_2 b_3 + \alpha(b_3 + b_2)$$

Die Matrix $A$ hat für $\gamma = 0.95$ und $\alpha = 0.1$ die Eigenwerte $\sigma(A) = \{0.985083, 1.00754\}$. Damit ist der Spektralradius $\rho(A) > 1$, und die wiederholte Anwendung dieser Doppelaktualisierung würde zur Divergenz führen.

Die Lernrate $\alpha = 0.1$ ist bereits relativ klein, es ist aber nicht ohne weiteres klar, dass es keine Lernrate $\alpha \neq 0$ gibt, für die der Spektralradius doch kleiner oder gleich 1 werden würde. Dies kann man aber sehen, wenn man die Matrix $(I + \alpha A_2)(I + \alpha A_3)$ schreibt als

$$(I + \alpha A_2)(I + \alpha A_3) = I + \frac{\alpha}{80} \begin{pmatrix} \alpha - 24 & 18 \\ \alpha - 20 & 18 \end{pmatrix}$$

Damit $\rho(A) > 1$ gilt, reicht es zu zeigen, dass die Matrix $C_\alpha$ immer einen positiven und einen negativen Eigenwert hat. Es gilt $\det(I - C_\alpha) = \lambda^2 + \lambda(6 - \alpha) - 72$ und damit hat $C_\alpha$ die reellen Eigenwerte

$$\lambda_{1/2} = \frac{1}{2} \left( \alpha - 6 \pm \sqrt{(\alpha - 6)^2 + 288} \right)$$

Da $\mathbb{R} \ni \sqrt{(\alpha - 6)^2 + 288} > |\alpha - 6|$ gilt, ist immer einer der Eigenwerte von $C_\alpha$ positiv und der andere negativ. Auf Grund von $\rho(A) = \max\{1 + \frac{\alpha}{80} \lambda_1, 1 + \frac{\alpha}{80} \lambda_2\}$ ist der Spektralradius der Iterationsmatrix $A$ für eine Lernrate $\alpha \neq 0$ immer größer als 1. Die Abhängigkeit des Spektralradius von der Lernrate $\alpha$ ist in Abbildung 4.3 graphisch dargestellt.

Die Analyse der Divergenz des TD[0] Verfahrens in Beispiel 10 zeigt sehr deutlich, dass der Einsatz von Funktionsapproximation problematisch sein kann, wenn man keine
4.2. DIVERGENZ DES TD[0] VERFAHRENS

Einschränkungen an die verwendeten Approximatoren stellt. Es wurde gezeigt, dass es keine Lernrate $\alpha \neq 0$ gibt, für die die wiederholte Anwendung der 2 Aktualisierungen beschränkt bleiben würde.

Die Vorgehensweise, die uns zu diesem Schluss gebracht hat, war die Analyse des Spektralradius von Produkten der beteiligten Matrizen. Bereits für zwei $2 \times 2$ Matrizen kann dieses relativ aufwendig werden. Die Analyse in Beispiel 10 war machbar, weil wir die Matrix $A = (I + \alpha A_2)(I + \alpha A_3)$ geschickt als Summe von 2 einfachen Matrizen schreiben konnten. Ansonsten hätte das volle charakteristische Polynom $\det(A - \lambda I)$ von $A$ studiert werden müssen, dass eine Abhängigkeit von $\alpha$ in bis zur vierten Potenz aufweist. Für größere Matrizen ist eine solche Analyse von Matrizenprodukten kaum möglich.

Wenn man in Beispiel 10 anstatt des Produktes $A = (I + \alpha A_2)(I + \alpha A_3)$ die gewichtete Summe

$$B = \frac{1}{2}(I + \alpha A_2) + \frac{1}{2}(I + \alpha A_3)$$

betrachten würde, so vereinfacht sich die Analyse erheblich. Nun reicht es die Eigenwerte der Matrix $\frac{1}{2}(A_2 + A_3)$ zu untersuchen. Diese sind gleich $\{-0.075, 0.0375\}$ und damit ist der Spektralradius $\rho(B) = \max\{1 - 0.075\alpha, 1 + 0.0375\alpha\}$ der Matrix $B$ für alle Lernraten $\alpha \neq 0$ größer als 1. Allerdings ist nicht sofort einleuchtend, was diese Aussage wirklich bedeutet. Wir werden im Abschnitt 4.3 sehen, dass $\rho(B) > 1$ ebenfalls eine Verletzung einer notwendigen Bedingung für die Beschränktheit von beliebigen TD[0] Aktualisierungen darstellt.

Die Verwendung der Matrix $B = I + \frac{\alpha}{2}(A_2 + A_3)$ kann durch gleichzeitiges Betrachten von mehreren Übergängen, und die anschließende simultane Fehlerminimierung, motiviert werden. Dies wird im Abschnitt 4.3 in voller Allgemeinheit als synchrones TD[0] Verfahren entwickelt. Wir betrachten hier die in Beispiel 10 vorgestellten Übergänge. Für den Übergang $0 \sim 0$ gilt es den Fehler $(v(0; w) - \tau_2)^2$ und für den Übergang $0.5 \sim 1$ den Fehler $(v(0.5; w) - \tau_3)^2$ zu minimieren, wobei

$$\tau_2 = \gamma v(0; w) + r(0), \quad \text{und} \quad \tau_3 = \gamma v(1; w) + r(0.5)$$

die einzelnen TD[0] Zielwerte darstellen. Um beide Fehler gleichzeitig zu minimieren betrachten wir eine gewichtete Summe beider Fehler

$$2E(w, \tau_2, \tau_3) = \frac{1}{2}((v(0; w) - \tau_2)^2 + (v(0.5; w) - \tau_3)^2)$$

Laut (2.17) und (4.2) ist

$$\frac{\partial E(w, \tau_2, \tau_3)}{\partial w} = \frac{1}{2} \left((\varphi(0)^T w - \tau_2)\varphi(0) + (\varphi(0.5)^T w - \tau_3)\varphi(0.5)\right)$$

und der Fehler $E(w, \tau_2, \tau_3)$ kann gegen die Richtung des Gradienten $\partial E/\partial w$ mit
Schrittweitenparameter bzw. Lernrate \( \alpha \) minimiert werden

\[
w^{k+1} = w^k - \alpha \frac{\partial E(w, \tau_2, \tau_3)}{\partial w} \\
= w^k - \frac{\alpha}{2} \left( (\varphi(0)^\top w^k - \tau_2)\varphi(0) + (\varphi(0.5)^\top w^k - \tau_3)\varphi(0.5) \right) \\
= \left( I + \frac{\alpha}{2}(\varphi(0)(\gamma\varphi(0) - \varphi(0))^\top) \right) w^k - \frac{\alpha}{2} r(0)\varphi(0) \\
+ \left( I + \frac{\alpha}{2}(\varphi(0.5)(\gamma\varphi(0.5) - \varphi(0.5))^\top) \right) w^k - \frac{\alpha}{2} r(0.5)\varphi(0.5) \\
= \left( I + \frac{\alpha}{2}(A_2 + A_3) \right) w^k - \frac{\alpha}{2} \left( r(0)\varphi(0) + r(0.5)\varphi(0.5) \right)
\]

(4.11)

Die Aktualisierungsvorschrift (4.11) hat als Iterationsmatrix \( I + \frac{\alpha}{2}(A_2 + A_3) \) was der weiter oben eingeführten Matrix \( B \) entspricht. Dort wurde auch gezeigt, dass die Matrix \( B \) für alle \( \alpha \neq 0 \) einen Spektralradius größer als 1 hat. Damit wird die wiederholte Anwendung der Aktualisierungsvorschrift (4.11) für jede Lernrate \( \alpha \neq 0 \) divergieren.

Zusammenfassend gesehen hat dieser Abschnitt gezeigt, dass die TD[0] Vorschrift für hintereinander ausgeführte Aktualisierungen divergieren kann, und zwar für alle möglichen Lernraten die ungleich 0 sind. Die dabei beteiligten Matrizen lassen sich auf Grund von Produktbildung relativ schwer untersuchen. Der Analyse eher zugänglich sind gewichtete Summen von Matrizen. Wir haben gezeigt, dass dies dem Fall von gleichzeitigem Minimieren der Fehler für mehrere Übergänge entspricht. Im Abschnitt 4.3 werden wir diesen synchronen Fall weiter verallgemeinern, und auch zeigen, dass dessen Beschränktheit notwendig für die Beschränktheit bei der allgemeinen asynchronen Vorgehensweise ist.

### 4.3 Synchrones Reinforcement Lernen


In [SM03a] wurden für eine spezielle Klasse von Matrizen hinreichende Kriterien für die Konvergenz des synchronen RL angegeben. Wir werden dieses Resultat mit Hilfe der Ergebnisse aus Kapitel 3 auf beliebige Matrizen und auf sowohl hinreichende als
4.3. SYNCHRONES REINFORCEMENT LERNEN


Grundlage für das synchrone TD[0] Verfahren bildet eine endliche Menge \( \Gamma \) von Übergängen innerhalb des Zustandsraumes \( S \) und den dazu gehörigen Belohnungswerten

\[
\Gamma = \{(g_i, h_i, r_i) \mid g_i, h_i \in S, r_i \in \mathbb{R}, i = 1 \ldots k\}. \tag{4.12}
\]

Dabei ist \( g_i \sim h_i \) ein Übergang mit dem Startzustand \( g_i \) und dem gemäß der durch eine Strategie \( \pi \) implizierten Systemwahr scheinlichkeit \( p(\cdot|g_i, \pi(g_i)) \) erfolgten Zielzustand \( h_i \). Die unmittelbare Belohnung für diesen Übergang wird mit \( r_i \) bezeichnet.

Außer der Endlichkeit von \( \Gamma \) gibt es keine Einschränkungen an die Menge \( \Gamma \). So müssen die Startzustände \( g_i \) nicht unbedingt alle verschiedenen sein. Das Vorkommen von z.B. den Übergängen \( g \sim h_1 \) und \( g \sim h_2 \) in \( \Gamma \) mit \( h_1 \neq h_2 \) spiegelt die stochastische Natur des zu Grunde liegenden Systems wider. Da die Menge \( \Gamma \) endlich ist, ist auch die Anzahl der unterschiedlichen Zustände in \( \Gamma \) endlich. Wir bezeichnen die in den Übergängen von \( \Gamma \) vorkommenden unterschiedlichen Zustände mit \( \{s_1, \ldots, s_n\} \subset S \).

Für jeden der in der Menge \( \Gamma \) vorkommenden Übergänge \( g_i \sim h_i \) gibt es gemäß der TD[0] Aktualisierung (4.7) den Zielwert \( \tau_i \) mit

\[
\tau_i = \gamma v(h_i; w) + r_i = \gamma \varphi(h_i)^\top w + r_i
\]

Ausgehend von den einzelnen Zielwerten wird die Fehlerfunktion

\[
2E(w) = \frac{1}{|\Gamma|} \sum_{i=1}^{|\Gamma|} (v(g_i; w) - \tau_i)^2 \tag{4.13}
\]

betrachtet. Der Faktor \( \frac{1}{|\Gamma|} \) ist zur Unabhängigkeit von der Anzahl der Übergänge enthalten.

