Universität Bielefeld
Fakultät für Mathematik

Diplomarbeit

Exponentielle Integratoren

Vorgelegt von: Simon Dieckmann
Betreuer: Prof. Dr. W.-J. Beyn

Bielefeld, November 2010

Inhaltsverzeichnis

Einleitung

Ein- und Mehrschrittverfahren sind das übliche Mittel zur Lösung von gewöhnlichen Differentialgleichungen, wenn keine explizite Lösungsformel bekannt ist. Eine wesentliche Rolle spielen sie bei Verwendung der sogenannten (vertikalen) Linienmethode zur Lösung parabolischer partieller Differentialgleichungen. Dabei wird das Anfangsrandwertproblem durch Semidiskretisierung in ein System gewöhnlicher Differentialgleichungen überführt, welches dann mittels eines Ein- oder Mehrschrittverfahrens gelöst werden kann.

Während analytische Lösungsmethoden vielfach nicht einsetzbar sind, ermöglichte die Entwicklung leistungsstarker Prozessoren die Berechnung numerischer Lösungen in hinnehmbarer Zeit. Explizite Runge-Kutta-Verfahren sind insbesondere bei Verwendung mehrerer parallel arbeitender Prozessoren eine gute Wahl, da die Berechnung der numerischen Lösung allein durch Additionen und Multiplikationen von Matrizen und Vektoren erfolgt. Allerdings führen sie zu hohem Aufwand, da wegen der schlechten Stabilitätseigenschaften die Zeitschrittweite sehr klein gewählt werden muss. Implizite Verfahren hingegen erfordern das Lösen von Gleichungssystemen, was Parallelisierung und Vektorisierung erschwert.

Die Suche nach einem Mittelweg führte in den 90er Jahren (siehe [38], [9], [19], [20]) zu den sogenannten **exponentiellen Integratoren**. Dies sind Verfahren, die auf der Approximation des Matrixexponentials oder ähnlicher Operatoren, den sogenannten Phi-Funktionen, beruhen. Der Operator selbst wird allerdings nicht berechnet, sondern lediglich das Produkt mit einem Vektor mittels Krylow-Unterraum-Approximation ausgewertet.

Die Idee Verfahren zu konstruieren, die im simplen Fall einer linearen Differentialgleichung mit konstanten Koeffizienten $u'(t) = Au(t)$ die exakte Lösung liefern, geht zumindest auf das Jahr 1958 zurück (siehe [16]). In den folgenden drei Jahrzehnten wurde auf Basis der Formel zur Variation der Konstanten eine Vielzahl von exponentiellen Integratoren konstruiert. Dem praktischen Nutzen stand jedoch die Schwierigkeit der Auswertung des Matrixexponentials entgegen.

Das erste Kapitel dieser Diplomarbeit basiert auf der Publikation [20] von M. Hochbruck, C. Lubich und H. Selhofer. Zunächst wird exemplarisch die Konstruktion des exponentiell angepassten Euler-Verfahrens mittels Variation der Konstanten motiviert, um dann die klassischen Konsistenz- und damit auch Konvergenzaussagen für die bereits in [19] eingeführten Einschrittverfahren herzuleiten.

Um im Vergleich zu impliziten Runge-Kutta- und Mehrschrittverfahren konkurrenzfähig zu sein, muss der Aufwand zur Berechnung der Krylow-Unterräume minimiert werden. Dies führt zu den sogenannten reduzierten Verfahren, deren bekanntester Vertreter exp4 ist. Deren Konstruktion erlaubt es, die Berechnung einiger Krylow-Unterräume durch Matrix-Vektor-Multiplikationen zu vereinfachen bzw. ganz zu ersetzen. Im letzten Teil des ersten Kapitels wird die Theorie, die hinter der Konstruktion der reduzierten Verfahren steht, analysiert.

Das zweite Kapitel, das auf der Publikation [19] von M. Hochbruck und C. Lubich basiert, thematisiert ebenfalls die Frage, warum exponentielle Integratoren wie exp4 trotz der aufwendigen Auswertung des Matrixexponentials konkurenzfähig sind. Mit Hilfe des numerischen Wertebereichs

$$\mathcal{F}(A) := \Big\{ \langle Az, z\rangle : z \in \mathcal{H},\ \langle z, z\rangle = 1 \Big\}$$

wird zunächst eine allgemeine Formel für die Konvergenz des Arnoldi-Verfahrens, das zu den bekanntesten Vertretern der Krylow-Unterraum-Approximationen zählt, hergeleitet.

Im Fall einer hermiteschen, semidefiniten Matrix erhält man ab einem bestimmten Punkt in der Iteration super-lineare Konvergenz. Sofern kein besserer Vorkonditionierer (siehe [32]) für das Verfahren der konjugierten Gradienten zur Verfügung steht, führt dies zu einem deutlichen Vorteil im Vergleich zu impliziten Verfahren. Im zweiten Beispiel ist der numerische Wertebereich in einer Kreisscheibe in der negativen komplexen Halbebene enthalten.

Krylow-Unterraum-Approximationen sind nur eine von vielen Möglichkeiten zur Auswertung des Matrixexponentials (siehe [33]). Eine Alternative sind Padé-Approximation mit dem Skalierung-Algorithmus aus [18]. Die Verfahren aus [19], [20] basieren allerdings auf der Auswertung von $\varphi\left(\tau f'(v)\right)$, wobei $u'(t) = f(u(t))$ die zu lösende Gleichung ist und $\tau > 0$ linear von der Schrittweite abhängt. Die Auswertung von

$$\varphi(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+1)!} = \begin{cases} \frac{e^z - 1}{z} & z \neq 0 \\ 1 & z = 0 \end{cases}$$

muss daher in jedem Schritt neu erfolgen. Bei Verwendung von Padé-Approximationen ist dies sehr aufwendig.

Inhalt des dritten Kapitels sind die exponentiellen Runge-Kutta-Verfahren für Evolutionsgleichungen der Form

$$u'(t) + Au(t) = g(t, u(t))$$

aus der Arbeit [21] von M. Hochbruck und A. Ostermann. Die Grundidee der exponentiellen Integratoren bleibt erhalten. Die Differentialgleichung ist in einen linearen, autonomen und in einen nicht-linearen Teil aufgeteilt. Die Verfahren erhält man wiederum auf Basis der Formel zur Variation der Konstanten. Der wesentliche Unterschied besteht allerdings darin, dass die Aufteilung in linearen und nicht-linearen Teil in jedem Zeitschritt dieselbe ist. Im Fall einer konstanten Schrittweite bedeutet dies, dass die Auswertung der Operatoren nur einmal durchgeführt werden muss.

Die Verwendung solcher Verfahren ist sinnvoll, wenn Spektralmethoden verwendet werden, die zu kleinen, vollbesetzten und unsymmetrischen oder zu diagonalen Matrizen führen. Beispiele, in denen exponentielle Integratoren bessere Ergebnisse als übliche Verfahren liefern, finden sich unter anderem in [26] und [30].

Die Konsistenz- und Konvergenzanalyse der exponentiellen Runge-Kutta-Verfahren wird im abstrakten Rahmen einer Evolutionsgleichung im Banachraum durchgeführt, so dass sich die Ergebnisse auf partielle Differentialgleichungen übertragen lassen. Die Definition der Phi-Funktionen basiert dabei auf der Theorie sektorieller Operatoren und analytischer

Halbgruppen aus [14]. Die Konvergenzanalyse beschränkt sich allerdings auf den zeitlichen Diskretisierungsfehler. Der zusätzliche Fehler der Raumdiskretisierung wird nicht betrachtet.

Gegenstand des vierten Kapitels ist das numerische Verhalten der exponentiellen Integratoren angewandt auf dünnbesetzte, hochdimensionale Systeme am Beispiel der Nagumo-Gleichung

$$U_t = U_{xx} + U(1-U)(U-\alpha) + \Phi(x,t),$$

wobei Standard-Finite-Differenzen zur Raumdiskretisierung verwendet werden.

Abgesehen von der Schwierigkeit, die passende Größe der Krylow-Unterräume zu wählen, liefern die Verfahren des ersten Kapitels mit Krylow-Unterraum-Approximation verhältnismäßig gute Ergebnisse. Die Verfahren des dritten Kapitels mit Padé-Approximation sind hingegen nicht konkurrenzfähig. Der Einwand, dass Spektralmethoden statt Differenzenapproximationen zu besseren Ergebnissen führen würden, ist zwar berechtigt, dennoch bleibt festzustellen, dass Padé-Approximationen im Falle hochdimensionaler Systeme zu aufwendig sind.

Die im vierten Kapitel verwendete und auf den ersten Blick sehr sinnvolle Idee, in diesem Fall auf die Verfahren aus [19], [20] zurückzugreifen, ist jedoch mit einem Problem behaftet. Die Konstruktion dieser Verfahren basiert auf den klassischen Ordnungsbedingungen für gewöhnliche Differentialgleichungen. Im Fall partieller Differentialgleichungen ist daher im Allgemeinen eine Ordnungsreduktion zu erwarten. Nach [23, Example 2.25.] kann exp4 als exponentielles Rosenbrock-Verfahren (siehe [24]) angesehen werden. Die zugehörigen Ordnungsbedingungen zeigen, dass es im abstrakten Rahmen des dritten Kapitels dieser Diplomarbeit höchstens konsistent der Ordnung 2 sein kann. Dabei liegt sogar die Vermutung nahe, dass dies für die gesamte Verfahrensklasse gilt.

In [1, Discussion] wurde bereits die Möglichkeit thematisiert, die exponentiellen Runge-Kutta-Verfahren aus [21] mit variablen Schrittweiten zu implementieren. Bei Verwendung der Padé-Approximation ist dies nicht sinnvoll, da diese in jedem Schritt neu berechnet werden müsste, bei Verwendung von Krylow-Unterraum-Approximationen hingegen schon. Eine wesentliche Frage ist daher, ob sich die Ergebnisse aus [21] auf den Fall variabler Schrittweiten übertragen lassen.

Im letzten Teil des dritten Kapitels wird unter Voraussetzung einer Schrittweitenbedingung der Form

$$\frac{h_j}{h_{j+1}} \leq C$$

Konvergenz der Ordnung 1 für das Nørsett–Euler-Verfahren im Fall variabler Schrittweiten gezeigt.

Der Inhalt dieser Diplomarbeit ist auf explizite Einschrittverfahren beschränkt. In Analogie zu den Runge-Kutta-Verfahren existieren implizite exponentielle Integratoren, die in [22] betrachtet werden. Die Idee von S. P. Nørsett, explizite Adams-Verfahren zu modifizieren, um A-Stabilität zu erreichen, führte zu den exponentiellen Mehrschrittverfahren. Außerdem sei noch bemerkt, dass es neben den exponentiellen Integratoren für parabolische Gleichungen mit glatten Lösungen spezielle Verfahren für hochoszillatorische Probleme gibt. Für einen größeren Überblick und eine Klassifizierung der exponentiellen Integratoren siehe auch [23].

Ich möchte mich bei Prof. Dr. Wolf-Jürgen Beyn für die Betreuung dieser Arbeit bedanken. Außerdem danke ich meinen Eltern, die mir dieses Studium ermöglicht haben.

1. Exponentielle Integratoren für steife Differentialgleichungen

1.1. Die Verfahren

In diesem Kapitel, das im Wesentlichen auf der Arbeit [20] von M. Hochbruck, C. Lubich und H. Selhofer basiert, wird eine Klasse von Verfahren zur numerischen Lösung autonomer Anfangswertaufgaben der Art

$$\begin{cases} u'(t) = f(u(t)), \\ u(t_0) = u_0 \end{cases} \tag{1.1}$$

mit $t_0 \in \mathbb{K},\ u_0 \in \mathbb{K}^N$ sowie

$$\begin{aligned} f : \mathcal{D} &\longrightarrow \mathbb{K}^N, \\ v &\longmapsto f(v) \end{aligned}$$

betrachtet, wobei $\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$, $\mathcal{D} \subseteq \mathbb{K}^N$ offen und f hinreichend glatt ist.

Die Verfahren werden also für den autonomen Fall konzipiert. Eine nicht-autonome Differentialgleichung kann durch Hinzunehmen der Gleichung $t' = 1$ autonomisiert werden. Siehe hierzu [15, 2.2].

Für gegebenes $t_E > t_0$ lässt sich die exakte Lösung von (1.1)

$$\begin{aligned} u : [t_0, t_E] &\longrightarrow \mathbb{K}^N, \\ t &\longmapsto u(t) \end{aligned}$$

mittels eines Einschrittverfahrens

$$u_{n+1} = u_n + hV(u_n, h)$$

auf einem Gitter $\Omega_h = \{t_j \in \mathbb{R} :\ \ t_j = t_0 + jh,\ j \in \mathbb{N}_0,\ t_j \leq t_E\}$ numerisch approximieren. Dabei ist

$$\begin{aligned} V : \mathbb{K}^N \times (0, \infty) &\to \mathbb{K}^N, \\ (v, h) &\mapsto V(v, h) \end{aligned}$$

die Verfahrensfunktion und $h > 0$ die Schrittweite. In diesem Kapitel wird die Schrittweite der Einfachheit halber als konstant angenommen, wenngleich Verfahren mit Schrittweitensteuerung häufig bessere Resultate liefern. In Kapitel 3 hingegen wird der nichtäquidistante Fall für ein einfaches Verfahren explizit betrachtet.

Neben dem Aufwand ist zur Charakterisierung eines Verfahrens von entscheidender Bedeutung, wie gut $u(t_n)$ mit $t_n \in \Omega_h$ durch u_n approximiert wird.

Zur Analyse des Konvergenzfehlers erweist es sich hierbei als nützlich, zunächst den Konsistenzfehler

$$\tau_h(t_n) = \frac{1}{h}\Big(u(t_{n+1}) - u(t_n)\Big) - V\big(u(t_n), h\big)$$

zu betrachten. Ein Verfahren heißt dabei konsistent der Ordnung p, wenn

$$\sup_{t_j \in \Omega_h} \|\tau_h(t_j)\| = \mathcal{O}(h^p)$$

gilt. Die numerische Lösung wird, wie bereits oben erwähnt, nur auf endlichem Gitter konstruiert. Dieses liegt im Intervall $[t_0, t_E]$, auf dem die exakte Lösung definiert ist. Der obige Konsistenzfehler muss also nur auf endlichem Intervall gleichmäßig wie $\mathcal{O}(h^p)$ gegen null konvergieren.

1.1.1. Exponentiell angepasstes Euler-Verfahren

Die Frage ist nun, wie man ein geeignetes Verfahren findet. Eine Idee basiert auf Linearisierung der rechten Seite der Differentialgleichung.

Sei t_n fest gewählt. Da f und u als hinreichend glatt angenommen werden, sind die auftretenden Ableitungen beschränkt, so dass für t mit $0 \leq t - t_n \leq h$ mittels Taylorentwicklung von f um $u(t_n)$

$$f\big(u(t)\big) = f\big(u(t_n)\big) + f'\big(u(t_n)\big)\Big(u(t) - u(t_n)\Big) + \mathcal{O}(h^2) \tag{1.2}$$

gilt.

Setzt man

$$\begin{aligned} b_n(t) &:= f\big(u(t)\big) - f'\big(u(t_n)\big)u(t), \\ A_n &:= f'\big(u(t_n)\big), \end{aligned}$$

dann erhält man

$$\begin{aligned} f\big(u(t)\big) &= f'\big(u(t_n)\big)u(t) + \Big(f\big(u(t)\big) - f'\big(u(t_n)\big)u(t)\Big) \\ &= A_n u(t) + b_n(t) \end{aligned}$$

und analog zur Linearisierung in (1.2) lässt sich nun $f\big(u(t)\big)$ als

$$\begin{aligned} f\big(u(t)\big) &= f\big(u(t_n)\big) + f'\big(u(t_n)\big)\Big(u(t) - u(t_n)\Big) + \mathcal{O}(h^2) \\ &= A_n u(t) + b_n(t_n) + \mathcal{O}(h^2) \end{aligned}$$

schreiben.

Entsprechend ist nun u die Lösung der nicht-autonomen linearen Differentialgleichung

$$u'(t) = A_n u(t) + b_n(t)$$

mit Anfangswert $u(t_n)$. An dieser Stelle sei bemerkt, dass im dritten Kapitel die semilinearen parabolischen Gleichungen $u'(t) = Au(t) + g(t, u(t))$ für die Konvergenzanalyse auf dieselbe Form gebracht werden, indem man $\tilde{g}(t) := g(t, u(t))$ setzt.

Mit Hilfe der Variation der Konstanten lässt sich nun $u(t_{n+1})$ als

$$u(t_{n+1}) = e^{(t_{n+1}-t_n)A_n}u(t_n) + \int_{t_n}^{t_{n+1}} e^{(t_{n+1}-t)A_n}b_n(t)\ dt \tag{1.3}$$

darstellen. Der Zusammenhang zur obigen Linearisierung besteht hier darin, dass $b_n(t)$ näherungsweise durch $b_n(t_n)$ ersetzt werden kann, wobei

$$b_n(t) - b_n(t_n) = \mathcal{O}(h^2)$$

gilt. Hieraus folgt mittels Standardabschätzung des Integrals

$$\begin{aligned}&\left\| \int_{t_n}^{t_{n+1}} e^{(t_{n+1}-t)A_n}\Big(b_n(t) - b_n(t_n)\Big)\ dt \right\| \\ &\leq h \cdot \sup_{0\leq\tau\leq h}\left\|e^{\tau A_n}\right\| \cdot \sup_{t_n\leq t\leq t_{n+1}} \|b_n(t) - b_n(t_n)\| = \mathcal{O}(h^3).\end{aligned}$$

Indem nun in (1.3) $b_n(t)$ als $b_n(t_n) + \big(b_n(t) - b_n(t_n)\big)$ geschrieben wird, erhält man daraus

$$\begin{aligned}u(t_{n+1}) &= e^{(t_{n+1}-t_n)A_n}u(t_n) + \int_{t_n}^{t_{n+1}} e^{(t_{n+1}-t)A_n}b_n(t_n)\ dt + \mathcal{O}(h^3) \\ &= e^{(t_{n+1}-t_n)A_n}u(t_n) + \sum_{k=0}^{\infty}\int_{t_n}^{t_{n+1}} \frac{(t_{n+1}-t)^k A_n^k}{k!}\ b_n(t_n)\ dt + \mathcal{O}(h^3) \\ &= e^{hA_n}u(t_n) + \sum_{k=0}^{\infty}\left[-\frac{(t_{n+1}-t)^{k+1}A_n^k}{(k+1)!}\ b_n(t_n)\right]_{t_n}^{t_{n+1}} + \mathcal{O}(h^3) \\ &= \sum_{k=0}^{\infty}\frac{h^k A_n^k}{k!}\ u(t_n) + \sum_{k=0}^{\infty}\frac{h^{k+1}A_n^k}{(k+1)!}\ b_n(t_n) + \mathcal{O}(h^3) \\ &= u(t_n) + hA_n\sum_{k=1}^{\infty}\frac{h^{k-1}A_n^{k-1}}{k!}\ u(t_n) + h\sum_{k=0}^{\infty}\frac{h^k A_n^k}{(k+1)!}\ b_n(t_n) + \mathcal{O}(h^3) \\ &= u(t_n) + h\sum_{k=0}^{\infty}\frac{(hA_n)^k}{(k+1)!}\Big(A_n u(t_n) + b_n(t_n)\Big) + \mathcal{O}(h^3).\end{aligned}$$

Definiert man schließlich

$$\begin{aligned}&\varphi : \mathbb{C} \longrightarrow \mathbb{C}, \\ &\varphi(z) = \sum_{k=0}^{\infty}\frac{z^k}{(k+1)!} = \begin{cases}\frac{e^z-1}{z} & z \neq 0 \\ 1 & z = 0,\end{cases}\end{aligned} \tag{1.4}$$

dann lässt sich dies als

$$u(t_{n+1}) = u(t_n) + h\varphi(hA_n)f(u(t_n)) + \mathcal{O}(h^3) \tag{1.5}$$

schreiben.

Natürlich ist es nun naheliegend, auf Basis von (1.5) ein Verfahren zu konstruieren, bei dem $u(t_n)$ durch u_n approximiert wird.

Das **exponentiell angepasste Euler-Verfahren** für (1.1) hat die Form

$$u_{n+1} = u_n + h\varphi(hf'(u_n))f(u_n),$$

wobei φ wie in (1.4) definiert ist.

Der Unterschied zum klassischen (expliziten) Euler-Verfahren

$$u_{n+1} = u_n + hf(u_n)$$

besteht hierbei lediglich im zusätzlichen Faktor $\varphi(hf'(u_n))$.

Aus der Darstellung (1.5) erhält man sofort die Konsistenzordnung für das exponentiell angepasste Euler-Verfahren. Für den Spezialfall einer linearen Differentialgleichung mit konstanten Koeffizienten

$$\begin{cases} u'(t) = \mathcal{A}u(t) + \mathcal{B}, \\ u(t_0) = u_0, \end{cases} \tag{1.6}$$

wobei $\mathcal{A} \in \mathbb{K}^{N\times N}$ und $\mathcal{B} \in \mathbb{K}^N$ ist, stimmen die vorausgegangenen Rechnungen auch ohne die $\mathcal{O}(h^3)$ Terme. Dies wird im folgenden Satz zusammengefasst.

Satz 1.1.1
Das exponentiell angepasste Euler-Verfahren ist konsistent der Ordnung 2 und liefert die exakte Lösung für lineare Differentialgleichungen mit konstanten Koeffizienten (1.6).

Beweis:
Für den Konsistenzfehler gilt nach (1.5)

$$\begin{aligned} \tau_h(t_n) &= \frac{1}{h}\Big(u(t_{n+1}) - u(t_n)\Big) - V\big(u(t_n), h\big) \\ &= \frac{1}{h}\Big(u(t_{n+1}) - u(t_n)\Big) - \varphi(hf'(u_n))f(u_n) \\ &= \mathcal{O}(h^2), \end{aligned}$$

was nach Definition die behauptete Konsistenzordnung liefert.

Ist f linear, so ist die Formel

$$u(t_{n+1}) = u(t_n) + h\varphi(h\mathcal{A})f(u(t_n))$$

exakt erfüllt. Aus $u_n = u(t_n)$ folgt dann $u_{n+1} = u(t_{n+1})$, so dass man zusammen mit $u_0 = u(t_0)$ induktiv die Behauptung erhält.

■

In Hinblick auf Kapitel 3 sei noch bemerkt, dass das obige Vorgehen voraussetzt, dass die Operatoren A_n beschränkt sind. Im Rahmen partieller Differentialgleichungen können wir dies allerdings nicht mehr annehmen.
Mittels Halbgruppen kann man allerdings die φ-Funktion auch für eine Klasse unbeschränkter Operatoren definieren. Die Zusammenhänge der beiden Definitionen werden im Anhang erläutert.

1.1.2. Weitere Verfahren

Um vom klassischen Euler-Verfahren ausgehend höhere Konsistenz- und damit auch Konvergenzordnungen zu erhalten, besteht eine Idee darin, in jedem Schritt die Funktion f an mehreren Stellen auszuwerten. Das Resultat sind die Runge-Kutta-Verfahren

$$\begin{aligned} u_{n+1} &= u_n + h\sum_{i=1}^{s} b_i k_i(u_n, h), \\ k_i(v,h) &= f\left(v + h\sum_{j=1}^{i-1} \alpha_{ij} k_j(v,h)\right), \quad i = 1, ..., s, \end{aligned}$$

wobei die Einträge der Runge-Kutta-Matrix $(\alpha_{ij})_{ij} \in \mathbb{R}^{s\times s}$ und die Gewichte $b_i \in \mathbb{R}$ mit $i = 1, ..., s$ die Koeffizienten des Verfahrens sind.

Diese Idee lässt sich natürlich auch auf exponentielle Integratoren übertragen, indem in Analogie zum exponentiell angepassten Euler-Verfahren bei einem gegebenen Runge-Kutta-Verfahren die Stufenwerte $k_i(u_n, h)$ mit $\varphi(hf'(u_n))$ oder allgemeiner mit $\varphi(\gamma hf'(u_n))$ multipliziert werden. Dabei ist $\gamma \in \mathbb{R}$.

Eine weitere Möglichkeit besteht darin, die ursprüngliche Differentialgleichung (1.1) wie schon bei der Herleitung des exponentiell angepassten Euler-Verfahrens als

$$u'(t) = f'(u(t_n))u(t) + \Big(f(u(t)) - f'(u(t_n))u(t)\Big)$$

zu schreiben und analog zur Konstruktion der Rosenbrock-Wanner-Verfahren ([12, IV.7]) auf die beiden Summanden Verfahren mit verschiedenen Runge-Kutta-Matrizen anzuwenden. In diesem Fall sind das $(\beta_{ij})_{ij}$ für den ersten und $(\alpha_{ij})_{ij}$ für den zweiten Term.

Dies führt dann zu den Stufenwerten

$$\begin{aligned} k_i(v,h) = \varphi(\gamma hf'(v)) \Bigg(& hf'(v)\sum_{j=1}^{i-1} \beta_{ij} k_j(v,h) \\ & + f\left(v + h\sum_{j=1}^{i-1} \alpha_{ij} k_j(v,h)\right) - hf'(v)\sum_{j=1}^{i-1} \alpha_{ij} k_j(v,h)\Bigg), \end{aligned}$$

so dass sich mit

$$\gamma_{ij} = \beta_{ij} - \alpha_{ij}$$

das entstandene Verfahren als

$$\begin{aligned} u_{n+1} &= u_n + h\sum_{i=1}^{s} b_i k_i(u_n, h), \\ k_i(v,h) &= \varphi(\gamma hf'(v)) \left(f\left(v + h\sum_{j=1}^{i-1} \alpha_{ij} k_j(v,h)\right) + hf'(v)\sum_{j=1}^{i-1} \gamma_{ij} k_j(v,h)\right) \end{aligned} \tag{1.7}$$

mit $i = 1, ..., s$ schreiben lässt.

Für gegebenes $\gamma \in \mathbb{R}$, zwei Matrizen $(\gamma_{ij})_{ij}, (\alpha_{ij})_{ij} \in \mathbb{R}^{s\times s}$ und die Gewichte b_i, wobei $i = 1, ..., s$ ist, liefert also (1.7) einen **exponentiellen Integrator** für die Anfangswertaufgabe (1.1).

Bemerkung:
Durch Wahl der Koeffizienten $(\gamma_{ij})_{ij} \equiv 0$ und $\gamma = 0$ erhält man ein Runge-Kutta-Verfahren. Insbesondere lässt sich die Verfahrensfunktion von (1.7) analog zu den Runge-Kutta-Verfahren als

$$V(v,h) = \sum_{i=1}^{s} b_i k_i(v,h)$$

schreiben.

Grundvoraussetzung:
Im weiteren Verlauf wird angenommen, dass $(\alpha_{ij})_{ij}$ und $(\gamma_{ij})_{ij}$ echte untere Dreiecksmatrizen sind, d.h. $\alpha_{ij} = \gamma_{ij} = 0$ für $i \leq j$. Die bei den Stufenwerten auftretenden Summen könnte man also auch bis s laufen lassen, wobei einige Summanden identisch 0 sind.

Wegen des Terms $\varphi(\gamma h f'(u_n))$ ist die Auswertung der k_i verglichen mit Runge-Kutta-Verfahren sehr aufwendig. Angesichts dieser Tatsache erscheint es ohnehin nicht besonders sinnvoll, implizite Verfahren von (1.7) zu betrachten. Daher ist die Beschränkung auf explizite Verfahren keine wirkliche Einschränkung. Implizite Verfahren speziell für semilineare Gleichungen finden sich in [22].

1.2. Ordnungsbedingungen

Das exponentiell angepasste Euler-Verfahren ist lediglich konsistent der Ordnung 2. Da die obige Klasse von Einschrittverfahren eingeführt wurde, um höhere Ordnungen zu erreichen, muss das Ziel nun darin bestehen, Ordnungsbedingungen in Abhängigkeit der Koeffizienten zu finden, so dass auf dieser Grundlage bessere Verfahren konstruiert werden können. Der folgende Satz aus der Numerik gewöhnlicher Differentialgleichungen ist dabei hilfreich.

Satz 1.2.1
Sei $U \subseteq \mathbb{R}^{N+1}$ eine offene Umgebung des Graphen der exakten Lösung

$$u : [t_0, t_E] \longrightarrow \mathbb{R}^N.$$

Seien $f \in \mathcal{C}^p(U; \mathbb{R}^N)$ und $V \in \mathcal{C}^p(U \times (0, h_0); \mathbb{R}^N)$ für ein $h_0 > 0$. Dann ist das Einschrittverfahren mit Verfahrensfunktion

$$\begin{aligned} V : \mathbb{R}^N \times (0,\infty) &\to \mathbb{R}^N, \\ (v,h) &\mapsto V(v,h) \end{aligned}$$

genau dann konsistent der Ordnung p, wenn für $q = 0, ..., p-1$

$$(q+1)\frac{\partial^q V}{\partial h^q}(u(t),0) = \frac{d^q}{dt^q}\Big[f(u(t))\Big]$$

gilt.

Die Konsistenzbedingungen kann man aus der Taylorentwicklung von exakter und numerischer Lösung (siehe hierzu beispielsweise [10],[41], [11]) ablesen.

Die Aussage ist direkt auf den Fall $\mathbb{K} = \mathbb{C}$ übertragbar. Falls f nicht holomorph ist, muss (1.1) ohnehin in ein reelles, $2N$-dimensionales System umgewandelt werden, um (1.7) anwenden zu können.

Der Satz 1.2.1 soll nun auf die Verfahren in (1.7) angewandt werden. Hierzu ist es zunächst zweckmäßig, φ genauer zu betrachten. Die Funktion lässt sich als Potenzreihe

$$\varphi(z) = \sum_{k=1}^{\infty} \frac{z^{k-1}}{k!}$$

um 0 entwickeln. Hieran kann man $\varphi(0) = 1$ und $\varphi'(0) = \frac{1}{2}$ sofort ablesen.

Für sehr kleines p lassen sich die Ordnungsbedingungen nun recht einfach finden.

Ordnung 1:
Im Fall $p = 1$ muss nur $q = 0$ betrachtet werden, so dass es genügt

$$V\big(u(t),0\big) = \sum_{i=1}^{s} b_i k_i\big(u(t),0\big)$$

zu berechnen und gleich $f\big(u(t)\big)$ zu setzen. Auswerten der k_i aus (1.7) an der Stelle $\big(u(t),0\big)$ liefert dabei

$$k_i\big(u(t),0\big) = \varphi(0) f\big(u(t)\big),$$

so dass man

$$V\big(u(t),0\big) = \sum_{i=1}^{s} b_i f\big(u(t)\big)$$

erhält. Die Konsistenzbedingung für die Ordnung $p = 1$ lautet daher

$$\sum_{i=1}^{s} b_i = 1.$$

Ordnung 2:
Unter der Annahme, dass für $q = 0$ die Gleichheit erfüllt ist, wird nun $q = 1$ betrachtet. Diesmal wird die partielle Ableitung der Verfahrensfunktion nach h benötigt. Differenziert man zunächst die k_i nach h, so erhält man

$$\begin{aligned}\frac{\partial k_i}{\partial h}(v,h) = {} & \varphi'(\gamma h f'(v))\gamma f'(v)\left(f\left(v + h\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\right) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)\right) \\ & + \varphi\big(\gamma h f'(v)\big)\ f'\left(v + h\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\right)\sum_{j=1}^{i-1}\alpha_{ij}\left(k_j(v,h) + h\frac{\partial k_j}{\partial h}(v,h)\right) \\ & + \varphi\big(\gamma h f'(v)\big)\ f'(v)\sum_{j=1}^{i-1}\gamma_{ij}\left(k_j(v,h) + h\frac{\partial k_j}{\partial h}(v,h)\right).\end{aligned}$$

Ausgewertet an der Stelle $\big(u(t),0\big)$ liefert dies

$$\frac{\partial k_i}{\partial h}\big(u(t),0\big) = \varphi'(0)\gamma f'\big(u(t)\big)f\big(u(t)\big) + \varphi(0)f'\big(u(t)\big)\sum_{j=1}^{i-1}\underbrace{(\alpha_{ij} + \gamma_{ij})}_{=\beta_{ij}}\ k_j\big(u(t),0\big)$$

und folglich ist

$$\frac{\partial V}{\partial h}(u(t),0) = \sum_{i=1}^{s} b_i \left(\frac{1}{2}\gamma + \sum_{j=1}^{i-1} \beta_{ij} \right) f'(u(t)) f(u(t)).$$

Da andererseits

$$\frac{d}{dt}\Big[f(u(t)) \Big] = f'(u(t)) f(u(t))$$

gilt, folgt als Ordnungsbedingung

$$2\Bigg(\underbrace{\sum_{i=1}^{s} b_i}_{=1} \frac{1}{2}\gamma + \sum_{i=1}^{s} b_i \sum_{j=1}^{i-1} \beta_{ij} \Bigg) = 1,$$

was sich zu

$$\sum_{i=1}^{s} b_i \sum_{j=1}^{i-1} \beta_{ij} = \frac{1}{2}(1-\gamma)$$

umformen lässt.

1.2.1. Linearer Fall

Bei höheren Ableitungen von $f(u(t))$ nach t tauchen auf Grund der Produktregel mehrere Summanden auf, was das Aufstellen der Ordnungsbedingungen erheblich erschwert. Im Fall einer linearen Differentialgleichung mit konstanten Koeffizienten ist die Situation allerdings auch für hohe Ordnungen noch recht einfach. Dies liegt insbesondere daran, dass die auftretenden Ableitungen eine einfache Form besitzen. Ist nämlich $f(u(t)) = \mathcal{A}u(t) + \mathcal{B}$, dann gilt

$$\frac{d^q}{dt^q}\Big[f(u(t)) \Big] = \mathcal{A}^q f(u(t)),$$

da höhere Ableitungen von f identisch 0 sind.

Um den Satz über Ordnungsbedingungen für diese spezielle Klasse von Anfangswertproblemen beweisen zu können, wird das folgende Lemma benötigt. Wesentlicher Bestandteil dieses Hilfssatzes ist das Einschrittverfahren

$$\begin{aligned} u_{n+1} &= y_m(u_n, h), \\ y_i(v,h) &= y_{i-1}(v,h) + h\varphi\Big(hf'(y_{i-1}(v,h)) \Big) f(y_{i-1}(v,h)), \quad i = 1, ..., m, \\ y_0(v,h) &= v, \end{aligned} \tag{1.8}$$

bei dem ein Schritt der Hintereinanderausführung von m Schritten des exponentiell angepassten Euler-Verfahrens entspricht.

Lemma 1.2.2
Sei $m \in \mathbb{N}_1$. Das Verfahren (1.8) mit Schrittweite $\tau > 0$ liefert für lineare Anfangswertaufgaben mit konstanten Koeffizienten (1.6) dieselbe numerische Lösung wie das Verfahren aus (1.7) mit Schrittweite $h = m\tau$, Stufenzahl $s = m$ und den Koeffizienten $\gamma = \frac{1}{m}$, $b_i = \frac{1}{m}$, $\alpha_{ij} = \frac{1}{m}$ für $i > j$, $\alpha_{ij} = 0$ für $i \leq j$ sowie $(\gamma_{ij})_{ij} \equiv 0$.

Beweis:
Zur Unterscheidung werden im Folgenden die durch (1.8) gegebenen Näherungen mit $\hat{u}_n$ bezeichnet, während (1.7) weiterhin u_n liefert.

Definiere für $i = 1, .., s$

$$\ell_i(v,\tau) := \varphi(\tau\mathcal{A}) f\left(v + \tau \sum_{j=1}^{i-1} \ell_j(v,\tau)\right).$$

Dies wird nun verwendet, um das Verfahren (1.8) geeignet umzuschreiben. Mittels Induktion nach k wird zunächst

$$y_k(v,\tau) = v + \tau \sum_{i=1}^{k} \ell_i(v,\tau) \tag{1.9}$$

gezeigt.

Induktionsanfang $k = 1$:
Im Fall $k = 1$ folgt die Behauptung aus

$$\begin{aligned} y_1(v,\tau) &= y_0(v,\tau) + \tau\varphi(\tau\mathcal{A}) f\Big(y_0(v,\tau)\Big) = v + \tau\varphi(\tau\mathcal{A}) f(v) \\ &= v + \tau \sum_{i=1}^{1} \ell_i(v,\tau). \end{aligned}$$

Induktionsschritt $k \to (k+1)$:
Nimmt man an, dass (1.9) für k nach Induktionsvoraussetzung gilt, so erhält man

$$\begin{aligned} y_{k+1}(v,\tau) &= y_k(v,\tau) + \tau\varphi(\tau\mathcal{A}) f\Big(y_k(v,\tau)\Big) \\ &\overset{IV}{=} v + \tau \sum_{i=1}^{k} \ell_i(v,\tau) + \tau\varphi(\tau\mathcal{A}) f\left(v + \tau \sum_{i=1}^{k} \ell_i(v,\tau)\right) \\ &= v + \tau \sum_{i=1}^{k} \ell_i(v,\tau) + \tau\ell_{k+1}(v,\tau) \\ &= v + \tau \sum_{i=1}^{k+1} \ell_i(v,\tau). \end{aligned}$$

Also gilt (1.9) für alle $k \in \mathbb{N}_1$.

Nun gilt es noch, das Verfahren aus (1.7) mit den im Satz genannten Parametern geeignet umzuformen. Die Stufenwerte lassen sich dabei als

$$k_i(v,h) = \varphi\Big(\tfrac{1}{m} h\mathcal{A}\Big) f\left(v + h \sum_{j=1}^{i-1} \tfrac{1}{m} k_j(v,h)\right)$$

schreiben. Durch Induktion nach i kann man daher

$$\ell_i(v,\tau) = k_i(v,h) \tag{1.10}$$

zeigen.

