Das Verfahren der sukzessiven Überrelaxation (SOR–Verfahren) bei periodischen Markov–Ketten

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik der

Universität Karlsruhe

genehmigte

DISSERTATION

von

Dipl.–Math.oec. Grischa Markus Freimann

aus Lörrach

Tag der mündlichen Prüfung: 08. Dezember 1999
Referent: Prof. Dr. W. Niethammer
Korreferent: Prof. Dr. R. Scherer

Ein besonderes Anliegen ist es mir, an dieser Stelle all denen meinen Dank auszudrücken, die durch ihre vielfältige Unterstützung das Entstehen dieser Arbeit überhaupt erst ermöglichten.

Besonders danken möchte ich Herrn Prof. Dr. W. Niethammer für die Betreuung dieser Arbeit. Er war für mich immer ansprechbar und bereit, mich nach Kräften zu unterstützen. Das äußerst angenehme Klima an dem von ihm geleiteten Institut war wesentliche Voraussetzung für einen erfolgreichen Abschluss der Arbeit.

Herrn Prof. Dr. R. Scherer danke ich für die freundliche Übernahme des Korreferates und für seine wertvollen Ratschläge vor dem Kolloquiumstermin.

Meinen jetzigen und ehemaligen Kollegen danke ich für die angenehme Atmosphäre und das Gemeinschaftsgefühl, das sie vermittelt haben.

Meinem Mathematiklehrer am Gymnasium, Herrn W. Haubensak, danke ich dafür, dass er mir das Interesse an der Mathematik vermittelt hat, das bis heute nicht erlahmt ist.

Der Kronmüller–Stiftung danke ich für die Förderung dieser Arbeit durch Gewährung eines zwölfmonatigen IPP–Stipendiums.

Meinen Freunden und Bundesbrüdern danke ich für viele schöne Stunden, die gerade in anstrengenden Zeiten in angenehmster Weise etwas von der Arbeit abgelenkt haben.

Der letzte und größte Dank gilt meinen Eltern, die mich auf meinem Weg immer wohlwollend begleitet und nach Kräften unterstützt haben.

Inhaltsverzeichnis

1 Einleitung

Die mathematische Modellierung unterschiedlichster Problemstellungen in vielerlei Wissenschaften führt auf sogenannte Markov–Ketten. Dabei ergibt sich ein System aus nicht nowendigerweise endlich vielen Zuständen, bei dem die Übergänge von einem Zustand in einen anderen nur vom augenblicklichen Zustand abhängen, nicht hingegen von früheren. Eine solche Modellierung findet man bei Warteschlangen (vgl. [57, Kapitel 9]) ebenso wie bei einer Reihe soziologischer Fragestellungen. Geburts–, Todes– und Immigrationsprozesse (vgl. [62]) lassen sich dadurch ebenso modellieren wie Untersuchungen der Markenwahl im Bereich des Marketing (vgl. [13]) oder Fragestellungen in der Verhaltensforschung ([35]). Mit der Verallgemeinerung zu Markov–Ketten höherer Ordnung, wobei bei einem Verfahren k–ter Ordnung der Übergang nicht nur vom letzten Zustand abhängt, sondern von den k letzten Zuständen, lassen sich Sprachvergleiche mittels aufeinanderfolgender Vokale und Konsonanten durchführen (diese entsprechen dabei den einzelnen Zuständen — vgl. [41]). Dabei sind für einfache polynesische Sprachen wie Lifu oder Samoa Markov–Ketten erster Ordnung hinreichend, für Latein oder Italienisch zweite Ordnung, wogegen und für Deutsch sogar dritte Ordnung notwendig wird (für Hebräisch wird es dann noch komplizierter). Neben vielen solchen Anwendungen in der Informatik, den Wirtschaftswissenschaften, der Biologie etc. lassen sich auch Spiele wie das berühmte Monopoly (vgl. [2]) modellieren. Diese Problemstellungen haben gemeinsam, dass die Suche nach dem stationären Wahrscheinlichkeitsverteilungsvektor eine herausragende Rolle spielt. Bei ergodischen Markov–Ketten, bei denen man von jedem Zustand letztlich in jeden anderen gelangen kann, gibt dieser Vektor an, wie wahrscheinlich das System die einzelnen Zustände annimmt, wenn erst einmal der Einfluss der Ausgangsverteilung beseitigt ist (d.h. im Prinzip wenn das System unendlich lange läuft — bei periodischen Markov–Ketten ist die Situation noch etwas anders).

In der Sprache der linearen Algebra entspricht dieses Problem der Suche nach dem Linkseigenvektor zum Eigenwert 0 einer singulären Matrix mit eindimensionalem Kern (falls die Übergänge zeitkontinuierlich verlaufen) oder nach dem Linkseigenvektor zum Eigenwert 1 einer stochastischen Matrix (wenn die Übergänge in diskreten Zeitschritten ablaufen). Der dabei jeweils berechnete Linkseigenvektor entspricht normiert gerade dem stationären Wahrscheinlichkeitsvektor.

Eine besondere Struktur weisen periodische Markov–Ketten auf; sie führen im zeitkontinuierlichen Fall auf eine Matrix, den sogenannten infinitesimalen Generator, die zyklisch ist und deren Jacobi–Matrix eine sehr spezielle Eigenwertstruktur aufweist, die in der Theorie und Praxis ausgenutzt werden kann.

Natürlich könnte man das Problem der Berechnung des Linkseigenvektors nach Transformation auf ein reguläres System mit dem Gauß–Algorithmus lösen. Gerade bei großen und dünn besetzten Matrizen führt dies aber zu hohem Speicheraufwand, sodass iterative Verfahren an Bedeutung gewinnen, die entweder auf ein ebenso transformiertes reguläres System oder direkt auf das singuläre System angewendet werden. Für p–periodische Markov–Ketten wird dabei das Verfahren der sukzessiven Überrelaxation (SOR–Verfahren) besonders bedeutsam, weil es (vgl. [64]) einen einfachen Zusammenhang zwischen den Eigenwerten der Jacobi–Matrix J und den Eigenwerten der SOR–Iterationsmatrix gibt, mit dessen Hilfe man auch zu Aussagen über die optimale Wahl des Relaxationsparameters und über den zugehörigen Konvergenzfaktor gelangt.

Erste Ergebnisse in dieser Richtung findet man in dem berühmten Buch *Matrix Iterative Analysis* von Varga 1992 (vgl. [64]). Er gibt an, wie sich der optimale Relaxationsfaktor berechnen lässt, falls das Spektrum von J^p nichtnegativ ist. Verallgemeinerungen dieser Aussagen stammen aus der Folgezeit, so für allgemein nichtnegatives oder nichtpositives Spektrum (vgl. Wild, Niethammer 1987 [65]) und für reelles Spektrum (vgl. Eiermann, Niethammer, Ruttan 1990 [11]). Zwei Arbeiten aus jüngster Zeit (vgl. Galanis, Hadjidimos und Noutsos [18] und [19]) schließen die verbleibende Lücke, nämlich das Problem des komplexen Spektrums, wobei zuerst der Fall eines komplexen Eigenwert–$2p$–Tupels der Jacobi–Matrix J behandelt wird, bevor der allgemeineren Fall betrachtet wird.

Überraschend in der langen Arbeit von Kontovasilis, Plemmons und Stewart 1991 (vgl. [34]) ist die Tatsache, dass bei infinitesimalen Generatoren p–periodischer Markov–Ketten, für die das Spektrum von J^p nichtnegativ ist, Relaxationsparameter auch außerhalb des Intervalls $(0, 2)$ zugelassen werden. Aus diesem Intervall muss aber nach dem Lemma von Kahan der Relaxationsparameter notwendigerweise gewählt werden, wenn man ein konvergentes Verfahren erhalten will. Darüberhinaus können neben dem klassischen optimalen Relaxationsparameter nach Varga noch weitere optimale existieren, unter Umständen sogar negative, die denselben Konvergenzfaktor erzielen. Diese weiteren optimalen Relaxationsparameter sind insgesamt weniger störungsanfällig als der klassische. Dafür wird nach der SOR–Iteration ein zusätzlicher Rekonstruktionsschritt notwendig. Dies liegt daran, dass die SOR–Iteration, die man hier wegen der Homogenität des linearen Gleichungssystems einfacher als Vektoriteration auffasst, auch mit den nicht–klassischen Relaxationsparametern konvergiert, allerdings gegen einen Vektor der nicht dem stationären Wahrscheinlichkeitsvektor entspricht, aus dem man diesen aber andererseits eindeutig konstruieren kann. Die Rede ist in diesem Zusammenhang von *extended convergence*; es wird im weiteren Verlauf der Arbeit auch der Begriff *verallgemeinerte Konvergenz* dafür verwendet.

Der weitere Aufbau dieser Arbeit ist wie folgt:

Im einleitenden Kapitel 2 wird das SOR–Verfahren bei nichtsingulären linearen Gleichungssystemen eingeführt. Klassische Ergebnisse und Konvergenzsätze werden vorgestellt. Ferner werden Begriffe von Varga wie derjenige der schwach zyklischen Matrix vom Index p und der p–zyklischen Matrix bereitgestellt und in Zusammenhang mit dem SOR–Verfahren gebracht.

Im Fall, dass das vorliegende lineare Gleichungssystem singulär ist, werden die Fälle eines konsistenten Systems, d.h. dass die rechte Seite im Bildraum der vorliegenden Matrix liegt, und eines inkonsistenten Systems unterschieden. Dabei ist der konsistente Fall hier von besonderer Bedeutung, weil die rechte Seite 0 als Eigenvektor der Matrix natürlich nichttrivial in deren Bildraum liegt.

Im letzten Abschnitt dieses Kapitels geht es allgemein um die Einführung von Markov–Ketten und die Bereitstellung wichtiger Begriffe in diesem Zusammenhang, wobei Grundbegriffe zeitkontinuierlicher und zeitdiskreter Markov–Ketten getrennt behandelt werden.

Thema des dritten Kapitels ist die Anwendung des SOR–Verfahrens bei p–periodischen Markov–Ketten, wenn das Spektrum von J^p reell ist. Es zeigt sich, dass ein Schritt des SOR–Verfahrens gerade äquivalent ist zu p Schritten eines p–Schritt–Relaxations–Verfahrens, weswegen man zur Betrachtung solcher Verfahren übergeht, weil hierbei die Hilfsmittel aus der Theorie semiiterativer Verfahren angewendet und brauchbare Ergebnissen hergeleitet werden können. Hierbei sind nun sogenannte Hypozykloide, eine spezielle Klasse von Rollkurven, von Bedeutung. Das p–Schritt–Relaxationsverfahren und damit auch das SOR–Verfahren ist genau dann konvergent, wenn alle Eigenwerte von J im Innengebiet eines dem Verfahren eindeutig zugeordneten Hyozykloids liegen. Im Falle singulärer konsistenter p–zyklischer Systemen reduziert sich das auf die Untersuchung, wann alle Eigenwerte außer dem zum Eigenwert 1 gehörenden p–Tupel der Jacobi–Matrix im Innengebiet beziehungsweise auf dem Rand dieses Hypozykloids liegen. Aus diesem Grund werden im Abschnitt 3.2 diese Rollkurven ausführlich betrachtet.

Der nächste Abschnitt enthält die Resultate von Varga, Wild, Niethammer, Eiermann und Ruttan (vgl. [64], [65] und [11]) über den optimalen Relaxationsparameter und den zugehörigen Konvergenzfaktor für nichtsinguläre Gleichungssysteme, wobei diese leicht modifiziert auch für singuläre Gleichungssysteme angewendet werden können.

Kapitel 3.4 führt die verallgemeinerte Konvergenz ([34]) ein und erläutert Modifikationen für das singuläre Gleichungssystem im Fall, dass der Relaxationsparame-

ter nicht in $(0, 2)$ liegt. Bei Konvergenz im klassischen Sinn müssen alle Eigenwerte der Iterationsmatrix außer dem Eigenwert 1 echt kleiner als 1 sein, während im Fall der verallgemeinerten Konvergenz der dominante Eigenwert der Iterationsmatrix, der im Allgemeinen größer als 1 ist, dem Eigenwert–p–Tupel 1 von J zugeordnet sein muss. Man kann das Verfahren als Vektoriteration auffassen, der im Anschluss ein Rekonstruktionsschritt folgt. Der Konvergenzfaktor hängt ab vom Quotienten aus subdominantem und dominantem Eigenwert, wobei der erstere einem anderen Eigenwert–p–Tupel von J als 1 zugeordnet sein muss. Im Prinzip müsste man bei der Vektoriteration eine regelmäßige Normierung in jedem Schritt durchführen, weil das Verfahren, wenn der dominante Eigenwert außerhalb des Einheitskreises liegt, sonst divergiert. In der Praxis der verallgemeinerten Konvergenz genügt es, am Schluss einen solchen Normierung– und Rekonstruktionsschritt durchzuführen.

Aufgrund der zuletzt genannten Bedingung wird in Abschnitt 3.5 untersucht, wo in der komplexen Ebene die anderen 1 zugeordneten Eigenwerte liegen. Diese Frage ist in der Arbeit von Kontovasilis, Plemmons und Stewart [34] nicht behandelt. Es wird hier gezeigt, dass für Relaxationsparameter größer als $\frac{p}{p-1}$ diese Eigenwerte sicher im Einheitskreis liegen (vgl. Korollar 3.14), für negative Relaxationsparameter konnte zumindest für kleine Werte von p gezeigt werden, dass der subdominante Eigenwert der SOR–Iterationsmatrix nicht dem Eigenwert–p–Tupel 1 von J zugeordnet ist (Satz 3.16 und Bemerkung 3.4), womit die oben genannte Bedingung erfüllt ist.

Im abschließenden Abschnitt dieses Kapitels folgen Abschätzungen, um wieviel schlechter das Gauß–Seidel–Verfahren höchstens ist als das optimale SOR–Verfahren. Diese Frage ist insofern von besonderer Wichtigkeit, als man im Allgemeinen das Spektrum der Jacobi–Matrix nicht kennt und somit — will man SOR mit optimalem Relaxationsparameter durchführen — aufwendig berechnen müsste.

Im dritten Kapitel wird von einem komplexen Spektrum von J^p ausgegangen. Dabei wird zunächst gezeigt, dass in diesem Fall die verallgemeinerte Konvergenz unter Umständen nicht gegen den stationären Wahrscheinlichkeitsvektor konvergiert. Konvergenz erfolgt gegen einen anderen Vektor, wenn der dominante Eigenwert nicht dem Eigenwert–p–Tupel 1 zugeordnet ist, was nur bei komplexem Spektrum auftreten kann. Bei mehreren komplexen Eigenwerten von J^p ist völlig unklar, zu welchem Tupel der dominante Eigenwert der SOR–Iterationsmatrix gehört.

Abschnitt 4.2 enthält die Ergebnisse von Galanis, Hadjidimos und Noutsos aus [18] und [19] für komplexes Spektrum von J^p und zwar aufgeteilt in das Einpunktproblem (nur ein komplexes $2p$–Tupel) und das Mehrpunktproblem. Diese sehr technischen Ergebnisse werden hier möglichst anschaulich wiedergegeben. We-

sentlich ist hierbei die Technik, dass man zunächst ein Hypozykloid sucht, das die Eigenwerte von J möglichst gut einschließt. Im regulären Fall müssen dabei alle Eigenwerte von J im Innengebiet und auf dem Rand des Hypozykloids liegen, während dies im singulären Fall für alle Eigenwerte bis auf das p–Tupel vom Betrag 1 gilt. Erweiterte Konvegenz ist hierbei nicht gemeint, es wird von klassischer Konvergenz ausgegangen.

Es folgen im nächsten Abschnitt einige Lemmata, die später dazu dienen die wesentlichen Ergebnisse aus Abschnitt 4.4 zu beweisen. Insbesondere werden die Grundlagen dafür gelegt zu zeigen, dass man in Abhängigkeit davon, ob p gerade oder ungerade ist, Teilintervalle der reellen Achse bijektiv in das Intervall $(0, \frac{p}{p-1})$ abbilden kann. Somit reicht die Kenntnis der Konvergenz des SOR–Verfahrens in diesem Intervall aus, um die erweiterte Konvergenz im Griff zu haben.

Dieser Beweis wird im Hauptabschnitt 4.4 geführt (vgl. Satz 4.12), der somit einerseits einen alternativen Beweis zu dem länglichen in [34] enthält, zum anderen aber auch eine Erweiterung dazu ist, da keine Einschränkungen mehr an das Spektrum von J^p getroffen werden. Im weiteren Verlauf werden eine Reihe Einzelfälle betrachtet. Schließlich folgt als zentrales Ergebnis, dass durch das verallgemeinerte SOR–Verfahren kein besserer Konvergenzfaktor zu erzielen ist als mit dem klassischen SOR–Verfahren (Satz 4.14).

Analog zu Abschnitt 3.6 wird in Abschnitt 4.5 untersucht, um wieviel schlechter es im komplexen Fall ist, wenn statt SOR mit optimalem Relaxationsparameter einfach das Gauß–Seidel–Verfahren durchgeführt wird. Man verliert nicht mehr als in den reellen Grenzfällen, nämlich dass das Spektrum von J^p nichtnegativ bzw. nichtpositiv ist (vgl. z.B. Satz 4.20). Außerdem wird der Frage nachgegangen, ob man aus der speziellen Herkunft der Matrix schon Aussagen über ihr Spektrum treffen kann, was aber weitgehend zu verneinen ist.

In Kapitel 5 werden zunächst an drei numerischen Beispielen die Aussagen der vorhergehenden Kapitel verdeutlicht. Insbesondere werden die Aussagen, dass man aus der Kenntnis des klassischen optimalen Relaxationsparameters und des Index p die weiteren optimalen Parameter berechnen kann und andererseits kein besserer Konvergenzfaktor erreicht wird, am Beispiel herausgestellt.

Zum Abschluss wird speziell noch einmal auf den Fall $p = 2$ eingegangen, der schon aus historischen Gründen, es handelt sich hierbei um nichts anderes als die PA–Matrizen von Young [67], von besonderer Bedeutung ist.

Bezeichnungen

In dieser Arbeit werden Matrizen mit großen lateinischen Buchstaben bezeichnet, Vektoren mit kleinen lateinischen und Skalare mit kleinen griechischen.

Eine Ausnahme bilden π, was, sofern es nicht ohnehin für die Kreiszahl steht, auch für den stationären Wahrscheinlichkeitsvektor verwendet wird und p, welches der Index der Zyklizität ist. P bezeichnet je nach Zusammenhang eine stochastische Matrix, eine Permutationsmatrix oder eine Projektion.

Alle Vektoren, die auftreten, sind, sofern nicht explizit anderes erwähnt wird, Spaltenvektoren.

Folgende Bezeichnungen sind für spezielle Größen reserviert :

Q	infinitesimaler Generator der zeitkontinuierlichen Markov–Kette (vgl. S. 24)
D	(Block–)Diagonalmatrix von Q^T (vgl. S. 9 und S. 11)
$-L$	untere (Block–)Dreiecksmatrix von Q^T (vgl. S. 9 und S. 11)
$-U$	obere (Block–)Dreiecksmatrix von Q^T (vgl. S. 9 und S. 11)
J	(Block–)Jacobi–Matrix von Q^T, $J = D^{-1}(L+U)$ (vgl. S. 11)
$\mathcal{L}_\omega$	SOR–Iterationsmatrix mit Relaxationsparameter ω (vgl. S. 10)
π^*	der stationäre Wahrscheinlichkeitsvektor $(\pi^*)^T Q = 0^T$ (vgl. S. 24)
ω	Relaxationsparameter (vgl. S. 10)
μ	Eigenwert von J (vgl. S. 14)
λ	Eigenwert von $\mathcal{L}_\omega$ (vgl. S. 14)
τ	$\tau = \sqrt[p]{\lambda}$ (vgl. Bemerkung 2.6 auf S. 15)
p	Index der Zyklizität (vgl. S. 12)

$$
\begin{aligned}
\sigma(A) &= \{\xi \in \mathbb{C} : \xi \text{ ist Eigenwert von } A\} \qquad \text{Spektrum von } A \\
\rho(A) &= \max_{\xi \in \sigma(A)} |\xi| \qquad \text{Spektralradius von } A \\
\tilde{\sigma}(A) &= \sigma(A) \setminus \{\xi : |\xi| = \rho(A)\} \\
\tilde{\rho}(A) &= \max_{\xi \in \sigma(A), |\xi| \neq \rho(A)} |\xi| \qquad \text{Subspektralradius von } A
\end{aligned}
$$

Generell bezeichnen Größen mit einem „$*$“ im Index (wie z.B. ω^*, η^*) jeweils optimale Größen, beim erweiterten SOR–Verfahren teilen sich diese auf in $\omega_0, \omega_+, \omega_-$, η_0, η_+, η_-, wobei der Index „0“ sich jeweils auf Größen, deren Relaxationsparameter im klassischen Intervall $(0, \frac{p}{p-1})$ liegt, bezieht, der Index „+“ auf Größen, deren Relaxationsparameter größer als $\frac{p}{p-1}$ ist, und analog der Index „$-$“ auf Größen mit negativem Relaxationsparameter.

2 Einführung

In diesem Kapitel werden zum einen bekannte Sätze für das SOR–Verfahren und das Block–SOR–Verfahren zusammengestellt, die im weiteren Verlauf der Arbeit von Bedeutung sind; insbesondere werden Voraussetzungen und klassische Ergebnisse für die Anwendung des SOR–Verfahrens für p–zyklische Matrizen bereitgestellt. Zum anderen wird durch die Einführung von Markov–Ketten der Bogen zur Berechnung des stationären Wahrscheinlichkeitsverteilungsvektors bei p–periodischen, (im Allgemeinen) zeitkontinuierlichen Markov–Ketten geschlagen.

2.1 SOR bei nichtsingulären linearen Gleichungssystemen

Das SOR–Verfahren

Gegeben sei das lineare Gleichungssystem

$$Ax = b \tag{1}$$

mit einer nichtsingulären Matrix $A \in \mathbb{R}^{n\times n}$ und der rechten Seite $b \in \mathbb{R}^n$.[1]

(1) soll in Fixpunktform $x = Tx + g$ überführt werden. Dazu sei auf die entsprechende Literatur verwiesen (z.B. [64, 21, 3, 45]). Die Fixpunktiteration $x_{k+1} = Tx_k + g$ konvergiert, falls $\|T\| < 1$ für eine Matrixnorm gilt und genau dann, wenn $\rho(T) < 1$ gilt.

Die Matrix A kann zerlegt werden gemäß

$$A = D - L - U \quad \text{bzw.} \quad A = M - N, \tag{2}$$

wobei D Diagonalmatrix, L untere und U obere Dreiecksmatrix bzw. M nichtsinguläre Matrix ist, und ergibt, falls D nichtsingulär ist, die Fixpunktform

$$x = D^{-1}(L + U)x + D^{-1}b \tag{3}$$

und das zugehörige Gesamtschrittverfahren oder Jacobi–Verfahren.

Werden die bereits berechneten Komponenten mitverwendet, gelangt man zum Einzelschritt– bzw. Gauß–Seidel–Verfahren mit der Fixpunktform

$$x = (D - L)^{-1}Ux + (D - L)^{-1}b. \tag{4}$$

[1] Prinzipiell lässt sich natürlich alles mit komplexen Matrizen und komplexen Vektoren einführen, im weiteren Verlauf werden aber nur reelle benötigt.

Das Gauß–Seidel–Verfahren lässt sich auch schreiben in der Form

$$x_{k+1} = x_k + h_{k+1}$$

mit Änderungsvektor

$$h_{k+1} = ((D-L)^{-1}U - I)x_k + (D-L)^{-1}b.$$

Mit Relaxationsparameter ω erhält man die Vorschrift

$$x_{k+1} = x_k + \omega h_{k+1},$$

welche aus einer Zerlgung der Matrix A resultiert gemäß

$$A = \underbrace{\frac{1}{\omega}D - L}_{=:M} - \underbrace{\left[\left(\frac{1}{\omega} - 1\right)D + U\right]}_{=:N}. \qquad (\omega \neq 0)$$

und heißt SOR–Verfahren (**S**uccessive **O**ver **R**elaxation)

$$\begin{aligned} x_{k+1} &= \mathcal{L}_\omega x_k + (D - \omega L)^{-1}b \qquad \text{mit} \\ \mathcal{L}_\omega &= (D - \omega L)^{-1}((1-\omega)D + \omega U). \end{aligned} \tag{5}$$

Ist $\omega > 1$ spricht man von Überrelaxation, für $\omega < 1$ von Unterrelaxation.

Lemma 2.1 (Lemma von Kahan) *Für den SOR–Operator $\mathcal{L}_\omega$ gilt für alle $\omega \neq 0$ $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$.*

Beweis:
Geht man vom normierten System $Ax = I - \tilde{L} - \tilde{U} = b$, $A \in \mathbb{C}^{n\times n}$ aus, so gilt für die Iterationsmatrix $\mathcal{L}_\omega$

$$\begin{aligned} \det(\mathcal{L}_\omega) &= \det((I - \omega\tilde{L})^{-1}((1-\omega)I + \omega\tilde{U}))) \\ &= \det(I - \omega\tilde{L})^{-1}\det((1-\omega)I + \omega\tilde{U})) \\ &= (1-\omega)^n, \end{aligned}$$

weil die Faktoren von $\mathcal{L}_\omega$ Dreiecksmatrizen sind.

Für die Eigenwerte λ_i von $\mathcal{L}_\omega$ gilt

$$|\det(\mathcal{L}_\omega)| = |(1-\omega)^n| = \prod_{i=1}^{n} |\lambda_i|,$$

woraus folgt

$$\rho(\mathcal{L}_\omega) = \max_{1 \le i \le n} |\lambda_i| \ge |1 - \omega|.$$

□

Für reelle Relaxationsparameter ω gilt damit

Korollar 2.2 *Das SOR–Verfahren kann nur für $0 < \omega < 2$ konvergent sein.*

Definition 2.1 *Sei $A \in \mathbb{R}^{n \times n}$ in der folgenden Weise partitioniert:*

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,N} \\ A_{2,1} & A_{2,2} & \dots & A_{2,N} \\ \vdots & & & \vdots \\ A_{N,1} & A_{N,2} & \dots & A_{N,N} \end{pmatrix}. \tag{6}$$

Ferner sei $D = \text{blockdiag}(A_{1,1}, A_{2,2}, \dots, A_{N,N})$ die Blockdiagonale von A.

Dann ist

$$J := -D^{-1}A + I \tag{7}$$

die Block–Jacobi–Matrix[2] *von A.*

Zerlegt man (6) gemäß

$$A = \hat{D} - \hat{L} - \hat{U}, \tag{8}$$

wobei $\hat{D} = \text{blockdiag}(A_{1,1}, A_{2,2}, \dots, A_{N,N})$ die Blockdiagonale von A, $\hat{L}$ untere Blockdreiecksmatrix und $\hat{U}$ obere Blockdreiecksmatrix ist, gelangt man zu den Blockvarianten der Verfahren (3), (4) und (5). Dabei muss jeweils vorausgesetzt werden, dass die Blockdiagonalmatrix invertierbar ist.

Für die Konvergenz des SOR–Verfahrens gibt es eine Reihe von Konvergenzsätzen:

Satz 2.3 (Reich–Ostrowski–Theorem) *[64, Theorem 3.6] Ist die Matrix A des linearen Gleichungssystems $Ax = b$ Hermitesch, so gilt: Das SOR–Verfahren konvergiert genau dann, wenn A positiv definit ist und $0 < \omega < 2$ erfüllt ist.*

[2] Wenn im weiteren Verlauf der Arbeit einfach nur von Jacobi–Matrix gesprochen wird, ist immer diese Block–Jacobi–Matrix gemeint.

Korollar 2.4 *[53, Satz 11.14] Für eine symmetrische, positiv definite Matrix* $A \in \mathbb{R}^{n\times n}$ *gilt*

$$\rho(\mathcal{L}_\omega) < 1 \quad \textit{für} \quad \omega \in (0,2).$$

Satz 2.5 *[64, Theorem 3.16] Ist* A *eine* M*–Matrix (d.h. ist* $a_{ij} \leq 0$ *für* $i \neq j$ *und* $A^{-1} \geq 0$*), so konvergiert das SOR–Verfahren für* $0 < \omega \leq 1$ *und es gilt*

$$1 > \rho(\mathcal{L}_{\omega_1}) \geq \rho(\mathcal{L}_{\omega_2}) \qquad \textit{für} \quad 0 < \omega_1 < \omega_2 < 1.$$

Bemerkung 2.1 *Bei* M*–Matrizen ist der Konvergenzfaktor des Gauß–Seidel–Verfahrens immer besser als der eines unterrelaxierten SOR–Verfahren.*

Schwach zyklische Matrizen vom Index p

In diesem Unterabschnitt sollen einige Begriffe bereitgestellt werden, so der der schwach zyklischen Matrix vom Index p und der p–zyklischen Matrix (vgl. [63, 64]).

Definition 2.2 *Eine Matrix* $B \in \mathbb{R}^{n\times n}$ *ist* schwach zyklisch vom Index p*, wenn es eine* $n \times n$*–Permutationsmatrix* P *gibt, sodass*

$$PBP^T = \begin{pmatrix} 0 & 0 & \dots & 0 & B_{1,p} \\ B_{2,1} & 0 & & 0 & 0 \\ 0 & B_{3,2} & \ddots & 0 & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & B_{p,p-1} & 0 \end{pmatrix}, \tag{9}$$

wobei die Diagonalblöcke quadratisch sind.

Bemerkung 2.2 *Eine Matrix kann schwach zyklisch von verschiedenen Indizes sein, so ist die Matrix (9) auch* 2*–zyklisch, wenn man die Blockdiagonalmatrix* $diag(B_{2,1}, \dots, B_{p,p-1})$ *als einen Block auffasst.*

Davon zu unterscheiden ist der Fall einer zyklischen Matrix, auch wenn beide Definitionen eng miteinander verbunden sind.

Definition 2.3 *Hat eine Matrix* $A \in \mathbb{R}^{n\times n}$ *die Gestalt* $A = I - B$ *mit einer Matrix* B*, die schwach zyklisch vom Index* p *ist, dann ist* A p–zyklisch.

Bemerkung 2.3 *Ist die Jacobi–Matrix $J = D^{-1}(L+U)$ einer Matrix A schwach zyklisch vom Index p, dann ist $A = D(I - D^{-1}(L+U))$ eine p–zyklische Matrix.*

Definition 2.4 *Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt* reduzibel, *wenn eine Permutationsmatrix P existiert, sodass*

$$PAP^T = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right) \tag{10}$$

mit quadratischen Matrizen A_{11} und A_{22} der Ordnung $n_1 > 0$ und $n_2 > 0$ ($n_1 + n_2 = n$). A heißt irreduzibel, wenn A nicht reduzibel ist.

Lemma 2.6 *Die Matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ ist genau dann irreduzibel, wenn es für jede beliebige Zerlegung $W = S \cup T$ der Menge $W := \{1, 2, \ldots, n\}$ mit $S \cap T = \emptyset$, $S \neq \emptyset$, $T \neq \emptyset$ ein Element $a_{ij} \neq 0$ gibt mit $i \in S$ und $j \in T$.*

Satz 2.7 *Sei $B \geq 0$ eine irreduzible, schwach zyklische Matrix vom Index p der Dimension $n \times n$. Dann hat B genau p Eigenwerte vom Betrag $\rho(B)$ der Form*

$$\rho(B) \exp\left[i\left(\frac{2\pi j}{p}\right)\right], \qquad 0 \leq j \leq p-1.$$

Ferner ist das Spektrum von B invariant unter Drehungen um den Winkel $2\pi/p$ (aber keinen kleineren Winkel).

Bemerkung 2.4 *Eigenwerte solcher Matrizen treten immer in p–Tupeln auf.*

Definition 2.5 *Ist eine Matrix (6) $A = D - L - U$ p–zyklisch, dann heißt A konsistent geordnet, wenn alle Eigenwerte der Matrix*

$$B(\alpha) = \alpha L + \alpha^{-(p-1)} U$$

für $\alpha \neq 0$ unabhängig von α sind.

Bemerkung 2.5 *Aus dieser Definition folgt, dass eine Matrix A der Gestalt*

$$A = \begin{pmatrix} A_{1,1} & 0 & 0 & \ldots & 0 & A_{1,p} \\ A_{2,1} & A_{2,2} & 0 & \ldots & 0 & 0 \\ 0 & A_{3,2} & A_{3,3} & \ldots & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & A_{p,p-1} & A_{p,p} \end{pmatrix} \tag{11}$$

mit quadratischen Diagonalblöcken konsistent geordnet ist.

Das SOR–Verfahren für p–zyklische Matrizen

Bei Diskretisierung elliptischer Randwertaufgaben wie der Konvektions–Diffusions–Gleichung mit konstanten Koeffizienten

$$\begin{aligned} -\Delta u(x,y) + s u_x(x,y) + t u_y(x,y) &= f(x,y) \qquad \text{in } (0,1)\times(0,1) =: Q \\ u(x,y) &= g(x,y) \qquad \text{für } (x,y) \in \partial Q \\ s,t \in \mathbb{R} &\qquad f,g \in C^1(0,1)^2 \end{aligned}$$

mit schachbrettartiger Nummerierung der Unbekannten gelangt man (nach Diagonalskalierung) auf eine 2–zyklische Matrix, für David Young (vgl. [67]) Ausgangspunkt seiner Betrachtungen über PA–Matrizen, die nichts anderes als 2–zyklische Matrizen sind. Bei Young finden sich Aussagen über die Berechnung des optimalen Relaxationsparameters und den Konvergenzfaktor.

Dabei wird aufgrund der Zyklizität — zunächst für 2–zyklische, dann für p–zyklische Matrizen — immer das Block–SOR–Verfahren betrachtet; auch wenn im Folgenden nur *SOR* steht, ist Block–SOR gemeint. Für p–zyklische Matrizen gewinnt man mit Hilfe von Satz 2.7 bzw. Bemerkung 2.4 den fundamentalen Satz von Varga (vgl. [64, Theorem 4.3]).

Satz 2.8 *Sei A nach (6) eine konsistent geordnete p–zyklische Matrix mit nichtsingulären Diagonalblöcken $A_{i,i}$ $1 \le i \le N$. Ist $\omega \neq 0$ und λ ein von Null verschiedener Eigenwert von $\mathcal{L}_\omega$ und erfüllt μ die Gleichung*

$$(\lambda + \omega - 1)^p = \lambda^{p-1}\omega^p\mu^p, \tag{12}$$

dann ist μ Eigenwert der Block–Jacobi–Matrix J zu A.

Ist umgekehrt μ Eigenwert von J und erfüllt λ die Formel (12), so ist λ Eigenwert von $\mathcal{L}_\omega$.

Durch diesen Satz ist der Zusammenhang zwischen den p–Tupeln der Eigenwerte von J und den Eigenwerten von $\mathcal{L}_\omega$ hergestellt.

Korollar 2.9 *Sei A nach (6) eine konsistent geordnete p–zyklische Matrix mit nichtsingulären Diagonalblöcken $A_{i,i}$ $1 \le i \le N$.*

Ist μ Eigenwert der Block–Jacobi–Matrix J, dann ist μ^p Eigenwert von $\mathcal{L}_1$. Ist umgekehrt λ Eigenwert von $\mathcal{L}_1$ und $\lambda = \mu^p$, dann ist μ Eigenwert von J.

(12) erweist sich als Verallgemeinerung der Formel von Young (vgl. [67, Theorem 2.2 u.a.]) für PA–Matrizen[3].

Entscheidend für die Konvergenzgeschwindigkeit des Iterationsverfahrens ist der Spektralradius der Iterationsmatrix; er entspricht gerade dem Konvergenzfaktor (vgl. auch Definition 2.7 und Bemerkung 2.7). Unter einem optimalen Relaxationsparameter versteht man denjenigen Relaxationsparameter, der den Spektralradius der Iterationsmatrix minimiert.

Bereits in [64, Theorem 4.4] gibt Varga die Antwort auf die Frage nach dem optimalen Relaxationsparameter – allerdings unter der Einschränkung, dass das Spektrum der p–ten Potenz der Jacobi–Matrix reell und nichtnegativ ist.