Das Ziel des synchronen TD[0] Verfahrens ist es, den Fehler (4.13) zu minimieren. Dazu verwenden wir den Gradienten \( \partial E/\partial w \), der laut (2.17) und (4.2) die folgende
Form

\[
\frac{\partial E(w)}{\partial w} = \frac{1}{|\Gamma|} \sum_{i=1}^{|\Gamma|} (\varphi(g_i) \top w - \tau_i) \varphi(g_i)
\]

\[
= -\frac{1}{|\Gamma|} \left( \sum_{i=1}^{|\Gamma|} \varphi(g_i) (\gamma \varphi(h_i) - \varphi(g_i) \top w + r_i \varphi(g_i)) \right)
\]

(4.14)

hat. Die Fehlerfunktion (4.13) wird nun mittels Gradientenabstiegs gegen die Richtung des Gradienten (4.14) minimiert

\[
w^{k+1} = w^k + \frac{\alpha}{|\Gamma|} \sum_{i=1}^{|\Gamma|} \varphi(g_i) (\gamma \varphi(h_i) - \varphi(g_i) \top w_k + \frac{\alpha}{|\Gamma|} \sum_{i=1}^{|\Gamma|} r_i \varphi(g_i))
\]

(4.15)

Die Darstellung (4.15) ist auf Grund der Summation von $|\Gamma|$ Matrizen nicht besonders handhabbar. Wir werden im Folgenden eine Darstellung entwickeln, die besser für die Untersuchung von Konvergenzeigenschaften der Vorschrift (4.15) geeignet ist. Es bezeichne

\[
\hat{d}_i = \frac{1}{|\Gamma|} \{ (g, h, r) \in \Gamma \mid g = s_i \}
\]

die relative Häufigkeit des Zustands $s_i$ in den Startzuständen von $\Gamma$, und

\[
\hat{p}_{ij} = \frac{\{ (g, h, r) \in \Gamma \mid g = s_i, h = s_j \}}{\{ (g, h, r) \in \Gamma \mid g = s_i \}}
\]

die relative Häufigkeit des Übergangs $s_i \leadsto s_j$ unter den Übergängen in $\Gamma$ mit Startzustand $s_i$. Die Definition von $\hat{p}_{ij}$ macht natürlich nur dann Sinn, wenn $\hat{d}_i > 0$ ist. Im Fall von $\hat{d}_i = 0$ wird $\hat{p}_{ij}$ ebenfalls auf 0 gesetzt. Mit Hilfe der Koeffizienten $\hat{d}_i$ und $\hat{p}_{ij}$ kann man (4.15) auch schreiben als

\[
w^{k+1} = \sum_{i=1}^n \hat{d}_i \sum_{j=1}^n \hat{p}_{ij} \left( I + \alpha \varphi(s_i) (\gamma \varphi(s_j) - \varphi(s_i)) \top \right) w_k + \sum_{i=1}^n \hat{d}_i r(s_i) \varphi(s_i)
\]

(4.16)

Man kann die Koeffizienten $\hat{d}_i$ und $\hat{p}_{ij}$ als Matrizen $\hat{D}$ und $\hat{P}$ zusammenfassen

\[
\hat{D} = \text{diag}(\hat{d}_1, \ldots, \hat{d}_n), \quad \hat{P} = (\hat{p}_{ij}) \in \mathbb{R}^{n \times n},
\]

(4.17)

wobei $\hat{D}$ eine nichtnegative Diagonalmatrix mit Spur 1 und $\hat{P}$ eine stochastische Matrix bilden. Unter Anwendung der Identität (3.27) kann man aus (4.16) die äquivalente Form

\[
w^{k+1} = \left( I - \alpha \Phi \top \hat{D} (I - \gamma \hat{P}) \Phi \right) w_k + \Phi \top \hat{D} r
\]

(4.18)

herleiten. Dabei ist $r = (r(s_1), \ldots, r(s_n)) \top$, und die Matrix $\Phi$ ist durch (3.26) mit $\varphi_i = \varphi(s_i)$ gegeben. Die Form (4.18) des synchronen TD[0] Verfahrens wird im weiteren Verlauf als die kanonische Form des synchronen TD[0] Verfahrens
4.3. SYNCHRONES REINFORCEMENT LERNEN

verwendet. Dabei sind die Konvergenzeigenschaften dieses Verfahrens maßgeblich von den Eigenschaften der generischen Matrix

\[ A_{TD} := \Phi^T \tilde{D}(I - \gamma \tilde{P})\Phi \quad (4.19) \]

abhängig. Generisch deshalb, weil die Matrizen \( \tilde{D} \) und \( \tilde{P} \) je nach Menge \( \Gamma \) der Übergänge variieren, und wir an Konvergenzeigenschaften von (4.18) unabhängig von einer speziellen Menge \( \Gamma \) von Übergängen interessiert sind.

Da es sich bei dem synchronen TD[0] Verfahren (4.18) um eine Matrixiteration handelt, können wir Satz 4 anwenden, um das Verhalten des synchronen TD[0] Verfahrens zu untersuchen. Dies fassen wir im folgenden Korollar zusammen

**Korollar 4** Das synchrone TD[0] Verfahren (4.18) konvergiert genau dann für eine geeignete Lernrate \( \alpha > 0 \) wenn die Matrix \( A_{TD} \) positiv stabil ist. Ist \( A_{TD} \) lediglich semi positiv stabil, so konvergiert das TD[0] Verfahren bedingt für ein geeignetes \( \alpha > 0 \) genau dann, wenn folgende Bedingungen

\[ \text{Ker}(A_{TD}^2) = \text{Ker}(A_{TD}) \quad \text{und} \quad \Phi^T \tilde{D}r \in \text{Im}(A_{TD}) \quad (4.20) \]

gelten. Ist die Matrix \( A_{TD} \) nicht semi (positiv) stabil, d.h. sie hat Eigenwerte mit positivem und negativem Realteil, so divergiert das synchrone TD[0] Verfahren für alle Lernraten \( \alpha \neq 0 \).

Beweis: Man kann immer eine Lernrate \( \alpha > 0 \) so wählen, dass die Matrix \( I - \alpha A_{TD} \) höchstens die 1 als Eigenwert mit Betrag 1 hat. Ferner ist die 1 nur dann ein Eigenwert von \( I - \alpha A_{TD} \), wenn \( A_{TD} \) singulär ist. Damit kann man im Falle der Beschränktheit der Iteration (4.18) für geeignete Lernraten \( \alpha \) auch die bedingte Konvergenz des synchronen TD[0] Verfahrens folgen.

Für die Beschränktheit der Iteration (4.18) ist im Fall einer singulären Matrix \( A_{TD} \) die Gleichheit des Eigenraums zum Eigenwert 1 mit dem korrespondierenden Hauptraum erforderlich (vgl. Satz 4). Da der Eigenraum von \( I - \alpha A_{TD} \) zum Eigenwert 1 gleich dem Kern von \( A_{TD} \) ist, ist der Hauptraum von \( I - \alpha A_{TD} \) zum Eigenwert 1 nur dann gleich dem korrespondierenden Eigenraum, wenn \( \text{Ker}(A_{TD}^2) = \text{Ker}(A_{TD}) \) gilt.

Nach diesen Überlegungen folgen alle Behauptungen dieses Korollars direkt aus Satz 4.

Die Bedingung (4.20) in Korollar 4 ist nicht immer einfach nachzuprüfen, und schon gar nicht für alle möglichen nichtnegativen Diagonalmatrizen \( \tilde{D} \) und stochastischen Matrizen \( \tilde{P} \). Wir wollen uns daher auf die Bedingung für die (semi) positive Stabilität der Matrix \( A_{TD} = \Phi^T \tilde{D}(I - \gamma \tilde{P})\Phi \) konzentrieren. Diese stellt auf jeden Fall eine notwendige Bedingung für die (bedingte) Konvergenz des synchronen TD[0] Verfahrens (4.18) dar. Es wird ferner in Korollar 5 gezeigt werden, dass die (semi) positive Stabilität von \( A_{TD} \) auch im asynchronen Fall eine wichtige Rolle spielt. Die Bedingung (4.20) ist eine Verfeinerung, die im synchronen Fall benötigt wird, falls \( A_{TD} \) semi positiv stabil ist, also insbesondere singulär ist.
Tabelle 4.1: Eigenschaften von Matrizen der Form $\Phi^\top \hat{D}(I - \gamma \hat{P})\Phi$.

<table>
<thead>
<tr>
<th>Eigenschaften</th>
<th>(semi) positiv stabil</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Phi$ orthogonal</td>
<td>ja, vgl. Lemma 7</td>
</tr>
<tr>
<td>$\Phi$ Untermatrix einer nichtnegativen orthogonalen Matrix</td>
<td>ja, vgl. Satz 5</td>
</tr>
<tr>
<td>$\Phi$ nichtnegative Untermatrix einer orthogonalen Matrix</td>
<td>nein, vgl. Beispiel 2</td>
</tr>
<tr>
<td>$\Phi$ substochastisch</td>
<td>nein, vgl. Beispiel 4</td>
</tr>
<tr>
<td>$\Phi$ mit Einträgen aus ${0, 1}$</td>
<td>nein, vgl. Beispiel 5</td>
</tr>
<tr>
<td>$\Phi$ mit Einträgen aus ${0, 1}$ und orthogonalen Spalten</td>
<td>ja, vgl. Satz 7</td>
</tr>
</tbody>
</table>

Im Abschnitt 3.3 von Kapitel 3 wurden Matrizen der Form $\Phi^\top \hat{D}(I - \gamma \hat{P})\Phi$ auf (semi) positive Stabilität untersucht. Das Ergebnis dieser Untersuchungen haben wir in Tabelle 4.3 zusammengefasst.

Es fällt auf, dass außer dem eher uninteressanten Fall einer orthogonalen Matrix $\Phi$, nur für eine Klasse von Matrizen gezeigt werden konnte, dass die (semi) positive Stabilität der Matrix $\Phi^\top \hat{D}(I - \gamma \hat{P})\Phi$ für alle möglichen Diagonalmatrizen $\hat{D}$ und stochastischen Matrizen $\hat{P}$ erhalten bleibt.

Es bleibt fraglich, ob eine andere große Klasse von Matrizen $\Phi$ existiert, für die ebenfalls die (semi) positive Stabilität erhalten bleibt. Sollte diese Frage in Zukunft entweder bejaht oder verneint werden können, so wird dies eine weitreichende Bedeutung auch für das asynchrone TD[0] Verfahren haben.

Wir werden jetzt zeigen, dass die (semi) positive Stabilität der Matrix $\Phi^\top \hat{D}(I - \gamma \hat{P})\Phi$ auch für das asynchrone TD[0] Verfahren wichtig ist. Laut Korollar 4 ist sie notwendig für die Beschränktheit des synchronen TD[0] Verfahrens und wir werden zeigen, dass die Beschränktheit des synchronen TD[0] Verfahrens notwendig für die Beschränktheit des TD[0] Verfahrens ist.

Den Schlüssel zu diesem Resultat bildet die Darstellung (3.26) von $\Phi$ und Satz 12. In der Tat ist diese Aussage eine unmittelbare Folgerung von Satz 12, so dass wir unser Hauptresultat als Korollar formulieren werden.

**Korollar 5** Gegeben sei das synchrone TD[0] Verfahren (4.18). Die Matrizen $\hat{D}$ und $\hat{P}$ sind dabei durch eine Menge $\Gamma$ (vgl. (4.12)) von Übergängen mittels (4.17) bestimmt.

Falls das synchrone TD[0] Verfahren divergiert, so divergiert auch das korrespondierende asynchrone TD[0] Verfahren (4.7) für bestimmte Folgen von Aktualisierungen. Dabei ist die Relation zwischen der Approximator Matrix $\Phi$ und den in (4.7) vorkommenden Merkmalsvektoren $\varphi(\cdot)$ durch (3.26) gegeben.