Induktionsanfang $i = 1$:

$$\ell_1(v,\tau) = \varphi(\tau\mathcal{A})f(v) = \varphi\Big(\tfrac{1}{m}h\mathcal{A}\Big)f(v) = k_1(v,h)$$

Induktionsschritt $(1,...,i-1) \to i \le s$:

$$\begin{aligned}\ell_i(v,\tau) &= \varphi(\tau\mathcal{A})f\left(v+\tau\sum_{j=1}^{i-1}\ell_j(v,\tau)\right) \overset{IV}{=} \varphi(\tau\mathcal{A})f\left(v+\tau\sum_{j=1}^{i-1}k_j(v,h)\right)\\ &= \varphi\Big(\tfrac{1}{m}h\mathcal{A}\Big)f\left(v+h\sum_{j=1}^{i-1}\tfrac{1}{m}k_j(v,h)\right) = k_i(v,h)\end{aligned}$$

Also gilt (1.10) für $i = 1,...,s$, woraus sich sofort

$$u_{n+1} = u_n + h\sum_{i=1}^{m}\frac{1}{m}k_i(u_n,h) = u_n + \tau\sum_{i=1}^{m}\ell_i(u_n,\tau)$$

ergibt.

Die folgende Induktion über n schließt nun den Beweis ab.

Induktionsanfang $n = 0$:
Nach Konstruktion ist $\hat{u}_0 = u_0$, so dass die Aussage für $n = 0$ gilt.

Induktionsschritt $n \to (n+1)$:
Nach Induktionsvoraussetzung gilt nun $\hat{u}_n = u_n$. Dann aber folgt

$$\hat{u}_{n+1} = y_m(\hat{u}_n,\tau) \overset{IV}{=} y_m(u_n,\tau) = u_n + \tau\sum_{i=1}^{m}\ell_i(u_n,\tau) = u_{n+1},$$

was zu zeigen war.

■

Satz 1.2.3
Ein exponentielles Einschrittverfahren wie in (1.7) approximiert die Lösung linearer Differentialgleichungen mit konstanten Koeffizienten (1.6) exakt, falls für alle $q \in \mathbb{N}_0$

$$\sum_{\substack{(i_0,...,i_q)\in\mathcal{J}_{q+1}\\ i_q<...<i_0}} b_{i_0}\prod_{j=0}^{q-1}\beta_{i_j,i_{j+1}} = \frac{1}{q+1}\prod_{j=0}^{q-1}\left(\frac{1}{q-j}-\gamma\right)$$

gilt, wobei $\mathcal{J}_j = \{0,...,s\}^j$ *ist.*

Beweis:
Die Beweisidee ist, die Formel aus Satz 1.2.1 für beliebig großes q nachzurechnen, also Konsistenz beliebig großer Ordnung zu zeigen. Da sowohl die exakte wie auch die numerische Lösung analytische Funktionen bzgl. h sind, folgt dann die Behauptung des Satzes.

Angewandt auf die Anfangswertaufgabe aus (1.6)

$$\begin{cases} u'(t) = \mathcal{A}u(t) + \mathcal{B}, \\ u(t_0) = u_0, \end{cases}$$

lässt sich die Verfahrensfunktion von (1.7) als

$$V(v,h) = \sum_{i_0=1}^{s} b_{i_0} k_{i_0}(v,h)$$

mit

$$\begin{aligned} k_{i_0}(v,h) &= \varphi(\gamma h\mathcal{A})\left(\mathcal{A}v + h\mathcal{A}\sum_{i_1=1}^{i_0-1}\alpha_{i_0,i_1}k_{i_1}(v,h) + \mathcal{B} + h\mathcal{A}\sum_{i_1=1}^{i_0-1}\gamma_{i_0,i_1}k_{i_1}(v,h)\right) \\ &= \varphi(\gamma h\mathcal{A})\left(f(v) + h\mathcal{A}\sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}k_{i_1}(v,h)\right) \end{aligned}$$

schreiben.

Der erste Teil des Beweises besteht nun darin, per Induktion über q zu zeigen, dass es für $m = 0, ..., q$ Polynome $Q_{q,m}(\gamma)$ mit Grad kleiner gleich $q - m$ gibt, so dass

$$\frac{\partial^q k_{i_0}}{\partial h^q}(u(t),0) = \left(Q_{q,0}(\gamma) + \sum_{m=1}^{q} Q_{q,m}(\gamma) \sum_{\substack{(i_1,\ldots,i_m)\in\mathcal{J}_m \\ i_m<\ldots<i_1<i_0}} \prod_{j=0}^{m-1}\beta_{i_j,i_{j+1}}\right)\mathcal{A}^q f(u(t))$$

gilt.

Induktionsanfang $q = 0$:
Bei der expliziten Berechnung der Konsistenzbedingungen für Ordnung 1 wurde bereits

$$k_{i_0}(u(t),0) = f(u(t))$$

gezeigt. Das gesuchte Polynom ist $Q_{0,0} = 1$.

Induktionsschritt $(0, ..., q-1) \to q$:
In Hinblick auf die folgende Rechnung sei zunächst bemerkt, dass

$$\begin{aligned} \sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}\frac{\partial^{\ell-1}k_{i_1}}{\partial h^{\ell-1}}(u(t),0) &= \sum_{m=0}^{\ell-1} Q_{\ell-1,m}(\gamma) \sum_{\substack{(i_1,\ldots,i_{m+1})\in\mathcal{J}_{m+1} \\ i_{m+1}<\ldots<i_1<i_0}} \prod_{j=0}^{m}\beta_{i_j,i_{j+1}}\mathcal{A}^{\ell-1}f(u(t)) \\ &= \sum_{m=1}^{\ell} Q_{\ell-1,m-1}(\gamma) \sum_{\substack{(i_1,\ldots,i_m)\in\mathcal{J}_m \\ i_m<\ldots<i_1<i_0}} \prod_{j=0}^{m-1}\beta_{i_j,i_{j+1}}\mathcal{A}^{\ell-1}f(u(t)) \end{aligned}$$

für $\ell = 1, ..., q$ aus der Induktionsvoraussetzung folgt.

Nun definiert man für $i_0 = 1, ..., s$ Funktionen

$$\hat{k}_{i_0}(v,h) := f(v) + h\mathcal{A}\sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}k_{i_1}(v,h),$$

die für $\ell \geq 1$ nach der Leibnizregel Ableitungen der Form

$$\frac{\partial^\ell \hat{k}_{i_0}}{\partial h^\ell}(v,h) = \binom{\ell}{1}\mathcal{A}\sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}\frac{\partial^{\ell-1}k_{i_1}}{\partial h^{\ell-1}}(v,h) + \mathcal{O}(h) \tag{1.11}$$

besitzen. Dann aber gilt

$$\begin{aligned}
\frac{\partial^q k_{i_0}}{\partial h^q}(u(t),0) &= \sum_{\ell=0}^{q}\binom{q}{\ell}\frac{d^{q-\ell}}{dh^{q-\ell}}\Big[\varphi(\gamma h\mathcal{A})\Big]_{h=0}\frac{\partial^\ell \hat{k}_{i_0}}{\partial h^\ell}(u(t),0)\\
&\overset{(1.11)}{=}\sum_{\ell=1}^{q}\binom{q}{\ell}\frac{1}{(q-\ell+1)}(\gamma\mathcal{A})^{q-\ell}\ell\mathcal{A}\sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}\frac{\partial^{\ell-1}k_{i_1}}{\partial h^{\ell-1}}(u(t),0)\\
&\quad+\frac{1}{(q+1)}(\gamma\mathcal{A})^q f(u(t))\\
&= \sum_{\ell=1}^{q}\hat{Q}_{q,\ell}(\gamma)\mathcal{A}^{q-\ell+1}\sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}\frac{\partial^{\ell-1}k_{i_1}}{\partial h^{\ell-1}}(u(t),0)+\frac{1}{(q+1)}(\gamma\mathcal{A})^q f(u(t))\\
&\overset{IV}{=}\sum_{\ell=1}^{q}\hat{Q}_{q,\ell}(\gamma)\sum_{m=1}^{\ell}Q_{\ell-1,m-1}(\gamma)\sum_{\substack{(i_1,\ldots,i_m)\in\mathcal{J}_m\\ i_m<\ldots<i_1<i_0}}\prod_{j=0}^{m-1}\beta_{i_j,i_{j+1}}\mathcal{A}^q f(u(t))\\
&\quad+\frac{1}{(q+1)}\gamma^q\mathcal{A}^q f(u(t))\\
&= \left(Q_{q,0}(\gamma)+\sum_{m=1}^{q}Q_{q,m}(\gamma)\sum_{\substack{(i_1,\ldots,i_m)\in\mathcal{J}_m\\ i_m<\ldots<i_1<i_0}}\prod_{j=0}^{m-1}\beta_{i_j,i_{j+1}}\right)\mathcal{A}^q f(u(t)),
\end{aligned}$$

wobei

$$\begin{aligned}
Q_{q,0}(\gamma) &= \frac{1}{(q+1)}\gamma^q,\\
Q_{q,m}(\gamma) &= \sum_{\ell=m}^{q}\hat{Q}_{q,\ell}(\gamma)Q_{\ell-1,m-1}(\gamma)
\end{aligned}$$

für $m = 1, ..., q$ und

$$\hat{Q}_{q,\ell}(\gamma) = \binom{q}{\ell}\frac{\ell}{(q-\ell+1)}\gamma^{q-\ell} = \binom{q}{\ell-1}\gamma^{q-\ell}$$

für $\ell = 1, ..., q$ ist. Die Gradbedingungen lassen sich durch Zählen der Grade überprüfen. Die Notation ist so gewählt, dass $\deg(Q_{\ell,m}) = \ell - m$ ist. Außerdem sei noch darauf hingewiesen, dass stets $Q_{q,q} = 1$ gilt. Damit ist die Induktion abgeschlossen.

Der zweite Teil des Beweises besteht nun darin, mittels Induktion über q zu zeigen, dass es Polynome $P_q(\gamma)$ mit Grad kleiner gleich q gibt, so dass sich die Konsistenzbedingungen als

$$\sum_{\substack{(i_0,\ldots,i_q)\in\mathcal{J}_{q+1}\\ i_q<\ldots<i_0}} b_{i_0}\prod_{j=0}^{q-1}\beta_{i_j,i_{j+1}} = P_q(\gamma)$$

schreiben lassen.

Induktionsanfang $q = 0$:
Für $q = 0$ wurde die Bedingung bereits explizit berechnet. Sie lautet

$$\sum_{i=1}^{s} b_i = 1$$

und hat folglich die geforderte Form, wobei $P_q(\gamma) = 1$ ist.

Induktionsschritt $(0, ..., q-1) \to (q)$:
Die zusätzliche Bedingung lautet

$$(q+1)\frac{\partial^q V}{\partial h^q}(u(t), 0) = \mathcal{A}^q f(u(t)). \tag{1.12}$$

Da sich die linke Seite nach dem ersten Teil des Beweises mittels

$$\begin{aligned}\frac{\partial^q V}{\partial h^q}(u(t), 0) &= \sum_{i_0=1}^{s} b_{i_0} \frac{\partial^q k_{i_0}}{\partial h^q}(u(t), 0) \\ &= \sum_{m=0}^{q} Q_{q,m}(\gamma) \sum_{\substack{(i_0,...,i_m)\in\mathcal{J}_m \\ i_m<...<i_0}} b_{i_0} \prod_{j=0}^{m-1} \beta_{i_j,i_{j+1}} \mathcal{A}^q f(u(t))\end{aligned}$$

geeignet umschreiben lässt, ist

$$\sum_{\substack{(i_0,...,i_q)\in\mathcal{J}_{q+1} \\ i_q<...<i_0}} b_{i_0} \prod_{j=0}^{q-1} \beta_{i_j,i_{j+1}} \mathcal{A}^q f(u(t)) = P_q(\gamma) \mathcal{A}^q f(u(t))$$

äquivalent zu (1.12), wobei

$$P_q(\gamma) = \frac{1}{q+1} - \sum_{m=0}^{q-1} Q_{q,m}(\gamma) P_m(\gamma)$$

ist. Das Polynom P_q hat maximal Grad q. Damit ist die Aussage mittels Induktion gezeigt.

Im letzten Teil des Beweises ist nun noch zu verifizieren, dass tatsächlich

$$P_q(\gamma) = \frac{1}{q+1} \prod_{j=0}^{q-1} \left(\frac{1}{q-j} - \gamma \right)$$

gilt. Durch seine q Nullstellen $1, \frac{1}{2}, ..., \frac{1}{q}$ ist das Polynom auf der rechten Seite bis auf eine Konstante bestimmt. Um die Nullstellen von P_q zu finden, nutzt man die in (1.1.1) gezeigte Tatsache, dass das exponentiell angepasste Euler-Verfahren für lineare Differentialgleichungen mit konstanten Koeffizienten die exakte Lösung liefert. Da bei diesem Verfahren $\gamma = 1$ und $(\beta_{ij})_{ij} \equiv 0$ ist, folgt $P_q(1) = 0$ für $q \geq 1$.

Um für $q > 1$ die anderen Nullstellen zu finden, betrachtet man für $m = 2, ..., q$ die Verfahren mit den Parametern $s = m$, $\gamma = \frac{1}{m}$, $b_i = \frac{1}{m}$, $\alpha_{ij} = \frac{1}{m}$ für $i > j$, $\alpha_{ij} = 0$ für $i \leq j$ und $(\gamma_{ij})_{ij} \equiv 0$, die nach Lemma 1.2.2 der Hintereinanderausführung von exponentiell angepassten Euler-Schritten entsprechen und daher ebenfalls die exakte Lösung liefern. Wegen $m \geq i_j \geq 1$ für $j = 0, ..., q$ ist $i_q < ... < i_0$ für $m = 2, ..., q$ nicht möglich, so dass

$$\sum_{\substack{(i_0,...,i_q)\in\mathcal{J}_{q+1} \\ i_q<...<i_0}} b_{i_0} \prod_{j=0}^{q-1} \beta_{i_j,i_{j+1}} = 0$$

folgt.

Daher muss $P_q\left(\frac{1}{m}\right) = 0$ für $m = 1, ..., q$ gelten. Dies liefert genau q Nullstellen.

Um nun die Gleichheit der Polynome zu erhalten, werden die Werte an der Stelle $\gamma = 0$ betrachtet. Dann ist $\varphi(\gamma h\mathcal{A}) = 1$, so dass sich die Ableitungen der Stufenwerte vereinfachen. Für $q \geq 1$ erhält man

$$\frac{\partial^q k_{i_0}}{\partial h^q}(u(t), 0) = \binom{q}{1}\mathcal{A}\sum_{i_1=1}^{i_0-1}\beta_{i_0,i_1}\frac{\partial^{q-1}k_{i_1}}{\partial h^{q-1}}(u(t), 0)$$

und Auflösen der Rekursion liefert

$$\frac{\partial^q k_{i_0}}{\partial h^q}(u(t), 0) = q!\mathcal{A}^q \sum_{\substack{(i_1,...,i_q)\in\mathcal{J}_q \\ i_q<...<i_1<i_0}} \prod_{j=0}^{q-1}\beta_{i_j,i_{j+1}} f(u(t)).$$

Aus der Ordnungsbedingung

$$(q+1)\frac{\partial^q V}{\partial h^q}(u(t), 0) = \mathcal{A}^q f(u(t))$$

wird dann

$$\sum_{\substack{(i_0,...,i_q)\in\mathcal{J}_{q+1} \\ i_q<...<i_0}} \prod_{j=0}^{q-1}\beta_{i_j,i_{j+1}}\mathcal{A}^q f(u(t)) = \tfrac{1}{(q+1)!}\mathcal{A}^q f(u(t)). \tag{1.13}$$

Daher folgt $P_q(0) = \frac{1}{(q+1)!}$ und die Aussage ist bewiesen.

■

Bemerkung:
Da die $\mathcal{A}^q f(u(t))$ so gewählt werden können, dass für Gleichheit in (1.13) die Koeffizienten übereinstimmen müssen, sind die hinreichenden Bedingungen zugleich notwendig.

Der Satz ist auch für nichtautonome Gleichungen der Art $u' = \mathcal{A}u + \mathcal{B}_1 + t\mathcal{B}_2$ anwendbar, da diese durch Autonomisierung auf die Form

$$\begin{pmatrix} t \\ u \end{pmatrix}' = \begin{pmatrix} 0 & 0 \\ \mathcal{B}_2 & \mathcal{A} \end{pmatrix}\begin{pmatrix} t \\ u \end{pmatrix} + \begin{pmatrix} 1 \\ \mathcal{B}_1 \end{pmatrix}$$

gebracht werden. Dies ist eine Gleichung wie in (1.6).

1.2.2. Nicht-linearer Fall

Ohne eine geeignete Notation erscheint es sehr schwierig, die Konsistenzbedingungen im nicht-linearen Fall zu charakterisieren. Daher gilt es zunächst, eine Indexmenge zu definieren, deren Elemente den bei der Ableitung von $f(u(\cdot))$ auftretenden Differentialen zugeordnet werden können.

Hierzu sei die Existenz eines Objektes $\circ$ vorausgesetzt, das als Wurzel bezeichnet wird. Definiere nun für $j \in \mathbb{N}_0$ die Mengen $\mathbb{B}_j$ rekursiv durch

$$\mathbb{B}_0 := \{\circ\},$$
$$\mathbb{B}_j := \left\{ [T_1, ..., T_k] : \ k \in \mathbb{N}_1, \ T_1, ..., T_k \in \bigcup_{m=0}^{j-1} \mathbb{B}_m, \ k + \sum_{\ell=1}^{k} \left(\sum_{m=0}^{j-1} m \cdot \mathbf{1}_{\mathbb{B}_m}(T_\ell) \right) = j, \right\},$$

wobei $[T_1, ..., T_k]$ das nicht-geordnete k-Tupel ist, das $T_1, ..., T_k$ enthält. Dabei sind zwei k-Tupel gleich, falls sie dieselben Objekte mit denselben Häufigkeiten enthalten. Dies liefert nun die gesuchte Indexmenge.

Definition 1.2.4
Die Vereinigung

$$\mathbb{B} := \bigcup_{j \in \mathbb{N}_0} \mathbb{B}_j$$

heißt Menge der ***Butcher-Bäume****.*

An dieser Stelle erscheint es sinnvoll, die Voraussetzung, dass f hinreichend glatt ist, zu präzisieren. Sei also $f \in \mathcal{C}^p(\mathcal{D}; \mathbb{K}^N)$ mit $p \in \mathbb{N}_1$.

Unter dieser Voraussetzung kann dann eine Abbildung

$$F_{f,p} : \bigcup_{j=0}^{p} \mathbb{B}_j \longrightarrow \mathcal{C}(\mathcal{D}; \mathbb{K}^N)$$

konstruiert werden, die im Folgenden einfach als F bezeichnet wird. Setze hierzu

$$F(\circ)(v) := f(v)$$

und definiere rekursiv

$$F\big([T_1, ..., T_k]\big)(v) := f^k(v)\big[F(T_1)(v), ..., F(T_k)(v)\big].$$

Nach dem Satz von Schwarz spielt die Reihenfolge bei der Differentiation keine Rolle, so dass die Zuordnung wohldefiniert ist.

Jedem Differential soll nun noch die natürliche Zahl zugeordnet werden, die aussagt, bei der wievielten Ableitung von $f(u(\cdot))$ das Differential auftritt. Das wäre z.B. 0 für $f(v)$ und 1 für $f'(v)f(v)$. Die folgende Definition präzisiert dies.

Definition 1.2.5
Der Ausdruck $\varrho\big(F(T)\big)$ heißt **Ordnung** *von $F(T)$, wobei $\varrho : \mathbb{F} \longrightarrow \mathbb{N}_0$ rekursiv durch*

$$\varrho\big(F(\circ)\big) := 0,$$
$$\varrho\Big(F\big([T_1, ..., T_k]\big)\Big) := k + \sum_{\ell=1}^{k} \varrho\big(F(T_\ell)\big)$$

definiert ist.

Wegen der Kommutativität der Addition ist dies wohldefiniert. Aus den obigen Konstruktionen ist ersichtlich, dass ein Differential der Ordnung j in $\mathbb{B}_j$ liegt, denn für $j \geq 1$ gilt

$$\begin{aligned}\mathbb{B}_j &= \left\{ [T_1, ..., T_k] : \ k \in \mathbb{N}_1, \ T_1, ..., T_k \in \mathbb{B}, \ k + \sum_{\ell=1}^{k} \varrho(F(T_\ell)) = j, \right\} \\ &= \left\{ [T_1, ..., T_k] : \ k \in \mathbb{N}_1, \ T_1, ..., T_k \in \mathbb{B}, \ \varrho\Big(F\big([T_1, ..., T_k]\big)\Big) = j, \right\}.\end{aligned}$$

Damit sind die im Folgenden benötigten Grundlagen bereitgestellt. Nun gilt es, die obige Notation zu verwenden, um die Ableitungen von $f(u(\cdot))$ in gewünschter Form darzustellen.

Lemma 1.2.6
Sei $k \in \mathbb{N}_0$. Dann existieren $c(T) \in \mathbb{N}_1$, so dass

$$\frac{d^k}{dt^k}\Big[f(u(t))\Big] = \sum_{T \in \mathbb{B}_k} c(T) F(T)(u(t))$$

gilt.

Beweis:
Im Beweis dieses Lemmas und des anschließenden Satzes wird die Multiindex Notation

$$a = (a_1, ..., a_k), \qquad |a| = \sum_{j=1}^{k} a_j, \qquad \langle a \rangle = \sum_{j=1}^{k} j a_j$$

verwendet. Außerdem wird statt $[\underbrace{a_1, ..., a_1}_{r_1 - mal}, ..., \underbrace{a_k, ..., a_k}_{r_k - mal}]$ einfach $\prod_{j=1}^{k} a_j^{r_j}$ geschrieben.

Nach der Formel von Faà di Bruno (siehe [12, Lemma II.2.8]) gilt

$$\frac{d^k}{dt^k}\Big[f(u(t))\Big] = \sum_{\langle a \rangle = k} \widetilde{c}(a) f^{(|a|)}(u(t)) \prod_{j=1}^{k} \Big(u^{(j)}(t)\Big)^{a_j}, \tag{1.14}$$

wobei die $\widetilde{c}(a) \in \mathbb{N}_1$ sind. Bei der k-ten Ableitung wird dabei stets über Multiindizes der Länge k summiert. Wegen

$$u^{(j+1)}(t) = \frac{d^j}{dt^j}\Big[f(u(t))\Big] \tag{1.15}$$

kann man daraus mittels Induktion nach k die zu zeigende Behauptung folgern.

Induktionsanfang $k = 0$:
Mit $c(\circ) = 1$ ist die Gleichung erfüllt.

Induktionsschritt $(0, ..., k-1) \to k \leq p$:
Sind nach Induktionsvoraussetzung

$$\frac{d^j}{dt^j}\Big[f(u(t))\Big] = \sum_{T \in \mathbb{B}_j} c(T) F(T)(u(t))$$

für $j = 0, ..., k-1$, so erhält man zusammen mit (1.14) und (1.15)

$$\frac{d^k}{dt^k}\Big[f(u(t))\Big] = \sum_{\langle a\rangle = k} \widetilde{c}_k(a) f^{(|a|)}(u(t)) \prod_{j=0}^{k-1} \Big(\sum_{T\in\mathbb{B}_j} c(T)F(T)(u(t))\Big)^{a_{j+1}}.$$

Nach obiger Konstruktion der Butcher-Bäume und zugehöriger Differentiale gilt wegen $\langle a\rangle = |a| + \sum_{j=1}^{k} a_j(j-1) = k$ und der Multilinearität der Ableitung, dass die rechte Seite eine Linearkombination der Differentiale aus $F(\mathbb{B}_k)$ ist. Andererseits taucht jedes Differential der Ordnung k in der Summe auf, wobei die Koeffizienten ungleich null sind.

■

Für jedes Differential gibt es nun genau eine Ordnungsbedingung. Genauer gesagt, gilt der folgende Satz, wobei wiederum $\beta_{ij} = \alpha_{ij} + \gamma_{ij}$ gesetzt wird.

Satz 1.2.7
Es existieren von den Koeffizienten α_{ij}, β_{ij} und b_i unabhängige Polynome $P_T(\gamma)$ mit Grad kleiner gleich $\varrho(F(T))$, so dass sich die Ordnungsbedingungen als

$$\sum_{i=1}^{s} b_i \Phi_{iT} = P_T(\gamma)$$

schreiben lassen, wobei die Φ_{iT} durch

$$\Phi_{i\circ} := 1$$

$$\Phi_{iT} := \begin{cases} \displaystyle\sum_{j_1,...,j_k=1}^{i-1} \prod_{m=1}^{k} \Big(\alpha_{i,j_m}\Phi_{j_m,T_m}\Big), & T = [T_1, ..., T_k] \text{ für } k \geq 2 \\ \displaystyle\sum_{j=1}^{i-1} \beta_{i,j}\Phi(T_1), & T = [T_1] \end{cases}$$

gegeben sind.

Bemerkung:
Die Ordnungsbedingungen der Rosenbrock-Wanner-Verfahren (siehe [12, IV.7]) haben genau dieselbe Form, wobei sich nur die Polynome $P_T(\gamma)$ unterscheiden. Aufgrund der speziellen Struktur ist der Beweis im Fall der Rosenbrock-Wanner-Verfahren allerdings ein wenig einfacher und kann nicht direkt übernommen werden.

Beweis:
Das Vorgehen ist im Wesentlichen dasselbe wie im linearen Fall. Der erste Teil des Beweises besteht darin, per Induktion über q zu zeigen, dass es für $T \in \mathbb{B}_q$ und $S \in \bigcup_{j=0}^{q} \mathbb{B}_j$ Polynome $Q_{T,S}(\gamma)$ mit Grad kleiner gleich $\varrho(F(T)) - \varrho(F(S)) = q - \varrho(F(S))$ gibt, so dass

$$\frac{\partial^q k_i}{\partial h^q}(u(t), 0) = \sum_{T\in\mathbb{B}_q} \sum_{\substack{S\in\mathbb{B}_\ell \\ 0\leq\ell\leq q}} Q_{T,S}(\gamma)\Phi_{iS}F(T)(u(t))$$

gilt.

Induktionsanfang $q = 0$:
Wegen $\Phi_{i\circ} = 1$, $k_i(u(t), 0) = f(u(t))$ und $F(\circ)(v) = f(v)$ gilt die Behauptung mit $Q_{\circ,\circ}(\gamma) = 1$. Dies entspricht erwartungsgemäß $Q_{0,0}(\gamma)$ aus dem Beweis des linearen Falls.

Induktionsschritt $(0, ..., q-1) \to q$:
Definiere analog zum linearen Fall

$$\hat{k}_i(v,h) := f\Big(v + h\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\Big) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)$$

für $i = 1, ..., s$.

Mit Hilfe der Formel von Faà di Bruno (siehe [12, Lemma II.2.8]) erhält man für die k-ten partiellen Ableitungen ($k = 1, ..., q$) nach h an der Stelle $h = 0$

$$\begin{aligned}
\frac{\partial^k \hat{k}_i}{\partial h^k}(v,0) &= \frac{d^k}{dh^k}\left[f\Big(v + h\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\Big) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)\right]_{h=0} \\
&= \sum_{\langle a\rangle = k} \widetilde{c}_k(a) f^{(|a|)}(v) \prod_{m=1}^{k}\left(\frac{d^{m-1}}{dh^{m-1}}\left[\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\right]_{h=0}\right)^{a_m} \\
&\quad + \sum_{\substack{\langle a\rangle = k \\ |a|=1}} \widetilde{c}_k(a) f'(v) \sum_{j=1}^{i-1}\left(\gamma_{i,j}\frac{\partial^{k-1}k_j}{\partial h^{k-1}}(v,0)\right) \\
&= \sum_{\substack{\langle a\rangle = k \\ |a|\neq 1}} \widetilde{c}_k(a) f^{(|a|)}(v) \prod_{m=1}^{k}\left(\sum_{j_1,...,j_k=1}^{i-1}\alpha_{i,j_m}\frac{\partial^{m-1}k_{j_m}}{\partial h^{m-1}}(v,0)\right) \\
&\quad + \sum_{\substack{\langle a\rangle = k \\ |a|=1}} \widetilde{c}_k(a) f'(v) \sum_{j=1}^{i-1}\left(\beta_{i,j}\frac{\partial^{k-1}k_j}{\partial h^{k-1}}(v,0)\right).
\end{aligned}$$

Man beachte dabei im letzten Schritt, dass

$$\sum_{m=1}^{k} m a_k = \langle a\rangle = k$$

gilt.

Da nach Induktionsvoraussetzung

$$\frac{\partial^m k_i}{\partial h^m}(u(t),0) = \sum_{T\in\mathbb{B}_m}\sum_{\substack{S\in\mathbb{B}_\ell \\ 0\le\ell\le m}} Q_{T,S}(\gamma)\Phi_{iS}F(T)(u(t))$$

für $m = 0, ..., k-1$ gilt, folgt aus der Definition der Φ_{iT}, dass es Polynome $\widetilde{Q}_{T,S}$ mit Grad kleiner gleich $\varrho(F(T)) - \varrho(F(S)) - 1 = k - 1 - \varrho(F(S))$ gibt, so dass

$$\frac{\partial^k \hat{k}_i}{\partial h^k}(u(t),0) = \sum_{T\in\mathbb{B}_k}\sum_{\substack{S\in\mathbb{B}_m \\ 1\le m\le k}} \widetilde{Q}_{T,S}(\gamma)\Phi_{iS}F(T)(u(t))$$

ist.

Mit der Leibnizregel folgt dann

$$\begin{aligned}\frac{\partial^q k_i}{\partial h^q}(u(t),0) &= \sum_{k=0}^{q}\binom{q}{k}\frac{d^{q-k}}{dh^{q-k}}\Big[\varphi\Big(\gamma h f'(u(t))\Big)\Big]_{h=0}\frac{\partial^k \hat{k}_i}{\partial h^k}(u(t),0)\\ &= \sum_{k=1}^{q}\binom{q}{k}\frac{1}{q-k+1}\Big(\gamma f'(u(t))\Big)^{q-k}\sum_{T\in\mathbb{B}_k}\sum_{\substack{S\in\mathbb{B}_m\\ 1\le m\le k}}\widetilde{Q}_{T,S}(\gamma)\Phi_{iS}F(T)(u(t))\\ &\quad+\frac{1}{q+1}\Big(\gamma f'(u(t))\Big)^{q}\\ &= \sum_{T\in\mathbb{B}_q}\sum_{\substack{S\in\mathbb{B}_\ell\\ 0\le\ell\le q}}Q_{T,S}(\gamma)\Phi_{iS}F(T)(u(t)),\end{aligned}$$

wobei sich die $Q_{T,S}$ aus der vorletzten Zeile ablesen lassen. Die Gradbedingungen kann man analog zum linearen Fall durch Nachzählen überprüfen. Dies vollendet die Induktion.

Abschließend sei bemerkt, dass $\widetilde{Q}_{T,T} = Q_{T,T} \neq 0$ ist und außerdem $\widetilde{Q}_{T,S} = Q_{T,S} = 0$ für $\varrho(F(T)) = \varrho(F(S))$ mit $T \neq S$ gilt.

Es bleibt noch zu zeigen, dass sich die Ordnungsbedingungen als

$$\sum_{i=1}^{s} b_i\Phi_{iT} = P_T(\gamma)$$

schreiben lassen. Der Beweis erfolgt durch Induktion über $q = \varrho(F(T))$.

Induktionsanfang $q = 0$:
Für $q = 0$ wurde die Ordnungsbedingung bereits explizit berechnet. Dabei ist $P_\circ(\gamma) = 1$.

Induktionsschritt $(0, ..., q-1) \to (q)$:
Nach Induktionsvoraussetzung sind die Ordnungsbedingungen für $0, ..., q-1$ erfüllt. Die zusätzlichen Ordnungsbedingungen ergeben sich aus

$$(q+1)\frac{\partial^q V}{\partial h^q}(u(t),0) = \frac{d^q}{dt^q}\Big[f(u(t))\Big],$$

wobei sich die linke Seite nach dem ersten Teil dieses Beweises als

$$\begin{aligned}\frac{\partial^q V}{\partial h^q}(u(t),0) &= \sum_{i=1}^{s} b_i\frac{\partial^q k_i}{\partial h^q}(u(t),0)\\ &= \sum_{i=1}^{s} b_i\sum_{T\in\mathbb{B}_q}\sum_{\substack{S\in\mathbb{B}_\ell\\ 0\le\ell\le q}}Q_{T,S}(\gamma)\Phi_{iS}F(T)(u(t))\end{aligned}$$

umschreiben lässt.

Für die rechte Seite erhält man nach Lemma 1.2.6

$$\frac{d^q}{dt^q}\Big[f(u(t))\Big] = \sum_{T\in\mathbb{B}_q} c(T)F(T)(u(t)).$$

Indem nun noch $\sum_{i=1}^{s} b_i \Phi_{iS}$ für S mit $\varrho\big(F(S)\big) < \varrho\big(F(T)\big) = q$ durch $P_S(\gamma)$ ersetzt wird, erhält man durch Koeffizientenvergleich

$$\sum_{i=1}^{s} b_i \Phi_{iT} = \underbrace{n(T)\Big(\frac{1}{q+1}c(T) - \sum_{\substack{S\in\mathbb{B}_\ell \\ 0\le\ell\le q-1}} Q_{T,S}(\gamma)P_S(\gamma)\Big)}_{=:P_T(\gamma)}$$

mit $n(T) \in \mathbb{N}_1$. Die Polynome P_T haben dabei im Maximalfall Grad $\varrho\big(F(T)\big) = q$. Damit ist die Aussage mittels Induktion gezeigt.

■

Bis zur Ordnung 4 sind die Φ_{iT} und $P_T(\gamma)$ in der folgenden Tabelle aufgelistet.

Differential $F(T)$	Φ_{iT}	$P_T(\gamma)$
f	1	1
$f'f$	$\sum_{j=1}^{i-1} \beta_{ij}$	$\frac{1}{2}(1-\gamma)$
$f''[f,f]$	$\sum_{j,k=1}^{i-1} \alpha_{ij}\alpha_{ik}$	$\frac{1}{3}$
$f'f'f$	$\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} \beta_{ij}\beta_{jk}$	$\frac{1}{3}(\frac{1}{2}-\gamma)(1-\gamma)$
$f'''[f,f,f]$	$\sum_{j,k,\ell=1}^{i-1} \alpha_{ij}\alpha_{ik}\alpha_{i\ell}$	$\frac{1}{4}$
$f''[f'f,f]$	$\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\sum_{\ell=1}^{i-1} \alpha_{ij}\beta_{jk}\alpha_{i\ell}$	$\frac{1}{8}-\frac{\gamma}{6}$
$f'f''[f,f]$	$\sum_{j=1}^{i-1}\sum_{k,\ell=1}^{j-1} \beta_{ij}\alpha_{jk}\alpha_{j\ell}$	$\frac{1}{12}-\frac{\gamma}{6}$
$f'f'f'f$	$\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\sum_{\ell=1}^{k-1} \beta_{ij}\beta_{jk}\beta_{k\ell}$	$\frac{1}{4}(\frac{1}{3}-\gamma)(\frac{1}{2}-\gamma)(1-\gamma)$

Beispiel:
Gegeben sei das zweistufige Verfahren mit den Koeffizienten $\gamma = \frac{1}{2}$, $\alpha_{21} = \alpha$, $\gamma_{21} = \frac{3}{4}\alpha^2 - \alpha$, $b_1 = 1 - \frac{1}{3\alpha^2}$, $b_2 = \frac{1}{3\alpha^2}$. Dabei ist $\alpha \in \mathbb{R}$ ein freier Parameter.

Folglich ist $\beta_{21} = \alpha_{21} + \gamma_{21} = \frac{3}{4}\alpha^2$. Mittels Nachrechnen erhält man nun

$$\begin{aligned}
\sum_{i=1}^{2} b_i &= 1 - \frac{1}{3\alpha^2} + \frac{1}{3\alpha^2} = 1,\\
\sum_{i=1}^{2}\sum_{j=1}^{i-1} b_i\beta_{ij} &= \frac{1}{3\alpha^2}\cdot\frac{3}{4}\alpha^2 = \frac{1}{4} = \frac{1}{2}(1-\gamma),\\
\sum_{i=1}^{2}\sum_{j,k=1}^{i-1} b_i\alpha_{ij}\alpha_{ik} &= \frac{1}{3\alpha^2}\cdot\alpha^2 = \frac{1}{3},\\
\sum_{i=1}^{2}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} b_i\beta_{ij}\beta_{jk} &= 0 = \frac{1}{3}\left(\frac{1}{2}-\gamma\right)(1-\gamma).
\end{aligned}$$

Das Verfahren ist also konsistent der Ordnung 3. Da unabhängig von der Wahl von α stets

$$\sum_{i=1}^{2}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1}\sum_{\ell=1}^{i-1}\alpha_{ij}\beta_{jk}\alpha_{i\ell} = 0 \neq \frac{1}{24} = \frac{1}{8} - \frac{\gamma}{6}$$

ist, sind nicht alle Bedingungen für Ordnung 4 erfüllt. Allerdings gilt für alle $q \geq 2$

$$\sum_{\substack{(i_0,\dots,i_q)\in\mathcal{J}_{q+1}\\ i_q<\dots<i_0}} b_{i_0}\prod_{j=0}^{q-1}\beta_{i_j,i_{j+1}} = 0 = \frac{1}{q+1}\prod_{j=0}^{q-1}\left(\frac{1}{q-j}-\gamma\right),$$

so dass das Verfahren nach Satz 1.2.3 die exakte Lösung von linearen Anfangswertaufgaben mit konstanten Koeffizienten liefert.