Satz 2.10 *Sei $A \in \mathbb{R}^{n\times n}$ eine konsistent geordnete, p–zyklische Matrix mit nichtsingulären Diagonalblöcken $A_{i,i}$, $1 \le i \le N$. Alle Eigenwerte von J^p seien reell und nichtnegativ, $0 \le \rho(J) < 1$. Sei ferner ω_0 die Lösung von*

$$(p-1)^{p-1}\omega^p(\rho(J))^p - p^p(\omega-1) = 0 \tag{13}$$

in $(1, \frac{p}{p-1})$. Dann gilt

1. *$\rho(\mathcal{L}_{\omega_0}) = (p-1)(\omega_0 - 1)$,*
2. *$\rho(\mathcal{L}_{\omega}) > \rho(\mathcal{L}_{\omega_0})$ für alle $\omega \neq \omega_0$.*

Darüberhinaus ist das Block–SOR–Verfahren konvergent für alle $0 < \omega < \frac{p}{p-1}$.

Bemerkung 2.6 *Mit der Substitution $\tau = \lambda^{1/p}$ findet man (12) in der Literatur (vgl. [65] und [34]) (durch Ziehen der p–ten Wurzel) auch in der Form*

$$\tau^p - \omega\mu\tau^{p-1} - (1-\omega) = 0, \tag{14}$$

Ist μ Eigenwert der Jacobi–Matrix und $\hat{\tau}$ Lösung von (14), dann ist $\hat{\tau}^p$ Eigenwert von $\mathcal{L}_\omega$.

[3] Diese sind gerade die 2–zyklischen Matrizen.

2.2 SOR bei singulären und inkonsistenten Gleichungssystemen

In der Praxis kommt es vor, dass ein lineares Gleichungssystem $Ax = b$ mit singulärer Matrix $A \in \mathbb{R}^{n\times n}$ vorliegt, das gelöst werden muss. Man unterscheidet die beiden Fälle

1. $b \in \text{bild}(A)$, (konsistentes lineares System),
2. $b \notin \text{bild}(A)$ (inkonsistentes lineares System).

Im Fall des inkonsistenten linearen Systems existiert keine Lösung des Systems d.h. es ist $\|b - Ax\| > 0$ für alle Vektoren $x \in \mathbb{R}^n$. Man geht dann im Allgemeinen dazu über, denjenigen Vektor x zu bestimmen, für den $\|b - Ax\|_2$ minimal wird. Zur Berechnung dieser Kleinste–Quadrate–Lösung kann beispielsweise die Gaußsche Normalengleichung $A^T Ax = A^T b$ gelöst werden.

Konsistente singuläre Systeme

Auf die entsprechende Literatur ([33, 61]) sei verwiesen.

Satz 2.11 *[33, Theorem 1] Das lineare Gleichungssystem $Ax = b$ besitze eine Lösung $\hat{x}$, und die Matrix M sei nicht singulär. Dann sind äquivalent*

1. *Für alle x_0 konvergiert die Folge $\{x_k\}_k$ aus $Mx_{k+1} = (M - A)x_k + b$ gegen eine Lösung von $Ax = b$.*
2. *Für alle e_0 konvergiert die Folge $\{e_k\}_k$ aus $e_k = Te_k =: (I - M^{-1}A)e_k$ gegen einen Vektor aus dem Kern von A.*
3. *Es gibt ein Komplement $\overline{\text{kern}}(A)$ von $\text{kern}(A)$ (d.h. $\text{kern}(A) \cap \overline{\text{kern}}(A) = \emptyset$ und $\text{kern}(A) \oplus \overline{\text{kern}}(A) = \mathbb{R}^n$), so dass gilt*
   (a) *$\text{kern}(A)$ und $\overline{\text{kern}}(A)$ sind bezüglich T invariante Unterräume, es gilt also $T(\text{kern}(A)) \subseteq \text{kern}(A)$ und $T(\overline{\text{kern}}(A)) \subseteq \overline{\text{kern}}(A)$.*
   (b) *$T|_{\text{kern}(A)}$ ist die Identität, und für alle $x \in \overline{\text{kern}}(A)$ gilt $T^k x \to 0$ für $k \to \infty$.*

   In Worten: T ist die Identität auf $\text{kern}(A)$ und konvergent auf $\overline{\text{kern}}(A)$.

Definition 2.6 *Eine n–dimensionale Matrix T heißt* konvergent für *eine n–dimensionale Matrix A, wenn T auf* $\text{kern}(A)$ *die Identität ist und konvergent auf* $\overline{\text{kern}}(A)$ *ist.*

Lemma 2.12 *[33, Lemma 1]*

1. *Die Folge von Potenzen $\{T^k\}_k$ einer quadratischen Matrix T konvergiert genau dann (gegen einen beliebigen Grenzwert), wenn T konvergent für eine (geeignete) Matrix A ist.*
2. *Äquivalent dazu konvergieren die Potenzen von T genau dann, wenn für eine Matrix A gilt*
$$\lim_{k\to\infty} T^k = P_{\text{kern}(A)},$$
wobei $P_{\text{kern}(A)}$ eine Projektion auf $\text{kern}(A)$ *ist.*

Im konkreten Fall wird Satz 2.11 zu folgendem Satz (vgl. [9, Theorem A]), wobei man hier von der Fixpunktformulierung

$$x = Tx + c$$

ausgeht, die aus $Ax = b$ durch die Zerlegung $A = M - (M - A)$ (M nichtsingulär) mit $T := I - M^{-1}A$ und $c := M^{-1}b$ gewonnen werden kann. Ist A singulär, so ist 1 dann sicher im Spektrum von T enthalten, d.h. es ist $\rho(T) \geq 1$.

Satz 2.13 *Das System $x = Tx + c$ sei singulär aber lösbar. Die Fixpunktiteration*

$$x_k = Tx_{k-1} + c \quad (k \geq 1); \qquad x_0 \in \mathbb{R}^n \tag{15}$$

konvergiert für alle x_0 genau dann gegen eine Lösung, wenn folgende drei Bedingungen erfüllt sind:

- $\rho(T) = 1$,
- $\lambda = 1$ *ist einziger Eigenwert von T mit Betrag* 1,
- $rang(I - T) = rang(I - T)^2$.

Darüberhinaus gilt (mit $x^* = \lim_{k\to\infty} x_k$*)*

$$\limsup_{k\to\infty}\left\{\sup_{x_0\neq x^*}\left[\frac{\|x^*-x_k\|}{\|x^*-x_0\|}\right]^{1/k}\right\} = \max\{|\lambda| : \lambda\in\sigma(T)\setminus\{1\}\}$$

für jede Vektornorm in R^n*.*

Definition 2.7 *Konvergiert ein Iterationsverfahren mit den Iterierten* $x_0, x_1, \ldots$ *gegen einen Grenzwert* x^**, so heißt*

$$\limsup_{k\to\infty}\left\{\sup_{x_0\neq x^*}\left[\frac{\|x^*-x_k\|}{\|x^*-x_0\|}\right]^{1/k}\right\} \tag{16}$$

der Konvergenzfaktor *des Verfahrens.*

Bemerkung 2.7

- *Da der Konvergenzfaktor ein asymptotisches Verhalten beschreibt, heißt (16)* asymptotischer Konvergenzfaktor *oder auch Konvergenzrate*[4]*.*
- *Für ein reguläres lineares Gleichungssystem entspricht (16) gerade dem Spektralradius der Iterationsmatrix.*

Aus der Theorie der semiiterativen Methoden (vgl. [9] — für die spezielle Anwendung auf Markov–Ketten vgl. auch [38]) folgt eine Variante des obigen Satzes.

Korollar 2.14 *Das System* $x = Tx + c$ *sei singulär aber lösbar und es sei* $c \in$ *bild*$((I-T)^q,\quad q = ind(I-T)$*. Die Fixpunktiteration (15) konvergiert für alle* $x_0 \in$ *bild*$(I-T)^q +$ *kern*$(I-T)$ *genau dann gegen eine Lösung, wenn folgende zwei Bedingungen erfüllt sind:*

- $\rho(T) = 1$*,*
- $\lambda = 1$ *ist einziger Eigenwert von* T *mit Betrag* 1*.*

[4] Nach [53, S. 587] ist diese definiert durch den negativen Logarithmus zur Basis 10 des Konvergenzfaktors.

Darüberhinaus gilt (mit $x^* = \lim_{k\to\infty} x_k$*)*

$$\limsup_{k\to\infty}\left\{\sup_{x_0\neq x^*}\left[\frac{\|x^*-x_k\|}{\|x^*-x_0\|}\right]^{1/k}\right\} = \max\{|\lambda| : \lambda\in\sigma(T)\setminus\{1\}\}$$

für jede Vektornorm in R^n*.*

Angewendet auf das SOR–Verfahren (5) und das lineare Gleichungssystem $Ax = b$ ergibt sich, dass $\mathcal{L}_\omega$ die Rolle der Matrix T übernimmt, ferner ist

$$\begin{aligned} I-\mathcal{L}_\omega &= I-(D-\omega L)^{-1}((1-\omega)D+\omega U)\\ &= (D-\omega L)^{-1}(D-\omega L)-(D-\omega L)^{-1}((1-\omega)D+\omega U)\\ &= (D-\omega L)^{-1}(\omega D-\omega L-\omega U)\\ &= \omega(D-\omega L)^{-1}A \end{aligned}$$

und $c = (D-\omega L)^{-1}b \in \text{bild}((\omega(D-\omega L)^{-1}A)^q$, weil das lineare System als konsistent vorausgesetzt war.

Für die Anwendung des SOR–Verfahrens auf solche Systeme folgt also, dass Konvergenz vorliegt, wenn der dominante Eigenwert von $\mathcal{L}_\omega$ 1 ist und kein weiterer Eigenwert vom Betrag 1 vorliegt. Der Konvergenzfaktor entspricht (asymptotisch) dem Betrag des subdominanten Eigenwertes.

Als Besonderheit ist noch die Anwendung des SOR–Verfahrens auf homogene Systeme $Ax = 0$ mit nichttrivialem kern(A) zu nennen:

Hier entspricht die Anwendung des SOR–Verfahrens

$$x_{k+1} = \mathcal{L}_\omega x_k$$

gerade der Durchführung der Vektoriteration bezüglich der SOR–Matrix $\mathcal{L}_\omega$. Als Lösung des Systems wird der zum betragsgrößten Eigenwert von $\mathcal{L}_\omega$ gehörende Eigenvektor gefunden. Die Konvergenzgeschwindigkeit wird durch den Betrag des Quotienten zwischen subdominantem und dominantem Eigenwert bestimmt. Ist A singulär und nach Satz 2.14 lösbar, so ist der Konvergenzfaktor der Vektoriteration mit $\xi := \max\{|\lambda| : \lambda\in\sigma(T)\setminus\{1\}\}$ dann gerade $\frac{|\xi|}{1} = |\xi|$.

Inkonsistente singuläre Systeme

Im Fall eines inkonsistenten Systems (vgl. [61]) sucht man anstelle der Lösung von

$$Ax = b$$

eine Lösung des Systems

$$Ax = AA^{-}b, \tag{17}$$

wobei A^{-} eine generalisierte Inverse von A ist, d.h. insbesondere $AA^{-}A = A$ gilt. (17) besitzt mit $x = A^{-}b$ auf jeden Fall eine Lösung.

Für iterative Verfahren

$$x_{k+1} = Tx_k + g, \quad \text{bzw.} \quad x_{k+1} = x_k + R(b - Ax_k), \qquad k = 0, 1, 2, \ldots \tag{18}$$

mit geeigneter Matrix $R \in \mathbb{R}^{n \times n}$ gilt

Korollar 2.15 *([61, Corollary 12]) Für jedes $b \in \mathbb{R}^n$ erzeugt (18) eine Lösung von (17) genau dann, wenn ein R mit $I - T = RA$ und $c = Rb$ existiert und für R gilt*

1. *$kern(RA) = kern(A)$,*
2. *$bild(RA) = bild(A)$ und*
3. *$\rho(T|_{bild(R)}) < 1$ oder T ist konvergent (d.h. $\lim_{k \to \infty} T^i$ existiert).*

Wie bereits erwähnt gelangt man durch Betrachtung von $A^T Ax = A^T b$ zu einer Kleinste–Quadrate–Lösung.

2.3 Grundlagen zu Markov–Ketten

Es werden die notwendigen Grundlagen über Markov–Ketten bereitgestellt. Insbesondere wird gezeigt, dass bei Berechnung des stationären Wahrscheinlichkeitsvektors ein homogenes lineares Gleichungssystem $Ax = 0$ mit singulärer Matrix A mit $\dim(\text{kern}(A)) = 1$ gelöst werden muss. Dabei ist die triviale Lösung $x = 0$ nicht von Interesse. Ferner geht es speziell um p–periodische Markov–Ketten. Verwiesen sei dabei auf [57].

Grundlegende Begriffe

Gegeben sei ein System mit einer (nicht notwendigerweise endlichen) Menge von Zuständen, in denen sich das System befinden kann. Zu jedem festen Zeitpunkt befinde sich das System in genau einem dieser Zustände. Ferner seien Übergangswahrscheinlichkeiten bzw. Übergangsraten zwischen diesen Zuständen gegeben. Es wird dabei immer angenommen, dass die Übergänge eine infinitesimal kleine Zeit in Anspruch nehmen.

Ein solches System heißt *Markov–Prozess*, wenn die weitere Entwicklung des Systems nur von dem Zustand abhängt, in dem sich das System augenblicklich befindet, nicht aber von früheren Zuständen. Die letztgenannte Eigenschaft heißt auch *Markov–Eigenschaft* und man spricht von der *Gedächtnislosigkeit des Prozesses*.

Ist der Zustandsraum diskret, spricht man von einer *Markov–Kette* (vgl. [57]).

In der Sprache der Stochastik ist ein Markov–Prozess eine Familie von Zufallsvariablen $\{X(t), t \in T\}$ auf gegebenem Wahrscheinlichkeitsraum mit der Markov–Eigenschaft

$$P(X(t) \leq x | X(t_0) = x_0, \ldots X(t_n) = x_n) \quad = \quad P(X(t) \leq x | X(t_n) = x_n)$$

für $t_0 < t_1 < \ldots < t_n < t$.

Hängen die Übergangswahrscheinlichkeiten der Markov–Kette nicht von der Zeit ab, spricht man von einer *homogenen* Markov–Kette.

Die Zustände der Markov–Kette lassen sich weiter charakterisieren:

- Ein Zustand, in den das System unendlich oft zurückkehrt, heißt *rekurrent*.
- Ein Zustand, für den die Wahrscheinlichkeit, dass das System nie in ihn zurückkehrt, größer als 0 ist, heißt *transient*.

- Ein Zustand, der nach Erreichen nicht mehr verlassen wird, heißt *absorbierend*.

Äquivalente Bezeichnungen gibt es jeweils für Mengen von Zuständen.

Ferner lässt sich das System weiter charakterisieren:

- Ein System, das genau nach kp ($k, p \in \mathbb{N}$) Schritten wieder in den Anfangszustand zurückkehrt, heißt *p–periodisch*.
- Ein System, in dem jeder Zustand von jedem Zustand aus erreichbar ist (eventuell nach Durchlaufen von Zwischenzuständen), heißt *ergodisch*.

Man unterscheidet *zeitdiskrete* ($T = \{0, 1, 2, \ldots\}$) und *zeitkontinuierliche* ($T = \{t : 0 \leq t \leq \infty\}$) Markov–Ketten.

Grundlagen zu zeitdiskreten Markov–Ketten

Im zeitdiskreten Fall sind die Übergänge durch (Einzelschritt–)Übergangswahrscheinlichkeiten charakterisiert. Dabei handelt es sich um bedingte Wahrscheinlichkeiten und es ist (im homogenen Fall)

$$P(X_{n+1} = j | X_n = i) = p_{ij}.$$

Im inhomogenen Fall sind die Größen jeweils noch von t abhängig.

Da es sich um Wahrscheinlichkeiten handelt, muss gelten

$$0 \leq p_{ij} \leq 1 \quad \text{und} \quad \sum_j p_{ij} = 1 \quad \forall i.$$

Daraus folgt, dass $P = (p_{ij})$ eine stochastische Matrix ist.

Befindet sich das System im n–ten Schritt im Zustand i, hat also die Wahrscheinlichkeitsverteilung $e_i := (0, \ldots, 0, \underbrace{1}_{i}, 0, \ldots, 0)^T$, so hat das System im $(n+1)$–ten Schritt die Wahrscheinlichkeitsverteilung

$$e_i^T P = p_{i\cdot}.$$

Ferner gelten die (zeitdiskreten) Chapman–Kolmogorov–Gleichungen

$$p_{ij}^{(n)} = \sum_k p_{ik}^{(l)} p_{kl}^{(n-l)} \quad 0 < l < n$$

bzw.

$$P^n = P^l P^{n-l}.$$

Ist das System zu Beginn in einer Ausgangsverteilung $\pi^T(0)$, so gilt also

$$\pi^T(n) = \pi^T(0) P^{(n)} = \pi^T(0) P^n.$$

Dabei soll die Schreibweise $p_{ij}^{(n)}$ die Wahrscheinlichkeit bezeichnen, dass man sich nach n Schritten im Zustand j befindet, wenn man sich momentan im Zustand i befindet. Entsprechend bezeichnet die Matrix P^n zunächst die Übergangswahrscheinlichkeitsmatrix nach n Schritten. Aufgrund der Chapman–Kolmogorov–Gleichungen ist dies aber nichts anderes als die p–te Potenz der Matrix P.

Aufgrund der Herleitung ist P eine nichtnegative, stochastische Matrix. Ist P ferner irreduzibel, kann man den Satz von Perron–Frobenius anwenden:

Satz 2.16 (Perron–Frobenius) *[64, Theorem 2.1] Sei $A \in \mathbb{R}^{n \times n} \geq 0$[5] und irreduzibel. Dann gilt:*

1. *Der Spektralradius $\rho(A)$ ist ein einfacher Eigenwert von A (der sogenannte* Perron–Eigenwert*). Außerdem ist $\rho(A) > 0$.*
2. *Zum Perron–Eigenwert gehört ein positiver Perron–Eigenvektor $x > 0$.*
3. *Jeder nichtnegative Eigenvektor von A ist Eigenvektor zum Perron–Eigenwert $\rho(A)$ (und ist damit echt positiv).*

Nach dem Satz von Perron–Frobenius hat P also einen Eigenwert $\rho(P) > 0$ mit positivem Eigenvektor.

Weil P stochastisch ist, gilt $Pe = e$, wobei e der Vektor aus lauter Einsen ist. Es gilt $e > 0$, woraus nach Teil 3 von Satz 2.16 sofort folgt, dass

$$\rho(P) = 1$$

[5] Die Schreibweise $A \geq 0$ soll bedeuten, dass A komponentenweise nichtnegativ ist. Analoges gilt auch für Vektoren, das heißt, dass $x > 0$ bedeutet, dass x komponentenweise positiv ist.

sein muss. Damit existiert auch ein

$$\pi^* > 0 \text{ mit} \quad (\pi^*)^T P = (\pi^*)^T, \tag{19}$$

also ein positiver Linkseigenvektor. In der Sprache der Markov–Ketten heißt dieser Vektor *stationärer (Wahrscheinlichkeits–)Verteilungsvektor* der Kette.

Grundlagen zu zeitkontinuierlichen Markov–Ketten

Da die Zeit hier als kontinierlich vorausgesetzt wird, muss dies auch bei der Aufstellung der Matrix beachtet werden. Die Übergänge sind nicht mehr durch Wahrscheinlichkeiten gekennzeichnet, sondern sie sind jeweils einer Wahrscheinlichkeitsverteilung unterworfen. Bei dieser muss es sich aufgrund der Markov–Eigenschaft um die Exponential–Verteilung handeln.

Man betrachtet statt der Übergangswahrscheinlichkeiten nun

$$q_{ij}(t) := \lim_{\Delta t \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t} \qquad i \neq j,$$

die sogenannten *Übergangsraten.*

Aufgrund der Tatsache, dass $\sum_j p_{ij} = 1$ sein muss, ergibt sich ferner

$$q_{ii}(t) = -\sum_{\substack{j \\ j \neq i}} q_{ij}(t). \tag{20}$$

Die Matrix $Q = (q_{ij})$ heißt *infinitesimaler Generator*. Es ist

$$Q(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t, t+\Delta t) - I}{\Delta t} \right\}.$$

Die Übergangsraten des infinitesimalen Generators[6] haben den Charakter von Ableitungen: die Wahrscheinlichkeit eines Übergangs von i nach j steigt mit der Größe von Δt ($q_{ij} > 0$), die Wahrscheinlichkeit des Verbleibens im Zustand i fällt ($q_{ii} < 0$).

[6] Im Weiteren wird diese Matrix auch einfach kurz als *Generator* bezeichnet.

Aufgrund der Herleitung (die Zeilensumme in jeder Zeile ist danach 0) ist die Matrix Q singulär und die Matrix

$$\tilde{Q} = Q + I$$

stochastisch[7], d.h. es existiert ein

$$\begin{array}{lll} x > 0 \text{ mit} & x^T(Q+I) & = x^T \\ \text{d.h.} & x^T Q & = 0^T. \end{array}$$

Dieser Vektor x mit $x^T Q = 0$ ist für zeitkontinuierliche Systeme der stationäre Wahrscheinlichkeitsvektor.

Das Analogon der Chapman–Kolmogorov–Gleichung lautet hier

$$\frac{\partial \pi(t)}{\partial t} = \pi^T(t) Q(t)$$

und es ist

$$\pi^T(t) = \pi^T(0) e^{Qt}.$$

Zur Berechnung bzw. Näherung von e^{Qt} vgl. als Übersicht [57], für die Anwendung von Krylov–Verfahren [29] und [30].

Insgesamt:

Zur Bestimmung des stationären Wahrscheinlichkeitsvektors muss das System

$$\pi^T P = \pi^T \quad \text{bzw.} \quad \pi^T Q = 0^T \tag{21}$$

gelöst werden, das heißt also ein homogenes lineares Gleichungssystem mit eindimensionalem Kern der Matrix. Insbesondere können damit alle Sätze und Ergebnisse aus Abschnitt 2.2 für konsistente singuläre Systeme angewendet werden.

p–periodische Markov–Ketten

Wie bereits erwähnt heißt ein System, das nach kp ($k, p \in \mathbb{N}$) Schritten wieder in den Ausgangszustand zurückkehrt, p–periodisch.

[7] Gegebenenfalls muss die Matrix vorher noch diagonalskaliert werden.

Insbesondere führt eine p–periodische Markov–Kette auf eine p–zyklische Übergangsmatrix (vgl. Definition 2.3) und damit können alle bekannten Sätze aus Abschnitt verwendet werden, insbesondere aber gilt:

- die von Null verschiedenen Eigenwerte der Jacobi–Matrix treten in p–Tupeln auf (Satz 2.7),
- insbesondere sind damit die p–ten Einheitswurzeln Eigenwerte dieser Matrix
- bei Anwendung des SOR–Verfahrens gilt Satz 2.8.

3 SOR bei p–periodischen Markov–Ketten mit reellem Spektrum

Es wird die SOR–Konvergenz bei Anwendung des Verfahrens auf konsistent geordnete p–zyklische Matrizen, wie sie bei Markov–Ketten auftreten, untersucht. Zunächst wird der Zusammenhang zwischen SOR bei p–zyklischen Matrizen und einem p–Schritt–Relaxationsverfahren betrachtet, wodurch man Bereiche der komplexen Ebene angeben kann, in denen das Spektrum der Jacobi–Matrix liegen muss, sodass man überhaupt Konvergenz erwarten kann. Von Bedeutung ist dabei, dass das lineare Gleichungssystem konsistent und singulär ist, was bei der Berechnung des stationären Wahrscheinlichkeitsvektors einer p–periodischen Markov–Kette sicher erfüllt ist.

Es erweist sich, dass das Spektrum der Jacobi–Matrix mit Ausnahme des p–Tupels vom Betrag 1 enthalten sein muss im Innenbereich eines Hypozykloids, weswegen die Betrachtung dieser Rollkurven sich im nächsten Abschnitt anschließt.

Mit Hilfe der Betrachtung von Hypozykloiden gelangt man zu Aussagen über die optimale Wahl des Relaxationsparameters (so in [65], [11], [18] und [19]) und über den zugehörigen optimalen Konvergenzfaktor.

Bei der Berechnung des stationären Wahrscheinlichkeitsvektors einer p–periodischen, zeitkontinuierlichen Markov–Kette liegt eine singuläre Matrix Q vor. Kapitel 3.4 widmet sich der Frage nach der Konvergenz und führt die *extended convergence* oder verallgemeinerte Konvergenz von Kontovasilis et al. (vgl. [34]) ein, wobei gezeigt wird, dass es sich dabei im Wesentlichen um Vektoriteration handelt.

Der nächste Abschnitt behandelt die Frage, wo beim erweiterten SOR–Verfahren neben 1 und dem dominanten Eigenwert die anderen dem p–Tupel 1 von J^p zugeordneten Eigenwerte liegen. Diese Frage ist insofern von Bedeutung, als dadurch gezeigt wird, dass sie mit der Konvergenz und der Konvergenzgeschwindigkeit nichts zu tun haben, weil sie betragsmäßig kleiner als der subdominante Eigenwert sind.

Alle diese bisherigen Abschnitte gehen davon aus, dass das Spektrum von J^p

1. reell (bzw. im Einzelfall nichtnegativ oder nichtpositiv ist) und
2. in einem bekannten Intervall $[\alpha, \beta]$ liegt.

Gerade der letzte Punkt wird in der Praxis nicht zu erfüllen sein. Deshalb wird im

letzten Abschnitt dieses Kapitels die Frage behandelt, um wieviel schlechter die Konvergenzgeschwindigkeit wird, wenn man das Gauß–Seidel–Verfahren verwendet und auf Relaxation verzichtet.

Das Ausgangssystem, das zu lösen ist, sei immer $Ax = b$ bzw. in Fixpunktform $x = Tx + c$.

3.1 p–Schritt Relaxation und SOR

Ziel muss zunächst sein, Aussagen über das Konvergenzverhalten und die Konvergenzgeschwindigkeit des SOR–Verfahrens zu treffen. Im Allgemeinen ist das nicht möglich, in speziellen Fällen — wie bei p–zyklischen Matrizen — aber sehr wohl.

Aus der Literatur (vgl. insbesondere [46]) ist bekannt, dass man für eine andere Klasse von Verfahren Konvergenzaussagen kennt, die auf den ersten Blick nichts mit dem SOR–Verfahren zu tun zu haben scheint.

Es handelt sich dabei um Verfahren der Klasse[8]

$$\mathcal{E}_p: \qquad y^{(k)} = \mu_0(Ty^{(k-1)} + c) + \mu_1 y^{(k-1)} + \ldots + \mu_p y^{(k-p)} \tag{22}$$

mit $k > p, y^{(0)}, \ldots, y^{(p-1)} \in \mathbb{R}^n, \mu_0 \neq 0$ und $\sum_{j=0}^{p} \mu_j = 1$.

Zu jedem Verfahren dieser Klasse gehört eine rationale Funktion

$$R(\varphi) = \frac{1 - \mu_1\varphi - \ldots - \mu_p\varphi^p}{\mu_0\varphi}. \tag{23}$$

Sei $\overline{D}_1$ die abgeschlossene Einheitskreisscheibe in der komplexen Ebene. Dann gilt nach [46, Theorem 1]

Satz 3.1 *Ist R schlicht in $\overline{D}_1$, dann gilt:*

Ist

$$\sigma(T) \subseteq U(R) := \overline{\mathbb{C}} \setminus R(\overline{D}_1),$$

dann konvergiert das Verfahren (22).

[8] Das ist ein Spezialfall der allgemeineren Euler–Verfahren.

Aus der Vorschrift (22) erkennt man, dass sich $y^{(k)}$ mittels eines Polynoms v_k in der Form

$$y^{(k)} = v_k(T)c$$

darstellen lässt.

Nach [46] ist der asymptotische Konvergenzfaktor in Abängigkeit der Matrix T und der Polynome v_k dann

$$\kappa(T) = \max_{\tau_i \in \sigma(T)} \overline{\lim_{k\to\infty}} |v_k(\tau_i)|^{1/k}. \tag{24}$$

Dies ist gerade der in Definition 2.7 eingeführte Konvergenzfaktor.

Damit gilt (vgl. auch [46, Corollary 2]):

Satz 3.2 *Ist R schlicht in D_η $(\eta > 1)$ und ist $\sigma(T) \subseteq U_\eta(R) := \overline{\mathbb{C}} \setminus R(D_\eta)$, so gilt*

$$\kappa(T) \leq \frac{1}{\eta},$$

d.h. dass (22) mindestens mit dem Konvergenzfaktor $\frac{1}{\eta}$ konvergiert.

Liegt mindestens ein Eigenwert von T auf $\partial U_\eta(R)$, so ist der Konvergenzfaktor genau $\frac{1}{\eta}$.

Durch die Setzungen $T := J, \mu_0 = \omega \in \mathbb{R}, \mu_2 = \ldots = \mu_{p-1} = 0, \mu_p = 1 - \omega$ erhält man die speziellere Klasse

$$y^{(k)} = \omega J y^{(k-1)} + (1-\omega) y^{(k-p)} + \omega c, \qquad k \geq p, \ldots \tag{25}$$

mit $c = D^{-1}b$ und gegebenen Startvektoren $y^{(0)}, \ldots, y^{(p-1)}$.

Dieser *p–Schritt–Relaxation* zugeordnet ist analog zu oben die rationale Funktion

$$R(\varphi) = R_{\omega,p}(\varphi) = \frac{1 - (1-\omega)\varphi^p}{\omega\varphi}. \tag{26}$$

Das dieser Funktion zugeordnete Verfahren gehört zur Menge $\mathcal{E}_p$, die diesbezüglichen Ergebnisse der Sätze 3.1 und 3.2 können also verwendet werden.

Die Ränder der mit Hilfe von (26) gegebenen Bereiche $U(R_{\omega,p}), U_\eta(R_{\omega,p})$ sind Hypozykloiden (zu dieser Klasse von Kurven vgl. auch das nächste Kapitel 3.2).

Die Gleichungen der Randkurve erhält man durch Einsetzen von $\eta e^{i\vartheta}, (0 \leq \vartheta < 2\pi)$ in (26).

Es ergibt sich:

$$\mathrm{Re}(R_{\omega,\eta,p}(\vartheta)) = \frac{1}{\omega}\left(\frac{1}{\eta}\cos(\vartheta) + (\omega-1)\eta^{p-1}\cos((p-1)\vartheta)\right) \quad (27)$$

$$\mathrm{Im}(R_{\omega,\eta,p}(\vartheta)) = -\frac{1}{\omega}\left(\frac{1}{\eta}\sin(\vartheta) - (\omega-1)\eta^{p-1}\sin((p-1)\vartheta)\right). \quad (28)$$

Durch Abbildung der abgeschlossenen Einheitskreisscheibe mittels (26) erhält man das abgeschlossene Außengebiet der durch (27) und (28) für $\eta = 1$ gegebene Kurve.

Das Verfahren (22) konvergiert also nicht, wenn ein Eigenwert von T in $R(\overline{D}_1)$ liegt. Umgekehrt konvergiert (22), wenn alle Eigenwerte in dem offenen Komplement $\overline{\mathbb{C}} \setminus R(\overline{D}_1)$ liegen.

Analog konvergiert das Verfahren nicht mit dem Konvergenzfaktor $\frac{1}{\eta}$, wenn ein Eigenwert von T in $R(D_\eta)$ liegt. Es konvergiert mindestens mit dem Konvergenzfaktor $\frac{1}{\eta}$, wenn alle Eigenwerte in dem abgeschlossenen Komplement $\overline{\mathbb{C}} \setminus R(D_1)$ liegen, und genau mit dem Konvergenzfaktor $\frac{1}{\eta}$, wenn ein Eigenwert auf dem Rand liegt.

Zur Veranschaulichung diene Abbildung 1, in der neben dem Einheitskreis die durch (27) und (28) gegebenen Kurven mit $p = 5$, $\omega = 1.1$ und $\eta = 1$ (durchgezogene Linie) bzw. $\eta = 1.1$ (gestrichelte Linie) eingezeichnet sind. Liegen die Eigenwerte von T innerhalb der durchgezogenen Linie (und nicht auf dem Rand), konvergiert das die p–Schritt–Relaxation. Liegen sie innerhalb der gestrichelten Linie, ist der Konvergenzfaktor mindestens $\frac{1}{1.1}$, liegt ein Eigenwert genau auf der gestrichelt gezeichneten Kurve, liegt genau dieser Konvergenzfaktor vor.

Wir kommen nun zum Zusammenhang der p–Schritt–Relaxation mit dem SOR–Verfahren:

Betrachtet man von den nach (25) berechneten, analog zur Matrix J partitionierten Vektoren $y^{k+1}, \ldots, y^{k+p}$ vom Vektor y^{k+s} die $s.$–te Block–Komponente, so sieht man, dass man durch Zusammenfassen dieser p Komponenten in einem Vektor genau einen SOR–Schritt — angewendet auf einen ähnlich zusammengesetzten Vektor aus $y^{k+1-p}, \ldots, y^k$ beschreibt (in [65, S.33f] wird das etwas ausführlicher ausgeführt).

Man erhält also den wichtigen Satz

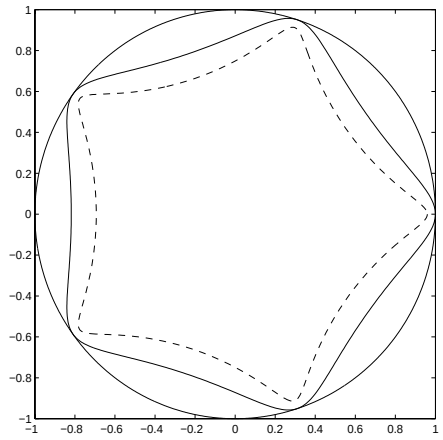

Abbildung 1: (27) und (28) mit $\omega = 1.1, \eta = 1$ und $\eta = 1.1$

Satz 3.3 *Die p–Schritt–Relaxation (25) auf ein lineares System $Ax = b$ mit schwach zyklischer Jacobi–Matrix J vom Index p angewendet konvergiert für beliebigen Startvektor genau dann gegen die Lösung des Systems, wenn das SOR–Verfahren konvergiert.*

Im Fall der Konvergenz konvergiert das SOR–Verfahren p mal schneller.

Man kann also für die Konvergenzanalyse des SOR–Verfahrens gerade die Ergebnisse aus den Sätzen 3.1 und 3.2 verwenden, wobei anstelle des Konvergenzfaktors $\frac{1}{\eta}$ aus Satz 3.2 nunmehr $\frac{1}{\eta^p}$ zu setzen ist.

3.2 Hypozykloide

In diesem Abschnitt sollen einige grundlegende Eigenschaften von Hypozykloiden zusammengefasst werden (vgl. auch [14][9], [52]).

Rollt ein kleinerer Kreis (Mittelpunkt M, Radius r) auf der Innenseite eines größeren (OBdA Mittelpunkt 0, Radius R) ab, so beschreibt jeder Punkt P mit festem Abstand h vom Mittelpunkt M eine Hypozykloide (vgl. Abbildungen 2 und 3) bzw. ein Hypozykloid.

Das Hypozykloid ist geschlossen, wenn das Verhältnis $\frac{R}{r}$ ganzzahlig ist. In diesem Fall besteht es aus $\frac{R}{r}$ kongruenten Bögen. Im Folgenden sollen nur solche Hypozykloide betrachtet werden.

Man unterscheidet drei Klassen von Hypozykloiden in Abhängigkeit von r und h:

(i) verkürztes Hypozykloid für $h < r$, d.h. der Punkt P liegt im Innern des rollenden Kreises,

(ii) gewöhnliches Hypozykloid für $h = r$, d.h. der Punkt P liegt auf dem Rand des rollenden Kreises,

(iii) gestrecktes Hypozykloid für $h > r$, d.h. der Punkt P liegt außerhalb des rollenden Kreises. In diesem Fall besitzt das Hypozykloid Schlaufen.