Insbesondere ist die (semi) positive Stabilität der Matrix $\Phi^\top \hat{D}(I - \gamma \hat{P})\Phi$ notwendig für die Beschränktheit des asynchronen TD[0] Verfahrens.

Beweis: die Behauptung folgt direkt aus Satz 12. ☐
Es bleibt anzumerken, dass das TD[0] Verfahren selbst im Fall von semi positiver Stabilität der Matrix $\Phi \top \hat{D}(I - \gamma \hat{P}) \Phi$ divergieren kann (vgl. Korollar 4). Wie sich in diesem Fall das asynchrone TD[0] Verfahren verhält hängt dann von den weiteren Bedingungen in Korollar 4 ab.

Sind erst mal alle notwendigen Kriterien für die Beschränktheit des synchronen TD[0] Verfahrens erfüllt, so ist noch nicht die Beschränktheit des asynchronen TD[0] Verfahrens gesichert. Dass die Lücke nicht sehr groß ist, zeigt sich anhand der Approximatorenklasse solcher $\Phi$, die Einträge aus $\{0, 1\}$ haben, und ferner orthogonale Spalten besitzen. Laut Tabelle 4.3 bzw. Satz 7 ist dann die Matrix $\Phi \top \hat{D}(I - \gamma \hat{P}) \Phi$ (semi) positiv stabil. Die Zeilen von $\Phi$ sind laut (3.26) gerade die Merkmalsvektoren $\varphi(i)$, und die obige Bedingung bedeutet, dass maximal ein Eintrag von $\varphi(i)$ gleich 1 ist, alle anderen sind gleich Null. Mit Satz 10 kann man nun auf die Beschränktheit des asynchronen TD[0] Verfahrens schließen. Satz 10 besagt in Wirklichkeit sogar mehr, denn nur die Startzustände müssen die strikte Bedingung an die Merkmalsvektoren $\varphi(i)$ erfüllen. Wir werden diesen Sachverhalt im Kapitel 5 genauer untersuchen.

### 4.4 Zusammenhang mit linearen Gleichungssystemen

Sowohl das synchrone als auch das asynchrone TD[0] Verfahren dienen dazu, aus beobachteten Übergängen den Wert einer Strategie iterativ anzunähern. Ohne die Verwendung von Funktionsapproximation gilt es das lineare Gleichungssystem (2.2) iterativ zu lösen. Da wir von einer festen Strategie $\pi$ ausgehen, werden wir die beteiligten Vektoren und Matrizen ohne besondere Kennzeichnung der Strategie schreiben

$$v = \gamma P v + r, \quad v, r \in \mathbb{R}^n \quad (4.21)$$

Unter Verwendung von Funktionsapproximation kann die Wertfunktion $v(\cdot)$ nur mittels der parametrisierten Version $v(\cdot; w)$ approximiert werden. Diese parametrisierte Version $v(\cdot; w)$ wird dabei durch die lineare Kombination (4.2) von Basisfunktionen $\{\phi_1, \ldots, \phi_m\}$ dargestellt. Man kann nun die zu den Basisfunktionen $\phi_i$ assozierte Approximatormatrix $\Phi$ durch

$$\Phi = \begin{pmatrix} \phi_1(1) & \cdots & \phi_m(1) \\ \vdots & \ddots & \vdots \\ \phi_1(n) & \cdots & \phi_m(n) \end{pmatrix} = \begin{pmatrix} \varphi(1)^\top \\ \vdots \\ \varphi(n)^\top \end{pmatrix} \in \mathbb{R}^{n \times m} \quad (4.22)$$

definieren. Dabei sind $\varphi(i)$ die durch (4.3) definierten Merkmalsvektoren. Die Darstellung der Matrix $\Phi$ mit Hilfe von Merkmalsvektoren offenbart die Ähnlichkeit zur Darstellung von $\Phi$ in (3.26).

Die parametrisierte Wertfunktion $v(\cdot; w)$ lässt sich nun in Matrixschreibweise darstellen

$$\begin{pmatrix} v(1; w) \\ \vdots \\ v(n; w) \end{pmatrix} = \Phi w \quad (4.23)$$
Durch Einsetzen von $v(\cdot; w)$ für $v(\cdot)$ in (4.21) erhält man das Gleichungssystem

$$\Phi w = \gamma P \Phi w + r$$

welches allerdings im Gegensatz zu (4.21) nicht immer lösbar ist. Daher ist man an einem Vektor $w$ interessiert, für welchen der euklidische Fehler

$$E(w, \tau) = \frac{1}{2} \| \Phi w - \tau \|_2^2$$

(4.24) mit $\tau = P \Phi w + r$ minimiert wird. Da $E(w, \tau)$ eine konvexe Funktion ist, wird der Fehler genau dann minimal, wenn der Gradient $\partial E/\partial w$ annulliert wird. Die Ableitung des Fehlers $E(w, \tau)$ ergibt

$$\frac{\partial}{\partial w} E(w, \tau) = \Phi^\top \Phi w - \Phi^\top \tau = \Phi^\top \Phi w - \Phi^\top (r + \gamma P \Phi w) = \left( \Phi^\top (I - \gamma P) \Phi \right) w - \Phi^\top r$$

Falls $\Phi$ vollen Spaltenrang hat, so ist die Matrix $\Phi^\top (I - \gamma P) \Phi$ regulär, und folglich ist das lineare Gleichungssystem

$$\Phi^\top (I - \gamma P) \Phi w = \Phi^\top r$$

(4.25) eindeutig lösbar. Durch die Aufteilung der Matrix $\Phi^\top (I - \gamma P) \Phi$ in $I + (I - \Phi^\top (I - \gamma P) \Phi)$ kann man das Gleichungssystem (4.25) mittels der Splitting Methode ([Mei99]) iterativ lösen. In dieser speziellen Form heißt die Splitting Methode Richardson Verfahren (vgl. [Mei99]). Dieses kann man durch Relaxation mit Relaxationsparameter $\alpha$ erweitern, und erhält

$$w^{k+1} = \left( I - \alpha \Phi^\top (I - \gamma P) \Phi \right) w^k + \alpha \Phi^\top r$$

(4.26) Die Vorschrift (4.26) gleicht nun bis auf die Verwendung der nichtnegativen Diagonalmatrix $\hat$ der iterativen Form des synchronen TD[0] Verfahrens (4.18). Durch die Verwendung des Fehlermaßes

$$E_D(w, \tau) = \| \Phi w - \tau \|_D^2, \quad \text{mit } \| x \|_D^2 = x^\top D x$$

in (4.26) kann man aber auch die Verwendung der Matrix $D$ als Gewichtungsmatrix in diese Analogie einfließen lassen.

Das synchrone TD[0] Verfahren entspricht also im Wesentlichen dem Vorgehen, dass anhand der aufgenommenen Übergänge die Übergangsmatrix $\hat P$ als Approximation an die wahre Systemmatrix $P$ aufgestellt wird. Ferner werden noch die Häufigkeiten der Startzustände in die Matrix $\hat D$ aufgenommen. Für dieses durch $\hat P$ und $\hat D$ approximierete System wird dann das Richardson Verfahren angewendet. Dieses braucht nicht immer zu konvergieren, selbst wenn das zu Grunde liegende lineare Gleichungssystem eindeutig lösbar ist.

Wir können diese Beobachtung analog zu Korollar 5 im folgenden Korollar zusammenfassen.
Korollar 6 Das Richardson Verfahren (4.26) löst das Gleichungssystem (4.25) für eine bestimmte Matrix $\Phi$ und beliebige stochastische Matrizen $P$ falls ein Relaxationsparameter $\alpha$ existiert, so dass
\[
\rho \left( I - \alpha \Phi^\top (I - \gamma P) \Phi \right) < 1
\]
für alle stochastischen Matrizen $P$ existiert. Die leicht abgeschwächte Bedingung
\[
\rho \left( I - \alpha \Phi^\top (I - \gamma P) \Phi \right) \leq 1
\]
is für die Beschränktheit des Richardson Verfahrens notwendig. Die Beschränktheit des Richardson Verfahrens ist wiederum für die Beschränktheit von beliebigen Folgen von asynchronen TD[0] Aktualisierungen (4.7) notwendig.

Beweis: die Behauptung folgt aus der oben erwähnten Äquivalenz des synchronen TD[0] Verfahrens und des Richardson Verfahrens für das Gleichungssystem (4.25). Die Folgerung über die Beschränktheit für asynchrone TD[0] Aktualisierungen folgt mittels dieser Äquivalenz direkt aus Korollar 5. \hfill $\Diamond$

Die Konvergenz des Richardson Verfahrens für beliebige stochastischen Matrizen $P$ ist damit im wesentlichen davon abhängig, ob die Matrix $\Phi^\top (I - \gamma P) \Phi$ die positive Stabilität der Matrix $I - \gamma P$ erhält. Die Tabelle 4.3 zeigt, dass dies bisher nur für eine Klasse von Matrizen $\Phi$ zugesichert werden konnte. Andere geläufige Klassen scheiden dagegen aus.

Ein möglicher Ausweg ist die Minimierung des Fehlers $E(w, \tau)$ aus (4.24) mit Berücksichtigung der Tatsache, dass $\tau$ in differenzierbarer Weise von $w$ abhängt. Dies ist Gegenstand des nächsten Abschnitts.

4.5 Residuelles Gradienten Verfahren

In diesem Abschnitt minimieren wir den den Fehler $E(w, \tau)$ aus (4.24), unter der Ausnutzung der Tatsache, dass $\tau$ eine differenzierbare Funktion von $w$ ist. Dies ist im Reinforcement Lernen nur bei der Strategieauswertung der Fall, bei der Wertiteration oder beim Q-Lernen enthält $\tau$ nicht differenzierbare Anteile. Da wir das TD[0] Verfahren vornehmlich deswegen untersuchen, weil es im Zusammenhang mit Funktionsapproximation die gleichen Probleme wie bei der Wertiteration oder beim Q-Lernen aufzeigt, ist dies ein sehr spezieller Ausweg, der nicht allgemein anwendbar ist.

Nichtdestotrotz kommt dieses Verfahren unter dem Namen Residuelles Gradienten Verfahren (RG Verfahren) in der Literatur vor ([Bai95]). Es lässt sich ferner gut mittels des allgemeinen Satzes 4 analyseren und die Untersuchungen aus Abschnitt 3.6 lassen Rückschlüsse auf die Vorteile des synchronen TD[0] Verfahrens gegenüber dem RG Verfahren in Bezug auf die Konvergenzgeschwindigkeit ziehen.