Wendet man ein Einschrittverfahren, das konsistent der Ordnung p ist, zur Lösung gewöhnlicher Differentialgleichungen an, so genügt schon die Lipschitz-Stetigkeit der rechten Seite und der Verfahrensfunktion, um Konvergenz der Ordnung p zu erhalten. Durch Überprüfen der Konsistenzbedingungen, wie im obigen Beispiel, kann man also direkt die Konvergenzordnung berechnen.

1.3. Reduzierte Verfahren

Analog zu den Runge-Kutta-Verfahren benötigt man für hohe Ordnungen auch eine hohe Stufenzahl. Dies hat allerdings wesentliche Konsequenzen für den Aufwand des Verfahrens.

Ist die Dimension N klein genug, kann man $\varphi(\gamma h f'(v))$ etwa durch Padé-Approximation auswerten. Bei einer Größenordnung von $N \sim 10^3$ ist dies allerdings zu aufwendig, so dass das Matrix-Vektor-Produkt mittels Krylow-Unterraum-Approximation berechnet und $\varphi(\gamma h f'(v))$ selbst nicht mehr ausgewertet wird. Die Krylow-Unterraum-Approximation ist Thema des zweiten Kapitels.

Um bei einem Verfahren der Form

$$\begin{aligned}
u_{n+1} &= u_n + h\sum_{i=1}^{s} b_i k_i(u_n,h),\\
k_i(v,h) &= \varphi(\gamma h f'(v))\left(f\left(v + h\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\right) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)\right)
\end{aligned}$$

einen Stufenwert $k_i(v,h)$ zu berechnen, muss $\varphi(\gamma h f'(v))$ mit einem Vektor multipliziert werden, der von f, v und h, den vorherigen Stufenwerten und den Koeffizienten des Verfahrens abhängt. Daher sind die s Vektoren im Allgemeinen voneinander verschieden, so dass s verschiedene Krylow-Unterräume berechnet werden, was die Verfahren sehr aufwendig macht. Die Lösung des Problems sind die reduzierten Verfahren, die nun betrachtet werden. Wie bisher wird dabei $\beta_{ij} = \alpha_{ij} + \gamma_{ij}$ gesetzt.

Die wesentliche Idee besteht darin, geeignete Koeffizienten zu bestimmen, so dass das zugehörige Verfahren umformuliert werden kann.

Von wesentlicher Bedeutung sind dabei die Eigenschaften der φ-Funktion. Die benötigte Rekursionsformel liefert das folgende Lemma.

Lemma 1.3.1
Sei $j \in \mathbb{N}_0$ und $z \in \mathbb{C}$. Dann gilt

$$\varphi((j+1)z) = \frac{j}{j+1}\Big(z\varphi(z)+1\Big)\varphi(jz) + \frac{1}{j+1}\varphi(z). \tag{1.16}$$

Beweis:
Da für $z \in \mathbb{C}$

$$e^z = z\varphi(z) + 1$$

ist, erhält man

$$\varphi((j+1)z) = \frac{e^{jz}e^z - 1}{(j+1)z} = \frac{1}{j+1}\frac{e^z-1}{z}\sum_{k=0}^{j} e^{kz} = \frac{1}{j+1}\varphi(z)\sum_{k=0}^{j}\Big(z\varphi(z)+1\Big)^k, \tag{1.17}$$

wobei ausgenutzt wird, dass $(e^z-1)\sum_{k=0}^{j} e^{kz}$ eine Teleskopsumme ist.
Zweimalige Anwendung dieser Identität und eine Indexverschiebung liefern dann

$$\begin{aligned}
\varphi\big((j+1)z\big) &\overset{(1.17)}{=} \frac{1}{j+1}\varphi(z)\left(\sum_{k=1}^{j}\Big(z\varphi(z)+1\Big)^k\right) + \frac{1}{j+1}\varphi(z) \\
&= \frac{j}{j+1}\Big(z\varphi(z)+1\Big)\frac{\varphi(z)}{j}\left(\sum_{k=0}^{j-1}\Big(z\varphi(z)+1\Big)^k\right) + \frac{1}{j+1}\varphi(z) \\
&\overset{(1.17)}{=} \frac{j}{j+1}\Big(z\varphi(z)+1\Big)\varphi(jz) + \frac{1}{j+1}\varphi(z),
\end{aligned}$$

was zu zeigen war.

■

Um die Notation zu vereinfachen, werden nun noch Hilfsvektoren

$$d_i(v,h) := f\big(y_i(v,h)\big) - f(v) - hf'(v)w_i(v,h) \tag{1.18}$$

definiert, wobei

$$y_i(v,h) := v + hw_i(v,h),$$
$$w_i(v,h) := \sum_{j=1}^{i-1} \alpha_{ij} k_j(v,h)$$

ist.

Bemerkung:
Die wesentliche Idee hinter den Hilfsvektoren ist die Taylorentwicklung

$$f(y_i(v,h)) = f(v) + hf'(v)w_i(v,h) + \mathcal{O}(h^2), \quad h \to 0.$$

Die Konsequenz ist, dass $d_i(v,h)$ für kleine Schrittweiten wesentlich kleiner als $f(v)$ wird. Beide Ausdrücke sind allerdings bei reduzierten Verfahren für die Berechnung der Stufenwerte von Bedeutung. In einem der Beispiele, die am Schluss dieses Kapitels zu finden sind, ist etwa $k_1(v,h) = \varphi\left(\frac{1}{3}hf'(v)\right) f(v)$ und $k_4(v,h) = \varphi\left(\frac{1}{3}hf'(v)\right) d_4(v,h)$.

Die Eigenschaft, dass die Norm von $d_4(v,h)$ sehr klein ist, führt dazu, dass beim Krylow-Unterraum-Verfahren nur wenige Iterationsschritte benötigt werden, um eine gewählte Fehlertoleranz zu unterschreiten. Letzteres geht aus den Abschätzungen des zweiten Kapitels hervor.

Eine nützliche Formel liefert das folgende Lemma.

Lemma 1.3.2
Seien $h > 0$ und $v \in \mathbb{K}^N$. Dann gilt für $i = 1, ..., s$

$$k_i(v,h) = k_1(v,h) + \varphi(\gamma hf'(v))d_i(v,h) + \varphi(\gamma hf'(v))hf'(v)\sum_{j=1}^{i-1}\beta_{ij}k_j(v,h) \tag{1.19}$$

mit $d_i(v,h)$ aus (1.18).

Beweis:
Dies folgt direkt aus der Definition der Stufenwerte und Hilfsvektoren, denn es gilt

$$\begin{aligned}
k_i(v,h) &= \varphi(\gamma hf'(v)) \left(f\left(v + h\sum_{j=1}^{i-1}\alpha_{ij}k_j(v,h)\right) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)\right) \\
&= \varphi(\gamma hf'(v)) \left(f(y_i(v,h)) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)\right) \\
&\overset{(1.18)}{=} \varphi(\gamma hf'(v)) \left(d_i(v,h) + f(v) + hf'(v)w_i(v,h) + hf'(v)\sum_{j=1}^{i-1}\gamma_{ij}k_j(v,h)\right) \\
&= k_1(v,h) + \varphi(\gamma hf'(v))d_i(v,h) + \varphi(\gamma hf'(v))hf'(v)\sum_{j=1}^{i-1}\underbrace{(\alpha_{ij}+\gamma_{ij})}_{=\beta_{ij}}k_j(v,h).
\end{aligned}$$

■

Hiermit sind die Grundlagen für den entscheidenden Satz geschaffen, auf dessen Grundlage ein Verfahren mit geeigneten Koeffizienten umformuliert werden kann.

Die Bedeutung des nun folgenden Satzes liegt darin, dass für die ersten n Stufen nur eine Krylow-Unterraum-Approximation durchgeführt werden muss. Dies ist darin begründet, dass man aus $k_1(v,h) = \varphi(\gamma h f'(v))f(v)$ die anderen $k_i(v,h) = \varphi(i\gamma h f'(v))f(v)$ mit der Rekursionsformel (1.16) berechnen kann.

Für die jeweils nächsten n Stufen ist die Situation ähnlich. Wiederum ist nur eine Krylow-Unterraum-Approximation durchzuführen. Ist $s = q \cdot n + r$, so müssen statt s also $q+1$ bzw. im Fall $r = 0$ sogar nur q Krylow-Unterräume berechnet werden.

Der Rechenaufwand wird zudem noch dadurch verringert, dass die Norm $d_{mn+1}(v,h)$ typischerweise wesentlich kleiner als die von $f(v)$ ist. In der Praxis bedeutet das, dass die Krylow-Unterraum-Approximation nach wesentlich weniger Schritten abgebrochen werden kann, da der Anteil am Gesamtfehler klein ist. Details zu den Abbruchkriterien finden sich in [20, 6.3].

Satz 1.3.3
Für die Stufenzahl s gelte $s = q \cdot n + r$ mit $q, n \in \mathbb{N}_1$ und $0 \leq r \leq n-1$. Sei außerdem $\gamma = \frac{1}{n}$. Dann existieren Koeffizienten α_{ij} und β_{ij} mit $j = 1, ..., i-1$, so dass für $i = 1, ..., n$

$$k_i(v,h) = \varphi(i\gamma h f'(v))f(v)$$

und unter der Voraussetzung $mn + i \leq s$ mit $m \in \mathbb{N}_1$

$$k_{mn+i}(v,h) = k_1(v,h) + \varphi(i\gamma h f'(v))d_{mn+1}(v,h)$$

gilt.

Bemerkung:
In [20, 5.1] wurde dieses Resultat zwar vorgestellt, allerdings nicht bewiesen. Der folgende Beweis liefert nicht nur die Existenzaussage, sondern ist konstruktiv. Folglich ist er ein nützliches Hilfsmittel, um geeignete Verfahren zu finden oder die Konstruktion von Verfahren nachzuvollziehen.

Beweis:
Im ersten Schritt dieses Beweises führt eine Induktion über i zur ersten Aussage des Satzes. Setze dazu für alle j und $i = 1, .., n$ zunächst $\alpha_{ij} = 0$.

Induktionsanfang $i = 1$:
Da das Verfahren explizit sein soll, ist zwangsläufig $\beta_{1j} = 0$. Damit gilt bereits

$$k_1 = \varphi(\gamma h f'(v))f(u) = \varphi(1 \cdot \gamma h f'(v))f(u).$$

Induktionsschritt $(i-1) \to i \leq n$:
Wegen

$$w_i(v,h) = \sum_{j=1}^{i-1} \alpha_{ij} k_j(v,h) = 0$$

ist

$$y_i(v,h) = v + w_i(v,h) = v$$

und damit

$$d_i(v,h) = f(y_i(v,h)) - f(v) - hf'(v)w_i(v,h) = 0$$

für $i = 1, ..., n$.

Mit Blick auf die Rekursionsformel (1.16) wählt man nun

$$\beta_{ij} = \begin{cases} \frac{i-1}{i}\beta_{i-1,j} & j \leq i-2 \\ \frac{i-1}{i}\gamma & j = i-1 \\ 0 & j \geq i. \end{cases} \tag{1.20}$$

Dann folgt

$$\begin{aligned}
k_i(v,h) &\overset{(1.19)}{=} k_1(v,h) + \varphi(\gamma hf'(v))\underbrace{d_i(v,h)}_{=0} + \varphi(\gamma hf'(v))hf'(v)\sum_{j=1}^{i-1}\beta_{ij}k_j(v,h) \\
&= \beta_{i,i-1}\varphi(\gamma hf'(v))hf'(v)k_{i-1}(v,h) + \underbrace{\varphi(\gamma hf'(v))f(v)}_{=k_1(v,h)} \\
&\quad + \varphi(\gamma hf'(v))hf'(v)\sum_{j=1}^{i-2}\beta_{ij}k_j(v,h) \\
&\overset{(1.20)}{=} \frac{i-1}{i}\gamma\varphi(\gamma hf'(v))hf'(v)k_{i-1}(v,h) + \frac{1}{i}\varphi(\gamma hf'(v))f(v) \\
&\quad + \frac{i-1}{i}\underbrace{\varphi(\gamma hf'(v))\left(f(v) + hf'(v)\sum_{j=1}^{i-2}\beta_{i-1,j}k_j(v,h)\right)}_{=k_{i-1}(v,h)} \\
&\overset{(1.19)}{=} \frac{i-1}{i}\Big(\gamma\varphi(\gamma hf'(v))hf'(v) + I\Big)k_{i-1}(v,h) + \frac{1}{i}\varphi(\gamma hf'(v))f(v) \\
&\overset{IV}{=} \left(\frac{i-1}{i}\Big(\gamma\varphi(\gamma hf'(v))hf'(v) + I\Big)\varphi((i-1)\gamma hf'(v)) + \frac{1}{i}\varphi(\gamma hf'(v))\right)f(v) \\
&\overset{(1.16)}{=} \varphi(i\gamma hf'(v))f(v),
\end{aligned}$$

was zu zeigen war.

Die zweite Aussage wird ebenfalls mittels Induktion über i bewiesen. Sei $m \geq 1$ beliebig.

Induktionsanfang $i = 1$:
Setze $\beta_{mn+1,j} = 0$ für $j = 1, ..., s$ und wähle beliebige $\alpha_{mn+1,j}$ für $j = 1, ..., mn$. Damit das Verfahren explizit ist, gilt natürlich $\alpha_{mn+1,j} = 0$ für $j \geq mn$.

Die Behauptung lässt sich nun leicht nachrechnen, denn

$$k_{mn+1}(v,h) = k_1(v,h) + \varphi(1 \cdot \gamma hf'(v))d_{mn+1}(v,h)$$

folgt sofort aus (1.19), da dort der letzte Summand wegfällt.

Induktionsschritt $(i-1) \to i \leq n$:
Setze zunächst

$$\alpha_{mn+i,j} = \alpha_{mn+1,j}$$

für $j = 1, ..., s$, woraus sofort

$$d_{mn+i} = d_{mn+1}$$

folgt. Wiederum mit Blick auf die Rekursionsformel (1.16) wählt man

$$\beta_{mn+i,j} = \begin{cases} \frac{i-1}{i}(\beta_{mn+i-1,j} - \delta_{1j} \cdot \gamma) & j \leq mn+i-2 \\ \frac{i-1}{i}\gamma & j = mn+i-1 \\ 0 & j \geq mn+i. \end{cases} \tag{1.21}$$

Dann folgt

$$\begin{aligned}
& k_{mn+i}(v,h) \\
\overset{(1.19)}{=}\ & k_1(v,h) + \varphi(\gamma h f'(v)) d_{mn+1}(v,h) + \varphi(\gamma h f'(v)) h f'(v) \sum_{j=1}^{mn+i-1} \beta_{mn+i,j} k_j(v,h) \\
=\ & k_1(v,h) + \varphi(\gamma h f'(v)) d_{mn+1}(v,h) + \varphi(\gamma h f'(v)) h f'(v) \beta_{mn+i,mn+i-1} k_{mn+i-1}(v,h) \\
& + \varphi(\gamma h f'(v)) h f'(v) \sum_{j=1}^{mn+i-2} \beta_{mn+i,j} k_j(v,h) \\
\overset{(1.21)}{=}\ & k_1(v,h) + \varphi(\gamma h f'(v)) d_{mn+1}(v,h) + \tfrac{i-1}{i} \varphi(\gamma h f'(v)) h f'(v) \sum_{j=1}^{mn+i-2} \beta_{mn+i-1,j} k_j(v,h) \\
& + \varphi(\gamma h f'(v)) h f'(v) \tfrac{i-1}{i} \gamma \Big(k_{mn+i-1}(v,h) - k_1(v,h) \Big) \\
\overset{(1.19)}{=}\ & \tfrac{1}{i} k_1(v,h) + \tfrac{i-1}{i} k_{mn+i-1}(v,h) + \tfrac{1}{i} \varphi(\gamma h f'(v)) d_{mn+1}(v,h) \\
& + \tfrac{i-1}{i} \varphi(\gamma h f'(v)) \gamma h f'(v) (k_{mn+i-1}(v,h) - k_1(v,h)) \\
\overset{IV}{=}\ & \tfrac{1}{i} k_1(v,h) + \tfrac{i-1}{i} \Big(k_1(v,h) + \varphi((i-1)\gamma h f'(v)) d_{mn+1}(v,h) \Big) \\
& + \tfrac{i-1}{i} \varphi(\gamma h f'(v)) \gamma h f'(v) \varphi((i-1)\gamma h f'(v)) d_{mn+1}(v,h) \\
& + \tfrac{1}{i} \varphi(\gamma h f'(v)) d_{mn+1}(v,h) \\
=\ & k_1(v,h) + \tfrac{i-1}{i} \Big(\varphi(\gamma h f'(v)) \gamma h f'(v) + I \Big) \varphi((i-1)\gamma h f'(v)) d_{mn+1}(v,h) \\
& + \tfrac{1}{i} \varphi(\gamma h f'(v)) d_{mn+1}(v,h) \\
\overset{(1.16)}{=}\ & k_1(v,h) + \varphi(i\gamma h f'(v)) d_{mn+1}(v,h),
\end{aligned}$$

was zu zeigen war.

■

Der Satz liefert genug freie Parameter, um Verfahren der Ordnung 4 zu konstruieren. Im Folgenden sind zwei Beispiele zu finden. Dabei wurde $\tilde{k}_i = k_i - k_1$ für $i \geq n$ gesetzt.

Beispiel:
Betrachte das durch

$$u_{n+1} = u_n + h\Big(k_2(u_n,h) + \tfrac{16}{27}\tilde{k}_3(u_n,h)\Big)$$

gegebene dreistufige Verfahren, wobei

$$\begin{aligned}
k_1(v,h) &= \varphi\left(\tfrac{1}{2}hf'(v)\right) f(v),\\
k_2(v,h) &= \varphi\left(hf'(v)\right) f(v),\\
\tilde{k}_3(v,h) &= \varphi\left(\tfrac{1}{2}hf'(v)\right) d_3(v,h)
\end{aligned}$$

und

$$\begin{aligned}
d_3(v,h) &= f(y_3(v,h)) - f(v) - hf'(v)w_3(v,h),\\
y_3(v,h) &= v + hw_3(v,h),\\
w_3(v,h) &= \tfrac{3}{8}\Big(k_1(v,h) + k_2(v,h)\Big)
\end{aligned}$$

ist. Nach [20, 5.2] ist dies ein Verfahren der Ordnung 4 und liefert die exakte Lösung von linearen Anfangswertaufgaben mit konstanten Koeffizienten.

Beispiel:
Betrachte das durch

$$u_{n+1} = u_n + h\Big(k_3(u_n,h) + \tilde{k}_4(u_n,h) - \tfrac{4}{3}\tilde{k}_5(u_n,h) + \tilde{k}_6(u_n,h) + \tfrac{1}{6}\tilde{k}_7(u_n,h)\Big)$$

gegebene siebenstufige Verfahren, wobei

$$\begin{aligned}
k_1(v,h) &= \varphi\left(\tfrac{1}{3}hf'(v)\right) f(v),\\
k_2(v,h) &= \varphi\left(\tfrac{2}{3}hf'(v)\right) f(v),\\
k_3(v,h) &= \varphi\left(hf'(v)\right) f(v),\\
\tilde{k}_4(v,h) &= \varphi\left(\tfrac{1}{3}hf'(v)\right) d_4(v,h),\\
\tilde{k}_5(v,h) &= \varphi\left(\tfrac{2}{3}hf'(v)\right) d_4(v,h),\\
\tilde{k}_6(v,h) &= \varphi\left(hf'(v)\right) d_4(v,h),\\
\tilde{k}_7(v,h) &= \varphi\left(\tfrac{1}{3}hf'(v)\right) d_7(v,h)
\end{aligned}$$

und

$$\begin{aligned}
d_4(v,h) &= f(y_4(v,h)) - f(v) - hf'(v)w_4(v,h),\\
d_7(v,h) &= f(y_7(v,h)) - f(v) - hf'(v)w_7(v,h)
\end{aligned}$$

sowie

$$\begin{aligned}
y_4(v,h) &= v + hw_4(v,h),\\
y_7(v,h) &= v + hw_7(v,h),\\
w_4(v,h) &= \tfrac{-7}{300}k_1(v,h) + \tfrac{97}{150}k_2(v,h) + \tfrac{37}{300}k_3(v,h),\\
w_7(v,h) &= \tfrac{59}{300}k_1(v,h) - \tfrac{7}{75}k_2(v,h) + \tfrac{269}{300}k_3(v,h) + \tfrac{2}{3}(\tilde{k}_4(v,h) + \tilde{k}_5(v,h) + \tilde{k}_6(v,h))
\end{aligned}$$

ist.

Nach [20, 5.2] ist dies ein Verfahren der Ordnung 4 und liefert die exakte Lösung von linearen Anfangswertaufgaben mit konstanten Koeffizienten. Da es dem vorherigen Verfahren bei differential-algebraischen Gleichungen und bei Verwendung einer inexakten Jacobi-Matrix überlegen ist und zudem eingebettete Verfahren zur Schrittweitensteuerung existieren, wurde es von den Autoren von [20] als **exp4** implementiert. Näheres findet sich dazu in [20, 6.-8.].

2. Approximation des Exponentialoperators

In diesem Kapitel wird die Approximation des Matrix-Exponentialoperators mittels Krylow-Unterraum-Verfahren betrachtet. Der Zusammenhang mit den exponentiellen Integratoren des vorherigen Kapitels ist dabei recht einfach einzusehen, wenn man die Frage nach dem Aufwand der Verfahren stellt. Um einen Ausdruck der Form $\varphi(\tau A)v$, bei der A eine Matrix, v ein Vektor und τ eine positiver, reeller Wert ist, zu berechnen, sind Padé-Approximationen grundsätzlich keine schlechte Wahl. Bei sehr großen Systemen, etwa im Fall $A \in \mathbb{C}^{1000\times 1000}$, werden sie allerdings zu aufwendig. Folglich muss ein anderer Weg gefunden werden.

Krylow-Unterraum-Verfahren sind iterative, numerische Verfahren, die eigentlich zur Lösung großer, dünnbesetzter, linearer Gleichungssysteme verwendet werden. Zu den bekanntesten Vertretern zählen das Arnoldi- und das Lanczos-Verfahren. Die wesentliche Idee ist in jedem Fall die Projektion in einen niedrigdimensionalen Raum.

Gallopoulos und Saad haben bereits in [9] und [38] Fehlerschranken für die Approximation von $e^{\tau A}v$ gefunden. Da die theoretischen Ergebnisse die wesentlich kleineren Fehler der numerischen Beispiele jedoch nicht erklären können, ist es sinnvoll, die Krylow-Unterraum-Verfahren noch einmal genauer zu analysieren.

Mit Hilfe des holomorphen Funktionalkalküls, d.h. insbesondere der Cauchyschen Integralformel angewandt auf Matrizen, wird im Folgenden das Problem, eine Schranke für die Approximation von $e^{\tau A}v$ zu finden, auf den Fall eines linearen Gleichungssystems zurückgeführt. Da die Fehlerschranken bei der Approximation von $e^{\tau A}v$ auch für $\varphi(\tau A)v$ gelten, ist es gerechtfertigt, $e^{\tau A}v$ statt $\varphi(\tau A)v$ zu betrachten.

Der Inhalt dieses Kapitels basiert hauptsächlich auf der Arbeit [19] von M. Hochbruck und C. Lubich. Für die benötigten Grundlagen zu Krylow-Unterraum-Verfahren betrachte man [18, Chapter 13] oder auch [13], [32]. Der Kalkül der Matrixfunktion wird im ersten Kapitel von [18] ausführlich behandelt.

Schließlich sei noch bemerkt, dass sich Druskin und Knizherman (siehe [5], [6], [28] und [29]) in der russischen Literatur mit der Thematik beschäftigt und ebenfalls Fehlerschranken gefunden haben, jedoch mit einer anderen Herangehensweise.

2.1. Grundlagen

Die Approximation von Funktionen mit Polynomen ist in der Mathematik ein vielfach verwendetes Mittel, so dass es naheliegend ist, auch e^A durch ein Polynom anzunähern. Da die Anwendung im vorherigen Kapitel die explizite Kenntnis von e^A gar nicht erfordert, ist eine Approximation der Form

$$e^A v \approx P_{k-1}(A)v$$

sinnvoll, wobei $\deg(P_{k-1}) \leq k-1$ für gegebenes $k \in \mathbb{N}_1$ ist. Gesucht ist also eine Näherung für das Produkt von e^A mit einem Vektor v, wobei das Resultat $P_{k-1}(A)v$ wiederum ein Vektor und insbesondere Element des folgenden Unterraums ist.

Definition 2.1.1
Seien $A \in \mathbb{C}^{N\times N}$, $v \in \mathbb{C}^N \setminus \{\mathbf{0}\}$ und $k \in \mathbb{N}_1$ gegeben. Der Untervektorraum

$$\mathcal{K}_k(A, v) = \operatorname{span}\{v, Av, \ldots, A^{k-1}v\}$$

von $\mathbb{C}^N$ heißt k-ter Krylow-Unterraum von A und v.

Entscheidend ist demnach, ein Element des Krylow-Unterraums zu finden, das $e^A v$ geeignet approximiert. Eine Methode hierzu ist die Arnoldi-Iteration aus [18, Algorithm 13.3]. Das folgende Lemma fasst die wichtigsten Eigenschaften zusammen.

Lemma 2.1.2
Seien eine Matrix $A \in \mathbb{C}^{N\times N}$ und ein Vektor $v \in \mathbb{C}^N$ mit $\|v\| = 1$ gegeben. Das Arnoldi-Verfahren angewandt auf A und $v_1 := v$ liefert die beiden Matrizen

$$V_m := \begin{pmatrix} v_1 & | & \ldots & | & v_m \end{pmatrix} \in \mathbb{C}^{N\times m}$$

und

$$H_m := \begin{pmatrix} h_{11} & \ldots & \ldots & h_{1m} \\ h_{21} & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{m,m-1} & h_{mm} \end{pmatrix} \in \mathbb{C}^{m\times m}$$

sowie $h_{m+1,m} \in \mathbb{C}$ mit den nachfolgenden Eigenschaften.

- *H_m ist eine obere Hessenbergmatrix, d.h. es gilt $h_{ij} = 0$ für $i > j+1$.*
- *Die Vektoren $v_1, ..., v_m$ sind bzgl. des kanonischen Skalarprodukts $\langle x, y\rangle := y^* x$ für $x, y \in \mathbb{C}^N$ orthonormal und es gilt*

  $$\operatorname{span}\{v_1, ..., v_m\} = \mathcal{K}_m(A, v),$$

  d.h. $\{v_i\}_{i=1}^m$ ist eine Orthonormalbasis des Krylow-Unterraums $\mathcal{K}_m(A, v)$.
- *Außerdem sind die beiden Identitäten*

  $$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T \tag{2.1}$$

  und

  $$V_m^* A V_m = H_m \tag{2.2}$$

  erfüllt. Bezüglich der Basis $\{v_i\}_{i=1}^m$ ist H_m demnach die darstellende Matrix der Projektion von A in den Unterraum $\mathcal{K}_m(A, v)$.

Beweis:
Für den Beweis siehe [18, 13.2.1].

■

Neben diesen für die folgende Fehleranalyse wesentlichen Aussagen sind noch zwei andere Dinge zu bemerken.

- Für die Arnoldi-Iteration ist es nicht notwendig, die Matrix A explizit zu kennen bzw. zu speichern. Grundsätzlich wäre es sogar möglich, die Matrix-Vektor-Produkte im Arnoldi-Algorithmus zu approximieren, etwa

$$f'(u)v \approx \frac{f(u+\delta v) - f(u)}{\delta},$$

  wobei $\delta \in \mathbb{R}$ hinreichend klein ist.

- Für $\tau \in \mathbb{R}$ gelten

$$V_m^* \tau A V_m = \tau H_m$$

  und

$$\mathcal{K}_m(\tau A, v) = \mathcal{K}_m(A, v).$$

  Daher kann der skalare Faktor bei der Durchführung der Arnoldi-Iteration ignoriert werden. Die Annahme $\|v\| = 1$ ist ebenso keine Einschränkung, da im allgemeinen Fall $v_1 = \frac{v}{\|v\|}$ gesetzt werden kann.

Wie bereits erwähnt dient das Arnoldi-Verfahren eigentlich zur Lösung linearer Gleichungssysteme. Aus diesem Grund ist es zur Fehlerabschätzung keineswegs sinnvoll, $e^{\tau A}$ bzw. $\varphi(\tau A)$ durch eine Potenzreihe darzustellen. Den benötigten Zusammenhang liefert hingegen der holomorphe Funktionalkalkül aus [18, Chapter 1]. Damit erhält man für ein Gebiet G und holomorphes $f : \overline{G} \longrightarrow \mathbb{C}$ unter der Bedingung $\sigma(A) \subseteq G$ die Formel

$$f(A) = \frac{1}{2\pi i} \int_{\partial G} f(z)(zI - A)^{-1} dz, \tag{2.3}$$

so dass das ursprüngliche Problem auf die bekannte (vgl. [37]) Approximation

$$(zI - A)^{-1} v \approx V_m (zI - H_m)^{-1} e_1 \tag{2.4}$$

zurückgeführt werden kann.

Allerdings erweist sich die Bedingung $\sigma(A) \subseteq G$ als zu schwach, da im Allgemeinen $\sigma(H_m) \nsubseteq \sigma(A)$ ist und folglich Elemente aus ∂G in $\sigma(H_m)$ liegen können. Dann wäre ∂G aber kein zulässiger Integrationsweg zur Bestimmung von $f(H_m)$ mittels (2.3) und (2.4). Der Ausweg besteht darin, die Voraussetzung zu verschärfen. Hierzu wird die folgende Definition benötigt.

Definition 2.1.3
Seien $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ ein Hilbertraum über $\mathbb{R}$ oder $\mathbb{C}$ und $B : \mathcal{H} \to \mathcal{H}$ ein beschränkter, linearer Operator. Dann bezeichnet man

$$\mathcal{F}(B) := \Big\{ \langle Bx, x\rangle : x \in \mathcal{H},\ \langle x, x\rangle = 1 \Big\}$$

als ***numerischen Wertebereich****.*

Im Fall $\mathcal{H} = \mathbb{C}^N$ und dem durch $\langle x, y\rangle = y^*x,\ x, y \in \mathbb{C}^N$ definierten Skalarprodukt ergibt sich

$$\mathcal{F}(B) = \left\{ \frac{x^*Bx}{x^*x} : x \in \mathbb{C}^N \setminus \{0\} \right\}.$$

Fordert man nun $\mathcal{F}(A) \subseteq G$, so ist $V_m f(H_m)e_1$ mittels (2.3) und (2.4) wohldefiniert. Der wesentliche Grund besteht in den beiden Inklusionen

$$\sigma(H_m) \subseteq \mathcal{F}(H_m) \subseteq \mathcal{F}(A),$$

die im Folgenden gezeigt werden.

Lemma 2.1.4
Seien die Voraussetzungen aus Definition 2.1.3 gegeben. Ist $\mathcal{H}$ endlich-dimensional, so gilt $\sigma(B) \subseteq \mathcal{F}(B)$.

Beweis:
Sei $\lambda \in \sigma(B)$. Dann gibt es ein $z \in \mathcal{H}$ mit $Bz = \lambda z$ und $\|z\| = 1$. Folglich ist

$$\langle Bz, z\rangle = \langle \lambda z, z\rangle = \lambda\langle z, z\rangle = \lambda$$

und damit ist $\lambda \in \mathcal{F}(A)$.

■

Folgerung 2.1.5
Für die Matrizen aus Lemma 2.1.2 gilt

$$\sigma(H_m) \subseteq \mathcal{F}(H_m) \subseteq \mathcal{F}(A).$$

Beweis:
Die erste Inklusion folgt direkt aus Lemma 2.1.4.

Für die zweite Inklusion werden hingegen die Ergebnisse aus Lemma 2.1.2 benötigt. Da die Spalten von V_m zueinander orthogonal sind, gilt

$$V_m^* V_m = I. \tag{2.5}$$

Dies impliziert

$$\|V_m\| = 1, \tag{2.6}$$

denn für alle $z \in \mathbb{C}^m$ mit $\|z\| = 1$ ist

$$\|V_m z\|^2 = (V_m z)^* V_m z = z^* V_m^* V_m z \overset{(2.5)}{=} z^* z = 1. \tag{2.7}$$

Außerdem gilt

$$z^* H_m z \overset{(2.2)}{=} z^* V_m^* A V_m z = \left(V_m z\right)^* A V_m z$$

für $z \in \mathbb{C}^m$. Ist nun $\lambda = z^* H_m z \in \mathcal{F}(H_m)$ mit $\|z\| = 1$, dann ist nach (2.7) auch $\|V_m z\| = 1$ und damit $\lambda = \left(V_m z\right)^* A V_m z \in \mathcal{F}(A)$.

■

Unter der Voraussetzung $\mathcal{F}(A) \subseteq G$ kann also die gewünschte Approximation

$$f(A)v \approx \frac{1}{2\pi i} \int_{\partial G} f(z) V_m (zI - H_m)^{-1} e_1 dz \tag{2.8}$$

verwendet werden.

2.2. Fehlerschranken für das Arnoldi-Verfahren

Im Folgenden seien stets $A \in \mathbb{C}^{N \times N}$, $v \in \mathbb{C}^N$ mit $\|v\| = 1$ und die zugehörigen Matrizen aus Lemma 2.1.2 gegeben. Mit $\|\cdot\|$ wird die euklidische Norm bzw. die zugehörige Operatornorm bezeichnet.

Außerdem wird die Notation durch

$$w(B) := \sup \{|\lambda| : \ \lambda \in \mathcal{F}(B)\}$$

vereinfacht. Man bezeichnet $w(B)$ als numerischen Radius.

2.2.1. Allgemeine Fehlerabschätzung

Die Formel (2.8) zeigt, dass der erste Schritt nun darin bestehen muss, die in (2.4) eingeführte Approximation

$$(zI - A)^{-1} v \approx V_m (zI - H_m)^{-1} e_1$$

zu analysieren. Hierzu erweist sich der folgende Abstandsbegriff als nützlich.

Seien $S \subseteq \mathbb{C}$ und $z \in \mathbb{C}$, dann wird der Abstand dist des Punktes z von der Menge S durch $\operatorname{dist}(z, S) := \inf \{|z - \xi| : \ \xi \in S\}$ definiert. Auf Basis dieser Definition liefert das folgende Lemma Abschätzungen für die Terme $\left\|(zI - A)^{-1}\right\|$ und $\left\|(zI - H_m)^{-1}\right\|$.

Lemma 2.2.1
Seien $B \in \mathbb{C}^{k \times k}$ für ein $k \in \mathbb{N}_1$ und $S \subseteq \mathbb{C}$ nichtleer, abgeschlossen und konvex. Seien außerdem $z \in \mathbb{C} \setminus S$ und $\mathcal{F}(B) \subseteq S$. Dann gilt

$$\left\|(zI - B)^{-1}\right\| \leq \operatorname{dist}(z, S)^{-1}.$$

Beweis:
Der in [39, 3.] eingeführte M-numerische Wertebereich ist eine Verallgemeinerung der klassischen Definition 2.1.3, welche dem Fall $M = 1$ entspricht. Die zu zeigende Behauptung folgt daher aus [39, Satz 4.1].

■

Das Lemma 2.2.1 wird jetzt verwendet, um eine Schranke für den Fehler der Näherung (2.4) zu finden. Um die Notation zu vereinfachen, definiere aber zunächst noch

$$d(S) := \inf \{|z - \xi| : \ \xi \in S, \ z \in \mathcal{F}(A)\} = \inf \{ \operatorname{dist}(z, S) : \ z \in \mathcal{F}(A)\}$$

für $S \subseteq \mathbb{C}$.

Bemerkung:
Im Gegensatz zum Hausdorff-Abstand wird zweimal das Infimum gebildet. Die Verwendung dieses Abstandsbegriffs ist damit begründet, dass im Folgenden $\mathcal{F}(A) \subseteq S$ ist und sichergestellt werden muss, dass der Abstand von $\mathcal{F}(A)$ zu jedem Randpunkt von S nach unten beschränkt ist.

Eine Schranke für $\|(zI-A)^{-1}v-V_m(zI-H_m)^{-1}e_1\|$ liefert nun das Lemma 2.2.2.

Lemma 2.2.2
Seien G ein Gebiet und $E \subseteq \mathbb{C}$ kompakt und konvex mit $\mathcal{F}(A) \subseteq E \subseteq G$. Außerdem seien $z \in \partial G$ und P_m ein Polynom mit $\deg(P_m) \leq m$ und $P_m(z) = 1$. Dann ist

$$\left\|(zI - A)^{-1}v - V_m(zI - H_m)^{-1}e_1\right\| \leq 2d(\partial G)^{-1} \cdot \|P_m(A)\|.$$

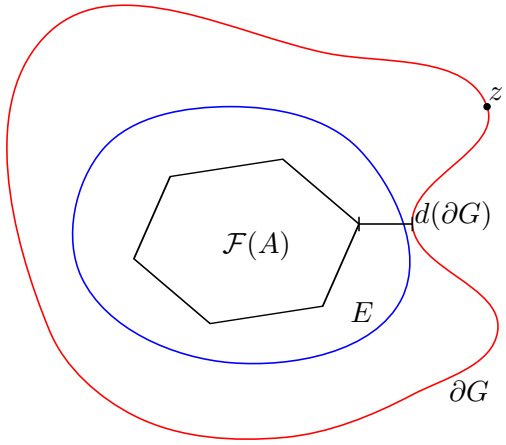


Abbildung 2.1.: Konstellation in Lemma 2.2.2

Beweis:
Da z eine Nullstelle von $P_m - 1$ ist, gibt es ein Polynom Q_m mit

$$P_m(X) - 1 = (z - X)Q_m(X).$$

Aus $\deg Q_m \leq m - 1$ folgt dabei $Q_m(A)v \in \text{span}\{v, Av, ..., A^{m-1}v\} = \mathcal{K}_m(A, v)$. Weil zudem die Spalten von V_m eine Basis von $\mathcal{K}_m(A, v)$ bilden, gibt es einen Vektor $y_m \in C^m$, so dass $V_m y_m = Q_m(A)v$ gilt.