Die Gleichung des Hypozykloids ist — mit $\vartheta \in [0, 2\pi)$ — aufgeteilt in Real– und Imagionärteil

$$x(\vartheta) = (R-r)\cos(\vartheta) + h\cos\left(\frac{R-r}{r}\vartheta\right) \tag{29}$$

$$y(\vartheta) = (R-r)\sin(\vartheta) - h\sin\left(\frac{R-r}{r}\vartheta\right). \tag{30}$$

Die Abbildungen 2 und 3 zeigen die drei Fälle. Gestrichelt ist jeweils die Kurve gezeichnet, die der Punkt P (gekennzeichnet durch einen Stern) beschreibt, wenn der kleine Kreis auf dem großen abrollt.

[9] In diesem Buch ist die Bezeichnungsweise allerdings nicht ganz dieselbe wie im Weiteren und in der neueren Literatur. Gestreckte Hypozykloide werden als Trochoide bezeichnet. Gewöhnliche und verkürzte Hypozykloide als Zykloidale. In der Literatur hat sich aber die Bezeichnung *Zykloide* durchgesetzt.

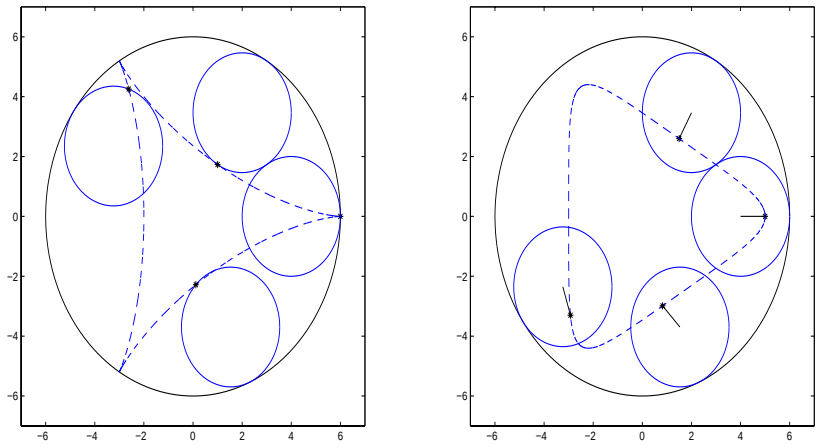

Abbildung 2: Gewöhnliches und verkürztes Hypozykloid

Ist $h > R - r$, so ist das Hypozykloid leer. Die Sprechweise *„das Hypozykloid ist leer“* soll bedeuten, dass das vom Hypozykloid umschlossene (Innen–)Gebiet leer ist.

Ferner gilt:

Satz 3.4 *Ist P ein Punkt des Hypozykloid, so gilt für den Betrag $|P|$ des Punktes, dass*

$$R - r - h \leq |P| \leq R - r + h.$$

Beweis:
Die Behauptung folgt unmittelbar aus der Gleichung unter Beachtung des Wertebereichs der Kreisfunktionen. □

Von besonderer Bedeutung in [18] sind auch die Halbachsen a und b eines Hypozykloids.

b heißt reelle, a imaginäre Halbache des Hypozykloids, und es gilt

$$x(\vartheta) = \frac{b+a}{2}\cos(\vartheta) + \frac{b-a}{2}\cos((p-1)\vartheta),$$

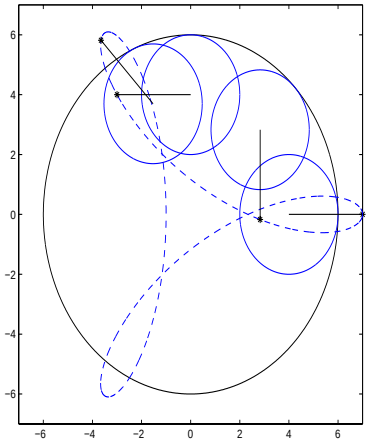

Abbildung 3: Gestrecktes Hypozykloid

$$y(\vartheta) \quad = \quad -\frac{b+a}{2}\sin(\vartheta) + \frac{b-a}{2}\sin((p-1)\vartheta).$$

Vergleicht man diese Darstellung mit der bisherigen, so gilt

$$\begin{aligned} a &= R - r - h, \\ b &= R - r + h. \end{aligned}$$

Die beiden Werte bezeichnen also nach Satz 3.4 anschaulich den Inkreis– bzw. Umkreisradius des Hypozykloids.

Wird $h > 0$ nicht vorausgesetzt[10], kann man feststellen, ob das Hypozykloid im Gegensatz zum gewöhnlichen Hypozykloid nach (27),(28) um den Winkel $\frac{\pi}{p}$ gedreht ist, nämlich genau dann, wenn $a > b$. Die Ecke eines gewöhnlichen Hypozykloids mit drei kongruenten Bögen liegt dann nicht mehr auf der reellen Achse sondern auf der imaginären.

[10] $h > 0$ ist in der Herleitung als Rollkurve zwar eine sinnvolle Voraussetzung, im Weiteren wird sich aber zeigen, dass es günstiger sein kann, auf diese Einschränkung zu verzichten.

Zusammenhang zwischen Hypozykloiden und der SOR–Konvergenz

Nach dem letzten Abschnitt liegt für ω Konvergenz mit einem Konvergenzfaktor nicht schlechter als $\left(\frac{1}{\eta}\right)^p$ vor, wenn alle Eigenwerte aus $\tilde{\sigma}(J^p)$ in dem Gebiet $U_\eta(\omega)$ liegen. Genau liegt dieser Konvergenzfaktor vor, wenn wenigstens ein Eigenwert aus $\tilde{\sigma}(J^p)$ auf der dieses Gebiet berandenden Kurve liegt, alle anderen — außer den zum p–fachen Eigenwert 1 gehörenden — müssen im Inneren des umrandeten Gebietes liegen.

Nach (27) und (28) (vgl. auch [65]) gelten für diese Randkurve, die durch $q_\omega(\eta e^{i\vartheta})$ beschrieben wird, die Gleichungen

$$\mathrm{Re}(q_\omega(\eta e^{i\vartheta})) = \frac{1}{\omega}\left(\frac{1}{\eta}\cos(\vartheta) + (\omega-1)\eta^{p-1}\cos((p-1)\vartheta)\right), \tag{31}$$

$$\mathrm{Im}(q_\omega(\eta e^{i\vartheta})) = -\frac{1}{\omega}\left(\frac{1}{\eta}\sin(\vartheta) - (\omega-1)\eta^{p-1}\sin((p-1)\vartheta)\right). \tag{32}$$

Dies ist gerade die Gleichung eines Hypozykloids (vgl. (29) und (30)), der Zusammenhang wird mit

$$R = \frac{p}{(p-1)\omega\eta}, \quad r = \frac{1}{(p-1)\omega\eta}, \quad h = \frac{\omega-1}{\omega}\eta^{p-1} \tag{33}$$

klar.

Für $\omega < 0$ (dies wird im späteren Abschnitt 3.4 benötigt) ist entsprechend

$$R = -\frac{p}{(p-1)\omega\eta}, \quad r = -\frac{1}{(p-1)\omega\eta}, \quad h = \frac{\omega-1}{\omega}\eta^{p-1}, \tag{34}$$

zu setzen bzw. für $\omega \in (0,1)$

$$R = \frac{p}{(p-1)\omega\eta}, \quad r = \frac{1}{(p-1)\omega\eta}, \quad h = \frac{1-\omega}{\omega}\eta^{p-1}. \tag{35}$$

In allen drei Fällen gilt[11]

$$\frac{R}{r} = \frac{\frac{p}{(p-1)\omega\eta}}{\frac{1}{(p-1)\omega\eta}} = p,$$

d.h. das Hypozykloid ist eine geschlossene Kurve und besteht aus p kongruenten Bögen (wie auch bei den Abbildungen aus diesem Kapitel).

[11] Hier am Beispiel $\omega > 1$.

In folgender Tabelle soll für die Fälle $\omega < 1$ und $\omega > 1$ zusammengefasst werden, für welche Beziehungen zwischen den Größen p, ω und η welche Art von Hypozykloid vorliegt:

Art des Hypozykloid	$\omega < 1$	$\omega > 1$
verkürzt $(h < r)$	$(p-1)(1-\omega) < \frac{1}{\eta^p}$	$(p-1)(\omega-1) < \frac{1}{\eta^p}$
gewöhnlich $(h = r)$	$(p-1)(1-\omega) = \frac{1}{\eta^p}$	$(p-1)(\omega-1) = \frac{1}{\eta^p}$
gestreckt $(R-r > h > r)$	$(p-1)(1-\omega) > \frac{1}{\eta^p}$	$(p-1)(\omega-1) > \frac{1}{\eta^p}$
leer $(h > R-r)$	$\frac{1}{\eta^p} < 1-\omega$	$\frac{1}{\eta^p} < \omega-1$

Die Bedingungen folgen unmittelbar, indem man für R, r und h die jeweils zutreffenden Größen gemäß (33), (34) oder (35) einsetzt.

Im Grenzfall $h = R - r$ ist das Innere des Hypozykloids (von den Schlaufen abgesehen, deren Inneres aber ohnehin nicht zum Innengebiet des Hypozykloids zu rechnen ist) zusammengezogen auf einen Punkt, für $h > R - r$ überlappen die Schlaufen und es gibt in diesem Sinne kein Inneres mehr (vgl. die beiden Abbildungen), das bedeutet insbesondere, dass das SOR–Verfahren mit dem Parameter ω nicht mit dem Konvergenzfaktor $\left(\frac{1}{\eta}\right)^p$ konvergiert, unabhängig davon, wo in der komplexen Ebene die Eigenwerte von J liegen.

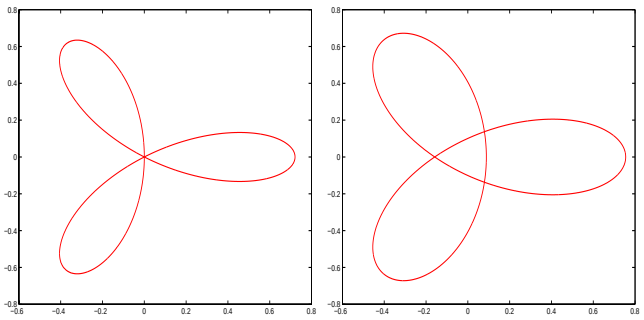

Abbildung 4: aus einem Punkt bestehendes Inneres und leeres Hypozykloid

Gesondert sollen noch kurz die Fälle $\omega = 0$ und $\omega = 1$ behandelt werden:

zunächst $\omega = 0$:

aus der Gleichung (12) von Varga (vgl. [64]) folgt die Beziehung

$$(\lambda - 1)^p = 0,$$

d.h. $\lambda = 1$ ist für jedes $\mu^p \in \sigma(J^p)$ p-facher Eigenwert der SOR–Matrix $\mathcal{L}_0$, d.h. das Verfahren konvergiert nicht. Der Fall $\omega = 0$ ist also irrelevant.

zu $\omega = 1$:

in diesem Fall reduzieren sich die Gleichungen des Hypozykloids zu

$$x(\vartheta) = \frac{\cos(\vartheta)}{\eta} \quad \text{bzw.} \quad y(\vartheta) = -\frac{\sin(\vartheta)}{\eta},$$

d.h. zu einem Kreis, die Eigenwerte von $\mathcal{L}_1$ sind wegen der Gleichung von Varga genau die Lösungen von $\lambda^p = \mu^p$.

Im Folgenden soll $H(\omega, \eta, p, .)$ jeweils die Bezeichnung für ein Hypozykloid mit p kongruenten Bögen und den Größen ω und η bezeichnen. ϑ läuft dabei von 0 bis 2π.

Beispiel 3.1 *Gegeben sei eine 3–zyklische Matrix A, deren Jacobi–Matrix mit Satz 2.7 die Eigenwerte* $\alpha e^{2ik\pi/3} \quad k = 0, 1, 2$ *mit* $\alpha \in \{1, 0.8, 0.4\}$ *besitzt.*

Nach Satz 2.10 ist der optimale Relaxationsparameter[12]

$$\omega_0 \approx 1.1013$$

und der zugehörige optimale Konvergenzfaktor

$$2(\omega_0 - 1) \approx 0.2026 \approx \frac{1}{\eta_0^3}, \qquad \textit{also} \qquad \eta_0 \approx 1.7025.$$

Abbildung 5 zeigt die Eigenwerte von J mit dem durch η_0 *und* ω_0 *bestimmten Hypozykloid.*

Man sieht, dass zum optimalen Relaxationsparameter ein gewöhnliches Hypozykloid gehört. Wählt man η *kleiner (d.h. geht man von einem besseren Konvergenzfaktor aus, als man hat), erhält man ein verkürztes Hypozykloid, dessen Inneres das zu* $\alpha = 0.8$ *gehörende Tripel nicht mehr enthält. Wählt man* η *größer (d.h. geht man von schlechterer Konvergenz aus, als man sie tatsächlich hat), bekommt man ein gestrecktes Hypozykloid.*

[12] Gerundet auf vier Dezimalen.

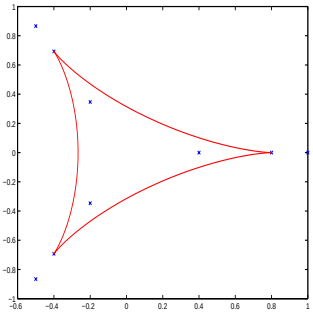

Abbildung 5: Eigenwerte von J und zu ω_0 und η_0 gehörendes Hypozykloid

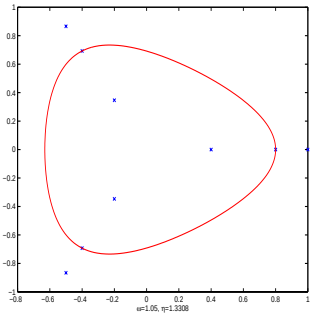

Abbildung 6: Hypozykloid mit $\omega = 1.05, \eta \approx 1.3308$

Wie sieht es nun aus, wenn nicht der optimale Relaxationsparameter verwendet wird, sondern z.B. $\omega = 1.05$ *? Auch zu diesem Relaxationsparameter gibt es einen eindeutig bestimmten Konvergenzfaktor (siehe Abbildung 6).*

In den nächsten Abbildungen sind weitere Hypozykloide zu $\omega = 1.05$ *und verschiedenen* η*–Werten gezeichnet. Man sieht, dass entweder keine Konvergenz mit diesem Konvergenzfaktor vorliegt (die zu* $\alpha = 0.8$ *gehörenden Eigenwerte sind außerhalb des Hypozykloid–Inneren und auch nicht auf dem Hypozykloid selbst) oder eigentlich bessere Konvergenz vorliegt als durch* η *angegeben (es liegen zwar alle Eigenwerte innerhalb des Hypozykloids, aber keine auf dem Rand).*

Zu jedem Relaxationsparameter ω *gibt es also genau ein* η*, das den tatsächlichen Konvergenzfaktor bezeichnet.*

Ähnliche Figuren ergeben sich (vgl. Abbildung 11), wenn man einen Relaxationsparameter wählt, der größer als der optimale ist (hier $\omega = 1.15$).

Bemerkung 3.1 *Zu bemerken ist, dass im Falle der Konvergenz des SOR–Verfahrens (auch des in Kapitel 3.4 verallgemeinerten Verfahrens) gilt, dass*

$$\eta = \left(\frac{1}{|\lambda_{subdominant}(\mathcal{L}_\omega)|} \right)^{\frac{1}{p}} \tag{36}$$

ist; dies folgt unmittelbar aus [34, Theorem 3.3], wenn man berücksichtigt, dass im Fall der Konvergenz der subdominante Eigenwert der SOR–Matrix gerade einem Eigenwert μ mit $|\mu| < 1$ der Jacobi–Matrix zugeordnet ist.

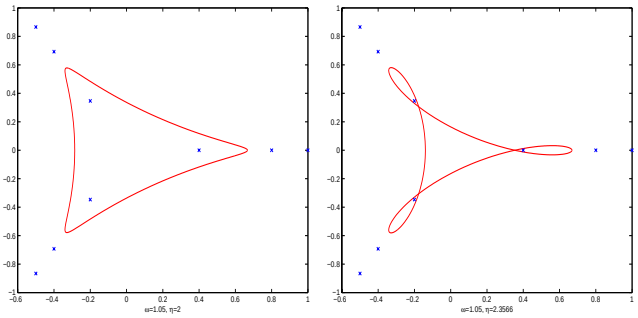

Abbildung 7: Hypozykloide für $\omega = 1.05$ und $\eta = 2$ bzw. $\eta = 2.3566$

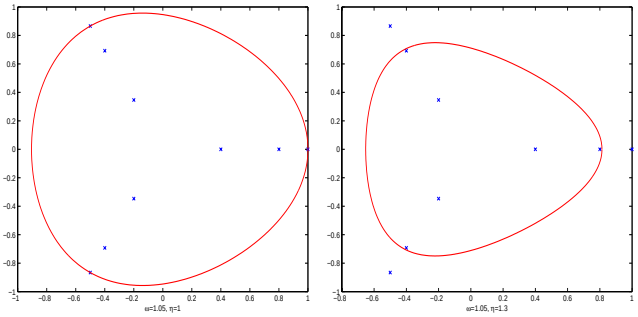

Abbildung 8: Hypozykloide für $\omega = 1.05$ und $\eta = 1$ bzw. $\eta = 1.3$

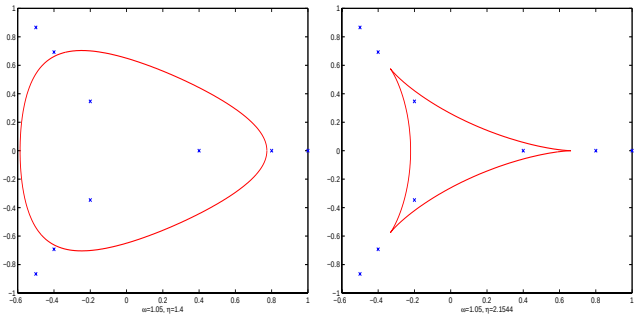

Abbildung 9: Hypozykloide für $\omega = 1.05$ und $\eta = 1.4$ bzw. $\eta = 2.1544$

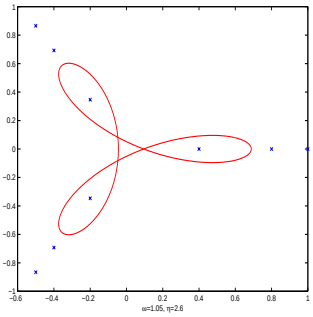

Abbildung 10: Hypozykloid für $\omega = 1.05$ und $\eta = 2.6$

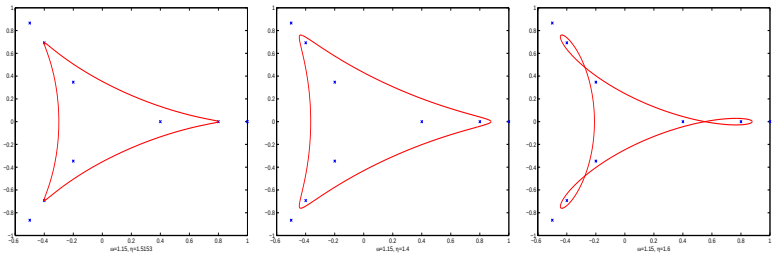

Abbildung 11: $\omega = 1.15, \quad \eta \approx 1.5153$ (optimal) $, 1.4, 1.6$

3.3 Aussagen über optimale Konvergenz

Die ersten Aussagen über optimale Konvergenz finden sich in dem bereits zitierten Satz 2.10.

Eine Erweiterten dieses Satzes um den Fall nichtpositiver Eigenwerte findet sich bei Wild und Niethammer (vgl. [65, Theorem 6]):

Korollar 3.5 *Die zu dem linearen Gleichungssystem $Ax = b$ gehörige Jacobi–Matrix sei schwach zyklisch vom Index p und die Eigenwerte von J^p seien nichtpositiv mit*

$$0 < \alpha = \rho(J) < \frac{p}{p-2}.$$

Der optimale Relaxationsparameter ω_0 des SOR–Verfahrens ist gegeben durch die eindeutige Lösung $\omega_0 \in (\frac{p-3}{p-2}, 1)$ der Gleichung

$$(p-1)^{p-1}\omega^p\alpha^p - p^p(1-\omega) = 0.$$

Für den Konvergenzfaktor ergibt sich in diesem Fall

$$\rho(\mathcal{L}_{\omega_0}) = (p-1)(1-\omega_0).$$

Dieser Satz ist — auch wenn historisch früher — eigentlich ein Korollar zu dem allgemeineren Satz, den in [11] Eiermann, Niethammer und Ruttan für den allgemeineren Fall $\sigma(J^p) \in [-\alpha^p, \beta^p]$ mit $\alpha, \beta \in \mathbb{R}_0^+$ herleiteten.

Dabei wurden Satz 2.10 und Korollar 3.5 zusammengefasst und präzisiert zu

Satz 3.6 *Die zu dem linearen Gleichungssystem $Ax = b$ gehörige Jacobi–Matrix sei schwach zyklisch vom Index p und die Eigenwerte von J^p seien enthalten in $[-\alpha^p, \beta^p]$ mit $0 \le \beta < 1$ und $0 \le \alpha < \frac{p}{p-2}$.*

1. *Ist $\alpha \le \frac{p-2}{p}\beta$ und $\rho(J) = \beta$, dann ist der optimale Relaxationsparameter ω_0 des p–zyklischen SOR–Verfahrens die eindeutige Lösung von*

   $$(p-1)^{p-1}\omega^p\beta^p - p^p(\omega-1) = 0$$

   im Intervall $(1, \frac{p}{p-1})$. Der Konvergenzfaktor ist

   $$\rho(\mathcal{L}_{\omega_0}) = (p-1)(\omega_0 - 1).$$

2. *Ist* $\beta \leq \frac{p-2}{p}\alpha$ *und* $\rho(J) = \alpha$*, dann ist der optimale Relaxationsparameter* ω_0 *des p–zyklischen SOR–Verfahrens die eindeutige Lösung von*

$$(p-1)^{p-1}\omega^p\beta^p - p^p(1-\omega) = 0$$

im Intervall $(\frac{p-3}{p-2}, 1)$*. Der Konvergenzfaktor ist*

$$\rho(\mathcal{L}_{\omega_0}) = (p-1)(1-\omega_0).$$

Ferner wird in [11] die noch verbleibende Lücke — nämlich die Lücke, dass $\frac{p-2}{p}\beta \leq \alpha < \beta$ oder $\beta < \alpha < \frac{p}{p-2}\beta$ ist — geschlossen:

Satz 3.7 *Die zu dem linearen Gleichungssystem* $Ax = b$ *gehörige Jacobi–Matrix sei schwach zyklisch vom Index p und die Eigenwerte von* J^p *seien enthalten in* $[-\alpha^p, \beta^p]$ *mit* $-\alpha^p, \beta^p \in \sigma(J^p)$*,* $0 \leq \beta < 1$ *und* $0 \leq \alpha < \frac{p}{p-2}$*.*

1. *Ist* $\frac{p-2}{p}\beta \leq \alpha < \beta$*, dann ist der optimale Relaxationsfaktor des p–zyklischen SOR–Verfahrens die eindeutige Lösung* ω_0 *von*

$$\left(\frac{\alpha+\beta}{2}\omega\right)^p - \frac{\alpha+\beta}{\beta-\alpha}(\omega-1) = 0$$

im Intervall $(1, 1+\frac{\beta-\alpha}{\alpha+\beta})$*. Für den optimalen Konvergenzfaktor ergibt sich*

$$\rho(\mathcal{L}_{\omega_0}) = \frac{\alpha+\beta}{\beta-\alpha}(\omega_0-1) = \left(\frac{\alpha+\beta}{2}\omega_0\right)^p.$$

2. *Ist* $\beta < \alpha < \frac{p}{p-2}\beta$*, dann ist der optimale Relaxationsfaktor des p–zyklischen SOR–Verfahrens die eindeutige Lösung* ω_0 *von*

$$\left(\frac{\alpha+\beta}{2}\omega\right)^p - \frac{\alpha+\beta}{\alpha-\beta}(1-\omega) = 0$$

im Intervall $(1-\frac{\alpha-\beta}{\alpha+\beta}, 1)$*. Für den optimalen Konvergenzfaktor ergibt sich*

$$\rho(\mathcal{L}_{\omega_0}) = \frac{\alpha+\beta}{\alpha-\beta}(1-\omega_0) = \left(\frac{\alpha+\beta}{2}\omega_0\right)^p.$$

3.4 SOR für periodische Markov–Ketten; die *Extended Convergence*

SOR für p–periodische Markov–Ketten

Für die Jacobi–Matrix, die aus einer zeitkontinuierlichen p–periodischen Markov–Kette resultiert, sowie für die Matrix, die aus einer zeitdiskreten p–periodischen Markov–Kette resultiert, gilt nach Satz 2.7, dass die Eigenwerte immer in p–Tupeln auftreten. Dabei hat nach Abschnitt 2.3 ein p–Tupel sicher Betrag 1, die anderen sind vom Betrage kleiner als 1.

Wendet man das SOR–Verfahren an, so sind nach Satz 2.8 jedem Eigenwert–p–Tupel von J gerade p Eigenwerte von $\mathcal{L}_\omega$ zugeordnet.

Dass einer der 1 zugeordneten Eigenwerte selber wieder 1 sein muss, folgt sofort aus (12). Damit SOR im klassischen Sinn konvergiert, müssen alle Eigenwerte betragsmäßig kleiner als 1 sein.

Der Eigenwert 1 von $\mathcal{L}_\omega$ stört die Konvergenz des SOR–Verfahrens nach Korollar 2.14 nicht, weil

- $x = \mathcal{L}_\omega x$ nach Kapitel 2.3 lösbar ist, weil $x^T Q = 0^T$ bzw. $x^T P = x^T$ nichttrivial lösbar ist,
- $0 \in \text{bild}((I - \mathcal{L}_\omega)^q, q = \text{ind}(I - \mathcal{L}_\omega)$, nach Kapitel 2.3 gilt dies für den stationären Vektor,
- $\rho(\mathcal{L}_\omega) = 1$ nach Voraussetzung,
- $\lambda = 1$ einziger Eigenwert von $\mathcal{L}_\omega$ mit Betrag 1 ist.

Man kann also von Konvergenz des SOR–Verfahrens ausgehen, wenn alle weiteren Eigenwerte betragsmäßig kleiner als 1 sind. Die Ergebnisse von Kapitel 3.3 behalten ihre Gültigkeit, wenn man den Spektralradius durch den Subspektralradius ersetzt.

Es ergeben sich zwei Fragen:

1. Wann sind die Eigenwerte von $\mathcal{L}_\omega$, die zu den Eigenwerten $|\mu| < 1$ von J gehören betragsmäßig kleiner als 1 ?
2. Ist gesichert, dass die weiteren $p - 1$ Eigenwerte, die zu 1 gehören, betragsmäßig kleiner als 1 sind ?

zu 1) Aufgrund des stetigen Abhängigkeit von $\mu^p \in \mathbb{R}$ in (12) sind die Eigenwerte, die zu $|\mu| < 1$ gehören, auf jeden Fall betragsmäßig kleiner als der größte zu 1 gehörende Eigenwert. Dieser ist bei Konvergenz des Verfahrens gerade 1, das heißt, dass die zu $|\mu| < 1$ gehörenden Eigenwerte die Konvergenz an sich nicht stören, sehr wohl aber Einfluss auf die Konvergenzgeschwindigkeit haben.

zu 2) Für das Intervall $(0, \frac{p}{p-1})$ gilt die Abschätzung, die im folgenden Kapitel 3.5 gezeigt wird. Auch diese $p-1$ Eigenwerte stören die Konvergenz nicht.

Bevor jetzt die Frage nach den anderen zu 1 gehörenden Eigenwerten beantwortet wird, soll zunächst die *extended convergence* oder *verallgemeinerte Konvergenz*, die von Kontovasilis, Plemmons und Stewart in [34] eingeführt wurde, vorgestellt werden. Damit kann insbesondere Frage 2) auch gleich in größerem Rahmen behandelt werden.

Die *Extended Convergence* von Kontovasilis, Stewart und Plemmons

In einer fast 70–seitigen Arbeit in LAA 154–156 überraschten Kontovasilis, Plemmons und Stewart 1991 (vgl. [34]) dadurch, dass sie für das SOR–Verfahren zur Lösung von (21) für p–periodische Markov–Ketten Relaxationsparameter außerhalb des durch das Lemma von Kahan (Lemma 2.1) zwingend vorgeschriebenen Intervalls $(0, 2)$ verwenden, ja sogar negative Relaxationsparameter.

Was dabei passiert, wird deutlicher, wenn das Verfahren

$$x_{k+1} = \mathcal{L}_\omega x_k$$

nicht aus der Sicht des SOR–Verfahrens sondern als Vektoriteration betrachtet wird.

Im Fall des klassischen SOR–Verfahrens (das heißt wenn der Relaxationsparameter im Intervall $(0, \frac{p}{p-1})$ liegt) ist — zumindest wenn das Spektrum von J^p reell ist — 1 einfacher dominanter Eigenwert von $\mathcal{L}_\omega$. Die Vektoriteration konvergiert dann gegen den dem Eigenwert 1 zugeordneten Eigenvektor, also gegen den stationären Wahrscheinlichkeitsvektor.

Liegt der Relaxationsparameter außerhalb dieses Intervalls, konvergiert die Vektoriteration gegen den Eigenvektor des dann dominanten Eigenwertes. Dieser kann

dabei dem Eigenwert–p–Tupel 1 von J^p zugeordnet sein oder aber einem anderen, welcher Fall nur bei komplexem Spektrum von J^p auftreten kann.

Ist der dominante Eigenwert von $\mathcal{L}_\omega$ dem Eigenwert–p–Tupel 1 zugeordnet, konvergiert das Verfahren nicht gegen den stationären Wahrscheinlichkeitsverteilungsvektor (dieser gehört ja zum Eigenwert 1 von $\mathcal{L}_\omega$). Aus dem Vektor, gegen den das Verfahren konvergiert, lässt sich aber der stationäre Vektor rekonstruieren.

Es gilt nämlich ([34, Theorem 3.2]) für p–zyklische Generatoren Q

Satz 3.8 *Ist $\psi = (\psi_1, \psi_2, \ldots, \psi_p)^T$ Eigenvektor von J zum Eigenwert μ, λ^p ein μ zugeordneter Eigenwert von $\mathcal{L}_\omega$, so ist der dazugehörige Eigenvektor*

$$\psi' = (\psi_1, \lambda\psi_2, \ldots, \lambda^{p-1}\psi_p)^T.$$

Dabei seien die Vektoren ψ und ψ' genauso partitioniert wie die Jacobi–Matrix J.

Dieser Satz erlaubt das Korollar

Korollar 3.9 *Ist der (betrags–)maximale Eigenwert von $\mathcal{L}_\omega$ dem Eigenwert 1 von J zugeordnet, dann lässt sich der stationäre Wahrscheinlichkeitsvektor aus dem Vektor rekonstruieren, gegen den das Verfahren konvergiert.*

Ist der dominante Eigenwert von $\mathcal{L}_\omega$ dagegen einem anderen Eigenwert–p–Tupel als 1 zugeordnet, konvergiert die Vektoriteration zwar gegen den zugehörigen Eigenvektor, aus diesem aber lässt sich der stationäre Wahrscheinlichkeitsvektor nicht mehr rekonstruieren. Das Verfahren ist zur Bestimmung des stationären Wahrscheinlichkeitsverteilungsvektors in diesem Fall untauglich.

Die Vektoriteration würde eigentlich in jedem Iterationsschritt einen Normierungsschritt einschließen (der von dem Rekonstruktionsschritt nach Satz 3.8 bzw. Korollar 3.9 vollkommen unterschiedlich ist), in der Praxis zeigt es sich jedoch, dass das Verfahren auch ohne die wiederholte Normierung und mit einem einzigen Rekonstruktionsschritt am Ende sehr gut (und stabil) konvergiert (und das, selbst wenn $\|\mathcal{L}_\omega x_i\|_2 = O(10^{30})$ ist — das Problem ist eher zu entscheiden, wann man durch den Rekonstruktionsschritt den stationären Wahrscheinlichkeitsverteilungsvektor hinreichend genau bestimmt hat).

Kontovasilis, Plemmons und Stewart nennen dieses Verfahren *extended convergence*. Im Weiteren wird dafür auch der Begriff *verallgemeinerte (SOR–)Konvergenz* benutzt.

Bei der Beschreibung der *verallgemeinerten SOR–Konvergenz* sind drei Fälle zu unterscheiden:

1. Ist 1 einfacher, dominanter Eigenwert, dann entspricht das Verfahren gerade dem klassischen SOR–Verfahren, d.h. es konvergiert gegen den gesuchten stationären Wahrscheinlichkeitsvektor.

2. Ist der dominante Eigenwert größer als 1, aber andererseits dem Eigenwert–p–Tupel 1 von J^p zugeordnet, konvergiert das verallgemeinerte SOR–Verfahren gegen einen Vektor (Konvergenz allerdings nicht im klassischen Sinn, sondern Konvergenz nach einem abschließenden Rekonstruktionsschritt), aus dem man den stationären Wahrscheinlichkeitsvektor rekonstruieren kann und auch muss.

3. Ist der dominante Eigenwert größer als 1, aber nicht dem Eigenwert–p–Tupel 1 von J^p zugeordnet, so konvergiert das Verfahren in keinem Fall gegen den stationären Wahrscheinlichkeitsverteilungsvektor. Dieser Fall tritt ohnehin nur bei komplexem Spektrum von J^p auf.

In [34, Kapitel 7] wird ein numerischer Algorithmus angegeben, mit dessen Hilfe im zweiten der drei obigen Fälle der stationäre Wahrscheinlichkeitsvektor ohne Kenntnis des betragsmaximalen Eigenwertes von $\mathcal{L}_\omega$ rekonstruiert werden kann.

Das Verfahren besitzt nach [34, Kapitel 5] für $\sigma(J^p) \in \mathbb{R}_0^+$ sicher dann optimalen Konvergenzfaktor, wenn ihm ein gewöhnliches Hypozykloid ($h = r$, vgl. Kapitel 3.2) zugeordnet ist, das die reelle Achse gerade im Punkt $\tilde{\rho}(J)$ schneidet. Ist $\omega > 1$, dann muss für die Parameter gelten (dies ist eine direkte Folgerung aus der Forderung $h = r$), dass

$$\frac{1}{\eta^p} = (p-1)(\omega - 1) \tag{37}$$

ist. Weil außerdem das Hypozykloid dann die positive reelle Achse im Punkt

$$x = \frac{p}{(p-1)\omega\eta}$$

schneidet (wegen (37) an der Stelle $\vartheta = 0$) und $x = \tilde{\rho}(J)$ gelten muss, folgt unmittelbar aus dieser Argumentation die Bestimmungsgleichung (13).

Durch analytische Argumente (wie die Descartessche Zeichenregel und das Verhalten des Vorzeichens der linken Seite) erkennt man, dass (13) zwei positive Lösungen ω_0 und ω_+ besitzt. Ist p ungerade, folgt aus den dann gültigen Beziehungen ebenfalls (13) und die Existenz einer negativen Lösung ω_- dieser Gleichung.

Durch analytische Umformungen und Einsetzen erhält man Intervalle, in denen diese Lösungen $\omega_0, \omega_+, \omega_-$ liegen müssen.

Es bezeichne

$$\alpha(\omega) := \max\{|\lambda| : \lambda^p - \omega\lambda^{p-1} - (1-\omega) = 0\}. \tag{38}$$

$\alpha^p(\omega)$ ist also der größte 1 zugeordnete Eigenwert von $\mathcal{L}_\omega$.

Damit lässt sich zeigen, dass die Konvergenzfaktoren für $\omega_0, \omega_+, \omega_-$ gleich sind. Es wird allgemeiner untersucht, mit welchem η man bei vorgegebenem ω den optimalen Konvergenzfaktor (denjenigen, der auch zu ω_0 gehört) erhält und ob dabei die Eigenwerte von J im Innengebiet des zugehörigen Hypozykloids liegen oder Eigenwerte auf der Randkurve liegen. Dadurch wird gezeigt, dass der Konvergenzfaktor schlechter ist als bei den optimalen Werten der Relaxationsparameter. Die Untersuchung wird teilintervallweise durchgeführt.