Die Abhängigkeit des Zielwertes $\tau$ von der Parameter $w$ wird durch die
Funktionsschreibweise $\tau(w)$ ausgedrückt. Damit ergibt sich

$$2 \frac{\partial}{\partial w} E(w, \tau(w)) = \frac{\partial}{\partial w} \| \Phi w - (r + \gamma P \Phi w) \|^2_2$$

$$= \frac{\partial}{\partial w} \| (I - \gamma P) \Phi w - r \|^2_2$$

$$= 2 \Phi^\top (I - \gamma P)^\top (I - \gamma P) \Phi w - 2 (I - \gamma P) \Phi r$$

(4.27)

Der Gradient der konvexen Funktion $E(w, \tau(w))$ wird zu Null, wenn das Gleichungssystem

$$\Phi^\top (I - \gamma P)^\top (I - \gamma P) \Phi w = (I - \gamma P) \Phi r$$

für ein $w$ lösbar ist. Auf dieses Gleichungssystem lässt sich wieder das Richardson Verfahren anwenden, und man erhält die folgende Iteration

$$w^{k+1} = \left( I - \alpha \Phi^\top (I - \gamma P)^\top (I - \gamma P) \Phi \right) w^k - \alpha (I - \gamma P) \Phi r$$

(4.28)

Hat die Matrix $\Phi$ vollen Spaltenrang, so lässt sich die Konvergenz von (4.28) leicht nachweisen. Interessanter ist der Fall, wenn $\Phi$ nicht den vollen Spaltenrang hat. Dann ist die Matrix $\Phi^\top (I - \gamma P)^\top (I - \gamma P) \Phi$ singulär, und die Iterationsmatrix $I - \alpha \Phi^\top (I - \gamma P)^\top (I - \gamma P) \Phi$ hat für jede Lernrate $\alpha$ den Eigenwert 1.

An dieser Stelle kommt aber Satz 4 zum Einsatz, dessen Folgerungen es ist, dass allgemeine Matrixiterationen mit Iterationsmatrix $(I - \alpha C^\top C)$ und Translationsvektor $\alpha Cr$ immer für eine geeignete Lernrate $\alpha$ konvergieren. Der Grenzwert kann allerdings vom Startvektor $w^1$ abhängen.

**Korollar 7** Für eine beliebige Matrix $C \in \mathbb{R}^{n \times m}$ und $w^k \in \mathbb{R}^n$, $r \in \mathbb{R}^m$ existiert eine Intervall von positiven Lernraten $\alpha$, so dass die Iteration

$$w^{k+1} = (I - \alpha C^\top C) w^k + \alpha Cr$$

(4.29)

für jeden Startwert $w^1 \in \mathbb{R}^n$ konvergiert. Der Grenzwert hängt allerdings vom Startwert $w^1$ ab, wenn die Matrix $C^\top C$ singulär ist.

Beweis: Die Iterationsmatrix wird kurz als $M_\alpha := I - \alpha C^\top C$ bezeichnet. Die Matrix $C^\top C$ ist positiv semidefinit, d.h. sie ist symmetrisch und hat nur reelle Eigenwerte. Ist $C^\top C$ regulär, so ist der Spektralradius der Iterationsmatrix $M_\alpha$ für alle $\alpha$ aus dem offenen Intervall $(0, \rho(C^\top C))$ kleiner als 1. Damit ist die Konvergenz der Iteration unabhängig vom Startwert gesichert, vgl. Satz 4.

Schwieriger ist der Fall, wenn $C^\top C$ singulär ist. Dann ist nämlich 0 ein Eigenwert von $C^\top C$ und damit ist 1 ein Eigenwert von $M_\alpha$. Laut Satz 4 müssen daher die Bedingungen $\text{Cond}(M_\alpha)$, $\text{CondE}(M_\alpha)$ (vgl. (3.7), (3.8)) und $C^\top r \in \text{Im}(I - M_\alpha) = \text{Im}(C^\top C)$ erfüllt sein, um die bedingte Konvergenz der Iteration (4.29) zu zeigen. Bedingung $\text{Cond}(M_\alpha)$ ist für $\alpha \in (0, 2\rho(C^\top C))$ erfüllt. Bedingung $\text{CondE}(M_\alpha)$ ist erfüllt, da die Matrix $C^\top C$ diagonalisiertbar ist, woraus unmittelbar folgt, dass die Eigenräume vom $C^\top C$ den jeweiligen Haupträumen gleichen. Die Matrix $M_\alpha$ hat die gleichen Eigen- und Haupträume wie $C^\top C$, woraus die Gültigkeit von $\text{CondE}(M_\alpha)$ folgt.
4.5. RESIDUELLES GRADIENTEN VERFAHREN

Es bleibt die Bedingung $C^\top r \in \text{Im}(C^\top C)$ zu zeigen. Wir zeigen dazu $r = C^\top C x$ mit $x = C^\dagger$, wobei $C^\dagger$ die Moore-Penrose Inverse der Matrix $C$ darstellt (vgl. [Bjö96, HJ85]). Nun gilt

$$C^\top C x = (C^\dagger C)^\top C^\top r$$
$$= ((C^\dagger C)^\top C)^\top r$$
$$
\equiv (C^\dagger C)^\top r$$

wobei in ($\ast$) bzw. ($\ast\ast$) die Eigenschaften $(C^\dagger C)^\top = C^\dagger C$ bzw. $C^\dagger C = C$ der Pseudoinversen benutzt wurden (vgl. [Bjö96, HJ85]). Damit sind für $\alpha \in (0, 2p(C^\top C))$ alle drei Bedingungen $\text{Cond}(M_\alpha)$, $\text{CondE}(M_\alpha)$ und $C^\top r \in \text{Im}(I - M_\alpha)$ erfüllt. Laut Satz 4 konvergiert daher die Iteration (4.29), wobei der Grenzwert von dem jeweiligen Startwert $w^1$ abhängen kann. \hfill $\Diamond$

Die Überlegungen in Korollar 7 gelten natürlich auch für eine synchrone Form des RG Verfahrens. Die Zusammenhänge sind hier analog wie beim synchronen TD Verfahren und der Lösung der Gleichungssystem (4.25) mittels des Richardson Verfahrens. Um das synchronen RG Verfahren zu erhalten, müssen beim Gradienten (4.14) die einzelnen Zielwerte $\tau_i$ analog zu (4.27) als differenzierbare Funktionen $\tau_i(w)$ behandelt werden. Die einzelnen Schritte der Herleitung lassen wir aus Analogiegründen aus, merken nur an, dass das synchronne RG Verfahren die Form

$$w^{k+1} = \left( I - \alpha \Phi \left( I - \gamma \hat{P} \right) \hat{D} (I - \alpha \Phi) \Phi \right) w^k - \alpha \left( I - \gamma \hat{P} \right) \Phi r$$

(4.30)

hat, wobei $\hat{D}$ und $\hat{P}$ wie in (4.17) definiert sind.

Korollar 7 hat damit auch gezeigt, dass zumindest bei der Strategieauswertung das synchronne RG Verfahren eine Alternative zum synchronen TD Verfahren darstellt. Was das für die asynchrone Version der RG Verfahrens bedeutet, werden wir im folgenden Unterabschnitt erörtern.

4.5.1 Asynchrones RG Verfahren

In diesem Abschnitt gehen wir auf die asynchrone Variante des RG Verfahrens ein. Der Unterschied zwischen dem asynchronen RG Verfahren und dem asynchronen TD[0] Verfahren liegt wie schon im synchronen Fall in der Bildung des Gradienten. Wiederum wird $\tau$ in (4.5) als differenzierbare Funktion $\tau(w)$ von $w$ aufgefasst. Man erhält

$$\frac{\partial E(w, \tau(w))}{\partial w} = \left( \gamma \varphi(h) - \varphi(g) \right)^\top w + r(g) \left( \gamma \varphi(h) - \varphi(g) \right)$$

Auch hier lassen sich alle Schritte zur Entwicklung des asynchronen RG Verfahrens anhand der Entwicklung des asynchronen TD[0] Verfahrens festmachen. Aus Analogiegründen lassen wir diese erneut weg, und kommen gleich zur der asynchronen RG Aktualisierung

$$w^{k+1} = \left( I - \alpha (\gamma \varphi(h) - \varphi(g) \gamma \varphi(h) - \varphi(g))^\top \right) w^k + \alpha r(g) (\gamma \varphi(h) - \varphi(g))$$

(4.31)
KAPITEL 4. KONVERGENZ VON APPROXIMATIVEM RL

Der Zusammenhang zwischen dem asynchronen RG Verfahren und der synchronen RG Version (4.30) entspricht dem Zusammenhang zwischen dem asynchronen TD[0] Verfahren und seiner synchronen Version. Insbesondere lassen sich Satz 12 und damit auch Korollar 5 übertragen. Die Kernaussage dieser Überlegung ist letztendlich, dass die Beschränktheit des synchronen RG Verfahrens zwingend notwendig für die Beschränktheit des asynchronen Verfahrens ist. Da Korollar 7 gezeigt hat, dass das synchronen RG Verfahren stets (bedingt) konvergiert, ist diese notwendige Voraussetzung für das asynchrone RG Verfahren immer erfüllt.

Es wird jetzt gezeigt, dass das asynchrone RG Verfahren bei konstanter Lernrate divergieren kann. Dies erfordert allerdings eine sehr spezielle Folge von Aktualisierungen. Um dieses Beispiel zu konstruieren, gehen wir davon aus, dass zu jedem Einheitsvektor $d$ immer 2 Zustände $g$ und $h$ existieren so dass $d = \gamma\varphi(h) - \varphi(g)$ gilt. Dies lässt sich z.B. erreichen, wenn der Zustandsraum $S$ gleich dem Intervall $[0, 1]$ gesetzt wird, und ein 2 dimensionaler Merkmalsvektor $\varphi(s) = (\phi_1(s), \phi_2(s))^T \in \mathbb{R}^2$ mit $\phi_1(s) = \phi_2(s) = s$ für alle $s \in S$ verwendet wird. Iteration (4.31) kann dann kurz als

$$w^{k+1} = (I - \alpha dd^T)w^k + \alpha rd$$

(4.32)
gerichtet werden. Zu jedem $w^k \in \mathbb{R}^2$ existiert damit ein Einheitsvektor $d$, der zu $w^k$ orthogonal ist. Die Anwendung der Vorschrift (4.32) liefert

$$w^{k+1} = (I - \alpha dd^T)w^k + \alpha rd = w^k + \alpha rd$$

Wegen der Orthogonalität von $w^k$ und $d$ gilt ferner

$$\|w^{k+1}\|_2^2 = \|w^k + \alpha rd\|_2^2 = \|w^k\|_2^2 + (\alpha r)^2$$

Die Anwendung der Vorschrift (4.32) verlängert damit den Ausgangsvektor $w$ um einen konstanten Faktor. Bei einem festen Belohnungswert $r$ bekommt man

$$\|w^{k+1}\|_2^2 = \|w^1\|_2^2 + k(\alpha r)^2 \rightarrow \infty$$

was eine divergente Folge darstellt.

Obiges Beispiel funktioniert nicht, wenn man nur endlich viele Vektoren $d$ zur Verfügung hat, was letztendlich einem endlichen Zustandsraum $S$ entspricht. Der Grund dafür liegt darin, dass wegen der wachsenden Länge von $w^{k+1}$ gegenüber $w^k$, aber nur einer beschränkten Ablenkung in Richtung des zu $w^k$ orthogonalen Vektors $d$, die Winkeldifferenz zwischen $w^{k+1}$ und $w^k$ für steigendes $k$ immer kleiner wird.

Sind nur endlich viele verschiedene $d$ zugelassen, so wird unter bestimmten schwachen Voraussetzungen die Folge (4.32) konvergieren. Wir wollen das hier aber nicht weiter vertiefen. Es ging uns vor allem darum zu zeigen, dass selbst die Konvergenz eines synchronen Verfahrens nicht unbedingt die Beschränktheit der asynchronen Version garantieren muss. Um so bemerkenswerter ist daher die Aussage von Satz 10 wo die Beschränktheit des Produktes von unendlich vielen verschiedenen Matrizen (die natürlich einer bestimmten Struktureinschränkung unterliegen) gezeigt wurde. Dies wird im Kapitel 5 auf die Klassen von Gitterapproximatern über unendlichen Zustandsräumen angewendet, um deren Stabilität zu zeigen.