Dies impliziert jedoch

$$\begin{aligned} P_m(A)v &= v - (zI - A)Q_m(A)v \\ &= v - (zI - A)V_m y_m. \end{aligned} \tag{2.9}$$

Damit ist bereits eine nützliche Darstellung für $P_m(A)$ gefunden worden. Nun gilt es die linke Seite der zu zeigenden Ungleichung geeignet umzuformen. Definiere hierzu

$$\Delta_m = (zI - A)^{-1} - V_m(zI - H_m)^{-1}V_m^*.$$

Da die Spalten von $V_m = (v_1, ..., v_m)$ orthogonal zueinander sind und $v = v_1$ ist, gilt $V_m^* v = V_m^* v_1 = e_1$. Daraus ergibt sich

$$\begin{aligned} \Delta_m v &= (zI - A)^{-1} v - V_m (zI - H_m)^{-1} V_m^* v \\ &= (zI - A)^{-1} v - V_m (zI - H_m)^{-1} e_1. \end{aligned} \tag{2.10}$$

Wegen $V_m^* V_m = I$ und $V_m^* A V_m = H_m$ ist außerdem

$$\begin{aligned} \Delta_m (zI - A) V_m &= V_m - V_m (zI - H_m)^{-1} V_m^* (zI - A) V_m \\ &= V_m - V_m (zI - H_m)^{-1} (z V_m^* V_m - V_m^* A V_m) \\ &= V_m - V_m (zI - H_m)^{-1} (zI - H_m) \\ &= V_m - V_m = 0. \end{aligned} \tag{2.11}$$

Zusammengefasst erhält man demnach

$$\begin{aligned} (zI - A)^{-1} v - V_m (z - H_m)^{-1} e_1 &\overset{(2.10)}{=} \Delta_m v \overset{(2.11)}{=} \Delta_m (v - (zI - A) V_m y_m) \\ &\overset{(2.9)}{=} \Delta_m P_m(A) v. \end{aligned}$$

Da $\|v\| = 1$ gilt, ist $\|P_m(A)v\| \leq \|P_m(A)\|$, was

$$\begin{aligned} \left\| (zI - A)^{-1} v - V_m (zI - H_m)^{-1} e_1 \right\| &\leq \|\Delta_m\| \cdot \|P_m(A) v\| \\ &\leq \|\Delta_m\| \cdot \|P_m(A)\| \end{aligned} \tag{2.12}$$

impliziert.

Nun muss nur noch der Faktor $\|\Delta_m\|$ abgeschätzt werden. Wegen $z \in \partial G$ und $E \subseteq G = G^\circ$ ist $z \notin E$. Aus $\mathcal{F}(H_m) \subseteq \mathcal{F}(A) \subseteq E$ folgt daher mit Lemma 2.2.1

$$\left\| (zI - A)^{-1} \right\| \leq \operatorname{dist}(z, \mathcal{F}(A))^{-1}$$

und

$$\left\| (zI - H_m)^{-1} \right\| \leq \operatorname{dist}(z, \mathcal{F}(H_m))^{-1} \leq \operatorname{dist}(z, \mathcal{F}(A))^{-1}.$$

Die Ungleichung $\operatorname{dist}(z, \mathcal{F}(A)) \leq \operatorname{dist}(z, \mathcal{F}(H_m))$ ist dabei eine direkte Konsequenz aus der Definition des Abstands über das Infimum und der Inklusion $\mathcal{F}(H_m) \subseteq \mathcal{F}(A)$.

Weil $\|V_m^*\| = \|V_m\| = 1$ ist, erhält man hieraus die Abschätzung

$$\begin{aligned} \|\Delta_m\| &= \left\| (zI - A)^{-1} - V_m (zI - H_m)^{-1} V_m^* \right\| \\ &\leq \left\| (zI - A)^{-1} \right\| + \|V_m\| \cdot \left\| (zI - H_m)^{-1} \right\| \cdot \|V_m^*\| \\ &\leq 2 \cdot \operatorname{dist}(z, \mathcal{F}(A))^{-1} \leq 2 d(\partial G)^{-1}, \end{aligned}$$

wobei auch hier die letzte Ungleichung eine direkte Konsequenz aus der Definition über das Infimum und der Voraussetzung $z \in \partial G$ ist.

Zusammen mit (2.12) liefert $\|\Delta_m\| \leq 2 d(\partial G)^{-1}$ schließlich die Behauptung des Satzes.

■

Die Fehlerschranke in Lemma 2.2.2 ist von der Wahl des Polynoms P_m abhängig, wobei die Suche nach einem geeigneten Vertreter zur Theorie der Faber-Polynome (siehe [4], [7], [40]) führt. Demnach existiert zu einer nicht-leeren, konvexen und kompakten Teilmenge E der komplexen Ebene eine biholomorphe Abbildung

$$\psi : \mathbb{C} \setminus \overline{D_1} \longrightarrow \mathbb{C} \setminus E$$

mit $\psi(w) = \rho w + \mathcal{O}(1),\ w \to \infty$ und $\rho > 0$. Dabei bezeichnet $\overline{D_1} = \{w \in \mathbb{C} :\ |w| \leq 1\}$ den Abschluss der Einheitskreisscheibe. Die Umkehrabbildung von ψ sei

$$\phi : \mathbb{C} \setminus E \longrightarrow \mathbb{C} \setminus \overline{D_1}. \tag{2.13}$$

Das n-te Faber-Polynom ist dann der Nebenteil von ϕ^n.
Wegen $\phi(z) = \frac{z}{\rho} + \mathcal{O}(1),\ z \to \infty$ ist dies ein Polynom vom Grad n, wobei $\left(\frac{z}{\rho}\right)^n$ der Term mit der höchsten Potenz ist.

Lemma 2.2.3
Sei wiederum $z \in \partial G$. Unter den zusätzlichen Voraussetzungen, dass E keine Linie und $\mathcal{F}(A) \subseteq E^\circ$ ist, gibt es ein Polynom P_m mit $\deg(P_m) \leq m$ und $P_m(z) = 1$, so dass die Abschätzung

$$\|P_m(A)\| \leq \frac{1}{2}\frac{\ell(\partial E)}{d(\partial E)}|\phi(z)|^{-m}$$

mit der zu E zugehörigen Funktion ϕ aus (2.13) gilt.

Bemerkung:
Im Fall $\mathcal{F}(A) \cap \partial E \neq \emptyset$ geht dies in die Aussage $\|P_m(A)\| \leq \infty$ über.

Beweis:
Unter obigen Voraussetzungen ist

$$P_m(A) = \frac{1}{2\pi \mathbf{i}} \int_{\partial E} P_m(\lambda)(\lambda I - A)^{-1} d\lambda.$$

Nach Lemma 2.2.1 ist $\left\|(\mu I - A)^{-1}\right\| \leq \operatorname{dist}(\mu, \mathcal{F}(A))^{-1} \leq d(\partial E)^{-1}$ für $\mu \in \partial E$, so dass

$$\|P_m(A)\| \leq \frac{1}{2\pi} \cdot \ell(\partial E) \cdot \max_{\mu \in E} |P_m(\mu)| \cdot d(\partial E)^{-1} \tag{2.14}$$

gilt. Sei $\phi_m(\omega)$ das zu E zugehörige Faber-Polynom von Grad m. Dann ist $\phi_m(\omega)$ der Nebenteil von $\phi(\omega)^m$ und damit gilt $\phi(\omega)^m = \phi_m(\omega) + \mathcal{O}(\frac{1}{\omega}),\ \ \omega \to \infty$.
Definiere nun

$$P_m(X) = \frac{\phi_m(X) - \phi_m(z) + \phi(z)^m}{\phi(z)^m},$$

so das insbesondere $P_m(z) = 1$ ist und daher die obigen Voraussetzungen an P_m erfüllt sind. Nach [31, Thm. 2] gilt für $\mu \in \mathbb{C} \setminus E$

$$|\phi_m(\mu) - \phi(\mu)^m| \leq 1.$$

Dann aber folgt

$$\max_{\mu \in \partial E} |\phi_m(\mu)| \leq 2$$

und da das Maximum nach dem Maximumsprinzip auf dem Rand angenommen wird, erhält man

$$\begin{aligned}\max_{\mu\in E}|P_m(\mu)| = \max_{\mu\in\partial E}|P_m(\mu)| &= \left(\max_{\mu\in\partial E}|\phi_m(\mu)+\phi_m(z)-\phi(z)^m|\right)\cdot|\phi(z)|^{-m}\\ &\le \left(\max_{\mu\in\partial E}|\phi_m(\mu)|+|\phi_m(z)-\phi(z)^m|\right)\cdot|\phi(z)|^{-m} \le 3|\phi(z)|^{-m}.\end{aligned}$$

Mittels Einsetzen in (2.14) erhält man die Aussage des Lemmas.

■

Da der Fall, dass E eine Linie ist, nicht in das Schema dieses Lemmas passt, wird dieser nun noch gesondert betrachtet. Außerdem liefert das nun folgende Lemma eine zusätzliche Schranke für den Fall, dass E eine Kreisscheibe ist.

Lemma 2.2.4
Seien G ein Gebiet, $E \subseteq \mathbb{C}$ eine Kreisscheibe oder Linie mit $\mathcal{F}(A) \subseteq E \subseteq G$ und $z \in \partial G$. Dann gibt es ein Polynom P_m mit $\deg(P_m) \le m$ und $P_m(z) = 1$, so dass die Abschätzung

$$\|P_m(A)\| \le 3\big|\phi(z)\big|^{-m}$$

mit der zu E zugehörigen Funktion ϕ aus (2.13) gilt.

Beweis:
Der Fall einer Kreisscheibe $\{\omega : |\omega - z_0| \le \rho\}$ wird in [3, Seite 3] betrachtet. Das zugehörige Faber-Polynom ϕ_m ist durch $\phi_m(z) = \left(\frac{z-z_0}{\rho}\right)^m$ gegeben.

Wähle nun als Polynom P_m

$$P_m(X) = \left(\frac{X-z_0}{z-z_0}\right)^m.$$

Mit Hilfe der Rechenregeln für den numerischen Wertebereich [39, (3.3)] ist

$$\begin{aligned}w\left(\frac{1}{z-z_0}(A-z_0I)\right) &= \sup\left|\mathcal{F}\left(\frac{1}{z-z_0}(A-z_0I)\right)\right| = \sup\left|\frac{\mathcal{F}(A)-z_0}{z-z_0}\right|\\ &\le \sup_{\mu\in E}\left|\frac{\mu-z_0}{z-z_0}\right|\end{aligned} \tag{2.15}$$

und weil der numerische Wertebereich nach Voraussetzung in der obigen Kreisscheibe enthalten ist, gilt für $\mu \in E$

$$\left|\frac{\mu-z_0}{\rho}\right| \le 1. \tag{2.16}$$

Da nach [3, Seite 3] erstens $\|A\| \le 2w(A)$ und zweitens $w(A^k) \le \big(w(A)\big)^k$ für $k \in \mathbb{N}_1$ gilt, erhält man

$$\begin{aligned}\|P_m(A)\| &\le 2w(P_m(A)) \le 2P_m\big(w(A)\big) \overset{(2.15)}{\le} 2\sup_{\mu\in E}|P_m(\mu)|\\ &= 2\sup_{\mu\in E}\left|\left(\frac{\mu-z_0}{\rho}\right)\right|^m\left|\left(\frac{\rho}{z-z_0}\right)\right|^m \overset{(2.16)}{\le} 2\big|\phi(z)\big|^{-m},\end{aligned}$$

was zu zeigen war.

Betrachte nun den Fall, dass E eine Strecke ist. Das Ziel ist hier zunächst, A mittels einer linearen Transformation in eine hermitesche Matrix B zu überführen.

Wegen $\sigma(A) \subseteq \mathcal{F}(A) \subseteq E$ gibt es $\alpha, \beta \in \mathbb{C}$, $\beta \neq 0$, so dass für die Eigenwerte

$$\lambda_j \in \{\alpha + x\beta : x \in \mathbb{R}\}$$

für $j = 1, ..., N$ gilt.

Definiere nun $B := \frac{1}{\beta}A - \alpha I$ und außerdem $B_1 := \frac{1}{2}(B + B^*)$ sowie $B_2 := \frac{1}{2i}(B - B^*)$. Dann ist $B = B_1 + \mathbf{i}B_2$ und B_1, B_2 sind hermitesche Matrizen.

Nach [39, 3.3] gilt $\mathcal{F}(B) = \frac{1}{\beta}\mathcal{F}(A) - \alpha$ und damit $\mathcal{F}(B) \subseteq \mathbb{R}$. Wegen B_1 hermitesch ist ebenso $\mathcal{F}(B_1) \subseteq \mathbb{R}$ und folglich gilt $\mathcal{F}(\mathbf{i}B_2) = \mathcal{F}(B) - \mathcal{F}(B_1) \subseteq \mathbb{R}$.
Andererseits ist B_2 hermitesch, so dass $\mathbf{i}v^*B_2v \in \mathbf{i}\mathbb{R}$ für $v \in \mathbb{C}^N$ ist. Also muss $v^*B_2v = 0$ und wegen v beliebig $B_2 = 0$ gelten. Damit ist $B = B_1 + 0$ hermitesch.

Zusammengefasst gilt also, dass sich A für gewisse $\alpha, \beta \in \mathbb{C}$ und eine hermitesche Matrix B als $A = \alpha I + \beta B$ schreiben lässt .

Folglich existiert eine Orthonormalbasis aus Eigenvektoren von B. Seien $\mu_1, ..., \mu_N$ die Eigenwerte von B und $v_1, ..., v_N$ die zugehörigen orthonormalen Eigenvektoren. A und B haben dieselben Eigenvektoren, die Eigenwerte von A sind $\alpha + \beta\mu_1, ..., \alpha + \beta\mu_N$, denn für $j = 1, ..., N$ gilt

$$Av_j = (\alpha I + \beta B)v_j = \alpha I v_j + \beta B v_j = \alpha v_j + \beta\mu_j v_j = (\alpha + \beta\mu_j)v_j.$$

Ist nun $v \in \mathbb{C}^N$ mit $\|v\| = 1$ gegeben, so gilt

$$\begin{aligned}\|Av\| &= \left\| A\left(\sum_{j=1}^{N}\langle v, v_j\rangle v_j\right)\right\| = \left\|\left(\sum_{j=1}^{N}\langle v, v_j\rangle(\alpha + \beta\mu_j)v_j\right)\right\| \\ &\leq \rho(A)\left\|\sum_{j=1}^{N}\langle v, v_j\rangle v_j\right\| = \rho(A)\end{aligned}$$

und das Bilden des Supremums über v liefert $\|A\| = \rho(A)$. Da auf Grund der Submultiplikativität der Matrixnorm $\|A^k\| \leq \|A\|^k \leq \rho(A)^k$ folgt, erhält man zusammen mit $\sigma(A) \subseteq E$, dass

$$\|P_m(A)\| \leq \left|P_m\big(\rho(A)\big)\right| \leq \max_{\mu \in E}|P_m(\mu)|$$

ist.

Gemäß [40, (3.12)] ist $\left|\phi_m(\mu) - \phi(\mu)^m\right| \leq 1$ für $\mu \in \mathbb{C} \setminus E$, so dass nach der Wahl

$$P_m(X) := \frac{\phi_m(X) - \phi_m(z) + \phi(z)^m}{\phi(z)^m}$$

genau wie im Beweis des vorherigen Satzes argumentiert werden kann.

■

Auf Basis der obigen Lemmata ist es nun möglich, den nachfolgenden Satz zu beweisen, der eine allgemeine Schranke für den Fehler der Arnoldi-Approximation liefert.

Satz 2.2.5
Seien die Annahmen aus Lemma 2.2.2 gegeben. Unter den Voraussetzungen von Lemma 2.2.3 gilt für den Fehler der Arnoldi-Approximation $\varepsilon_m = \|f(A)v - V_m f(H_m)e_1\|$ *die Abschätzung*

$$\varepsilon_m \leq \frac{M}{2\pi} \int_{t_A}^{t_E} |f(\gamma(\theta))| \cdot |\phi(\gamma(\theta))|^{-m} \cdot |\gamma'(\theta)| \, d\theta,$$

wobei $M = \frac{\ell(\partial E)}{d(\partial E)d(\partial G)}$ *und* $\gamma : [t_A, t_E] \longrightarrow \partial G$ *eine geeignete Parametrisierung von* ∂G *ist.*

Unter den Voraussetzungen von Lemma 2.2.4 erhält man dieselbe Abschätzung jedoch mit der Konstanten $M = \frac{6}{d(\partial G)}$.

Beweis:
Setzt man die Ungleichung aus Lemma 2.2.4 bzw. Lemma 2.2.3 in Lemma 2.2.2 ein, so erhält man

$$\|(zI - A)^{-1}v - V_m(zI - H_m)^{-1}e_1\| \leq M \cdot |\phi(z)|^{-m}$$

für $z \in \partial G$. Die Behauptung des Satzes folgt daraus, da

$$\begin{aligned}
\varepsilon_m &= \|f(A)v - V_m f(H_m)e_1\| \\
&= \left\| \frac{1}{2\pi i} \int_{\partial G} f(z)(zI - A)^{-1}v dz - \frac{1}{2\pi i} \int_{\partial G} f(z) V_m (zI - H_m)^{-1} e_1 dz \right\| \\
&= \left\| \frac{1}{2\pi i} \int_{\partial G} f(z) \left((zI - A)^{-1}v - V_m(zI - H_m)^{-1}e_1 \right) dz \right\| \\
&\leq \frac{1}{2\pi} \int_{t_A}^{t_E} |f(\gamma(\theta))| \cdot \left\| \Big((\gamma(\theta))I - A \Big)^{-1} v - V_m \Big((\gamma(\theta))I - H_m \Big)^{-1} e_1 \right\| \cdot |\gamma'(\theta)| \, d\theta
\end{aligned}$$

ist.

■

2.2.2. Fehlerabschätzung für hermitesche Matrizen

Auf Basis des allgemeinen Resultats in Satz 2.2.5 können nun spezifische Fehlerschranken für verschiedene Klassen von Matrizen bestimmt werden. Der nun folgende Satz beschreibt den Fall einer hermiteschen Matrix. Die Einschränkung, dass A keine positiven Eigenwerte besitzen darf, ist nicht so restriktiv, wie es zunächst scheint, da man statt A eine Matrix $B = A + \alpha I$ für ein genügend großes $\alpha \in \mathbb{R}$ betrachten kann. Entscheidend ist hierbei der Zusammenhang

$$e^{\tau B}v - V_m e^{\tau \widetilde{H}_m} e_1 = \left(e^{\tau A + \alpha I}v - V_m e^{\tau H_m + \alpha I} e_1 \right) = e^{\tau \alpha} \left(e^{\tau A}v - V_m e^{\tau H_m} e_1 \right).$$

Dabei sei $\widetilde{H}_m$ die Hessenbergmatrix, die bei der Krylow-Unterraum-Approximation von B gebildet wird.

Satz 2.2.6
Sei A eine hermitesche, negativ semidefinite Matrix mit $\sigma(A) \subseteq [-4\rho, 0]$ für gegebenes $\rho > 0$. Dann gilt für den Fehler der Arnoldi-Approximation von $e^{\tau A}v$

$$\varepsilon_m \leq 10e^{-\frac{m^2}{5\rho\tau}},$$

falls $\sqrt{4\rho\tau} \leq m \leq 2\rho\tau$ ist und

$$\varepsilon_m \leq 10(\rho\tau)^{-1}e^{-\rho\tau}\left(\frac{e\rho\tau}{m}\right)^m,$$

falls $m \geq 2\rho\tau$ ist.

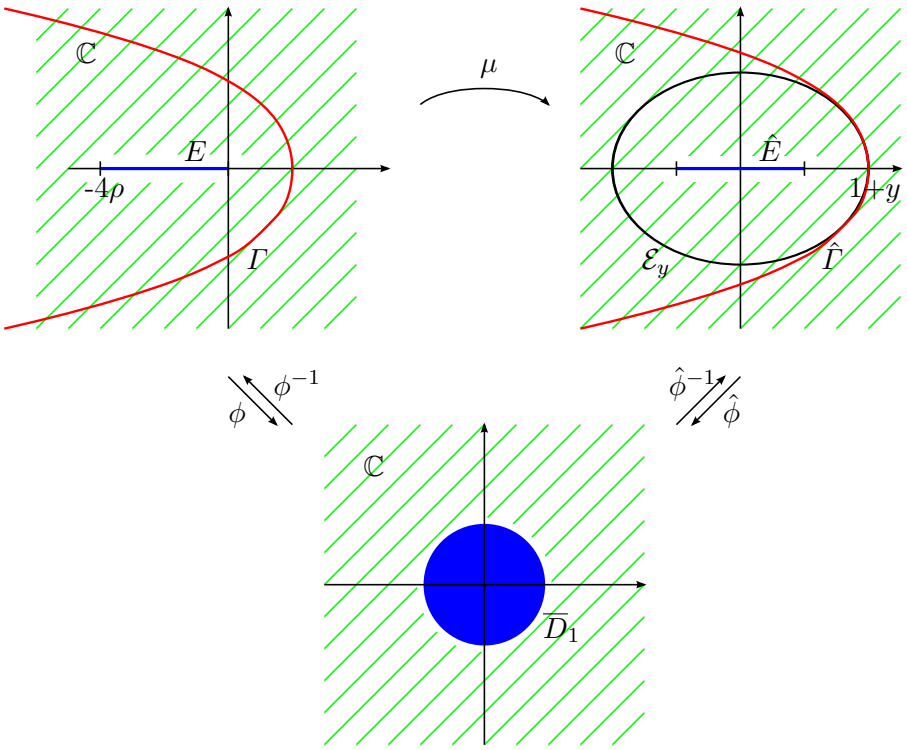


Abbildung 2.2.: Konstellation in Satz 2.2.6

Beweis:
Da der numerische Wertebereich $\mathcal{F}(A)$ im hermiteschen Fall das kleinste Intervall ist, das alle Eigenwerte enthält (siehe [27]), gilt $\mathcal{F}(A) \subseteq [-4\rho, 0]$ und es ist zweckmäßig, Satz 2.2.5 mit $E = [-4\rho, 0]$ anzuwenden.

Um die Rechnungen im Folgenden zu erleichtern, definiert man $\hat{E} = [-1, 1]$ und führt die affin-lineare Transformation

$$\begin{aligned} \mu : \mathbb{C} &\longrightarrow \mathbb{C} \\ z &\longmapsto 1 + \frac{z}{2\rho} \end{aligned}$$

ein, die $E = [-4\rho, 0]$ auf $\hat{E} = [-1, 1]$ abbildet.

Die Anwendung von Satz 2.2.5 erfordert es, eine biholomorphe Funktion ϕ bzw. $\hat{\phi}$ zu finden, die $\mathbb{C}\setminus E$ bzw. $\mathbb{C}\setminus\hat{E}$ auf das Komplement der abgeschlossenen Einheitskreisscheibe $\mathbb{C}\setminus\overline{D_1}$ abbildet. Die nötigen Grundlagen aus der Funktionentheorie werden im Anhang bereit gestellt. Danach existiert eine holomorphe Abbildung

$$g : \{z^2-1 : z \in \mathbb{C}\setminus[-1,1]\} \longrightarrow \mathbb{C}$$

mit $\left(g(z^2-1)\right)^2 = z^2-1$ für alle $z \in \mathbb{C}\setminus[-1,1]$, so dass die Funktion

$$\begin{aligned}\hat{\phi} : \mathbb{C}\setminus[-1,1] &\longrightarrow \mathbb{C}\setminus\overline{D_1}\\ z &\longmapsto z+g(z^2-1)\end{aligned}$$

biholomorph ist und die Eigenschaft $\hat{\phi}(z) = 2z + \mathcal{O}(1),\ z\to\infty$ besitzt.

Ist $\hat{\phi}$ bekannt, so erhält man ϕ durch Transformation, d.h.

$$\begin{aligned}\phi : \mathbb{C}\setminus E &\longrightarrow \mathbb{C}\setminus\overline{D_1},\\ z &\longmapsto \hat{\phi}(\mu(z)) = \hat{\phi}\left(1+\frac{z}{2\rho}\right),\end{aligned}$$

denn diese Abbildung erfüllt nach Konstruktion die geforderten Voraussetzungen. Insbesondere gilt $\phi(z) = \frac{z}{\rho} + \mathcal{O}(1),\ z\to\infty$.

Nun ist noch ein Integrationsweg zu wählen. Betrachte hierzu zunächst für $x>0$ die Ellipsen $\mathcal{E}_x$ mit der Parametrisierung

$$\begin{aligned}\xi_x : [0,2\pi) &\longrightarrow \mathcal{E}_x,\\ \theta &\longmapsto (1+x)\cos(\theta) + \mathbf{i}\sqrt{x^2+2x}\sin(\theta)\end{aligned}$$

und zudem die Parabel $\hat{\Gamma}$, die durch

$$\begin{aligned}\hat{\gamma} : \mathbb{R} &\longrightarrow \hat{\Gamma}\\ \theta &\longmapsto (1+y)\left(1-\tfrac{1}{2}\theta^2\right) + \mathbf{i}\sqrt{y^2+2y}\ \theta\end{aligned}$$

gegeben ist. Dabei ist $y>0$ ein freier Parameter, der noch zu wählen ist.

Aus der Taylorentwicklung von sin und cos erkennt man, dass $\hat{\Gamma}$ die Ellipse $\mathcal{E}_y$ oskuliert, wobei allerdings $\mathcal{E}_y \cap \hat{\Gamma} = \{1+y\}$ gilt. Ist nämlich $\xi_y(\theta_1) = \hat{\gamma}(\theta_2)$ mit $\theta_1 \neq 0$, dann impliziert $1-\frac{1}{2}\theta_2^2 = \cos(\theta_1) > 1-\frac{1}{2}\theta_1^2$ die Bedingung $|\theta_1| > |\theta_2|$. Wegen $|\theta_2| = |\sin(\theta_1)| < |\theta_1|$ ist dies ein Widerspruch. Die Parabel verläuft also in den übrigen Punkten außerhalb der Ellipse.

Als Integrationsweg Γ wird nun das Urbild von $\hat{\Gamma}$ unter der Transformation μ gewählt, d.h. es gilt $\gamma(\theta) = 2\rho(\hat{\gamma}(\theta)-1)$, was $\gamma'(\theta) = 2\rho\hat{\gamma}'(\theta)$ impliziert. Außerdem ist $\phi(\gamma(\theta)) = \hat{\phi}(\hat{\gamma}(\theta))$ erfüllt und wegen $\mathcal{F}(A)\subseteq E$ gilt

$$\delta := \inf_{\lambda\in\Gamma}\inf_{x\in E}|\lambda-x| \geq \inf_{\lambda\in\Gamma}\inf_{z\in\mathcal{F}(A)}|\lambda-z| =: d,$$

so dass man mit Satz 2.2.5 die Ungleichung

$$\begin{aligned}\varepsilon_m &\leq \frac{6}{2\pi d}\int_{-\infty}^{\infty}\left|e^{\gamma(\theta)\tau}\right|\cdot\left|\phi(\gamma(\theta))\right|^{-m}\cdot\left|\gamma'(\theta)\right|d\theta\\ &\leq \frac{2\rho}{\delta}\int_{-\infty}^{\infty}\left|e^{2\rho\tau\left(\hat{\gamma}(\theta)-1\right)}\right|\cdot\left|\hat{\phi}(\hat{\gamma}(\theta))\right|^{-m}\cdot\left|\hat{\gamma}'(\theta)\right|d\theta\end{aligned}$$

erhält.

Die vier auftretenden Faktoren gilt es nun geeignet abzuschätzen. Hierbei ist zunächst zu zeigen, dass $\delta = 2\rho y$ gilt. Dieser Wert entspricht genau dem Abstand des Scheitelpunkts der Parabel zur Strecke. Da sich die Transformation μ aus einer Translation, unter der der Abstand invariant ist, und einer Streckung zusammensetzt, genügt es stattdessen

$$\inf_{\lambda\in\hat{\Gamma}}\inf_{x\in\hat{E}}|\lambda - x| = y$$

zu zeigen.
Für $x \in \hat{E} = [-1,1]$ erhält man aus der Definition von $\hat{\gamma}$

$$\begin{aligned}
|\hat{\gamma}(\theta) - x|^2 &= \Big((1-x) + y - (1+y)\tfrac{1}{2}\theta^2\Big)^2 + (y^2+2y)\theta^2 \\
&= (1-x)^2 + y^2 + (1+y)^2\tfrac{1}{4}\theta^4 + 2(1-x)y \\
&\quad - (1-x)(1+y)\theta^2 - y(1+y)\theta^2 + (y^2+2y)\theta^2 \\
&\geq y^2 + (1-x)^2 + (1+y)^2\tfrac{1}{4}\theta^4 - (1-x)(1+y)\theta^2 \\
&\geq y^2,
\end{aligned}$$

wobei der letzte Schritt durch Substitution aus einer binomischen Formel folgt. Dies beweist die Behauptung, dass der minimale Abstand tatsächlich im Scheitelpunkt der Parabel angenommen wird.

Beim zweiten Faktor erhält man wegen $1 + y \geq 1$

$$\begin{aligned}
\left|e^{2\rho\tau\left(\hat{\gamma}(\theta)-1\right)}\right| &= \left|e^{2\rho\tau\left((1+y)\left(1-\frac{1}{2}\theta^2\right)+\mathbf{i}\sqrt{2y+y^2}\theta-1\right)}\right| \\
&= e^{2\rho\tau\left(y-(1+y)\frac{1}{2}\theta^2\right)} \\
&= e^{2\rho\tau y}\cdot e^{-(1+y)\rho\tau\theta^2} \\
&\leq e^{2\rho\tau y}\cdot e^{-\rho\tau\theta^2}.
\end{aligned}$$

Als Nächstes gilt es $\left|\hat{\phi}(\hat{\gamma}(\theta))\right|$ wegen der negativen Potenz nach unten abzuschätzen. Sei $\hat{\phi}(\hat{\gamma}(\theta)) = Re^{\mathbf{i}\alpha}$ mit $R > 1$. Dann erhält man

$$\begin{aligned}
\hat{\phi}(\hat{\gamma}(\theta)) = \hat{\phi}^{-1}(Re^{\mathbf{i}\alpha}) = \hat{\psi}(Re^{\mathbf{i}\alpha}) &= \frac{1}{2}\left(Re^{\mathbf{i}\alpha} + \frac{1}{R}e^{-\mathbf{i}\alpha}\right) \\
&= \frac{1}{2}\left(\left(R+\frac{1}{R}\right)\cos(\alpha) + \mathbf{i}\left(R-\frac{1}{R}\right)\sin(\alpha)\right)
\end{aligned}$$

und folglich ist $\hat{\phi}(\hat{\gamma}(\theta)) \in \mathcal{E}_{\frac{1}{2}\left(R+\frac{1}{R}\right)} = \mathcal{E}_{\hat{\phi}^{-1}(R)}$.

Da die Parabel $\hat{\Gamma}$ bis auf den Berührpunkt außerhalb der Ellipse $\mathcal{E}_y$ verläuft, gilt für gegebenes $\lambda \in \hat{\Gamma} \setminus \{1+y\}$, dass λ Element einer Ellipse $\mathcal{E}_x$ mit $x > y$ ist, denn die Ellipsen überdecken ganz $\mathbb{C} \setminus [-1,1]$ und haben untereinander keine Schnittpunkte. Wählt man nun $R > 1$ so, dass $|\hat{\phi}(\lambda)| = R$ gilt und definiert $r := |\hat{\phi}(1+y)|$, dann folgt aus $\frac{1}{2}\left(R+\frac{1}{R}\right) > \frac{1}{2}\left(r+\frac{1}{r}\right)$, dass $R > r$ ist.

Dies liefert nun die Abschätzung

$$\left|\hat{\phi}(\lambda)\right| \geq \left|\hat{\phi}(1+y)\right| = r \quad \forall\lambda \in \hat{\Gamma}.$$

Für den letzten Faktor folgt aus der Wahl von $\hat{\gamma}'(\theta)$ die Ungleichung

$$|\hat{\gamma}'(\theta)| = \left|(-1)\cdot(1+y)\theta + \mathbf{i}\sqrt{2y+y^2}\right| \leq (1+y)|\theta| + \sqrt{2y+y^2}.$$

Zusammengefasst erhält man also

$$\begin{aligned}
\varepsilon_m &\leq \frac{2\rho}{\delta}\int_{-\infty}^{\infty}\left|e^{2\rho\tau\left(\hat{\gamma}(\theta)-1\right)}\right|\cdot\left|\hat{\phi}(\hat{\gamma}(\theta))\right|^{-m}\cdot|\hat{\gamma}'(\theta)|\,d\theta \\
&\leq y^{-1}e^{2\rho\tau y}r^{-m}\int_{-\infty}^{\infty}e^{-\rho\tau\theta^2}\left((1+y)|\theta| + \sqrt{2y+y^2}\right)d\theta \\
&= e^{2\rho\tau y}r^{-m}\int_{0}^{\infty}e^{-\rho\tau\theta^2}\cdot 2y^{-1}\left((1+y)\theta + \sqrt{2y+y^2}\right)d\theta \\
&= e^{2\rho\tau y}r^{-m}\Big[-e^{-\rho\tau\theta^2}\frac{1+y}{y\rho\tau}\theta\Big]_0^{\infty} + \int_0^{\infty}2e^{-\rho\tau\theta^2}\sqrt{\frac{2+y}{y}}\,d\theta \\
&= e^{2\rho\tau y}r^{-m}\left(\frac{1+y}{\rho\tau y} + \sqrt{\frac{(2+y)\pi}{\rho\tau y}}\right),
\end{aligned}$$

wobei verwendet wurde, dass für $a > 0$

$$\int_0^{\infty}e^{-ax^2}dx = \frac{1}{2\sqrt{a}}\sqrt{\pi}$$

gilt.

Unter der zusätzlichen Voraussetzung $y \leq \frac{1}{2}$ soll nun eine geeignete Schranke für r gefunden werden. Hierbei ist entscheidend, dass das Polynom $0,0784 + 0,304x + 0,295x^2$ keine reellen Nullstellen besitzt und daher stets größer 0 ist. Außerdem ist

$$\sqrt{y^2+2y} - \sqrt{2y} = \frac{y^2}{\sqrt{y^2+2y}+\sqrt{2y}} \geq \frac{y^2}{\sqrt{1,25}+1}.$$

Daher gilt

$$\begin{aligned}
0,035y^2 &\leq 0,0784y - 0,304y\sqrt{y} + 0,33y^2 \\
&\leq 0,04\sqrt{2y} + 0,0784y - \frac{(0,96)^3 y\sqrt{2y}}{3} + \left(\frac{1}{\sqrt{1,25}+1} - \frac{(0,96)^4}{6}\right)y^2 \\
&\leq 1 + y + \sqrt{y^2+2y} - \sum_{k=0}^{4}\frac{\left(0,96\sqrt{2y}\right)^k}{k!}.
\end{aligned}$$

Wegen

$$\sum_{k=5}^{\infty}\frac{\left(0,96\sqrt{2y}\right)^k}{k!} \leq (0,96)^5\left(\frac{1}{5!}+\frac{1}{6!}+\frac{2}{7!}\right)4y^2 \leq 0,034y^2$$

folgt insgesamt

$$r = \left|\hat{\phi}(1+y)\right| = 1 + y + \sqrt{y^2+2y} \geq e^{0,96\sqrt{2y}}.$$

Wählt man nun $y = \frac{m^2}{8(\rho\tau)^2}$, dann ist $y \leq \frac{1}{2}$ äquivalent zu $m \leq 2\rho\tau$ und man erhält

$$\begin{aligned}\varepsilon_m &\leq e^{2\rho\tau y} r^{-m} \left(\frac{1+y}{\rho\tau y} + \sqrt{\frac{(2+y)\pi}{\rho\tau y}} \right) \\ &\leq \left(\frac{8\rho\tau}{m^2} + \frac{1}{\rho\tau} + \sqrt{\left(\frac{16\rho\tau}{m^2} + \frac{1}{\rho\tau} \right) \pi} \right) e^{\frac{1-2\cdot 0{,}96}{4\rho\tau} m^2} \\ &\leq \left(12\frac{\rho\tau}{m^2} + 8\frac{\sqrt{\rho\tau}}{m} \right) e^{-\frac{0{,}92}{4\rho\tau} m^2}.\end{aligned}$$

Falls wie vorausgesetzt $\sqrt{4\rho\tau} \leq m$ ist, liefert dies

$$\varepsilon_m \leq 7e^{-\frac{0{,}92}{4\rho\tau} m^2},$$

so dass der erste Teil des Satzes bewiesen ist.