Insgesamt erhält man

Satz 3.10 *Wenn das Block–SOR–Verfahren zur Lösung von $Q^T x = 0$ eingesetzt wird, die assoziierte Block–Jacobi–Matrix J die Form (9) besitzt, die Eigenwerte von J^p nichtnegativ sind und $\tilde{\rho}(J) = \max\{|\mu| : \mu \in \sigma(J), |\mu| < 1\}$ bezeichnet, dann gilt:*

1. *Ist $\tilde{\rho}(J) > 0$ und ω_0 und ω_+ die einzigen positiven Nullstellen von (13) in*

$$\left(1, \frac{p}{p-1}\right) \quad \text{bzw.} \quad \left(\frac{p}{p-1}\tilde{\rho}(J)^{-p/(p-1)}, \frac{p^{p/(p-1)}}{p-1}\tilde{\rho}(J)^{-p/(p-1)}\right),$$

   dann erreichen beide Werte den optimalen Konvergenzfaktor von $(p-1)(\omega_0 - 1)$. Ist p ungerade, gilt dies zusätzlich für die eindeutige negative Lösung von (13) in

$$\left(-\frac{p^2}{p-1}\tilde{\rho}(J)^{-p/(p-1)}, -\frac{p}{p-1}\tilde{\rho}(J)^{-p/(p-1)}\right).$$

   Im klassichen Sinn konvergiert das Verfahren dabei nur für ω_0.

2. *Ist $\tilde{\rho}(J) = 0$, erhält man für $\omega = 1$ als einzigen Wert den optimalen Konvergenzfaktor 0 und das SOR–Verfahren konvergiert im klassischen Sinn.*

Bemerkung 3.2 *Eine bessere Konvergenz als mit dem klassischen optimalen Relaxationsparameter erreicht man nicht. Verwendet man dennoch das erweiterte SOR–Verfahren, dann deshalb, weil dieses unempfindlicher auf Störungen des Relaxationsparameters reagiert.*

Im reellen Fall ist klar, dass der dominante Eigenwert von $\mathcal{L}_\omega$ dem Eigenwert–p–Tupel 1 zugeordnet ist; aufgrund der Herleitung muss aber für die Konvergenz sichergestellt sein, dass der subdominante Eigenwert von $\mathcal{L}_\omega$ einem anderen Eigenwert–p–Tupel zugeordnet ist. Kontovasilis, Plemmons und Stewart treffen in [34] dazu keine Aussagen. Deswegen wird im nächsten Abschnitt das Verhalten der anderen zum Eigenwert–p–Tupel 1 gehörenden Eigenwerte untersucht.

3.5 Lage der anderen 1 zugeordneten Eigenwerte der SOR–Matrix

Wie bereits erwähnt, ist es von besonderer Bedeutung, wo in der komplexen Ebene diejenigen Eigenwerte der SOR–Matrix liegen, die dem Eigenwert–p–Tupel 1 von J^p zugeordnet sind. Einer dieser Eigenwerte ist sicher 1, für $\omega < 0$ und $\omega > \frac{p}{p-1}$ wird — zumindest wenn die Eigenwerte von J^p reell sind — ein weiterer zum dominanten Eigenwert. Es ist aber für die Konvergenz wichtig sicherzustellen, dass die anderen $p-2$ Eigenwerte keinen Einfluss auf den subdominanten Eigenwert von $\mathcal{L}_\omega$ nehmen, dass dieser also wirklich einem Eigenwert $|\mu| < 1$ von J zugeordnet ist.

Für $\omega > \frac{p}{p-1}$ zeigt sich, dass die weiteren 1 zugeordneten Eigenwerte alle im Einheitskreis enthalten sind, für $\omega < 0$ lässt sich zumindest für kleine p zeigen, dass sie kleiner sind als der größte zu μ gehörende Eigenwert.

Die Beweisführung ist unabhängig davon, ob die Eigenwerte reell oder komplex sind. Dazu werden einige Hilfsmittel bereitgestellt.

Die Schur–Transformierte und der Satz von Schur–Cohn

Die nötigen Hilfsmittel zum Beweis des Satzes finden sich in [27, Kapitel 6.8]:

Definition 3.1 *Für ein Polynom* $p(z) = \sum_{k=0}^{n} a_k z^k$ *vom Grad* $n \geq 1$ *ist das Polynom vom Grad* $n-1$

$$\begin{aligned} Tp(z) &:= \overline{a_0} p(z) - a_n p^*(z) && (39) \\ &= \sum_{k=0}^{n-1} (\overline{a_0} a_k - a_n \overline{a_{n-k}}) z^k && (40) \end{aligned}$$

die Schurtransformierte *von* p.

Es gilt für die iterierten Schurtransformierten

$$T^k p := T(T^{k-1} p).$$

Setzt man

$$\gamma_k := T^k p(0), \qquad k = 1, \ldots, n, \tag{41}$$

so gelten folgende Aussagen

Satz 3.11 *Ist $p \neq 0$ ein Polynom vom Grad n. Dann liegen alle Nullstellen von p außerhalb des geschlossenen Einheitskreises genau dann, wenn*

$$\gamma_k > 0, \quad k = 1, \ldots, n.$$

Satz 3.12 *Ist p ein Polynom vom Grad n, γ_k wie in (41) mit $\gamma_k \neq 0$, $k = 1, 2, \ldots, n$. Bezeichnet man diejenigen Indizes, für die $\gamma_k < 0$ ist mit k_j, $j = 1, \ldots, m$ mit $k_1 < k_2 < \ldots < k_m$, so ist die Zahl $h(p)$ der Nullstellen von p, die innerhalb des Einheitskreises liegen, gegeben durch*

$$h(p) = \sum_{j=1}^{m} (-1)^{j-1}(n + 1 - k_j). \tag{42}$$

Die Berechnung der Zahlen γ_k heißt dabei auch *Schur–Cohn–Algorithmus.*

Anwendung auf die Gleichung der p–Schritt–Relaxation

Gegenstand dieses Abschnitts ist die Formel von Varga in der Form (14), weil diese einfacher zu handhaben ist (vgl. auch Bemerkung 2.6).

Satz 3.13 *Für die um den Eigenwert 1 reduzierte Gleichung (14)*

$$\frac{\overbrace{\tau^p - \omega\tau^{p-1} - (1-\omega)}^{=:g_\omega(\tau)}}{\tau - 1} = \tau^{p-1} + (1-\omega)\sum_{k=0}^{p-2}\tau^k =: \tilde{g}_\omega(\tau)$$

lautet die $p-1$–te Schur–Transformierte

$$\begin{aligned} T^{p-1}(\tilde{g}_\omega(\tau)) = \\ \omega^{(2^{p-2})}(\omega-2)^{(2^{p-4})}(2\omega-3)^{(2^{p-5})}(3\omega-4)^{(2^{p-6})} \\ \ldots[(p-3)\omega-(p-2)][(p-1)\omega-p]. \end{aligned} \tag{43}$$

Der Koeffizient von τ in $T^{p-2}\tilde{g}_\omega(\tau)$ ist

$$\begin{aligned} \omega^{(2^{p-3})}(\omega-1)(\omega-2)^{(2^{p-5})}(2\omega-3)^{(2^{p-6})}(3\omega-4)^{(2^{p-7})} \\ \ldots[(p-4)\omega-(p-3)]. \end{aligned} \tag{44}$$

Dabei sind jeweils die Glieder des Produktes zu berücksichtigen für die jeweils der Exponent von 2 *größer oder gleich* 0 *ist.*

Beweis:
Der Beweis erfolgt mittels vollständiger Induktion über p.

Zum Induktionsanfang betrachtet man die Fälle $p = 2$, $p = 3$ (dieser spielt in gewisser Weise eine Sonderrolle) und $p = 4$:

$p = 2$: Spaltet man $\tau - 1$ von $\tau^2 - \omega\tau - (1 - \omega)$ ab, ergibt sich

$$\tau + (1 - \omega) = \tilde{g}_\omega(\tau).$$

und dann

$$T\tilde{g}_\omega(\tau) = (1 - \omega)^2 - 1 = \omega(\omega - 2).$$

Dies ist die Behauptung des Satzes.

$p = 3$: Aus

$$\tilde{g}_\omega(\tau) = \tau^2 + (1 - \omega)\tau + (1 - \omega)$$

folgt

$$\begin{aligned} T\tilde{g}_\omega(\tau) &= [(1 - \omega)^2 - 1] + [(1 - \omega)^2 - (1 - \omega)]\tau \\ &= \omega(\omega - 1)\tau + \omega(\omega - 2) \\ T^2\tilde{g}_\omega(\tau) &= \omega^2(\omega - 2)^2 - \omega^2(\omega - 1)^2 \\ &= -\omega^2(2\omega - 3). \end{aligned}$$

Auch dies ist die Behauptung des Satzes, unter Berücksichtigung, dass $(p - 3)\omega - (p - 2) = 0\omega - 1 = -1$ ist.

$p = 4$: Aus

$$\tilde{g}_\omega(\tau) = \tau^3 + (1 - \omega)\tau^2 + (1 - \omega)\tau + (1 - \omega)$$

folgt

$$\begin{aligned} T\tilde{g}_\omega(\tau) &= [(1 - \omega)^2 - (1 - \omega)]\tau^2 + [(1 - \omega)^2 - (1 - \omega)]\tau + [(1 - \omega)^2 - 1] \\ &= \omega(\omega - 1)\tau^2 + \omega(\omega - 1)\tau + \omega(\omega - 2) \\ T^2\tilde{g}_\omega(\tau) &= [\omega^2(\omega - 1)(\omega - 2) - \omega^2(\omega - 1)^2]\tau + [\omega^2(\omega - 2)^2 - \omega^2(\omega - 1)^2] \\ &= -\omega^2(\omega - 1)\tau - \omega^2(2\omega - 3) \\ T^3\tilde{g}_\omega(\tau) &= \omega^4(2\omega - 3)^2 - \omega^4(\omega - 1)^2 \\ &= \omega^4(3\omega - 4)(\omega - 2). \end{aligned}$$

In allen drei Fällen sieht man, dass auch der Koeffizient von τ in der linearen Gleichung stimmt, wobei für $p = 2$ das dann leere Produkt 1 gesetzt wird.

Sei die Behauptung wahr für p; dann ist zu zeigen, dass sie auch für $p+1$ stimmt.

Es gilt

$$\begin{aligned} T^p\tilde{g}_\omega(\tau) &= \left(\omega^{(2^{p-2})}\right)^2 \left((\omega-2)^{(2^{p-4})}\right)^2 \ldots [(p-3)\omega-(p-2)]^2[(p-1)\omega-p]^2 \\ &\quad - \left(\omega^{(2^{p-2})}\right)^2 \left((\omega-2)^{(2^{p-4})}\right)^2 \ldots [(p-3)\omega-(p-2)]^2(\omega-1)^2 \\ &= \left(\omega^{(2^{p-2})}\right)^2 \left((\omega-2)^{(2^{p-4})}\right)^2 \ldots [(p-3)\omega-(p-2)]^2 \\ &\quad \left[\underbrace{[(p-1)\omega-p]^2-(\omega-1)^2}_{=:\chi}\right] \end{aligned}$$

mit

$$\begin{aligned} \chi &= [(p-1)\omega-p]^2-(\omega-1)^2 \\ &= (p-1)^2\omega^2-2p(p-1)\omega+p^2-\omega^2+2\omega-1 \\ &= [(p-2)\omega-(p-1)](p\omega-(p+1)). \end{aligned}$$

Zusammengesetzt ergibt sich

$$\begin{aligned} T^p\tilde{g}_\omega(\tau) &= \left(\omega^{(2^{p-2})}\right)^2 \left((\omega-2)^{(2^{p-4})}\right)^2 \ldots [(p-3)\omega-(p-2)]^2 \\ &\quad \left[[(p-1)\omega-p]^2-(\omega-1)^2\right] \\ &= \omega^{(2^{p-1})}(\omega-2)^{(2^{p-3})} \ldots [(p-3)\omega-(p-2)]^2 \\ &\quad [(p-2)\omega-(p-1)][p\omega-(p+1)]. \end{aligned}$$

Die Koeffizienten von τ^k sind in T^l, $1 \le l \le p-1$ für $k > 0$ alle dieselben aufgrund der besonderen Gestalt des Polynoms $\tilde{g}$ und der Vorschrift (40) der Schurtransformation. Deshalb gilt für den neuen Koeffizienten von τ in $T^{p-1}\tilde{g}_\omega(\tau)$

$$\begin{aligned} v &= \omega^{(2^{p-3})}(\omega-2)^{(2^{p-5})} \ldots [(p-2)\omega-(p-1)] \\ &\quad \omega^{(2^{p-3})}(\omega-2)^{(2^{p-5})} \ldots [(p-4)\omega-(p-3)](\omega-1) \\ &\quad - \left(\omega^{(2^{p-3})}\right)^2 \left((\omega-2)^{(2^{p-5})}\right)^2 \ldots [(p-4)\omega-(p-3)]^2(\omega-1)^2 \end{aligned}$$

$$
\begin{aligned}
&= \left(\omega^{(2^{p-3})}\right)^2 \left((\omega-2)^{(2^{p-5})}\right)^2 \ldots [(p-4)\omega-(p-3)]^2(\omega-1) \\
&\quad [[(p-2)\omega-(p-1)]-(\omega-1)] \\
&= \left(\omega^{(2^{p-3})}\right)^2 \left((\omega-2)^{(2^{p-5})}\right)^2 \ldots [(p-4)\omega-(p-3)]^2(\omega-1) \\
&\quad [(p-3)\omega-(p-2)], \\
&= \omega^{(2^{p-2})}(\omega-2)^{(2^{p-4})} \ldots [(p-4)\omega-(p-3)]^2[(p-3)\omega-(p-2)](\omega-1).
\end{aligned}
$$

was genau der Behauptung entspricht.

Damit ist der Induktionsschluss vollzogen und der Satz insgesamt gezeigt. □

Bemerkung 3.3 *Die ersten Schur–Transformierten sind für $\tilde{g}_\omega^{(k)}$, wenn k den Grad von $g(\tau)=(\tau-1)\tilde{g}(\tau)$ bezeichnet,*

$$
\begin{aligned}
T\tilde{g}_\omega^{(2)}(\tau)) &= \omega(\omega-2) \\
T^2\tilde{g}_\omega^{(3)}(\tau)) &= -\omega^2(2\omega-3) \\
T^3\tilde{g}_\omega^{(4)}(\tau)) &= \omega^4(\omega-2)(3\omega-4) \\
T^4\tilde{g}_\omega^{(5)}(\tau)) &= \omega^8(\omega-2)^2(2\omega-3)(4\omega-5) \\
T^5\tilde{g}_\omega^{(6)}(\tau)) &= \omega^{16}(\omega-2)^4(2\omega-3)^2(3\omega-4)(5\omega-6)
\end{aligned}
$$

Satz 3.13 in Verbindung mit Satz 3.12 ermöglicht Aussagen über die Lage der Nullstellen von g und damit über die Lage der Eigenwerte von $\mathcal{L}_\omega$.

Korollar 3.14 *Für $\omega > 2$ liegen alle Lösungen von $g_\omega(\tau)=0$ bis auf zwei innerhalb des Einheitskreises, eine liegt außerhalb, eine ist exakt 1.*

Insbesondere sind damit alle zu 1 gehörenden Eigenwerte von $\mathcal{L}_\omega$ bis auf zwei im Einheitskreis, einer ist exakt 1, einer (falls das Spektrum von J^p reell ist, der dominante) liegt außerhalb des Einheitskreises.

Beweis:
Für $\omega > 2$ ist $T^2(\tilde{g}_\omega(\tau)) < 0$ und für $3 \le k \le p-1$ und $k=1$ ist nach Satz 3.13 offensichtlich $T^k(\tilde{g}_\omega(\tau)) > 0$.

Mit Satz 3.12 (vgl. auch [27, Theorem 6.8c]) ergibt sich die Zahl der Nullstellen von $\tilde{g}$ im Einheitskreis als

$$h(\tilde{g}) = (-1)^0(n+1-k_0) = (p-1)+1-2 = p-2,$$

wobei $k_1 = 2$ in der Notation von Henrici.

Das bedeutet, dass $p-2$ Nullstellen von $\tilde{g}$ im Einheitskreis liegen, eine außerhalb. Nimmt man die bereits abgespaltene Nullstelle $\tau = 1$ hinzu (geht also von $\tilde{g}$ zu g über), folgt die Behauptung. □

Korollar 3.15 *Für $\omega < 0$ liegen alle Nullstellen von $\tilde{g}(\tau)$ außerhalb des Einheitskreises.*

Beweis:
In (43) sind nur zwei Faktoren mit einem Exponenten versehen, der keine Zweier–Potenz ist, nämlich

$$[(p-3)\omega - (p-2)] \quad \text{und} \quad [(p-1)\omega - p].$$

Diese beiden Faktoren werden für $\omega < 0$ negativ, während alle anderen wegen der Exponenten positiv bleiben.

Das bedeutet, dass für das reduzierte Polynom $\tilde{g}(\tau)$ gilt

$$\gamma_k > 0, \quad k = 1, \ldots, p-1,$$

und nach Satz 3.11, dass alle Nullstellen außerhalb des Einheitskreises liegen müssen. □

Für negative Relaxationsparameter ergibt sich

Satz 3.16 *Für $p = 3$ und $\omega < 0$ ist auf jeden Fall für $\omega < -1 - \sqrt{5}$ der größte zu μ mit $|\mu| < 1$ gehörende Eigenwert größer als 2 zu 1 gehörende Eigenwerte, d.h. der subdominante Eigenwert von $\mathcal{L}_\omega$ gehört zu μ.*

Beweis:
Für $\mu = 0$ ist $1 - \omega$ p–facher Eigenwert von $\mathcal{L}_\omega$.

Es genügt zu zeigen, dass 2 zu 1 gehörende Eigenwerte kleiner sind als $1 - \omega$ bzw. dass eine Lösung von

$$g(\tau) = (1-\omega)^{2/3}\tau^2 + (1-\omega)^{4/3}\tau + (1-\omega)$$

im Einheitskreis liegt (dazu vgl. [27, S. 491ff]), wenn wieder das Polynom

$$\frac{\tau^p - \tau^{p-1} - (1-\omega)}{\tau - 1}$$

betrachtet wird.

Mittels der Substitution $\xi := (1-\omega)^{1/3}$ ist

$$g(\tau) = \xi^2\tau^2 + \xi^4\tau + \xi^3,$$

und die ersten beiden Schur–Transformierten lauten

$$\begin{aligned} Tg(\tau) &= (\xi^7 - \xi^6)\tau + (\xi^6 - \xi^4) \\ T^2g(\tau) &= -\xi^{14} + 2\xi^{13} - 2\xi^{10} + \xi^8. \end{aligned}$$

Wegen $\omega < 0$ folgt

$$\begin{aligned} 1-\omega &> 1 \\ (1-\omega)^{2/3} &> 1 \\ \xi^2 &> 1 \\ \xi^6 > \xi^4, & \end{aligned}$$

also

$$Tg(0) > 0.$$

Ferner ist

$$\lim_{\xi\to\infty} T^2g(\tau)(\xi) = -\infty$$

und $T^2g(\tau)(\xi)$ besitzt nach der Vorzeichenregel von Descartes genau eine oder drei positive Nullstellen. 1 ist, wie man durch Nachrechnen sieht, doppelte Nullstelle, die dritte ist $\frac{1}{2}(1+\sqrt{5})$:

$$\begin{aligned} T^2g\left(\frac{1}{2}(1+\sqrt{5})\right) &= -\left(\frac{1}{2}(1+\sqrt{5})\right)^{14} + 2\left(\frac{1}{2}(1+\sqrt{5})\right)^{13} \\ &\quad -2\left(\frac{1}{2}(1+\sqrt{5})\right)^{10} + \left(\frac{1}{2}(1+\sqrt{5})\right)^{8} \\ &= -\frac{1}{16384}(1+\sqrt{5})^{14} + \frac{1}{4096}(1+\sqrt{5})^{13} \\ &\quad -\frac{1}{512}(1+\sqrt{5})^{10} + \frac{1}{256}(1+\sqrt{5})^{8} \\ &= 0. \end{aligned}$$

Mit $\xi = (1-\omega)^{1/3}$ folgt

$$\begin{aligned}\left(\frac{1}{2}+\frac{1}{2}\sqrt{5}\right)^5 &= 1-\omega \\ \frac{1}{8}+\frac{3}{8}\sqrt{5}+\frac{15}{8}+\frac{5}{8}\sqrt{5} &= 1-\omega \\ \omega &= -1-\sqrt{5},\end{aligned}$$

was der Behauptung entspricht.

□

Figur 12 zeigt die Beträge der zu 1 gehörenden Eigenwerte (für $\omega > -3$ ein konjugiert komplexes Paar) sowie die Gerade $1-\omega$ (gestrichelte Linie). Ferner ist der Wert $-1-\sqrt{5}$ bezeichnet.

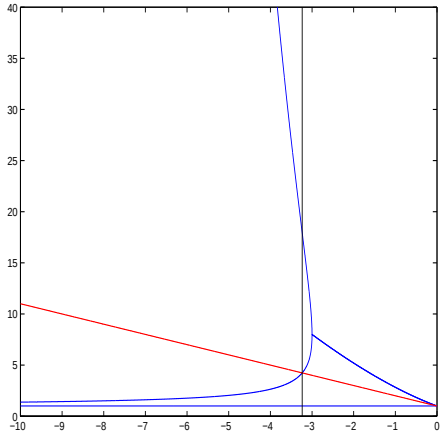

Abbildung 12: Eigenwerte zu 1 und $1-\omega$ in Abhängigkeit von ω

Für $\mu > 0$ rückt der entsprechende Wert näher an die -3 heran.

Bemerkung 3.4 *Für $p = 4$ lässt sich analog zeigen, dass $1-\omega$ größer ist als alle zu 1 gehörenden Eigenwerte bis auf den größten.*

Für $p > 4$ werden die Ausdrücke sehr unübersichtlich, so ist für $p = 5$ $T^4 g(\tau)(\xi)$

mit analogem $\xi = (1-\omega)^{1/5}$ *ein Polynom* 104*–ten Grades, bei dem* 37 *der Koeffizienten von Null verschieden sind.*

Bemerkung 3.5 *Mit dem von Locher in [36] bzw. in [37] vorgeschlagenen Algorithmus ergibt sich ebenfalls für die Fälle* $p = 3$, $p = 5$ *problemlos, für die Fälle* $p = 7$ *und* $p = 9$ *mit deutlich erhöhtem Rechenaufwand, dass, sofern* $\omega < 0$ *hinreichend weit von* 0 *entfernt ist, alle Nullstellen von* $\tau^p - \omega\tau^{p-1} - (1-\omega)$ *bis auf eine in einem Kreis um* 0 *mit Radius* $\sqrt[p]{1-\omega}$ *liegen.*

Auch damit ist klar, dass in dem Bereich, in dem dieses gilt, der subdominante Eigenwert von $\mathcal{L}_\omega$ *zu* $\mu \neq 1$ *gehört, da zu* $\mu = 0$ *schon der Eigenwert* $1-\omega$ *gehört und alle zu* 1 *gehörenden Eigenwerte bis auf einen in einem Kreis um* 0 *mit Radius* $1-\omega$ *liegen.*

Eine allgemeine Formel bzw. einen allgemeinen Beweis legt aber auch dieser Algorithmus, der auf Tschebyscheff–Polynomen und Sturmschen Ketten basiert, nicht nahe.

3.6 Lage bei unbekanntem Spektrum der Jacobi–Matrix

Voraussetzung für alle Untersuchungen in [65], [11], [18] und bei den Betrachtungen über das verallgemeinerte SOR–Verfahren (nach [34]) sowie auch in den Kapiteln 3.2 und 3.3 ist, dass Informationen über die Lage des Spektrums der Jacobi–Matrix bekannt sind.

Bei fehlenden Informationen stellen sich zwei Fragen:

1. Mit welchem Relaxationsparameter rechnet man ?
2. Um wieviel wird der Konvergenzfaktor dadurch schlechter ?

Die Antwort auf die erste Frage ist, dass man das Gauß–Seidel–Verfahren anwenden sollte, sofern man nicht ohnehin auf andere Verfahren wie z.B. Krylov–Verfahren (vgl. auch [28]) übergeht. Es ist ohne weitere Kenntnisse über das Spektrum nicht möglich, etwas darüber zu sagen, ob Über– oder Unterrelaxation zu besseren Konvergenzfaktoren führt, geschweige denn, dass man etwas über den optimalen Parameter aussagen kann.

Dass das Gauß–Seidel–Verfahren optimal ist, wenn die Eigenwerte in einem Kreis um 0 enthalten sind (und nichts anderes weiß man letztlich), folgt nach [46] und ist im letzten Abschnitt des Kapitels 4.5 ausgeführt.

Aussagen darüber, um wieviel das Gauß–Seidel–Verfahren asymptotisch schlechter ist als das SOR–Verfahren mit optimalem Relaxationsparameter, lassen sich treffen und zwar unterschieden in die Fälle $\sigma(J) \in \mathbb{R}^+$ (bzw. $\mathbb{R}^-$), $\sigma(J) \in \mathbb{R}$ und — im späteren Kapitel 4.5 — $\sigma(J) \in \mathbb{C}$.

Fall $\sigma(J^p) \in \mathbb{R}_0^+$ oder $\sigma(J^p) \in \mathbb{R}_0^-$:

Satz 3.17 *Sind die Eigenwerte von J^p positiv, ist das Gauß–Seidel–Verfahren höchstens um den Faktor*

$$\left(\frac{p}{p-1}\right)^p \tag{45}$$

schlechter als das SOR–Verfahren mit optimalem Parameter ω^*.

Sind die Eigenwerte von J^p nichtpositiv, so beträgt dieser Faktor

$$\left(\frac{p}{p-2}\right)^p. \tag{46}$$

Beweis:
(i) Eigenwerte μ^p von J^p nichtnegativ

Nach Varga ([64]) besitzt das SOR–Verfahren mit optimalem ω den Konvergenzfaktor

$$\rho^* = (p-1)(\omega^* - 1), \tag{47}$$

wobei ω^* Lösung ist von

$$\mu^p \omega^p = p^p (p-1)^{1-p} (\omega - 1). \tag{48}$$

Dann gilt unter Beachtung von $\omega^* > 1$

$$\begin{aligned} \left|\frac{\rho_1}{\rho^*}\right| &= \left|\frac{\mu^p}{(p-1)(\omega^*-1)}\right| \\ &\overset{(4)}{=} \frac{p^p(\omega^*-1)}{(p-1)^p(\omega^*)^p(\omega^*-1)} = \left(\frac{p}{p-1}\right)^p \frac{1}{(\omega^*)^p} < \left(\frac{p}{p-1}\right)^p. \end{aligned} \tag{49}$$

(ii) Eigenwerte μ^p von J^p nichtpositiv

Nach ([65]), ([11]) besitzt das SOR–Verfahren mit optimalem ω den Konvergenzfaktor

$$\rho^* = (p-1)(1-\omega^*), \tag{50}$$

wobei ω^* Lösung ist von

$$-\mu^p \omega^p = p^p (p-1)^{1-p} (1-\omega). \tag{51}$$

Dann gilt unter Beachtung von $\omega^* > \frac{p-2}{p-1}$ (vgl. [65])

$$\begin{aligned} \left|\frac{\rho_1}{\rho^*}\right| &= \left|\frac{\mu^p}{(p-1)(\omega^*-1)}\right| \\ &\overset{(4)}{=} \frac{p^p(1-\omega^*)}{(p-1)^p(\omega^*)^p(1-\omega^*)} = \left(\frac{p}{p-1}\right)^p \frac{1}{(\omega^*)^p} < \left(\frac{p}{p-2}\right)^p \end{aligned} \tag{52}$$

□

Bemerkung 3.6 *Insbesondere gelten*

$$\left(\frac{p}{p-2}\right)^p > e^2$$

und

$$\lim_{p\to\infty}\left(\frac{p}{p-2}\right)^p = e^2 \approx 7.3891.$$

Fall $\sigma(J^p) \in \mathbb{R}$:

Satz 3.18 *Sind die Eigenwerte von J^p reell, d.h. $\sigma(J^p) \in [-\alpha^p, \beta^p]$, und ist ferner $\beta > \alpha$, so ist das Gauß–Seidel–Verfahren höchstens um den Faktor*

$$\left(\frac{2\max\{\alpha,\beta\}}{\alpha+\beta}\right)^p$$

schlechter als das SOR–Verfahren mit optimalem Relaxationsparameter.

Ist dagegen $\alpha > \beta$ so ist es höchstens um den Faktor

$$\left(\frac{\max\{\alpha,\beta\}}{\beta}\right)^p$$

schlechter.

Beweis:
Nach [11] ist der optimale Relaxationsparameter die Lösung von

$$\left(\frac{\alpha+\beta}{2}\omega\right)^p - \frac{\alpha+\beta}{\beta-\alpha}(\omega-1) = 0$$

in $(1, 1+(\beta-\alpha)/(\alpha+\beta))$ (falls $\beta > \alpha$) bzw. $(1-(\alpha-\beta)/(\alpha+\beta), 1)$ (falls $\alpha > \beta$). Für den optimalen Konvergenzfaktor gilt

$$\rho(\mathcal{L}_\omega^*) = \frac{\alpha+\beta}{\beta-\alpha}(\omega^*-1) = \left(\frac{\alpha+\beta}{2}\omega^*\right)^p.$$

Der subdominante Eigenwert von $\mathcal{L}_1$ gehört auf jeden Fall zum größeren der beiden Parameter α, β.

Damit gilt

$$\begin{aligned}\left|\frac{\rho_1}{\rho^*}\right| &= \left|\frac{\max^p\{\alpha,\beta\}}{\frac{\alpha+\beta}{\beta-\alpha}(\omega^*-1)}\right| \\ &= \left|\frac{2^p\max^p\{\alpha,\beta\}}{(\alpha+\beta)^p(\omega^*)^p}\right| \\ &< \left(\frac{2\max\{\alpha,\beta\}}{\alpha+\beta}\right)^p.\end{aligned}$$

Die Abschätzung für $\alpha > \beta$ ergibt sich, wenn man berücksichtigt, dass in diesem Fall $\omega^* > 1 - \frac{\alpha-\beta}{\alpha+\beta} = \frac{2\beta}{\alpha+\beta}$ gilt. □

Bemerkung 3.7 *Aus beiden Ergebnissen folgt, dass für $\alpha = \beta$ das Gauß–Seidel–Verfahren optimal ist.*

4 SOR bei p–periodischen Markov–Ketten mit komplexem Spektrum

In diesem Kapitel wird der allgemeinere Fall $\sigma(J^p) \subseteq \mathbb{C}$ betrachtet. Kernpunkt ist dabei die *extended convergence*.

Dazu wird zunächst aufgezeigt dass dieser Fall im Vergleich zum bisherigen besondere Schwierigkeiten birgt.

Aussagen von Hadjidimos et al. über die Wahl des optimalen Relaxationsparameters beim klassischen SOR–Verfahren werden im zweiten Abschnitt vorgestellt und erläutert (vgl. [18] und [19]).

Im dritten Abschnitt werden einige vorbereitende Lemmata gezeigt, die im Hauptabschnitt dieses Kapitels benötigt werden. In diesem, dem vierten Abschnitt, wird die *extended convergence* für Markov–Ketten behandelt, deren Jacobi–Matrix beliebiges Spektrum besitzt. Es wird untersucht, unter welchen Bedingungen weitere optimale Relaxationsparameter existieren und wo diese liegen. Es zeigt sich insbesondere, dass kein besserer Konvergenzfaktor als der im klassischen Fall optimale erreicht werden kann.

Im letzten Abschnitt soll abgeschätzt werden, um wieviel langsamer es ist, wenn statt SOR mit optimalem Relaxationsparameter einfach Gauß–Seidel durchgeführt wird.

4.1 Vorbetrachtungen

Wie in Korollar 3.9 bereits erwähnt, konvergiert die *extended convergence* nur gegen den gesuchten stationären Wahrscheinlichkeitsvektor, wenn der betragsgrößte Eigenwert von $\mathcal{L}_\omega$ dem Eigenwert–p–Tupel 1 von J^p zugeordnet ist.

Im Fall, dass $\sigma(J^p) \subseteq \mathbb{R}$ ist, gehört $\lambda_{\max}(\mathcal{L}_\omega)$ immer zum Eigenwert–p–Tupel 1 von J^p (vgl. auch [57]). Im Fall, dass es Eigenwerte $\mu^p \in \mathbb{C}\setminus\mathbb{R}$ von J^p gibt (vgl. das entsprechende Beispiel aus [57]; dieses Beispiel ist auch Gegenstand von Beispiel 4.1 und von Abschnitt 5.1), kann $\lambda_{\max}(\mathcal{L}_\omega)$ auch einem anderen Eigenwert μ mit $|\mu| < 1$ zugeordnet sein.

In diesem Fall ist das Verfahren nicht anwendbar. Die *extended convergence* als Vektoriteration betrachtet konvergiert zwar nach wie vor gegen den Eigenvektor x von $\lambda_{\max}(\mathcal{L}_\omega)$. Gehört $\lambda_{\max}(\mathcal{L}_\omega)$ aber nicht zum Eigenwert–p–Tupel 1 von J^p, so lässt sich aus x nicht mehr der gesuchte Eigenvektor π zum Eigenwert 1 von J rekonstruieren (zur Rekonstruktion vgl. [34]), d.h. diese Fälle, die ausgeschlossen werden müssen, sind ohnehin irrelevant.

In diesem Fall ergibt sich für das Intervall $(0, \frac{p}{p-1})$, dass für bestimmte Relaxationsparameter die komplexen Eigenwerte nicht im Innengebiet des Hypozykloids mit $\eta = 1$ liegen können, dass also das SOR–Verfahren nach Satz 3.1 nicht konvergieren kann. Liegt das Einpunktproblem (d.h. es existiert nur ein komplexex Eigenwert–$2p$–Tupel von J^p) mit $z = \alpha + i\beta$ als komplexem Eigenwert vor, ergeben sich der Grenzparameter ω, für den der wenigstens zweifache dominante Eigenwert sowohl der 1 als auch einem μ mit $|\mu| < 1$ zugeordnet ist, und das zugehörige ϑ als eine der Lösungen des nichtlinearen Gleichungssystems

$$\frac{\cos(\vartheta)}{\omega} + \frac{\omega - 1}{\omega}\cos((p-1)\vartheta) = \alpha \tag{53}$$

$$-\frac{\sin(\vartheta)}{\omega} + \frac{\omega - 1}{\omega}\sin((p-1)\vartheta) = \beta. \tag{54}$$

Hierbei beschreibt eine Lösung des Gleichungssystems ein gestrecktes Hypozykloid (ist damit nach Satz 4.1 uninteressant), die andere dagegen ist die interessante Lösung.

Für das erste Beispiel ergeben sich die beiden Hypozykloide aus Abbildung 13.

Zwei Beispiele

Beispiel 4.1 *Ein erstes Beispiel für die entstehende Problematik erwähnt Stewart in seinem Buch* Introduction to the Numerical Solution of Markov Chains *[57, Kapitel 7.2.2].*

Es gibt, wie Abbildung 14 zeigt, zwei Bereiche, in denen $\lambda_{\max}(\mathcal{L}_\omega)$ nicht zum Eigenwert 1 von J^p gehört:

Es gibt ein größeres Intervall $I_1 \approx [-3.0244, -0.5957]$ und ein kleineres Intervall $I_2 \approx [1.3943, 1.6281]$, so dass die extended convergence nicht den gewünschten Vektor berechnet, wenn $\omega \in I_1 \cup I_2$ gewählt wird.

Die beiden unteren Intervallgrenzen sind dabei gekoppelt durch die später eingeführte Gleichung (65).