---

3 Dabei ist vorausgesetzt, dass die Belohnungswerte $r$ beschränkt sind.

4 Eine mögliche Voraussetzung ist, dass die Intervalle zwischen dem Auftreten des gleichen Vektors $d$ nicht unbeschränkt wachsen.
4.6 Konvergenzgeschwindigkeit des TD[0] und RG Verfahrens im Vergleich

In diesem Abschnitt widmen wir uns der Frage, ob das synchrone TD[0] Verfahren evtl. doch Vorteile gegenüber dem synchrone RG Verfahren hat. Insbesondere geht es um die Konvergenzgeschwindigkeit beider Verfahren. Im Abschnitt 3.6 von Kapitel 3 haben wir uns mit der Lage von Eigenwerten von Matrizen der Form (3.32) beschäftigt. Die Matrix $A_{TD}(\alpha, P)$ aus (3.32) entspricht der Iterationsmatrix des synchronen TD[0] Verfahrens (4.18) wenn die Matrizen $\Phi$ und $D$ jeweils der Identitätsmatrix $I$ gleichen. Gleiches gilt für die Matrix $A_{RG}(\alpha, P)$ aus (3.32) und das synchrone RG Verfahren (4.30). Damit lässt sich die Aussage von Satz 13 direkt auf das synchrone TD[0] bzw. RG Verfahren ohne Funktionsapproximation anwenden.

**Korollar 8** Betrachte das synchrone TD[0] Verfahren (4.18) und das synchrone RG Verfahren (4.30), in beiden Fällen seien die Approximationsmatrix $\Phi$ und die Gewichtungsmatrix $D$ gleich der Identitätsmatrix $I$. Dies entspricht dem Fall der Tabellenrepräsentation, in dem alle Zustände gleichmäßig besucht werden. Falls die Matrix $I - \gamma P$ die Voraussetzungen von Lemma 11 erfüllt, dann konvergiert das synchrone TD[0] Verfahren schneller oder im Extremfall gleich schnell wie das synchrone RG Verfahren.

Beweis: die Behauptung folgt unmittelbar aus Satz 13.

Die empirischen Untersuchungen aus Abschnitt 3.6 in Kapitel 3 lassen vermuten, dass die Klassse der Matrizen $P$, für die das Korollar 8 gültig bleibt, auf alle stochastischen Matrizen ausgedehnt werden kann. Damit ist im Fall der Tabellenrepräsentation das synchrone TD[0] Verfahren dem synchrone RG Verfahren vorzuziehen. Die Lage ändert sich, sobald $\Phi$ nicht der Identitätsmatrix entspricht. Die Konvergenz ist im TD[0] Fall nicht mehr garantiert, ferner zeigt eine Untersuchung in [Sch03], dass bei dem bekannten Beispiel von Baird [Bai95] für bestimmte (kleine) Diskontierungsfaktoren $\gamma$ das RG Verfahren schneller konvergiert.

4.7 Approximative Wertiteration

In diesem Abschnitt gehen wir auf des Verfahren der Wertiteration im Zusammenhang mit Funktionsapproximation ein. Die Vorschrift für Wertiteration (2.11) wird unter Anwendung von Funktionsapproximation zu

$$v(g; w) \leftarrow \max_{a \in A} \left( r(g, a) + \gamma \sum_{h \in S} p(h|g, a)v(h; w) \right)$$

Der Parameter $w$ kann analog zur approximativen TD[0] Aktualisierung durch Minimierung des Fehlers (4.4) mit

$$\tau = \max_{a \in A} \left( r(s, a) + \gamma \sum_{h \in S} p(h|g, a) v(h; w) \right) = \varphi(h)^T w$$
angepasst werden. Unter Anwendung des Gradienten (4.5) und von Relaxation erhält man die folgende Vorschrift für die approximative Wertiteration

\[ w^{k+1} = w^k - \alpha (\varphi(g)^T w^k - \tau) \varphi(g) \]

\[ = w^k - \alpha \left( \varphi(g)^T w^k - \max_{a \in A} \left( r(g, a) + \gamma \sum_{h \in S} p(h|g, a) \varphi(h)^T w^k \right) \right) \varphi(g) \]

\[(4.33)\]

Das Maximum in (4.33) wird für eine bestimmte Aktion \(a\) angenommen. Man erhält damit mit der abkürzenden Schreibweise \(p_h := p(h|g, a)\) und \(r := r(g, a)\)

\[ w^{k+1} = \left( I + \alpha \varphi(g) \left( \sum_{h \in S} p_h \varphi(h) - \varphi(g) \right)^T \right) w^k + \alpha r \varphi(g) \]

\[ = \sum_{h \in S} p_h \left( (I + \alpha \varphi(g) (\gamma \varphi(h) - \varphi(g))^T) w^k + \alpha r \varphi(g) \right) \]

\[(4.34)\]

Wegen \(\sum_{h \in S} p_h = 1\) entspricht die asynchrone Aktualisierung der Wertiteration mit Funktionsapproximation (4.34) einer konvexen Summe von asynchronen TD[0] Aktualisierungen (4.7). Wenn man ferner in (4.34) \(\sum_{h \in S} p_h \varphi(h)\) durch ein einzelnes \(\varphi(h)\) ersetzt, dann ergibt sich sogar direkt die asynchrone TD[0] Aktualisierung (4.7). Dies ist insbesondere bei deterministischen Systemen der Fall, so dass wir für deterministische Systeme aus der Beschränktheit des asynchronen TD[0] Verfahrens ohne weiteres Zutun auf die Beschränktheit der asynchronen Wertiteration (4.34) schließen können.

Ist andererseits für eine Klasse von \(\varphi(h) \in B \subset \mathbb{R}^n\) die Beschränktheit des asynchronen TD[0] Verfahrens gesichert, so überträgt sich dies auf die asynchrone Wertiteration, wenn aus \(\varphi(h), \varphi(h') \in B\) und \(\lambda \in [0, 1]\) immer

\[(1 - \lambda)\varphi(h) + \lambda \varphi(h') \in B\]

folgt, also wenn \(B\) eine konvexe Menge ist. Man sieht auch, dass von dieser Forderung nur die Zielzustände betroffen sind. Der Merkmalsvektor des Ausgangszustands \(g\) wird nicht weiter verändert. Diese Feststellung fassen wir im folgenden Lemma zusammen

**Lemma 13** Sei das asynchrone TD[0] Verfahren (4.7) für beliebige Zustandsübergänge \(g \sim h \in S\) beschränkt. Dann bleibt auch die asynchrone Wertiteration (4.33) beschränkt, falls die Menge der Merkmalsvektoren

\[ \varphi(S) = \{ \varphi(h) \mid h \in S \} \]

konvex ist.

Beweis: siehe obige Herleitung.

\[\diamond\]

\[\text{Man kann auch eine Teilmenge von } S \text{ nehmen, falls nicht jeder Zustand ein Zielzustand sein kann.}\]
4.8 Approximatives Q-Lernen

Ähnlich zur Herleitung der Aktualisierungsvorschrift für die approximative Wertiteration (4.33) kann man auch beim approximativen Q-Lernen vorgehen. Dazu wird eine approximierte Version \( q(\cdot; w) \) der Q-Funktion \( q(\cdot) \) benötigt

\[
q(s, a; w) = \sum_{i=1}^{m} w_i \phi_i(s, a) = \varphi(s, a)^{\top} w
\]

wobei die Basisfunktionen \( \phi_i \) diesmal nicht nur den Zustandsraum \( S \) sondern das kartesische Produkt \( S \times A \) in die reellen Zahlen abbilden. Man kann auch das kartesische Produkt von \( S \) und \( A \) als einen erweiterten Zustandsraum \( S' := S \times A \) auffassen. Die Herleitung der Aktualisierungsvorschrift für das approximative Q-Lernen ergibt sich analog zur Herleitung der TD[0] Vorschrift (4.7) oder der Vorschrift (4.33) der Wertiteration. Man erhält für einen Zustandsübergang \( g \sim h \) unter der Verwendung der Aktion \( a \) die folgende Aktualisierungsvorschrift:

\[
w_{k+1} = \left( I - \alpha \varphi(g, a) \varphi(g, a)^{\top} \right) w_k + \alpha \gamma \max_{b \in A} \left( \varphi(h, b)^{\top} w_k \right) \varphi(g, a) + \alpha r(g, a) \varphi(g, a)
\]

(4.35)

Das Maximum \( \max_{b \in A} \varphi(h, b)^{\top} w_k \) wird für eine Aktion \( c = \arg\max_{b \in A} \varphi(h, b)^{\top} w_k \) angenommen. D.h. nach Wahl dieser maximierenden Aktion \( c \) ergibt sich die Vorschrift

\[
w_{k+1} = \left( I + \alpha \varphi(g, a) (\gamma \varphi(h, c) - \varphi(g, a))^{\top} \right) w_k + \alpha r(g, a) \varphi(g, a)
\]

Diese Vorschrift gleicht nun der für das asynchrone TD[0] Verfahren (4.7). Der einzige Unterschied besteht darin, dass der erweiterte Zustandsraum \( S' := S \times A \) statt dem normalen Zustandsraum \( S \) verwendet wurde. Damit können wir die Stabilität von approximativem Q-Lernen an die Beschränktheit des TD[0] Verfahrens auf dem erweiterten Zustandsraum \( S' \) koppeln.

Lemma 14 Sei das asynchrone TD[0] Verfahren (4.7) auf dem erweiterten Zustandsraum \( S' = S \times A \) beschränkt. Dann bleibt auch das asynchrone Q-Lernen (4.35) beschränkt.

Beweis: siehe obige Herleitung.

4.9 Zusammenfassung

Beispiel 10 ist ein elementares Beispiel mit nur 3 Zuständen, 2 Basisfunktionen und 2 Zustandsübergängen, welches zeigt, dass das asynchrone TD[0] Verfahren für jede Lernrate \( \alpha \neq 0 \) divergieren kann. Dass bekannteste Beispiel für die Divergenz des TD[0] Verfahrens stammt aus [Bai95] und wurde in [Sch03] eingehend untersucht. Beispiel 10 ist viel kleiner, außerdem zeigen wir die Divergenz beim asynchronen TD[0] Verfahren, und nicht nur in der synchronen Variante wie in [Bai95, Sch03]. Allerdings motiviert dieses Beispiel die Untersuchung des synchronen TD[0] Verfahrens, da dieses besser algebraisch zugänglich ist, und wie Korollar 5 gezeigt hat, eng mit dem asynchronen TD[0] Verfahren verknüpft ist.


Im Abschnitt 4.4 haben wir einen neuen Zugang zum synchronen TD[0] Verfahren vorgestellt. Wir zeigen, dass es sich als approximierte Form des Richardson Verfahrens zur Lösung des Gleichungssystems (4.25) außen lässt. Dieses Gleichungssystem kann aufgestellt werden, wenn man die stochastische Systemmatrix \( P \) kennt. Es wurde gezeigt, dass das synchronen TD[0] Verfahren analog zu diesem Vorgehen ist, allerdings wird die typischerweise unbekannte Systemmatrix \( P \) durch ihre angenäherte Variante \( \hat{P} \) ersetzt.