Für den zweiten Teil muss der Fall $m \geq 2\rho\tau$ betrachtet werden. Wählt man $r = \frac{m}{\rho\tau}$, dann ist

$$y = \frac{1}{2}\left(r + \frac{1}{r} \right) - 1 = \frac{m}{2\rho\tau} + \frac{\rho\tau}{2m} - 1$$

und damit

$$\rho\tau y = \frac{m}{2} + \frac{(\rho\tau)^2}{2m} - \rho\tau \geq \frac{2\rho\tau}{2} + \frac{\rho\tau}{2} - \rho\tau \geq \frac{\rho\tau}{4}.$$

Als Fehlerabschätzung erhält man folglich

$$\begin{aligned}\varepsilon_m &\leq e^{2\rho\tau y} r^{-m} \left(\frac{1+y}{\rho\tau y} + \sqrt{\frac{(2+y)\pi}{\rho\tau y}} \right) \\ &\leq \left(\frac{4}{\rho\tau} + \frac{1}{\rho\tau} + \sqrt{\left(\frac{8}{\rho\tau} + \frac{1}{\rho\tau} \right) \pi} \right) e^m e^{\frac{(\rho\tau)^2}{m}} e^{-2\rho\tau} \left(\frac{\rho\tau}{m} \right)^m \\ &\leq \left(\frac{5}{\rho\tau} + 3\sqrt{\pi}\sqrt{\frac{1}{\rho\tau}} \right) e^{\frac{(\rho\tau)^2}{m}} e^{-2\rho\tau} \left(\frac{e\rho\tau}{m} \right)^m.\end{aligned}$$

Dies ist schon die zu zeigende Behauptung

$$\varepsilon_m \leq 10(\rho\tau)^{-1} e^{-\rho\tau} \left(\frac{e\rho\tau}{m} \right)^m,$$

denn es gilt

$$\begin{aligned}\left(\frac{5}{\rho\tau} + 3\sqrt{\pi}\sqrt{\frac{1}{\rho\tau}} \right) e^{\frac{(\rho\tau)^2}{m}} e^{-2\rho\tau} &\leq \left(\frac{5}{\rho\tau} + 3\sqrt{\pi}\sqrt{\frac{1}{\rho\tau}} \right) e^{-\frac{\rho\tau}{2}} e^{-\rho\tau} \\ &\leq 5(\rho\tau)^{-1} e^{-\rho\tau} \left(1 + \sqrt{2\rho\tau} e^{-\frac{\rho\tau}{2}} \right) \\ &\leq 10(\rho\tau)^{-1} e^{-\rho\tau}.\end{aligned}$$

■

Bemerkung:
Zu Vergleichszwecken, sei bemerkt, dass nach [19]

$$\|x - x_m\| \leq 2\sqrt{1+4\rho\tau}\left(1 - \frac{2}{\sqrt{1+4\rho\tau}}\right)^m$$

für den Fehler des Verfahrens der konjugierten Gradienten angewandt auf $(I - \tau A)x = v$ gilt.

Wird der Fehler auf einer halblogarithmischen Darstellung logarithmisch gegen m aufgetragen, so entspricht dies lediglich einer Graden.

Dies führt zu einem Vorteil der exponentiellen Integratoren gegenüber impliziten Verfahren, der jedoch durch eine gute Vorkonditionierung des Verfahrens der konjugierten Gradienten gegebenenfalls aufgehoben werden kann.

2.2.3. Beispiel mit radial beschränktem numerischen Wertebereich

Dieser Satz behandelt den Fall, dass der numerische Wertebereich von A in der abgeschlossenen Kreisscheibe mit Mittelpunkt $-\rho$ und Radius ρ enthalten ist. Im Gegensatz zum vorherigen Satz ist das zugehörige Faber-Polynom hier sehr einfach zu bestimmen.

Satz 2.2.7
Sei A eine Matrix mit $\mathcal{F}(A) \subseteq E := \{z : |z + \rho| \leq \rho\}$ für gegebenes $\rho > 0$. Dann gilt für den Fehler der Arnoldi-Approximation von $e^{hA}v$ die Abschätzung

$$\varepsilon_m \leq 12e^{-\rho\tau}\left(\frac{e\rho\tau}{m}\right)^m,$$

falls $m \geq 2\rho\tau$ ist.

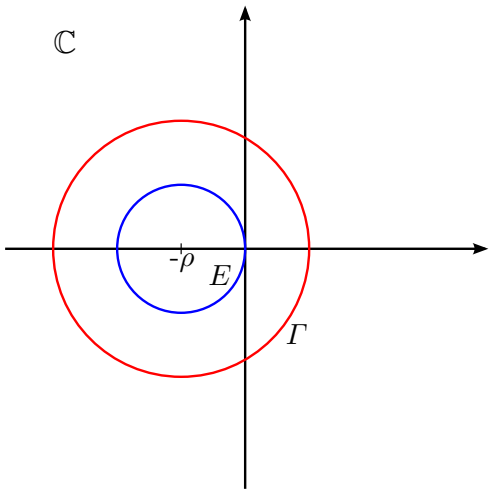


Abbildung 2.3.: Konstellation in Satz 2.2.7

Beweis:
Wählt man

$$\begin{aligned}\gamma : [0,1] &\longrightarrow \Gamma,\\ t &\longmapsto r\rho e^{2\pi i t} - \rho,\end{aligned}$$

so beschreibt Γ einen Kreis um $-\rho$ mit Radius $r\rho$, der E umschließt. Folglich ist Satz 2.2.5 mit Γ als Integrationsweg anwendbar.

Setze nun $r = \frac{m}{\rho\tau} \geq 2$. Wegen $\mathcal{F}(A) \subseteq E$ ist $d(\Gamma) = \text{mindist}(\mathcal{F}(A), \Gamma) \geq r\rho - \rho$ und damit $M = \frac{6}{d(\Gamma)} \leq \frac{6}{\rho(r-1)}$. Für den Fehler der Arnoldi-Approximation von $e^{hA}v$ bedeutet dies, dass

$$\begin{aligned}\varepsilon_m &\leq \frac{6}{2\pi\rho(r-1)} \int_\Gamma \left|e^{\tau\lambda}\right| \cdot |\phi(\lambda)|^{-m} \cdot |d\lambda| \\ &= \frac{6}{2\pi\rho(r-1)} \int_0^1 \left|e^{\tau(r\rho\exp(2\pi i t)-\rho)}\right| \cdot \left|\phi(r\rho e^{2\pi i t} - \rho)\right|^{-m} \cdot \left|r\rho\ 2\pi i e^{2\pi i t}\right| dt \\ &= \frac{6}{2\pi\rho(r-1)} \int_0^1 e^{\tau(r\rho-\rho)} \cdot r^{-m} \cdot 2\pi r\rho\ dt \\ &= 6\frac{r}{r-1} e^{\tau\rho(r-1)}\ r^{-m} \\ &= 6\underbrace{\frac{1}{1-\frac{\rho\tau}{m}}}_{\leq 2} e^m e^{-\tau\rho} \left(\frac{\rho\tau}{m}\right)^m\end{aligned}$$

ist.

■

3. Semilineare parabolische Gleichungen

Schwerpunkt dieses Kapitels sind Konvergenzaussagen für eine Klasse exponentieller Integratoren, die auf Evolutionsgleichungen der Form $u'(t) + Au(t) = g(t, u(t))$ angewandt werden. Im Gegensatz zum linearen Operator A, der im Allgemeinen unbeschränkt ist, wird für g eine Lipschitzbedingung angenommen.

Der typische Fall ist eine semilineare parabolische Differentialgleichung, die mittels Ortsdiskretisierung in ein hochdimensionales System gewöhnlicher Differentialgleichungen überführt werden kann. Auf Grund der Steifheit des Systems ist die klassische Ordnungstheorie allerdings nicht mehr hinreichend, um die Konvergenz der Verfahren zu beschreiben. Die steifen Ordnungsbedingungen werden daher ganz allgemein im abstrakten Rahmen einer Evolutionsgleichung im Banachraum ermittelt, so dass Verfahren, die diese Bedingungen erfüllen, nicht von einer Ordnungsreduktion betroffen sind.

Die betrachteten Verfahren besitzen zwei wesentliche Eigenschaften. Erstens gehen die exponentiellen Integratoren im nicht-steifen Fall $A \equiv 0$ in explizite Runge-Kutta-Verfahren über und zweitens liefern sie im Fall $g \equiv 0$ die exakte Lösung. Beides sind typische Charakteristika exponentieller Integratoren für semilineare parabolische Gleichungen.

Dieses Kapitel orientiert sich im Wesentlichen an der Arbeit [21] von M. Hochbruck und A. Ostermann. Für die theoretischen Grundlagen siehe [14] von D. Henry. In [21] wurde die Schrittweite der Einfachheit halber als konstant angenommen. Im letzten Teil dieses Kapitels wird am Beispiel des Nørsett–Euler-Verfahrens gezeigt, dass die Konvergenzanalyse auf den Fall einer variablen Schrittweite ausgeweitet werden kann. Zudem sei bemerkt, dass im Fall der zuvor betrachteten Verfahren zweiter Ordnung ähnlich vorgegangen werden kann.

3.1. Standardbeispiel

Das Standardbeispiel einer parabolischen Differentialgleichung ist die eindimensionale Wärmeleitungsgleichung

$$U_t = KU_{xx}$$

mit den homogenen Dirichletschen Randbedingungen $U(a, t) = 0$ und $U(b, t) = 0$ für gegebene $a, b \in \mathbb{R}$ sowie einer Anfangsbedingung $U(x, 0) = U_0(x)$. Dabei ist $K > 0$ eine Konstante. Gesucht ist nun eine Lösung $U(x, t)$ für $t > 0$ und $x \in [a, b]$.

Um die Differentialgleichung abstrakt im Banachraum schreiben zu können, definiert man einen linearen Operator A durch

$$Av = -Kv_{xx}$$

für $v \in \mathcal{D}(A) = H_0^1(a, b) \cap H^2(a, b)$.

Bemerkung:
Genauso wäre es möglich gewesen, $Av = +Kv_{xx}$ zu setzen. Die getroffene Wahl richtet sich an [21] und [14].

Da sich die Wärmeleitungsgleichung auf diese Weise als

$$u'(t) + Au(t) = 0$$

mit $u = [x \mapsto U(\cdot, x)]$ schreiben lässt, ist es naheliegend, dass man die Lösung in Analogie zur gewöhnlichen Differentialgleichung durch

$$u(t) = e^{-tA}u(0)$$

ausdrücken möchte.

Eine Definition von e^{-tA} über die Reihe der Exponentialfunktion scheitert zwar an der Unbeschränkheit von A, andererseits ist A in $L^2(a, b)$ dicht definiert, wegen

$$\langle Av, w\rangle = -K\int_a^b v_{xx}(x)w(x)dx = K\int_a^b v_x(x)w_x(x)dx = -K\int_a^b v(x)w_{xx}(x)dx = \langle v, Aw\rangle$$

selbstadjungiert und besitzt ein reines Punktspektrum. Daher kann $e^{-tA}v$ mittels Spektralzerlegung als

$$e^{-tA}v = e^{-tA}\sum_k c_k v_k = \sum_k c_k e^{-\lambda_k t}v_k$$

definiert werden, wobei λ_k die Eigenwerte von A und v_k die zugehörigen Eigenfunktionen sind. Die Reihe konvergiert für $t \geq 0$, da das Spektrum von A wegen

$$\langle Av, v\rangle = -K\int_a^b v_{xx}(x)v(x)dx = K\int_a^b v_x(x)v_x(x)dx \geq 0$$

nach unten beschränkt ist.

Auf diese Weise erhält man eine Halbgruppe beschränkter Operatoren $(e^{-tA})_{t\geq 0}$. Die obigen Voraussetzungen an A sind jedoch sehr restriktiv. Die Verallgemeinerung führt zum Begriff des sektoriellen Operators.

3.2. Grundlagen aus der Funktionalanalysis

Definition 3.2.1
Ein linearer Operator

$$B : \mathcal{D}(B) \subseteq X \longrightarrow X$$

heißt ***sektoriell****, wenn B abgeschlossen ist, $\mathcal{D}(B)$ dicht in X liegt und es $\phi \in (0, \frac{\pi}{2})$, $M \geq 1$ sowie $a \in \mathbb{R}$ gibt, so dass*

$$\Sigma_{a,\phi} = \{z \in \mathbb{C} : \ \phi \leq |\arg(z - a)| \leq \pi, \ z \neq a\} \subseteq \mathbb{C} \setminus \sigma(B)$$

ist und

$$\left\|(zI - B)^{-1}\right\| \leq \frac{M}{|z - a|} \quad \forall z \in \Sigma_{a,\phi}$$

gilt.

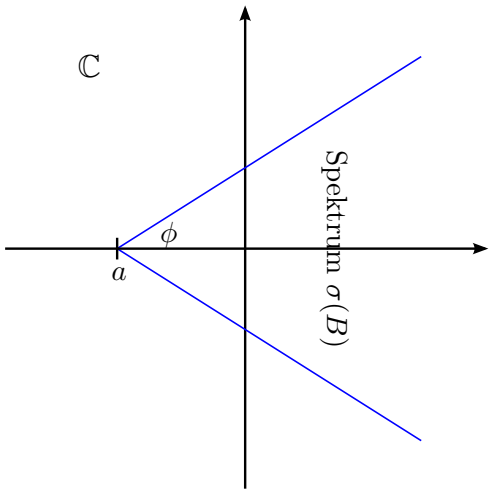


Abbildung 3.1.: Sektorieller Operator

Nach [14, Example 2] ist ein nach unten beschränkter, selbstadjungierter, dicht definierter Operator in einem Hilbertraum sektoriell. Folglich ist $A : \mathcal{D}(A) \longrightarrow L^2(a,b)$ aus obigem Beispiel ein sektorieller Operator.

Da ein sektorieller Operator B im Allgemeinen unbeschränkt ist, ist es wiederum nicht gerechtfertigt, e^{-tB} über die Potenzreihe zu definieren. Einen Ausweg liefert die folgende Definition.

Definition 3.2.2
Eine Familie $\{T(t)\}_{t\geq 0}$ *von beschränkten, linearen Operatoren*

$$T(t) : X \longrightarrow X,$$

welche die Eigenschaften

1. $T(0) = I$,
2. $T(s+t) = T(s)T(t) \quad s,t \geq 0$,
3. $\lim\limits_{t\searrow 0} T(t)x = x \quad \forall x \in X$,
4. $t \longmapsto T(t)x$ *ist für alle* $x \in X$ *analytisch auf* $(0,\infty)$

erfüllt, heißt ***analytische Halbgruppe****.*

Der Operator

$$L : \mathcal{D}(L) \longrightarrow X,$$
$$x \longmapsto \lim_{t\searrow 0} \frac{T(t)x - x}{t}$$

mit $\mathcal{D}(L) = \left\{ \lim\limits_{t\searrow 0} \dfrac{T(t)x - x}{t} \text{ existiert} \right\}$ *heißt* ***infinitesimaler Erzeuger*** *von* $\{T(t)\}_{t\geq 0}$.

Die Exponentialfunktion kann nun auf dieser Grundlage als analytische Halbgruppe definiert werden. Wie bereits im zweiten Kapitel wird auch an dieser Stelle der holomorphe Funktionalkalkül verwendet.

Lemma 3.2.3
Sei B ein sektorieller Operator. Dann ist $-B$ infinitesimaler Erzeuger der analytischen Halbgruppe $\{e^{-tB}\}_{t\geq 0}$ mit

$$e^{-tB} = \frac{1}{2\pi \mathbf{i}} \int_{\Gamma} (z+B)^{-1} e^{zt} dz,$$

wobei Γ ein geeigneter Integrationsweg sei.

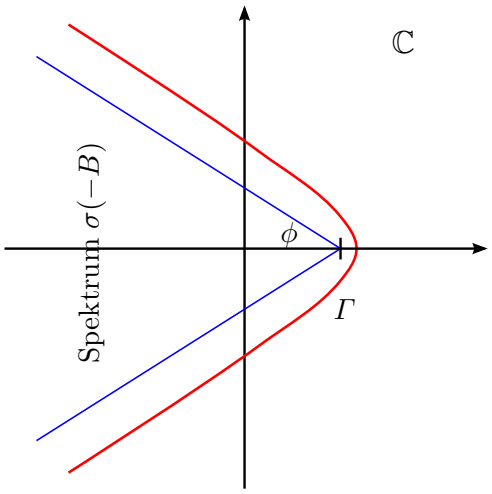


Abbildung 3.2.: Integrationsweg

Beweis:
Siehe Theorem 1.3.4 in [14].

■

Um die Annahmen an g und A formulieren zu können, wird im Folgenden auf einen in [14] eingeführten Raum zurückgegriffen. Das Lemma 3.2.4 fasst die wesentlichen Resultate zusammen.

Lemma 3.2.4
Seien B ein sektorieller Operator mit Sektor $\Sigma_{a,\phi}$, $\alpha \in [0,1)$ und $\omega > -a$. Dann sind Potenzen von $\widetilde{B} := B + \omega I$ wohldefiniert und die Menge $\mathcal{D}(\widetilde{B}^{\alpha}) \subseteq X$ mit der Norm $v \longmapsto \|\widetilde{B}v\|$ ist ein Banachraum. Ferner führt eine andere Wahl von ω zu einer äquivalenten Norm.

Beweis:
Siehe 1.4 und insbesondere Theorem 1.4.8 in [14].

■

Damit sind die benötigten Grundlagen bereitgestellt, so dass nun die wesentlichen Annahmen folgen können.

3.3. Annahmen der Konvergenztheorie und Verfahren

Sei $(X, \|\cdot\|)$ ein Banachraum. Da keine Verwechselungsgefahr besteht, wird die zugehörige Operatornorm ebenfalls mit $\|\cdot\|$ bezeichnet. Im Folgenden werden nicht-autonome Anfangswertaufgaben der Form

$$\begin{cases} u'(t) + Au(t) = g(t, u(t)), \\ u(0) = u_0 \end{cases} \tag{3.1}$$

behandelt, wobei

$$\begin{aligned} A : \mathcal{D}(A) &\longrightarrow X, \\ v &\longmapsto Av \end{aligned}$$

ein linearer Operator mit Definitionsbereich $\mathcal{D}(A) \subseteq X$ und

$$\begin{aligned} g : \mathbb{R} \times X &\longrightarrow X, \\ (t, v) &\longmapsto g(t, v) \end{aligned}$$

eine im Allgemeinen nicht-lineare Abbildung ist. Möglicherweise ist g nur auf einer geeigneten Teilmenge definiert. Für die analytischen Grundlagen siehe auch [14, Chapter 3].

Erste Annahme (A1):
$A : \mathcal{D}(A) \longrightarrow X$ sei ein sektorieller Operator mit Sektor $S_{a,\phi}$, $\alpha \in [0, 1)$ und $\omega > -a$. Außerdem sei der Banachraum $(Y, \|\cdot\|_Y)$ durch $Y := \mathcal{D}(\widetilde{A}^\alpha) \subseteq X$ und $\|v\|_Y := \|\widetilde{A}v\|$ gegeben.

Zweite Annahme (A2):
Die AWA (3.1) besitze eine hinreichend glatte, eindeutige Lösung $u : [0, T] \to Y$.
$\mathsf{g} : [0, T] \times Y \longrightarrow X$ erfülle in einer Umgebung der exakten Lösung u eine Lipschitzbedingung, d.h. es existieren $R \in \mathbb{R}, L = L(R, T) > 0$ mit

$$\|g(t, v) - g(t, w)\| \leq L \|v - w\|_Y$$

für alle $t \in [0, T]$ und $v, w \in Y$, die $\max(\|v - u(t)\|_Y, \|w - u(t)\|_Y) \leq R$ erfüllen. Falls die Lipschitzbedingung nicht global ist, wird die Schrittweite stets als klein genug angenommen.
Außerdem sei $\tilde{g} : [0, T] \longrightarrow X,\ \tilde{g}(t) := g(t, u(t))$ hinreichend oft Fréchet-differenzierbar.

Für gegebenes $T \in (0, \infty)$ soll die Lösung von (3.1)

$$\begin{aligned} u : [0, T] &\longrightarrow X, \\ t &\longmapsto u(t) \end{aligned}$$

auf dem endlichen Intervall $[0, T]$ mittels expliziter Einschrittverfahren der Art

$$\begin{aligned} u_{n+1} &= e^{-hA} u_n + h \sum_{i=1}^{s} b_i(-hA) \cdot g\big(t_n + c_i h, k_i(t_n, u_n, h)\big), \\ k_i(t, v, h) &= e^{-c_i hA} v + h \sum_{j=1}^{i-1} a_{ij}(-hA) \cdot g\big(t + c_j h, k_i(t, v, h)\big), \quad i = 1, ..., s \end{aligned} \tag{3.2}$$

approximiert werden.

Im Gegensatz zu den Runge-Kutta-Vefahren sind die Koeffizienten $a_{ij}(\cdot)$ und $b_i(\cdot)$ keine Konstanten, sondern Abbildungen. Bei der Definition kann wiederum auf die Halbgruppen-Theorie zurückgegriffen werden. Für die c_i gelte die gewohnte Bedingung $0 \leq c_i < 1$.

Analog zum Butcher-Tableau eines Runge-Kutta-Verfahren können die Koeffizienten als

$$\begin{array}{c|ccc} c_1 & a_{11}(\cdot) & \dots & a_{1s}(\cdot) \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1}(\cdot) & \dots & a_{ss}(\cdot) \\ \hline & b_1(\cdot) & \dots & b_s(\cdot) \end{array}$$

zusammengefasst werden, wobei einige Einträge wegfallen, da die Verfahren explizit sind. Im Fall $A = 0$ sind die Koeffizienten konstant und man erhält das gewohnte Butcher-Tableau des zu Grunde liegenden Runge-Kutta-Verfahrens. Dies rechtfertigt es, (3.2) als exponentielles Runge-Kutta-Verfahren zu bezeichnen.

3.4. Klassische Ordnung

Wenngleich der klassische Konvergenzbegriff im Fall partieller Differentialgleichungen ungeeignet ist, erscheint es trotzdem zu Vergleichszwecken nützlich, die klassische Ordnung der Verfahren aus (3.2) zu analysieren. Außerdem ermöglicht dies geeignete Kandidaten für die $b_i(\cdot)$ und $a_{ij}(\cdot)$ zu finden, die dann im abstrakten Rahmen über eine Halbgruppe definiert werden können. Hierzu sei g in einer Umgebung der exakten Lösung p-mal stetig differenzierbar, wobei p die betrachtete Konsistenzordnung ist.

Die Verfahrensfunktionen hat die Form

$$V(t, v, h) = -\varphi(-hA)Av + \sum_{i=1}^{s} b_i(-hA) \cdot g\big(t + c_i h, k_i(t, v, h)\big),$$

wobei φ die aus dem ersten Kapitel bekannte Funktion

$$\begin{aligned} \varphi : \mathbb{C} &\longrightarrow \mathbb{C}, \\ \varphi(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+1)!} &= \begin{cases} \frac{e^z - 1}{z} & z \neq 0 \\ 1 & z = 0 \end{cases} \end{aligned}$$

ist. Dies folgt aus

$$\begin{aligned} u_{n+1} &= e^{-hA} u_n + h \sum_{i=1}^{s} b_i(-hA) \cdot g\big(t_n + c_i h, k_i(t_n, u_n, h)\big) \\ &= u_n + \Big(e^{-hA} - I\Big) u_n + h \sum_{i=1}^{s} b_i(-hA) \cdot g\big(t_n + c_i h, k_i(t_n, u_n, h)\big) \\ &= u_n - h\varphi(-hA)Au_n + h \sum_{i=1}^{s} b_i(-hA) \cdot g\big(t_n + c_i h, k_i(t_n, u_n, h)\big) \end{aligned}$$

und der Definition der Verfahrensfunktion durch $u_{n+1} = u_n + hV(t_n, u_n, h)$.

Die Ordnungsbedingungen für $p = 1$ und $p = 2$ lassen sich nun wie folgt berechnen.

Ordnung 1:
Wegen $k_i(t, u(t), 0) = u(t)$ und $\varphi(0) = 1$ ist

$$V(t, u(t), 0) = -Au(t) + \sum_{i=1}^{s} b_i(0) \cdot g(t, u(t)).$$

Die Konsistenzbedingung für Ordnung 1 lautet daher

$$\sum_{i=1}^{s} b_i(0) = 1.$$

Ordnung 2:
Der Einfachheit halber seien die Koeffizienten so gewählt, dass die Bedingungen

$$c_i = \sum_{j=1}^{i-1} a_{ij}(0), \quad i = 1, ..., s$$

erfüllt sind. Dies ist eine typische Forderung an Runge-Kutta-Verfahren, um die Transformation auf ein autonomes System zu ermöglichen.

Diesmal wird die partielle Ableitung der Verfahrensfunktion nach h benötigt, welche sich als

$$\begin{aligned}\frac{\partial V}{\partial h}(t, v, h) = \varphi'(-hA)A^2 v &- \sum_{i=1}^{s} b_i'(-hA)A \cdot g(t + c_i h, k_i(t, v, h)) \\ &+ \sum_{i=1}^{s} b_i(-hA) \cdot \frac{\partial g}{\partial t}(t + c_i h, k_i(t, v, h))c_i \\ &+ \sum_{i=1}^{s} b_i(-hA) + \frac{\partial g}{\partial v}(t + c_i h, k_i(t, v, h))\frac{\partial k_i}{\partial h}(t, v, h)\end{aligned}$$

schreiben lässt. Ausgewertet an der Stelle $(t, u(t), 0)$ erhält man

$$\begin{aligned}\frac{\partial V}{\partial h}(t, u(t), 0) = \frac{1}{2}A^2 u(t) &- A\sum_{i=1}^{s} b_i'(0) \cdot g(t, u(t)) \\ &+ \sum_{i=1}^{s} b_i(0) \cdot \left(\frac{\partial g}{\partial t}(t, u(t))c_i + \frac{\partial g}{\partial v}(t, u(t))\frac{\partial k_i}{\partial h}(t, u(t), 0)\right)\end{aligned}$$

und mit

$$\frac{\partial k_i}{\partial h}(t, u(t), 0) = -c_i Au(t) + \sum_{j=1}^{i-1} a_{ij}(0) \cdot g(t, u(t))$$

folgt

$$\begin{aligned}\frac{\partial V}{\partial h}(t, u(t), 0) &= \frac{1}{2}A\Big(Au(t) - 2\sum_{i=1}^{s} b_i'(0) \cdot g(t, u(t))\Big) + \sum_{i=1}^{s} b_i(0)c_i \cdot \frac{\partial g}{\partial t}(t, u(t)) \\ &\quad + \frac{\partial g}{\partial v}(t, u(t))\Big(-\sum_{i=1}^{s} b_i(0)c_i Au(t) + \sum_{i=1}^{s} b_i(0)\sum_{j=1}^{i-1} a_{ij}(0) \cdot g(t, u(t))\Big) \\ &= \frac{1}{2}A\Big(Au(t) - 2\sum_{i=1}^{s} b_i'(0) \cdot g(t, u(t))\Big) + \sum_{i=1}^{s} b_i(0)\sum_{j=1}^{i-1} a_{ij}(0)\frac{\partial g}{\partial t}(t, u(t)) \\ &\quad + \sum_{i=1}^{s} b_i(0)\sum_{j=1}^{i-1} a_{ij}(0)\frac{\partial g}{\partial v}(t, u(t))\Big(g(t, u(t)) - Au(t)\Big).\end{aligned}$$

Andererseits gilt wegen $f(t,v) = -Av + g(t,v)$

$$\begin{aligned}\frac{d}{dt}\Big[f(t,u(t))\Big] &= -Au'(t) + \frac{\partial g}{\partial t}(t,u(t)) + \frac{\partial g}{\partial v}(t,u(t))u'(t)\\ &= A\Big(Au(t) - g(t,u(t))\Big) + \frac{\partial g}{\partial t}(t,u(t)) + \frac{\partial g}{\partial v}(t,u(t))\Big(g(t,u(t)) - Au(t)\Big).\end{aligned}$$

Daher lauten die zusätzlichen Ordnungsbedingungen für Ordnung 2

$$\sum_{i=1}^{s} b_i'(0) = \frac{1}{2} \text{ und } \sum_{i=1}^{s} b_i(0) \sum_{j=1}^{i-1} a_{ij}(0) = \frac{1}{2}.$$

3.5. Steife Ordnung

Nun wird die Evolutionsgleichung (3.1) wieder abstrakt im Banachraum $(X, \|\cdot\|)$ betrachtet. Die klassische Ordnungstheorie basiert auf der Taylorentwicklung von exakter und numerischer Lösung. Um die Differenz abzuschätzen, wird dabei vorausgesetzt, dass $\|hA\| \to 0$ für $h \to 0$ gilt. Bei der Lösung partieller Differentialgleichungen gilt dies allerdings nur für eine fest gewählte Ortsdiskretisierung.

3.5.1. Konsistenzanalyse

Bei der Analyse der klassischen Ordnung trat die Funktion φ auf, wobei es allerdings genügte, den Wert der Ableitungen im Nullpunkt zu kennen. Um φ und ähnliche Funktionen auch auf unbeschränkte Operatoren anwenden und damit in der folgenden Konvergenzanalyse nutzen zu können, ist nun die Theorie der analytischen Halbgruppen hilfreich.

Definiere für $j \in \mathbb{N}_0$ und $t > 0$

$$\varphi_j(-tB) = \begin{cases} e^{-tB}, & j = 0 \\ \dfrac{1}{t^j}\displaystyle\int_0^t e^{-(t-\tau)B}\frac{\tau^{j-1}}{(j-1)!}d\tau, & j > 0. \end{cases} \tag{3.3}$$

Für $j = 1$ und beschränkte Operatoren gilt dann $\varphi_1 = \varphi$ (siehe Anhang).
Definiere nun noch

$$\psi_j(z) := \varphi_j(z)c_i^j - \sum_{k=1}^{s} b_k(z)\frac{c_k^{j-1}}{(j-1)!} \tag{3.4}$$

für $j \in \mathbb{N}_1$ und für $j, i \in \mathbb{N}_1$ setze

$$\psi_{ji}(z) := \varphi_j(c_i z)c_i^j - \sum_{k=1}^{i-1} a_{ik}(z)\frac{c_k^{j-1}}{(j-1)!}. \tag{3.5}$$

Während die φ_j eigenständige Funtionen sind, hängen die ψ_j und ψ_{ji} von der Wahl des Verfahrens ab.

Mit Hilfe der ψ_j und ψ_{ji} lassen sich die klassischen Ordnungsbedingungen vereinfacht ausdrücken, denn es gelten die Äquivalenzen

$$\begin{aligned}
\psi_1(0) = 0 \Leftrightarrow \varphi_1(0) &= \sum_{k=1}^{s} b_k(0), \\
\psi_1'(0) = 0 \Leftrightarrow \varphi_1'(0) &= \sum_{k=1}^{s} b_k'(0), \\
\psi_2(0) = 0 \Leftrightarrow \varphi_2(0) &= \sum_{k=1}^{s} b_k(0)c_k, \\
\psi_{1,i}(0) = 0 \Leftrightarrow \varphi_1(0)c_i &= \sum_{k=1}^{i-1} a_{ik}(0), \quad i = 1, ..., s.
\end{aligned}$$

Wegen $\varphi_1(0) = 1$ und $\varphi_1'(0) = \varphi_2(0) = \frac{1}{2}$ erhält man auf der rechten und damit auch auf der linken Seite die klassischen Bedingungen für Ordnung 2.

Mit Blick auf die klassischen Ordnungsbedingungen liegt es nahe, die Koeffizienten der Verfahren als Linearkombinationen der Phi-Funktionen aus (3.3) zu wählen. Eine wichtige Abschätzung liefert daher Lemma 3.5.1, das aus der Annahme (A1) folgt.

Lemma 3.5.1
(A1) impliziert für $j \in \mathbb{N}_0$ die Abschätzung

$$\|t^\gamma \widetilde{A}^\gamma \varphi_j(-tA)\| \leq C, \quad \gamma \geq 0,$$

welche gleichmäßig für $0 \leq t \leq T$ gilt.

Beweis:
Für den Fall $j = 0$, also e^{-tA} siehe [22, Lemma 1]. Der Fall $j > 0$ lässt sich nun darauf zurückführen. Ist $\tau \in [0, t]$, so ist $t - \tau \in [0, T]$ und es gilt $\|(t-\tau)^\gamma \widetilde{A}^\gamma e^{-(t-\tau)}\| \leq C$. Daraus ergibt sich

$$\begin{aligned}
\|t^\gamma \widetilde{A}^\gamma \varphi_j(-tA)\| &\leq \frac{t^\gamma}{t^j} \int_0^t \left\| \widetilde{A}^\gamma e^{-(t-\tau)A} \right\| \frac{\tau^{j-1}}{(j-1)!} d\tau \leq C \cdot \frac{t^\gamma}{t^j} \int_0^t (t-\tau)^{-\gamma} \frac{\tau^{j-1}}{(j-1)!} d\tau \\
&\leq C \cdot \frac{t^\gamma}{t(j-1)!} \int_0^t (t-\tau)^{-\gamma} d\tau = C \cdot \frac{t^\gamma}{t(j-1)!} t^{1-\gamma} \leq C.
\end{aligned}$$

■

Auf dieser Grundlage ist es nun möglich, den Konvergenzfehler $e_h(t_n) = u_n - u(t_n)$ zu analysieren. Definiere jedoch zunächst für $n \in \mathbb{N}_1$

$$\delta_h(t_n) := u(t_n) - \left(e^{-hA} u(t_{n-1}) + h \sum_{i=1}^{s} b_i(-hA) \tilde{g}(t_{n-1} + c_i h) \right). \tag{3.6}$$

Damit ist $\delta_h(t_n)$ der lokale Fehler des Verfahrens (3.2).

Bemerkung:
Im ersten Kapitel wurde der Konsistenzfehler als

$$\tau_h(t_n) = \frac{1}{h}\Big(u(t_{n+1}) - u(t_n)\Big) - V\big(u(t_n), h\big)$$

eingeführt.

Der Vorteil besteht darin, dass Konsistenz- und Konvergenzfehler dieselbe Ordnung besitzen. Im Rahmen partieller Differentialgleichungen kann die dazu notwendige Stabilität allerdings nicht ohne weiteres vorausgesetzt werden. Bei den notwendigen Rechnungen zur Konvergenzanalyse erweist sich die mit h durchmultiplizierte Form (3.6) als praktischer. Man beachte dabei, dass

$$\delta_h(t_{n+1}) = u(t_{n+1}) - u(t_n) - hV\big(u(t_n), h\big) = h\tau_h(t_n)$$

gilt.

Die Verfahren, die im Folgenden betrachtet werden, erfüllen die Bedingung $\psi_j(-hA) = 0$ für $1 \le j \le r$ mit $r \in \mathbb{N}_1$. Dies ist eine Verallgemeinerung der klassischen Ordnungsbedingungen, die $\psi_j(0) = 0$ voraussetzen. Sei daher

$$\delta_h^{[r]}(t_n) := \delta_h(t_n) - \sum_{j=0}^{r-1} h^{j+1}\psi_{j+1}(-hA)\tilde{g}^{(j)}(t_{n-1}) \tag{3.7}$$

mit ψ_j aus (3.4). Vorrangiges Ziel ist es nun, eine Schranke für $\delta_h^{[r]}(t_{n+1})$ zu finden. Lemma 3.5.2 liefert genau dies.

Lemma 3.5.2
Der zuvor definierte Term $\delta_h^{[r]}(t_{n+1})$ besitzt die Darstellung

$$\begin{aligned}\delta_h^{[r]}(t_{n+1}) = &\int_0^h e^{-(h-\tau)A}\int_0^\tau \frac{(\tau-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma d\tau \\ &- h\sum_{i=1}^{s} b_i(-hA)\int_0^{c_ih}\frac{(c_ih-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma.\end{aligned}$$

Beweis:
Die exakte Lösung der Differentialgleichung (3.1) im Punkt $t_{n+1} = t_n + h$ lässt sich mit Hilfe der Formel zur Variation der Konstanten berechnen. Man erhält

$$u(t_n+h) = e^{-hA}u(t_n) + \int_0^h e^{-(h-\tau)A}\tilde{g}(t_n+\tau)d\tau, \tag{3.8}$$

so dass

$$\begin{aligned}\delta_h(t_{n+1}) &\overset{(3.6)}{=} u(t_n+h) - \left(e^{-hA}u(t_n) + h\sum_{i=1}^{s} b_i(-hA)\tilde{g}(t_n+c_ih)\right) \\ &\overset{(3.8)}{=} \int_0^h e^{-(h-\tau)A}\tilde{g}(t_n+\tau)d\tau - h\sum_{i=1}^{s} b_i(-hA)\tilde{g}(t_n+c_ih)\end{aligned}$$

folgt. Mit der Taylor-Formel

$$\tilde{g}(t_n+\tau) = \sum_{j=0}^{r-1}\frac{\tau^j}{j!}\tilde{g}^{(j)}(t_n) + \int_0^\tau \frac{(\tau-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma$$

ergibt dies

$$\begin{aligned}\delta_h(t_{n+1}) \;=\; &\int_0^h e^{-(h-\tau)A}\left(\sum_{j=0}^{r-1}\frac{\tau^j}{j!}\tilde{g}^{(j)}(t_n) + \int_0^\tau\frac{(\tau-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma\right)d\tau \\ &- h\sum_{i=1}^{s} b_i(-hA)\left(\sum_{j=0}^{r-1}\frac{(c_ih)^j}{j!}\tilde{g}^{(j)}(t_n) + \int_0^{c_ih}\frac{(c_ih-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma\right).\end{aligned}$$

Da nach (3.3) mit $t = h$

$$h^{j+1}\varphi_{j+1}(-hA) = \int_0^h e^{-(h-\tau)A}\frac{\tau^j}{j!}d\tau$$

gilt und damit

$$\psi_{j+1}(z) = \int_0^h e^{-(h-\tau)A}\frac{\tau^j}{j!}d\tau - \sum_{i=1}^{s} b_i(-hA)\frac{c_i^j}{j!}$$

ist, folgt

$$\begin{aligned}\delta_h^{[r]}(t_{n+1}) &\overset{(3.7)}{=} \delta_h(t_{n+1}) - \sum_{j=0}^{r-1} h^{j+1}\psi_{j+1}(-hA)\tilde{g}^{(j)}(t_n)\\ &= \int_0^h e^{-(h-\tau)A}\int_0^\tau \frac{(\tau-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma d\tau\\ &\quad - h\sum_{i=1}^{s} b_i(-hA)\int_0^{c_ih}\frac{(c_ih-\sigma)^{r-1}}{(r-1)!}\tilde{g}^{(r)}(t_n+\sigma)d\sigma,\end{aligned}$$

was zu zeigen war.