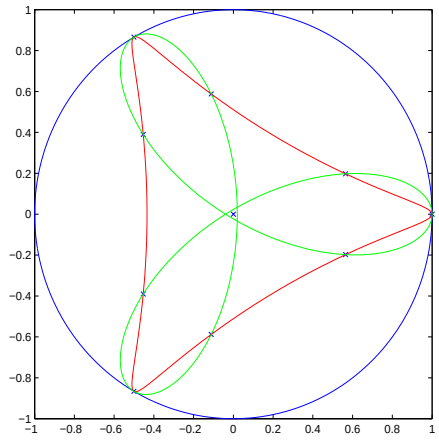

Abbildung 13: Die zu den Lösungen von (53) und (54) gehörenden Hypozykloide

Noch unangenehmer ist das Vielpunktproblem, bei dem nicht nur ein komplexes $2p$–Tupel von Eigenwerten von J vorliegt sondern mehrere. Als Verdeutlichung mag folgendes Beispiel dienen:

Beispiel 4.2 *Die Jacobi Matrix J^p besitze die Eigenwerte $\lambda_1 = 1$, $\lambda_2 \approx -0.182 + 0.61i$ und $\lambda_3 \approx -0.4885 + 0.3217i$. Die Eigenwerte von J sind in Abbildung 15 aufgetragen zusammen mit den jeweiligen Kreisen um den Ursprung, auf denen sie liegen.*

In diesem Fall ist zunächst unklar, welchem der Eigenwerttupel von J^p der dominante Eigenwert der SOR–Matrix zugeordnet ist (vgl. Abbildung 16), dem Eigenwert λ_1 (durchgezogene Linie), λ_2 (gestrichelte Linie) oder auch dem Eigenwert λ_3 (gestrichelte und gepunktete Linie). Man sieht zudem, dass keineswegs die klassische Konvergenz für $(0, \frac{p}{p-1})$ vorliegt (auch keine extended convergence*) und dass man, auch wenn klassische Konvergenz vorliegt, nicht ohne weiteres sagen kann, zu welchem Eigenwert–Tupel von J^p der subdominante Eigenwert von $\mathcal{L}_\omega$ gehört.*

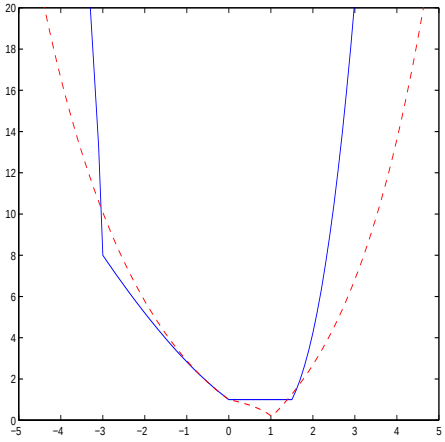

Abbildung 14: Betrag des größten zu 1 (durchgezogen) und zu $|\mu| < 1$ (gestrichelt) gehörenden Eigenwertes

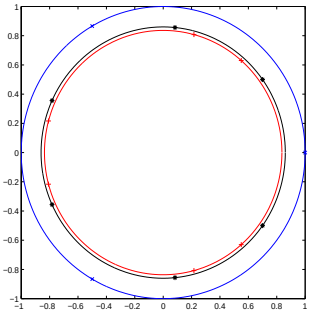

Abbildung 15: Die Eigenwerte der Jacobi–Matrix J.

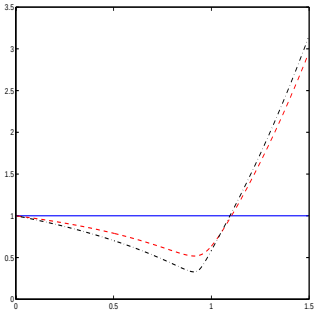

Abbildung 16: betragsgrößte Eigenwerte von $\mathcal{L}_\omega$ im Beispiel 4.2

4.2 Optimalitätsaussagen von Galanis, Hadjidimos und Noutsos

Zwei Arbeiten ([18] und — gerade Anfang dieses Jahres erschienen — [19]) von Galanis, Hadjidimos und Noutsos beinhalten Untersuchungen über den optimalen Relaxationsparameter für den Fall, dass das Spektrum von J^p — abgesehen vom Eigenwert 1 — komplex ist. Die verwendete Technik ist die, dass man unter allen möglichen Hypozykloiden, in deren Innengebiet bzw. auf deren Rand die Eigenwerte von J liegen, dasjenige sucht, dem der kleinste Konvergenzfaktor (und damit das größte η) zugeordnet ist. Diese Arbeiten beziehen sich dabei ausschließlich auf das klassische SOR–Verfahren, die *extended convergence* spielt keine Rolle.

Ein wesentlicher Punkt bei der Konvergenzuntersuchung des SOR–Verfahrens ist die Beschränkung auf den Winkel $\mathcal{S}$ zwischen dem Strahl $-\pi/p$ und der reellen Achse, weil aufgrund von Satz 2.7 das Spektrum invariant ist unter Drehungen um den Winkel $2\pi/p$ (vergleiche auch Abbildung 17) und von jedem $2p$–Tupel[13] genau ein Eigenwert in S liegt.

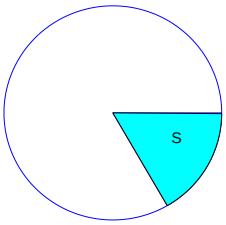

Abbildung 17: Segment $\mathcal{S}$ für $p = 3$

Liegt im Segment $\mathcal{S}$ genau ein Eigenwert (dieser Fall ist in [18] untersucht), sucht man ausgehend von den beiden gewöhnlichen Hypozykloiden durch diesen Punkt das Hypozykloid, das dem optimalen Konvergenzfaktor zugeordnet ist. Die Untersuchung bezieht sich ausschließlich auf Hypozykloide, was insofern ausreichend ist, als bei einem konvergenten Verfahren alle Eigenwerte — außer dem 1 zugeordneten p–Tupel — im Innengebiet bzw. auf dem Rand dieser Rollkurven liegen müssen.

Liegen mehrere Eigenwerte in $\mathcal{S}$, ist zu untersuchen, wann Eigenwerte im Innen-

[13] Da die Generatormatrix als reell vorausgesetzt war, ist das Spektrum ferner achsensymmetrisch zur rellen Achse.

gebiet von Hypozykloiden bezüglich anderer Eigenwerte liegen. Dieser Fall ist in [19] recht technisch und konstruktiv behandelt; im Weiteren wird der Inhalt dieser Arbeit möglichst anschaulich mit eigenen Worten wiedergegeben.

Bezeichnung: *Gegeben sei das Spektrum von J^p. Ein Hypozykloid $H(\omega, \eta, p, .)$ heißt* bestes Hypozykloid *oder* optimales Hypozykloid*, wenn $\tilde{\sigma}(J^p)$ im Innern des Hypozykloids liegt und wenigstens einer dieser Eigenwerte auf dem Hypozykloid liegt, d.h. wenn*

$$\tilde{\sigma}(J) \subseteq \overline{U}_\eta(\omega)$$

liegt und es kein $\omega \in (0, \frac{p}{p-1})$ ($\omega \notin [0, \frac{p}{p-1}]$) gibt, sodass diese Bedingung für ein größeres (kleineres) zugehöriges η erfüllt ist.

Dieses optimale Hypozykloid ist dem optimalen Relaxationsparameter zugeordnet.

Die Technik, die in [18] und [19] angewendet wird, ist die, dass man unter allen Hypozykloiden, die durch den (oder die) in diesem Segment liegenden Punkt P (bzw. Punkte $P_1, \ldots, P_k$) mit den Polarkoordinaten r und ψ (bzw. $r_1, \ldots, r_k$ und $\psi_1, \ldots, \psi_k$) und den kartesischen Koordinaten $\alpha = r\cos(\psi), \quad \beta = r\sin(\psi)$ (bzw. $\alpha_1, \ldots, \alpha_k$ und $\beta_1, \ldots, \beta_k$) gehen, dasjenige wählt, das den besten Kovergenzfaktor erzielt.

Dabei ist zunächst folgende Feststellung wichtig, die sich aus [18, Theorem 3.2 und Theorem 3.3] ergibt und die Anzahl der überhaupt infrage kommender Hypozykloide zum Teil deutlich verkleinert.

Satz 4.1 *Seien zwei Hypozykloide, die beide einem konvergenten SOR–Verfahren zugeordnet sind, mit gemeinsamer imaginärer (reeller) Halbachse a (b) gegeben, wobei eines verkürzt, das andere gestreckt ist. Dann gilt für die Spektralradien der zugehörigen SOR–Verfahren $\rho(\mathcal{L}_{\omega_{vk}})$ (verkürztes Hypozykloid) bzw. $\rho(\mathcal{L}_{\omega_{gt}})$ (gestrecktes Hypozykloid) die Beziehung*

$$\rho(\mathcal{L}_{\omega_{vk}}) < \rho(\mathcal{L}_{\omega_{gt}}). \tag{55}$$

Mit anderen Worten: es genügt bei der Untersuchung konvergenter SOR–Verfahren und insbesondere bei der Suche nach dem optimalen Relaxationsparameter, wenn man gewöhnliche Hypozykloide (die für die Grenzfälle nichtnegativer oder nichtpositiver Eigenwerte von J^p immer die optimalen Hypozykloide sind) und verkürzte Hypozykloide betrachtet.

Die Halbachsen der Hypozykloide durch P haben dann die Form

$$b = \frac{r}{\cos(p\vartheta/2)} \cos\left[\left(\frac{p}{2}-1\right)\vartheta - \psi\right], \tag{56}$$

$$a = \frac{r}{\sin(p\vartheta/2)} \sin\left[\left(\frac{p}{2}-1\right)\vartheta - \psi\right], \tag{57}$$

wobei ϑ den entsprechenden Wert in den Gleichungen des Hypozykloids (27) bzw. (28) bezeichnet, der dem Punkt P zugeordnet ist.

Das Einpunktproblem

Liegt ein komplexes $2p$–Tupel vor, dann gibt es im bereits angesprochenen Segment $\mathcal{S}$ genau einen Punkt P und es gehen genau zwei gewöhnliche Hypozykloide mit $0 < \vartheta_{II} < -\psi < \vartheta_I < \frac{\pi}{p}$ durch P. Diese sind durch die Halbachsen a_I und b_I bzw. a_{II} und b_{II} gegeben. Durch ϑ_I und ϑ_{II} ist dabei jeweils die Lage von P auf dem Hypozykloid bestimmt.

Abbildung 18 zeigt die beiden gewöhnlichen Hypozykloide im Fall $p = 3$ mit $z = \alpha + i\beta = 0.7 + 0.1i$.

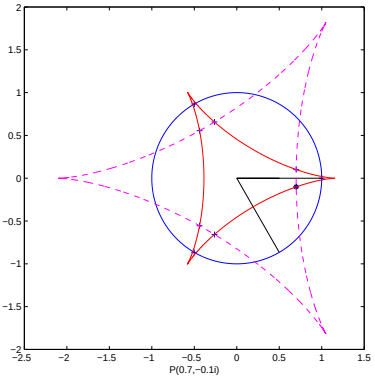

Abbildung 18: Gewöhnliche Hypozykloide, die durch P gehen

Es zeigt sich, dass die Größen a, b und ϑ teilintervallweise bijektiv aufeinander abgebildet werden können und dass $\frac{\partial a}{\partial b}$ bzw. $\frac{\partial \vartheta}{\partial b}$ monoton sind, sodass sinnvollerweise von $\vartheta(b)$ etc. gesprochen werden kann.

Damit kann nach folgendem Satz dasjenige $\hat{\vartheta}$ berechnet werden, mit dem nach

(56) und (57) das optimale Hypozykloid berechnet werden kann.

Satz 4.2 *Gelte $b_{II} < 1$ und sei $x_0 = x_0(b)$ Lösung von $(b-a)x^p - 2x + b + a = 0$ ($x := \frac{1}{\eta} = \tilde{\rho}^{1/p}(\mathcal{L}_\omega)$) im Intervall $[b_{II}, \min\{1, b_I\}]$. Dann existiert ein eindeutiger Wert von b (ϑ) in $(b_{II}, \min\{1, b_I\})$ ($(\vartheta_{II}, \vartheta(\min\{1, b_I\}))$), bei dem $x_0 = x_0(b)$ ($x_0 = x_0(\vartheta)$) sein Minimum annimmt. Der zugehörige Wert ϑ_0 ist gegeben als eindeutige Lösung von*

$$\left(r \frac{\sin(p\vartheta) - p\sin(\vartheta+\psi)\cos[(p-1)\vartheta - \psi]}{\sin[(p-1)\vartheta - \psi] - (p-1)\sin(\vartheta+\psi)\cos(p\vartheta)} \right)^p = \frac{\sin[(p-1)\vartheta - \psi]\cos(p\vartheta) - (p-1)\sin(\vartheta+\psi)}{\sin[(p-1)\vartheta - \psi] - (p-1)\sin(\vartheta+\psi)\cos(p\vartheta)} \tag{58}$$

in $(\vartheta_{II}, \vartheta(\min\{1, b_I\}))$ und der Wert x_0^p ist gegeben durch

$$x_0^p = \frac{\sin[(p-1)\hat{\vartheta} - \psi]\cos(p\hat{\vartheta}) - (p-1)\sin(\hat{\vartheta}+\psi)}{\sin[(p-1)\hat{\vartheta} - \psi] - (p-1)\sin(\hat{\vartheta}+\psi)\cos(p\hat{\vartheta})}.$$

Bemerkung 4.1 *Der optimale Relaxationsparameter berechnet sich aus den Parametern a_0, b_0 und x_0 dann mittels*

$$\omega_0 = \frac{2x_0}{a_0 + b_0}.$$

Zeichnet man das so erhaltene optimale Hypozykloid noch in die obige Figur ein, ergibt sich Abbildung 19.

Das Mehrpunktproblem

Auch hierbei beschränkt man sich auf das Kreissegment $\mathcal{S}$ zwischen dem Strahl $-\pi/p$ und der reellen Achse, wobei in diesem Fall k komplexe Zahlen $z_1, \ldots, z_k$ in diesem Segment liegen sollen (mit den Polarkoordinaten r_i, ψ_i und den kartesischen Koordinaten (α_i, β_i), $\quad i = 1, \ldots, k$).

Ein wichtiger Satz in diesem Zusammenhang ist (vgl. [19, Theorem 3.2])

Satz 4.3 *Zwei verschiedene zur reellen Achse symmetrische Hypozykloide mit p kongruenten Bögen können in dem Segment zwischen dem Strahl $-\pi/p$ und der reellen Achse höchstens einen gemeinsamen Punkt haben.*

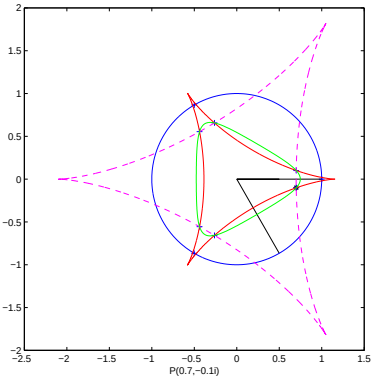

Abbildung 19: Gewöhnliche Hypozykloide, die durch P gehen, und das optimale

Daraus folgt, dass durch zwei feste Punkte in diesem Segment also höchstens ein Hypozykloid gehen kann.

Satz 4.4 *Seien $P_1(r_1, \psi_1)$ und $P_2(r_2, \psi_2)$ zwei Punkte mit $r_1, r_2 > 0$ und $-\frac{\pi}{p} < \psi_2 < \psi_1 < 0$. Seien $H_{I,1}$ und $H_{II,1}$ mit den Halbachsen $(a_{I,1}, b_{I,1})$ bzw. $(a_{II,1}, b_{II,1})$ die beiden gewöhnlichen Hypozykloide durch P_1. Entsprechend seien $H_{I,2}$ und $H_{II,2}$ die beiden gewöhnlichen Hypozykloide durch P_2. Die Punkte $A_{a_{I,1}}$ bzw. $A_{a_{II,1}}$ seien durch die Polarkoordinaten $(a_{I,1}, -\frac{\pi}{p})$ und $(a_{II,1}, -\frac{\pi}{p})$ gegeben. Liegt P_2 im Innern des kurvilinearen Dreiecks $A_{a_{I,1}} P_1 A_{a_{II,1}}$, dann existiert ein eindeutiges Hypozykloid, das durch P_1 und P_2 geht.*

Die folgenden Graphiken demonstrieren diesen Sachverhalt. Abbildung 20 zeigt dabei den Punkt P im Segment zwischen $-\pi/p$ und der rellen Achse sowie die beiden zugehörigen gewöhnlichen Hypozykloide.

Figur 21 zeigt als Ausschnitt das entscheidende Segment, außerdem sind die Punkte $A_{a_{I,1}}$ und $A_{a_{II,1}}$ eingezeichnet, insbesondere ist also das kurvilineare Dreieck aus Satz 4.4 zu erkennen.

In den abschließenden Figuren in Abbildung 22 ist in diesem kurvilinearen Dreieck ein weiterer Punkt P_2 gewählt und zusätzlich das entsprechende eindeutige durch P_1 und P_2 gehende Hypozykloid eingezeichnet.

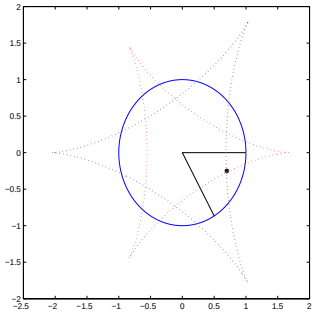

Abbildung 20: P und die zugehörigen gewöhnlichen Hypozykloide

Satz 4.5 *Die Punkte P_1 und P_2 sollen die Voraussetzungen von Satz 4.4 erfüllen, ferner sei $b_{II,1} < 1$. Enthält das optimlale Hypozykloid H_1 bezüglich P_1 (bzw. das optimale Hypozykloid H_2 bezüglich P_2) den Punkt P_2 (bzw. P_1) im Abschluss des Inneren, ist dieses Hypozykloid das optimale bezüglich den Punkten P_1 und P_2. Enthält keines dieser Hypozykloide H_1 und H_2 den jeweils anderen Punkt, ist das optimale Hypozykloid dasjenige $H_{1,2}$ bezüglich beiden Punkten.*

Die Aufgabe, für das Mehrpunktproblem den optimalen Relaxationsparameter zu bestimmen, wird konstruktiv gelöst, indem im Wesentlichen jeweils zwei der Punkte und die zugehörigen Hypozykloide H_1, H_2 und $H_{1,2}$ unter Berücksichtigung von Satz 4.3 und Satz 4.5 verglichen werden.

Die Algorithmen, die in [17] angegeben sind, sind sehr störungsanfällig. Man muss bis zu einer größtmöglichen Genauigkeit iterieren, um das gesuchte verkürzte Hypozykloid zu erhalten, wohingegen man bei kleinen Störungen bereits Divergenz erhält.

Beim Zwei–Punkt–Problem kann aufgrund von Beziehungen der reellen Halbachsen das optimale Hypozykloid bestimmt und aus dessen Parametern der optimale Relaxationsparameter und der Konvergenzfaktor gewonnen werden.

Bemerkung 4.2 *Für die reellen Halbachsen b_1, b_2 und $b_{1,2}$ gilt:*

Ist $b_1 \leq b_{1,2}$, ist das Hypozykloid H_1 nur bezüglich P_1 optimal.

Ist $b_{1,2} \leq b_2$, so ist das Hypozykloid H_2 nur bezüglich P_2 optimal.

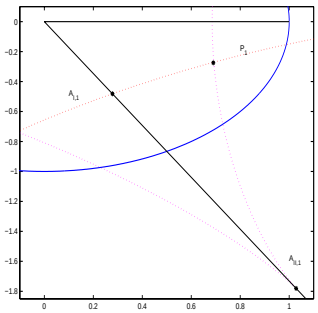

Abbildung 21: Das kurvilineare Dreieck aus Satz 4.4

Andernfalls ist das Hypozykloid $H_{1,2}$ zu P_1 und P_2 optimal.

Abbildung 23 zeigt neben den beiden komplexen Eigenwerten jeweils die optimalen Hypozykloide bezüglich nur eines Eigenwertes sowie das optimale Hypozykloid bezüglich beider Eigenwerte.

Nach Satz 4.4 ist P_1 der näher an der reellen Achse gelegene Punkt. Man sieht, dass offenbar $b_1 > b_{1,2}$ ist, außerdem ist $b_2 < b_{1,2}$, das heißt das Hypozykloid bezüglich beiden Punkten bezeichnet das optimale Verfahren. Dies ist auch anschaulich sofort klar, weil der jeweils andere Punkt nicht im Innengebiet bzw. auf dem Rand des nur für einen Punkt optimalen Hypozykloids liegt.

Beim Mehrpunktproblem werden zunächst alle komplexen Eigenwerte, die keinen Einfluss auf den optimalen Relaxationsparameter haben, aus der weiteren Betrachtung ausgeschlossen. Von den verbleibenden Eigenwerten betrachtet man alle möglichen Kombinationen von Zweipunktproblemen. Dadurch erhält man zuletzt ein optimales Hypozykloid und einen optimalen Relaxationsparameter.

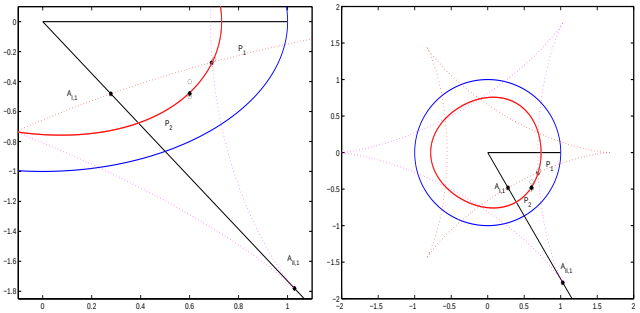

Abbildung 22: Die vollständige Figur mit dem optimalen Hypozykloid

Vorbereitende Schritte beim Mehrpunktproblem

Im Weiteren werden zunächst die Schritte zur Vereinfachung des Problems vorgestellt. Die Begründung, warum diese Schritte durchgeführt werden und sinnvoll sind, wird im Anschluss gegeben.

Im Segment zwischen dem Strahl $-\pi/p$ und der positiven reellen Achse liegen k komplexe Zahlen $z_1, \ldots, z_k$ mit den Polardarstellungen $r_j e^{i\psi_j} \quad j = 1, \ldots, k$ und $0 \geq \psi_1 \geq \psi_2 \geq \ldots \geq \psi_k \geq -(\pi/p)$.

1. Haben mehrere z_j denselben Winkel ψ_j, werden alle bis auf dasjenige mit dem größten Radius r_j entfernt.

   Man erhält dadurch eine Menge von k_1 komplexen Zahlen im Segment $\mathcal{S}$.

2. Alle Punkte müssen, damit SOR konvergiert, enthalten sein im kurvilinearen Dreieck OBA mit $O(0,0), B(1,0), A(\frac{p-2}{p}, -\frac{\pi}{p})$ ausgeschlossen[14] der Kurve BA.

   Ist dies der Fall, erhält man für jeden Punkt P_j das gewöhnliche Hypozykloid $H_{II,j}$ mit der reellen Halbachse $b_{II,j}$. Ist ein $b_{II,j} \geq 1$, dann existiert kein konvergentes SOR–Verfahren.

[14] In [19] ist der Punkt A fälschlicherweise als $A(\frac{p}{p-2}, -\frac{\pi}{p})$ angegeben. Dieser Punkt liegt aber nicht auf dem durch $(1,0)$ gehenden Hypozykloid.

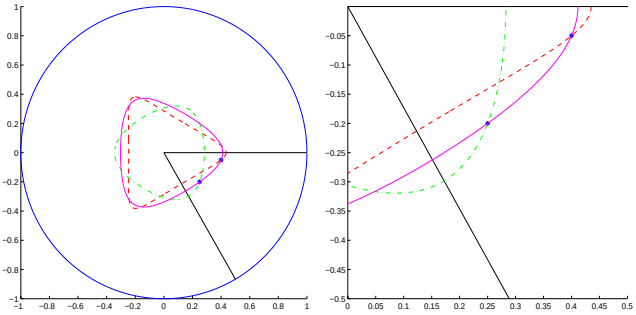

Abbildung 23: Das Zweipunktproblem, rechts ein Ausschnitt

3. Existiert ein konvergentes SOR–Verfahren, ordnet man die Punkte nach nichtsteigender reeller Halbachse $b_{II,1} \geq b_{II,2} \geq \ldots \geq b_{II,k_I}$. Gilt Gleichheit, wird die Zahl mit kleinstem ψ_j behalten und alle anderen entfernt, sodass k_2 Eigenwerte bleiben.

   Man bestimmt dann die gewöhnlichen Hypozykloide $H_{I,j}$ mit den reellen Halbachsen $b_{I,j}$ $(j = 1, \ldots, k_2)$ durch diese k_2 Punkte.

   Ist für ein j $b_{I,j-1} \geq b_{I,j}$, so wird P_j entfernt.

   Danach bleiben k_3 Eigenwerte übrig.

Als nächstes sollen die drei vorbereitenden Schritte begründet werden, was in [19] etwas kurz kommt.

ad 1.) Aufgrund der Gestalt der Hypozykloide ist klar, dass mit $re^{i\psi}$ auch $\tilde{r}e^{i\psi}$ mit $\tilde{r} < r$ im Innern des Hypozykloids enthalten ist; daher können solche Zahlen von vorneherein entfernt werden.

ad 2.) Der Grund dafür, dass alle Zahlen, die nicht in dem kurvilinearen Dreieck liegen, herausfallen, ist, dass in diesem Fall die reelle Halbachse des optimalen Hypozykloids größer oder gleich 1 wird. Damit ist, weil die reelle Halbachse entweder Inkreis– oder Umkreisradius des Hypozykloids ist, klar, dass auch die Eigenwerte von J mit Betrag 1 innerhalb des Hypozykloids liegen, was zu einem nicht konvergenten Verfahren führt oder aber zu einem Verfahren, für das es ein besseres Verfahren mit anderem Relaxationsparameter und kleinerem Konvergenzfaktor gibt.

Abbildung 24 zeigt das kurvilineare Dreieck.

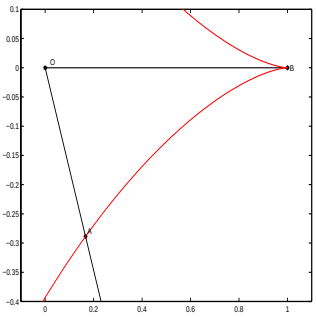

Abbildung 24: Das kurvilineare Dreieck OBA

ad 3.) Besitzen mehrere Punkte dieselbe reelle Halbachse, führt aus geometrischen Gründen derjenige Punkt mit kleinstem Winkel ψ_j zum ungünstigsten Hypozykloid, weswegen alle anderen weggelassen werden können.

Dass Punkte P_j mit $b_{I,j-1} \geq b_{I,j}$ weggelassen werden, ist wiederum aus geometrischen Gründen einsichtig.

Für die verbleibenden Eigenwerte P_j, $1 \leq j \leq k_3$ gilt dann

$$b_{II,k_3} < b_{II,k_3-1} < \ldots < b_{II,1} \leq b_{I,1} < b_{I,2} < \ldots < b_{I,k_3} \quad \text{und } b_{II,1} < 1. \tag{59}$$

Bestimmung des optimalen Hypozykloids

Das optimale Hypozykloid wird nun bestimmt, indem die k_3 optimalen Hypozykloide H_j bezüglich der Punkte P_j $(j = 1, \ldots, k_3)$ und die $\binom{k_3}{2}$ optimalen Hypozykloide H_{j_1,j_2} bezüglich der Punktepaare P_{j_1}, P_{j_2}, $1 \leq j_1 < j_2 \leq k_3$ bestimmt werden, was mit Hilfe der Lösung des Zweipunktproblems möglich ist.

Es gilt dann (vgl. [19, Theorem 5.1 und Theorem 5.2])

Satz 4.6 *Die vorbereitenden Schritte seien wie beschrieben durchgeführt und (59) gelte. Dann existiert die Lösung des Minimierungsproblems*

$$\min_{b \in [b_{II,1}, \min\{b_{I,k},1\}]} \max\{x_j(b), \quad j = 1, \ldots, k_3\}. \tag{60}$$

Ferner ist die Lösung gegeben durch die Elemente eines der k_3 optimalen Hypozykloide H_j durch die Punkte P_j, $j = 1, \ldots, k_3$ oder eines der $\binom{k_3}{2}$ Hypozykloide H_{j_1,j_2} durch die Punktepaare P_{j_1}, P_{j_2}, $1 \leq j_1 < j_2 \leq k_3$.

Das optimale Hypozykloid ist eindeutig.

4.3 Einige Lemmata

Im Weiteren wird wieder die *extended convergence* von Kontovasilis, Plemmons und Stewart behandelt (vgl. [34]).

Dabei werden nach einer Vorbemerkung vorbereitende Lemmata gezeigt, die für den Beweis der Kernaussagen in Kapitel 4.4 benötigt werden:

Bemerkung 4.3 *Die Hypozykloide $H(\omega, \eta, p, .)$ und $H(\omega, \eta, p, . + \pi)$ liegen in der komplexen Ebene gleich.*

Das liegt daran, dass für die Gestalt des Hypozykloids nur die Parameter ω, η und p von Bedeutung sind, ϑ aber nur den genauen Punkt auf dem Hypozykloid bezeichnet.

Lemma 4.7 *Für das zu festem $\omega \in \mathbb{R} \setminus \{0\}$ gehörende Hypozykloid $H(\omega, \eta, p, .)$ mit größtmöglichem (Standard–SOR) bzw. kleinstmöglichem (*extended convergence*) η gilt*

$$\frac{1}{\eta^p} > 1 - \omega. \tag{61}$$

Beweis:
Für $\omega > 1$ ist die Behauptung wegen $\eta > 0$ trivial.

Andernfalls nehmen wir an, es gelte

$$\frac{1}{\eta^p} \leq 1 - \omega,$$

d.h. für $\omega \in (0, 1)$

$$\frac{1}{\omega\eta} \leq \frac{1-\omega}{\omega}\eta^{p-1}$$

und für $\omega \in \mathbb{R}^-$

$$-\frac{1}{\omega\eta} \leq \frac{\omega-1}{\omega}\eta^{p-1},$$

was in beiden Fällen gleichbedeutend mit der Beziehung

$$R - r \leq h$$

ist, wenn R, r, h die Größen des Hypozykloids bezeichnen (vgl. Abschnitt 3.2).

Also haben wir es mit einem leeren oder auf einen Punkt zusammengezogenen Hypozykloid zu tun. Dies bezeichnet aber kein konvergentes Verfahren, d.h. es kann sich im Widerspruch zur Voraussetzung nicht um ein optimales Hypozykloid handeln. □

Lemma 4.8 *Ist $\omega < 0$, so liegt der zu $\vartheta_0 := 0$ gehörende Punkt des Hypozykloids auf der negativen rellen Achse, ansonsten auf der positiven.*

Beweis:
Es gelten

$$\mathrm{Im}(H(\omega,\eta,p,\vartheta_0)) = y(\vartheta_0) = -\frac{1}{\omega\eta}\underbrace{\sin(\vartheta_0)}_{=0} + \frac{\omega-1}{\omega}\eta^{p-1}\underbrace{\sin((p-1)\vartheta_0)}_{=0} = 0$$

und

$$\mathrm{Re}(H(\omega,\eta,p,\vartheta_0)) = \frac{1}{\omega\eta} + \frac{\omega-1}{\omega}\eta^{p-1}.$$

Für $\omega < 0$ ist

$$\frac{1}{\omega\eta} + \frac{\omega-1}{\omega}\eta^{p-1} < 0$$

äquivalent zu

$$\frac{1}{\omega\eta} < \frac{1-\omega}{\omega}\eta^{p-1}$$

bzw. zu

$$\frac{1}{\eta^p} > 1-\omega,$$

was wegen Lemma 4.7 richtig ist.

Für $\omega > 0$ ist entsprechend

$$\frac{1}{\omega\eta} + \frac{\omega-1}{\omega}\eta^{p-1} > 0$$

äquivalent zu

$$\frac{1}{\eta^p} > 1-\omega,$$

was wiederum wegen Lemma 4.7 richtig ist. □

Das folgende Lemma stellt den Kern des Beweises zu den Beziehungen der optimalen Relaxationsparameter untereinander in Kapitel 4.4 dar:

Lemma 4.9 *Sei* $f(x) := \frac{x-1}{x^p}, p \in \mathbb{N}$. *Für* f *gilt:*

1. f *hat ein lokales Maximum für* $\hat{x} := \frac{p}{p-1}$ *mit* $f(\hat{x}) = \frac{(p-1)^{p-1}}{p^p}$.
2. f *ist in* $(0, \frac{p}{p-1})$ *streng monoton wachsend mit* $\lim_{x \to 0+} f(x) \to -\infty$.
3. f *ist in* $(\frac{p}{p-1}, \infty)$ *streng monoton fallend mit* $\lim_{x \to \infty} f(x) = 0$.
4. *An der Stelle* $x = 0$ *liegt für gerades* p *bzw. ungerades* p *ein Pol ohne bzw. mit Vorzeichenwechsel vor.*
5. *In* $(-\infty, 0)$ *ist* f *streng monoton, und es gilt* $\lim_{x \to -\infty} f(x) = 0$.

Beweis:

1. Es ist
$$f'(x) = \frac{x - (x-1)p}{x^{p+1}}$$
und
$$f''(x) = \frac{-2px + (x-1)p^2 + (x-1)p}{x^{p+2}}.$$
Das heißt
$$f'(x) = 0 \Leftrightarrow (1-p)x = -p \Leftrightarrow \hat{x} = -\frac{p}{1-p} = \frac{p}{p-1}.$$
Ferner ist[15]
$$f''\left(\frac{p}{p-1}\right) \simeq -2\frac{p^2}{p-1} + \frac{p^2}{p-1} + \frac{p}{p-1} = \frac{p-p^2}{p-1} < 0,$$
d.h. es liegt ein relatives Maximum vor. Wie man leicht durch Nachrechnen sieht, ist $f(\hat{x}) = \frac{(p-1)^{p-1}}{p^p}$.

[15] Die Schreibweise $A \simeq B$ soll bedeuten, dass A und B dasselbe Vorzeichen besitzen.

2. Wegen $x^{p+1} > 0$ in $(0, \frac{p}{p-1})$ genügt es, den Zähler des Bruches zu betrachten.

$$f'(x) = (1-p)x + p > 0 \Leftrightarrow x < -\frac{p}{1-p} = \frac{p}{p-1}$$

und es gilt

$$\lim_{x \to 0+} \frac{x-1}{x^p} = \lim_{x \to 0+} -\frac{1}{x^p} \to -\infty.$$

3. Analog ist für $x > \frac{p}{p-1}$ $f'(x) < 0$ und es gilt[16]

$$\lim_{x \to \infty} \frac{x-1}{x^p} = \lim_{x \to \infty} \frac{1}{px^{p-1}} = 0.$$

4. Sei $x < 0$ und $x \to 0$.

Für gerade p ist $\frac{x-1}{x^p} < 0$, d.h.

$$\lim_{x \to 0-} \frac{x-1}{x^p} \to -\infty,$$

d.h. es liegt (mit 2.) ein Pol ohne Vorzeichenwechsel vor.

Für ungerade p ist $\frac{x-1}{x^p} > 0$, d.h.

$$\lim_{x \to 0-} \frac{x-1}{x^p} \to \infty,$$

d.h. es liegt (mit 2.) ein Pol mit Vorzeichenwechsel vor.

5. Für $x \in (-\infty, 0)$ ist

$$\text{Zähler}(f'(x)) = \underbrace{\underbrace{(1-p)}_{<0}\underbrace{x}_{<0}}_{>0} + p > 0.$$

Ferner ist x^{p+1} für ungerade p größer, für gerade p kleiner als 0. Das heißt, f wächst streng monoton, falls p ungerade ist, und fällt streng monoton, falls p gerade ist.

Außerdem gilt

$$\lim_{x \to -\infty} \frac{x-1}{x^p} = \lim_{x \to -\infty} \frac{1}{px^{p-1}} = 0.$$

□

[16] mit der Regel von de l'Hospital

Die folgende Tabelle zeigt zusammengefasst den Wertebereich von f in Abhängigkeit von p:

$x \in$	p gerade	p ungerade
$(0,1)$	$(-\infty,0)$	$(-\infty,0)$
$\left(1,\frac{p}{p-1}\right)$	$\left(0,\frac{(p-1)^{p-1}}{p^p}\right)$	$\left(0,\frac{(p-1)^{p-1}}{p^p}\right)$
$\left(\frac{p}{p-1},\infty\right)$	$\left(0,\frac{(p-1)^{p-1}}{p^p}\right)$	$\left(0,\frac{(p-1)^{p-1}}{p^p}\right)$
$(-\infty,0)$	$(-\infty,0)$	$(0,\infty)$

Besonders wichtig ist, dass in den einzelnen Teilintervallen f jeweils eine streng monotone Funktion ist, d.h. die Abbildung $f : x \to f(x)$ ist — bezogen auf die einzelnen vier Teilintervalle — bijektiv.