In den Abschnitten 4.5 und 4.5.1 sind wir auf das RG Verfahren eingegangen. In Korollar 7 zeigen wir einen gegenüber [SM03a] vereinfachten Beweis, dass das synchrone RG Verfahren immer (bedingt) konvergiert. Wir argumentieren ähnlich wie beim TD[0] Verfahren, dass die Beschränktheit des synchronen RG Verfahrens notwendig für die Beschränktheit des asynchronen RG Verfahrens ist. Allerdings konstruieren wir ein Beispiel, in dem das asynchrone RG Verfahren mit Funktionsapproximation divergiert. Dieses Beispiel zeigt auf, dass es durchaus noch Untersuchungsbedarf gibt, um die Verbindung zwischen

- Im Abschnitt 4.6 sind wir auf die Konvergenzgeschwindigkeit des synchronen TD$[0]$ und RG Verfahrens eingegangen. Unter Anwendung der Ergebnisse aus Abschnitt 3.6 zeigen wir für eine etwas größere Klasse von Systemen als in [Sch03, SM03b], dass das synchrone TD$[0]$ Verfahrens nachweislich schneller als das RG Verfahren konvergiert. Eine weitere Verallgemeinerung dieses Ergebnisses wurde anhand der empirischen Untersuchungen in Abschnitt 3.6 vermutet, konnte aber bisher nicht nachgewiesen werden.

- In den Abschnitten 4.7 und 4.8 sind wir auf mögliche Konsequenzen eingegangen, die die Beschränktheit des approximativen TD$[0]$ Verfahrens auf die Verfahren der Wertiteration und des Q-Lernens hat. Für die Wertiteration haben wir gezeigt, dass wenn die Merkmalsvektoren eine bestimmte Eigenschaft haben (vgl. Korollar 13), dann folgt aus der Beschränktheit des TD$[0]$ Verfahrens auch die Beschränktheit des Verfahrens der Wertiteration. Die geforderte Eigenschaft hat gewisse Ähnlichkeit mit der Eigenschaft für Approximatore, die für die so genannte fitted Wertiteration in [Gor95] gefordert werden. Allerdings folgt bei uns aus dieser Eigenschaft die Beschränktheit der asynchronen Wertiteration. In [Gor95] wird dagegen für eine spezielle Form der Wertiteration (fitted value iteration) die Konvergenz gezeigt wird.
Kapitel 5

Gitterbasierte Approximatoren


In Abbildung 5.1 sehen wir die Hierarchie der von uns betrachteten Gitterapproxim atoren. Der einfachste Fall ist der stückweise konstante Gitterapproximator. Für diesen Typ von Gitterapproxim atoren werden wir zeigen, dass sowohl das asynchrone TD[0] als auch das Verfahren der Wertiteration bei konstanter Lernrate beschränkt bleiben.

Eine Verallgemeinerung von stückweise konstanten Gitterapproxim atoren bilden

<table>
<thead>
<tr>
<th>Abschnitt 5.1</th>
<th>Abschnitt 5.2</th>
<th>Abschnitt 5.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stückweise konstante Gitterapproxim atoren</td>
<td>Stückweise lineare Gitterapproxim atoren</td>
<td>CMAC Approximator</td>
</tr>
</tbody>
</table>

Abbildung 5.1: Verschiedene gitterbasierte Approxim atoren.
5.1 STÜCKWEISE KONSTANTE GITTERAPPROXIMATOREN

Abbildung 5.2: Voronoi Kachelung eines regelmäßigen Gitters in $\mathbb{R}^2$.

stückweise lineare oder allgemeiner stückweise stetige Gitterapproximatoren. Bei dieser Klasse von Gitterapproximatoren kann das synchrone TD[0] und damit auch das asynchrone TD[0] Verfahren divergieren. Allerdings kann man bei einer Einschränkung der Menge von Startzuständen auf die Gitterstützstellen Divergenz sowohl im asynchronen TD[0] Fall als auch bei der Wertiteration vermeiden.


Wir werden die Gitterapproximatoren ausgehend von der allgemeinen Form eines linearen Funktionsapproximators $\phi : S \rightarrow \mathbb{R}$

$$\phi(x) = \sum_{i=1}^{m} w_i \phi_i(x)$$  \hspace{1cm} (5.1)

mit den Basisfunktionen $\phi_i : S \rightarrow \mathbb{R}$, behandeln. Dies ist insbesondere bei den stückweise linearen Gitterapproximatoren etwas gewöhnungsbedürftig, erlaubt uns aber die volle Bandbreite unserer Ergebnisse auszuschöpfen.

Einem Gitterapproximator liegt immer eine Kachelung oder Triangulierung des Zustandsraumes $S$ zugrunde, wobei der Zustandsraum $S$ hier als eine Teilmenge des $\mathbb{R}^n$ anzusehen ist. Zu jeder Basisfunktion $\phi_i$ korrespondiert ein ausgewählter Punkt $p_i$ des $\mathbb{R}^n$. In der Praxis liegen die Punkte $p_i$ oft auf einem regelmäßigen Gitter, was aber keine zwingende Voraussetzung ist.

5.1 Stückweise konstante Gitterapproximatoren

Der einfachste Gitterapproximator ist stückweise konstant, und zwar auf den Voronoizellen der Punkte $p_i$. Eine Voronoizelle $V(p_i)$ ist definiert durch

$$V(p_i) = \{ x \in \mathbb{R}^n \mid ||x - p_i|| < ||x - p_j|| \text{ für } i \neq j \} \cup$$
$$\{ x \in \mathbb{R}^n \mid ||x - p_i|| \leq ||x - p_j|| \text{ für } i < j \}$$  \hspace{1cm} (5.2)
es ist also die Menge aller \( x \in \mathbb{R}^n \), die näher an dem Punkt \( p_i \) als an einem der anderen Punkte \( p_j \) liegen. Bei gleichem Abstand erhält als Konvention der Punkt mit dem kleineren Index den Vorzug. Die Norm, die in (5.2) verwendet wird ist frei wählbar, in den meisten Fällen wird es sich aber um die Euklidische Norm \( || \cdot ||_2 \) handeln. In Abbildung 5.2 ist sind die Voronoi Zellen für ein regelmäßiges Gitter im \( \mathbb{R}^2 \) skizziert.

Aus der Definition von Voronoizellen (5.2) ist klar, dass in Bezug auf eine Menge von Punkten \( \{p_1, \ldots, p_m\} \) je zwei Voronoizellen disjunkt sind: \( V(p_i) \cap V(p_j) = \emptyset \) falls \( i \neq j \). Ferner ist \( \mathbb{R}^n = \bigcup_{i=1}^{n} V(p_i) \). Wir definieren nun

\[
\phi_i(x) = 1_{V(p_i)} = \begin{cases} 
1 & x \in V(p_i) \\
0 & x \notin V(p_i)
\end{cases}
\]

Die Funktion \( \phi = \sum_{i=1}^{m} w_i \phi_i \) ist damit eine stückweise konstante Funktion. Im eindimensionalen Fall \( \mathbb{R}^1 \) handelt es sich um Treppenfunktionen. In Abbildung 5.3 ist eine solche Treppenfunktion dargestellt.

### 5.1.1 Eigenschaften beim approximativen RL

Für den Merkmalsvektor \( \varphi(x) = (\phi_1(x), \ldots, \phi_m(x))^{\top} \) (vgl. (4.3)) eines stückweise konstanten Approximators gilt

\[
\phi_i(x) \in \{1, 0\} \text{ und } \sum_{i=1}^{m} \phi_i(x) \leq 1
\]

Der Merkmalsvektor \( \varphi(x) \) hat damit an maximal einer Stelle den Wert 1, alle anderen Stellen sind gleich 0. Die Anwendung einer asynchronen TD[0] Aktualisierung (4.7) auf einen Zustandsübergang \( g \rightarrow h \) mit Belohnungssignal \( r(g) \) produziert

\[
w = \left( I + \alpha \varphi(g)(\gamma \varphi(h) - \varphi(g))^{\top} \right) w + \alpha r(g)\varphi(g)
\]

Man kann obige Vorschrift auch in homogener Form schreiben. So kommt man auf Iterationsmatrizen, die den Voraussetzungen von Satz 10 genügen, da ein
5.1. STÜCKWEISE KONSTANTE GITTEPARPROXIMATOREN

Abbildung 5.4: Stückweise lineare Funktion von einer Treppenfunktion.

Merkmalsvektor \( \varphi(h) \) maximal an einer Stelle den Eintrag 1 hat, und ansonsten nur 0 Einträge enthält. Damit ist die Beschränktheit des Vektors \( w \) auch bei beliebigen Anwendungen der TD[0] Vorschrift gesichert, was wir im folgende Korollar zusammenfassen

**Korollar 9** Das asynchrone TD[0] Verfahren (4.7) bleibt bei der Anwendung von stickweise konstanten Gitterapproximatoren für Lernraten \( \alpha \in [0, 1) \) beschränkt.

Beweis: siehe obige Herleitung und Satz 10.

Korollar 9 garantiert damit für stickweise konstante Gitterapproximatoren Beschränktheit im allgemeinsten Fall, nämlich dem asynchronen TD[0] Verfahren. Damit ist gleichzeitig klar, dass auch das synchrone TD[0] Verfahren beschränkt bleiben muss. Die Beschränktheit im synchronen Fall kann man aber auch unabhängig von Korollar 9 bzw. Satz 10 aus Satz 7 ableiten. Wenn man nämlich die Merkmalsvektoren entsprechend (4.22) bzw. (4.22) zu einer Approximatormatrix \( \Phi \) zusammenfasst, dann erfüllt diese die Voraussetzungen von Satz 7.

Kehren wir aber nochmal zum Korollar 9 zurück. Satz 10 erlaubt es, dass die Merkmalsvektoren \( \varphi(h) \) der Zielzustände \( h \) auch beliebige nichtnegative Einträge haben, so lange deren Summe kleiner oder gleich 1 ist. In unserem Fall sind also auch konvexe Kombinationen der Merkmalsvektoren \( \varphi(h) \) von Satz 10 als neuer Merkmalsvektor eines Zielzustands erlaubt. Wenn man sich die Herleitung von Lemma 13 anschaut, dann reicht diese Bedingung aus, um die Beschränktheit des asynchronen Wertiterations Verfahrens zu zeigen.

**Korollar 10** Das asynchrone Verfahren der Wertiteration (4.34) bleibt bei der Anwendung von stickweise konstanten Gitterapproximatoren für Lernraten \( \alpha \in [0, 1) \) beschränkt.


Die Aussagen von Korollar 9 und 10 zeigen, dass stickweise konstante Gitterapproximatoren beim approximativen RL keine Probleme bereiten. Das Problematische an ihnen ist eher die Tatsache, dass man anhand der Funktionswerte

5.2 Stückerweise lineare Gitterapproximatoren

Wie demonstrieren zuerst die Idee im eindimensionalen Fall, und verallgemeinern diese später auf höhere Dimensionen. Im eindimensionalen Fall verbinden wir einfach die Funktionswerte an den Stellen \( p_i \) mit Liniensegmenten, und definieren damit eine neue Funktion, siehe Abbildung 5.4. Dieses Vorgehen kann man auch mit Hilfe von Basisfunktionen ausdrücken. Man ersetzt die Indikatorfunktion \( 1_{V(p_i)} \), durch eine Hutfunktion

\[
\phi_i(x) = \begin{cases} 
\frac{x-p_{i-1}}{p_i-p_{i-1}} & \text{falls } i > 1 \text{ und } x \in [p_{i-1}, p_i] \\
\frac{p_{i+1}-x}{p_{i+1}-p_i} & \text{falls } i < m \text{ und } x \in [p_i, p_{i+1}] \\
0 & \text{sonst}
\end{cases}
\]

Diese Definition gilt für alle Basisfunktionen \( \phi_i \) mit \( 1 \leq i \leq m \). Die am weitesten links stehende Basisfunktion \( \phi_1 \) wird dabei nur durch den rechten Teil einer Hutfunktion definiert, links von \( p_1 \) ist die Funktion gleich 0. Analog verhält es sich mit der Hutfunktion \( \phi_m \), alle anderen Basisfunktionen werden durch vollständige Hutfunktionen repräsentiert. In Abbildung 5.5 haben wir als Beispiel die Transformation der Indikatorfunktion \( 1_{V(p_2)} \) zu der korrespondierenden Hutfunktionen \( \phi_2 \) dargestellt.