■

Die Darstellung des Konsistenzfehlers in Lemma 3.5.2 mittels der Formel zur Variation der Konstanten ermöglicht es, den folgenden Konsistenzsatz zu beweisen. Hierbei wird wiederum auf den Raum $(Y, \|\cdot\|_Y)$ mit $Y = \mathcal{D}(\widetilde{A}^\alpha) \subseteq X$ und die zugehörige Norm $\|v\|_Y := \|\widetilde{A}v\|$ zurückgegriffen.

Satz 3.5.3
Seien $0 < \nu \leq 1$ und $\widetilde{A}^{\nu-1}\tilde{g}^{(r)} \in L^\infty(0,T;Y)$. Dann gelten

$$h^{1-\nu}\left\|\delta_h^{[r]}(t_{n+1})\right\|_Y + \left\|\widetilde{A}^{\nu-1}\delta_h^{[r]}(t_{n+1})\right\|_Y \leq Ch^{r+1}\sup_{t_n\leq t\leq t_{n+1}}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y$$

und

$$\left\|\sum_{j=0}^{n-1} e^{-jhA}\delta_h^{[r]}(t_{n-j})\right\|_Y \leq Ch^r\sup_{0\leq t\leq t_n}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y$$

gleichmäßig für $0 \leq t_n \leq T$. Insbesondere hängt C für festes T nicht von n oder h ab.

Beweis:
Aus Lemma 3.5.1 folgt

$$\begin{aligned}\left\|(h-\tau)^{1-\nu}\widetilde{A}^{1-\nu}e^{-(h-\tau)A}\right\| + \left\|e^{-(h-\tau)A}\right\| \leq C,\\ \left\|(h\widetilde{A})^{1-\nu}b_i(-hA)\right\| + \left\|b_i(-hA)\right\| \leq C,\end{aligned} \tag{3.9}$$

wobei daran erinnert sei, dass die Koeffizienten der Verfahren grundsätzlich als Linearkombinationen der Phi-Funktionen (3.3) gewählt werden.

Mit Lemma 3.5.2 erhält man nun

$$
\begin{aligned}
&h^{1-\nu}\left\|\delta_h^{[r]}(t_{n+1})\right\|_Y + \left\|\widetilde{A}^{\nu-1}\delta_h^{[r]}(t_{n+1})\right\|_Y \\
\leq\ & \int_0^h \left(h^{1-\nu}\left\|\widetilde{A}^{1-\nu}e^{-(h-\tau)A}\right\| + \left\|e^{-(h-\tau)A}\right\|\right)\int_0^\tau \frac{(\tau-\sigma)^{r-1}}{(r-1)!}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t_n+\sigma)\right\|_Y d\sigma d\tau \\
&+ h\sum_{i=1}^s \left(\left\|(h\widetilde{A})^{1-\nu}b_i(-hA)\right\| + \left\|b_i(-hA)\right\|\right)\int_0^{c_ih}\frac{(c_ih-\sigma)^{r-1}}{(r-1)!}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t_n+\sigma)\right\|_Y d\sigma \\
\overset{(3.9)}{\leq}& C\int_0^h\left(h^{1-\nu}(h-\tau)^{\nu-1}+1\right)\int_0^\tau\frac{(\tau-\sigma)^{r-1}}{(r-1)!}d\sigma d\tau \sup_{t_n\leq t\leq t_{n+1}}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y \\
&+ Ch\sum_{i=1}^s\int_0^{c_ih}\frac{(c_ih-\sigma)^{r-1}}{(r-1)!}d\sigma\sup_{t_n\leq t\leq t_{n+1}}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y \\
\leq\ & \frac{C}{(r-1)!}\left(\left(h^{1-\nu}\underbrace{\left[(\tau-h)^\nu\right]_{\tau=0}^{\tau=h}}_{=h^\nu}+h\right)h^r + h^{r+1}\sum_{i=1}^s c_i^r\right)\sup_{t_n\leq t\leq t_{n+1}}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y.
\end{aligned}
$$

Damit ist bereits

$$
h^{1-\nu}\left\|\delta_h^{[r]}(t_{n+1})\right\|_Y + \left\|\widetilde{A}^{\nu-1}\delta_h^{[r]}(t_{n+1})\right\|_Y \leq Ch^{r+1}\sup_{t_n\leq t\leq t_{n+1}}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y \tag{3.10}
$$

gezeigt. Zusammen mit der Dreiecksungleichung und erneut Lemma 3.5.1 folgt daraus wegen

$$
\begin{aligned}
\left\|\sum_{j=0}^{n-1}e^{-jhA}\delta_h^{[r]}(t_{n-j})\right\|_Y &\leq \left\|\delta_h^{[r]}(t_n)\right\|_Y + \sum_{j=1}^{n-1}\left\|\widetilde{A}^{1-\nu}e^{-jhA}\widetilde{A}^{\nu-1}\delta_h^{[r]}(t_{n-j})\right\|_Y \\
&\leq h^{\nu-1}h^{1-\nu}\left\|\delta_h^{[r]}(t_n)\right\|_Y + C\sum_{j=1}^{n-1}(jh)^{\nu-1}\left\|\widetilde{A}^{\nu-1}\delta_h^{[r]}(t_{n-j})\right\|_Y \\
&\overset{(3.10)}{\leq} Ch^{r+1}\sum_{j=1}^{n-1}(jh)^{\nu-1}\sup_{t_{n-j-1}\leq t\leq t_{n-j}}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y \\
&\leq Ch^r\sup_{0\leq t\leq t_n}\left\|\widetilde{A}^{\nu-1}\tilde{g}^{(r)}(t)\right\|_Y
\end{aligned}
$$

auch die zweite Ungleichung. Im letzten Schritt wurde

$$
h\sum_{j=1}^{n-1}(jh)^{\nu-1} \leq \int_0^T t^{\nu-1}dt = \frac{T^\nu}{\nu}
$$

verwendet, wobei die linke Seite der Gleichung eine Untersumme des Integrals ist.

■

3.5.2. Konvergenzanalyse

Das Ziel muss es nun sein, die gewonnenen Erkenntnisse zu nutzen, um Aussagen über die Konvergenz zu gewinnen. Lemma 3.5.4 ist dabei der erste Schritt.

Lemma 3.5.4
Für den Konvergenzfehler $e_h(t_{n+1})$ gilt

$$e_h(t_{n+1}) = e^{hA}e_h(t_n) + h\sum_{i=1}^{s} b_i(-hA)\Big(g\big(t_n + c_ih, k_i(t_n, u_n, h)\big) - \tilde{g}(t_n + c_ih)\Big) - \sum_{j=0}^{r-1} h^{j+1}\psi_{j+1}(-hA)\tilde{g}^{(j)}(t_n) - \delta_h^{[r]}(t_{n+1}).$$

Beweis:
Aus den Definitionen der obigen Terme folgt

$$\begin{aligned} e_h(t_{n+1}) &= u_{n+1} - u(t_{n+1}) \\ &\overset{(3.2)}{=} e^{-hA}u_n + h\sum_{i=1}^{s} b_i(-hA)\cdot g\big(t_n + c_ih, k_i(t_n, u_n, h)\big) - u(t_{n+1}) \\ &\overset{(3.6)}{=} e^{-hA}e_h(t_n) + h\sum_{i=1}^{s} b_i(-hA)\Big(g\big(t_n + c_ih, k_i(t_n, u_n, h)\big) - \tilde{g}(t_n + c_ih)\Big) \\ &\quad - \delta_h(t_{n+1}) \\ &\overset{(3.7)}{=} e^{-hA}e_h(t_n) + h\sum_{i=1}^{s} b_i(-hA)\Big(g\big(t_n + c_ih, k_i(t_n, u_n, h)\big) - \tilde{g}(t_n + c_ih)\Big) \\ &\quad - \sum_{j=0}^{r-1} h^{j+1}\psi_{j+1}(-hA)\tilde{g}^{(j)}(t_n) - \delta_h^{[r]}(t_{n+1}). \end{aligned}$$

■

Auf die Formel in Lemma 3.5.4 können die Lipschitz-Bedingung von g und nach Auflösen der Rekursion das folgende diskrete Gronwall-Lemma aus [22] angewandt werden.

Lemma 3.5.5
Seien $h>0$, $M \in \mathbb{N}_1$ und $Mh \leq T$. Außerdem seien $\varepsilon_n > 0$ für $n = 1, ..., M$ gegeben, so dass für gewisse $0 \leq \rho < 1$ und $a, b \geq 0$

$$\varepsilon_n \leq ah\sum_{\nu=1}^{n-1} t_{n-\nu}^{-\rho}\varepsilon_\nu + b$$

gilt. Dann ist

$$\varepsilon_n \leq Cb,$$

wobei C nur von ρ, a und T abhängt.

Beweis:
Siehe Lemma 4 in [22].

■

3.5.3. Konvergenz des Nørsett–Euler-Verfahrens

Betrachte nun den Fall $s = 1$. Für ein explizites Verfahren müssen $c_1 = 0$ und $a_{11}(-tA) \equiv 0$ gelten. Fordert man zudem in Rückblick auf die Fehlerdarstellung in Lemma 3.5.4, dass $\psi_1(-tA) \equiv 0$ erfüllt ist, so folgt $b_1(-tA) = \varphi_1(-tA) =: \varphi(-tA)$ und alle Koeffizienten sind bestimmt.

Das Nørsett–Euler-Verfahren für (3.1) hat die Form

$$u_{n+1} = e^{-hA}u_n + h\varphi(-hA)g(t_n, u_n). \tag{3.11}$$

Im linearen Fall stimmt dieses Verfahren mit dem exponentiell angepassten Eulerverfahren aus dem ersten Kapitel überein. Der Unterschied besteht darin, dass nicht die Ableitung der gesamten rechten Seite von (3.1), sondern nur des linearen Teils verwendet wird.

Satz 3.5.6 zeigt nun, dass das Nørsett–Euler-Verfahren angewandt auf (3.1) konvergent der Ordnung 1 ist.

Satz 3.5.6
Die numerische Lösung u_n von (3.1) werde mit dem Nørsett–Euler-Verfahren (3.11) berechnet. Außerdem seien $\tilde{g} : [0, T] \longrightarrow X$ (Fréchet-)differenzierbar und ein $\beta \in (0, 1]$ gegeben, so dass $\widetilde{A}^{\beta-1}\tilde{g}' \in L^\infty\big((0, T); V\big)$ gilt, wobei wiederum $\widetilde{A} = A + \omega I$ ist.
Dann gilt die Fehlerabschätzung

$$\|u_n - u(t_n)\|_Y \leq C \cdot h \sup_{0\leq t\leq t_n} \left\|\widetilde{A}^{\beta-1}\tilde{g}'(t)\right\|_Y$$

gleichmäßig für $0 \leq nh \leq T$.

Beweis:
Setzt man in Lemma 3.5.4 die Parameter des Nørsett-Euler-Verfahrens ein, so erhält man

$$e_h(t_n + 1) = e^{-hA}e_h(t_n) + h\varphi(-hA)\Big(g(t_n, u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}), \tag{3.12}$$

wobei wiederum $e_h(t_n) = u_n - u(t_n)$ der Konvergenzfehler ist.

Mittels Induktion nach n kann man nun zeigen, dass

$$e_h(t_n) = h\sum_{j=0}^{n-1} e^{-(n-j-1)hA}\varphi(-hA)\Big(g(t_j, u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j}) \tag{3.13}$$

die explizite Darstellung der $e_h(t_n)$ für $n \in \mathbb{N}_1$ ist.

Induktionsanfang $n = 1$:
Wegen $e_h(t_0) = u_0 - u(t_0) = 0$ folgt

$$\begin{aligned} e_h(t_1) = e_h(t_{0+1}) &\overset{(3.12)}{=} 0 + h\varphi(-hA)\Big(g(t_0, u_0) - f(t_0)\Big) - \delta_h(t_{0+1}) \\ &= h\sum_{j=0}^{1-1} e^{-(1-j-1)hA}\varphi(-hA)\Big(g(t_j, u_j) - f(t_j)\Big) - \sum_{j=0}^{1-1} e^{-jhA}\delta_h(t_{1-j}). \end{aligned}$$

Induktionsschritt $n \to (n+1)$:
Nimmt man nach Induktionsvoraussetzung an, dass (3.13) für n gilt, so erhält man

$$\begin{aligned}
e_h(t_{n+1}) &\overset{(3.12)}{=} e^{-hA}e_h(t_n) + h\varphi(-hA)\Big(g(t_n,u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}) \\
&\overset{IV}{=} e^{-hA}\left(h\sum_{j=0}^{n-1} e^{-(n-j-1)hA}\varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j})\right) \\
&\quad + h\varphi(-hA)\Big(g(t_n,u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}) \\
&= e^{-hA}h\sum_{j=0}^{n-1} e^{-(n-j-1)hA}\varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big) \\
&\quad + h\varphi(-hA)\Big(g(t_n,u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}) - e^{-hA}\sum_{j=1}^{n} e^{-(j-1)hA}\delta_h(t_{n-(j-1)}) \\
&= h\sum_{j=0}^{n} e^{(n+1-j-1)hA}\varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n} e^{-jhA}\delta_h(t_{n+1-j}),
\end{aligned}$$

was die Induktion abschließt. Bevor das Gronwall-Lemma angewandt werden kann, sind noch ein paar Abschätzungen nötig. Dabei wird zum einen die Lipschitz-Bedingung an g

$$\|g(t_n,u_n) - \tilde{g}(t_n)\| = \|g(t_n,u_n) - g(t_n,u(t_n))\| \le L\|u_n - u(t_n)\|_Y = L\|e_h(t_n)\|_Y \tag{3.14}$$

und zum anderen die Ungleichung

$$k^{-x} \le \left(\frac{k+1}{2}\right)^{-x}, \tag{3.15}$$

die für $k \in \mathbb{N}_0$ und $x \ge 0$ gilt, benutzt.

Zusammen mit der Definition von $\|\cdot\|_Y$, Lemma 3.5.1 und Satz 3.5.3 erhält man dann

$$\begin{aligned}
\|e_h(t_n)\|_Y &\overset{(3.13)}{=} \left\|h\sum_{j=0}^{n-1} e^{-(n-j-1)hA}\varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j})\right\|_Y \\
&\le h\sum_{j=0}^{n-1}\left\|\widetilde{A}^{\alpha}e^{-(n-j-1)hA}\varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big)\right\| + \left\|\sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j})\right\|_Y \\
&\overset{(3.14)}{\le} h\sum_{j=0}^{n-1}\left\|\widetilde{A}^{\alpha}e^{-(n-j-1)hA}\varphi(-hA)\right\| \cdot L\|e_h(t_j)\|_Y + \left\|\sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j})\right\|_Y \\
&\overset{(3.15)}{\le} C\left(h\sum_{j=0}^{n-2} t_{n-j}^{-\alpha}\|e_h(t_j)\|_Y + h^{1-\alpha}\|e_h(t_{n-1})\|_Y\right) + \left\|\sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j})\right\|_Y \\
&\le C\left(h\sum_{j=1}^{n-1} t_{n-j}^{-\alpha}\|e_h(t_j)\|_Y + h\sup_{0\le t\le t_n}\|\widetilde{A}^{\beta-1}\tilde{g}'(t)\|_Y\right).
\end{aligned}$$

Wendet man schließlich das Gronwall-Lemma 3.5.5 mit $b = Ch\sup_{0\le t\le t_n}\|\widetilde{A}^{\beta-1}\tilde{g}'(t)\|_Y$, $\rho = \alpha$ und $\varepsilon_\nu = \|e_h(t_\nu)\|_Y$ an, so folgt

$$\|e_h(t_n)\|_Y \le Ch\sup_{0\le t\le t_n}\|\widetilde{A}^{\beta-1}\tilde{g}'(t)\|_Y.$$

■

3.5.4. Konvergenz der Ordnung 2

Der Fall des Nørsett–Euler-Verfahrens ist verhältnismäßig einfach zu analysieren, da das Verfahren einstufig ist. Bei mehrstufigen Verfahren muss hingegen zusätzlich der Fehler an Zwischenstellen untersucht werden.

Für $i = 1, ..., s$ und $n \in \mathbb{N}_1$ ist der globale Fehler an den Zwischenstellen durch

$$E_i(t_n, u_n, h) = k_i(t_n, u_n, h) - u(t_n + c_i h) \tag{3.16}$$

und der lokale Fehler durch

$$\Delta_i(t_n, h) := u(t_n + c_i h) - \left(e^{-c_i hA} u(t_n) + h \sum_{k=1}^{i-1} a_{ik}(-hA)\tilde{g}(t_n + c_k h) \right) \tag{3.17}$$

gegeben. Außerdem sei noch

$$\Delta_i^{[r]}(t_n, h) := \Delta_i(t_n, h) - \sum_{j=0}^{r-1} h^{j+1} \psi_{j+1,i}(-hA)\tilde{g}^{(j)}(t_n)$$

mit $\psi_{j,i}$ aus (3.5). Dies ist dieselbe Konstruktion wie bei den $\delta_h(t_n)$. Die Beweise verlaufen folglich analog. Von wesentlicher Bedeutung ist es wiederum, eine Formel zu finden, mit der der globale auf den lokalen Fehler zurückgeführt werden kann.

Lemma 3.5.7
Sei $\tilde{g}$ r-mal (Fréchet-)differenzierbar. Für $i = 1, ..., s$ ist dann

$$\begin{aligned} E_i(t_n, u_n, h) = e^{-c_i hA} e_h(t_n) + h \sum_{j=1}^{i-1} a_{ij}(-hA) \Big(g\big(t_n + c_j h, k_j(t_n, u_n, h)\big) - \tilde{g}(t_n + c_j h) \Big) \\ - \sum_{j=0}^{r-1} h^{j+1} \psi_{j+1,i}(-hA)\tilde{g}^{(r)}(t_n) - \Delta_i^{[r]}(t_n, h). \end{aligned}$$

Beweis:
Der Beweis verläuft analog zu dem von Lemma 3.5.4.

■

Nun gilt es noch, die Terme $\Delta_{ni}^{[r]}$, $r \in \mathbb{N}$ abzuschätzen. Satz 3.5.8 liefert die gesuchte Schranke. Es ergibt sich eine ähnliche Darstellung wie in Satz 3.5.3.

Satz 3.5.8
Sei $0 < \nu \leq 1$ und $\widetilde{A}^{\nu-1}\tilde{g}^{(r)} \in L^\infty([0,1]; Y)$. Dann gilt

$$h^{1-\nu} \left\| \Delta_i^{[r]}(t_n, h) \right\|_Y + \left\| \widetilde{A}^{\nu-1} \Delta_i^{[r]}(t_n, h) \right\|_Y \leq C h^{r+1} \sup_{0 \leq \tau \leq 1} \left\| \widetilde{A}^{\nu-1} \tilde{g}^{(r)}(t_n + \tau h) \right\|_Y$$

gleichmäßig für $0 \leq t_n \leq T$.

Beweis:
Der Beweis verläuft analog zu dem von Satz 3.5.3.

■

Satz 3.5.9
Die numerische Lösung u_n von (3.1) werde mit einem Verfahren (3.2) berechnet, das $\psi_1(-hA) = \psi_2(-hA) = \psi_{12}(-hA) = 0$ erfüllt. Außerdem sei $f : [0,T] \longrightarrow X$ zweimal (Fréchet)-differenzierbar.

1. *Sind $\widetilde{A}^{\beta-1}\tilde{g}'(t) \in L^\infty\big((0,T);Y\big)$ und $\widetilde{A}^{\kappa-1}\tilde{g}''(t) \in L^\infty\big((0,T);Y\big)$ für $\beta,\kappa \in (0,1]$, dann gilt die Fehlerabschätzung*

$$\|u_n - u(t_n)\|_Y \le C\Big(h^{1+\beta} \sup_{0\le t\le t_n} \Big\|\widetilde{A}^{\beta-1}\tilde{g}'(t)\Big\|_Y + h^2 \sup_{0\le t\le t_n} \Big\|\widetilde{A}^{\kappa-1}\tilde{g}''(t)\Big\|_Y\Big).$$

2. *Sind $\tilde{g}'(t) \in L^\infty\big((0,T);Y\big)$ und $\widetilde{A}^{\kappa-1}\tilde{g}''(t) \in L^\infty\big((0,T);Y\big)$, so ist das Verfahren folglich konsistent der Ordnung 2.*

Beweis:
Löst man die Rekursion in Lemma 3.5.4, so erhält man

$$\begin{aligned} e_h(t_n) = h\sum_{j=0}^{n-1} e^{-(n-j-1)hA} \sum_{i=1}^{s} b_i(-hA)\Big(g\big(t_j + c_i h, k_i(t_j,u_j,h)\big) - \tilde{g}(t_j + c_i h)\Big) \\ - \sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j}) \end{aligned} \tag{3.18}$$

als explizite Darstellung der $e_h(t_n)$.

Der Beweis mittels Induktion nach n ist derselbe wie in Satz 3.5.6, wobei man lediglich

$$\varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big)$$

durch

$$\sum_{i=1}^{s} b_i(-hA)\Big(g\big(t_j + c_i h, k_i(t_j,u_j,h)\big) - \tilde{g}(t_j + c_i h)\Big)$$

ersetzt.

Auch hier wird die Lipschitzbedingung an g

$$\begin{aligned} \big\|g(t_n + c_i h, k_i(t_n,u_n,h)) - \tilde{g}(t_n + c_i h)\big\| &\le L\|k_i(t_n,u_n,h) - u(t_n + c_i h)\|_Y \\ &= L\|E_i(t_n,u_n,h)\|_Y \end{aligned} \tag{3.19}$$

verwendet, um den Konvergenzfehler abzuschätzen.

Zusammen mit der Definition von $\|\cdot\|_Y$, Lemma 3.5.1 und Satz 3.5.3 erhält man

$$\begin{aligned}
\|e_h(t_n)\|_Y &\overset{(3.18)}{=} \Big\| h\sum_{j=0}^{n-1} e^{-(n-j-1)hA} \sum_{i=1}^{s} b_i(-hA)\Big(g\big(t_j+c_ih, k_i(t_j,u_j,h)\big) - \tilde g(t_j+c_ih)\Big)\Big\|_Y \\
&\quad + \Big\| \sum_{j=0}^{n-1} e^{-jhA}\delta_h(t_{n-j})\Big\|_Y \\
&\le h\sum_{j=0}^{n-1}\sum_{i=1}^{s} \Big\| \widetilde A^{\alpha} e^{-(n-j-1)hA} b_i(-hA)\Big(g\big(t_j+c_ih, k_i(t_j,u_j,h)\big) - \tilde g(t_j+c_ih)\Big)\Big\| \\
&\quad + Ch^2 \sup_{0\le t\le t_n} \|\widetilde A^{\kappa-1}\tilde g''(t)\|_Y \\
&\overset{(3.19)}{\le} C\Big(h\sum_{j=0}^{n-2} \big((n-j-1)h\big)^{-\alpha} \|E(t_j,u_j,h)\|_Y + h^{1-\alpha}\|E(t_{n-1},u_{n-1},h)\|_Y\Big) \\
&\quad + Ch^2 \sup_{0\le t\le t_n} \|\widetilde A^{\kappa-1}\tilde g''(t)\|_Y \\
&\overset{(3.15)}{\le} C\Big(h\sum_{j=0}^{n-1} t_{n-j}^{-\alpha}\|E(t_j,u_j,h)\|_Y + h^2 \sup_{0\le t\le t_n} \|\widetilde A^{\kappa-1}\tilde g''(t)\|_Y\Big),
\end{aligned}$$

wobei $E(t_j,u_j,h) := \max\limits_{2\le i\le s} E_i(t_j,u_j,h)$ ist.

Im Gegensatz zum Nørsett–Euler-Verfahren kann nun noch nicht das diskrete Gronwall-Lemma angewandt werden, da der Fehler an den Zwischenstellen auftaucht.

Nach Lemma 3.5.7 gilt jedoch wegen $\psi_{12}(-hA) = 0$

$$\begin{aligned}
E_i(t_n,u_n,h) &= e^{-c_ihA}e_h(t_n) + h\sum_{j=1}^{i-1} a_{ij}(-hA)\Big(g\big(t_n+c_jh, k_j(t_n,u_n,h)\big) - \tilde g(t_n+c_jh)\Big) \\
&\quad - \Delta_i(t_n,h)
\end{aligned}$$

und nach Satz 3.5.8 ist

$$\|\Delta_i(t_n,h)\|_Y \le Ch^{1+\beta} \sup_{0\le\tau\le1} \big\|\widetilde A^{\beta-1}\tilde g'(t_n+\tau h)\big\|_Y,$$

was zusammen

$$\|E(t_j,u_j,h)\|_Y \le C\left(\|e_h(t_n)\|_Y + h\|E(t_j,u_j,h)\|_Y + h^{1+\beta}\sup_{0\le\tau\le1}\big\|\widetilde A^{\beta-1}\tilde g'(t_n+\tau h)\big\|_Y\right)$$

und damit auch

$$\|E(t_j,u_j,h)\|_Y \le C\left(\|e_h(t_n)\|_Y + h^{1+\beta}\sup_{0\le\tau\le1}\big\|\widetilde A^{\beta-1}\tilde g'(t_n+\tau h)\big\|_Y\right)$$

impliziert.

Insgesamt erhält man also die Abschätzung

$$\begin{aligned}
\|e_h(t_n)\|_Y &\le Ch\sum_{j=0}^{n-1} t_{n-j}^{-\alpha}\|e_h(t_n)\|_Y \\
&\quad + C\Big(h^{1+\beta}\sup_{0\le\tau\le1}\big\|\widetilde A^{\beta-1}\tilde g'(t_n+\tau h)\big\|_Y + h^2\sup_{0\le t\le t_n}\|\widetilde A^{\kappa-1}\tilde g''(t)\|_Y\Big),
\end{aligned}$$

so dass man das diskrete Gronwall-Lemma (3.5.5) mit $\rho = \alpha$, $\varepsilon_\nu = \|e_h(t_\nu)\|_Y$ sowie $b = C\Big(h^{1+\beta} \sup\limits_{0\le\tau\le1} \big\|\widetilde{A}^{\beta-1}\tilde{g}'(t_n+\tau h)\big\|_Y + h^2 \sup\limits_{0\le t\le t_n} \|\widetilde{A}^{\kappa-1}\tilde{g}''(t)\|_Y\Big)$ anwenden kann. Dies liefert die zu zeigende Behauptung.

■

Beispiel:
Gesucht ist ein Verfahren der Ordnung 2 mit Stufenzahl $s = 2$. Die Ordnungsbedingungen lauten in diesem Fall

$$b_1 + b_2 \equiv \varphi_1,\ b_2c_2 \equiv \varphi_2 \text{ und } a_{21} \equiv c_2\varphi_{1,2}$$

mit $\varphi_{1,2}(-hA) := \varphi_1(-c_2hA)$.

Die expliziten Verfahren, die diese Bedingungen erfüllen, lassen sich als

$$\begin{array}{c|cc} 0 & 0 & 0 \\ c_2 & c_2\varphi_{1,2} & 0 \\ \hline & \varphi_1 - \frac{1}{c_2}\varphi_2 & \frac{1}{c_2}\varphi_2 \end{array} \tag{3.20}$$

schreiben, wobei $c_2 \neq 0$ ein freier Parameter ist.

3.6. Variable Schrittweiten

Nun wird die Schrittweite nicht mehr als konstant angenommen. Daher steht h im Folgenden für die Schrittweitenfolge $h = (h_0, ..., h_{M-1})$ mit $h_j > 0$ für $j = 0, ..., M-1$, wobei $h_n = t_{n+1} - t_n$ die Schrittweite zwischen den beiden Stützstellen t_n und t_{n+1} auf dem nicht-äquidistanten Gitter $\Omega_h = \{0 = t_0 < t_1 < ... < t_M = T\}$ ist.

Die Notation lässt sich vereinfachen, indem man $t_{nkj} := t_n - t_{k-j}$ und $t_{nk} := t_{nk0} = t_n - t_k$ setzt. Ensprechend ist $t_{n00} = t_n$, so dass in t_{nkj} Nullen am Ende weggelassen werden können.

Außerdem wird $h_{*j} := \max\limits_{0\le k\le j-1} h_k$ gesetzt. Unter $h \to 0$ ist demnach $h_{*M} \to 0$ zu verstehen, wobei die Zahl der Schritte M natürlich von h abhängt.

Um nun Konvergenz auch im Falle variabler Schrittweiten zu zeigen, müssen ein paar Beweise modifiziert werden. Insbesondere das diskrete Gronwall-Lemma (3.5.5) setzt eine konstante Schrittweite voraus und muss daher ersetzt werden.

Lemma 3.6.1
Seien $(h_0, ..., h_{M-1}) \in \mathbb{R}^M$ mit $h_j > 0$ für $j = 0, ..., M-1$ und $M \in \mathbb{N}_1$. Außerdem seien $\varepsilon_n > 0$ für $n = 1, ..., M$ gegeben, so dass für gewisse $0 \le \rho < 1$ und $a, b \ge 0$

$$\varepsilon_n \le a \sum_{k=1}^{n-1} h_k t_{nk}^{-\rho} \varepsilon_k + h_{*n} b$$

gilt. Dann ist

$$\varepsilon_n \le C h_{*n} e^{Et_n} b$$

für $n = 1, ..., M$. Dabei hängen C und E nur von ρ, a und T ab.

Beweis:
Der Beweis erfolgt mittels Induktion nach n.

Induktionsanfang $n = 1$:
Die leere Summe ist identisch null, so dass sofort die Behauptung folgt.

Induktionsschritt $(1, ..., n-1) \to n \leq M$:
Das weitere Vorgehen orientiert sich nun am Beweis von Lemma 4.7 in [2]. Danach gilt für $E > 0$ die Abschätzung

$$\sum_{k=1}^{n-1} h_k t_{nk}^{-\rho} e^{-Et_{nk}} \leq \int_0^{t_n} e^{-E(t_n-t)}(t_n - t)^{-\rho} dt \leq BE^{\rho-1}, \tag{3.21}$$

wobei $B = \displaystyle\int_0^\infty e^{-\tau}\tau^{-\rho} d\tau$ ist. Die rechte Ungleichung basiert dabei auf der Substitution $\tau = E(t_n - t)$, während links eine Untersumme des Integrals steht.

Wähle nun E groß genug, so dass

$$1 + aBE^{\rho-1}C \leq C \tag{3.22}$$

ist.

Dann folgt

$$\begin{aligned}
\varepsilon_n &\leq a \sum_{k=1}^{n-1} h_k t_{nk}^{-\rho} \varepsilon_k + h_{*n} b \overset{IV}{\leq} a \sum_{k=1}^{n-1} h_k t_{nk}^{-\rho} C h_{*k} e^{Et_k} b + h_{*n} b \\
&\leq a \sum_{k=1}^{n-1} h_k t_{nk}^{-\rho} e^{-Et_{nk}} e^{Et_n} C h_{*n} b + h_{*n} b \overset{(3.21)}{\leq} aBE^{\rho-1} e^{Et_n} C h_{*n} b + h_{*n} b \\
&\overset{(3.22)}{\leq} C h_{*n} e^{Et_n} b.
\end{aligned}$$

■

Der Einfachheit halber wird nur das Nørsett–Euler-Verfahren betrachtet. Dies vereinfacht die Rechnungen, da das Verfahren einstufig ist. Im Fall variabler Schrittweite lässt es sich als

$$u_{n+1} = e^{-h_n A} u_n + h_n \varphi(-h_n A) g(t_n, u_n) \tag{3.23}$$

schreiben.

Lemma 3.6.2
Für den Konvergenzfehler des Nørsett–Euler-Verfahrens gilt die Rekursion

$$e_h(t_{n+1}) = e^{-h_n A} e_h(t_n) + h_n \varphi(-h_n A)\Big(g(t_n, u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}),$$

wobei

$$\delta_h(t_{n+1}) = u(t_{n+1}) - \Big(e^{-h_n A} u(t_n) + h_n \varphi(-h_n A)\tilde{g}(t_n)\Big) \tag{3.24}$$

der lokale Fehler des Nørsett–Euler-Verfahrens im Fall variabler Schrittweiten ist.

Beweis:
Der Beweis erfolgt wie im Fall konstanter Schrittweite. Aus (3.24) erhält man

$$\begin{aligned} e_h(t_{n+1}) &= u_{n+1} - u(t_{n+1}) \\ &\overset{(3.23)}{=} e^{-h_n A}u_n + h_n\varphi(-h_nA)g(t_n,u_n) - u(t_{n+1}) \\ &\overset{(3.24)}{=} e^{-h_nA}e_h(t_n) + h_n\varphi(-h_nA)\Big(g(t_n,u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}). \end{aligned}$$

■

Satz 3.6.3 liefert nun die wesentlichen Abschätzungen für den lokalen Diskretisierungsfehler $\delta_h(t_n)$ bei Verwendung variabler Schrittweiten.

Satz 3.6.3
Sei $0 < \nu \leq 1$ mit $\widetilde{A}^{\nu-1}\tilde{g}' \in L^\infty(0,T;Y)$. Dann gelten

$$h_n^{1-\nu}\,\|\delta_h(t_{n+1})\|_Y + \left\|\widetilde{A}^{\nu-1}\delta_h(t_{n+1})\right\|_Y \leq Ch_n^2 \sup_{t_n\leq t\leq t_{n+1}} \left\|\widetilde{A}^{\nu-1}\tilde{g}'(t)\right\|_Y$$

und

$$\left\|\sum_{j=0}^{n-1} e^{-t_{nnj}A}\delta_h(t_{n-j})\right\|_Y \leq Ch_{*n} \sup_{0\leq t\leq t_n} \left\|\widetilde{A}^{\nu-1}\tilde{g}'(t)\right\|_Y$$

gleichmäßig für $0 \leq t_n \leq T$. Insbesondere hängt C für festes T nicht von n oder h ab.

Beweis:
Wegen $t_{n+1} = t_n + h_n$ ist der erste Teil analog zu Satz 3.5.3 mit h_n statt h. Dabei gilt

$$u(t_n + h_n) = e^{-h_nA}u(t_n) + \int_0^{h_n} e^{-(h_n-\tau)A}\tilde{g}(t_n+\tau)d\tau, \tag{3.25}$$

so dass zusammen mit

$$\begin{aligned} \tilde{g}(t_n+\tau) &= \tilde{g}(t_n) + \int_0^\tau \tilde{g}'(t_n+\sigma)d\sigma, \\ \varphi(-tA) &= \frac{1}{t}\int_0^t e^{-(t-\tau)A}d\tau \end{aligned} \tag{3.26}$$

die Fehlerdarstellung

$$\begin{aligned} \delta_h(t_{n+1}) &\overset{(3.24)}{=} u(t_n+h_n) - \Big(e^{-h_nA}u(t_n) + h_n\varphi(-h_nA)\tilde{g}(t_n)\Big) \\ &\overset{(3.25)}{=} \int_0^{h_n} e^{-(h_n-\tau)A}\tilde{g}(t_n+\tau)d\tau - h_n\varphi(-h_nA)\tilde{g}(t_n) \\ &\overset{(3.26)}{=} \int_0^{h_n} e^{-(h_n-\tau)A}\int_0^\tau \tilde{g}'(t_n+\sigma)d\sigma d\tau \end{aligned}$$

folgt. Dies impliziert

$$\|\delta_h(t_{n+1})\|_Y \leq \int_0^{h_n} \left\|\widetilde{A}^{1-\nu}e^{-(h_n-\tau)A}\right\| \int_0^\tau \left\|\widetilde{A}^{\nu-1}\tilde{g}'(t_n+\sigma)\right\|_Y d\sigma d\tau, \tag{3.27}$$

so dass man mit

$$\left\|(h_n-\tau)^{1-\nu}\widetilde{A}^{1-\nu}e^{-(h_n-\tau)A}\right\| + \left\|e^{-(h_n-\tau)A}\right\| \le C \tag{3.28}$$

(siehe Lemma 3.5.1) die erste Abschätzung durch

$$\begin{aligned}
&h_n^{1-\nu}\left\|\delta_h(t_{n+1})\right\|_Y + \left\|\widetilde{A}^{\nu-1}\delta_h(t_{n+1})\right\|_Y \\
&\overset{(3.27)}{\le} \int_0^{h_n}\left(h_n^{1-\nu}\left\|\widetilde{A}^{1-\nu}e^{-(h_n-\tau)A}\right\| + \left\|e^{-(h_n-\tau)A}\right\|\right)\int_0^{\tau}\left\|\widetilde{A}^{\nu-1}\tilde{g}'(t_n+\sigma)\right\|_Y d\sigma d\tau \\
&\overset{(3.28)}{\le} C\cdot h_n\int_0^{h_n}\left(h_n^{1-\nu}(h_n-\tau)^{\nu-1}+1\right)d\tau \sup_{t_n\le t\le t_{n+1}}\left\|\widetilde{A}^{\nu-1}\tilde{g}'(t)\right\|_Y
\end{aligned}$$

erhält. Zusammen mit der Dreiecksungleichung und erneut Lemma 3.5.1 folgt daraus wegen

$$\begin{aligned}
\left\|\sum_{j=0}^{n-1}e^{-t_{nnj}A}\delta_h(t_{n-j})\right\|_Y &\le \left\|\delta_h(t_n)\right\|_Y + \sum_{j=1}^{n-1}\left\|\widetilde{A}^{1-\nu}e^{-t_{nnj}A}\widetilde{A}^{\nu-1}\delta_h(t_{n-j})\right\|_Y \\
&\le h_{n-1}^{\nu-1}h_{n-1}^{1-\nu}\left\|\delta_h(t_n)\right\|_Y + C\cdot\sum_{j=1}^{n-1}t_{nnj}^{\nu-1}\left\|\widetilde{A}^{\nu-1}\delta_h(t_{n-j})\right\|_Y \\
&\le C\cdot\sum_{j=1}^{n-1}h_{n-j-1}^2 t_{nnj}^{\nu-1}\sup_{0\le t\le t_n}\left\|\widetilde{A}^{\nu-1}\tilde{g}'(t)\right\|_Y \\
&\le C\cdot h_{*n}\sup_{0\le t\le t_n}\left\|\widetilde{A}^{\nu-1}\tilde{g}'(t)\right\|_Y
\end{aligned}$$

auch die zweite Ungleichung. Im letzten Schritt wurde

$$\sum_{j=1}^{n-1}h_{n-j-1}t_{nnj}^{\nu-1} \le \int_0^T (T-t)^{\nu-1}dt \le C$$

und im vorletzten Schritt neben der bereits bewiesenen Ungleichung aus dem ersten Teil die Identität $t_{n,n,1} = t_n - t_{n-1} = h_{n-1}$ verwendet.