Im Fall, dass p ungerade ist, ist aufgrund obiger Tabelle derjenige Wert $\hat{x} < 0$ von Bedeutung, für den

$$\frac{x-1}{x^p} = \frac{(p-1)^{p-1}}{p^p}$$

ist, weil f Werte, die größer als $\frac{(p-1)^{p-1}}{p^p}$ sind, nur für $\hat{x} < x < 0$ annehmen kann[17].

Lemma 4.10 *Sei p ungerade. Es existiert ein eindeutiges $\hat{x} \in (-\infty,0)$, sodass*

$$\frac{x-1}{x^p} > \frac{(p-1)^{p-1}}{p^p}$$

für alle $x \in (\hat{x},0)$ gilt.

$\hat{x}$ ist die eindeutige Nullstelle von

$$(p-1)^{p-1}x^p - p^p(x-1) \tag{62}$$

in $(-\infty,0)$.

Beweis:
Sei p ungerade, $x < 0$. Dann ist

[17] Dieser Punkt $\hat{x}$ taucht auch in [34, Theorem 4.2] als ω^* auf, ohne dass dort aber näher auf die Besonderheit dieses Punktes eingegangen wird.

$$\frac{x-1}{x^p} > \frac{(p-1)^{p-1}}{p^p} \Leftrightarrow 0 < (p-1)^{p-1}x^p - (x-1)p^p =: h(x).$$

Es ist

$$h(x) = (p-1)^{p-1}x^p - p^p x + p^p$$

d.h. (da p ungerade ist)

$$h(-x) = \underbrace{-(p-1)^{p-1}|x^p|}_{<0}\underbrace{+p^p|x|}_{>0}\underbrace{+p^p}_{>0}.$$

Nach der Descartesschen Regel muss die Differenz aus der Anzahl der Vorzeichenwechsel (hier einer) und der der Nullstellen gerade und nichtnegativ sein, d.h. es muss genau eine Nullstelle von $h(-x)$ in $[0,\infty)$, d.h. genau eine Nullstelle von $h(x)$ in $(-\infty,0)$, nämlich eben $\hat{x}$ geben. □

Bemerkung 4.4 *Der Wert dieser Nullstelle $\hat{x}$ hängt nur noch von p ab. Folgende Tabelle zeigt die Werte der ersten ungeraden p:*

p	$\hat{x}$
3	−3.00000000
5	−2.06328649
7	−1.74043099
9	−1.57444451
11	−1.47245030
13	−1.40302411
15	−1.35251013

Für $0 > x > \hat{x}$ sind die zu $\mu^p = 1$ gehörenden größten Eigenwerte außer 1 konjugiert komplex, wie man am Beispiel $p = 3$ sieht:

die Gleichung von Varga (vgl. [64])

$$(\lambda + \omega - 1)^3 = \lambda^2\omega^3$$

besitzt die Lösungen

$$\lambda_1 = 1, \lambda_{2/3} = (\omega - 1)\left(\frac{1}{2}\omega^2 + \frac{1}{2}\omega - 1 \pm \frac{1}{2}\omega\sqrt{\underbrace{\omega^2 + 2\omega - 3}_{=:d(\omega)}}\right).$$

Die Nullstellen von d sind

$$\omega_1 = 1, \omega_2 = -3 = \hat{\omega},$$

und da d eine nach oben geöffnete Parabel beschreibt, sind λ_2 *und* λ_3 *für* $\omega \in (-3, 1)$ *konjugiert komplex.*

Dieser kleine, scheinbar unwichtige Teil verursacht nun noch einige Mühen (so auch schon in [34]). Es gilt für diesen speziellen Fall folgendes Lemma:

Lemma 4.11 *Ist p ungerade, so kann es in* $(\hat{\omega}, 0)$ *keinen Relaxationsparameter* ω *mit optimalem Konvergenzfaktor geben.*

Beweis:
Sei p ungerade.

Betrachtet wird zunächst der Punkt $\hat{\omega}$ selbst. Für $\hat{\omega}$ gilt nach (62)

$$1 - \hat{\omega} = -\frac{(p-1)^{p-1}}{p^p}\hat{\omega}^p. \tag{63}$$

Die größte Nullstelle von

$$(\lambda + \omega - 1)^p = \lambda^{p-1}\omega^p\mu^p$$

wächst mit $|\mu|$, ist betragsmäßig somit auf jeden Fall größer als die Lösung von

$$(\lambda + \omega - 1)^p = 0,$$

d.h. $1 - \omega$, wenn $\mu \to 0$ geht.

Daraus folgt, wenn man den Konvergenzfaktor aus Sicht der Vektoriteration betrachtet und (63) für den subdominanten Eigenwert von $\mathcal{L}_\omega$ verwendet, dass

$$\left|\frac{\lambda_{\text{subdominant}}}{\lambda_{\text{dominant}}}\right| \geq \frac{\frac{(p-1)^{p-1}}{p^p}\hat{\omega}^p}{\frac{(p-1)^{p-1}}{p^p}\hat{\omega}^p} = \frac{1}{p-1}, \tag{64}$$

wenn man für den Nenner [34, Theorem 4.2(b)(ii)], wo der exakte dominante Eigenwert bestimmt ist für den Fall, dass dieser dem Eigenwert 1 von J^p zugeordnet ist, verwendet.

Für $\omega = \hat{\omega}$ genügt es also, wenn man zeigt, dass der optimale Konvergenzfaktor kleiner oder gleich $\frac{1}{p-1}$ ist.

Gilt für den subdominanten Eigenwert μ^p von J^p, dass $\mu^p < \frac{1}{p-1}$, ist die Behauptung trivialerweise gezeigt, da dann sogar das Gauß–Seidel–Verfahren eine schnellere Konvergenz ermöglicht, da beim Gauß–Seidel–Verfahren (12) sich zu

$$\lambda^p = \lambda^{p-1}\mu^p$$

vereinfacht.

Ist $\mu^p \geq \frac{1}{p-1}$, so wächst der zu μ^p gehörende Eigenwert der SOR–Matrix noch, das heißt, auch hier wird die Konvergenz schlechter.

Insgesamt: für $\omega = \hat{\omega}$ ist der Konvergenzfaktor schlechter als der optimale.

Klar ist, dass für $\omega \to 0$ alle Eigenwerte gegen 1 konvergieren, das heißt, der Konvergenzfaktor wird beliebig schlecht.

Zu $\omega > \hat{\omega}$ gehört das konjugiert komplexe Eigenwertpaar

$$\pm\gamma e^{i\cos^{-1}(\pm t^*)}$$

mit geeigneten γ und t^* (vgl. [34, Theorem 4.2]).

Für $\omega > \hat{\omega}$ bleibt der Zähler von (64) eine untere Schranke für den subdominanten Eigenwert, der Nenner hingegen nimmt ab, d.h. der ganze Bruch wird größer, d.h. der Konvergenzfaktor ist schlechter (d.h. größer) als für $\hat{\omega}$.

□

4.4 Optimalität

Ziel dieses Kapitels ist es zu zeigen, dass man — auch im Fall eines komplexen Spektrums von J^p — durch die *extended convergence* keinen besseren Konvergenzfaktor erreichen kann als mit dem klassischen SOR–Verfahren. Für den Spezialfall $\tilde{\sigma}(J^p) \in \mathbb{R}^+$ ist dies in [34] gezeigt, für den Fall $\tilde{\sigma}(J^p) \in \mathbb{R}$ ergibt es sich (vgl. auch [57]) durch analoge Überlegungen. In [34] werden dabei in Teilintervallen das Verhalten des betragsmaximalen zu 1 gehörenden Eigenwertes sowie einer weiteren Größe $\gamma(\omega)$ untersucht.

Im Weiteren lehnt sich die Vorgehensweise eher an die in [19] an, indem direkt die Hypozykloiden betrachtet werden. Die Technik besteht dabei darin, dass die Teilintervalle außerhalb $I := (0, \frac{p}{p-1})$ bijektiv auf die Teilintervalle $(0,1)$ bzw. $(1, \frac{p}{p-1})$ abgebildet werden und dadurch gezeigt werden kann, dass die *extended convergence* bis auf das Intervall $(\hat{\omega}, 0)$, das ja bereits im vorigen Kapitel untersucht ist, gerade auf die klassische Konvergenz zurückgeführt werden kann.

Es wird gezeigt, dass für jedes $\omega \notin I$ entweder ein $\tilde{\omega} \in I$ existiert, das den gleichen Konvergenzfaktor ergibt, oder der Konvergenzfaktor auf jeden Fall schlechter ist als mit dem optimalen Konvergenzfaktor $\omega_0 \in I$. Zusammengenommen ergibt dieses gerade die Behauptung, dass durch erweiterte Konvergenz kein besserer Konvergenzfaktor erreicht werden kann.

Es zeigt sich ferner, dass es einzig von der Lage des klassischen optimalen Relaxationsparameters und von der Tatsache, ob p gerade oder ungerade ist, abhängt, ob bzw. wo es weitere optimale Relaxationsparameter gibt, die denselben Konvergenzfaktor ergeben wie der klassische optimale Relaxationsparameter. Die Frage, wo in der komplexen Ebene das Spektrum von J^p liegt, ist dabei vollkommen irrelevant.

Zentraler Satz für die weitere Vorgehensweise ist dabei

Satz 4.12 *Zwei Hypozykloide $H_1(\omega_1, \eta_1, p, \vartheta), H_2(\omega_2, \eta_2, p, \vartheta)$ mit $\omega_1, \omega_2 > 0, \omega_1 \neq \omega_2$ sind gleich, wenn*

$$\frac{\omega_1 - 1}{\omega_1^p} = \frac{\omega_2 - 1}{\omega_2^p} \tag{65}$$

und

$$\eta_2 = \frac{|\omega_1|\eta_1}{|\omega_2|} \tag{66}$$

gilt. Ist $\omega_1 < 0$, so sind die Hypozykloide $H_1(\omega_1, \eta_1, p, \vartheta + \pi)$ und $H_2(\omega_2, \eta_2, p, \vartheta)$

gleich.

Außerdem sind auch die Innengebiete der Hypozykloiden identisch.

Beweis:
Die Gleichheit der Hypozykloide wird einzeln für den Real– bzw. den Imaginärteil gezeigt. Zunächst zum Realteil $x(\vartheta)$:

Nach Lemma 4.8 ist für $\omega > 0$ für $\vartheta = 0$

$$\frac{1}{\omega\eta}\cos(\vartheta) + \frac{\omega - 1}{\omega}\eta^{p-1}\cos((p-1)\vartheta) > 0,$$

also $x(0) > 0$ und für $\omega < 0$ ist — ebenfalls nach Lemma 4.8 — $x(0) < 0$, wobei jeweils $y(0) = 0$ gilt. Deswegen muss man im Fall, dass $\omega_1 < 0$ ist, das Hypozykloid $H_1(\omega_1, \eta_1, p, \vartheta + \pi)$ (eigentlich allgemeiner $H_1(\omega_1, \eta_1, p, \vartheta + (2k+1)\pi) \quad k \in \mathbb{Z}$) betrachten.

Betrachtet man zunächst

Fall I $\omega_1, \omega_2 > 0$:

$$\frac{1}{\omega_1\eta_1}\cos(\vartheta) + \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1}\cos((p-1)\vartheta) \stackrel{!}{=}$$
$$\frac{1}{\omega_2\eta_2}\cos(\vartheta) + \frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1}\cos((p-1)\vartheta) \qquad \forall\vartheta$$

Dies ist gleichbedeutend mit

$$\underbrace{\left(\frac{1}{\omega_1\eta_1} - \frac{1}{\omega_2\eta_2}\right)}_{=0 \text{ nach (66)}}\cos(\vartheta) = \left(\frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1} - \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1}\right)\cos((p-1)\vartheta)$$

d.h.

$$0 = \left(\frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1} - \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1}\right)\cos((p-1)\vartheta).$$

Division durch $\cos((p-1)\vartheta$ ist zulässig, wenn $\cos((p-1)\vartheta) \neq 0 \Leftrightarrow (p-1)\vartheta \neq (2k+1)\frac{\pi}{2} \Leftrightarrow \vartheta \neq \frac{2k+1}{2(p-1)}\pi$ mit $k \in \mathbb{Z}$, d.h. für alle ϑ bis auf endlich viele.

Für diese ϑ muss dann gelten

$$\frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1} \stackrel{!}{=} \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1},$$

was mit (66) bedeutet

$$\frac{\omega_2 - 1}{\omega_2} \quad \frac{\omega_1^{p-1}\eta_1^{p-1}}{\omega_2^{p-1}} = \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1},$$

also

$$\frac{\omega_2 - 1}{\omega_2^p} = \frac{\omega_1 - 1}{\omega_1^p},$$

also Gleichung (65), die nach Voraussetzung gilt.

Für $\vartheta = \frac{2k+1}{2(p-1)}\pi$ folgt die Gleichheit dann aufgrund der Tatsache, dass die Funktion $\cos(\vartheta)$ stetig ist.

Fall II Sei OBdA $\omega_1 < 0, \omega_2 > 0$:

Nach Bemerkung 4.3 und Lemma 4.8 ist gleichzusetzen

$$\frac{1}{\omega_1\eta_1}\cos(\vartheta + \pi) + \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1}\cos((p-1)(\vartheta + \pi)) \stackrel{!}{=}$$
$$\frac{1}{\omega_2\eta_2}\cos(\vartheta) + \frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1}\cos((p-1)\vartheta) \qquad \forall\vartheta$$

und damit – wegen $\cos(\vartheta + \pi) = -\cos(\vartheta)$, also insbesondere auch $\cos((p-1)(\vartheta + \pi)) = (-1)^{p-1}\cos((p-1)\vartheta)$ —

$$-\frac{1}{\omega_1\eta_1}\cos(\vartheta) + \frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1}(-1)^{p-1}\cos((p-1)\vartheta) \stackrel{!}{=}$$
$$\frac{1}{\omega_2\eta_2}\cos(\vartheta) + \frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1}\cos((p-1)\vartheta) \qquad \forall\vartheta.$$

Dies führt auf

$$\underbrace{\left(\frac{1}{\omega_1\eta_1} + \frac{1}{\omega_2\eta_2}\right)}_{=0 \text{ nach (66)}}\cos(\vartheta) = \left((-1)^{p-1}\frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1} - \frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1}\right)\cos((p-1)\vartheta)$$

und damit

$$\frac{\omega_2 - 1}{\omega_2}\eta_2^{p-1} \stackrel{!}{=} (-1)^{p-1}\frac{\omega_1 - 1}{\omega_1}\eta_1^{p-1}$$

bzw. mit (66)

$$\frac{\omega_2 - 1}{\omega_2} \left(\frac{|\omega_1|\eta_1}{\omega_2} \right)^{p-1} = (-1)^{p-1} \frac{\omega_1 - 1}{\omega_1} \eta_1^{p-1}$$

und damit

$$\frac{\omega_2 - 1}{\omega_2^p} = (-1)^{p-1} \frac{\omega_1 - 1}{\omega_1 |\omega_1|^{p-1}}.$$

Ist p gerade, so wird dies zu ($\omega_1 < 0$!)

$$\frac{\omega_2 - 1}{\omega_2^p} = (-1)^{p-1} \frac{\omega_1 - 1}{-\omega_1^p} = \frac{\omega_1 - 1}{\omega_1^p}$$

und analog für ungerade p

$$\frac{\omega_2 - 1}{\omega_2^p} = \frac{\omega_1 - 1}{\omega_1 |\omega_1|^{p-1}} = \frac{\omega_1 - 1}{\omega_1^p},$$

also in beiden Fällen wiederum zu Gleichung (65).

Für $\vartheta = \frac{2k+1}{2(p-1)}\pi$ folgt die Gleichheit dann wiederum aus Stetigkeitsgründen.

Im zweiten Teil muss noch gezeigt werden, dass nicht nur die Realteile übereinstimmen, sondern auch die Imaginärteile:

Die Argumentation geht dabei vollkommen analog derjenigen für die Realteile, d.h. für den obigen Fall I

$$\underbrace{\left(\frac{1}{\omega_2 \eta_2} - \frac{1}{\omega_1 \eta_1} \right)}_{=0 \text{ nach (66)}} \sin(\vartheta) = \left(\frac{\omega_2 - 1}{\omega_2} \eta_2^{p-1} - \frac{\omega_1 - 1}{\omega_1} \eta_1^{p-1} \right) \sin((p-1)\vartheta)$$

und für Fall II unter Berücksichtigung der Identität $\sin(\vartheta + \pi) = -\sin(\vartheta)$.

Es muss nun noch gezeigt werden, dass nicht nur die Randkurve identisch ist, sondern insbesondere auch das Innengebiet, das durch die Hypozykloiden berandet wird.

Es ist für ω_0, η_0 und $z \in \overline{D}_1$ die Menge $\mathbb{C} \setminus \frac{1-(1-\omega_0)\eta_0^p z^p}{\omega_0 \eta_0 z}$ das Innengebiet des Hypozykloids (vgl. [34, Gl. (39)]).

Gezeigt werden muss, dass auch für die zugeordneten Werte ω_1, η_1 und $z \in \overline{D}_1$ die Menge $\mathbb{C} \setminus \frac{1-(1-\omega_1)\eta_1^p z^p}{\omega_1 \eta_1 z}$ das Innengebiet des Hypozykloids bezeichnet.

Fall I $\omega_0, \omega_1 > 1$

(66) wird dann zu

$$\eta_1 = \frac{\omega_0 \eta_0}{\omega_1},$$

also zu

$$\omega_1 \eta_1 = \omega_0 \eta_0.$$

Betrachtet man

$$q(\omega, \eta) = \frac{1 - (1-\omega)\eta^p z^p}{\omega \eta z},$$

so ist

$$\begin{aligned}
\frac{1-(1-\omega_0)\eta_0^p z^p}{\omega_0 \eta_0 z} &= \frac{1-(1-\omega_0)\eta_0^p z^p}{\omega_1 \eta_1 z} && \text{[mit (66)]} \\
&= \frac{1-\frac{(1-\omega_1)\omega_0^p}{\omega_1^p}\eta_0^p z^p}{\omega_1 \eta_1 z} && \text{[mit (65)]} \\
&= \frac{1-\frac{(1-\omega_1)\omega_0^p}{\omega_1^p}\frac{\omega_1^p \eta_1^p}{\omega_0^p} z^p}{\omega_1 \eta_1 z} && \text{[mit (66)]} \\
&= \frac{1-(1-\omega_1)\eta_1^p z^p}{\omega_1 \eta_1 z},
\end{aligned}$$

d.h. die Außengebiete der Hypozykloiden sind identisch (und damit insbesondere auch die Innengebiete).

Fall II $\omega_0 > 1, \omega_1 < 0$, p ungerade

(der Fall, dass p gerade ist, ist irrelevant, weil dann die Gleichung (65) nicht lösbar ist):

In diesem Fall ist $\eta_1 = -\frac{\omega_0 \eta_0}{\omega_1}$ und es folgt

$$\begin{aligned}
\frac{1-(1-\omega_0)\eta_0^p z^p}{\omega_0 \eta_0 z} &= \frac{1-(1-\omega_0)\eta_0^p z^p}{-\omega_1 \eta_1 z} \\
&= \frac{1-(1-\frac{(1-\omega_1)\omega_0^p}{\omega_1^p}\eta_0^p z^p}{-\omega_1 \eta_1 z} \\
&= \frac{1-\frac{(1-\omega_1)\omega_0^p}{\omega_1^p}\frac{(-\omega_1)^p \eta_1^p}{\omega_0^p} z^p}{-\omega_1 \eta_1 z} \\
&= \frac{1+(1-\omega_1)\eta_1^p z^p}{-\omega_1 \eta_1 z}
\end{aligned}$$

$$= \frac{1-(1-\omega_1)\eta_1^p \hat{z}^p}{\omega_1 \eta_1 \hat{z}},$$

wenn $z = \rho e^{i\varphi}$, $\hat{z} = \rho e^{i(\varphi-\pi)} e^{i\pi} = -\rho e^{i(\varphi-pi)}$ gesetzt wird (unter Beachtung der Tatsache, dass p ungerade ist). Es ergibt sich also wieder genau die Drehung des Hypozykloids, die bereits in Satz 4.12 erwähnt wurde.

Fall III $0 < \omega_0 < 1, \omega_1 < 0$, p gerade:

Dieser Fall lässt sich vollkommen analog zu Fall II zeigen. Auch hierbei tritt wiederum die Drehung auf.

übrige Fälle: In allen anderen Fällen ist das aus (65) und (66) bestehende System nicht lösbar, die Fälle sind also irrelevant.

□

Die Tatsache, dass zwei Hypozykloiden gleich sind, ist an und für sich nicht besonders interessant; der nächste Satz zeigt allerdings, dass in diesem Fall auch der Konvergenzfaktor der beiden Verfahren identisch ist.

Der Beweis dieses Satzes verwendet wesentlich die Darstellung der Eigenwerte von $\mathcal{L}_\omega$ als p-te Potenz der Nullstellen von (14), sowie die Darstellung von $\alpha(\omega) := \max\{|\lambda| : \lambda^p \text{ Eigenwert von } \mathcal{L}_\omega \text{ zum Eigenwert 1 von } J\}$ aus Definition 38.

Für $\omega \notin (0, \frac{p}{p-1})$ muss berücksichtigt werden, dass — aus Sicht der Vektoriteration — nicht mehr 1 dominanter Eigenwert ist sondern eine Zahl, deren Betrag größer als 1 ist, α^p bezeichnet den Betrag dieser Zahl.

Satz 4.13 *Sei $\sigma(J^p) \subseteq \mathbb{C}$ gegeben, $\omega_1 \in (0, \frac{p}{p-1})$ beliebig, $\frac{1}{\eta_1^p}$ der zugehörige Konvergenzfaktor. Durch Satz 4.12 — speziell durch die Formeln (65) und (66) — seien die Parameter eines deckungsgleichen Hypozykloids $H(\omega_2, \eta_2, p, .)$ gegeben. Dann ist mit $\alpha(\omega) = \max\{|\lambda| : \lambda^p$ Eigenwert von $\mathcal{L}_\omega$ zum Eigenwert 1 von $J\}$ (vgl. auch Definition 38)*

$$\left(\frac{1}{\eta_2 \alpha(\omega_2)}\right)^p$$

der Konvergenzfaktor des zugehörigen SOR–Verfahrens bezüglich ω_2 und es gilt

$$\left(\frac{1}{\eta_2 \alpha(\omega_2)}\right)^p = \frac{1}{\eta_1^p}, \tag{67}$$

d.h. beide Verfahren besitzen denselben Konvergenzfaktor.

Beweis:
Es ist (nach Bemerkung 3.1)

$$\eta_2 = \left(\frac{1}{|\lambda_{\text{subdominant }}(\omega_2)|} \right)^{\frac{1}{p}}.$$

Damit gilt

$$\begin{aligned} \left(\frac{1}{\eta_2 \alpha(\omega_2)} \right)^p &= \frac{|\lambda_{\text{subdominant}}(\omega_2)|}{\alpha^p(\omega_2)} \\ &= \frac{|\lambda_{\text{subdominant}}(\omega_2)|}{|\lambda_{\text{dominant}}(\omega_2)|}, \end{aligned}$$

was nach der Theorie der Vektoriteration den Konvergenzfaktor bezeichnet. Es ist also nur Gleichung (67) zu zeigen.

Das zu $\alpha(\omega)$ gehörende λ werde als $\lambda_{\max}$ bezeichnet.

Zu zeigen ist also

$$\frac{1}{\eta_2 \alpha(\omega_2)} = \frac{1}{\eta_1}$$

bzw.

$$\alpha(\omega_2) = \frac{\eta_1}{\eta_2}.$$

Sei $\omega_1 > 1$. Für $\omega_2 > \frac{p}{p-1}$ ist nach [34, Theorem 4.1] der betragsgrößte zu 1 gehörende Eigenwert positiv.

Hier muss also gelten

$$\begin{aligned} f_{\omega_2}(\alpha(\omega_2)) &= \left(\frac{\eta_1}{\eta_2} \right)^p - \omega_2 \left(\frac{\eta_1}{\eta_2} \right)^{p-1} - (1 - \omega_2) \\ &= \left(\frac{\omega_2}{\omega_1} \right)^p - \omega_2 \left(\frac{\omega_2}{\omega_1} \right)^{p-1} - (1 - \omega_2) \\ &= \left(\frac{\omega_2}{\omega_1} \right)^p (1 - \omega_1) - (1 - \omega_2). \end{aligned}$$

Daraus folgt, dass $f_{\omega_2}(\alpha(\omega_2)) = 0$ ist, wenn

$$\frac{\omega_2^p}{\omega_1^p} = \frac{1 - \omega_2}{1 - \omega_1}, \tag{68}$$

d.h. wenn (65) gilt.

Betrachtet man hier die Nullstellen von

$$\begin{aligned} f_\omega(\lambda) &= \lambda^p - \omega\lambda^{p-1} - (1-\omega) \\ &= (\lambda-1)\underbrace{\left(\lambda^{p-1} + (1-\omega)\sum_{i=0}^{p-2}\lambda^i\right)}_{=:\tilde{f}_\omega(\lambda)}, \end{aligned}$$

so kann man sich auf die Nullstellen von $\tilde{f}_\omega(\lambda)$ beschränken.

Nach [27, Theorem 6.4l] ist die positive Lösung σ von

$$\sigma^{p-1} = |a_0| + |a_1|\sigma + \ldots + |a_{p-2}|\sigma^{p-2}$$

Inklusionsradius für $\tilde{f}_{\omega_2}(\lambda)$. Wegen $\omega > 1$ ist dieses σ nichts anderes als die positive Nullstelle von $\tilde{f}_{\omega_2}(\lambda)$, d.h. alle weiteren Nullstellen liegen in einem Kreis um 0 mit Radius $\frac{\eta_1}{\eta_2}$, womit die Behauptung für diesen Fall vollständig bewiesen ist[18].

Sei nun $\omega_2 < 0$ und p ungerade (für gerade p ist nichts zu zeigen, weil dann (65) nicht lösbar ist.

Hier ist nach [34, Theorem 4.2] der betragsgrößte zu 1 gehörende Eigenwert negativ. Setzt man $-\frac{\eta_1}{\eta_2}$ in $f_{\omega_2}(\lambda)$ ein, so ergibt sich, weil in diesem Fall $\eta_2 = -\frac{\omega_1\eta_1}{\omega_2}$ gilt, dass auch hier eine Nullstelle von f vorliegt, wenn (65) erfüllt ist.

Nach Satz 3.16 und den Bemerkungen 3.4 und 3.5 sind $p-1$ zu 1 gehörende Eigenwerte kleiner als der größte zu $|\mu| < 1$ gehörende. Insbesondere genügt es also zu zeigen, dass

$$\frac{\eta_1^p}{\eta_2^p} > 1 - \omega_2$$

erfüllt ist.

Wegen (66) ist dies gleichbedeutend mit

$$\frac{|\omega_2|^p}{1-\omega_2} > \omega_1^p,$$

[18] Die Behauptung folgt auch sofort, wenn man Korollar 3.14 zum Beweis heranzieht. Wegen $\frac{\eta_1}{\eta_2} > 1$ liegen — neben dem Eigenwert, der exakt 1 ist — alle anderen Lösungen der Gleichung im Innern des Einheitskreises.

was wegen $\omega_2 < \hat{\omega}$ und $\omega_2 < \frac{p}{p-1}$ offenbar erfüllt ist.

Damit ist auch für diesen Fall gezeigt, dass $\frac{\eta_1}{\eta_2}$ wie im obigen Fall Inklusionsradius ist.

Im Fall $\omega_1 < 1$ folgt für gerade p und $\omega_2 < 0$ analog wie der eben durchgeführte Fall. Für gerade p ist nichts zu beweisen.

Damit ist der Satz vollständig gezeigt. □

Insbesondere gilt das zentrale Ergebnis

Satz 4.14 *Ein besserer Konvergenzfaktor als beim SOR–Verfahren mit klassischem optimalen Relaxationsparameter kann nicht erreicht werden.*

Beweis:
Die Intervalle $(\frac{p}{p-1}, \infty)$ und $(-\infty, \hat{\omega})$ können nach Satz 4.12 und 4.13 bijektiv auf $(0,1)$ oder $(1, \frac{p}{p-1})$ abgebildet werden, wobei jeweils gleiche Konvergenzfaktoren der durch (65) gekoppelten Relaxationsparameter erreicht werden.

Angenommen es gäbe einen Parameter $\tilde{\omega}$, der bessere Konvergenz als ω_0 bewirken würde. Dann wäre diesem durch (65) ein $\omega_0 \neq \omega' \in (0, \frac{p}{p-1})$ zugeordnet, das ebenfalls bessere Konvergenz als ω_0 erzielen würde – im Wiederspruch zur Optimalität von ω_0.

Dass in $(\hat{\omega}, 0)$ kein besserer Konvergenzfaktor erreicht werden kann, ist zusätzlich bereits in Lemma 4.11 gezeigt worden. □

Bemerkung 4.5 *Aus Satz 4.13 und 4.12 folgt nun aber insbesondere sofort, dass einzig die Lösbarkeit von Gleichung (65) Aussagen darüber gibt, ob ein weiterer Relaxationsparameter denselben Konvergenzfaktor erreicht wie ein gegebener Relaxationsparameter ω.*

Für die weiteren Betrachtungen gilt nach den angegebenen Stellen in der Literatur die folgende Generalvoraussetzung.

Generalvoraussetzung:

Für $\omega \in (0, \frac{p}{p-1})$ ist in den Fällen

- $\sigma(J^p) \in \mathbb{R}^+$ (vgl. [65])
- $\sigma(J^p) \in \mathbb{R}$ (vgl. [11])
- $\sigma(J^p) \in \mathbb{C}$ – Einpunktproblem (vgl. [18])
- $\sigma(J^p) \in \mathbb{C}$ – Mehrpunktproblem (vgl. [19])

Existenz und Lage des optimalen Relaxationsparameters ω_0 sowie das zugehörige η_0 bekannt.

Die zugehörigen Sätze sind in den Abschnitten 3.3 und 4.2 zitiert.

Zur Bezeichnungsweise ist anzumerken, dass — analog zu den Bezeichnungsweisen in [34] — ω_0 jeweils den optimalen Relaxationsparameter in $(0, \frac{p}{p-1})$ bezeichnet, ω_+ den entsprechenden für $\omega > \frac{p}{p-1}$ und ω_- denjenigen für $\omega < 0$.

Satz 4.15 *Für $\omega_0 \in (0,1)$ existiert für gerade p ein ω_-, das denselben Konvergenzfaktor ergibt; es existiert kein analoges ω_+. Ist p ungerade, so existiert neben ω_0 kein weiterer optimaler Relaxationsparameter.*

Beweis:
Nach Satz 4.12 ist

$$0 > \frac{\omega_0 - 1}{\omega_0^p} \stackrel{!}{=} \frac{\omega - 1}{\omega^p} =: g(\omega) \tag{69}$$

zu lösen. Genau dann existiert nämlich ein ω, was dasselbe Hypozykloid erzeugt, d.h. insbesondere den gleichen Konvergenzfaktor.

Sei $\omega > \frac{p}{p-1}$: dann ist $g(\omega) > 0$ und die Gleichung (69) wegen $g(\omega_0) < 0$ unlösbar. Andererseits existiert aber zu jedem $\omega \in (\frac{p}{p-1}, \infty)$ ein $\tilde{\omega} \in (1, \frac{p}{p-1})$, welches dasselbe Hypozykloid erzeugt (vgl. Lemma 4.9). Dieses führt nach Voraussetzung zu schlechterer Konvergenz als $H_0(\omega_0, \eta_0, p, \vartheta)$.

Sei nun $\omega < 0$, d.h. $\omega - 1 < 0$. Dann ist (69) genau für gerade p (d.h. $\omega^p > 0$) lösbar, nicht aber für ungerade p.

Ist p gerade, so existiert zu jedem $\omega \in (-\infty, 0)$ ein $\tilde{\omega} \in (0,1)$, das dasselbe

Hypozykloid erzeugt. Umgekehrt existiert also ein $\omega_- \in (-\infty, 0)$, das dasselbe Hypozykloid erzeugt wie ω_0, d.h. optimal ist.

Ist p ungerade, so ist (69) unlösbar. Zu jedem $\omega \in (\frac{p}{p-1}, \infty)$ existiert aber ein $\tilde{\omega} \in (1, \frac{p}{p-1})$, das dasselbe Hypozykloid erzeugt. Dieses gehört nach Voraussetzung zu einem Relaxationsparameter, der schlechtere Konvergenz als ω_0 bewirkt, d.h. es existiert kein ω_+.

Mit analoger Argumentationsweise erhält man, dass für $\omega \in (-\infty, \hat{\omega})$ ($\hat{\omega}$ analog zu $\hat{x}$ aus Lemma 4.10) kein optimaler Relaxationsparameter existieren kann. Was zu zeigen bleibt, ist die Tatsache, dass auch für $\omega \in (\hat{\omega}, 0)$ kein optimaler Relaxationsparameter existiert.

Sei also nun $0 > \omega > \hat{\omega}$. Nach Lemma 4.11 liegt in diesem Fall ein gestrecktes Hypozykloid vor, dessen innere Schnittpunkte mit dem Einheitskreis die Einheitswurzeln sind, d.h. insbesondere dass eine der Halbachsen echt größer als 1 ist, das Hypozykloid ist also schlechter als das optimale, welches nur Halbachsen kleiner als 1 besitzt. □

Der folgende Satz deckt den zweiten Fall, nämlich $\omega_0 \in (1, \frac{p}{p-1})$ ab.

Satz 4.16 *Ist $\omega_0 \in (1, \frac{p}{p-1})$, so existieren für ungerade p sowohl ein ω_- als auch ein ω_+, für gerade p existiert nur ein ω_+. Diese ω_- bzw. ω_+ erzeugen — soweit existent — dasselbe Hypozykloid und damit denselben Konvergenzfaktor wie ω_0.*

Beweis:
Zu lösen ist wieder

$$0 < \frac{\omega_0 - 1}{\omega_0^p} \stackrel{!}{=} \frac{\omega - 1}{\omega^p} =: g(\omega).$$

Nach Lemma 4.9 existiert in $(\frac{p}{p-1}, \infty)$ eine Lösung, also ein ω_+ unabhängig von p. Zugleich wird das Intervall $(1, \frac{p}{p-1})$ bijektiv auf das Intervall $(\frac{p}{p-1}, \infty)$ abgebildet, sodass dieses ω_+ der einzige optimale Relaxationsparameter in $(\frac{p}{p-1}, \infty)$ ist.

Ist p gerade, so wird $(-\infty, 0)$ bijektiv auf $(0, 1)$ abgebildet, das heißt auf den Bereich, in dem die Parameter liegen, die zu schlechterer Konvergenz als ω_0 führen. In diesem Fall gibt es also keinen weiteren optimalen Relaxationsparameter.

Ist dagegen p ungerade, so existiert — ebenfalls mit Lemma 4.9 — ein eindeutiges ω_-, und aufgrund der Bijektivität kann es im Intervall $(-\infty, \hat{x})$ kein weiteres

optimales ω geben. Zu untersuchen bleibt das Intervall $[\hat{x}, 0)$. Hier existiert aber nach Lemma 4.11 und den Bemerkungen im Beweis zu Satz 4.15 ebenfalls kein optimaler Relaxationsparameter. □

Korollar 4.17 *Sei $\sigma(J^p) \in [-\alpha^p, \beta^p] \subseteq \mathbb{R}$ mit $\alpha, \beta \in \mathbb{R}_0^+$. Dann gilt für die Verteilung der optimalen Relaxationsparameter folgende Tabelle:*

	$\alpha < \beta$		$\alpha > \beta$	
$\omega \in$	p gerade	p ungerade	p gerade	p ungerade
$(0,1)$	—		ω_0	
$(1, \frac{p}{p-1})$	ω_0		—	
$(\frac{p}{p-1}, \infty)$	ω_+	ω_+	—	—
$(-\infty, 0)$	—	ω_-	ω_-	—

Beweis:
Gemäß [11] gilt für $\alpha > \beta$, dass $\omega_0 \in (0,1)$ ist, für $\beta > \alpha$, dass $\omega_0 \in (1, \frac{p}{p-1})$ ist. Die Behauptung folgt dann unmittelbar aus der Anwendung von Satz 4.15 und 4.16. □

Bemerkung 4.6 *Dieselbe Tabelle ließe sich für $\sigma(J^p) \in \mathbb{C}$ auch aufstellen, indem man — ohne weitere Voraussetzungen an das Spektrum von J^p — unterscheidet, ob $\omega_0 \in (0,1)$ oder $\omega_0 \in (1, \frac{p}{p-1})$ liegt. Alle anderen Einträge wären genauso verteilt.*

Bemerkung 4.7 *Das ω_- aus Satz 4.16 ist dasselbe, was im Fall, dass $\sigma(J^p) \in \mathbb{R}^+$ ist, in [34] bestimmt wird:*

nach Satz 4.15 gilt

$$\frac{\omega - 1}{\omega^p} = \frac{\omega_0 - 1}{\omega_0^p}$$

und (nach [65])

$$(p-1)^{p-1}\mu^p\omega_0^p - p^p(\omega_0 - 1) = 0$$

d.h.

$$\underbrace{\frac{p^p}{(p-1)^{p-1}\mu^p}}_{=:C,\quad \text{unabh. von } \omega} \qquad \frac{\omega_0 - 1}{\omega_0^p} = C\,\frac{\omega_0 - 1}{\omega_0^p} = C\,\frac{\omega - 1}{\omega^p},$$

d.h. auch ω ist Nullstelle der Bestimmungsgleichung, d.h. insbesondere dass $\omega = \omega_-$ aus [34] ist. Gleiches gilt analog für ω_+.

Zusammenfassung:

Für gerade p gibt es zu jedem $\omega \in (-\infty, 0)$ ein $\tilde{\omega} \in (0, 1)$ mit demselben Konvergenzfaktor. $\tilde{\omega}$ kann gemäß (65) und (66) berechnet werden, falls ω bekannt ist, und umgekehrt.

Für ungerade p gibt es zu jedem $\omega \in (-\infty, \hat{\omega})$ ein $\tilde{\omega} \in (1, \frac{p}{p-1})$, das denselben Konvergenzfaktor erzeugt. Für $\omega \in (\hat{\omega}, 0)$ ist das SOR–Verfahren schlechter als das entsprechende SOR–Verfahren mit dem optimalen ω_0.

Für $\omega \in (\frac{p}{p-1}, \infty)$ existiert ein $\tilde{\omega} \in (1, \frac{p}{p-1})$, das denselben Konvergenzfaktor erzeugt.

Damit ist insbesondere gezeigt, dass es in Abhängigkeit der Lage von ω_0 und von der Tatsache, ob p gerade oder ungerade ist, außer den durch Satz 4.12 optimalen Relaxationsfaktoren keine weiteren optimalen gibt. Dies gilt unabhängig von der Art des Spektrums von J^p, d.h. insbesondere auch für $\sigma(J^p) \subseteq \mathbb{C}$.

4.5 Lage bei unbekanntem Spektrum der Jacobi–Matrix II

Einpunktproblem

Dieser Fall ist erheblich komplizierter als die in Abschnitt 3.6 behandelten Fälle, weil die Formeln, die auch in [18] angegeben sind, nur schwer zu handhaben sind.

Zunächst soll untersucht werden, wann das Gauß–Seidel–Verfahren das optimale Verfahren ist, d.h. wann man durch Relaxation nur verschlechtert.

Wann ist Gauß–Seidel optimal ?

Gauß–Seidel ist optimal, wenn das Hypozykloid zu einem Kreis entartet, d.h. wenn

$$a = b$$

gilt. Mit (56) und (57) (vgl. auch [18, Lemma 2.1]) wird dies zu

$$\frac{r}{\cos(p\vartheta/2)} \cos\left[\left(\frac{p}{2}-1\right)\vartheta - \psi\right] = \frac{r}{\sin(p\vartheta/2)} \sin\left[\left(\frac{p}{2}-1\right)\vartheta - \psi\right],$$

d.h.

$$\frac{\sin(p\vartheta/2)}{\cos(p\vartheta/2)} = \frac{\sin\left[\left(\frac{p}{2}-1\right)\vartheta - \psi\right)}{\cos\left[\left(\frac{p}{2}-1\right)\vartheta - \psi\right)},$$

also

$$\tan(p\vartheta/2) = \tan\left(\frac{p\vartheta}{2} - \vartheta - \psi\right).$$

Wegen $\vartheta \in (0, \frac{\pi}{p})$ und $-\psi \in (0, \frac{\pi}{p})$ führt das insgesamt auf

$$a = b = r \qquad \vartheta = -\psi. \tag{70}$$

Setzt man dies in die Bestimmungsgleichung (58) aus Satz 4.2 ein, so erhält man

$$\left(r\frac{\sin(-p\psi)}{\sin(-p\psi)}\right)^p = \frac{\sin(-p\psi)\cos(-p\psi)}{\sin(-p\psi)} = \cos(-p\psi),$$

also

$$r^p = \cos(-p\psi) = \rho^p(\mathcal{L}_\omega)$$

bzw.

$$\psi = -\frac{\arccos(r^p)}{p}. \tag{71}$$

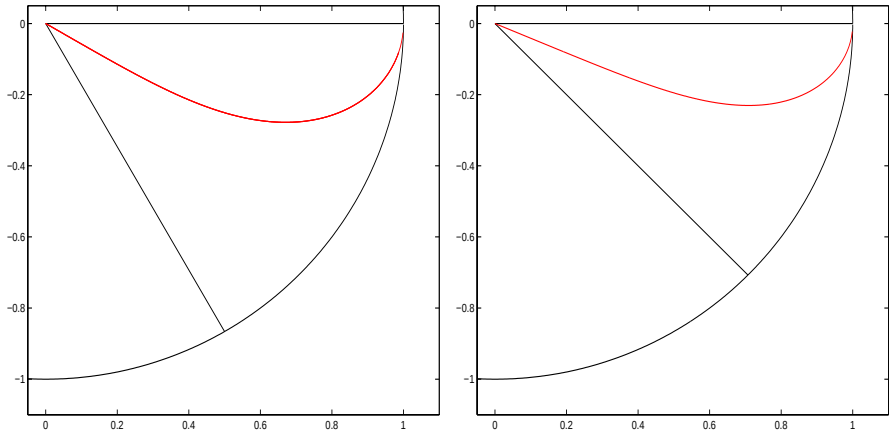

Abbildung 25: (71) für $p = 3$ und $p = 4$

Abbildung 25 zeigt die Kurve (71) für die Fälle $p = 3$ und $p = 4$.

Für $r = 0$ ist $\psi = -\frac{\pi}{2p}$, für $r = 1$ ist $\psi = 0$. Liegt die komplexe Zahl oberhalb dieser Kurve, muss überrelaxiert werden, liegt sie unterhalb dieser Kurve, sollte man unterrelaxieren. Das folgt sofort aus der Eindeutigkeit von (71), wenn man die Grenzfälle $\tilde{\sigma}(J^p) \in \mathbb{R}^+$ bzw. $\tilde{\sigma}(J^p) \in \mathbb{R}^-$ berücksichtigt.

Nach Satz 3.17 ist die Abschätzung für $\tilde{\sigma}(J^p) \in \mathbb{R}^-$ die gröbere, so dass man folgenden Satz formulieren kann:

Satz 4.18 *Betrachtet man das Einpunktproblem, das heißt neben den Eigenwerten auf dem Einheitskreis liegt ein (komplexer) Eigenwert innerhalb des durch den Einheitskreis, die reelle Achse und den Strahl $-\pi/p$ gegebenen Gebietes.*

Das Gauß–Seidel–Verfahren ist dann höchstens um den Faktor

$$\left(\frac{p}{p-2}\right)^p \tag{72}$$

schlechter als das SOR–Verfahren mit dem optimalen Parameter ω^*.

Beweis:
Es muss gezeigt werden, dass der Quotient

$$\left| \frac{\rho(\mathcal{L}_1)}{\rho(\mathcal{L}_{\omega^*})} \right| \tag{73}$$

in den beiden Teilgebieten oberhalb und unterhalb der durch (71) gegebenen Kurve größer oder gleich dem Quotienten in den beiden Grenzfällen ist bzw. dass $\rho(\mathcal{L}_{\omega^*})$ größer ist als im Fall, dass J^p nichtnegative bzw. nichtpositive Eigenwerte hat. Dazu betrachtet man jeweils die Hypozykloiden, die ein konvergentes Verfahren beschreiben.

$a > b$: Dies bedeutet Unterrelaxation. Betrachtet man zunächst den Grenzfall, dass — außer dem p–fachen Eigenwert 1 — J^p nichtpositive Eigenwerte hat. Insbesondere hat J einen Eigenwert $-\alpha \in \mathbb{R}^-$, der maximalen Betrag der Eigenwerte auf der negativen reellen Achse hat.

Es gilt für das zugehörige optimale Hypozykloid, dass

$$a = \alpha = \frac{r}{\sin(p\pi/2p)} \sin\left[\left(\frac{p}{2} - 1\right) \frac{p}{2} \frac{\pi}{p} + \frac{\pi}{p}\right],$$

die imaginäre Halbachse also durch diesen subdominanten Eigenwert bestimmt wird. Ferner ist (nach [18, Lemma 2.1])

$$\begin{aligned} b &= \frac{r}{\cos(p\vartheta/2)} \cos\left[\left(\frac{p}{2} - 1\right) \vartheta\right] \\ &= \frac{r \cos\left(\frac{p}{2}\frac{\pi}{p}\right)}{\cos\left(\frac{p}{2}\frac{\pi}{p}\right)} \\ &= \frac{-r\left(\frac{p}{2} - 1\right) \sin\left(\frac{p}{2}\frac{\pi}{p}\right)}{-\frac{p}{2}\sin\left(\frac{p}{2}\frac{\pi}{p}\right)} \\ &= r\left(\frac{p}{2} - 1\right)\frac{2}{p} = r - \frac{2r}{p} \\ &= \frac{p-2}{p} r. \end{aligned}$$

Insbesondere gilt damit

$$\frac{a}{b} = \frac{p}{p-2}.$$

Gemäß [18, Lemma 4.1] gilt für verkürzte Hypozykloide (und nur solche kommen überhaupt in Betracht)

$$\frac{a}{b} \in \left(\frac{p-2}{p}, \frac{p}{p-2} \right). \tag{74}$$

Sicherlich ist $a \geq r$ für ein konvergentes Verfahren[19] und wegen (74) ist damit

$$b \geq \frac{p-2}{p} r.$$

Nach [18, Lemma 4.4] gilt aber für den Spektralradius der SOR–Matrix

$$\rho(\mathcal{L}_\omega) \geq b^p,$$

falls $b < a$ ist. Damit gilt insgesamt

$$\rho^{1/p}(\mathcal{L}_\omega) \geq b \geq \frac{p-2}{p} r,$$

d.h. wir haben höchstens genauso schnelle Konvergenz. Insbesondere kann damit die Abschätzung (72) für nichtpositive Eigenwerte verwenden.

$a < b$: Dieser Fall folgt vollkommen analog, wenn man a und b vertauscht, bzw. wenn man berücksichtigt, dass in diesem Fall in (57) und (56) $\vartheta = 0$ zu setzen ist.

□

Bemerkung 4.8 *Die Abschätzung (72) ist eine echte, nicht sehr scharfe obere Schranke. In der Praxis ist der Quotient (73) meist sehr viel kleiner.*

Im Fall, dass für das optimale SOR–Verfahren im Verhältnis der Halbachsen der Hypozykloiden $a > b$ gilt, lässt sich auch folgender Satz formulieren, wenn $\mu = \alpha + i\beta$ ($\alpha > 0, \beta < 0$) die komplexe Zahl im Segment $\mathcal{S}$ ist:

Satz 4.19 *Sei $p > 2$. Für das Einpunktproblem mit $\mu^p \in \mathbb{C}$ ist, falls für den optimalen Relaxationsparameter gilt, dass $a > b$ (dieses sind die Halbachsen des*

[19] Der fragliche Punkt muss ja im Innern oder auf dem Rand des Hypozykloids liegen, was mit $a < r$ nicht zu erfüllen ist.

zugehörigen optimalen Hypozykloids) ist, das Gauß– Seidel–Verfahren höchstens

$$\left|\frac{\mu^p}{Re(\mu)^p \sin^p(\frac{\pi}{p})}\right|$$

schlechter als das SOR–Verfahren mit optimalem ω.

Beweis:
Die zitierten Formeln stammen ausnahmslos aus [18]. Insbesondere kann, weil die Figur bei Drehungen um den Winkel $\frac{2\pi}{p}$ konstant bleibt, $\vartheta \in (0, \frac{\pi}{p})$ vorausgesetzt werden.

Es ist

$$\begin{aligned}
\left|\frac{\mu^p}{x_0^p}\right| &\leq \left|\frac{\mu^p}{b^p}\right| && \text{(nach Lemma 4.4 ist } x_0 = \rho^{1/p}(\mathcal{L}_\omega) > b) \\
&= \left|\frac{\mu^p \cos^p(\vartheta\frac{p}{2})}{(\alpha\cos(\frac{p-2}{2}\vartheta) + \beta\sin(\frac{p-2}{2}\vartheta))^p}\right| && \text{(nach Lemma 2.1)} \\
&\leq \left|\frac{\mu^p}{(\alpha\cos(\frac{p-2}{2}\vartheta) + \beta\sin(\frac{p-2}{2}\vartheta))^p}\right| \\
&\leq \left|\frac{\mu^p}{\alpha^p\cos^p(\frac{p-2}{2}\vartheta)}\right| \\
&\leq \left|\frac{\mu^p}{\alpha^p\cos^p(\frac{p-2}{2}\frac{\pi}{p})}\right| && \text{(weil der Cosinus in } (0, \tfrac{\pi}{p}) \text{ fällt)} \\
&= \left|\frac{\mu^p}{\alpha^p\sin^p(\frac{\pi}{p})}\right|,
\end{aligned}$$

wobei die letzte Gleichung wegen der Gleichheit $\cos(\frac{p-2}{p}\frac{\pi}{2}) = \cos(\frac{\pi}{2} - \frac{\pi}{p}) = -\sin(-\frac{\pi}{p}) = \sin(\frac{\pi}{p})$ gilt. □

Bemerkung 4.9 *Es ist klar, dass diese Abschätzung für größere Werte von p wertlos ist, da dann* $\sin(\frac{\pi}{p})$ *klein wird, in diesem Fall ist die Abschätzung aus Satz 4.18 besser.*

Mehrpunktproblem

Aufgrund der Argumentation im letzten Unterabschnitt und weil ein Hypozykloid, das bezüglich zwei Punkten optimal ist, kein schnelleres Verfahren bezeichnen kann als eines, das nur bezüglich einem Punkt optimal ist (vgl. die entsprechenden Überlegungen in Kapitel 4.2), gilt auch in diesem Fall Abschätzung (72), wobei in diesem Fall der wirkliche Quotient (73) im Allgemeinen noch kleiner ist als im vorherigen Fall.

Zusammenfassend kann man feststellen:

Satz 4.20 *Führt man statt dem SOR–Verfahren mit optimalem Relaxationsfaktor einfach das Gauß–Seidel–Verfahren durch, so ist das Gauß–Seidel–Verfahren höchstens um den Faktor*

$$\left(\frac{p}{p-2}\right)^p$$

schlechter als SOR mit optimalem Relaxationsparameter.

Optimalität bei unbekanntem Spektrum — Eigenwerte der zu einer Markov–Kette gehörenden Matrix

In den vorigen Kapiteln werden immer nur Verfahren der Art

$$\mathcal{E}_p: \qquad y^{(k)} = \mu_0(Ty^{(k-1)} + c) + \mu_1 y^{(k-1)} + \ldots + \mu_p y^{(k-p)} \tag{75}$$

mit $k > p, y^{(0)}, \ldots, y^{(p-1)} \in \mathbb{R}^n, \mu_0 \neq 0$ und $\sum_{j=0}^{p} \mu_j = 1$ bei der Suche nach dem optimalen Verfahren berücksichtigt.

Der Klasse $\mathcal{E}_p$ sind dabei Funktionen

$$f(\varphi) = \frac{\mu_0 \varphi}{1 - \mu_1\varphi - \mu_2\varphi^2 - \ldots - \mu_p\varphi^p} \tag{76}$$

zugeordnet und man spricht auch davon, dass f zu $\mathcal{E}_p$ gehört ($f \in \mathcal{E}_p$).

Die Klasse $\mathcal{E}_p$ ist dabei nur eine Teilmenge der größeren Klasse der Euler–Funktionen (vgl. auch [46]):

Definition 4.1 *Eine (komplexwertige) Funktion g heißt* Euler–Funktion *($g \in \mathcal{E}$), wenn eine offene Umgebung U des abgeschlossenen Einheitskreises existiert, so dass g in U meromorph und schlicht ist und ferner $g(0) = 0$ und $g(1) = 1$ erfüllt ist.*

Bemerkung 4.10 *Die Funktionen der Klasse $\mathcal{E}_p$ sind damit offenbar Eulerfunktionen, die Bedingung $f(1) = 1$ ist wegen $\sum_{j=0}^{p} \mu_j = 1$ erfüllt, insbesondere ist mit R nach (23) und (26) auch jeweis $\frac{1}{R}$ Euler–Funktionen.*

Als Frage stellt sich, welches Verfahren aus der Klasse $\mathcal{E}$ asymptotisch optimal ist, wenn das Spektrum der Jacobi–Matrix bekannt ist. Dass es ein solches Verfahren geben muss, folgt aus [46, Theorem 8]:

Mit

Definition 4.2 *Es sei $\mathcal{M}$ die Menge aller kompakten Mengen U, die mehr als einen Punkt enthalten, deren Komplement einfach zusammenhängend ist und die 1 enthält.*

gilt nämlich

Satz 4.21 *Für jedes $U \in \mathcal{M}$ existiert eine Eulerfunktion, die optimal ist bezogen auf U.*

Im Fall, dass $p = 2$ ist und das Spektrum von J^2 reell ist, ist das entsprechende SOR–Verfahren mit optimalem Relaxationsparameter auch das optimale Euler–Verfahren.

Das Spektrum von J ist in diesem Fall — abgesehen von den Eigenwerten 1 und -1 — enthalten in dem Intervall $[-\mu, \mu]$ und nach Satz 2.10 gilt für den optimalen Relaxationsparameter

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \mu^2}}$$

und der entsprechende Konvergenzfaktor ist

$$\frac{1 - \sqrt{1 - \mu^2}}{1 + \sqrt{1 - \mu^2}} = \frac{\mu^2}{(1 + \sqrt{1 - \mu^2})^2}. \tag{77}$$

Gemäß [46, §9, Example 1] ist der Konvergenzfaktor des asymptotisch optimalen Euler–Verfahrens gerade

$$\frac{\mu}{1 + \sqrt{1 - \mu^2}},$$

also gerade die Quadratwurzel von (77), was auch aus Satz 3.3 folgt. Das SOR–Verfahren mit ω_0 ist also genau das asymptotisch optimale Euler–Verfahren.

[46, §9, Example 4] zeigt, dass man ähnliches im Fall, dass $p = 4$ bei nichtnegativen Eigenwerten von J^p ist, durchführen kann.

Im allgemeinen Fall ist aber unklar, welches das asymptotisch optimale Euler–Verfahren ist. Diese Problematik wird verstärkt dadurch, dass man im Allgemeinen bei einer gegebenen Markov–Kette keine Aussagen über das Spektrum von J bzw. J^p vorliegen hat.

Aus diesem Grund stellt sich die Frage, ob es allgemeinere Aussagen darüber gibt, welche Eigenwerte ein p–zyklischer Generator oder eine p–zyklische stochastische Matrix hat.

Was kann man über die Eigenwerte der zu einer Markov–Kette gehörenden Matrix sagen ?

Geht man von einer zeitdiskreten Markov–Kette aus, so erhält man eine schwach zyklische stochastische Matrix mit Index p, in der die Übergangswahrscheinlichkeiten stehen.

Aber selbst in diesem (verglichen mit dem zeitkontinuierlichen) einfacheren Fall, lässt sich über die Verteilung der Eigenwerte fast nichts aussagen.

Es gilt (vgl. [15, Satz 4.9])

Satz 4.22 *Sei Σ_n die Menge der Eigenwerte aller stochastischen $n \times n$–Matrizen und für $k \geq 2$ sei*

$$Z_k = \left\{ \sum_{j=0}^{k-1} c_j e^{\frac{2\pi i}{k} j} \quad \middle| \quad c_j \geq 0, \sum_{j=0}^{k-1} c_j = 1 \right\}$$

die konvexe Hülle der Punkte

$$1, e^{\frac{2\pi i}{k}}, \ldots, e^{\frac{2\pi i}{k}(k-1)},$$

also die Menge der Punkte innerhalb des abgeschlossenen regulären k–Ecks, das den Mittelpunkt in 0 und eine Ecke in 1 hat.

Dann gilt:

$$\bigcup_{k=2}^{n} Z_k \subseteq \Sigma_n.$$

Für große n nähert sich $\bigcup_{k=2}^{n} Z_k$ dem Einheitskreis, d.h. man kann für die Matrix, die aus einer Markov–Kette resultiert, nur voraussetzen, dass, wenn man die Eigenwerte vom Betrag 1 vernachlässigt, idealerweise die Eigenwerte in einem Kreis mit Radius r liegen.

Angesichts dieser Vorausinformation ist aber das Gauß–Seidel–Verfahren optimal, Relaxation lohnt sich nicht.

Satz 4.23 *Bezüglich der Information, dass alle Eigenwerte einer p–zyklischen, stochastischen Matrix in einem Kreis mit Radius r liegen, ist das Gauß–Seidel–Verfahren optimal.*

Beweis:
Nach [46, Theorem 8] existiert zu jeder kompakten Menge U, die aus mehr als einem Punkt besteht und deren Komplement die 1 enthält und einfach zusammenhängend ist, eine Eulerfunktion, die optimal bezüglich U ist.

In diesem Fall ist $U = \overline{D}_r$ und das Komplement ist $\tilde{U} = \overline{\mathbb{C}} \backslash U$. Durch die Abbildung

$$\tilde{p}(z) := \frac{r}{z}$$

wird die offene Einheitskreisscheibe auf $\tilde{U}$ abgebildet, wobei $\tilde{p}(0) = \infty$ ist. Zugleich ist $\tilde{p}(r) = 1$, d.h. dass der Wert s aus dem Beweis von [46, Theorem 8] gerade r ist.

Nach dem Beweis von [46, Theorem 8] setzt man nun

$$q(z) := \frac{1}{\tilde{p}(sz)} = z$$

und die Funktion $q(z) = z$ ist optimal bezüglich U.

Verglichen mit der (allgemeineren) Euler–Funktion

$$\frac{\omega z}{1 - (1 - \omega) z^p}$$

entspricht q genau dem Fall, dass

$$\omega = 1, \quad 1 - \omega = 0$$

ist, also gerade dem Gauß–Seidel–Verfahren. □

5 Numerische Beispiele und der Fall $p = 2$

Zur Verdeutlichung der Ergebnisse in den vorhergehenden Abschnitten sollen zum Abschluss ein paar Beispiele betrachtet werden. Außerdem soll der Fall $p = 2$ noch einmal gesondert behandelt werden.

Alle Berechnungen im Folgenden wurden mit MATLAB durchgeführt.

5.1 Beispiel 1: Beispiel von Stewart

Als erstes Beispiel werde die bereits in Abschnitt 4.1 erwähnte Matrix aus [57, Kapitel 7] betrachtet. Die Eigenwerte von J sind

$$\begin{aligned}
&-0.50000000000000 + 0.86602540378444i\\
&-0.50000000000000 - 0.86602540378444i\\
&\ \ 1.00000000000000\\
&-0.45354596039504 + 0.39037004425246i\\
&-0.45354596039504 - 0.39037004425246i\\
&-0.11129739500157 + 0.58796734561215i\\
&-0.11129739500157 - 0.58796734561215i\\
&\ \ 0.56484335539661 + 0.19759730135968i\\
&\ \ 0.56484335539661 - 0.19759730135968i\\
&\ \ 0.00000000000000
\end{aligned}$$

Als optimalen Relaxationsparameter[20] in $(0, \frac{3}{2})$ liefern die Programme, die nach den Sätzen aus [18] erstellt wurden, $\omega_0 = 1.01560396140845$. Der zugehörige optimale Konvergenzfaktor ist $\frac{1}{\eta_0^3} = 0.20621251191482$, es ist $\eta_0 = 1.69262855211170$.

Abbildung 26 zeigt die Lage der Eigenwerte sowie das optimale Hypozykloid.

Nach Satz 4.16 existiert ($p = 3$ ungerade, $\omega_0 > 1$) sowohl ein ω_+ als auch ein ω_-.

Durch Rechnung erhält man nach (65)

$$\begin{aligned}
\omega_+ &= 7.63834785042622\\
\omega_- &= -8.65395181183487
\end{aligned}$$

[20] In diesem Fall und auch im Weiteren sind die Zahlenwerte jeweils die auf die letzte Dezimale gerundeten Werte, die MATLAB liefert.

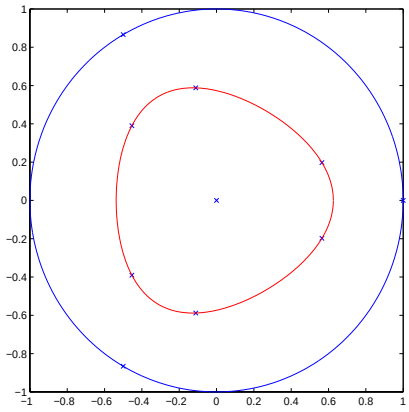

Abbildung 26: Das Spektrum der Stewart–Matrix und das optimale Hypozykloid.

Die zugehörigen Werte von η sind gemäß (66):

η_0	1.69262855211170
η_+	0.22505393789073
η_-	0.19864222728474

Der betragsgrößte Eigenwert von $\mathcal{L}_{\omega_+}$ ist $\lambda^+_{\max} = 425.4270871774196$, derjenige von $\mathcal{L}_{\omega_-}$ ist $\lambda^-_{\max} = -618.6859579520631$.

Damit gilt für den Konvergenzfaktor von ω_+:

$$\rho(\omega_+) = \left(\frac{1}{\eta_+ \sqrt[3]{425.4270871774196}} \right)^3 = 0.20621251191482.$$

Und entsprechend für ω_-:

$$\rho(\omega_-) = \left(\frac{1}{\eta_- \sqrt[3]{|-618.6859579520631|}} \right)^3 = 0.20621251191482,$$

d.h. es ergibt sich derselbe Konvergenzfaktor, der sich auch für ω_0 ergibt. Auch das Hypozykloid ist in allen drei Fällen dasselbe.

Abbildung 27 zeigt links den Konvergenzfaktor in Abhängigkeit von $\omega \in [7, 8]$, ω_+ ist dabei durch den senkrechten Strich bezeichnet.

Eine ähnliche Figur ergibt sich rechts für $\omega \in [-9, -8]$, hierbei ist ω_- durch den senkrechten Strich bezeichnet. Man beachte bei beiden Figuren, dass der Wertebereich der Ordinate nur das Intervall $[0.2062, 0.2065]$ ist.

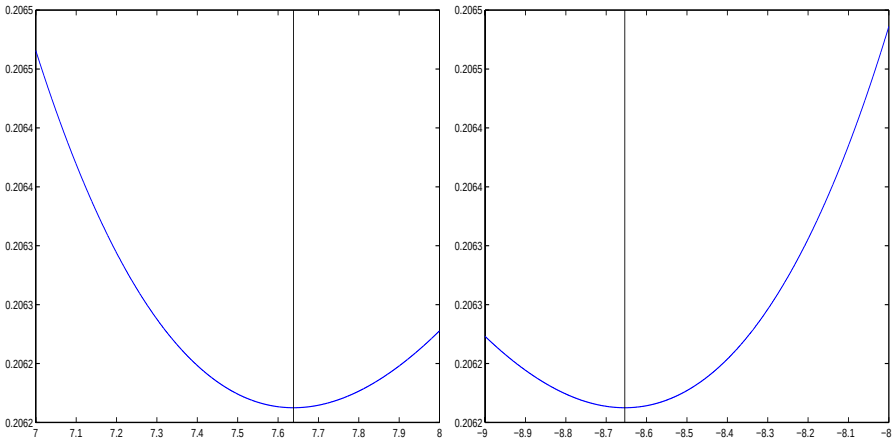

Abbildung 27: Konvergenzfaktor in Abhängigkeit vom Relaxationsparameter

Die „Langzeitentwicklung“ des Konvergenzfaktors zeigt Abbildung 28 (man beachte, dass die ω–Werte der Abszisse logarithmisch aufgetragen sind). Auch bei diesen Figuren sind die optimalen Relaxationsparameter eingetragen. Insbesondere sieht man, dass sich der Konvergenzfaktor nur wenig ändert, wenn man statt der optimalen Relaxationsparameter Werte wählt, die kleiner als ω_- oder größer als ω_+ sind. Die Parameter sind deutlich unempfindlicher gegen Störungen.

Bemerkung:

Die einfachste Art den Konvergenzfaktor zu bestimmen ist bei diesen kleinen Matrizen, das verallgemeinerte SOR–Verfahren als Vektoriteration aufzufassen, das heißt

$$\rho(\omega) = \frac{|\lambda_{\text{subdominant}}(\mathcal{L}_\omega)|}{|\lambda_{\max}(\mathcal{L}_\omega)|}$$

zu bestimmen.

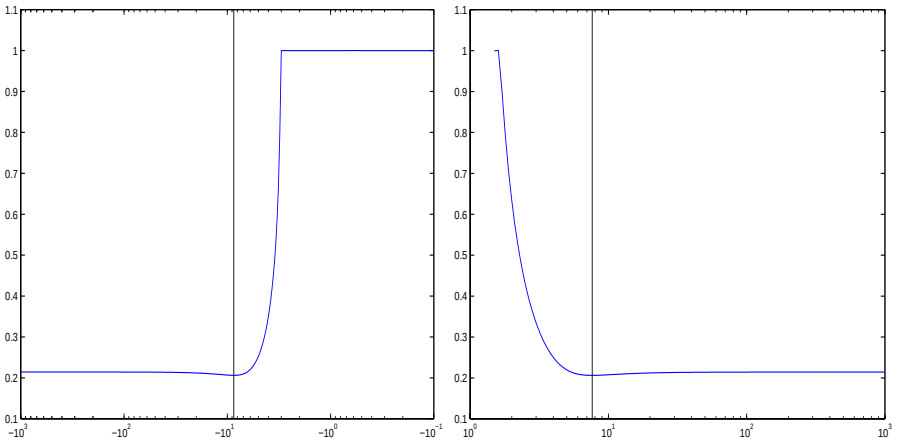

Abbildung 28: Abb. 27 mit größeren Relaxationsparameterwerten

Die *extended convergence* konvergiert in diesem Fall zunächst gegen einen Vektor, der Eigenvektor zum betragsgrößten Eigenwert ist. Ist dieser dem Eigenwert 1 von J^p zugeordnet, so läßt sich nach [34, Theorem 3.2] der zum Eigenwert 1 gehörende Eigenvektor rekonstruieren.

Problematisch ist der Fall, dass der dominante Eigenwert nicht dem Eigenwert 1 von J^p zugeordnet ist. Diese Bereiche sind in Abschnitt 4.1 besprochen.

5.2 Beispiel 2: Beispiel mit $p = 4$ und $\omega_0 < 1$

Gegeben sei die zufällig erzeugte Matrix[21] Q

$$Q = \begin{pmatrix}
-0.82 & 0.12 & 0.18 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0.16 & 0.18 & 0.19 \\
0.13 & -0.97 & 0.12 & 0.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0.15 & 0.07 & 0.03 \\
0.10 & 0.05 & -0.85 & 0.12 & 0 & 0 & 0 & 0 & 0 & 0 & 0.19 & 0.20 & 0.25 \\
0.19 & 0.24 & 0.20 & -0.83 & 0 & 0 & 0 & 0 & 0 & 0 & 0.14 & 0.09 & 0.15 \\
0.21 & 0.09 & 0.12 & 0.18 & -0.93 & 0.06 & 0.04 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.02 & 0.20 & 0.01 & 0.01 & 0.21 & -0.74 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.17 & 0.28 & 0.21 & 0.21 & 0.16 & 0.11 & -0.83 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.03 & 0.17 & 0.42 & -0.78 & 0.24 & 0.17 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.29 & 0.25 & 0.04 & 0.16 & -0.90 & 0.04 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.23 & 0.15 & 0.12 & 0.22 & 0.03 & -0.59 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.04 & 0.27 & 0.09 & -0.91 & 0.06 & 0.19 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.12 & 0.10 & 0.06 & 0.12 & -0.82 & 0.15 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.24 & 0.25 & 0.23 & 0.26 & 0.23 & -0.96
\end{pmatrix}$$

Man beachte, dass dies bereits die transponierte Matrix ist, d.h. dass $Qx = 0$ zu lösen ist.

Die Eigenwerte der zugehörigen (Block–)Jacobimatrix J sind

$$\begin{aligned}
&-1.00000000000000 \\
&0.00000000000000 + 1.00000000000000i \\
&0.00000000000000 - 1.00000000000000i \\
&1.00000000000000 \\
&-0.06664621730017 + 0.06664621730017i \\
&-0.06664621730017 - 0.06664621730017i \\
&0.06664621730017 + 0.06664621730017i \\
&0.06664621730017 - 0.06664621730017i \\
&-0.07122245281794 \\
&-0.00000000000000 + 0.07122245281794i \\
&-0.00000000000000 - 0.07122245281794i \\
&0.07122245281794 \\
&0.00000000000000,
\end{aligned}$$

d.h. es gilt $\sigma(J^4) \in [-0.7891544513e\text{–}4, 0.2573165410e\text{–}4]$.

Das optimale ω ist nach Satz 3.7

$$\omega_0 = 0.99999347851453$$

und es gilt

$$\eta_0 = 12.08653781717591$$

[21] Hier gerundet auf 2 Nachkommastellen.

bzw.

$$\rho(\omega_0) = 4.6858926934162 11e - 05.$$

Weil $p = 4$ gerade ist und $\omega_0 < 1$ gilt, existiert ein $\omega_- \in (-\infty, 0)$ mit demselben Konvergenzfaktor.

Die Rechnung nach (65) und (66) ergibt

$$\omega_- = -53.85311990663319$$

und

$$\eta_- = 0.22443377497812.$$

Da $\lambda^-_{\max} = 8.411138872797858e + 06$ ist, ergibt sich analog zum Kapitel 5.1

$$\rho(\omega_-) = \left(\frac{1}{\eta_- \sqrt[4]{\lambda^-_{\max}}} \right)^4 = 4.685892693415551e - 05.$$

Dass die beiden Werte sich in den letzten vier Dezimalen unterscheiden, ist auf mehrfache Rundungsfehler – bei der MATLAB–Eigenwertberechnung sowie bei der Lösung von (65) und (66) – zurückzuführen.

Die äquivalente Figur zu Abbildung 26 ist Abbildung 29, wobei der entscheidende Bereich gesondert hervorgehoben wird.

Die Entwicklung des Konvergenzfaktors im Intervall $[-54, -53]$ zeigt die Abbildung 30, ω_- ist dabei durch einen senkrechten Strich bezeichnet.

Die analogen Figuren für die größeren Intervalle sind Inhalt von Abbildung 31.

Man sieht, dass für $|\omega| \to \infty$ sich der Konvergenzfaktor demselben Wert nähert, für $\omega < 0$ von unten, für $\omega > 0$ von oben.