Wir werden jetzt zu stückweise linearen Approximatoren im \( \mathbb{R}^n \) übergehen. Auch hier gehen wir von einer Menge von Punkten \( P = \{p_1, \ldots, p_m\} \subset \mathbb{R}^n \) aus, allerdings werden die Punkte so gewählt, dass sie auf den Eckpunkten von \( n \) dimensionalem Simplex liegen. Der Approximator selbst wird auf der konvexen Hülle \( \text{conv}(P) \) der Punkte in \( P \) definiert.

Ein \( n \) dimensionales Simplex ist gegeben durch die konvexe Hülle von \( n + 1 \) affin unabhängigen Punkten. Dabei sind \( n + 1 \) Punkte \( \{q_1, \ldots, q_{n+1}\} \) affin unabhängig, falls aus

\[
\sum_{i=1}^{n+1} \alpha_i q_i = 0 \quad \text{und} \quad \sum_{i=1}^{n+1} \alpha_i = 0, \quad \alpha_i \in \mathbb{R}
\]

immer \( \alpha_i = 0, \ i = 1, \ldots, n + 1 \) folgt. Obige Bedingung ist äquivalent dazu, dass die Menge von Vektoren \( \{q_1 - q_{n+1}, \ldots, q_n - q_{n+1}\} \) linear unabhängig ist, also eine Basis.
5.2. STÜCKWEISE LINEARE GITTERAPPROXIMATOREN

des \( \mathbb{R}^n \) darstellt (vgl. [HUL01]). Dies erklärt auch, warum man von \( n \) dimensional
en Simplexen spricht. Ein Simplex, das durch \( Q = \{ q_1, \ldots, q_{n+1} \} \) aufgespannt wird,
bezeichnen wir mit \( S(q_1, \ldots, q_{n+1}) \) oder \( S(Q) \). Mehrere \( n \) dimensionale Simplexe, die
die konvexe Hülle \( \text{conv}(P) \) von \( P \) überdecken, und von Punkten aus \( P \) aufgespannt
werden, werden eine Triangulierung von \( P \) genannt. Eine solche Triangulierung muss
bestimmte zusätzliche Bedingungen erfüllen, die wir in der folgenden Definition
zusammenfassen.

**Definition 4** Eine Triangulierung \( S = \{ S_1, \ldots, S_k \} \) der Punkte \( P = \{ p_1, \ldots, p_m \} \)
ist eine Überdeckung der konvexen Hülle \( \text{conv}(P) \) von \( P \) durch Simplexe \( S \) aus \( S \)

\[
\text{conv}(P) = \bigcup_{S \in S} S_i
\]

so dass

1. Jedes Simplex \( S \in S \) ist \( n \) dimensional, und wird durch Punkte aus \( P \)
aufgespannt, d.h. \( S = S(q_1, \ldots, q_{n+1}) \) und die Menge \( \{ q_1, \ldots, q_{n+1} \} \subset P \) ist
   affin unabhängig.

2. Zu jedem Punkt \( p \in P \) existiert mindestens ein Simplex \( S \in S \), das durch
   \( p \) und \( n \) weitere Punkte \( \{ q_1, \ldots, q_n \} \subset P \) aufgespannt wird, d.h. \( S = S(q_1, \ldots, p, \ldots, q_n) \).

3. Die Schnittmenge von je zwei Simplexen enthält kein \( n \)-dimensionales Simplex.
   Insbesondere ist \( S_i \cap S_j \notin S \), für \( i \neq j \).

\[ \Diamond \]

Liegt ein Punkt \( p \) in einem Simplex \( S = S(q_1, \ldots, q_{n+1}) \) so kann er eindeutig durch
die konvexe Kombination

\[
p = \sum_{i=1}^{n+1} \lambda_i q_i, \quad \lambda_i \geq 0 \text{ und } \sum_{i=1}^{n+1} \lambda_i = 1
\]
dargestellt werden. Die Koeffizienten \( \lambda_i \) heißen auch baryzentrische Koordinaten von
\( p \) in \( S \).

Falls eine Triangulierung \( S \) von \( P \) gegeben ist, so kann man jedem \( p_i \in P \) eine
Basisfunktion \( \phi_i : \text{conv}(P) \rightarrow \mathbb{R} \) zuordnen. Wir definieren dazu für ein \( x \in \text{conv}(P) \)
die Menge \( S(x) \subset S \) als die Menge der Simplexe die \( x \) enthalten. Insbesondere ist
\( S(p_i) \) die Menge der Simplexe \( S \) in \( S \), die von \( p_i \) aufgespannt werden, d.h. für die
\( S = S(p_1, q_1, \ldots, q_n) \) mit \( \{ q_1, \ldots, q_n \} \subset P \) gilt. Nun kann man die zu \( p_i \) gehörige
Basisfunktion definieren. Als Trägermenge hat \( \phi_i \) die Vereinigung aller Simplexe, die
\( p_i \) enthalten

\[
\text{supp}(\phi_i) = \bigcup_{S \in S(p_i)} S
\]

Außerhalb von \( \text{supp}(\phi_i) \) ist \( \phi_i \) identisch 0. Liegt dagegen ein \( x \in \text{supp}(\phi_i) \), so existiert
ein Simplex \( S(q_1, \ldots, q_{n+1}) \in S(p_i) \), so dass \( x \) in \( S \) liegt. Dieses Simplex hat \( p_i \) als
KAPITEL 5. GITTERBASIERTE APPROXIMATOREN

Abbildung 5.6: Trägermenge einer Basisfunktion über 5 Simplexen.

einen seiner Eckpunkte, d.h. \( p_i = q_k \) für ein bestimmtes \( k \). Der Punkt \( x \) lässt sich als konvexe Kombination dieser Eckpunkte darstellen

\[
x = \sum_{j=1}^{n+1} \lambda_j q_j, \quad p_i = q_k
\]

\( \phi_i(x) \) wird nun gleich \( \lambda_k \) gesetzt, also gleich der baryzentrischen Koordinate, die \( x \) in Bezug auf \( p_i \) hat. Zusammenfassend erhält man

\[
\phi_i(x) = \begin{cases} 
\lambda_k & \text{falls } x = \sum_{j=1}^{n+1} \lambda_j q_j, \quad x \in S(q_1, \ldots, q_n = p_i, \ldots, q_{n+1}) \\
0 & \text{sonst}
\end{cases} \quad (5.3)
\]

Man kann sich leicht überlegen, dass die Definition in (5.3) wohldefiniert ist, d.h. falls \( x \in S = S(p, q_1, \ldots, q_n) \in \mathcal{S} \) und \( x \in S' = S(p, q_1', \ldots, q_n') \in \mathcal{S} \), dann hat \( x \) in Bezug auf \( p \) jeweils die gleiche baryzentrische Koordinate. Der Punkt \( x \) liegt dann nämlich in dem Simplex \( S \cap S' = S(p, r_1, \ldots, r_l) \) mit \( \{r_1, \ldots, r_l\} \subset \{q_1, \ldots, q_n\} \cap \{q_1', \ldots, q_n'\} \). Die Dimension von diesem Simplex ist \( l < n \). In jedem der Simplexe \( S, S' \) und \( S \cap S' \) hat \( x \) in Bezug auf Punkte, die nicht zu \( \{p, r_1, \ldots, r_l\} \) gehören, die baryzentrische Koordinate 0. In Bezug auf \( \{p, r_1, \ldots, r_l\} \) sind die baryzentrischen Koordinaten eindeutig. Insgesamt ergibt sich die Wohldefiniertheit der Basisfunktion. Mit obiger Herleitung ist auch die Stetigkeit der Basisfunktionen leicht herleitbar. Auf jedem Simplex ist die Funktion linear und daher stetig, und auf einer Schnittmenge von zwei Simplexen stimmen die beiden linearen Funktionen überein.


In der Praxis verwendet man meistens Punkte, die an den Koordinatenachsen ausgerichtet sind, und somit ein achsenparalleles Gitter bilden. Das hat den Vorteil, dass zum Ausrechnen der baryzentrischen Koordinaten kein aufwendiges
5.2. STÜCKWEISE LINEARE GITTERAPPROXIMATOREN

Abbildung 5.7: Trägermenge einer Basisfunktion in einer Kuhn-Triangulierung.


Bei der Kuhn-Triangulierung definiert das durch die Punkte \( P = \{ p_1, \ldots, p_m \} \) aufgespannte achsenparallele Gitter ebenfalls achsenparallele Quader, deren Eckpunkte in \( P \) liegen. Die Zugehörigkeit eines Punktes \( x = (x_1, \ldots, x_n)^T \) zu einem bestimmten Quader lässt sich einfach anhand der Koordinaten \( x_i \) festlegen. Jeder Quader besteht allerdings aus \( n! \) Kuhn-Simplexen und es bleibt zu klären, wie man zu einem Punkt \( x \) die Eckpunkte des zugehörigen Kuhn-Simplexes findet. OBdA. sei \( x \in [0, 1]^n \), es liegt also im Einheitsquader \([0, 1]^n \) mit den Eckpunkten \( q \in Q = \{0,1\}^n \). Jedes Kuhn-Simplex hat die Punkte \((0, \ldots, 0)^T\) und \((1, \ldots, 1)^T\) als Eckpunkte. Die Permutation \( \pi : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\} \) mit

\[
\pi(i) = \left\{ j \mid x_j > x_i \text{ oder } x_j \geq x_i \land j \geq i \right\}
\]

ordnet der größten Koordinate von \( x \) den Wert 1 zu, der zweitgrößten den Wert 2 zu usw., bis die kleinste Koordinate den Wert \( n \) zugeordnet bekommt. Bei Gleichheit von zwei Koordinaten entscheidet der Index selbst.

Ein \( q = (q(1), \ldots, q(n))^T \in Q \) gehört nun zum Kuhn-Simplex von \( x \), falls aus \( q(i) < q(j) \) immer \( \pi(i) > \pi(j) \) folgt. Man kann das auch so auffassen, dass ausgehend von \((0, \ldots, 0)^T\) erst der Index der größten \( x \) Koordinate (also \( \pi^{-1}(1) \)) auf 1 verändert wird, dann wird zusätzlich der Index der zweitgrößten Koordinate (also \( \pi^{-1}(2) \)) auf 1 verändert usw. Nach \( n \) Schritten gelangt man so immer zum letzten Eckpunkt \((1, \ldots, 1)^T\). Die so entstandene Folge bezeichnen wir mit \( \{r_1, \ldots, r_{n+1}\} \subset Q \). Es bleibt zu zeigen, dass \( x \) in dem Simplex \( S(r_1, \ldots, r_{n+1}) \) liegt. Setze dazu \( \alpha_{x(i)} = x_i \),

\footnote{Dieses Problem lässt sich auch durch das Vorberechnen der entsprechenden inversen Matrizen für jedes Simplex effizient lösen.}
Abbildung 5.8: Visualisierung von Beispiel 10 mit Hilfe von 2 Übergängen auf einem stückweise linearen Gitterapproximator mit 2 Basisfunktionen $\phi_1$ und $\phi_2$.