■

Damit sind die Voraussetzungen geschaffen, um Konvergenz der Ordnung 1 auch im Fall variabler Schrittweiten zu zeigen.

Satz 3.6.4
Die numerische Lösung u_n von (3.1) werde mit dem Nørsett–Euler-Verfahren (3.23) berechnet. Außerdem sei ein $\beta \in (0,1]$ gegeben, so dass $\widetilde{A}^{\beta-1}\tilde{g}' \in L^\infty((0,T);V)$ gilt mit $\widetilde{A} = A + \omega I$. Ferner sei die Schrittweitenbedingung

$$\frac{h_j}{h_{j+1}} \le C$$

erfüllt.

Dann gilt die Fehlerabschätzung

$$\|u_n - u(t_n)\|_Y \le C\cdot h_{*n}\sup_{0\le t\le t_n}\left\|\widetilde{A}^{\beta-1}\tilde{g}'(t)\right\|_Y$$

gleichmäßig für $0 \le t_n \le T$.

Beweis:
Nach Lemma 3.6.2 gilt

$$e_h(t_{n+1}) = e^{-h_n A} e_h(t_n) + h_n \varphi(-h_n A)\Big(g(t_n, u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}), \tag{3.29}$$

wobei $e_h(t_n) = u_n - u(t_n)$ der Konvergenzfehler an der Stelle t_n ist.

Mittels Induktion nach n kann man nun zeigen, dass

$$e_h(t_n) = \sum_{j=0}^{n-1} h_j e^{-t_{n,j+1}A} \varphi(-h_j A)\Big(g(t_j, u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n-1} e^{-t_{nnj}A} \delta_h(t_{n-j}) \tag{3.30}$$

die explizite Darstellung der $e_h(t_n)$ für $n \in \mathbb{N}_0$ ist.

Dabei werden die Eigenschaften der t_{nkj}

$$\begin{aligned} h_n + t_{nj} &= t_{n+1} - t_n + t_n - t_j = t_{n+1,j}, \\ h_n + t_{nnj} &= t_{n+1} - t_n + t_n - t_{n-j} = t_{n+1,n,j} = t_{n+1,n+1,j+1} \end{aligned} \tag{3.31}$$

und

$$t_{nn0} = t_{nn} = t_n - t_n = 0 \tag{3.32}$$

verwendet.

Induktionsanfang $n = 0$:
Wegen $u(t_0) = t_0$ ist $e_h(t_0) = u_0 - u(t_0) = 0$ und da die leeren Summen ebenfalls identisch null sind, gilt die Gleichheit.

Induktionsschritt $n \to (n+1)$:
Nimmt man nach Induktionsvoraussetzung an, dass (3.30) für n gilt, so folgt

$$\begin{aligned} e_h(t_{n+1}) &= e^{-h_n A} e_h(t_n) + h_n \varphi(-h_n A)\Big(g(t_n, u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}) \\ &\overset{IV}{=} e^{-h_n A}\left(\sum_{j=0}^{n-1} h_j e^{-t_{n,j+1}A} \varphi(-h_j A)\Big(g(t_j, u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n-1} e^{-t_{nnj}A} \delta_h(t_{n-j})\right) \\ &\quad + h_n \varphi(-h_n A)\Big(g(t_n, u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}) \\ &\overset{(3.31)}{=} \sum_{j=0}^{n-1} h_j e^{-t_{n+1,j+1}A} \varphi(-hA)\Big(g(t_j, u_j) - \tilde{g}(t_j)\Big) \\ &\quad + h_n \varphi(-hA)\Big(g(t_n, u_n) - \tilde{g}(t_n)\Big) - \delta_h(t_{n+1}) - \sum_{j=1}^{n} e^{-t_{n+1,n+1,j}A} \delta_h\big(t_{n-(j-1)}\big) \\ &\overset{(3.32)}{=} \sum_{j=0}^{n} h_j e^{-t_{n+1,j+1}A} \varphi(-hA)\Big(g(t_j, u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n} e^{-t_{n+1,n+1,j}A} \delta_h\big(t_{n+1-j}\big). \end{aligned}$$

Damit ist die Induktion abgeschlossen.

Aufgrund der Lipschitz-Bedingung von g ist

$$\|g(t_n, u_n) - \tilde{g}(t_n)\| = \|g(t_n, u_n) - g(t_n, u(t_n))\| \le L\|u_n - u(t_n)\|_Y = L\|e_h(t_n)\|_Y. \tag{3.33}$$

Außerdem gilt für $j \in \mathbb{N}_0$ nach Voraussetzung

$$\frac{h_j}{h_{j+1}} \leq C$$

und damit für $j \leq n-2$ auch

$$\frac{t_{nj}}{t_{n,j+1}} = \frac{t_n - t_j}{t_n - t_{j+1}} = \frac{t_n - t_j}{t_n - t_j - h_j} \leq C. \tag{3.34}$$

Zusammen mit der Definition von $\|\cdot\|_Y$, Lemma 3.5.1 und Satz 3.6.3 erhält man daher

$$\begin{aligned}
\|e_h(t_n)\|_Y &\overset{(3.30)}{=} \Big\| \sum_{j=0}^{n-1} h_j e^{-t_{n,j+1}A} \varphi(-h_j A)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big) - \sum_{j=0}^{n-1} e^{-t_{nnj}A} \delta_h(t_{n-j}) \Big\|_Y \\
&\leq \sum_{j=0}^{n-1} \Big\| h_j \widetilde{A}^{\alpha} e^{-t_{n,j+1}A} \varphi(-hA)\Big(g(t_j,u_j) - \tilde{g}(t_j)\Big)\Big\| + \Big\| \sum_{j=0}^{n-1} e^{-t_{nnj}A} \delta_h(t_{n-j}) \Big\|_Y \\
&\overset{(3.33)}{\leq} \sum_{j=0}^{n-1} h_j \Big\| \widetilde{A}^{\alpha} e^{-t_{n,j+1}A} \varphi(-hA)\Big\| \cdot L\|e_h(t_j)\|_Y + \Big\| \sum_{j=0}^{n-1} e^{-t_{nnj}A} \delta_h(t_{n-j}) \Big\|_Y \\
&\leq C\Bigg(\sum_{j=1}^{n-2} h_j t_{n,j+1}^{-\alpha} \|e_h(t_j)\|_Y + h_{n-1}^{1-\alpha} \|e_h(t_{n-1})\|_Y + h_{*n} \sup_{0\leq t\leq t_n} \|\widetilde{A}^{\beta-1}\tilde{g}'(t)\|_Y \Bigg) \\
&\overset{(3.34)}{\leq} C\Bigg(\sum_{j=1}^{n-1} h_j t_{nj}^{-\alpha} \|e_h(t_j)\|_Y + h_{*n} \sup_{0\leq t\leq t_n} \|\widetilde{A}^{\beta-1}\tilde{g}'(t)\|_Y \Bigg)
\end{aligned}$$

und mit dem Gronwall-Lemma 3.6.1 folgt schließlich

$$\|e_h(t_n)\|_Y \leq C h_{*n} \sup_{0\leq t\leq t_n} \|\widetilde{A}^{\beta-1}\tilde{g}'(t)\|_Y.$$

■

Bemerkung:
Die Fehlerkonstante enthält den Term e^{Et_n} aus dem Gronwall-Lemma 3.6.1 als Faktor. Der Satz liefert daher keine Aussage über das Langzeitverhalten.

4. Numerische Experimente

Gegenstand dieses Kapitels ist das numerische Verhalten der exponentiellen Integratoren am Beispiel der Nagumo-Gleichung

$$U_t = U_{xx} + U(1-U)(U-\alpha) + \Phi(x,t). \tag{4.1}$$

Alan Lloyd Hodgkin und Andrew Fielding Huxley präsentierten 1952 in [25] ihre Experimente an den Riesenaxonen von Tintenfischen. Ein Axon ist Teil einer Nervenzelle und dient zur Weiterleitung elektrischer Impulse. Das mathematische Modell, das insbesondere zur Simulation von Aktionspotentialen verwendet werden kann, bezeichnet man als Hodgkin-Huxley-Modell.

Das FitzHugh-Nagumo-Modell, das von Richard FitzHugh 1961 in [8] und J. Nagumo 1962 in [35] ausgearbeitet wurde, ist eine vereinfachte Version des Hodgkin-Huxley-Modells. Bezüglich des Zusammenhangs mit der Nagumo-Gleichung (4.1) sei insbesondere auf [34] verwiesen.

4.1. Ohne räumlichen Diskretisierungsfehler

Zunächst wird nur der Fehler der Zeitdiskretisierung betrachtet und zu diesem Zweck $\Phi(x,t)$ so gewählt, dass $U(x,t) = \Big(x - \sin(t)\Big) \cdot \Big(1 - (x - \sin(t))\Big)$ die exakte Lösung ist.

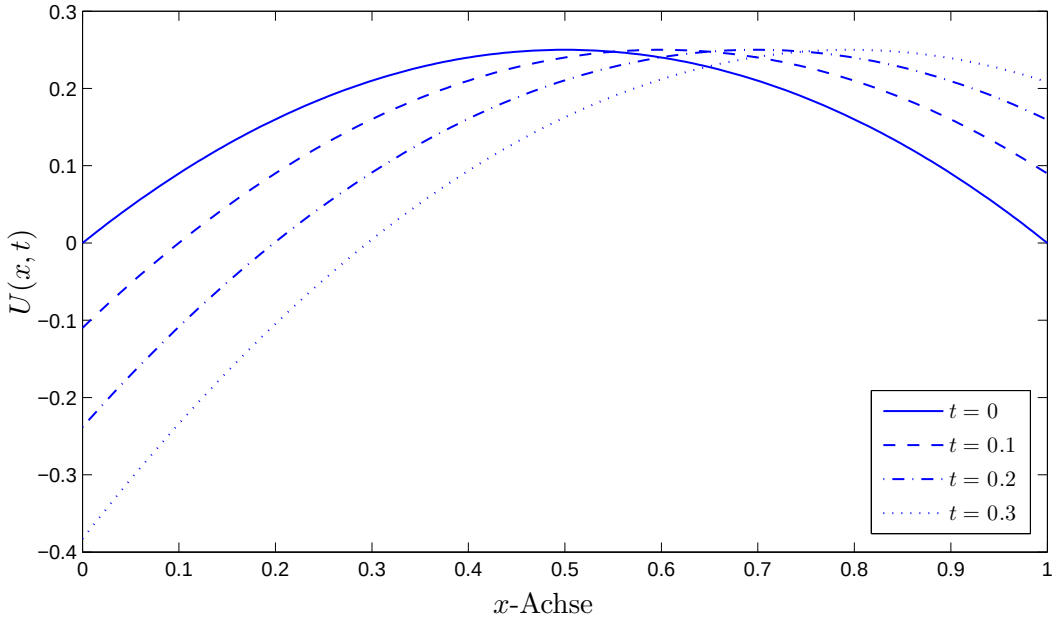


Abbildung 4.1.: Exakte Lösung der Differentialgleichung

Zum Zeitpunkt $t = 0$ erhält man die in Abbildung 4.1 (mittels der durchgezogenen Kurve) dargestellte Parabel $U(x, 0) = x(1 - x)$ und für $t > 0$ führen die Terme $\sin(t)$ dazu, dass sich diese sinusförmig im Raum bewegt.

Gesucht ist nun eine numerische Lösung der gegebenen Differentialgleichung auf dem Rechteck $[0, 1]\times[0, 1]$, wobei man die Anfangs- und die zeitabhängigen Dirichletschen Randbedingungen aus der exakten Lösung $U(x, 0) = x(1-x)$ sowie $U(0, t) = -\sin(t)(1+\sin(t))$ und $U(1, t) = (1 - \sin(t))\sin(t)$ erhält. Außerdem sei noch bemerkt, dass im Folgenden stets $\alpha = \frac{1}{4}$ ist.

Da unter den Voraussetzungen $F \in \mathcal{C}^4[a, b]$ und $x, x + h, x - h \in [a, b]$ die Abschätzung

$$\left|\frac{F(x-h) - 2F(x) + F(x+h)}{h^2} - F''(x)\right| \leq Ch^2 \max_{\zeta\in[a,b]} \left|F^{(4)}(\zeta)\right|$$

für die numerische Differentiation gilt, spielt der Fehler der Raumdiskretisierung, die mittels Standard-Finite-Differenzen durchgeführt wird, keine Rolle. Der Vollständigkeit halber sei bemerkt, dass in diesem Beispiel $\Delta x = \frac{1}{100} = \frac{1}{N+1}$ ist. Damit sind $x_j = \frac{j}{100}$ mit $j = 1, ..., 99$ die inneren Punkte der räumlichen Diskretisierung.

Insgesamt liefert dies ein Liniensystem

$$\begin{cases} u'(t) = f(t, u(t)), \\ u(0) = u_0 \end{cases} \tag{4.2}$$

wobei

$$f(t, v) = \underbrace{\frac{1}{\Delta x^2}\begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}}_{=:A} v + \frac{1}{\Delta x^2}\begin{pmatrix} U(0,t) \\ 0 \\ \vdots \\ 0 \\ U(1,t) \end{pmatrix} + \begin{pmatrix} (v_1 - v_1^2)\left(v_1 - \frac{1}{4}\right) + \Phi(x_1, t) \\ \vdots \\ \vdots \\ (v_N - v_N^2)(v_N - \frac{1}{4}) + \Phi(x_N, t) \end{pmatrix}$$

mit

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix} \in \mathbb{R}^N$$

und

$$u_0 = \begin{pmatrix} U(x_1, 0) \\ \vdots \\ U(x_N, 0) \end{pmatrix} \in \mathbb{R}^N$$

ist.

Die Zeitdiskretisierung wird nun einerseits mit dem Nørsett–Euler-Verfahren und andererseits mit dem exponentiellen Runge-Kutta-Verfahren der Ordnung 2, das im dritten Kapitel (siehe 3.20) konstruiert wurde, durchgeführt. Der freie Parameter ist dabei $c_2 = 1$.

Um die Konvergenz der beiden Verfahren untersuchen zu können, muss man noch eine Norm wählen. In endlich-dimensionalen Räumen sind Normen zwar äquivalent, allerdings hängen die zugehörigen Konstanten von der Dimension des Raums ab.

Als Normen werden im Folgenden die Maximumsnorm und die diskrete L^2-Norm gewählt, die für $v \in \mathbb{R}^N$ durch

$$\|v\|_{L^\infty} := \max_{j=1,\dots,N} |v_j|$$

und

$$\|v\|_{L^2} := \sqrt{\Delta x \sum_{j=1}^{N} v_j^2}$$

gegeben sind. Letzteres entspricht der gewöhnlichen L^2-Norm einer Intervalltreppenfunktion mit den Werten $v_1, \dots, v_N$ und der Intervallbreite Δx.

Ist $b - a = N \cdot \Delta x$, das heißt das Intervall $[a, b]$ ist in N Teilintervalle der Länge Δx zerlegt, so gilt

$$\|v\|_{L^2} = \sqrt{\Delta x \sum_{j=1}^{N} v_j^2} \leq \sqrt{N \cdot \Delta x \Big(\max_{j=1,\dots,N} |v_j| \Big)^2} = \sqrt{b-a} \cdot \|v\|_{L^\infty}.$$

Bei Verwendung des Raumintervalls $[0, 1]$ muss folglich die L^2-Norm kleiner oder gleich der L^∞-Norm sein, was sich in Abbildung 4.2 bestätigt.

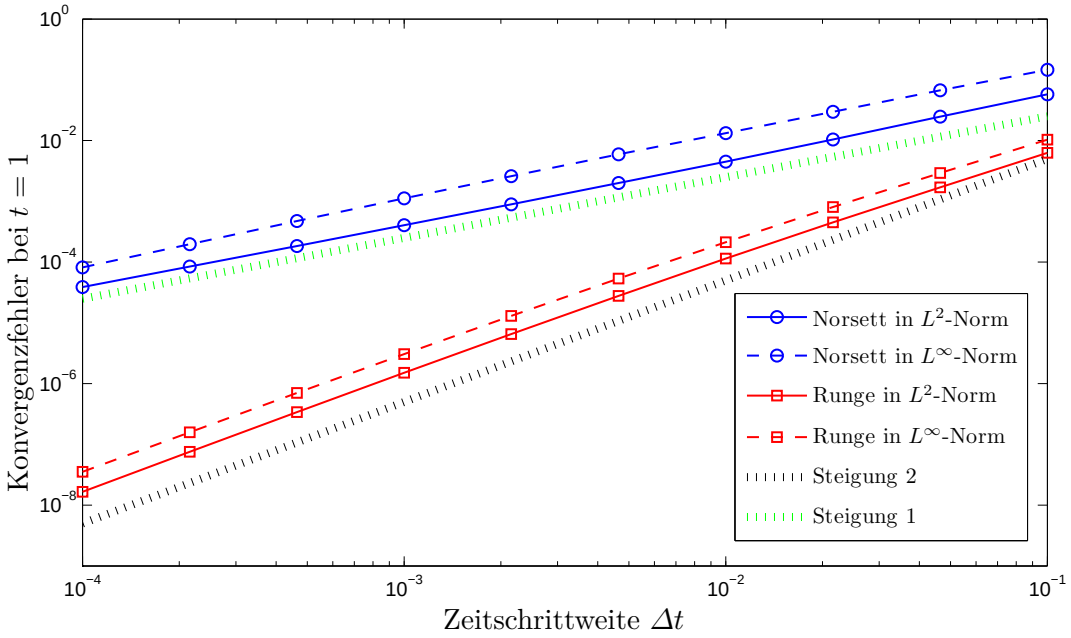


Abbildung 4.2.: Konvergenzverhalten des Nørsett-Euler-Verfahrens und des exponentiellen Runge-Kutta-Verfahrens aus (3.20)

Gegen die Zeitschrittweite ist der Konvergenzfehler an der Stelle $t = 1$ in L^∞-Norm und L^2-Norm aufgetragen. Die durchgezogene Kurve steht dabei für die L^2-Norm.

Der Konvergenzfehler des Nørsett–Euler-Verfahrens wird durch die Kreise dargestellt, während die Quadrate den Konvergenzfehler des exponentiellen Runge-Kutta-Verfahrens der Ordnung 2 symbolisieren. Diese Konventionen werden auch weiterhin beibehalten.

Aus der Steigung der gepunkteten Referenzgeraden ist ersichtlich, dass die beiden Verfahren bezüglich beider Normen die erwarteten experimentellen Konvergenzordnungen 1 bzw. 2 aufweisen.

4.2. Mit zeitlichem und räumlichem Diskretisierungsfehler

Im vorherigen Beispiel wurde die Anfangsrandwertaufgabe so gestellt, dass nur der zeitliche Diskretisierungsfehler beachtet werden musste. Wenngleich dies zu analytischen Zwecken nützlich sein kann, ist zu bemerken, dass es sich um keine realistische Annahme handelt. Aus diesem Grund werden im nächsten Beispiel sowohl der zeitliche als auch der räumliche Diskretisierungsfehler von Bedeutung sein.

4.2.1. Beispiel mit einer wandernden Welle

Im weiteren Verlauf sei $\Phi(x,t) = 0$, so dass man die Differentialgleichung

$$U_t = U_{xx} + U(1-U)(U-\alpha)$$

erhält. Wiederum wird der Fall $\alpha = \frac{1}{4}$ betrachtet. Die exakte Lösung lautet nun

$$U(x,t) = \frac{1}{1+\exp(-\frac{x-ct}{\sqrt{2}})},$$

wobei $c = -\sqrt{2}(\frac{1}{2}-\alpha)$ ist. Dies ist eine wandernde Welle mit Profil $v(\xi) = \dfrac{1}{1+\exp(-\frac{\xi}{\sqrt{2}})}$.

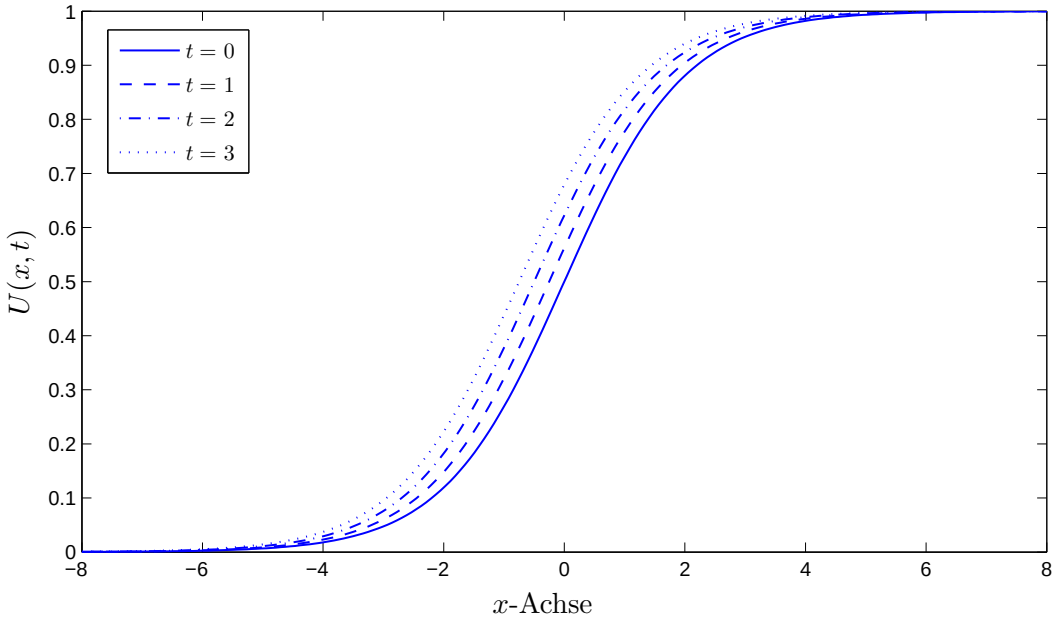


Abbildung 4.3.: Exakte Lösung der Nagumo-Gleichung

Für $x \to (-\infty)$ und $t \in [0,1]$ geht $U(x,t)$ exponentiell gegen 0 und für $x \to \infty$ gegen 1. Da das Ortsintervall $[-85, 85]$ groß genug ist, kann man die Dirichlet-Randbedingungen $u(-85,t) = 0$ und $u(85,t) = 1$ verwenden, ohne dass der zusätzliche Fehler numerisch bedeutsam ist.

Die Anfangsbedingung ist hingegen wie im vorherigen Beispiel durch die exakte Lösung zum Zeitpunkt $t = 0$ gegeben. Das Zeitintervall sei stets $[0,1]$, da man auf größeren Zeitintervallen qualitativ ohnehin dieselben Resultate erhält.

Diese Anfangsrandwertaufgabe wird nicht mehr verändert, variiert werden nur die numerischen Parameter, wobei jedoch immer Standard-Finite-Differenzen das Mittel zur Raumdiskretisierung bleibt.

4.2.2. Ergebnisse der Verfahren mit Padé-Approximation

Um einen ersten Eindruck vom Zusammenwirken der beiden Diskretisierungsfehler zu erhalten, wird das Nørsett–Euler-Verfahren mit verschiedenen Zeitschrittweiten angewandt, wobei die Raumschrittweite $\Delta x = \frac{1}{5}$ bzw. $\Delta x = \frac{1}{10}$ ist.

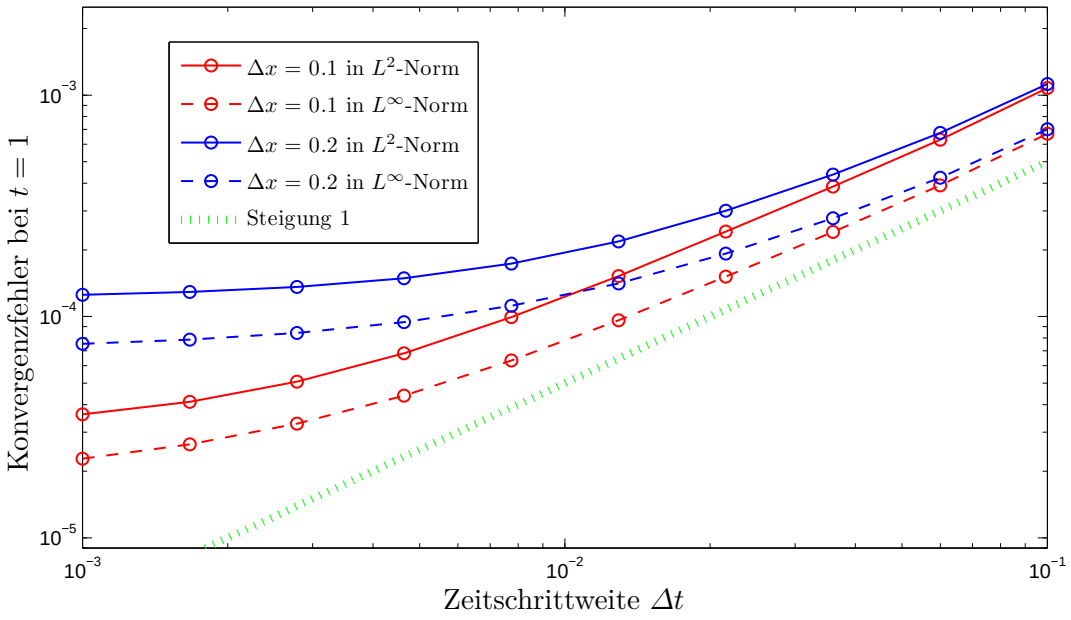


Abbildung 4.4.: Konvergenzverhalten des Nørsett–Euler-Verfahrens

Gegen die Zeitschrittweite ist der Konvergenzfehler an der Stelle $t = 1$ in der L^∞-Norm und in der L^2-Norm aufgetragen. Die durchgezogene Kurve steht dabei wiederum für die L^2-Norm, die wegen der großen Intervallbreite oberhalb der L^∞-Norm liegt.

Für große Zeitschrittweiten überwiegt der Fehler der Zeitdiskretisierung. Die Kurven verlaufen dort in etwa parallel zur gepunkteten Geraden mit Steigung 1. Für kleine Zeitschrittweiten überwiegt der Fehler der Ortsdiskretisierung. Die Kurve, die zur Raumschrittweite $\Delta x = \frac{1}{5}$ gehört, verläuft dort deutlich oberhalb der Kurve zur Raumschrittweite $\Delta x = \frac{1}{10}$. Ab einem gewissen Punkt kann der Konvergenzfehler nicht mehr durch Verkleinerung der Zeitschrittweite verringert werden.

Dies tritt bei Verfahren höherer Ordnung in verschärfter Form auf. Qualtitativ erhält man beim Verfahren der Ordnung 2 dasselbe Resultat wie beim Nørsett–Euler-Verfahren. Die Zeitskala in Abbildung 4.5 ist allerdings von 10^{-2} bis 10^{-0} gewählt, während sie in Abbildung 4.4 von 10^{-3} bis 10^{-1} geht. Unter denselben Voraussetzungen erhält man beim Verfahren der Ordnung 2 schon etwa ab $\Delta t = 10^{-1}$ keine weitere Verringerung des Konvergenzfehlers mehr.

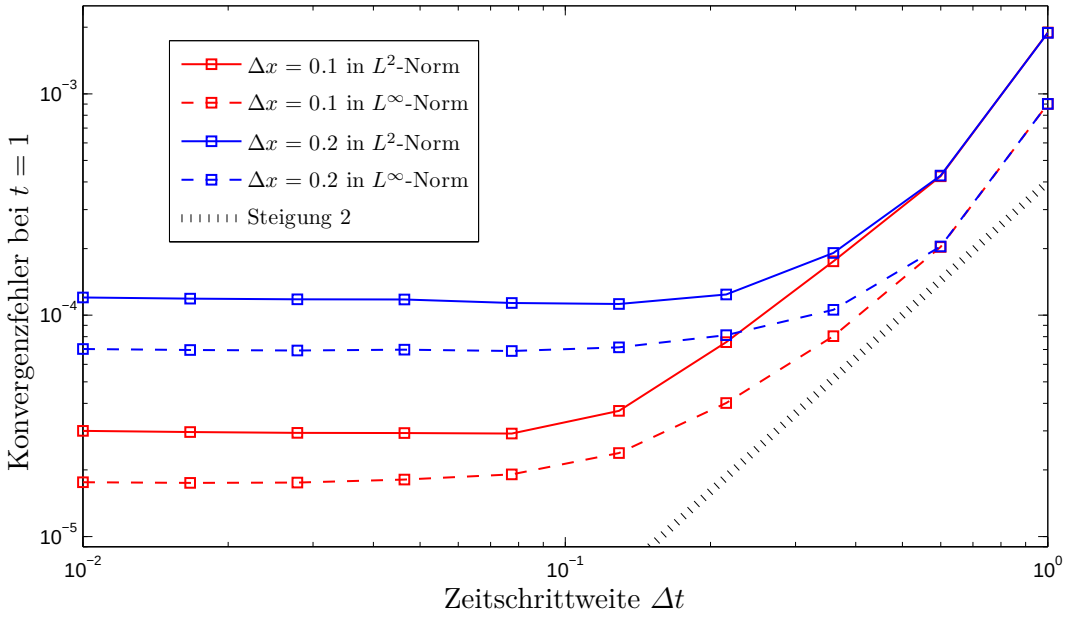
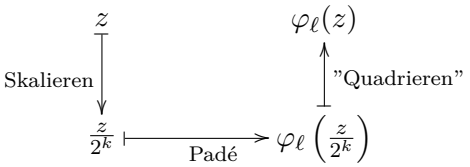


Abbildung 4.5.: Konvergenzverhalten des exponentiellen Runge-Kutta-Verfahrens der Ordnung 2 aus (3.20)

Die einzige Möglichkeit ist demnach, die Raumschrittweite zu verringern, was allerdings nicht ganz unproblematisch ist.

Die Berechnung der $\varphi_\ell(\Delta t A)$ wird mittels Padé-Approximation durchgeführt. Der erste Schritt besteht dabei darin, die Matrix $\Delta t A$ so zu skalieren, dass $\left\| \frac{\Delta t A}{2^k} \right\|_\infty < 1$ gilt.



Der zweite Schritt besteht in der eigentlichen Padé-Approximation $\varphi_\ell \approx \frac{N_d^\ell}{D_d^\ell}$ mit

$$N_d^\ell(z) = \frac{d!}{(2d+\ell)!} \sum_{i=0}^{d} \left(\sum_{j=0}^{i} \frac{(2d+\ell-j)!(-1)!}{j!(d-j)!(\ell+i-j)!} \right) z^i$$

$$D_d^\ell(z) = \frac{d!}{(2d+\ell)!} \sum_{i=0}^{d} \frac{(2d+\ell-i)!}{i!(d-i)!} (-z)^i.$$

Dabei ist d sowohl Zähler- als auch Nennergrad und $2d$ die Ordnung der Padé-Approximation. Die angegebenen Koeffizienten stammen aus [1].

Im dritten Schritt wird schließlich k-mal zurück skaliert. Im Fall von $\varphi_0(z) = e^z$ geschieht dies durch Quadrieren, da $e^{2z} = e^z e^z$ ist. Nähere Details zur Padé-Approximation des Exponentialoperators finden sich in [17] und [18].

Allgemein ergeben sich nach [1] die Formeln

$$\varphi_{2\ell}(2z) = \frac{1}{2^{2\ell}} \left(\varphi_\ell(z)\varphi_\ell(z) + \sum_{j=\ell+1}^{2\ell} \frac{2}{(2\ell - j)!} \varphi_j(z) \right),$$
$$\varphi_{2\ell+1}(2z) = \frac{1}{2^{2\ell+1}} \left(\varphi_\ell(z)\varphi_{\ell+1}(z) + \frac{1}{\ell!}\varphi_{\ell+1}(z) + \sum_{j=\ell+2}^{2\ell+1} \frac{2}{(2\ell + 1 - j)!} \varphi_j(z) \right),$$

die für $\varphi_1(z) = \varphi(z)$ und $\varphi_0(z) = e^z$ den Rekursionen

$$\varphi(2z) = \frac{1}{2}(e^z + 1)\varphi(z),$$
$$e^{2z} = e^z e^z$$

aus [20, 6.4] entsprechen.

Ist nun $\|\Delta t A\|_\infty > 1$, so muss die Matrix zur Berechnung von $e^{\Delta t A}$ k-mal quadriert werden. Da die Eigenschaft, dass A dünn-besetzt ist, bei der Padé-Approximation verloren geht, ist dies sehr aufwendig.

In der folgenden Grafik ist die benötigte Zeit zur Berechnung von $e^{\Delta t A}$ gegen die Zeitschrittweite Δt aufgetragen.
An der Stelle $\|\Delta t A\|_\infty = 1$ steigt die Rechenzeit erheblich, da skaliert werden muss.

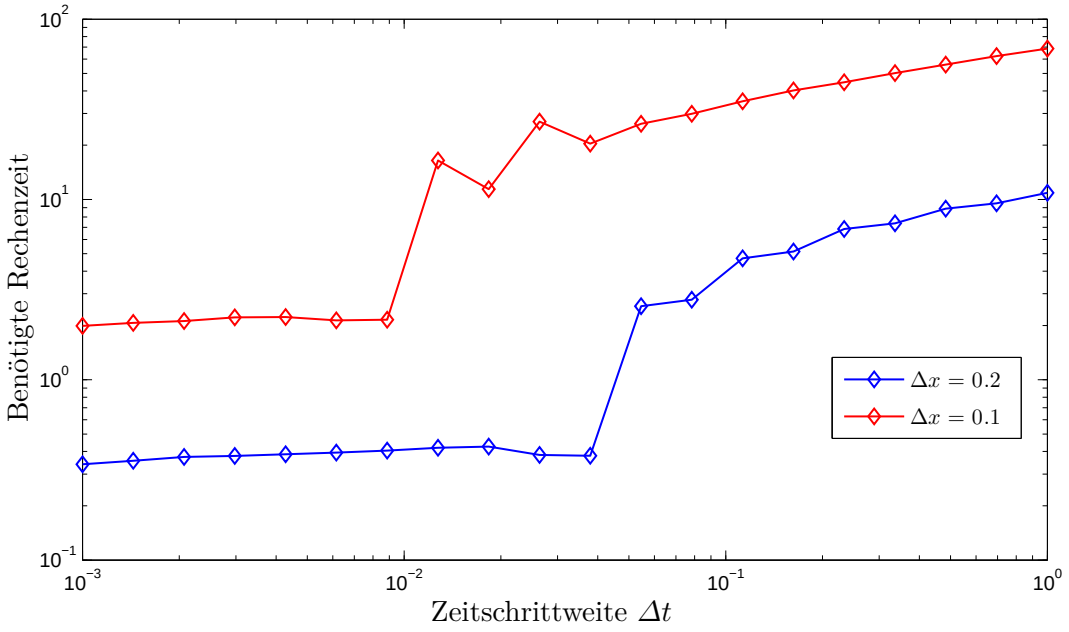


Abbildung 4.6.: Padé-Approximation des Exponentialoperators

Aus der Theorie der Runge-Kutta-Verfahren ist das Problem bekannt, dass eine Verkleinerung der Raumschrittweite Δx zur einer größeren Steifheit führt. Dieses Phänomen findet sich hier wieder, denn möchte man auf das Skalieren verzichten, so muss man die Zeitschrittweite klein genug wählen und erhält eine unpraktische Schrittweitenbedingung wie bei expliziten Runge-Kutta-Verfahren. Für die übrigen $\varphi-$Funktionen erhält man qualitativ dasselbe Resultat.

Fairerweise sollte an dieser Stelle allerdings noch der Fall betrachtet werden, dass das Zeitintervall sehr groß ist und entsprechend viele Zeitschritte nötig sind. Da die Berechnung der $\varphi-$Funktionen nur einmal durchgeführt werden muss, nimmt der Anteil am Gesamtaufwand mit einer Vergrößerung des Zeitintervalls ab. Doch selbst wenn einmalige Rechnungen vernachlässigt werden können, bleibt das Problem, dass in jedem Schritt Multiplikationen mit vollbesetzten Matrizen durchgeführt werden müssen.

Anhand dieser Grafik ist das Verhältnis von 100 Schritten des Nørsett–Euler-Verfahrens mit Schrittweite $\Delta t = \frac{1}{100}$ in Abhängigkeit der Raumschrittweite ersichtlich.