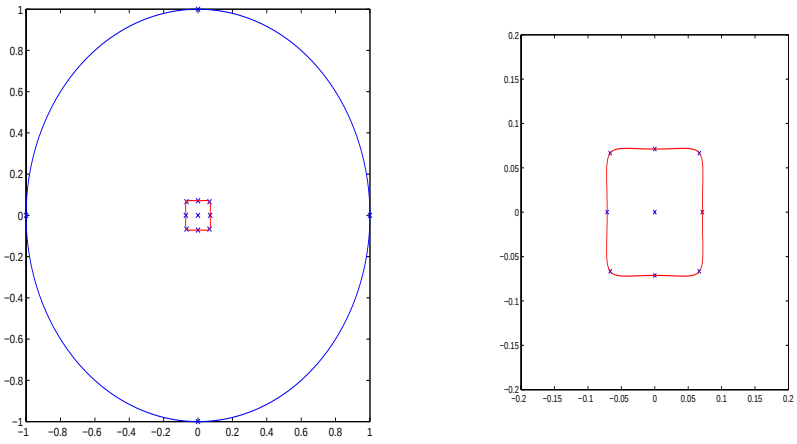

Abbildung 29: Spektrum und optimales Hypozykloid der Matrix aus Bsp.2

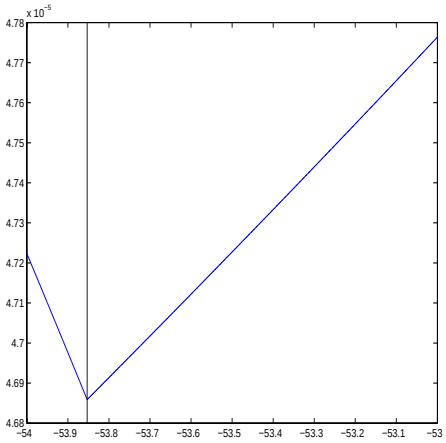

Abbildung 30: Konvergenzfaktor in Abhängigkeit vom Relaxationsparameter Bsp.2: $\omega \in [-54, 53]$

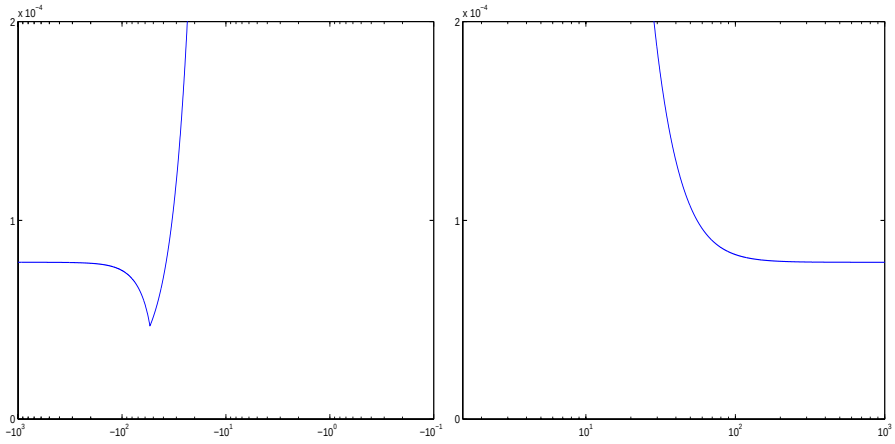

Abbildung 31: Abb. 30 mit größeren Intervallen

5.3 Beispiel 3: p ungerade, $\omega_0 < 1$

Gegeben sei die (wiederum zufällig erzeugte) Matrix[22]

$$Q = \begin{pmatrix}
-0.83 & 0.07 & 0.08 & 0.23 & 0 & 0 & 0 & 0.15 & 0.23 & 0.06 \\
0.23 & -0.99 & 0.10 & 0.29 & 0 & 0 & 0 & 0.01 & 0.04 & 0.11 \\
0.04 & 0.13 & -0.85 & 0.21 & 0 & 0 & 0 & 0.24 & 0.23 & 0.24 \\
0.02 & 0.37 & 0.18 & -0.99 & 0 & 0 & 0 & 0.21 & 0.24 & 0.09 \\
0.27 & 0.10 & 0.12 & 0.16 & -0.74 & 0.24 & 0.15 & 0 & 0 & 0 \\
0.14 & 0.02 & 0.23 & 0.02 & 0.15 & -0.70 & 0.22 & 0 & 0 & 0 \\
0.14 & 0.30 & 0.15 & 0.08 & 0.05 & 0.07 & -0.96 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.24 & 0.10 & 0.26 & -0.81 & 0.08 & 0.19 \\
0 & 0 & 0 & 0 & 0.03 & 0.08 & 0.17 & 0.06 & -0.88 & 0.15 \\
0 & 0 & 0 & 0 & 0.28 & 0.20 & 0.15 & 0.15 & 0.07 & -0.84
\end{pmatrix}.$$

Die Eigenwerte der Jacobi–Matrix J ergeben sich zu

$$\begin{array}{l}
-0.50000000000000 + 0.86602540378444i \\
-0.50000000000000 - 0.86602540378444i \\
1.00000000000000 \\
-0.12545003651055 \\
0.06272501825527 + 0.10864291852382i \\
0.06272501825527 - 0.10864291852382i \\
-0.09810042003577 \\
0.04905021001788 + 0.08495745587290i \\
0.04905021001788 - 0.08495745587290i \\
0.00000000000000
\end{array},$$

d.h. es ist (nach Satz 3.5)

$$\begin{aligned}
\omega_0 &= 0.99970776797775 \\
\eta_0 &= 11.96044138968262 \\
\rho(\omega_0) &= 5.844648241254521e-04
\end{aligned}$$

Das Spektrum von J sowie das zum optimalen Relaxationsparameter gehörende Hypozykloid ist aus den beiden Schaubildern in Abbildung 32 ersichtlich:

[22] Auch hier ist $Qx = 0$ zu lösen.

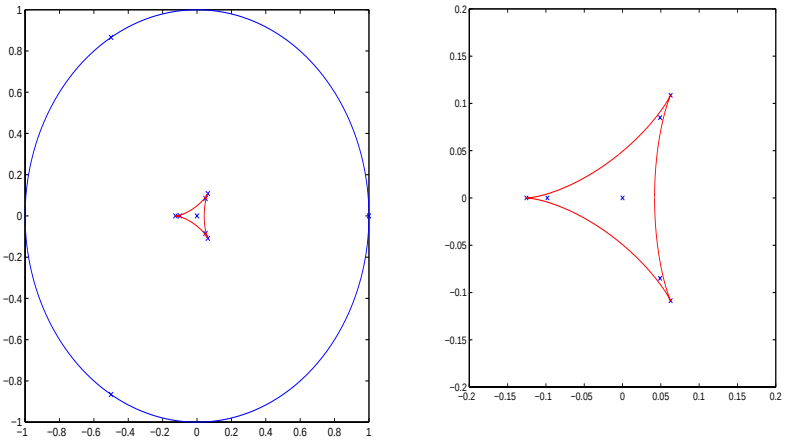

Abbildung 32: Spektrum von J in Bespiel 3 mit dem optimalen Hypozykloid

Nach Satz 4.15 bzw. 4.17 aus Kapitel 4.4 ist dies das einzige optimale ω. Die Abbildungen 33 und 34 zeigen den Konvergenzfaktor in Abhängigkeit vom Relaxationsparameter in den Intervallen $[-1000, 0)$, $(\frac{3}{2}, 1000]$ und $(0, \frac{3}{2})$. Der zu ω_0 gehörende Konvergenzfaktor ist dabei durch eine waagrechte Linie gekennzeichnet.

Man sieht, dass die Konvergenzfaktoren in den Teilintervallen $[-1000, 0)$ und $[\frac{3}{2}, 1000]$ schlechter sind als der optimale Konvergenzfaktor und dass es somit außer ω_0 keinen optimalen Konvergenzfaktor gibt.

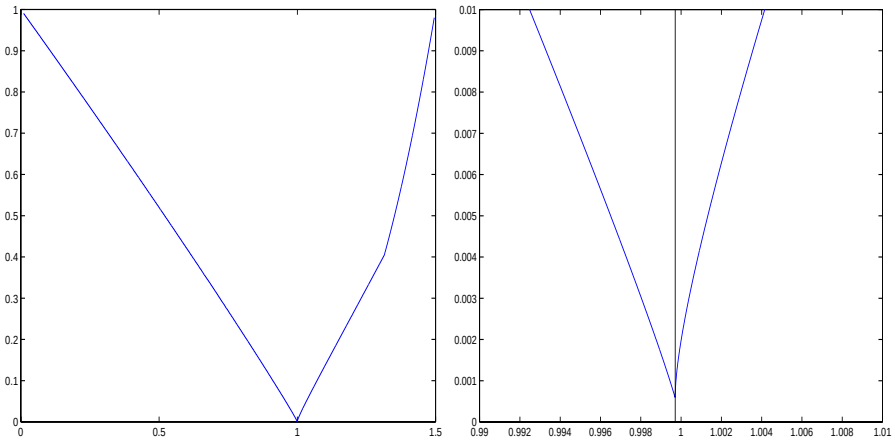

Abbildung 33: zu 30 analoge Abbildung

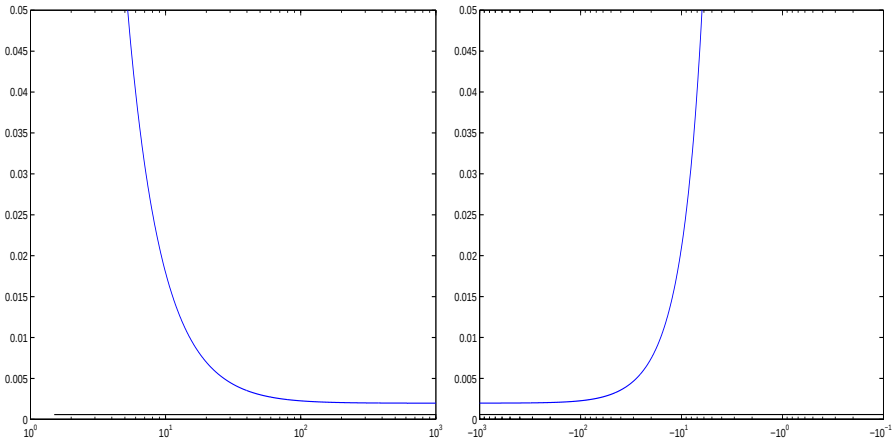

Abbildung 34: zu Abb.31 analoge Abbildung

5.4 Der Fall $p = 2$ und ein Beispiel

Eine Sonderrolle spielt — nicht nur aus historischen Gründen — der Fall $p = 2$. Die Hypozykloide werden in diesem Fall zu Ellipsen.

Dieser Fall ist unter anderem auch deswegen interessant, weil Pierce, Hadjidimos und Plemmons in [50] folgenden Satz zeigten:

Satz 5.1 *Die p–zyklische Matrix A werde zu einer q–zyklischen Form mit $2 \leq q < p$ repartitioniert. Ferner gelte*

1. $\sigma(J_p^p) \geq 0$ *und* $0 < \rho(J_p) < 1$ *oder*
2. $\sigma(J_p^p) \leq 0$ *und* $0 < \rho(J_p) < \frac{p}{p-2}$.

Dann gilt $\rho(\mathcal{L}_{\omega_q}) < \rho(\mathcal{L}_{\omega_p})$, d.h. die optimale q–zyklische SOR–Iteration ist asymptotisch schneller als die optimale p–zyklische.

Hieraus folgt sofort

Korollar 5.2 *Unter den Voraussetzungen von Satz 5.1 ist stets die 2–zyklische Iteration optimal.*

Dass Satz 5.1 unter allgemeineren Voraussetzungen an das Spektrum von J bzw. J^p nicht gilt, zeigen Eiermann, Niethammer und Ruttan in [11].

Für eine 2–zyklische Matrix sind die zu 1 gehörenden Eigenwerte von $\mathcal{L}_\omega$ 1 und $(\omega - 1)^2$.

Zu $\mu \neq 1$ gehören

$$\lambda_{1/2} = 1 - \omega + \frac{1}{2}\omega^2\mu^2 \pm \frac{1}{2}\omega\mu\sqrt{4 - 4\omega + \omega^2\mu^2}.$$

Das optimale $\omega \in (0, 2)$ ist — für $\tilde{\sigma}(J^2) \in \mathbb{R}^-$ bzw. $\tilde{\sigma}(J^2) \in \mathbb{R}^+$ — die Nullstelle von

$$\omega^2\mu^2 - 4\omega + 4 = 0 \tag{78}$$

in $(0, 2)$. Das heißt, für dieses ω gilt

$$\mu^2 = \frac{4\omega - 4}{\omega^2}.$$

Der Konvergenzfaktor ist damit

$$\rho(\omega_0) = \left| \frac{1 - \omega + \frac{1}{2}\omega^2\mu^2}{1} \right| = |1 - \omega + \frac{1}{2}(4\omega - 4)| = |\omega - 1|.$$

Entsprechend gilt für ω_- bzw. ω_+ (in der Formel gelte $x \in \{+, -\}$)

$$\rho(\omega_x) = \left| \frac{1 - \omega_x + \frac{1}{2}(4\omega_x - 4)}{(\omega_x - 1)^2} \right| = \left| \frac{\omega_x - 1}{(\omega_x - 1)^2} \right| = \left| \frac{1}{\omega_x - 1} \right|.$$

Es sind die Lösungen von

$$\mu^2\omega^2 - 4\omega + 4 = 0$$

die Parameter

$$\omega_{1,2} = \frac{2}{\mu^2} \pm \frac{2}{\mu^2}\sqrt{1 - \mu^2}.$$

Je nachdem, ob $\mu^2 > 0$ oder $\mu^2 < 0$ ist, ist der eine oder andere Wert der betragsmäßig größere. Unabhängig davon ist aber

$$\frac{1}{\frac{2}{\mu^2}(1 \pm \sqrt{1 - \mu^2}) - 1} = \frac{2}{\mu^2}(1 \mp \sqrt{1 - \mu^2}) - 1,$$

weil

$$1 = \frac{4}{\mu^4}(1 - (1 - \mu^2)) - \frac{2}{\mu^2}(1 \pm \sqrt{1 - \mu^2}) - \frac{2}{\mu^2}(1 \mp \sqrt{1 - \mu^2}) + 1 = \frac{4}{\mu^2} - \frac{4}{\mu^2} + 1.$$

Das heißt insbesondere, dass

$$\rho(\omega_x) = \left| \frac{1}{\omega_x - 1} \right| = |\omega_0 - 1|$$

ist; das heißt, ω_x hat denselben Konvergenzfaktor wie ω_0.

Dass keine weiteren optimalen Relaxationsparameter existieren, folgt wie im allgemeinen Fall.

Im Fall, dass $\tilde{\sigma}(J^2) \in [-\alpha^2, \beta^2], \quad \alpha, \beta \in \mathbb{R}^+$ ist (vgl. auch [11]) Gleichung (78) zu ersetzen durch

$$\left(\frac{\alpha + \beta}{2} \right)^2 \omega^2 - \frac{\alpha + \beta}{\beta - \alpha}\omega + \frac{\alpha + \beta}{\beta - \alpha} = 0. \tag{79}$$

Die Rechnungen werden dadurch deutlich aufwendiger und unübersichtlicher, lassen sich aber im Prinzip genauso durchführen.

Ein Beispiel

Nehmen wir an, die Jacobi–Matrix einer 2–periodischen Markov–Kette habe folgende Eigenwerte

$$\begin{aligned} \mu_1 &= 1 \\ \mu_2 &= -1 \end{aligned}$$

$$\begin{aligned} \mu_3 &= \tfrac{1}{2}i \\ \mu_4 &= -\tfrac{1}{2}i \end{aligned} \quad \text{beziehungsweise} \quad \begin{aligned} \hat{\mu}_3 &= \tfrac{1}{2} \\ \hat{\mu}_4 &= -\tfrac{1}{2} \end{aligned}$$

Dann sind die zu μ_1 und μ_2 gehörenden Eigenwerte der SOR–Matrix

$$\begin{aligned} \lambda_1(\omega) &= 1 \\ \lambda_2(\omega) &= (\omega - 1)^2 \end{aligned}$$

Die zu μ_3 und μ_4 gehörenden Eigenwerte sind die Lösungen von

$$\lambda^2 + \lambda(2(\omega - 1) + \frac{1}{4}\omega^2) + (\omega - 1)^2 = 0,$$

d.h.

$$\lambda_{3/4} = -\frac{1}{8}\omega^2 + (1 - \omega) \pm \frac{1}{2}\omega\sqrt{\frac{1}{16}\omega^2 + \omega - 1}. \tag{80}$$

Bezogen auf $\hat{\mu}_3$ und $\hat{\mu}_4$ ergibt sich

$$\hat{\lambda}_{3/4} = \frac{1}{8}\omega^2 - (\omega - 1) \pm \frac{1}{2}\omega\sqrt{\frac{1}{16}\omega^2 - \omega + 1}. \tag{81}$$

Nach [65] ist das optimale ω Lösung von

$$\omega^2\mu_3^2 - 4(\omega - 1) = 0 \tag{82}$$

bzw.

$$\omega^2\hat{\mu}_3^2 - 4(\omega - 1) = 0. \tag{83}$$

Die Lösungen dieser beiden Gleichungen sind aber genau die Werte von ω, für die der Ausdruck unter der Wurzel in den Gleichungen (80) bzw. (81) zu 0 wird.

Zum ersten Fall ($\mu_{3/4} = \pm\frac{1}{2}i$):

es ist $\omega_0 \in (0,1)$, was wegen $\mu_3^2 < 0$ klar ist. Außerdem besitzt die Gleichung (82) eine zweite Nullstelle. Diese muss wegen

$$g(-\omega) = \underbrace{\omega^2 \underbrace{\mu_3^2}_{<0}}_{<0} + \underbrace{4|\omega|}_{>0} + 4 = 0$$

aufgrund der Descartesschen Zeichenregel negativ sein, d.h. es existiert ein ω_-.

Zum zweiten Fall ($\hat{\mu}_{3/4} = \pm\frac{1}{2}$): hier ist $\omega_0 \in (1,2)$ und es muss aufgrund derselben Zeichenregel eine weitere positive Nullstelle geben, d.h. ein ω_+. Es ist nämlich

$$\hat{g}(\omega) = \underbrace{\omega^2 \hat{\mu}_3^2}_{>0} \underbrace{-4\omega}_{<0} \underbrace{+4}_{>0} = 0,$$

und da $\hat{g}$ mit ω_0 mindestens eine positive Nullstelle hat, muss es noch eine zweite geben.

Nach Satz 4.12 ist die zweite Nullstelle die Lösung von

$$\frac{x-1}{x^2} = \frac{\omega_0 - 1}{\omega_0^2},$$

d.h.

$$0 = x^2 - \frac{\omega_0^2}{\omega_0 - 1}x + \frac{\omega_0^2}{\omega_0 - 1},$$

also

$$\begin{aligned} x_{1/2} &= \frac{\omega_0^2}{2(\omega_0-1)} \pm \sqrt{\frac{\omega_0^4}{4(\omega_0-1)^2} - \frac{\omega_0^2}{\omega_0-1}} \\ &= \frac{\omega_0^2}{2(\omega_0-1)} \pm \frac{\omega_0}{2(\omega_0-1)}\sqrt{\omega_0^2 - 4\omega_0 + 4} \\ &= \frac{\omega_0^2 \pm \omega_0^2 \mp 2\omega_0}{2(\omega_0-1)} \end{aligned}$$

d.h.

$$\begin{aligned} x_1 &= \omega_0 \\ x_2 &= \frac{\omega_0}{\omega_0 - 1} \end{aligned}$$

Man sieht, dass die Lösung je nach Lage von ω_0 negativ oder positiv ist. Das entspricht gerade auch den Ergebnissen nach Korollar 4.17.

Im ersten Fall sind die optimalen Relaxationsparameter[23]

$$\begin{aligned} \omega_0 &= \sqrt{80} - 8 \approx 0.9442719 \\ \omega_- &= -8 - \sqrt{80} \approx -16.9442719, \end{aligned}$$

in zweiten Fall

$$\begin{aligned} \omega_0 &= 8 - \sqrt{48} \approx 1.07179677 \\ \omega_+ &= 8 + \sqrt{48} \approx 14.92820323, \end{aligned}$$

und es ist

$$\frac{\sqrt{80} - 8}{\sqrt{80} - 8 - 1} = \frac{(\sqrt{80} - 8)(\sqrt{80} + 9)}{-1} = -8 - \sqrt{80}$$

bzw.

$$\frac{8 - \sqrt{48}}{8 - \sqrt{48} - 1} = \frac{8 - \sqrt{48}}{7 - \sqrt{48}} = 8 + \sqrt{48}.$$

Dies entspricht natürlich auch den Herleitungen in [42], [43]:

danach ist für den zweiten Fall[24]

$$\begin{aligned} \omega_0 &= \frac{2}{1 + \sqrt{1 - \overline{\mu}^2}} \\ &= \frac{2}{1 + \sqrt{1 - \frac{1}{4}}} \\ &= \frac{2}{1 + \frac{1}{2}\sqrt{3}} \\ &= \frac{2(1 - \frac{1}{2}\sqrt{3})}{\frac{1}{4}} \\ &= 8 - 4\sqrt{3} = 8 - \sqrt{48} \end{aligned}$$

[23] Die Schreibweise $a \approx b$ im Folgenden soll bedeuten, dass b gerade a gerundet auf die letzte Dezimale entspricht.

[24] In [42], [43] wird ω_0 als ω_- bezeichnet.

und

$$\begin{aligned}\omega_+ &= \frac{2}{1-\sqrt{1-\overline{\mu}^2}} \\ &= \frac{2}{1-\sqrt{1-\frac{1}{4}}} \\ &= \frac{2}{1-\frac{1}{2}\sqrt{3}} \\ &= \frac{2(1+\frac{1}{2}\sqrt{3})}{1-\frac{3}{4}} \\ &= 8+4\sqrt{3} = 8+\sqrt{48}.\end{aligned}$$

Betrachten wir nun den Konvergenzfaktor gemäß den Ausführungen in [42].

Zunächst für den zweiten Fall:

es ist (nach [42, Proposition 1])

$$\rho(\omega_0) = \frac{1-\sqrt{1-\overline{\mu}^2}}{1+\sqrt{1-\overline{\mu}^2}} = \frac{1-\sqrt{1-\frac{1}{4}}}{1+\sqrt{1+\frac{1}{4}}} = \frac{(1-\frac{1}{2}\sqrt{3})^2}{1-\frac{3}{4}} = 7-\sqrt{48}$$

und

$$\rho(\omega_+) = \rho(\omega_0).$$

Zu untersuchen bleibt der erste Fall:

Es ist

$$\begin{aligned}\rho(\omega_0) &= \frac{|\lambda_{\text{subdominant}}|}{|\lambda_{\max}|} \\ &= \frac{9-4\sqrt{5}}{1} \\ &= 9-4\sqrt{5},\end{aligned}$$

wobei $9-4\sqrt{5}$ die (doppelte) Lösung von

$$(\lambda+\sqrt{80}-9)^2 = -\frac{1}{4}\lambda(\sqrt{80}-8)^2$$

ist.

Für ω_- ist

$$|\lambda_{\text{subdominant}}| = 9 + 4\sqrt{5}$$

als (doppelte) Lösung von

$$(\lambda - 9 - \sqrt{80})^2 = -\frac{1}{4}\lambda(-8 - \sqrt{80})^2$$

und

$$|\lambda_{\max}| = 161 + 72\sqrt{5}$$

als Lösung (neben 1) von

$$(\lambda - 9 - \sqrt{80})^2 = \lambda(-8 - \sqrt{80})^2.$$

Damit ist

$$\begin{aligned} \rho(\omega_-) &= \frac{9 + 4\sqrt{5}}{161 + 72\sqrt{5}} \\ &= (9 + 4\sqrt{5})(161 - 72\sqrt{5}) \\ &= 9 - 4\sqrt{5}, \end{aligned}$$

d.h. es ergibt sich derselbe Konvergenzfaktor (vgl. auch Satz 4.13).

Literatur

[1] L. V. Ahlfors. *Complex Analysis.* McGraw–Hill, New York, 3rd edition, 1979.

[2] R. B. Ash and R. L. Bishop. Monopoly as a Markov Process. *Mathematics Magazine*, 45:26–29, 1972.

[3] O. Axelsson. *Iterative Solution Methods.* Cambridge University Press, Cambridge, 1994.

[4] G. Barker and R. J. Plemmons. Convergent Iterations for Computing Stationary Distributions of Markov Chains. *SIAM J. Alg. Disc. Meth.*, 7(3):390–398, 1986.

[5] M. Benidir. On the Root Distribution of General Polynomials with Respect to the Unit Circle. *Signal Processing*, 53:75–82, 1996.

[6] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.

[7] J. W. Bishir and D. W. Drewes. *Mathematics in the Behavioral and Social Sciences.* Harcourt, Brace & World, New York, 1970.

[8] P. J. Courtois and P. Semal. Block Iterative Algorithms for Stochastic Matrices. *Linear Algebra and its Applications*, 76:59–70, 1986.

[9] M. Eiermann, I. Marek, and W. Niethammer. On the Solution of Singular Linear Systems of Algebraic Equations by Semiiterative Methods. *Numerische Mathematik*, 53(3):265–283, 1988.

[10] M. Eiermann and W. Niethammer. On the Construction of Semiiterative Methods. *SIAM J. Numer. Anal.*, 20(6):1153–1160, 1983.

[11] M. Eiermann, W. Niethammer, and A. Ruttan. Optimal Successive Overrelaxation Iterative Methods for P–Cyclic matrices. *Numerische Mathematik*, 57:593–606, 1990.

[12] J. Eisenfeld and N. Mitra. Relationship between Compartmental Matrices, M–Matrices, and Markov Chains with Application in Economics. *Journal of Mathematical Analysis and Applications*, 194:529–547, 1995.

[13] L. Fahrmeir, H. Hanfmann, and F. Ost. *Stochastische Prozesse.* Hanser, München, Wien, 1981.

[14] K. Fladt. *Analytische Geometrie spezieller ebener Kurven.* Akademische Verlagsgesellschaft, Frankfurt am Main, 1962.

[15] F.-J. Fritz, B. Huppert, and W. Willems, editors. *Stochastische Matrizen.* Springer–Verlag, Berlin, Heidelberg, New York, 1979.

[16] S. Galanis and A. Hadjidimos. Best Cyclic Repartitioning for Optimal Successive Overrelaxation Convergence. *SIAM J. Matrix Anal. Appl.*, 13(1):102–120, 1992.

[17] S. Galanis, A. Hadjidimos, and D. Noutsos. A Young–Eidson's Type Algorithm for Complex p–Cyclic SOR Spectra. unveröffentlicht.

[18] S. Galanis, A. Hadjidimos, and D. Noutsos. Optimal p-Cyclic SOR for Complex Spectra. *Linear Algebra and its Applications*, 263:233–260, 1997.

[19] S. Galanis, A. Hadjidimos, and D. Noutsos. A Young–Eidson's Type Algorithm for Complex p–Cyclic SOR Spectra. *Linear Algebra and its Applications*, 286(1–3):87–106, 1999.

[20] G. Golub, A. Greenbaum, and M. Luskin, editors. *Recent Advances in Iterative Methods*, volume 60 of *The IMA Volumes in Mathematics and Its Applications.* Springer–Verlag, New York, Berlin, Heidelberg, 1994.

[21] G. H. Golub and C. van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore, 1983.

[22] M. Gutknecht, W. Niethammer, and R. Varga. k-Step Iterative Methods for Solving Nonlinear Systems of Equations. *Numerische Mathematik*, 48:699–712, 1986.

[23] W. Hackbusch. *Iterative Lösung großer schwach besetzter Gleichungssysteme.* Teubner, Stuttgart, 1986.

[24] A. Hadjidimos. On the Optimization of the Classical Iterative Schemes for the Solution of Complex Singular Linear Systems. *SIAM J. Alg. Disc. Meth.*, 6(4):555–566, 1985.

[25] A. Hadjidimos, D. Noutsos, and M. Tzoumas. Exact SOR Convergence Regions for a General Class of P-Cyclic Matrices. *BIT*, 35(4):469–487, 1995.

[26] A. Hadjidimos and R. J. Plemmons. Analysis of p–Cyclic Iterations for Markov Cains. In C. D. Meyer and R. J. Plemmons, editors, *Linear Algebra, Markov Chains, and Queueing Models*, volume 48 of *The IMA Volumes in Mathematics and Its Applications*, pages 111–124. Springer–Verlag, New York, Berlin, Heidelberg, 1993.

[27] P. Henrici. *Applied and Computational Complex Analysis I.* Pure & Applied Mathematics. John Wiley & Sons, New York, London, Sydney, Toronto, 1974.

[28] M. Hochbruck. *Lanczos– und Krylov–Verfahren für nicht–Hermitesche lineare Systeme.* PhD thesis, Universität Karlsruhe, Karlsruhe, Mai 1992.

[29] M. Hochbruck and C. Lubich. On Krylov Subspace Approximations on the Matrix Exponential Operator. Technical report, Sonderforschungsbereich 382 Verfahren und Algorithmen zur Simulation physikalischer Prozesse auf Höchstleistungsrechnern, Januar 1995.

[30] M. Hochbruck and C. Lubich. On Krylov Subspace Approximations on the Matrix Exponential Operator. *SIAM Journal on Numerical Analysis*, 34(5):1911–1925, 1997.

[31] A. Householder. *The Numerical Treatment of A Single Nonlinear Equation.* International Series in Pure and Applied Mathematics. McGraw–Hill, New York, 1970.

[32] D. L. Isaacson and R. W. Madsen. *Markov Chains — Theory and Applications.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, London, Sydney, Toronto, 1976.

[33] H. B. Keller. On the Solution of Singular and Semidefinite Linear Systems by Iteration. *J. SIAM Numerical Analysis*, 2(2):281–290, 1965.

[34] K. Kontovasilis, R. J. Plemmons, and W. J. Stewart. Block Cyclic SOR for Markov Chains With p–Cyclic Infinitesimal Generator. *Linear Algebra and its Applications*, 154–156:145–223, 1991.

[35] T. Kratochwill. *Single Subject Research.* Academic Press, New York, 1978.

[36] F. Locher. A Stability Test for Real Polynomials. *Numerische Mathematik*, 66:33–40, 1993.

[37] F. Locher and M.-R. Skrzipek. An Algorithm for Locating All Zeros of a Real Polynomial. *Computing*, 54:359–375, 1995.

[38] I. Marek and D. B. Szyld. Iterative and Semi–Iterative Methods for Computing Stationary Probability Vectors of Markov Operators. *Mathematics of Computation*, 61(204):719–731, 1993.

[39] The MathWorks, Inc., Conchituate Place, 24 Prime Park Way, Natick, Mass. 01760. *Matlab Reference Guide*, August 1992.

[40] C. D. Meyer and R. J. Plemmons, editors. *Linear Algebra, Markov Chains, and Queueing Models*, volume 48 of *The IMA Volumes in Mathematics and Its Applications*. Springer–Verlag, New York, Berlin, Heidelberg, 1993.

[41] E. B. Newmann. The Pattern of Vowels and Consonants in Various Languages. *American Journal of Psychology*, 64:369–379, 1951.

[42] W. Niethammer. Markov Chains and Extended Convergence of SOR. Technical Report 98/3, Universität Karlsruhe, Fakultät für Mathematik, 1998.

[43] W. Niethammer. A Note on the Extended Convergence of SOR for Two–periodic Markov Chains. *Linear Algebra and its Applications*, to appear.

[44] W. Niethammer, J. de Pillis, and R. S. Varga. Convergence of Block Iterative Methods Applied to Sparse Least–Squares Problems. *Linear Algebra and its Applications*, 58:327–341, 1984.

[45] W. Niethammer and M. Eiermann. Numerische Lösung von Gleichungssystemen. Technical report, Studientext der Fernuniversität Hagen, Sommersemester 1998.

[46] W. Niethammer and R. S. Varga. The Analysis of k–Step Iterative Methods for Linear Systems from Summability Theory. *Numerische Mathematik*, 41:177–206, 1983.

[47] W. Niethammer and R. S. Varga. Relaxation Methods for Non–Hermitian Linear Systems. *Result. Math.*, 16(3/4):308–320, 1989.

[48] J. M. Ortega. *Numerical Analysis*. Classics in Applied Mathematics. SIAM, Philadelphia, 1990.

[49] B. Philippe, Y. Saad, and W. J. Stewart. Numerical Methods in Markov Chain Modeling. *Operations Research*, 40(6):1156–1179, 1992.

[50] D. J. Pierce, A. Hadjidimos, and R. J. Plemmons. Optimality Relationship for p–Cyclic SOR. *Numerische Mathematik*, 56:635–643, 1990.

[51] Y. Saad. Preconditionded Krylov Subspace Methods for the Numerical Solution of Markov Chains. In W. J. Stewart, editor, *Computations with Markov Chains*, pages 49–64. Kluwer Academic Publishers, Raleigh, 1995.

[52] G. Salmon. *Higher Plane Curves*. Chelsea Publishing Company, New York, 3rd edition, 1960.

[53] H. R. Schwarz. *Numerische Mathematik*. Teubner–Verlag, Stuttgart, 1997.

[54] W. J. Stewart. Direct Methods for the Numerical Solution of Markov Chains. Technical Report 89-04, North Carolina State University, 1989.

[55] W. J. Stewart. Markov Chains and Stochastic Matrices. Technical Report 89-01, North Carolina State University, 1989.

[56] W. J. Stewart, editor. *Numerical Solution of Markov Chains.* Marcel Dekker, Inc., New York, Basel, Hongkong, 1991.

[57] W. J. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, 1994.

[58] W. J. Stewart, editor. *Computations with Markov Chains.* Kluwer Academic Publishers, Boston, London, Dordrecht, 1995. Proceedings of the 2nd International Workshop on the Numerical Solution of Markov Chains.

[59] J. Stoer. *Einführung in die Numerische Mathematik I.* Springer–Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[60] J. Stoer and R. Bulirsch. *Einführung in die Numerische Mathematik II.* Springer–Verlag, Berlin, Heidelberg, New York, 197.

[61] K. Tanabe. Charakterization of Linear Stationary Iterative Processes for Solving a Singular System of Linear Equations. *Numerische Mathematik*, 22:349–359, 1974.

[62] S. Tavaré. The Genealogy of the Birth, Death, and Immigration Process. In M. W. Feldmann, editor, *Mathematical Evolutionary Theory*, pages 41–56. Princeton University Press, Princeton, New Jersey, 1989.

[63] R. S. Varga. p–Cyclic Matrices: A Generalization of the Young–Frankel Successive Overrelaxation Scheme. *Pacific J. Math.*, 9:617–628, 1959.

[64] R. S. Varga. *Matrix Iterative Analysis.* Prentice Hall, London, Sydney, Toronto, Tokio, 1962.

[65] P. Wild and W. Niethammer. Over– and Underrelaxation for Linear Systems with Weakly Cyclic Jacobi Matrices of Index p. *Linear Algebra and its Applications*, 91:29–52, 1987.

[66] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

[67] D. M. Young. *Iterative Solution of Large Linear Systems.* Computer Science and Applied Mathematics. Academic Press, New York, London, 1971.

Lebenslauf

Grischa Markus Freimann

10.07.1969	Geboren Freiburg im Breisgau Eltern: Dr. Hans Freimann und Gerda Freimann, geb. Masanneck
1975–1976	Grundschule in Freiburg–Kappel
1976–1979	Schulzentrums Steinen
1979–1988	Hebel–Gymnasiums Lörrach
03.05.1988	Abschluss: Abitur
01.10.1988–15.03.1989	Wehrdienst in Marburg an der Lahn und Weingarten
WS 1989/90– WS 1994/95	Studium der Wirtschaftsmathematik an der Universität Karlsruhe
16.12.1994	Diplom
1992–1994	Wissenschaftliche Hilfskraft am Institut für Praktische Mathematik
01.01.1995–31.12.1996	Wiss. Angestellter am IWRMM der Universität Karlsruhe
01.01.1997–31.12.1997	IPP–Stipendiat der Fakultät für Mathematik an der Universität Karlsruhe
seit 01.01.1998	Wiss. Angestellter am Institut für Praktische Mathematik der Universität Karlsruhe