Nach (5.4) gilt $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$. Aus der Definition der Punkte $r_i$ sieht man

$$x = r_1 + \alpha_1 (r_2 - r_1) + \alpha_2 (r_3 - r_2) + \ldots + \alpha_n (r_{n+1} - r_n)$$

$$= (1 - \alpha_1) r_1 + (\alpha_1 - \alpha_2) r_2 + \ldots + (\alpha_{n-1} - \alpha_n) r_n + \alpha_n r_{n+1}$$

Da $\alpha_i \geq \alpha_{i+1}$ gilt, sind alle $\lambda_i \geq 0$, ferner bildet die Summation über alle $\lambda_i$ eine Teleskopsumme, so dass $\sum_{i=1}^{n+1} \lambda_i = 1$ gilt. Damit ist $x$ als eine konvexe Kombination der Punkte $\{r_1, \ldots, r_{n+1}\}$ darstellbar, und liegt somit in dem Simplex $S(r_1, \ldots, r_{n+1})$.

5.2.1 Eigenschaften beim approximativen RL

Wie bereits erwähnt haben stückweise lineare Gitterapproximatore Vorteile gegenüber stückweise konstanten Gitterapproximatoren, da der Approximator innerhalb der Simplexe linear und damit differenzierbar ist. Damit lässt sich für die meisten Punkte eine Steigung der approximierten Wertfunktion ablesen. Leider sind die Eigenschaften in Bezug auf Stabilität beim RL bei stückweise linearen Gitterapproximaten nicht so gut wie im stückweise konstanten Fall. Bereits das Einführungsbeispiel 10 im Kapitel 4 hat gezeigt, dass sowohl das asynchrone als auch das synchrone TD[0] Verfahren bei einem System mit lediglich 3 Zuständen und 2 Übergängen für alle $\alpha \neq 0$ divergiert. Dass dieses Beispiel als stückweise linearer Gitterapproximator aufgefasst werden kann, ist in Abbildung 5.8 visualisiert.

Es stellt sich nun die Frage, ob man abgeschwächte Aussagen für die Beschränktheit des stückweise linearen Gitterapproximators machen kann. Wenn man sich nur auf die Stützstellen der Basisfunktionen beschränkt, dann entspricht ein stückweise linearer Gitterapproximator genau einem stückweise konstanten Gitterapproximator. D.h. wenn man die Übergänge so einschränkt, dass sie auf die Stützstellen des Approximators fallen, dann ist laut Korollar 9 auf jeden Fall die Beschränktheit beim asynchronen TD[0] Verfahren gesichert. Sofern man sich die Ausgangszustände $g$ bei Übergängen $g \cap h$ noch einigermaßen aussuchen kann, so ist es für die Zielzustände i.a. unmöglich, da diese von dem jeweiligen Systemmodell abhängen. Das folgende Korollar zeigt aber, dass man genau diesen Fall noch abdecken kann.
Korollar 11 Das asynchrone TD[0] Verfahren (4.7) als auch das asynchrone Verfahren der Wertiteration (4.34) bleiben bei der Anwendung von stückweise linearen Gitterapproximatoren für genügend kleine Lernraten α beschränkt, falls die Ausgangszustände \( q \) von Übergängen \( g \sim h \) auf Stützstellen der Basisfunktionen des Approximators liegen. Die Lage der Zielzustände \( h \) ist dagegen für die Beschränktheit unerheblich.

Beweis: Ähnlich wie bei der Herleitung von Korollar 9 und Korollar 10 kann man direkt Satz 10 einsetzen um die Behauptung zu zeigen. \( \diamond \)

5.3 CMAC Approximator

In diesem Abschnitt betrachten wir den CMAC Approximator (Cerebellar Model Articulator Controller) [Alb81, Sut96]. Er zeichnet sich dadurch aus, dass mehrere stückweise konstante Gitterapproximatoren überlagert und miteinander kombiniert werden. In Abbildung 5.9 sehen wir einen Vergleich zwischen einem stückweise konstanten Gitterapproximator und dem CMAC Approximator. Der CMAC Approximator verwendet in diesem Beispiel zwei Gitterapproximatoren, die jeweils nur eine Dimension unterteilen, entlang der anderen aber keine Unterteilung vornehmen. Insgesamt kommt man so auf 6 Basisfunktionen \( \{\phi_1, \ldots, \phi_6\} \) wohingegen der stückweise konstante Gitterapproximator 9 Basisfunktionen benötigt \( \{\psi_1, \ldots, \psi_9\} \). Besonders bei höheren Dimensionen macht sich das bemerkbar: der stückweise konstante Gitterapproximator benötigt exponentiell viele Basisfunktionen, beim CMAC Approximator hängt die Anzahl der Basisfunktionen linear von der Dimension ab. Diese Ersparnis wird durch Korrelation zwischen den verschiedenen Basisfunktionen erreicht, und wir werden zeigen, dass dies zur Divergenz führen kann.

In [MS04, Sch03] wurde gezeigt, dass eine einzelne immer wieder wiederholte TD[0] Aktualisierung im Zusammenhang mit dem CMAC Approximator konvergiert. Dieses Resultat lässt sich auch aus Lemma 3 ableiten. Damit scheint der CMAC Approximator als potenzieller Kandidat, um auch im Fall von asynchronem TD[0] Verfahren stabil zu sein. Wir werden jetzt zeigen, dass dies nicht der Fall ist. Wir werden nämlich ein einfaches Beispiel vorstellen, für das es beim synchronen TD[0] Verfahren zusammen mit dem CMAC Approximator zur Divergenz kommt. Damit wird laut Korollar 5 auch das asynchrone TD[0] Verfahren für bestimmte Folgen von Aktualisierungen divergieren. In Abbildung 5.10 ist unser Beispiel dargestellt. Der CMAC Approximator hat lediglich zwei Basisfunktionen \( \phi_1 \) und \( \phi_2 \). Beim synchronen TD[0] Verfahren kommen drei Zustände \( j, g \) und \( h \) in drei Übergängen vor

\[
j \sim h, \quad g \sim h \quad \text{und} \quad h \sim h
\]

Für die Merkmalsvektoren der Zustände \( j, g \) und \( h \) gilt

\[
\varphi(j) = (\phi_1(j), \phi_2(j))^\top = (1, 0)^\top \\
\varphi(g) = (\phi_1(g), \phi_2(g))^\top = (0, 1)^\top \\
\varphi(h) = (\phi_1(h), \phi_2(h))^\top = (1, 1)^\top
\]
Abbildung 5.9: Vergleich zwischen einem stückweise konstanten Gitterapproximator und dem CMAC Approximator.

womit sich für die Approximatorenmatrix Φ laut (3.26) bzw. (4.22) die Darstellung

\[ Φ = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \]

ergibt. Die Übergänge \( j \sim h, g \sim h \) und \( h \sim h \) implizieren laut (4.17) die Matrizen \( \hat{D} \) und \( \hat{P} \) mit

\[ \hat{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{und} \quad \hat{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \]

Im Beispiel 5 haben wir aber gesehen, dass bei dieser Wahl der Matrizen \( \hat{D}, \hat{P} \) und Φ die Matrix \( Φ^T \hat{D}(I - γ\hat{P})Φ \) sowohl einen negativen als auch einen positiven Eigenwert hat. Damit ist aber für jede Lernrate \( α \neq 0 \) der Spektralradius der Matrix \( I - αΦ^T \hat{D}(I - γ\hat{P})Φ \) größer als 1, was laut Korollar 4 die Divergenz des synchronen TD[0] Verfahrens für diese Wahl von Übergängen bedeutet. Wie bereits erwähnt, existieren damit laut Korollar 5 auch für jede Lernrate \( α \neq 0 \) divergente Folgen von asynchronen TD[0] Aktualisierungen. An dieser Stelle sieht man erneut die Nützlichkeit der Herangehensweise zur Untersuchung des asynchronen TD[0] Verfahrens über das synchrone TD[0] Verfahren. Im synchronen Fall muss man lediglich die Eigenwerte einer Matrix untersuchen, im asynchronen Fall ist es nicht trivial zu zeigen, dass für alle Lernraten \( α \neq 0 \) immer eine divergente Folge von asynchronen TD[0] Aktualisierungen existiert.
5.4 Zusammenfassung


- Für die stückweise linearen Gitter konnte nur eine abgeschwächte Aussage über die Verträglichkeit mit dem asynchronen TD[0] Verfahren gegeben werden. Nur für Übergänge mit Startzuständen auf Stützstellen von Basisfunktionen lässt sich divergentes Verhalten des asynchronen TD[0] Verfahrens vermeiden. Allerdings müssen die Zustände nicht einer endlichen und einmal festgelegten Menge entstammen. Unter der gleichen Einschränkung wie für das asynchrone TD[0] Verfahren bleibt auch das Verfahren der Wertiteration beschränkt. Die erzielten Resultate weisen durchaus Ähnlichkeit zu den Ergebnissen aus [Gor95, Gor99], in denen so genannte Averager als Funktionsapproximatoren studiert werden. In [Gor95, Gor99] wurde allerdings das Verfahren der Wertiteration...
modifiziert (fitted value iteration), um einen Konvergenznachweis zu führen. Unser Ergebnis bezieht sich auf das normale Verfahren der Wertiteration, mit der oben erwähnten Einschränkung in Bezug auf die Startzustände.

Kapitel 6

Zusammenfassung und Ausblick


Das approximative TD[0] Verfahren wurde bereits in [MS04] als Folge von speziellen affinen Abbildungen (so genannten schießen Projektionen) aufgefasst. In dieser Arbeit studieren wir diese affinen Abbildungen zum ersten Mal in homogener Form, und schlagen damit eine Brücke zum relativ jungen Gebiet von unendlichen Matrixprodukten [DL92, BW92, BE97]. Diese Verbindung erweist sich als sehr fruchtbar, zeigt aber auch eventuelle Grenzen, die bei der Untersuchung der Beschränktheit von approximativen RL gesetzt sind, da sich ausgerechnet die Frage nach der Beschränktheit von unendlichen Matrixprodukten als unentscheidbar erweist [BT00].

Als erstes positives Resultat der Verbindung zu unendlichen Matrixprodukten ist Satz 10 anzusehen. Es gelingt hier eine ganze Klasse von speziellen schießen


Diese Arbeit hat viele neue Aspekte bei der Untersuchung von approximativem RL eingebracht, und ermöglicht damit zukünftig weiterführende Untersuchungen. Nachdem die Notwendigkeit der Beschränktheit beim synchronen TD[0] Verfahren für die Beschränktheit der asynchronen Variante gezeigt werden konnte, wäre es interessant zu wissen, ob man unter Umständen auch hinreichende Bedingungen für die Beschränktheit des asynchronen TD[0] Verfahrens aufstellen kann, die sich rein algebraisch nachweisen lassen.


Das weiter oben zitierte Ergebnis von Korollar 2, welches die Beschränktheit des synchronen TD[0] Verfahrens für bestimmte Systeme unabhängig vom
Literaturverzeichnis


LITERATURVERZEICHNIS