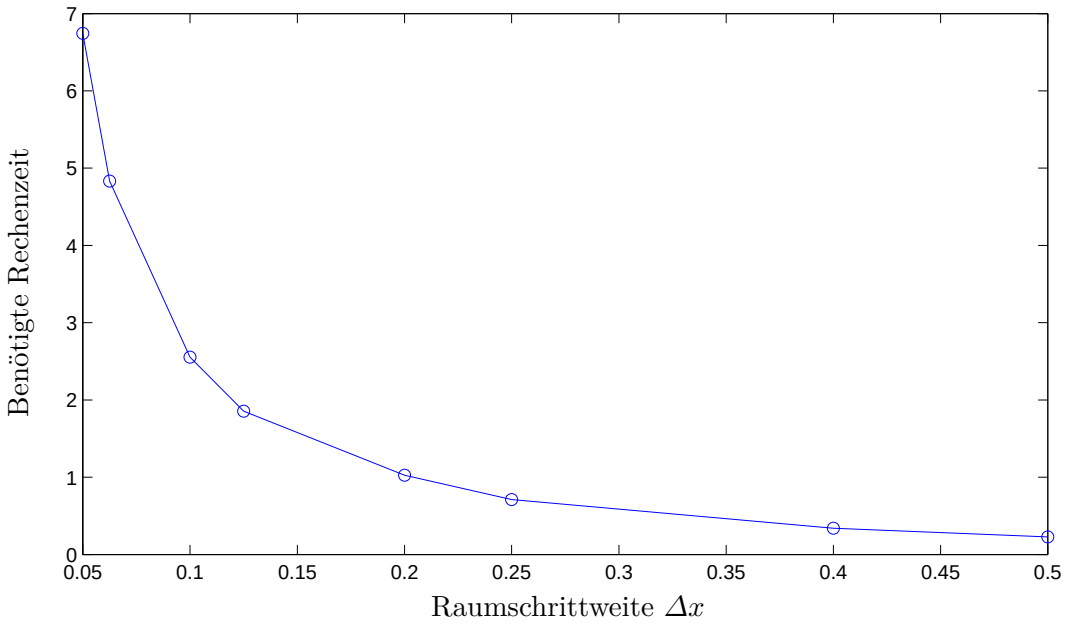


Abbildung 4.7.: Laufzeit des Nørsett–Euler-Verfahrens mit $\Delta t = \frac{1}{100}$ unter Vernachlässigung der Matrixberechnungen

Wie zu erwarten ist, steigt die Rechenzeit quadratisch mit der Raumschrittweite an, denn die Eigenschaft, dass die Matrix ΔtA dünn-besetzt ist, überträgt sich nicht auf $e^{\Delta tA}$ bzw. $\varphi(\Delta tA)$. Bei einem impliziten Runge-Kutta-Verfahren steigt der Aufwand hingegen nur linear.

Ein Ausweg könnte in Analogie zu den Verfahren aus dem ersten Kapitel darin bestehen, dass $e^{\Delta tA}$ und $\varphi(\Delta tA)$ selbst gar nicht benötigt werden, sondern nur das Resultat nach Anwendung auf einen Vektor, wenngleich dies in jedem Schritt für neue Vektoren ausgewertet werden müsste.

Eine Idee besteht dabei darin, die Padé-Approximation $P_d^\ell(z) \approx \dfrac{N_d^\ell(z)}{D_d^\ell(z)}$ zu faktorisieren, so dass zur Berechnung von $e^{\Delta t A}v$ und $\varphi(\Delta t A)v$ nur noch d Matrix-Vektor-Multiplikationen der Form $(A-wI)v$ durchgeführt und d lineare Gleichungssysteme der Form $(A-wI)x = v$ gelöst werden müssen. Im Falle komplexer Nullstellen $a \pm bi$ empfiehlt es sich allerdings, diese zu quadratischen Termen $z^2 - 2a + a^2 + b^2$ zusammenzufassen.

Das Problem besteht nun allerdings darin, dass bei dieser modifizierten Padé-Approximation ein Skalieren der Matrix nicht möglich ist.

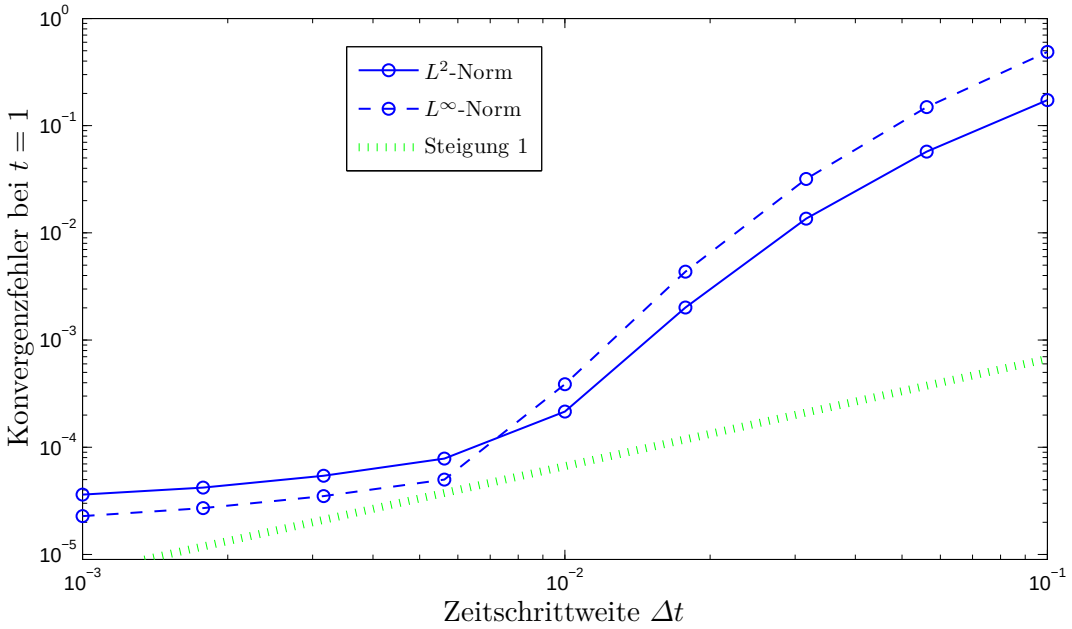


Abbildung 4.8.: Konvergenzverhalten des Nørsett–Euler-Verfahrens mit modifizierter Padé-Approximation

Das Resultat ist die zuvor bereits erwähnte Schrittweitenbeschränkung $\|\Delta t A\|_\infty < 1$. Wählt man eine größere Zeitschrittweite, so werden Konsistenz- und Konvergenzfehler des Verfahrens sehr groß.

Mit Blick auf die lineare Theorie bestünde noch die Möglichkeit, ein paar Schritte lang die Bedingung $\|\Delta t A\|_\infty < 1$ einzuhalten oder ein implizites Runge-Kutta-Verfahren zu verwenden. Die Hoffnung, dass dann der Anteil der Eigenvektoren mit sehr kleinen (also betragsmäßig großen) Eigenwerten an der Lösung vernachlässigbar ist, wird jedoch von der Nagumo-Gleichung nicht erfüllt, so dass dies das Problem nicht behebt. Zum Zeitpunkt $t = 1$ genügt ein einziger Schritt von der exakten Lösung ausgehend, um ein Resultat wie in Abbildung 4.8 zu erhalten.

4.2.3. Ergebnisse der Verfahren mit Krylow-Unterraum-Approximation

Eine andere Möglichkeit besteht darin, Krylow-Unterraum-Approximationen zu verwenden. Da diese in jedem Schritt durchgeführt werden müssen, ist die vom Zeitschritt unabhängige Aufspaltung in linearen und nicht-linearen Term möglicherweise unnötig.

Im Gegensatz zum Nørsett–Euler-Verfahren wird beim exponentiell angepassten Euler-Verfahren

$$u_{n+1} = u_n + h\varphi(hf'(u_n))f(u_n)$$

die Jacobi-Matrix der gesamten rechten Seite der Differentialgleichung (4.2) verwendet. Der Abbildung 4.9 ist das Konvergenzverhalten mit Raumschrittweite $\Delta x = \frac{1}{100}$ zu entnehmen.

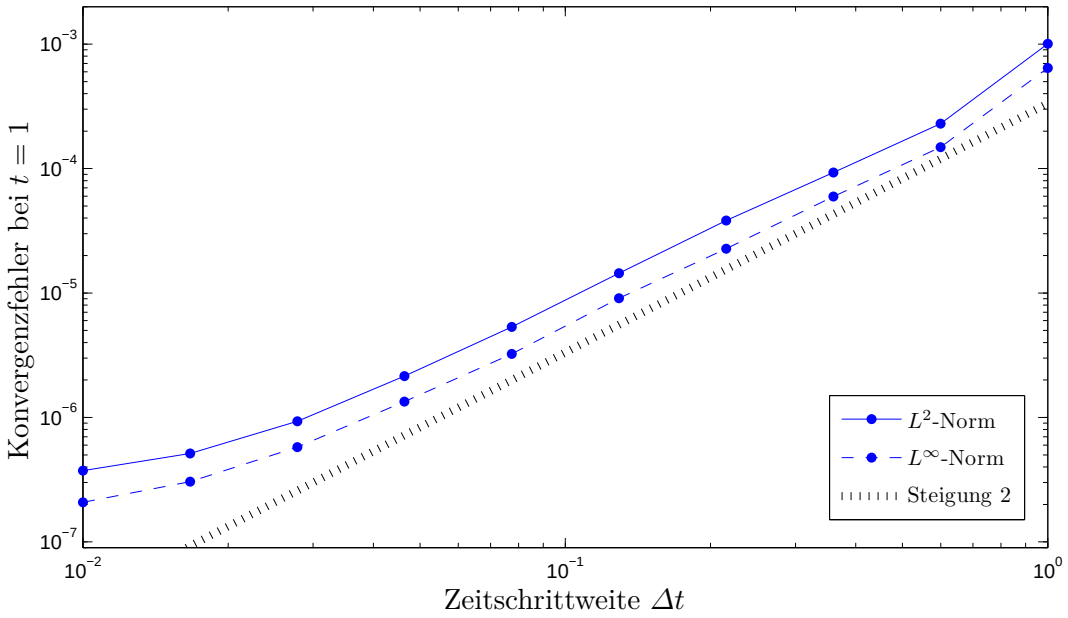


Abbildung 4.9.: Konvergenzverhalten des exponentiell angepassten Euler-Verfahrens

Das Resultat ist Konvergenz der Ordnung 2. Die Krylow-Unterraum-Approximation profitiert dabei davon, dass die Matrix dünn-besetzt ist, so dass Speicherbedarf und Aufwand verhältnismäßig gering sind. Nun bleibt noch zu überprüfen, ob Verfahren höherer Ordnung mit größeren Zeitschrittweiten und dementsprechend geringerem Aufwand bessere Ergebnisse liefern.

Mit den Standardeinstellungen sind die Fehler von exp4 (siehe Kapitel 1) allerdings in einer Größenordnung von 10^{-5}. Verkleinert man die Ortsschrittweite von $\frac{1}{100}$ auf $\frac{1}{200}$, steigt der Fehler sogar. Erst du genaues Anpassen von relativer und absoluter Toleranz sowie der maximalen Größe der Krylow-Unterräume kann man Fehler um 10^{-8} erhalten.

Die wesentliche Schwierigkeit besteht also darin, die Größe der Krylow-Unterräume zu wählen. Entscheidend ist dabei vor allem die Approximation von $\varphi\left(\gamma hf'(v)\right)f(v)$, da sich die übrigen Krylow-Unterräume entweder daraus bilden lassen oder aber nur einen kleinen Anteil am Gesamtfehler haben. Dies wurde in Kapitel 1 thematisiert.

Die Konsequenz daraus ist, dass sich die Aussagen zum exponentiell angepassten Euler-Verfahren auf andere Verfahren übertragen lassen. In Abbildung 4.9 ist die maximale Größe der Krylow-Unterräume $m_{\max} = 60$. Dies ist für die gewählten Raum- und Zeitschrittweiten demnach groß genug.

Für $m_{\max} = 50$ und $m_{\max} = 40$ ergibt sich unter denselben Bedingungen hingegen dieses Bild.

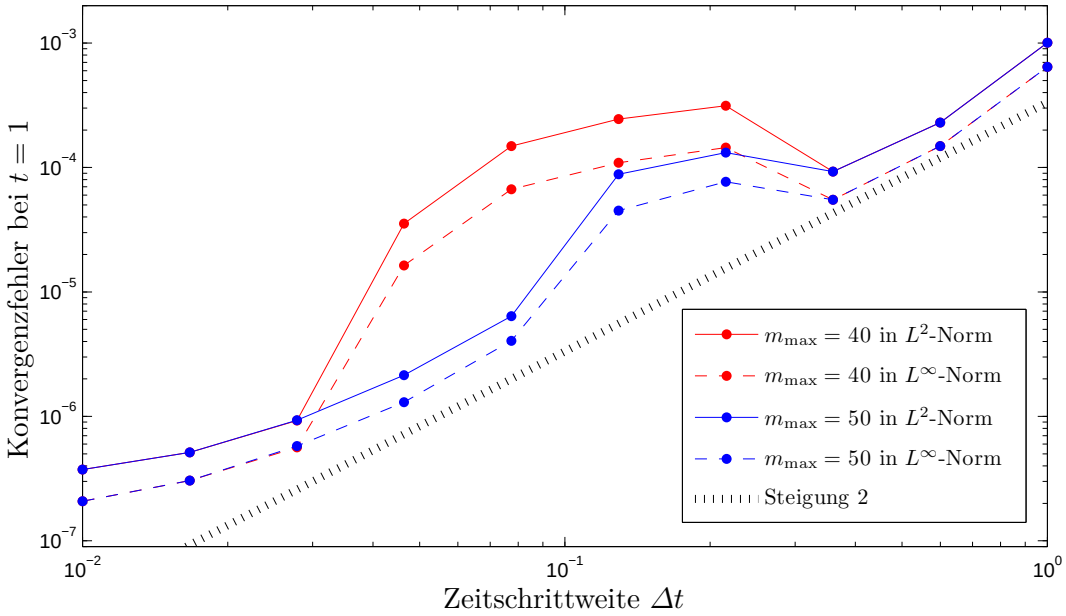


Abbildung 4.10.: Konvergenzverhalten des exponentiell angepassten Euler-Verfahrens mit mittelgroßen Krylowräumen

Für fest gewähltes m erhält man demnach wiederum eine Schrittweitenbedingung. Die entscheidende Frage ist daher, ob sich große Krylow-Räume lohnen oder besser mit einer kleinen Zeitschrittweite gerechnet werden sollte.

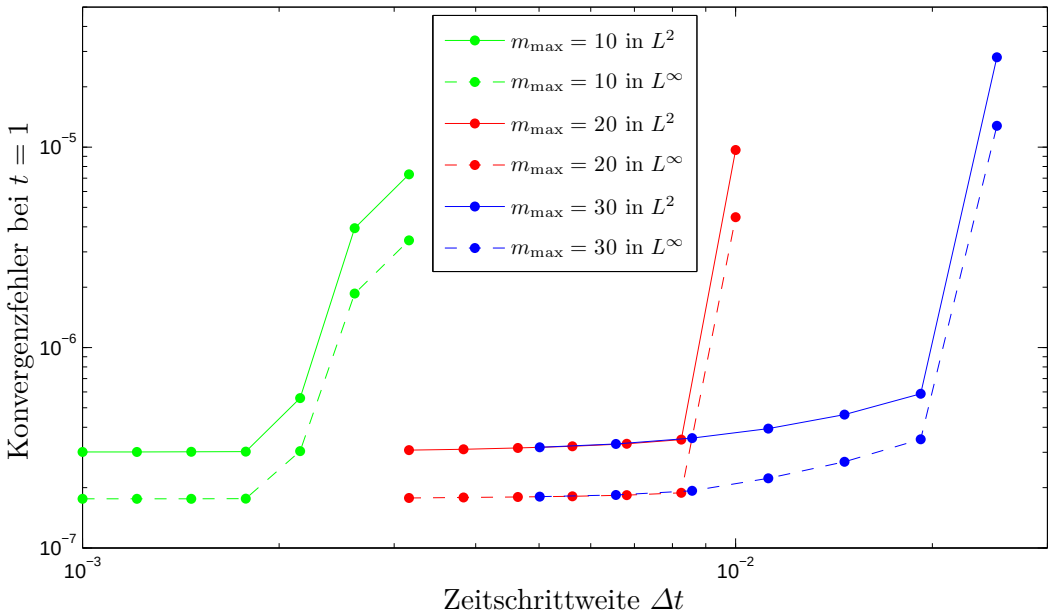


Abbildung 4.11.: Konvergenzverhalten des exponentiell angepassten Euler-Verfahrens mit kleinen Krylowräumen

Für $m_{max} \in \{10, 20, 30\}$ und Raumschrittweite $\Delta x = \frac{1}{100}$ sowie den Zeitschrittweiten aus Abbildung 4.11 ist in Abbildung 4.12 die Laufzeit des exponentiell angepassten Euler-Verfahrens gegen den Konvergenzfehler aufgetragen.

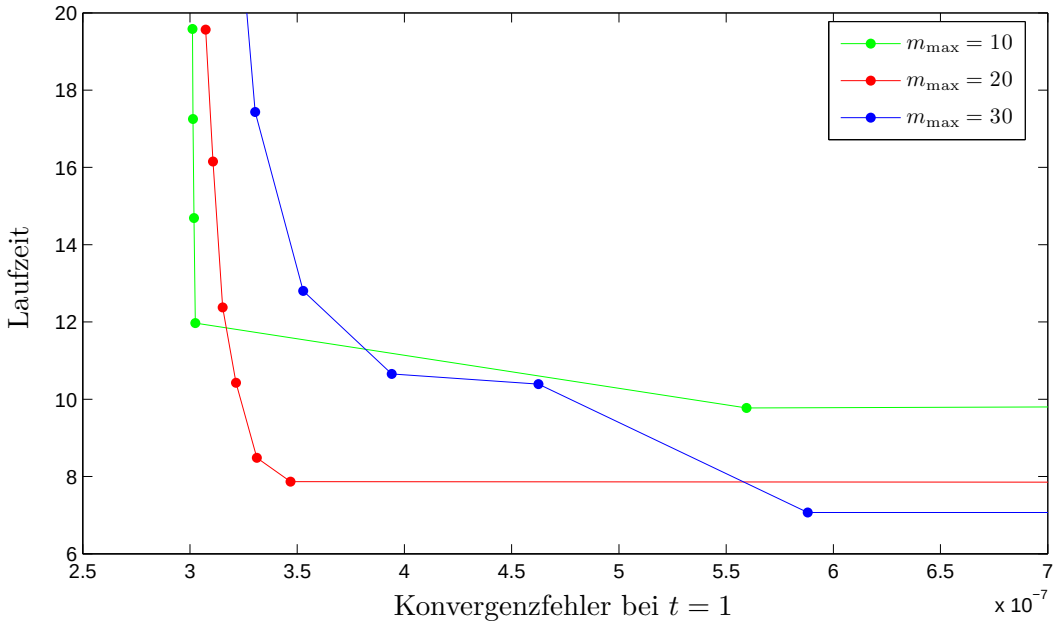


Abbildung 4.12.: Vergleich des Konvergenzfehlers in der L^2-Norm mit der Laufzeit des exponentiell angepassten Euler-Verfahrens

Das Resultat ist nicht besonders eindeutig. Diese Schwierigkeit, die Krylow-Unterraum-Approximationen im richtigen Moment zu stoppen, erklärt schließlich die starke Abhängigkeit der exponentiellen Integratoren von den Toleranzparametern.
Details zur Schrittweitensteuerung und zu den Abbruchkriterien der Krylow-Unterraum-Approximation finden sich in [20, 6.1-6.3]. Der Vollständigkeit halber sei noch bemerkt, dass das Resultat in der L^∞-Norm qualitativ dasselbe ist.

Schlussbemerkung:
Es lässt sich nicht leugnen, dass exponentielle Integratoren bei einer Vielzahl von Problemen im Vergleich zu den Standard-Integratoren konkurrenzfähig sind. Zumindest im Fall parabolischer Gleichungen mit glatten Lösungen sind etwaige Vorteile aber nicht signifikant. Implizite Mehrschritt- und Runge-Kutta-Verfahren sowie die Vorkonditionierung linearer Gleichungssysteme sind allerdings sehr ausgereift, so dass die Hürde für neue Verfahren in jedem Fall hoch ist.

Gute Vorkonditionierer für die Berechnung des Matrixexponentials könnten im Falle hochdimensionaler Probleme den Aufwand exponentieller Integratoren mit Krylow-Unterraum-Approximation deutlich verringern. In diesem Zusammenhang ist auf [42] zu verweisen. Für den Einsatz in der Praxis bedarf es zudem einer robusten Implementierung der exponentiellen Integratoren.

A. Anhang

A.1. Phi-Funktionen

Bei der Analyse der Konsistenz- und Konvergenzbedingungen trat im ersten Kapitel die φ-Funktion auf, die mittels Fallunterscheidung und über eine Potenzreihe definiert wurde. Im dritten Kapitel wurden für $j \in \mathbb{N}_0$ die Halbgruppen $\varphi_j(-tA)$ verwendet. An dieser Stelle wird gezeigt, dass die beiden Herangehensweisen für beschränkte Operatoren übereinstimmen.

Definiere hierzu zunächst rekursiv die Funktionen $\varphi_j : \mathbb{C} \longrightarrow \mathbb{C}$ durch

$$\varphi_0(z) = e^z$$

und für $j \in \mathbb{N}_0$

$$\varphi_{j+1}(z) = \begin{cases} \dfrac{\varphi_j(z) - \dfrac{1}{j!}}{z} & z \neq 0 \\ \dfrac{1}{(j+1)!} & z = 0. \end{cases}$$

Für $j = 1$ entspricht das der Definition (1.4) aus dem ersten Kapitel.

Lemma A.1.1
Die obigen φ_j sind ganze Funktionen, die

$$\varphi_j(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+j)!}$$

erfüllen. Für $j \in \mathbb{N}_1,\ t > 0$ und einen beschränkten Operator A gilt außerdem

$$\varphi_j(-tA) = \frac{1}{t^j} \int_0^t e^{-(t-\tau)A} \frac{\tau^{j-1}}{(j-1)!} d\tau.$$

Beweis:
Zeige zunächst durch Induktion nach j die Identität

$$\varphi_j(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+j)!}.$$

Damit erhält man automatisch die Holomophie auf ganz $\mathbb{C}$.

Induktionsanfang $j = 0$:
Die Gleichheit gilt wegen

$$\varphi_0(z) = e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}.$$

Induktionsschritt $j \to (j+1)$:
Für $z = 0$ ist die zu zeigende Aussage ohnehin erfüllt. Sei also $z \neq 0$. Dann folgt

$$\varphi_{j+1}(z) = \frac{\varphi_j(z) - \frac{1}{j!}}{z} \overset{IV}{=} \frac{1}{z}\left(\sum_{k=0}^{\infty} \frac{z^k}{(k+j)!} - \frac{1}{j!}\right) = \sum_{k=0}^{\infty} \frac{z^k}{(k+j+1)!},$$

wobei der letzte Schritt eine Indexverschiebung beinhaltet. Damit ist die Induktion abgeschlossen.

Die Darstellung der $\varphi_j(-tA)$ kann man nun ebenfalls durch Induktion nach j zeigen.

Behauptung:
Für $j \in \mathbb{N}_1,\ i \in \mathbb{N}_0$

$$\frac{1}{t^j} \int_0^t \frac{(\tau - t)^i}{i!} \frac{\tau^{j-1}}{(j-1)!} d\tau = \frac{(-t)^i}{(i+j)!} \tag{A.1}$$

Induktionsanfang $j = 1$:

$$\frac{1}{t} \int_0^t \frac{(\tau - t)^i}{i!} d\tau = \frac{1}{t}\left[\frac{(\tau - t)^{i+1}}{(i+1)!}\right]_{\tau=0}^{\tau=t} = \frac{(-t)^i}{i!}$$

Induktionsschritt $j \to (j+1)$:

$$\begin{aligned}
\frac{1}{t^{j+1}} \int_0^t \frac{(\tau - t)^i}{i!} \frac{\tau^j}{j!} d\tau &= \frac{1}{t^{j+1}}\left[\frac{(\tau - t)^{i+1}}{(i+1)!} \frac{\tau^j}{j!}\right]_{\tau=0}^{\tau=t} - \frac{1}{t^{j+1}} \int_0^t \frac{(\tau - t)^{i+1}}{(i+1)!} \frac{\tau^{j-1}}{(j-1)!} d\tau \\
&= 0 - \frac{1}{t} \cdot \frac{1}{t^j} \int_0^t \frac{(\tau - t)^{i+1}}{(i+1)!} \frac{\tau^{j-1}}{(j-1)!} d\tau \overset{IV}{=} -\frac{1}{t} \frac{(-t)^{i+1}}{(i+1+j)!} \\
&= \frac{(-t)^i}{(i+j+1)!}
\end{aligned}$$

Die Behauptung ist folglich durch Induktion bewiesen.

Nun aber gilt

$$\frac{1}{t^j} \int_0^t e^{-(t-\tau)A} \frac{\tau^{j-1}}{(j-1)!} d\tau \overset{(A.1)}{=} \sum_{k=0}^{\infty} \frac{(-t)^k}{(k+j)!} A^k = \varphi_j(-tA),$$

was zu zeigen war.

■

A.2. Hilfsmittel aus der Funktionentheorie

Zunächst sei an dieses klassische Resultat der Funktionentheorie erinnert.

Lemma A.2.1 *(Biholomorphiekriterium)*
Sei $U \in \mathbb{C}$ offen, $f : U \longrightarrow \mathbb{C}$ holomorph und injektiv.
Dann gilt $f : U \longrightarrow f(U)$ ist biholomorph, d.h. bijektiv und $f^{-1} : f(U) \to U$ ist holomorph.

Seine Bedeutung besteht im Folgenden darin, dass statt einer Funktion f, die sich möglicherweise nicht ohne Weiteres explizit angegeben lässt, die Umkehrfunktion betrachtet werden kann.

Diese Idee findet bereits im Beweis des nächsten Lemmas Anwendung. Dieses Lemma kann als Spezialfall des Riemannschen Abbildungssatzes angesehen werden, wobei die Abbildung zusätzlich als ein Zweig der Funktion $z+\sqrt{z^2-1}$ charakterisiert wird.

Lemma A.2.2
Es existiert eine holomorphe Abbildung

$$g : \{z^2 - 1 : z \in \mathbb{C} \setminus [-1, 1]\} \longrightarrow \mathbb{C}$$

mit $\left(g(z^2-1)\right)^2 = z^2 - 1$ für alle $z \in \mathbb{C} \setminus [-1, 1]$, so dass die Funktion

$$\begin{aligned} \hat{\phi} : \mathbb{C} \setminus [-1, 1] &\longrightarrow \mathbb{C} \setminus \overline{D_1} \\ z &\longmapsto z + g(z^2 - 1) \end{aligned}$$

biholomorph ist und die Eigenschaft $\hat{\phi}(z) = 2z + \mathcal{O}(1),\ z \to \infty$ besitzt.

Beweis:
Betrachte zunächst die Funktion

$$\begin{aligned} \hat{\psi} : \mathbb{C} \setminus \overline{D_1} &\longrightarrow \mathbb{C} \\ w &\longmapsto \frac{1}{2}\left(w + \frac{1}{w}\right). \end{aligned}$$

Als rationale Funktion, deren Nenner für alle $w \in \mathbb{C} \setminus \overline{D_1}$ ungleich Null ist, ist $\hat{\psi}$ im gesamten Definitionsbereich holomorph. Außerdem ist die Abbildung injektiv, denn es gilt

$$\hat{\psi}(w_1) = \hat{\psi}(w_2) \Rightarrow w_1 + \frac{1}{w_1} = w_2 + \frac{1}{w_2} \Rightarrow w_1 - w_2 = \frac{w_1 - w_2}{w_1 w_2},$$

so dass aus der Gleichheit der Bilder $w_1 = w_2$ oder $w_1 w_2 = 1$ folgt, wobei der zweite Fall auszuschließen ist, da ψ nur auf $\mathbb{C} \setminus \overline{D_1}$ definiert ist und damit $|w_1| > 1$ und $|w_2| > 1$ ist.

Surjektivität liegt allerdings nicht vor, genauer gilt für das Bild

$$\hat{\psi}(\mathbb{C} \setminus \overline{D_1}) = \mathbb{C} \setminus [-1, 1].$$

Dies kann man wie folgt einsehen. Für $z \in \mathbb{C}$ gilt $z \in \hat{\psi}(\mathbb{C} \setminus \overline{D_1})$ genau dann, wenn die Gleichung $w^2 - 2zw + 1 = 0$, die man durch Umformen von $z = \hat{\psi}(w)$ erhält, eine Lösung

w_0 mit $|w_0| > 1$ besitzt.
Ist $z \in [-1,1]$, d.h. z ist reell mit $|z| \leq 1$, so ist dies nicht der Fall, denn

$$\left(z \pm \mathbf{i}\sqrt{1-z^2}\right)^2 = z^2 \pm 2z\mathbf{i}\sqrt{1-z^2} + z^2 - 1 = 2z\left(z \pm \mathbf{i}\sqrt{1-z^2}\right) - 1$$

impliziert, dass die beiden Lösungen durch $w_\pm = z \pm \mathbf{i}\sqrt{1-z^2}$ gegeben sind, wobei allerdings $|w_+| = |w_-| = 1$ gilt. Daher gilt $\hat{\psi}(\mathbb{C} \setminus \overline{D_1}) \subseteq \mathbb{C} \setminus [-1,1]$.

Ist andererseits $w \in \mathbb{C}$ mit $|w| = 1$ gegeben, dann gibt es $\alpha \in [0, 2\pi)$, so dass $w = e^{\mathbf{i}\alpha}$ ist. Dann aber gilt

$$\hat{\psi}(w) = \frac{1}{2}\left(e^{\mathbf{i}\alpha} + e^{-\mathbf{i}\alpha}\right) = \cos(\alpha) \in [-1,1].$$

Dies hat zur Konsequenz, dass es für $z \in \mathbb{C} \setminus [-1,1]$ kein w mit $|w| = 1$ geben kann, so dass $z = \hat{\psi}(w)$ ist. Nach dem Fundamentalsatz der Algebra besitzt die äquivalente Gleichung $w^2 - 2zw + 1 = 0$ aber zwei Lösungen w_1, w_2. Da $\hat{\psi}(w_1) = \hat{\psi}(w_2)$ nun $w_1 = w_2$ oder $w_1 w_2 = 1$ impliziert, muss (genau) eine der beiden Lösungen vom Betrag größer 1 sein, also in $\mathbb{C} \setminus \overline{D_1}$ liegen, was zu zeigen war.

Eine bijektive, holomorphe Funktion besitzt automatisch eine holomorphe Umkehrfuktion. Sei daher $\hat{\phi}$ durch

$$\begin{aligned} \hat{\phi} : \mathbb{C} \setminus [-1,1] &\longrightarrow \mathbb{C} \setminus \overline{D_1} \\ z &\longmapsto \hat{\psi}^{-1}(z) \end{aligned}$$

erklärt.

Die Bedingung $\hat{\psi}(\hat{\phi}(z)) = z$, die sich zu $\left(\hat{\phi}(z)\right)^2 - 2z\hat{\phi}(z) + 1 = 0$ umformen lässt, impliziert mittels binomischer Formel $\left(\hat{\phi}(z) - z\right)^2 = z^2 - 1$. Daher lässt sich $\hat{\phi}$ explizit als $\hat{\phi}(z) = z + g(z^2 - 1)$ schreiben, wobei die holomorphe Funktion g die Eigenschaft $\left(g(z^2-1)\right)^2 = z^2 - 1$ für alle $z \in \mathbb{C} \setminus [-1,1]$ besitzt.

Wegen

$$\lim_{z\to\infty} \frac{\left(g(z^2-1)\right)^2}{z^2} = \lim_{z\to\infty} \frac{z^2-1}{z^2} = 1$$

sind nur ± 1 die möglichen Häufungspunkte von $\lim\limits_{z\to\infty} \frac{g(z^2-1)}{z}$. Allerdings scheidet (-1) aus, da dies für die entsprechende Teilfolge $\hat{\phi}(z) = z + g(z^2-1) \longrightarrow 0$ implizieren würde. Damit gilt $\hat{\phi}(z) = 2z + \mathcal{O}(1),\ z \to \infty$.

■

Literaturverzeichnis

[1] H. Berland, B. Skaflestad, and W. Wright, *EXPINT — A MATLAB package for exponential integrators*, ACM Trans. Math. Softw., 33 (2007), pp. 4–es.

[2] W.-J. Beyn and B. M. Garay, *Estimates of variable stepzise Runge-Kutta methods for sectorial evolution equations with nonsmooth data*, Appl. Numer. Math., 41 (2002), pp. 369–400.

[3] F. F. Bonsall and J.Duncan, *Numerical Ranges of Operators on Normed Spaces and of Elements of Normed Algebras*, Cambridge University Press, Cambridge, UK, 1971.

[4] J. H. Curtiss, *Faber polynomials and the Faber series*, The American Mathematical Monthly, 78 (1971), pp. 577–596.

[5] V. L. Druskin and L. A. Knizhnerman, *Two polynomial methods of calculating functions of symmetric matrices*, U.S.S.R. Comput. Math. and Math. Phys., 29 (1989), pp. 112–121.

[6] ——, *Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl., 2 (1995), pp. 205–217.

[7] M. Eiermann, *On semiiterative methods generated by Faber polynomials*, Numer. Math., 56 (1989), pp. 139–156.

[8] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophysical J., 1 (1961), pp. 445–466.

[9] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 1236–1264.

[10] R. Grigorieff, *Numerik gewöhnlicher Differentialgleichungen*, Teubner, Stuttgart, 1972.

[11] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I*, Springer-Verlag, New York, 1993.

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II*, Springer-Verlag, Berlin, 1996.

[13] M. Hanke-Bourgeois, *Grundlagen der Numerischen Mathamtik und des Wissenschaftlichen Rechnens*, Vieweg/Teubner, 3. ed., 2009.

[14] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer, Berlin, 1981.

[15] M. Hermann, *Numerik gewöhnlicher Differentialgleichungen*, Oldenbourg Wissenschaftsverlag, München, 2004.

[16] J. Hersch, *Contribution à la méthode des équations aux différences*, Z. Angew. Math. Phys., 9 (1958), pp. 129–180.

[17] N. J. Higham, *The scaling and squaring method for the matrix exponential revisited*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1179–1193.

[18] ——, *Functions of matrices: Theory and Computation*, SIAM, Philadelphia, 2008.

[19] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.

[20] M. Hochbruck, C. Lubich, and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comp., 19 (1998), pp. 1552–1574.

[21] M. Hochbruck and A. Ostermann, *Explicit exponential Runge-Kutta methods for semilinear parabolic problems*, SIAM J. Numer. Anal., 43 (2005), pp. 1069–1090.

[22] ——, *Exponential Runge-Kutta methods for parabolic problems*, Appl. Numer. Math., 53 (2005), pp. 323–339.

[23] ——, *Exponential integrators*, Acta Numerica, 19 (2010), pp. 209–286.

[24] M. Hochbruck, A. Ostermann, and J. Schweitzer, *Exponential Rosenbrock-type methods*, SIAM J. Numer. Anal., 47 (2009), pp. 786–803.

[25] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, Journal of Physiology, 117 (1952), pp. 500–544.

[26] A.-K. Kassam and L. N. Trefethen, *Fourth-order time-stepping for stiff PDEs*, SIAM J. Sci. Comput., 26 (2005), pp. 1214–1233.

[27] R. Kippenhahn, *Über den wertevorrat einer matrix*, Mathematische Nachrichten, 6 (1951), pp. 193–228.

[28] L. A. Knizhnerman, *Calculation of functions of unsymmetric matrices using Arnoldi's method*, U.S.S.R. Comput. Math. and Math. Phys., 31 (1991), pp. 1–9.

[29] ——, *Error bounds in Arnoldi's method: The case of a normal matrix*, U.S.S.R. Comput. Math. and Math. Phys., 32 (1992), pp. 1199–1211.

[30] S. Krogstad, *Generalized integrating factor methods for stiff PDEs*, J. Comput. Phys., 203 (2005), pp. 72–88.

[31] T. Kövari and C. Pommerenke, *On faber polynomials and faber expansions*, Math. Z., 99 (1967), pp. 193–206.

[32] A. Meister, *Numerik linearer Gleichungssysteme*, Vieweg, 1999.

[33] C. B. Moler and C. F. van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty five years later*, SIAM Review, 45 (2003), pp. 3–49.

[34] J. D. Murray, *Mathematical Biology I*, Springer-Verlag, New York, 2002.

[35] J. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. IRE, 50 (1962), pp. 2061–2070.

[36] S. P. Nørsett, *An a-stable modification of the adams-bashforth methods.*, Lecture Notes in Mathematics, 109 (1969), p. 214–219.

[37] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[38] ——, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.

[39] M. N. Spijker, *Numerical ranges and stability estimates*, Appl. Numer. Math., 13 (1993), pp. 241–249.

[40] G. Starke and R. S. Varga, *A hybrid Arnoldi-Faber iterative method for nonsymmetric systems of linear equations*, Numer. Math., 64 (1993), pp. 213–240.

[41] J. H. Stetter:, *Analysis of Discretization Methods in Ordinary Differential Equations*, Springer-Verlag, New York, 1973.

[42] J. van den Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comp., 27 (2006), pp. 1438–1457.

Symbolverzeichnis

$\mathcal{O}$	Landau-Symbol
arg	Argument einer komplexen Zahl
$\mathcal{C}^p$	Raum der p-mal stetig differenzierbaren Funktionen
$\frac{d}{dt}\Big[f(u(t))\Big]$	Ableitung der Verkettung $f \circ u$
$\mathcal{D}(A)$	Definitionsbereich von A
$\mathbf{1}_{\mathbb{B}_m}$	Indikatorfunktion, charakteristische Funktion
$\sigma(B)$	Spektrum von B
f'	Fréchet-Ableitung von f
H_0^1	Sobolev-Raum der schwach differenzierbaren Funktionen mit kompaktem Träger
H^2	Sobolev-Raum der zweimal schwach differenzierbaren Funktionen
I	Identitätsoperator
$\mathbf{i}$	Imaginäre Einheit
L^∞	Lebesgue-Raum der wesentlich beschränkten Funktionen
L^2	Lebesgue-Raum der quadratintegrierbar Funktionen
$(\alpha_{ij})_{ij}$	Matrix mit den Einträgen α_{ij}
$\mathbb{N}_1$	Menge der natürlichen Zahlen ohne Null
$\mathbb{N}_0$	Menge der natürlichen Zahlen mit Null
U_t	Partielle Ableitung $\frac{\partial U}{\partial t}$
u_n	Numerische Lösung $u_h(\cdot)$ an der Stelle t_n
$\mathbb{C}$	Menge der komplexen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen