TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Mathematik Institut für Numerik, Professur für Numerik partieller Differentialgleichungen

Optimierung & Numerik — Teil 2

Numerik gewöhnlicher Differentialgleichungen

Prof. Dr. Oliver Sander

Sommersemester 2020

Name : Eric Kunze
E-Mail : eric.kunze@mailbox.tu-dresden.de
Datum : 25. Juni 2020

Das vorliegende Skript basiert nahezu vollständig auf dem Vorlesungsskript von Prof. Dr. Oliver Sander für die Vorlesung „Optimierung und Numerik“ an der Technischen Universität Dresden. Das aktuelle Original ist unter
`https://gitlab.mn.tu-dresden.de/osander/skript-numerik`
zu finden.

Inhaltsverzeichnis

1 Steife Differentialgleichungen und implizite Verfahren

1.1 Wiederholung: Gewöhnliche Differentialgleichungen und Anfangswertprobleme

Das folgende Unterkapitel wiederholt ein paar wichtige Konzepte aus dem vorangegangenen Kapitel. Es existiert hauptsächlich als Vorlage für eine Vorlesung, die mit steifen Differentialgleichungen anfängt, und deshalb etwas Vorwissen wiederholen muss. Beim Lesen dieses Dokuments kann es übersprungen werden.

1.1.1 Gewöhnliche Differentialgleichungen

Gewöhnliche Differentialgleichung:

$$x_i' = f_i(t, x_1, \dots, x_d), \quad i = 1, \dots, d$$

wobei $(t, x) \in \mathbb{R} \times \mathbb{R}^d$ und $f_i \colon \Omega \to \mathbb{R}^d$, $\Omega \subseteq \mathbb{R} \times \mathbb{R}^d$ offen.

- Die Variable t ist häufig als Zeit interpretierbar, man spricht daher häufig von **Evolutionsproblemen**.
- x heißt **Zustandsvektor**.
- $\mathbb{R}^d$ mit $x \in \mathbb{R}^d$ heißt **Zustandsraum**.
- $\mathbb{R} \times \mathbb{R}^d$ heißt **erweiterter Zustandsraum**.

Achtung: Traditionell verwendet man das gleiche Symbol für Zustände $x \in \mathbb{R}^d$ und Funktionen in den Zustandsraum $x : \mathbb{R} \to \mathbb{R}^d$. Nicht verwirren lassen!

Beispiele

1. Lineare skalare Differentialgleichung (Radioaktiver Zerfall): Finde $x : \mathbb{R} \to \mathbb{R}$ so dass $x' = -kx$.

2. Allgemeiner: Finde $x : \mathbb{R} \to \mathbb{R}^d$ so dass

$$x' = -Ax, \qquad A \in \mathbb{R}^{d \times d}.$$

3. Höhere Ableitungen können wegtransformiert werden.

1.1.2 Existenz und Eindeutigkeit

Auch AWPs können mehr als eine Lösung haben.

Beispiel. Betrachte $x' = \sqrt{|x|}$, $x(0) = 0$. Die Funktion $f(x) = \sqrt{|x|}$ ist stetig auf $\mathbb{R}$, es existiert also eine Lösung, z.B.: $x(t) = 0$ für alle t. Eine weitere Lösung ist aber auch $x(t) = \frac{1}{4}t^2$ für $t > 0$. Es kommt noch schlimmer: Für ein $c > 0$, definiere

$$\tilde{x}(t) := \begin{cases} 0 & \text{falls } 0 \leq t \leq c \\ \frac{1}{4}(t-c)^2 & \text{falls } c < t \end{cases}$$

Auch $\tilde{x}$ löst das Problem. Es gibt also unendlich viele Lösungen!

Mit einer einfachen Zusatzforderung an f kann man Eindeutigkeit erhalten.

Definition. Die Abbildung $f \in C(\Omega, \mathbb{R}^d)$ heißt auf Ω bzgl. x lokal **Lipschitz-stetig**, wenn zu jedem $(t_0, x_0) \in \Omega$ ein offener Zylinder

$$Z \colon (t_0 - \tau, t_0 + \tau) \times B_\rho(x_0) \subset \Omega$$

existiert, in dem eine Lipschitzbedingung

$$|f(t,x) - f(t,\overline{x})| \leq L|x - \overline{x}| \qquad \forall (t,x), (t,\overline{x}) \in Z$$

mit Konstante L gilt.

Bemerkung. *Falls $f(t,x)$ nach x ableitbar ist, dann ist es auch bzgl. x lokal Lipschitz-stetig.*

Satz 1.1 (Picard–Lindelöf). Betrachte das Anfangswertproblem

$$x' = f(t,x) \qquad x(t_0) = x_0$$

auf dem erweiterten Zustandsraum $\Omega \subset \mathbb{R} \times \mathbb{R}^d$ mit $(t_0, x_0) \in \Omega$. f sei stetig und bzgl. x lokal Lipschitz-stetig.

Dann besitzt das AWP eine eindeutige Lösung.

1.1.3 Evolution und Phasenfluss

Falls die Bedingungen des Satzes von Picard–Lindelöf gelten, so kann man eine elegante neue Notation einführen.

Sei $(t_0, x_0) \in \Omega$. Bezeichne mit $J_{\max}(t_0, x_0)$ das maximale Zeitintervall, auf dem eine Lösung des dazugehörigen AWPs existiert. Zu jedem Anfangswert (t_0, x_0) gibt es eine eindeutige Lösung, d.h. zu jedem AW (t_0, x_0) ist der Wert $x(t)$ für alle $J_{\max}(t_0, x_0)$ eindeutig bestimmt.

Definition. Für alle $t_0, t \in J_{\max}(t_0, x_0)$ heißt

$$\Phi^{t,t_0} \colon x_0 \mapsto x(t)$$

Evolution der Differentialgleichung $x' = f(t, x)$.

Die Evolution einer Differentialgleichung ist wohldefiniert, weil das AWP für jedes x_0 eine eindeutige Lösung hat. Man kann also schreiben: $x(t) = \Phi^{t,t_0} x_0$.

Für autonome Gleichungen $x' = f(x)$ kann man die Abhängigkeit von t_0 weglassen (wähle immer $t_0 = 0$). Der Evolutionsoperator $\Phi^t x_0 = x(t)$ heißt dann „Phasenfluss“.

1.1.4 Das explizite Euler-Verfahren

Ziel. *Finde eine numerische Approximation der Lösung $x \in C^1([t_0, T], \mathbb{R}^d)$ des AWPs*

$$x' = f(t, x), \qquad x(t_0) = x_0$$

Vorgehensweise.

- Unterteile das Intervall $[t_0, T]$ durch $n + 1$ Zeitpunkte

$$t_0 < t_1 < t_2 < \ldots < t_n = T. \tag{1.1}$$

  Die Menge der Zeitpunkte heißt **Gitter** $\Delta := \{t_0, t_1, \ldots, t_n\}$
- **Schrittweite**: $\tau_j := t_{j+1} - t_j$ für $j = 0, \ldots, n-1$
- Maximale Schrittweite: $\tau_\Delta = \max_{j=0,\ldots,n-1} \tau_j$

Wir suchen eine Gitterfunktion

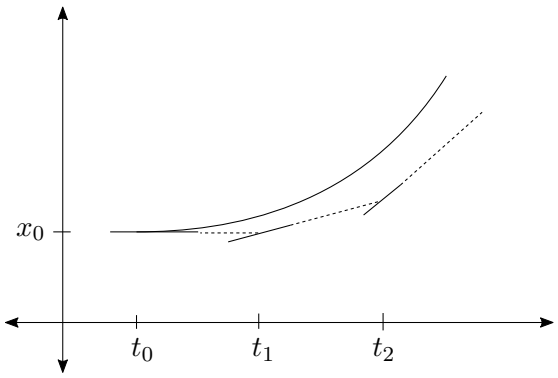
$$x_\Delta \colon \Delta \to \mathbb{R}^d,$$

welche die Lösung des AWPs an den Gitterpunkten möglichst gut approximiert.

Bemerkung. *Manchmal interpretieren wir so ein x_Δ auch als eine Funktion $[t_0, T] \to \mathbb{R}^d$, die die Werte an den Gitterpunkten linear interpoliert.*

Unsere Hoffnung ist natürlich, dass für immer feinere Gitter (wenn also τ_Δ immer kleiner wird) der Unterschied zwischen x und x_Δ immer kleiner wird.

Das explizite Euler-Verfahren nach L. Euler (1786), auch Eulersches Polygonzugverfahren



1. $x_\Delta(t_0) = x_0$

2. Für $t \in [t_j, t_{j+1}]$:
$$x_\Delta(t) = x_\Delta(t_j) + (t - t_j) f(t_j, x_\Delta(t_j))$$

3. Insbesondere:
$$x_\Delta(t_{j+1}) = x_\Delta(t_j) + \tau_j f(t_j, x_\Delta(t_j))$$

Keine Gleichungsysteme zu lösen $\longrightarrow$ das Verfahren ist explizit.

Beachte: Berechnung durch eine Zweiterm-Rekursion

1. $x_\Delta(t_0) = x_0$
2. $x_\Delta(t_{j+1}) = \Psi^{t_{j+1}, t_j} x_\Delta(t_j), j = 0, 1, \ldots, n-1$

mit Ψ unabhängig von Δ.

Die Funktion Ψ heißt **diskrete Evolution** des expliziten Euler-Verfahrens.

1.1.5 Konsistenz

Der Fehler der Lösung, also der Unterschied $x - x_\Delta$, besteht aus zwei Beiträgen:

- Jeder einzelne Schritt produziert einen Fehler.
- Da man nach dem ersten Schritt immer von einem fehlerbehafteten Wert startet, bekommt man auch falsche Ableitungen f.

Mit **Konsistenz** bezeichnet man das *lokale* Verhältnis zwischen der Evolution Φ und der diskreten Evolution Ψ.

Definition. Sei $(t, x) \in \Omega$. Die Differenz
$$\varepsilon(t, x, \tau) = \Phi^{t+\tau, t} x - \Psi^{t+\tau, t} x$$
heißt **Konsistenzfehler** von Ψ.

Eine diskrete Evolution heißt **konsistent**, wenn der Konsistenzfehler keine konstanten oder linearen Terme in τ hat. Formaler definiert man das wie folgt:

Definition. Eine diskrete Evolution Ψ heißt konsistent, falls
$$\begin{aligned} \Psi^{t,t} x &= x && \text{für alle } (t, x) \in \Omega, \\ \frac{d}{d\tau} \Psi^{t+\tau, t} x \Big|_{\tau=0} &= f(x, t) && \text{für alle } (t, x) \in \Omega. \end{aligned}$$

Man kann konsistente Evolutionen einfach charakterisieren. Dazu sei die diskrete Evolution $\Psi^{t+\tau, t} x$ bzgl. τ stetig differenzierbar. Dann sind äquivalent ([DB08, Lemma 4.4]):

- Ψ ist konsistent
- Ψ hat die Darstellung $\Psi^{t+\tau, t} x = x + \tau \psi(t, x, \tau)$. ψ heißt **Inkrementfunktion**. ψ ist stetig in $\tau = 0$, und $\psi(t, x, 0) = f(t, x)$.
- Für den Konsistenzfehler gilt $\varepsilon(t, x, \tau) = o(\tau)$ $(\tau \to 0)$.

Beispiel. Das explizite Euler-Verfahrenm $\Psi^{t+\tau, t} x = x + \tau f(t, x)$ ist konsistent, denn $\frac{\partial \Psi}{\partial \tau} = \frac{\partial}{\partial \tau}(x + \tau f(t, x)) = f(t, x)$.

Wir wollen eine *quantitative* Version von Konsistenz.

Definition. Eine diskrete Evolution Ψ besitzt Konsistenzordnung p, wenn es eine Konstante $C > 0$ unabhängig von t und x gibt, so dass

$$\varepsilon(t, x, \tau) \leq C\tau^{p+1}.$$

Hinweis. *Ja, der Exponent ist wirklich $p + 1$!*

Beispiel. Das Euler-Verfahren hat die Konsistenzordnung 1.

1.1.6 Konvergenz

Konsistenz ist ein lokales Phänomen, d.h. es betrachtet den Fehler nur in der Nähe eines festen t. Wir betrachten jetzt, wie gut eine Lösung $x \in C^1([t_0, T])$ insgesamt approximiert wird und hoffen, dass für kleinere $\tau_\Delta = \max \tau_j$ die Approximation immer besser wird und das schnell!

Definition. Der Vektor der Approximationsfehler auf dem Gitter Δ

$$\varepsilon_\Delta \colon \Delta \to \mathbb{R}^d, \qquad \varepsilon_\Delta(t) = x(t) - x_\Delta(t)$$

heißt **Gitterfehler**. Seine Norm

$$\|\varepsilon_\Delta\|_\infty = \max_{t \in \Delta} |\varepsilon_\Delta(t)|$$

heißt **Diskretisierungsfehler**.

Definition. Zu jedem Gitter Δ auf $[t_0, T]$ sei eine Gitterfunktion x_Δ gegeben. Die Familie dieser Gitterfunktionen konvergiert mit Ordnung $p \in \mathbb{N}$ gegen $x \in C^1([t_0, T])$, falls eine Konstante $C > 0$ existiert, so dass

$$\|\varepsilon_\Delta\|_\infty \leq C\tau_\Delta^p$$

für alle τ_Δ klein genug.

Konvergenz hängt eng mit dem Konsistenzfehler zusammen. Das ist praktisch: Der Konsistenzfehler lässt sich häufig direkt am Verfahren ablesen.

Satz 1.2 ([DB08, Satz 4.10]). Sei ψ lokal Lipschitz-stetig in x. Die diskrete Evolution sei konsistent mit Ordnung p, d.h.

$$\varepsilon(t, x, \tau) := x(t + \tau) - \Psi^{t+\tau,t} x(t) = \mathcal{O}(\tau^{p+1}).$$

Dann definiert die diskrete Evolution Ψ für alle Gitter Δ mit hinreichend kleiner Zeitschrittweite τ_Δ eine Gitterfunktion x_Δ zum Anfangswert $x_\Delta(t_0) = x_0$. Die Familie dieser Gitterfunktionen konvergiert mit Ordnung p gegen die Lösung x des AWPs, d.h.

$$\|\varepsilon_\Delta\|_\infty := \max_{t \in \Delta} |x(t) - x_\Delta(t)| = \mathcal{O}(\tau_\Delta^p)$$

Beispiel (Explizites Euler-Verfahren). $x_{j+1} = x_j + \tau_j f(t_j, x_j)$

Das Verfahren hat Konsistenzordnung 1 (der lokale Fehler verhält sich wie τ_Δ). Die Inkrementfunktion $\psi(t,x,\tau) = f(t,x)$ ist lokal Lipschitz-stetig in x.

$\Rightarrow$ Verfahren konvergiert mit Ordnung 1, d.h. $\|\varepsilon_\Delta\|_\infty = \mathcal{O}(\tau_\Delta^1)$.

$\Rightarrow$ halber Fehler bedeutet doppelte Anzahl von Zeitschritten.

$\Rightarrow$ doppelte Anzahl von f-Auswertungen $\implies$ doppelter Aufwand

1.1.7 Explizite Runge–Kutta-Verfahren

Klassische Konstruktion von Verfahren mit einer höheren Konsistenzordnung.

Insbesondere: Hohe Ordnung ohne Ableitungen von f

Allgemein hat man ein s-stufiges explizites Runge-Kutta-Verfahren:

1. $k_i := f\left(t + c_i\tau, x + \tau \sum_{j=1}^{i-1} a_{ij} k_j\right) \quad \forall i = 1, \ldots, s$
2. $\Psi^{t+\tau,t} x = x + \tau \sum_{i=1}^{s} b_i k_i.$

Die Größen $k_i = k_i(t,x,\tau)$ heißen Stufen des Verfahrens.

Koeffizienten:

$$A = \begin{pmatrix} 0 & & & & \\ a_{21} & 0 & & & \\ \vdots & \ddots & & & \\ a_{s1} & a_{s2} & \ldots & a_{s,s-1} & 0 \end{pmatrix} \in \mathbb{R}^{s\times s} b = (b_1, \ldots, b_s)$$
$$c = (c_1, \ldots, c_s)$$

Traditionell notiert man die Koeffizienten im Butcher-Schema:

$$\begin{array}{c|c} c^T & A \\ \hline & b \end{array}$$

Beispiel. ▪ explizites Euler-Verfahren $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$

▪ Verfahren von Runge $\begin{array}{c|cc} 0 & 0 & \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$

▪ Das klassische 4-stufige Runge-Kutta-Verfahren:

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

- Eine Funktionsauswertung pro Stufe
- $s + s + \frac{(s-1)s}{2}$ Parameter bei s Stufen

Bei geschickter Wahl der Koeffizienten erhält man Verfahren höherer Ordnung.

Lemma 1.1. Ein explizites Runge–Kutta-Verfahren ist genau dann konsistent für alle $f \in C(\Omega, \mathbb{R}^d)$, wenn $\sum_{i=1}^{s} b_i = 1$.

Außerdem betrachten wir nur folgende Vereinfachung:

Lemma 1.2. Ein explizites Runge–Kutta-Verfahren ist genau dann invariant unter Autonomisierung, wenn es konsistent ist, und

$$c_i = \sum_{j=1}^{i-1} a_{ij} \qquad \text{für } i = 1, \dots, s$$

erfüllt.

1.2 Steife Differentialgleichungen

Wir betrachten das explizites Euler-Verfahren

$$x_{k+1} = x_k + \tau f\left(t_k, x_k\right)$$

Das Verfahren ist konvergent mit Ordnung 1.

Löse damit das Anfangswertproblem

$$x' = \lambda x \qquad x(0) = 1 \qquad (\lambda \in \mathbb{R})$$

Die folgenden Bilder zeigen Rechnung für $\lambda = 1$, mit verschiedenen Schrittweiten[1]:

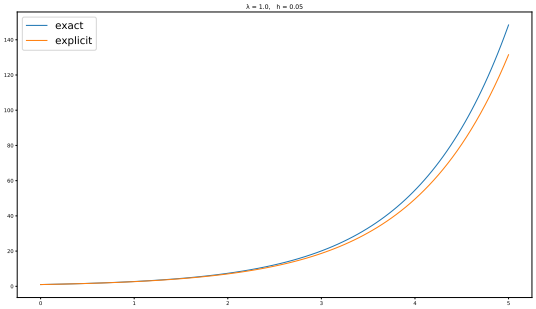


Abbildung 1.1: $\lambda = 1$, $h = 0{,}05$

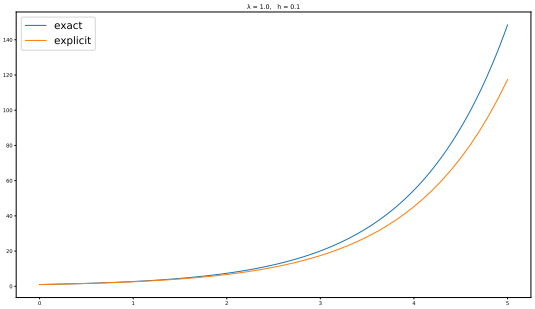


Abbildung 1.2: $\lambda = 1$, $h = 0{,}1$

[1] Die Tatsache dass die orange Linien nicht bis zum Ende des Zeitintervalls gehen sind Zeichen schlechter Programmierung, aber kein mathematisches Problem.

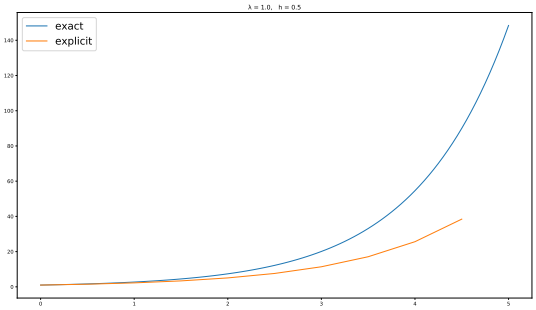


Abbildung 1.3: $\lambda = 1$, $h = 0{,}5$

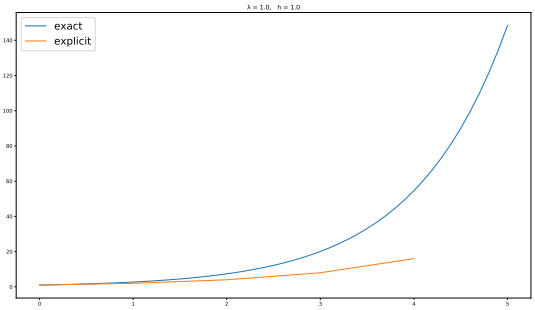
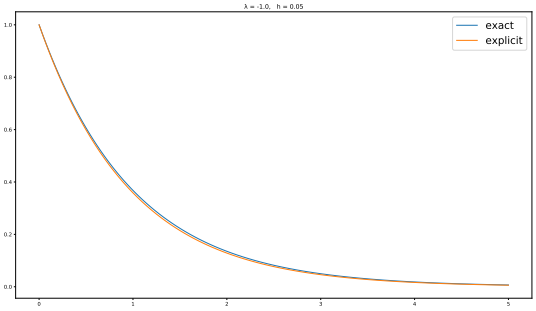
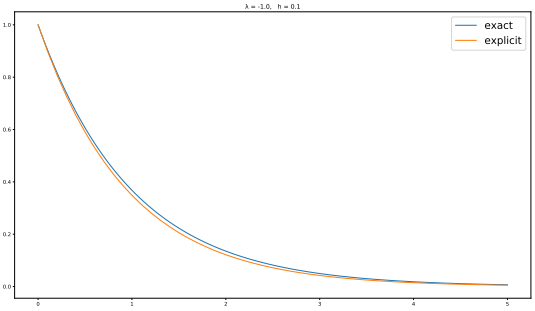


Abbildung 1.4: $\lambda = 1$, $h = 1{,}0$

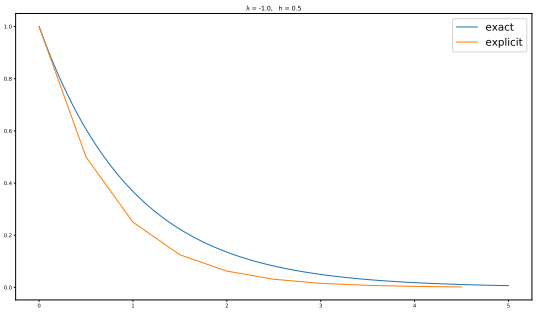
Als nächstes probieren wir ein negatives λ, hier z.B. $\lambda = -1$:



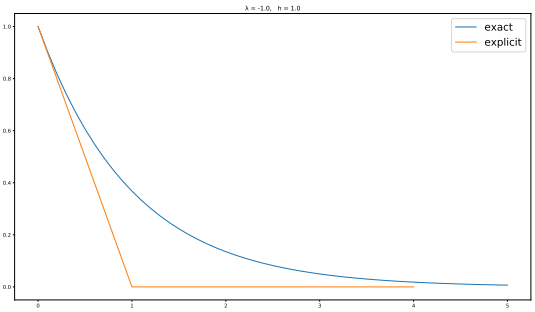
$\lambda = -1$, $h = 0{,}05$



$\lambda = -1$, $h = 0{,}1$



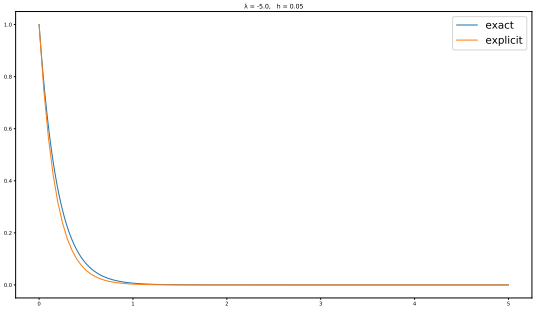
$\lambda = -1$, $h = 0{,}5$



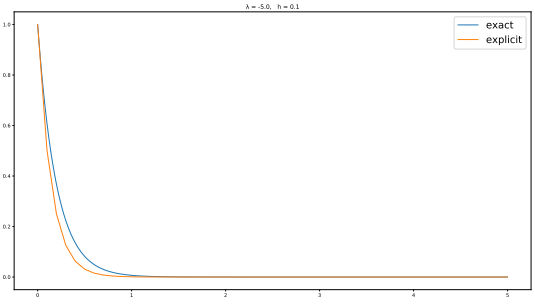
$\lambda = -1$, $h = 1{,}0$

Die letzte Rechnung sieht ein wenig seltsam aus, aber der Rest ist okay.

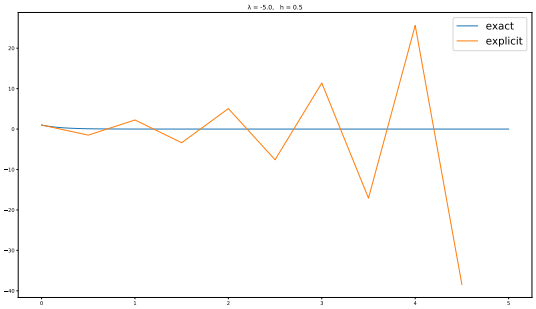
Probieren wir noch $\lambda = -5$:



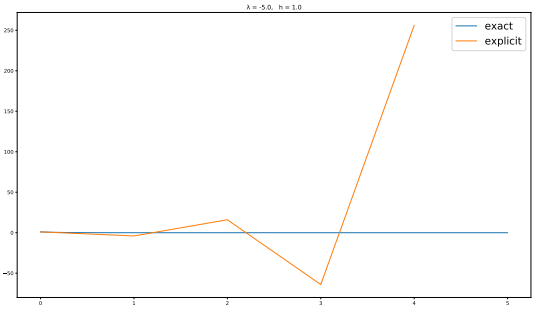
$\lambda = -5$, $h = 0{,}05$



$\lambda = -5$, $h = 0{,}1$



$\lambda = -5,\ h = 0{,}5$



$\lambda = -5,\ h = 1{,}0$

Fazit:

- Keine Überraschungen, falls $\lambda \geq 0$
- Falls $\lambda < 0$, dann produziert das explizite Euler-Verfahren nur dann qualitativ richtige Ergebnisse, falls der Zeitschritt $\tau < 1/|\lambda|$ ist.
- Für $\lambda < 0$, $\tau > 1/|\lambda|$ ist das Verfahren instabil.

Rechnen wir diese Beobachtungen nach: Für den expliziten Euler gilt

$$x_{k+1} = x_k + \tau f\left(t_k, x_k\right) = x_k + \tau\lambda x_k = \left(1 + \lambda\tau\right) x_k = \left(1 + \lambda\tau\right)^{k+1} x_0$$

Fall 1: $\lambda > 0$:

- Lösung $x(t) = \exp\left(\lambda t\right)$ monoton steigend in t
- Diskrete Lösung monoton steigend in k, da $1 + \lambda\tau > 1$

Fall 2: $\lambda < 0$

- Lösung $x(t) = \exp\left(\lambda t\right)$ monoton fallend und positiv
- x_k monoton fallend und positiv nur dann, wenn $0 < 1 + \lambda\tau < 1 \iff \tau \leq 1/|\lambda|$

Falls $\lambda < 0$ und $\tau > \frac{1}{|\lambda|}$

- Diskrete Lösung oszilliert.

Falls $\lambda < 0$ und sogar $\tau > \frac{2}{|\lambda|}$

- Diskrete Lösung ist unbeschränkt.

1.2.1 Steifheit und Kondition

Für Euler- und Runge-Kutta-Verfahren hatten wir Konvergenzaussagen der Form

$$\|\varepsilon_\Delta\|_\infty \leq C\tau_\Delta^p$$

bewiesen.

Problem: Diese Aussagen sind **asymptotisch**.

- Der Fehler wird kleiner, wenn wir ein hinreichend kleines τ_Δ weiter verkleinern.
- Die Konstante C ist aber unbekannt: Wir wissen nicht, **wie** klein τ_Δ sein muss, um eine vernünftige Genauigkeit zu erzielen.
- Man kann C nur in seltenen Fällen exakt ausrechnen.

Wir wollen stattdessen **qualitativ** verstehen, wann C groß sein kann.

Angenommen, wir erhalten für ein gegebenes τ_Δ eine gute Approximation der Lösung. Dann können wir davon ausgehen, dass das nicht zufällig so ist: Für einen leicht gestörten Anfangswert $x_0 + \delta x_0$ erwarten wir dann auch eine gute Approximation der gestörten Lösung.

Erinnerung: Intervallweise Kondition eines AWPs:

$$x' = f(t, x) \qquad x(t_0) = x_0 \qquad t \in [t_0, T].$$

Eine Störung der Eingabedaten $x_0 \mapsto x_0 + \delta x_0$ führt zu einer Störung der Lösung $x(t) \mapsto x(t) + \delta x(t)$ für alle $t \in [t_0, T]$.

Definition. Die **intervallweise Kondition** $\kappa[t_0, T]$ ist die kleinste Zahl, für die

$$\|\delta x\|_\infty \leq \kappa[t_0, T] \cdot \|\delta x_0\|.$$

Analog führen wir eine **diskrete Kondition** κ_Δ ein: die Auswirkung einer Störung des Anfangswerts auf eine von einem numerischen Verfahren erzeugte Gitterfunktion

$$\|\delta x_\Delta\|_\infty \leq \kappa_\Delta \cdot \|\delta x_0\|.$$

Wenn ein Verfahren für x_0 und $x_0 + \delta x_0$ (mit δx_0 klein) vernünftige Lösungen liefert, dann muss

$$\kappa_\Delta \approx \kappa[t_0, T]$$

gelten. Umgekehrt bedeutet $\kappa_\Delta \gg \kappa[t_0, T]$, dass das Verfahren völlig unbrauchbar ist, denn es reagiert auf kleine Störungen völlig anders als das eigentliche Problem. Das Gitter ist dann noch zu grob, da für jedes konvergente Verfahren

$$\kappa_\Delta \to \kappa[t_0, T] \qquad \text{für } \tau_\Delta \to 0.$$

gilt. Die Beziehung

$$\kappa_\Delta \approx \kappa[t_0, T]$$

ist eine qualitative Minimalforderung an ein Verfahren und die Wahl des Zeitschritts.

Definition (Steifheit). Für die bisher vorgestellten Verfahren gibt es Anfangswertprobleme, für die $\kappa_\Delta \approx \kappa[t_0, T]$ erst für sehr kleine τ_Δ gilt. Solche Probleme nennt man **steif**.

Ungewöhnlich: Es gibt keine mathematisch präzise Definition des Begriffs „steif“. Eine Verfahrensklasse klassifiziert die Probleme!

1.2.2 Beispiel: Das Modellproblem mit explizitem Euler

Wir betrachten wieder das Modellproblem

$$x' = \lambda x, \qquad x(0) = 1.$$

Wie ist die Kondition dieses AWPs?

Die Lösung des AWPs ist gegeben durch

$$x(t) = 1 \cdot e^{\lambda t}.$$

Betrachte jetzt stattdessen einen gestörten Startwert: $x(0) = 1 + d$. Damit erhalten wir die Lösung

$$x(t) = (1 + d)\exp(\lambda t) = \exp(\lambda t) + d\exp(\lambda t),$$

also ist $\delta x = d\exp(\lambda t)$ die Störung des Resultats. Es folgt

- $\kappa[0, T] = e^{\lambda T}$ falls $\lambda \geq 0$,
- $\kappa[0, T] = 1$ falls $\lambda \leq 0$.

Die diskrete Kondition des expliziten Euler-Verfahrens

$$x_\Delta(t_{k+1}) = (1 + \tau\lambda)x_\Delta(t_k) = (1 + \tau\lambda)^{k+1}x_0$$

ist linear in x_0. Deshalb gilt

$$\kappa_\Delta = \max_{0\leq k\leq n-1} |1 + \tau\lambda|^{k+1}$$

Fall 1: $\lambda \geq 0$: Dann ist $\kappa_\Delta = (1 + \tau\lambda)^n$. Wegen $1 + \tau\lambda \leq e^{\tau\lambda}$ gilt

$$\kappa_\Delta = (1 + \tau\lambda)^n \leq \exp(n\tau\lambda) = e^{\lambda T}$$

Also ist $\kappa_\Delta \approx \kappa[0, T]$, das AWP ist nichtsteif.

Fall 2: $\lambda < 0$:

$$\kappa_\Delta = \max_{0\leq k\leq n-1} |1 - \tau_\Delta|\lambda||^{k+1}$$

Falls $\tau < \frac{2}{|\lambda|}$ so ist $\kappa_\Delta \leq 1 = \kappa[0, T]$. Andererseits gilt für $\tau_\Delta \gg 2/|\lambda|$

$$\kappa_\Delta = |1 - \tau|\lambda||^n \gg 1 = \kappa[0, T].$$

Das Problem ist steif.

1.2.3 Stabilität

Wir betrachten noch einmal das vorige gestörte AWP

$$x' = \lambda x, \qquad x(0) = 1 + d$$

Wie verhält sich die Störung $d\exp(\lambda t)$? Dazu betrachten wir die drei Fälle:

Fall 1: $\lambda > 0$: Die Störung wächst exponentiell mit t. Das Lösen der Gleichung für große t ist kaum sinnvoll, bzw. sehr schwierig.

Fall 2: $\lambda = 0$: Die Störung bleibt für alle t in konstanter Größe erhalten.

Fall 3: $\lambda < 0$: Für große t wird die Störung „von alleine“ immer kleiner!

Betrachtet man die Auswirkung von Störungen nicht auf ein beschränktes Intervall $[t_0, T]$, sondern für alle Zeiten $[t_0, \infty)$, dann spricht man statt von Kondition meistens von **Stabilität**.

Die obige Dreiteilung ist typisch. Wir machen daraus eine Definition.

Definition. Sei (t_0, x_0) so, dass $\Phi^{t,t_0}x_0$ für alle $t \geq t_0$ existiert. Die Lösung des AWPs heißt

(1) **instabil**, falls weder (2) noch (3) gelten.

(2) **(Lyapunov)-stabil**, falls zu jedem $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass

$$\|\Phi^{t,t_0}x - \Phi^{t,t_0}x_0\| \leq \varepsilon$$

für alle $t \geq t_0$ und $\|x - x_0\| \leq \delta$,

(3) **asymptotisch stabil**, falls es zusätzlich ein $\delta_0 > 0$ gibt, so dass

$$\lim_{t\to\infty} \|\Phi^{t,t_0}x - \Phi^{t,t_0}x_0\| = 0$$

falls $\|x - x_0\| \leq \delta_0$,

Bemerkung. *Dieser Stabilitätsbegriff hat nichts mit der Stabilität von Algorithmen zu tun. Es kann anspruchsvoll bis zu schwierig sein, die Stabilität von DGL zu bestimmen.*

1.2.4 Das implizite Euler-Verfahren

Expliziter Euler:

$$x_{k+1} = x_k + \tau f\left(t_k, x_k\right)$$

Impliziter Euler:

$$x_{k+1} = x_k + \tau f\left(t_{k+1}, x_{k+1}\right)$$

Implizit bedeutet, dass in jedem Schritt ein Gleichungssystem gelöst werden muss.

Betrachten wir wieder das AWP

$$x' = \lambda x \qquad x(0) = 1 \qquad (\lambda \in \mathbb{R})$$

Implizites Euler-Verfahren:

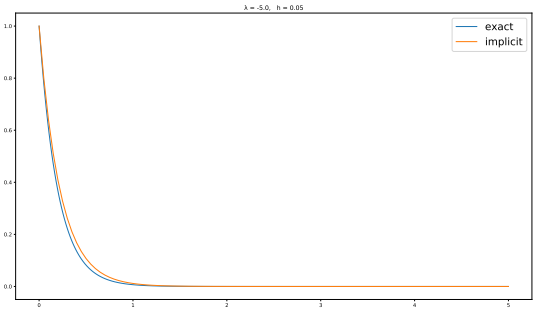
$$x_{k+1} = x_k + \tau f\left(t_{k+1}, x_{k+1}\right) = x_k + \tau\lambda x_{k+1}$$
$$\implies x_{k+1} = \frac{x_k}{1 - \tau\lambda} = \left(\frac{1}{1 - \tau\lambda}\right)^{k+1} x_0$$

Wenn $\lambda < 0$, so ist

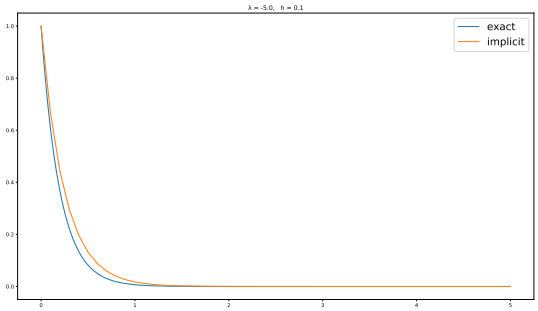
$$0 < \frac{1}{1 - \tau\lambda} < 1$$

für alle $\tau > 0$. Das Verfahren ist für alle $\tau > 0$ stabil.

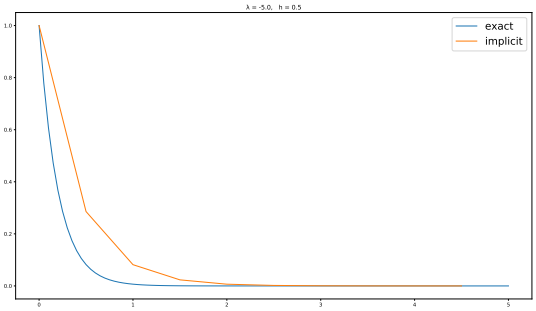
Das probieren wir wieder numerisch aus. Hier ist das implizite Euler-Verfahren für $x' = \lambda x$ mit $\lambda = -5$:



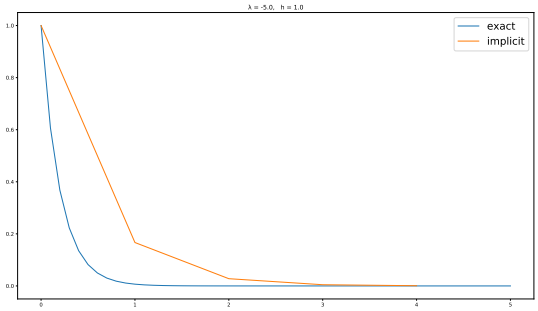
$\lambda = -5,\ h = 0{,}05$



$\lambda = -5,\ h = 0{,}1$



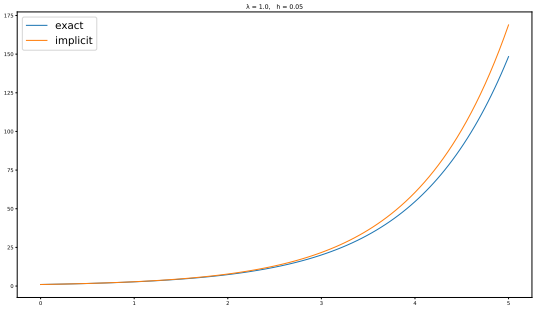
$\lambda = -5,\ h = 0{,}5$



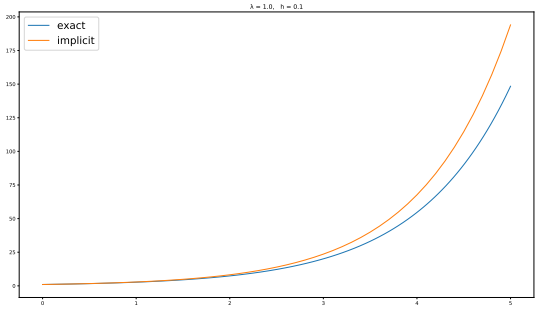
$\lambda = -5,\ h = 1{,}0$

Die letzte Rechnung (mit $h = 1{,}0$) ist zwar nicht mehr sonderlich präzise, aber stabil ist sie.

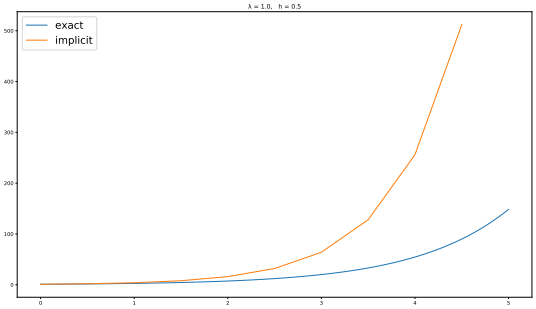
Ist jetzt alles gut? Nein, denn es steht zu vermuten dass wir für positive λ Probleme kriegen. Und in der Tat, für $\lambda = 1$ sieht das Ergebnis nicht mehr so rosig aus:



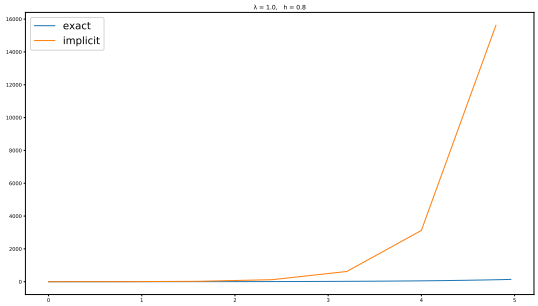
$\lambda = 1,\ h = 0{,}05$



$\lambda = 1,\ h = 0{,}1$



$\lambda = 1,\ h = 0{,}5$



$\lambda = 1,\ h = 0{,}8$

Man beachte hier die wechselnde Skalierung der vertikalen Achse. Die Zeitschrittweite im letzten Bild ist 0,8 statt wie bisher 1,0, weil der Wert 1,0 im Verfahren zu einer Division durch Null führt.

1.3 Stabilität von Einschrittverfahren

Das explizite Euler-Verfahren wird für die lineare Gleichung

$$x' = \lambda x, \qquad x(t_0) = x_0$$

instabil, wenn $\lambda < 0$ und der Zeitschritt τ zu groß ist.

1.3.1 Stabilität von linearen, autonomen, homogenen Differentialgleichungen

Wir verallgemeinern das jetzt und betrachten *lineare*, autonome, homogene Systeme

$$x' = Ax \qquad x(0) = x_0 \in \mathbb{R}^d \qquad (A \in \mathbb{R}^{d\times d})$$

Satz 1.3. Die Lösung dieses AWPs ist

$$x(t) = \exp(tA)x_0$$

wobei

$$\exp(tA) := \sum_{k=0}^{\infty} \frac{(tA)^k}{k!}$$

Diese Reihe konvergiert gleichmäßig auf jedem kompakten Zeitintervall.

Stabilität heißt Störungen im Startwert führen auf beschränkte Störungen in der Lösung (für $t \to \infty$).

Für lineare Gleichungen $x' = Ax$ mit $x(0) = x_0 + \delta_0$ ist die Lösung

$$x_\delta(t) = \exp(tA)\,(x_0 + \delta_0) = \exp(tA)x_0 + \exp(tA)\delta_0$$

d.h. die Störung löst das AWP $x' = Ax, \qquad x(0) = \delta_0$.

Lemma 1.3 ([DB08, Lemma 3.20]). Die Lösung x eines *linearen*, homogenen AWPs ist genau dann stabil, wenn

$$\sup_{t\geq 0}\|x(t)\| < \infty.$$

Sie ist asymptotisch stabil, falls $\|x(t)\| \xrightarrow{t\to\infty} 0$.

Beispiel. $x' = \lambda x$ ist stabil für $\lambda \leq 0$, und asymptotisch stabil für $\lambda < 0$.

Satz 1.4 ([DB08, Satz 3.23]). Die Lösung des AWPs

$$x' = Ax \qquad x(0) = x_0 \qquad (A \in \mathbb{C}^{d\times d})$$

ist genau dann stabil, wenn

- der Realteil aller Eigenwerte nicht positiv ist und
- falls λ ein Eigenwert von A mit $\mathrm{Re}(\lambda) = 0$, so hat λ die gleiche algebraische und geometrische Vielfachheit.

Die Lösung ist asymptotisch stabil, falls $\mathrm{Re}(\lambda) < 0$ für alle Eigenwerte λ von A.

1.3.2 Stabilität von linearen autonomen Rekursionen

Wunsch: Die von einem numerischen Verfahren erzeugte Folge x_k soll diese Stabilitätseigenschaften *erben*. Beim expliziten Euler-Verfahren

$$x_{k+1} = \Psi^\tau x_k = x_k + \tau A x_k = (I + \tau A)\, x_k$$

war das nicht der Fall. Beim impliziten Euler-Verfahren

$$x_{k+1} = x_k + \tau A x_{k+1} \implies x_{k+1} = (I - \tau A)^{-1} x_k$$

jedoch schon!

In Verallgemeinerung betrachten wir nun Verfahren der Form

$$x_{k+1} = \Psi^\tau x_k = R(\tau A)\, x_k$$

mit Polynomen P und Q, sodass

$$R(\tau A) = \frac{P(\tau A)}{Q(\tau A)}$$

x_{k+1} berechnet sich als Lösung des *linearen* Gleichungssystems

$$Q(\tau A)x_{k+1} = Q(\tau A)(\Psi^\tau x_k) = P(\tau A)x_k$$

Die rationalen Funktionen werden als Approximationen der Evolution $\Phi^\tau = \exp(\tau A)$ verwendet.

Damit $R(A)$ wohldefiniert ist muss $Q(A)$ invertierbar sein – $Q(A)$ darf also nicht den Eigenwert Null haben. Wir betrachten nun eine Verallgemeinerung der entsprechenden Bedingung für rationale Funktionen in $\mathbb{C}$:

Lemma 1.4. Eine rationale Funktion $r\colon z \mapsto \frac{p(z)}{q(z)}$ ist genau dort nicht definiert (bzw. hat genau dort Polstellen), wo $q(z) = 0$ ist.

Satz 1.5 ([DB08, Satz 3.42]). Für eine Matrix $A \in \mathbb{C}^{d\times d}$ ist $R(A)$ genau dann definiert, wenn kein Eigenwert von A Pol von R ist.

Satz 1.6 ([DB08, Satz 3.33]). Die lineare Iteration $x_{k+1} = Bx_k$ mit $B \in \mathbb{C}^{d\times d}$ ist genau dann *stabil*, wenn

- $|\lambda| \leq 1$ für alle Eigenwerte λ von B und
- Falls λ Eigenwert von B mit $|\lambda| = 1$, so hat λ gleiche algebraische und geometrische Vielfachheit.

Die Iteration ist *asymptotisch stabil*, falls $|\lambda| < 1$ für alle Eigenwerte λ von B.

1.3.3 Stabilitätsfunktionen

Wir müssen also die Eigenwerte von

$$B = R(\tau A)$$

betrachten. Kann man sie als Funktion von τ und den Eigenwerte von A berechnen?

Kurioserweise gilt:

Satz 1.7 ([DB08, Satz 3.42, Forts.]). Sei $\sigma(A)$ das Spektrum von A. Dann ist

$$\sigma(R(A)) = R(\sigma(A))$$

Mit anderen Worten: λ ist ein Eigenwert von A genau dann wenn $R(\lambda)$ ein Eigenwert von $R(A)$ ist.

Dabei ist jetzt $R(\lambda)$ die formale rationale Funktion R angewandt auf die komplexe Zahl λ.

Definition. Für ein gegebenes Einschrittverfahren heißt die dazugehörige Funktion $R\colon \mathbb{C} \to \mathbb{C}$ **Stabilitätsfunktion** des Verfahrens.

Vererbung von Stabilität heißt damit: Wenn

$$\operatorname{Re}(\lambda) \leq 0 \qquad \text{für alle Eigenwerte } \lambda \text{ von } A$$

dann soll auch

$$|R(\tau\lambda)| \leq 1 \qquad \text{für alle Eigenwerte } \lambda \text{ von } A$$

(plus Zusatzbedingungen für den Fall $\operatorname{Re}\lambda = 0$) gelten.

Definition. Die Menge

$$S = \{z \in \mathbb{C} \mid |R(z)| \leq 1\}$$

heißt **Stabilitätsgebiet** von R.

Beispiel. Explizites Euler-Verfahren:

$$x_{k+1} = \underbrace{(I + \tau A)}_{R(\tau A)} x_k \Rightarrow R(z) = 1 + z$$

Stabilitätsgebiet: $S = \{z \in \mathbb{C}\colon |1 + z| \leq 1\}$.

Damit ein Verfahren stabil ist muss $\tau\lambda \in S$ für alle $\lambda \in \sigma(A)$ sein.

Im Falle der skalaren Gleichung mit $\lambda \in \mathbb{R}_{<0}$ und dem Euler-Verfahren führt das auf die Bedingung $\tau \leq \frac{2}{|\lambda|}$.

Für eine graphische Darstellung der Stabilitätsgebiete expliziter Runge-Kutta-Verfahren siehe [DB08, Seite 238]. Folgendes Detail sticht ins Auge.

Lemma 1.5 ([DB08, Lemma 6.5]). Für jede konsistente rationale Approximation von exp gilt $0 \in \partial S$.

Also:

- Die Lösung $x(t) = \exp(tA)x_0$ ist stabil, wenn alle Eigenwerte von A in der linken Halbebene von $\mathbb{C}$ liegen.
- Ein numerisches Verfahren $\Psi^\tau = R(\tau A)$ ist stabil, wenn alle Eigenwerte von τA in S liegen (plus Zusatzbedingungen am Rand von S).

Folgende Eigenschaft ist deshalb wünschenswert:

Definition. Ein Einschrittverfahren heißt **A-stabil**, falls sein Stabilitätsgebiet die negative komplexe Halbebene enthält.

In diesem Fall gibt es keine Schrittweitenbeschränkung!

Explizite Verfahren können aber nicht A-stabil sein.

Lemma 1.6. Die Flüsse aller expliziten Runge-Kutta-Verfahren (für lineare Gleichungen) sind Polynome in τA, also $\Psi^\tau x = P(\tau A)x$.

Beweis. Wir zeigen dass für jedes $i \leq s$ der Ausdruck τk_i ein formales Polynom in τA ist. Dann folgt die Behauptung aus

$$\Psi^{t+\tau,t}x = x + \tau \sum_{i=1}^{s} b_i k_i = (\tau A)^0 x + \sum_{i=1}^{s} b_i \tau k_i.$$

Beweis mit vollständiger Induktion: Betrachte ein RK-Verfahren für $x' = Ax$:

$$k_i = f(t + c_i\tau, x + \tau \sum_{j=1}^{i-1} a_{ij}k_j) = A\Big[x + \tau \sum_{j=1}^{i-1} a_{ij}k_j\Big]$$

- $\tau k_1 = \tau Ax$ ist Polynom in τA.
- Seien τk_j Polynome für alle $j < 1$. Dann ist

$$\tau k_i = \tau Ax + \tau A \sum_{j+1}^{i-1} a_{ij}\tau k_j$$

ein Polynom in τA. □

Lemma 1.7 ([DB08, Lemma 6.11]). Das Stabilitätsgebiet von Polynomen ist kompakt.

Beweis. Für jedes Polynom P vom Grad ≥ 1 gilt $|P(z)| \xrightarrow{z\to\infty} \infty$. Also ist S beschränkt. □

Implizite Verfahren können A-stabil sein: z.B. das implizite Euler-Verfahren:

$$R(z) = \frac{1}{1-z}$$

mit dem Stabilitätsgebiet: $S = \{z \in \mathbb{C} \mid |1 - z| \geq 1\} \supset \mathbb{C}$.

Das Ziel für die Zukunft lautet jetzt, A-stabile Verfahren hoher Ordnung zu konstruieren.

1.4 Implizite Runge–Kutta-Verfahren

Wir betrachten jetzt wieder allgemeine nichtlineare, nicht-autonome Anfangswertprobleme

$$x' = f(t, x) \qquad x(t_0) = x_0$$

Wie können wir stabile Verfahren hoher Konsistenzordnung konstruieren?

Definition (Butcher 1964). Unter einem **allgemeinen Runge-Kutta-Verfahren** (kurz: RK-Verfahren) verstehen wir ein Verfahren der Form

$$k_i = f\Big(t + c_i\tau, x + \tau \sum_{j=1}^{s} a_{ij} k_j\Big), \qquad i = 1, \dots, s, \tag{i}$$

$$\Psi^{t+\tau,t} x = x + \tau \sum_{j=1}^{s} b_j k_j \tag{ii}$$

Im Unterschied zu expliziten RK-Verfahren geht die Summe in (i) über alle s Stufen, nicht nur über die ersten $i-1$.

Man fasst die Koeffizienten wieder in zwei Vektoren $b, c \in \mathbb{R}^s$ und eine Matrix $\mathcal{A} \in \mathbb{R}^{s\times s}$ zusammen. Darstellung der Koeffizienten wieder im Butcher-Schema:

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}$$

Beispiel (Implizites Euler-Verfahrens).

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Ein Verfahren ist explizit, wenn $\mathcal{A}$ eine strikte untere Dreiecksmatrix ist. Ansonsten muss in jedem Schritt ein nichtlineares Gleichungssystem gelöst werden.

Frage: Unter welchen Bedingungen ist dieses Gleichungssystem eindeutig lösbar (für hinreichend kleine τ?) Nur dann kann ja von einem wohldefinierten Zeitintegrationsverfahren gesprochen werden.

Um diese Frage zu klären schreiben wir das Verfahren zunächst in der sogenannten **symmetrischen Form** auf. Definiere

$$g_i = x + \tau \sum_{j=1}^{s} a_{ij} k_j, \qquad i = 1, \dots, s$$

Dann gilt

$$g_i = x + \tau \sum_{j=1}^{s} a_{ij} f(t + c_j\tau, g_j) \qquad i = 1, \dots, s$$

$$\Psi^{t+\tau,t} x = x + \tau \sum_{j=1}^{s} b_j f(t + c_j\tau, g_j)$$

Satz 1.8 ([DB08, Satz 6.28]). Die Abbildung $f \in C(\Omega, \mathbb{R}^d)$ sei auf $\Omega \subset \mathbb{R} \times \mathbb{R}^d$ bezüglich x lokal Lipschitz-stetig. Für ein implizites RK-Verfahren gibt es zu jedem $(x,t) \in \Omega$ ein $\tau_* > 0$, und *eindeutige* stetige Funktionen

$$g_i \colon (-\tau_*, \tau_*) \to \mathbb{R}^d \qquad (i = 1, \ldots, s)$$

so dass:

1. $g_i(0) = x$ für $i = 1, \ldots, s$
2. Für $|\tau| < \tau_*$ genügen die Vektoren $g_i(\tau)$, $i = 1, \ldots, s$, den Bestimmungsgleichungen des impliziten RK-Verfahrens.

Für den Beweis brauchen wir einen besonderen, parameterabhängigen Fixpunktsatz. Den Beweis dazu findet man bei [Die60, Satz 10.1.1].

Satz 1.9. Es seien E_1 und E_2 zwei Banach-Räume. U und V seien offene Kugeln in E_1 (bzw. E_2) jeweils um 0; der Radius von V sei β. Sei F eine stetige Abbildung von $U \times V$ nach E_2, so dass $\|F(\tau, y_1) - F(\tau, y_2)\| \leq \theta \cdot \|y_1 - y_2\|$ für $\tau \in U$, $y_1, y_2 \in V$ und θ eine Konstante mit $0 \leq \theta < 1$. Falls $\|F(\tau, 0)\| < \beta(1-\theta)$ für alle $\tau \in U$, existiert dann eine eindeutige Abbildung $g : U \to V$ so dass

$$g(\tau) = F(\tau, g(\tau))$$

für alle $\tau \in U$, und g ist stetig in U.

Beweis (Satz 1.8).

- Sei $(t_0, x_0) \in \Omega$ fest gewählt.
- f ist lokal Lipschitz-stetig bezüglich x. D.h. es gibt Parameter $\tau_1, \rho, L > 0$ so dass $|f(t,x) - f(t,\bar{x})| < L|x - \bar{x}|$ für alle $(t,x), (t,\bar{x}) \in (t_0 - \tau_1, t_0 + \tau_1) \times B_\rho(x_0) \subset \Omega$.
- Weiterhin gilt $|f(t, x_0)| < M$ für alle $t \in (t_0 - \tau_1, t_0 + \tau_1)$ (zur Not wird dafür τ_1 verkleinert).
- Wir wählen jetzt ein $0 < \theta < 1$ (Das wird das θ aus dem Satz von Dieudonné).
- Schreibe das RK-System als parameterabhängige Fixpunktgleichung $g(\tau) = F(\tau, g(\tau))$ mit

$$\begin{aligned} g &:= (g_1, \ldots, g_s)^T, \\ F(\tau, g) &:= \big(F_1(\tau, g), \ldots, F_s(\tau, g)\big)^T, \\ F_i(\tau, g) &= x_0 + \tau \sum_{j=1}^{s} a_{ij} f(t_0 + c_j \tau, g_j), \qquad i = 1, \ldots s. \end{aligned}$$

- g ist aus $E_2 = \mathbb{R}^{s \cdot d}$. Wähle dort die Norm $\|g\| = \max_{1 \leq i \leq s} |g_i|$.
- Notation: $g_* := (x_0, \ldots, x_0) \in \mathbb{R}^{s \cdot d}$.
- Jetzt definieren wir die Kugeln in $E_1 = \mathbb{R}$ und $E_2 = \mathbb{R}^{s \cdot d}$

$$U = (-\tau_*, \tau_*), \qquad V = \{g \in \mathbb{R}^{s \cdot d} \mid \|g - g_*\| < \rho\}$$

- Damit ist $F : U \times V \to \mathbb{R}^{s \cdot d}$ wohldefiniert und stetig.

- Wir zeigen als nächstes die Lipschitz-Stetigkeit von F im zweiten Argument: Für alle $(\tau, g), (\tau, \bar{g}) \in U \times V$ gilt

$$\begin{aligned}
\|F(\tau,g) - F(\tau,\bar{g})\| &= \|(F_1(\tau,g) - F_1(\tau,\bar{g}), \dots, F_s(\tau,g) - F_s(\tau,\bar{g}))^T\| \\
&= \max_{1\le i\le s} |F_i(\tau,g) - F_i(\tau,\bar{g})| \\
&= \max_{1\le i\le s} \Big| x_0 + \tau \sum_{j=1}^{s} a_{ij} f(t_0 + c_j\tau, g_j) - x_0 - \tau \sum_{j=1}^{s} a_{ij} f(t_0 + c_j\tau, \bar{g}_j) \Big| \\
&\le \tau \max_{1\le i\le s} \sum_{j=1}^{s} \Big[|a_{ij}| \cdot |f(t_0 + c_j\tau, g_j) - f(t_0 + c_j\tau, \bar{g}_j)| \Big] \\
&\le \tau \underbrace{\max_{1\le i\le s} \sum_{j=1}^{s} |a_{ij}|}_{=\|\mathcal{A}\|_\infty} \cdot \max_{1\le j\le s} |f(t_0 + c_j\tau, g_j) - f(t_0 + c_j\tau, \bar{g}_j)| \\
&\le \tau_* \|\mathcal{A}\|_\infty \max_{1\le j\le s} \Big| f(t_0 + c_j\tau, g_j) - f(t_0 + c_j\tau, \bar{g}_j) \Big|.
\end{aligned}$$

- Da f lokal Lipschitz-stetig im zweiten Argument ist, gilt

$$\begin{aligned}
\|F(\tau,g) - F(\tau,\bar{g})\| &\le \tau_* \|\mathcal{A}\|_\infty L \max_{1\le j\le s} |g_j - \bar{g}_j| \\
&= \tau_* \|\mathcal{A}\|_\infty L \|g - \bar{g}\| \\
&\le \theta \|g - \bar{g}\|
\end{aligned}$$

  wenn $\tau_* \le \frac{\theta}{L\|\mathcal{A}\|_\infty}$.

- Ähnlich zeigt man

$$\begin{aligned}
\|F(\tau,g_*) - g_*\| &= \|F(\tau,g_*) - F(0,g_*)\| \\
&= \max_{1\le i\le s} \Big| x_0 + \tau \sum_{j=1}^{s} a_{ij} f(t_0 + c_j\tau, x_0) - x_0 - 0 \Big| \\
&= \max_{1\le i\le s} \Big| \tau \sum_{j=1}^{s} a_{ij} f(t_0 + c_j\tau, x_0) \Big| \\
&\le \tau_* \|\mathcal{A}\|_\infty \max_{1\le j\le s} |f(t_0 + c_j\tau, x_0)| \\
&< \tau_* \|\mathcal{A}\|_\infty M \\
&\le \rho(1-\theta),
\end{aligned}$$

  falls $\tau_* \le \frac{\rho(1-\theta)}{\|\mathcal{A}\|_\infty M}$.

  (Der Fixpunktsatz fordert $\|F(\tau,0) - 0\| \le \rho(1-\theta)$, weil dort die Kugeln um 0 zentriert sind, und nicht wie hier um g_*.)

Wende jetzt den parameterabhängigen Fixpunktsatz 1.9 an. Er liefert

- Es existieren eindeutige $g(\tau) \in V$ für alle $\tau \in U = (-\tau_*, \tau_*)$, so dass

$$g(\tau) = F(\tau, g(\tau)).$$

- $g(\tau)$ ist stetig, insbesondere ist $g(0) = g_*$. □

Implizite RK-Verfahren sind also für kleine τ wohldefiniert. Wie sieht es mit Konsistenz und Stabilität aus?

Erinnerung: Konsistenztheorie für explizite RK-Verfahren

- Entwickle Φ und Ψ als Taylorreihen
- Wähle die Koeffizienten $b, c, \mathcal{A}$ so, dass möglichst viele Terme aus der Taylorreihe von Φ reproduziert werden.

Im Prinzip funktioniert das für implizite Verfahren genauso. Jedoch ist alles etwas komplizierter: Es müssen mehr Koeffizienten bestimmt werden; aber auch alles etwas einfacher: Für eine gegebene Ordnung p erhält man die gleiche Anzahl von Bestimmungsgleichungen wie im expliziten Fall [DB08, Satz 4.24]. Man hat aber mehr Freiheitsgrade, um diese zu erfüllen.

Satz 1.10. Unter den Bedingungen des obigen Satzes gilt:

- Die Evolution Ψ ist genau dann konsistent, wenn
$$\sum_{i=1}^{s} b_i = 1$$
- Ist $f \in C^p(\Omega, \mathbb{R}^d)$, so ist auch $\Psi^{t+\tau,t}x$ in τ p-fach differenzierbar.

Wie für explizite Verfahren zeigt man: Ein implizites Verfahren ist genau dann invariant unter Autonomisierung, wenn es konsistent ist und

$$c_i = \sum_{j=1}^{s} a_{ij} \qquad \text{für } i = 1, \dots, s$$

gilt.

Welche Ordnung kann man maximal erzielen? → Dafür braucht man die Stabilitätsfunktion.

Lemma 1.8. Die Stabilitätsfunktion R eines s-stufigen RK-Verfahrens $(b, \mathcal{A})$ ist durch
$$R(z) = 1 + zb^T(I - z\mathcal{A})^{-1}(1, \dots, 1)^T$$
gegeben. Die Funktion R kann in eindeutiger Weise als
$$R(z) = \frac{P(z)}{Q(z)}$$
dargestellt werden, wobei P, Q teilerfremde Polynome höchstens s-ten Grades mit $P(0) = Q(0) = 1$ sind.

Beweis. Übung. □

Erinnerung: Ein s-stufiges explizites RK-Verfahren hat höchstens die Konsistenzordnung $p \leq s$.

Lemma 1.9. Ein s-stufiges implizites RK-Verfahren besitze für alle $f \in C^\infty(\Omega, \mathbb{R}^d)$ die Konsistenzordnung $p \in \mathbb{N}$. Dann gilt $p \leq 2s$.

Beweis.
- Betrachte das AWP $x' = x$, $x(0) = 1$ mit Lösung $x(\tau) = e^\tau$ und RK-Approximation Ψ^τ.

- Das Verfahren ist konsistent mit Konsistenzordnung p, also gilt

$$\Psi^\tau 1 - \Phi^\tau 1 = R(\tau) - e^\tau = \mathcal{O}(\tau^{p+1})$$

- $R = \frac{P}{Q}$ ist Quotient zweier Polynome mit Ordnung jeweils $\leq s$. Es folgt $p \leq \deg P + \deg Q \leq 2s$.

Warum? Angenommen es gäbe Polynome P, Q mit $\deg P \leq k$ und $\deg Q \leq j$ sowie $k + j < p$. Das hieße $\frac{P(z)}{Q(z)} - e^z = \mathcal{O}(z^{k+j+2})$. Multiplikation mit $Q(z)$liefert dann $P(z) - Q(z)e^z = \mathcal{O}(z^{k+j+2})$. Daraus folgt $P = Q = 0$ ein Widerspruch! (Beweis: Übung. [DB08, Lemma 6.4]) □

1.5 Kollokationsverfahren

Wie konstruiert man jetzt konkrete RK-Verfahren? Die folgende Idee ist unabhängig von den RK-Verfahren entwickelt worden. Dass man dadurch implizite RK-Verfahren erhält wurde erst in den 1970ern entdeckt.

Betrachte

$$x' = f(t, x)$$

Seien $(t, x) \in \Omega$ und eine Schrittweite τ gegeben. Gesucht ist nun ein Schritt einer diskreten Evolution $\Psi^{t+\tau,t}x$. **Idee:**

Idee.

- *Wähle s Stützstellen im Intervall $(t, t + \tau)$*

$$t + c_i\tau \qquad 0 \leq c_1 < c_2 < \cdots < c_s \leq 1$$

- *Konstruiere ein Polynom $u \in P_s^d$, das*
  1. *den Anfangswert $u(t) = x$ erfüllt und*
  2. *die Differentialgleichung an den Stützstellen erfüllt*

$$u'(t + c_i\tau) = f(t + c_i\tau, u(t + c_i\tau)) \qquad (i = 1, \ldots, s)$$

  3. *Setze*

$$\Psi^{t+\tau,t}x := u(t + \tau)$$

Ein Bild!

*Diese Bedingungen nennen wir **Kollokationsbedingungen**.*

Einziger Parameter des Verfahrens sind die Stützstellen $c_1, \ldots, c_s$. Wir haben $s + 1$ Bedingungen an ein Polynom s-ten Grades. Wir *vermuten*, dass ein eindeutiges u existiert (zumindest für kleine τ). Klar ist das nicht, denn die Gleichungen für u sind nichtlinear!

Einfacher Ausweg: Wir interpretieren das Verfahren als implizites RK-Verfahren. Dann liefert Satz 1.8 Existenz und Eindeutigkeit.

- Angenommen es existiere eine Lösung $u \in P_s^d$.
- Sei $\{L_1, \ldots, L_s\}$ die Lagrange-Basis von P_{s-1} bezüglich der c_i, also

$$L_i(c_j) = \delta_{ij} \qquad (i, j = 1, \ldots, s)$$

- u' ist in P^d_{s-1} und hat Lagrange-Darstellung

$$u'(t+\theta\tau) = \sum_{j=1}^{s} \underbrace{u'(t+c_j\tau)}_{k_j:=} L_j(\theta) = \sum_{j=1}^{s} k_j L_j(\theta). \tag{1.2}$$

- Wir integrieren und nutzen die Kollokationsbedingung 1: $u(t) = x$

$$\begin{aligned} u(t+c_i\tau) &= u(t) + \int_t^{t+c_i\tau} u'(s)\,ds \\ &= x + \tau \int_0^{c_i} u'(t+\theta\tau)\,d\theta \\ &= x + \tau \int_0^{c_i} \sum_{j=1}^{s} k_j L_j(\theta)\,d\theta \\ &= x + \tau \sum_{j=1}^{s} k_j \underbrace{\int_0^{c_i} L_j(\theta)\,d\theta}_{a_{ij}:=} \\ &= x + \tau \sum_{j=1}^{s} a_{ij} k_j. \end{aligned}$$

- Das setzen wir in die Kollokationsbedingung 2 ein:

$$k_i = f\Big(t + c_i\tau, x + \tau \sum_{j=1}^{s} a_{ij} k_j\Big) \qquad (i = 1, \ldots, s)$$

- Abschließend benutzen wir die Kollokationsbedingung 3:

$$\begin{aligned} \Psi^{t+\tau,t} x &= u(t+\tau) \\ &= x + \tau \int_0^1 u'(t+\theta\tau)\,d\theta \\ &= x + \tau \int_0^1 \sum_{j=1}^{s} k_j L_j(\theta)\,d\theta \qquad \text{(wegen (1.2))} \\ &= x + \tau \sum_{j=1}^{s} b_j k_j \end{aligned}$$

  mit

$$b_j = \int_0^1 L_j(\theta)\,d\theta \qquad (j = 1, \ldots, s)$$

Wir erhalten tatsächlich ein RK-Verfahren mit

$$a_{ij} = \int_0^{c_i} L_j(\theta)\,d\theta, \qquad L_i(\theta) := \frac{\prod_{j\neq i}(c_j - \theta)}{\prod_{j\neq i}(c_j - c_i)} \qquad (i, j = 1, \ldots, s)$$

$$b_j = \int_0^1 L_j(\theta)\,d\theta \qquad (j = 1, \ldots, s)$$

Diese Größen hängen nur von den c_i ab.

- Die Stufen k_i sind gerade die Ableitungen von u an den Stützstellen c_i:

$$k_i = u'(t + c_i\tau) \qquad (i = 1, \ldots, s)$$

- Deutliche Reduktion der Komplexität: Nur noch s Freiheitsgrade $c_1, \dots, c_s$ statt bisher $2s + s^2$ Freiheitsgrade $c, b, \mathcal{A}$.
- Durch Satz 1.8 bekommen wir die Existenz einer eindeutigen Lösung für das Kollokationsproblem!

Aber sind alle Kollokations-Verfahren auch *gute* RK-Verfahren?

Erinnerung — Satz 4.18 aus [DB08]: Ein RK-Verfahren besitzt

- genau dann die Konsistenzordnung 1, wenn $\sum_{i=1}^{s} b_i = 1$
- genau dann die Konsistenzordnung 2, wenn zusätzlich $\sum_{i=1}^{s} b_i c_i = \frac{1}{2}$

Lemma 1.10 ([DB08, 6.37]). Die Koeffizienten eines durch Kollokation definierten RK-Verfahrens $(b, c, \mathcal{A})$ erfüllen

$$\sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k} \qquad k = 1, \dots, s. \tag{1.3}$$

Insbesondere sind solche Verfahren also konsistent ($k = 1$) und von mindestens zweiter Ordnung wenn $s \geq 2$.

Beweis. Nach Definition der b_j gilt

$$\sum_{j=1}^{s} b_j c_j^{k-1} = \sum_{j=1}^{s} \int_0^1 c_j^{k-1} L_j(\theta)\, d\theta.$$

$\sum_{j=1}^{s} c_j^{k-1} L_j(\theta)$ ist gerade die Lagrange-Darstellung des Polynoms θ^{k-1}. Deshalb

$$\sum_{j=1}^{s} b_j c_j^{k-1} = \int_0^1 \theta^{k-1}\, d\theta = \frac{1}{k}$$ □

Es besteht ein enger Zusammenhang zwischen RK-Verfahren und Quadraturformeln.

Lemma 1.11. Interpretiere die b_j $(j = 1, \dots, s)$ als Gewichte einer Quadraturformel mit Stützstellen c_j. Aus (1.3) folgt, dass diese Quadraturformel für Polynome höchstens $(s-1)$-ten Grades exakt ist.

Beweis. Sei $\pi \in P_{s-1}$, $\pi(\theta) = \sum_{j=0}^{s-1} \alpha_j \theta^j$. Dann ist

$$\begin{aligned}
\sum_{i=1}^{s} b_i \pi(c_i) &= \sum_{i=1}^{s} b_i \sum_{j=1}^{s} \alpha_{j-1} c_i^{j-1} \\
&= \sum_{j=1}^{s} \alpha_{j-1} \sum_{i=1}^{s} b_i c_i^{j-1} = \sum_{j=1}^{s} \alpha_{j-1} \frac{1}{j} = \sum_{j=1}^{s} \alpha_{j-1} \int_0^1 \theta^{j-1}\, d\theta \\
&= \int_0^1 \sum_{j=1}^{s} \alpha_{j-1} \theta^{j-1}\, d\theta = \int_0^1 \pi(\theta)\, d\theta
\end{aligned}$$ □

Wir werden sehen:

- Die Konsistenzordnung eines Kollokationsverfahrens wird im Wesentlichen durch die Eigenschaften dieser Quadraturformel bestimmt.
- Man kann aus bekannten Quadraturformeln Kollokationsverfahren gleicher Konsistenzordnung konstruieren.

Satz 1.11 ([DB08, 6.40]). Ein durch Kollokation erzeugtes implizites RK-Verfahren $(b, c, \mathcal{A})$ besitzt die Konsistenzordnung p für rechte Seiten $f \in \mathcal{C}^p(\Omega, \mathbb{R}^d)$ genau dann, wenn die durch Stützstellen c und Gewichte b gegebene Quadraturformel die Ordnung p besitzt.

Beweis (Skizze). Teil 1: RK-Verfahren hat Ordnung $p \quad \Rightarrow \quad$ Quadraturformel hat Ordnung p

- $(b, c, \mathcal{A})$ besitze Konsistenzordnung $p \Rightarrow$ Der Fehler bei einem Schritt ist $\mathcal{O}(\tau^{p+1})$.
- Insbesondere können wir Gleichungen der Form $x' = f(t)$ integrieren mit Lösung ist $x(t+\tau) = x(t) + \int_t^{t+\tau} f(s)\,ds$.
- Runge-Kutta-Verfahren dafür:

$$\Psi^{t+\tau,t}x = x + \tau \sum_{i=1}^{s} b_i k_i = x + \tau \sum_{i=1}^{s} b_i f(t + c_i\tau)$$

- Konsistenzfehler

$$\Psi^{t+\tau,t}x - \Phi^{t+\tau,t}x = x + \tau \sum_{i=1}^{s} b_i f(t + c_i\tau) - x - \int_t^{t+\tau} f(s)\,ds$$

  hat Ordnung p, d.h.

$$\tau \sum_{i=1}^{s} b_i f(t + c_i\tau) - \int_t^{t+\tau} f(s)\,ds = \mathcal{O}(\tau^{p+1})$$

- Das ist aber gerade die Definition davon dass die Quadraturformel von p-ter Ordnung ist ([DB08, Lemma 6.39]).

Teil 2: Quadraturformel hat Ordnung $p \Rightarrow$ RK hat Ordnung p

- Wir betrachten jetzt wieder allgemeine $x' = f(t, x)$.
- Sei τ so klein, dass das Kollokationspolynom $u \in P_s$ existiert. Betrachte u als Lösung einer Störung des AWPs

$$x'(\bar{t}) = f(\bar{t}, x(\bar{t})), \qquad x(t) = x.$$

  Dazu wird die rechte Seite gestört!

- Konkret löst u das AWP

$$u'(\bar{t}) = f(\bar{t}, u(\bar{t})) + [\underbrace{u'(\bar{t}) - f(\bar{t}, u(\bar{t}))}_{=:\delta f(\bar{t})}], \qquad u(t) = x$$

Plan: Schließe aus der Größe von δf auf den Fehler $x(t+\tau) - u(t+\tau)$. Das ist aber gerade der Konsistenzfehler $\Psi^{t+\tau,t}x - \Phi^{t+\tau,t}x$!

Ideen dabei:

- δf verschwindet an den Stützstellen der Quadraturformel
- Wird auch an den anderen Punkten klein bleiben

Wir benutzen ein allgemeines Resultat aus der Störungstheorie gewöhnlicher Differentialgleichungen.

Satz 1.12 (Aleksejew, Gröber ([DB08, Satz 3.4])). Es existiert eine beliebig häufig differenzierbare matrixwertige Funktion $M(\bar{t}, \sigma)$, so dass

$$x(t+\tau) - u(t+\tau) = \int_t^{t+\tau} M(t+\tau, \sigma)\, \delta f(\sigma)\, d\sigma$$

Schätze nun das Integral mit der Quadraturformel ab

$$x(t+\tau) - u(t+\tau) = \tau \sum_{j=1}^{s} b_j M(t+\tau, t+c_j\tau)\, \delta f(t+c_j\tau) + \mathcal{O}(\tau^{p+1})$$

u ist aber Kollokationspolynom. Deshalb ist

$$\delta f(t+c_j\tau) = u'(t+c_j\tau) - f(t+c_j\tau, u(t+c_j\tau)) = 0 \qquad \forall j = 1, \ldots, s.$$

Also folgt

$$x(t+\tau) - u(t+\tau) = \mathcal{O}(\tau^{p+1})$$ □

Bei diesem Argument muss man aber vorsichtig sein. Die Konstante in $\mathcal{O}(\tau^{p+1})$ hängt von höheren Ableitungen von $M(t+\tau, s)\delta f(s)$ nach s ab. Dieser Ausdruck hängt aber von u ab und u wiederum hängt von τ ab!

Es geht aber trotzdem alles gut: siehe dazu [DB08, Lemma 6.41], ca. 2 Seiten lang.

Was haben wir gelernt?

- In jedem Kollokationsverfahren steckt eine Quadraturformel der Ordnung $s \leq p \leq 2s$. Diese Ordnung wird an des Kollokations-Verfahren vererbt.
- Das betrifft nur den Fehler am Ende eines Zeitschritts, d.h.

$$x(t+\tau) - u(t+\tau) = \mathcal{O}(\tau^{p+1})$$

- Gleichzeitig hat man in Form von u auch eine Approximation für $x(\sigma)$ für alle σ **zwischen** t und $t+\tau$. Für die gilt ([DB08, Lemma 6.41]):

$$\max_{t \leq \sigma \leq t+\tau} |x(\sigma) - u(\sigma)| = \mathcal{O}(\tau^{s+1})$$

- Also schlechter als am Intervallende (da $s \leq p$) – Diesen Effekt nennt man **Superkonvergenz**.

1.5.1 Gauß-Verfahren

Wir bauen ein implizites RK-Verfahren hoher Ordnung:

- Wähle eine möglichst gute Quadraturformel.
- Konstruiere das dazugehörige Kollokationsverfahren.

Das Optimum für s Stützstellen sind Quadraturregeln der Ordnung $p = 2s$, d.h. Regeln die Polynome bis zum Grad $2s-1$ exakt integrieren.

Erinnerung: Ist eine Quadraturformel

$$\int_0^1 \phi(t)\, dt \approx \sum_{i=1}^{s} b_i \phi(c_i)$$

exakt für Polynome des Grades $2s-1$, so sind die Stützstellen

$$0 < c_1 < \ldots < c_s < 1$$

eindeutig definiert als die Nullstellen des s-ten Legendre-Polynoms P_s.

Definition (Legendre-Polynom).

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\ &\vdots \end{aligned}$$

Dies sind die Standarddarstellungen bzgl. des Intervals $[-1, 1]$. Für unsere Zwecke muss noch auf $[0, 1]$ umtransformiert werden.

Kollokationsverfahren mit diesen Stützstellen werden **Gauß-Verfahren** genannt.

Aus unserem Satz 1.11 folgt direkt das folgende Resultat zur Konsistenzordnung von Gauß-Verfahren.

Satz 1.13 ([DB08, Satz 6.43]). Für $f \in \mathcal{C}^{2s}(\Omega, \mathbb{R}^d)$ besitzt das s-stufige Gauß-Verfahren die Konsistenzordnung $p = 2s$.

Zusätzlich wollen wir aber auch A-Stabilität.

Satz 1.14 ([DB08], Satz 6.44). Jedes Gauß-Verfahren ist A-stabil.

Das beweisen wir auf einem kleinen Umweg über B-Stabilität.

1.6 Dissipative Differentialgleichungen

Wir müssen noch beweisen, dass Gauß-Verfahren A-stabil sind.

Das machen wir über einen Umweg: Wir führen einen neuen, stärkeren Stabilitätsbegriff ein, der direkt auf nichtlineare Gleichungen zielt. Dann zeigen wir, dass Gauß-Verfahren sogar in diesem stärkeren Sinne stabil sind.

Bei bisherigen Stabilitätsuntersuchungen haben wir uns auf lineare Differentialgleichungen

$$x' = \lambda x \tag{1.4}$$

beschränkt. Die waren genau dann stabil, wenn $\lambda \leq 0$.

Das kann man anders formulieren: Die Gleichung (1.4) ist stabil, wenn die rechte Seite $f(t, x) = \lambda x$ monoton fallend in x ist.

Das verallgemeinern wir jetzt für allgemeine, autonome Gleichungen

$$x' = f(x) \qquad \text{mit } x(t) \in \mathbb{R}^d$$

Definition. Eine Abbildung $f : \Omega_0 \to \mathbb{R}^d$ heißt **monoton fallend** oder **dissipativ** bzgl. eines Skalarproduktes $\langle \cdot, \cdot \rangle$, wenn für alle $x, \bar{x} \in \Omega_0$

$$\langle f(x) - f(\bar{x}), x - \bar{x} \rangle \leq 0$$

gilt.

Als nächstes definieren wir eine Variante des Begriffs „stabile Differentialgleichung“.

Definition. Ein Phasenfluss $\Phi^t : \Omega_0 \to \Omega_0$ heißt **nichtexpansiv**, wenn

$$|\Phi^t x - \Phi^t \bar{x}| \leq |x - \bar{x}|$$

für alle $x, \bar{x} \in \Omega_0$ und alle zulässigen t.

Lemma 1.12. Sei $x' = f(x)$ Differentialgleichung auf Ω_0 mit lokal Lipschitz-stetigem f.

$$\Phi \text{ nichtexpansiv} \quad \Leftrightarrow \quad f \text{ dissipativ}$$

Beweis. Betrachte die Funktion

$$\chi(t) \coloneqq |\Phi^t x - \Phi^t \bar{x}|^2 = \langle \Phi^t x - \Phi^t \bar{x}, \Phi^t x - \Phi^t \bar{x} \rangle$$

Ableiten nach t ergibt

$$\begin{aligned} \chi'(t) &= \langle (\Phi^t x)' - (\Phi^t \bar{x})', \Phi^t x - \Phi^t \bar{x} \rangle + \langle \Phi^t x - \Phi^t \bar{x}, (\Phi^t x)' - (\Phi^t \bar{x})' \rangle \\ &= 2 \langle (\Phi^t x)' - (\Phi^t \bar{x})', \Phi^t x - \Phi^t \bar{x} \rangle \\ &= 2 \langle f(\Phi^t x) - f(\Phi^t \bar{x}), \Phi^t x - \Phi^t \bar{x} \rangle. \end{aligned}$$

(⇐) Sei f dissipativ. Dann ist

$$\chi(t) = \chi(0) + \int_0^t 2 \underbrace{\langle f(\Phi^s x) - f(\Phi^s \bar{x}), \Phi^s x - \Phi^s \bar{x} \rangle}_{\leq 0} \, ds \leq \chi(0)$$

(⇒) Sei Φ nichtexpansiv. Dann ist $\chi(t) \leq \chi(0)$ für alle hinreichend kleinen t, d.h. χ ist monoton fallend bei $t = 0$. Somit ist $\chi'(0) = 2 \langle f(x) - f(\bar{x}), x - \bar{x} \rangle \leq 0$. □

Nichtexpansivität ist eine Eigenschaft, die man eventuell vererben möchte.

Definition (Butcher 1975). Ein Verfahren heißt **B-stabil**, wenn es für dissipative, hinreichend glatte rechte Seiten einen nichtexpansiven diskreten Phasenfluss erzeugt, also

$$|\Psi^\tau x - \Psi^\tau \bar{x}| \leq |x - \bar{x}|$$

für alle zulässigen $x, \bar{x}, \tau$.

Dieses Konzept ist stärker als A-Stabilität.

Lemma 1.13 ([DB08, Satz 6.50]). B-stabile Runge-Kutta-Verfahren sind A-stabil.

Beweis. Betrachte das komplexe AWP

$$x' = \lambda x, \; x(0) = 1, \; \lambda \in \mathbb{C}, \qquad \operatorname{Re} \lambda \leq 0$$

(Bei Systemen steht dieses AWP stellvertretend für einen Eigenwert.) Das AWP ist stabil. Ist die rechte Seite dissipativ?

- Reellifizierung: $x = u + iv$, $\lambda = \alpha + i\beta$

$$x' = \lambda x \quad \Leftrightarrow \quad \begin{pmatrix} u \\ v \end{pmatrix}' = \underbrace{\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}}_{=:A} \begin{pmatrix} u \\ v \end{pmatrix}$$

- Test auf Dissipativität von $f(x) = Ax$:

$$\begin{aligned} \langle Ax - A\bar{x}, x - \bar{x} \rangle = \langle A\tilde{x}, \tilde{x} \rangle &= (\tilde{u}, \tilde{v}) \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \\ &= \alpha(\tilde{u}^2 + \tilde{v}^2) \\ &\leq 0, \quad \text{da } \alpha = \operatorname{Re}\lambda \leq 0, \ f \text{ ist dissipativ!} \end{aligned}$$

Das Verfahren ist B-stabil. Also erhält man für diese dissipative rechte Seite einen nichtexpansiven diskreten Phasenfluss

$$\begin{aligned} |x - \bar{x}| \geq |\Psi^\tau x - \Psi^\tau \bar{x}| &= \underbrace{|R(\tau A)x - R(\tau A)\bar{x}|}_{(R\text{: Stabilitätsfunktion des Verfahrens})} \\ &= |R(\tau\lambda)(x - \bar{x})| = |R(\tau\lambda)| \cdot |x - \bar{x}| \end{aligned}$$

($|\cdot|$ ist als komplexer Betrag multiplikativ.) Daraus folgt $R(\tau\lambda) \leq 1$ für alle $\tau \geq 0$.

- Für $\tau = 1$ erhält man $|R(\lambda)| \leq 1$, also

$$\lambda \in S \coloneqq \{z \in \mathbb{C} : \ |R(\lambda)| \leq 1\}, \text{ das Stabilitätsgebiet.}$$

- λ ist in $\mathbb{C}_-$ beliebig

$\Rightarrow$ Das Verfahren ist A-stabil. □

Statt der A-Stabilität von Gauß-Verfahren zeigen wir:

Satz 1.15 ([DB08, 6.51]). Gauß-Verfahren sind B-stabil.

Beweis.
- Die rechte Seite f sei dissipativ und hinreichend glatt.
- Zu zeigen: Der diskrete Phasenfluss einen Gauß-Verfahrens ist nichtexpansiv.
- Wähle $x, \bar{x} \in \Omega_0$. Sofern τ klein genug ist, existieren die Kollokationspolynome $u, \bar{u} \in P_s$, mit

$$u(0) = x, \ u(\tau) = \Psi^\tau x, \qquad \bar{u}(0) = \bar{x}, \ \bar{u}(\tau) = \Psi^\tau \bar{x}$$

- Betrachte die Differenz

$$q(\theta) = |u(\theta\tau) - \bar{u}(\theta\tau)|^2$$

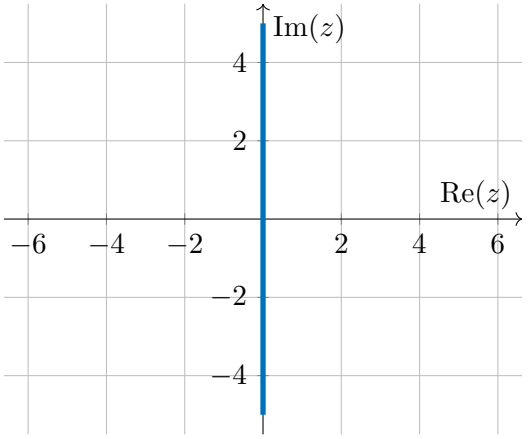
- Hauptsatz der Integralrechnung:

$$\begin{aligned} |\Psi^\tau x - \Psi^\tau \bar{x}|^2 = q(1) &= q(0) + \int_0^1 q'(\theta)\, d\theta \\ &= |x - \bar{x}|^2 + \int_0^1 q'(\theta)\, d\theta \end{aligned}$$

- Es ist also zu zeigen, dass
$$\int_0^1 q'(\theta)\,d\theta \leq 0$$
- Aber $q(\theta) = |u(\theta\tau) - \bar{u}(\theta\tau)|^2$ ist ein Polynom in θ vom Grad höchstens $2s$. Also ist q' ein Polynom vom Grad $2s-1$. Dafür ist Gauß-Quadratur exakt:
$$\int_0^1 q'(\theta)\,d\theta = \sum_{j=1}^{s} b_j q'(c_j).$$
- Wir zeigen jetzt, dass $q'(c_j) \leq 0$ für alle $j = 1, \cdots, s$. Da für alle Gauß-Quadraturformeln $b_j \geq 0\ \forall j$ gilt, folgt dann die Behauptung.
- Es gilt $q'(\theta) = 2\langle u'(\theta\tau) - \bar{u}'(\theta\tau), u(\theta\tau) - \bar{u}(\theta\tau)\rangle$
- Da $u, \bar{u}$ Kollokationspolynome sind, folgt weiter:
$$q'(c_j) = 2\langle f(u(c_j\tau)) - f(\bar{u}(c_j\tau)), u(c_j\tau) - \bar{u}(c_j\tau)\rangle\ \forall j = 1, \ldots, s$$
- Diese Ausdrücke sind alle ≤ 0, da f dissipativ ist. □

Achtung: Nicht alle A-stabilen Verfahren sind B-stabil!

Beispiel. Die Stabilitätsfunktion $R(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}$ beschreibt A-stabile Verfahren.



- Diese Stabilitätsfunktion gehört zur impliziten Mittelpunktsregel
$$\Psi^{t+\tau,t}x = \xi, \quad \xi = x + \tau f\Big(t + \frac{\tau}{2}, \frac{x+\xi}{2}\Big).$$
- Das ist ein Gauß-Verfahren ($s = 1$), also B-stabil.
- Die selbe Stabilitätsfunktion gehört außerdem zur impliziten Trapezregel
$$\Psi^{t+\tau,t}x = \xi, \quad \xi = x + \frac{\tau}{2}\Big[f(t+\tau,\xi) + f(t,x)\Big].$$
Dieses Verfahren ist *nicht* B-stabil!

Beweisskizze.

- Betrachte $x' = f(x)$ (skalar) mit $f(x) = \begin{cases} |x|^3, & x \leq 0 \\ -x^2, & x > 0 \end{cases}$

- f ist C^1, monoton fallend, also dissipativ.
- $x \equiv 0$ ist Fixpunkt der Gleichung und der impliziten Trapezregel. Wenn die implizite Trapezregel B-stabil sein soll, muss also

$$|\Psi^\tau x - \underbrace{\Psi^\tau 0}_{=0}| \leq |x - 0|$$

  für alle $x \in \mathbb{R},\ \tau > 0$ gelten (Das Verfahren existiert für alle τ).
- Allerdings erhält man für $x = -2$, $\tau = \frac{36}{7}$ gerade $\Psi^\tau x = 2{,}5$. $\Rightarrow$ Widerspruch!

1.7 Linear-implizite Einschrittverfahren

Wir haben Verfahren konstruiert, die hohe Ordnung haben, und trotzdem A-stabil sind, z.B. das Gauß-Verfahren; es gibt aber noch andere. Diese Verfahren sind implizit. Zum Berechnen des nächsten Zeitschritts muss ein Gleichungssystem gelöst werden.

- Falls f linear ist, so ist dieses Gleichungssystem linear. Das ist okay.
- Falls f nichtlinear ist, so ist Gleichungssysteme ebenfalls nichtlinear. Das kann ganz schön teuer werden!

Können wir A-stabile Verfahren konstruieren, für die bei jedem Schritt nur ein lineares Gleichungssystem gelöst werden muss, selbst wenn f nichtlinear ist?

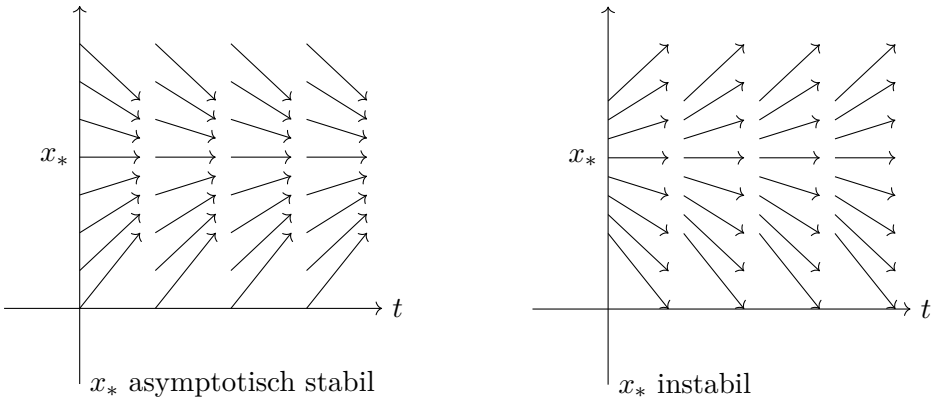
1.7.1 Stabilität von Fixpunkten

Wir wollen einen alternativen Stabilitätsbegriff für autonome nichtlineare Differentialgleichungen $x' = f(x)$ untersuchen.

Definition. Ein Zustand $x_* \in \Omega_0$ heißt Fixpunkt der Gleichung, wenn $f(x_*) = 0$, bzw. wenn $\Phi^t x_* = x_*$ für alle t ist.

Definition. Ein Fixpunkt x_* heißt asymptotisch stabil, wenn ein $\epsilon > 0$ existiert, so dass $\lim_{t\to\infty} \Phi^t x_0 = x_*$ für alle $x_0 \in \Omega_0$ mit $\|x_* - x_0\| < \epsilon$.

Beispiel.



Man erkennt an den Bildern, dass die asymptotische Stabilität von x_* mit der Ableitung von f in (der Nähe von) x_* zusammenhängt.

Satz 1.16 ([DB08, 3.30]). Sei $x_* \in \Omega_0$ Fixpunkt von $x' = f(x)$, und f sei stetig differenzierbar. Falls

$$\nu(Df(x_*)) < 0$$

so ist x_* asymptotisch stabiler Fixpunkt

Erinnerung: ν ist die Spektralabzisse, der größte Realteil aller Eigenwerte.

Zwischenfazit: Um die asymptotische Stabilität von Fixpunkten zu untersuchen, reicht es, sich die Linearisierung um x_* anzuschauen!

Wir betrachten jetzt zusätzlich die um x_* linearisierte Differentialgleichung

$$(x - x_*)' = x' = Df(x_*)(x - x_*). \tag{1.5}$$

Idee. *Wenn $Df(x_*)$ das Stabilitätsverhalten von x_* qualitativ richtig beschreibt, dann enthält die **lineare** Gleichung (1.5) vielleicht schon alle „schwierigen" (im Sinne der Stabilität) Aspekte von $x' = f(x)$ in der Nähe von x_*?*

Betrachte ein beliebiges Einschrittverfahren. Sei

- Ψ^τ diskreter Fluss für das Ausgangsproblem
- Ψ_*^τ diskreter Fluss für das linearisierte Problem $x' = Df(x_*)(x - x_*)$.

Definition. Ein Einschrittverfahren heißt **invariant gegen Linearisierung** um einen Fixpunkt x_*, wenn

1. $\Psi^\tau x_* = x_* \ \forall \tau > 0$ (τ zulässig) $\quad \rightarrow$ der Fixpunkt der Differentialgleichung ist auch Fixpunkt des numerischen Verfahrens für die nichtlineare Gleichung.
2. $\Psi_*^\tau x = x_* + R(\tau Df(x_*))(x - x_*)$ mit einer rationalen Funktion R, die nur vom Verfahren abhängt; d.h. für das linearisierte Problem degeneriert das Verfahren zu einer rationalen Approximation der Exponentialfunktion.
3. $D_x \Psi^\tau x|_{x=x_*} = \Psi_*^\tau$ für alle zulässigen τ $\quad \rightarrow \Psi_*^\tau$ ist Linearisierung von Ψ^τ.

Zum Beispiel sind alle expliziten RK-Verfahren in diesem Sinne invariant. Solch ein Verfahren heißt A-stabil, falls R A-stabil ist.

Invariante Verfahren retten den Zusammenhang zwischen der asymptotischen Stabilität eines Fixpunkts x_* und der Linearisierung dort ins Diskrete:

Satz 1.17 ([DB08, 6.23]). Sei Ψ^τ, Ψ_*^τ ein gegen Linearisierung invariantes Einschrittverfahren. Sei $\tau_c \geq 0$ die maximale Schrittweite, so dass Ψ_*^τ die asymptotische Stabilität erbt. Dann ist x_* asymptotisch stabiler Fixpunkt der Rekursion

$$x_{n+1} = \Psi^\tau x_n \qquad n = 0, 1, 2, \ldots$$

für alle $\tau < \tau_c$.

Beispiel. Skalare Differentialgleichung $x' = \lambda(1 - x^2) \qquad (\lambda > 0)$

- Fixpunkte: $x_s = 1$ (asymptotisch stabil) und $x_u = -1$ instabil
- Linearisierte Gleichung in x_s:

$$x' = f'(x_s)(x - x_s) = -2\lambda x_s(x - x_s) = -2\lambda(x - 1)$$

- Explizites Euler-Verfahren dafür stabil, wenn $\tau < 1/\lambda$
- Es folgt: x_s ist auch asymptotisch stabiler Fixpunkt des expliziten Euler-Verfahrens für die nichtlineare Gleichung

$$x_{n+1} = x_n + \tau f(x_n) = x_n + \tau\lambda(1 - x_n^2).$$

  Aber wie gesagt nur falls $\tau < 1/\lambda$.

Und nicht vergessen: x_s ist nur dann Attraktor, wenn man nah genug dran startet. Für dieses Beispiel heißt das:

- Kontinuierlich: $x_0 > -1$
- Euler: $x_0 \in [0, \frac{5}{4}]$.

1.7.2 Linear-implizite Runge–Kutta-Verfahren

Idee. *Behandle nur den linearen Teil von f implizit.*

Für festes $\bar{x} \in \Omega_0$ schreibe die Differentialgleichung als

$$x'(t) = Jx(t) + (f(x(t)) - Jx(t)) \qquad J = Df(\bar{x}) \in \mathbb{R}^{d\times d}$$

Hier ist $\bar{x}$ beliebig; in der Praxis linearisiert man um den Zustand zum vorigen Zeitschritt.

Wende das implizite Euler-Verfahren auf den ersten Term an, und das explizite Euler-Verfahren auf den Rest.

$$\Psi^\tau x = \xi + \tau(f(x) - Jx), \qquad \xi = x + \tau J\xi$$

Das ist das linear-implizite oder **semi-implizite Euler-Verfahren**. Wir haben nur ein *lineares* Gleichungssystem pro Schritt, aber sind trotzdem A-stabil.

Betrachten wir nun allgemein **linear-implizite Runge-Kutta-Verfahren**

$$\Psi^\tau x = x + \tau \sum_{j=1}^{s} b_j k_j$$

$$k_i = J\Big(x + \tau \sum_{j=1}^{i} \beta_{ij} k_j\Big) + \Big[f\Big(x + \tau \sum_{j=1}^{i-1} \alpha_{ij} k_j\Big) - J\Big(x + \tau \sum_{j=1}^{i-1} \alpha_{ij} k_j\Big)\Big]$$

für $i = 1, \dots, s$.

Hinweis. *Der obere Summationsindex des impliziten Teils ist i, nicht s.*

Dadurch kann der Phasenfluss durch wiederholtes Lösen *linearer* Gleichungssysteme berechnet werden.

1. $J = Df(x)$
2. $(I - \tau\beta_{ii}J)k_i = \tau \sum_{j=1}^{i-1}(\beta_{ij} - \alpha_{ij})Jk_j + f\Big(x + \tau \sum_{j=1}^{i-1} \alpha_{ij} k_j\Big)$ für $i = 1, \dots, s$
3. $\Psi^\tau x = x + \tau \sum_{j=1}^{s} b_j k_j$

Solche Verfahren heißen **lineare-implizite RK-Verfahren** oder **Rosenbrock-Verfahren**.

Koeffizienten: $A = (\alpha_{ij}) \in \mathbb{R}^{s\times s},\ B = (\beta_{ij}) \in \mathbb{R}^{s\times s},\ b = (b_1 \dots, b_s)$

Wählt man die β_{ii} alle gleich, so haben die s Gleichungssysteme in (2) alle die gleiche Matrix und es reicht eine LR-Zerlegung, um alle s Gleichungssysteme zu lösen.

Die Frage, ob sich die linearen Gleichungssysteme tatsächlich immer lösen lassen, ist einfacher als für den allgemeinen impliziten Fall:

Lemma 1.14. Sei $\beta \geq 0$ und $J \in \mathbb{R}^{d\times d}$. Die Matrix $I - \tau\beta J$ ist für alle $0 \leq \tau \leq \tau_*$ invertierbar. Dabei hängt τ_* von der Spektralabzisse $\nu(J)$ ab:

$$\tau_* = \infty \text{ für } \nu(J) \leq 0, \qquad \tau_* = \frac{1}{\beta\nu(J)} \text{ für } \nu(J) > 0.$$

Beweis. Zu zeigen: Unter den gegebenen Voraussetzungen hat $I - \tau\beta J$ nicht den Eigenwert 0. Nach Satz (1.7) über rationale Funktionen ist aber

$$\sigma(I - \tau\beta J) = 1 - \tau\beta\sigma(J).$$

Deshalb zu zeigen: J hat keinen Eigenwert λ mit $1 - \tau\beta\lambda = 0$.

Fall 1: $\nu(J) \leq 0$, d.h. insbesondere $\operatorname{Re}(\lambda) \leq 0$:

$$\operatorname{Re}(1 - \tau\beta\lambda) = 1 - \tau\beta \operatorname{Re}(\lambda) \geq 1 \quad \Rightarrow \quad 1 - \tau\beta\lambda \neq 0$$

Fall 2: $0 < \operatorname{Re}(\lambda) \leq \nu(J)$:

$$\operatorname{Re}(1 - \tau\beta\lambda) = 1 - \tau\beta \operatorname{Re}(\lambda) \geq 1 - \tau\beta\nu(J)$$

Also > 0 wenn $\tau < \frac{1}{\beta\nu(J)}$ □

Der Satz sagt also: Die steifen Anteile der Differentialgleichung (d.h. die nichtpositiven Eigenwert von J) beeinflussen nicht die Lösbarkeit des Gleichungssystems.

Für autonome *lineare* Probleme ist das Verfahren offensichtlich äquivalent zum impliziten Runge-Kutta-Verfahren $(b, (\beta_{ij}))$. Es hat also die selbe Stabilitätsfunktion.

Die Konstruktion der Bedingungsgleichungen funktioniert ähnlich wie bei expliziten RK-Verfahren.

1.8 Erhalt erster Integrale

Betrachte die autonome Differentialgleichung $x' = f(x)$ auf einem Phasenraum Ω_0.

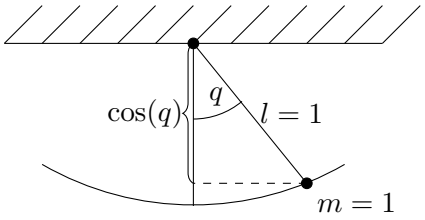
Definition. Eine Funktion $\mathcal{E}\colon \Omega_0 \to \mathbb{R}$ heißt **erstes Integral**, wenn

$$\mathcal{E}(\Phi^t x) = \mathcal{E}(x)$$

für alle $x \in \Omega_0$ und alle zulässigen t gilt.

Alternative Bezeichnungen: Invariante, Erhaltungsgröße, engl.: constant of motion

Beispiel. Mathematisches (Faden-)pendel



Bewegungsgleichungen für Winkel q:

$$\ddot{q} + \frac{g}{l} \sin q = 0$$

bzw. als System erster Ordnung:

$$\dot{p} = -mgl \sin q$$
$$\dot{q} = \frac{1}{ml^2} p$$

Erhält $\mathcal{E}(p, q) = \frac{1}{2} \frac{1}{ml^2} p^2 - mgl \cos q$ (die totale Energie).

Beispiel. Betrachte ein System mit N Partikeln

- $q_i \in \mathbb{R}^3,\ i = 1, \ldots, N$ Positionen, $p_i \in \mathbb{R},\ i = 1, \ldots, N$ Impulse
- m_i: Massen
- Paarweise Interaktion über Kräfte, die vom Abstand abhängen.

Bewegungsgleichungen:

$$q_i' = \frac{p_i}{m_i}, \qquad p_i' = \sum_{j=1}^{N} \nu_{ij}(q_i - q_j)$$

mit

$$\nu_{ij}(y) = -\nu_{ji}(-y)$$

daraus folgt insbesondere dass $\nu_{ii} = 0$.

Die Bewegungsgleichungen erhalten den Gesamtimpuls $P = \sum_{i=1}^{N} p_i$, denn

$$\frac{d}{dt} \sum_{i=1}^{N} p_i = \sum_{i=1}^{N} p_i' = \sum_{i=1}^{N} \sum_{j=1}^{N} \nu_{ij}(q_i - q_j) = 0$$

Ebenso: Der Gesamtdrehimpuls $L = \sum_{i=1}^{N} q_i \times p_i$

$$\begin{aligned} \frac{d}{dt} \sum_{i=1}^{N} q_i \times p_i &= \sum_{i=1}^{N} q_i' \times p_i + \sum_{i=1}^{N} q_i \times p_i' \\ &= \sum_{i=1}^{N} \frac{1}{m_i} \underbrace{p_i \times p_i}_{=0} + \sum_{i=1}^{N} \sum_{j=1}^{N} q_i \times \nu_{ij}(q_i - q_j) \\ &= 0. \end{aligned}$$

Klassifikation der Erhaltungsgrößen: (für diese Beispiele)

- Impuls: linear

- Drehimpuls: quadratisch
- Energie beim Fadenpendel: nichtlinear

Man hätte nun gerne numerische Verfahren, die erste Integrale erhalten.

Zunächst eine einfache Charakterisierung mit Hilfe von f:

Lemma 1.15 ([DB08, 6.56]). Sei f lokal Lipschitz-stetig. Eine Funktion $\mathcal{E} \in C^1(\Omega_0, \mathbb{R})$ ist genau dann erstes Integral, wenn

$$\nabla \mathcal{E}(x) \cdot f(x) = 0$$

für alle $x \in \Omega_0$.

Beweis. Die Kettenregel liefert $0 = \frac{d}{dt}\mathcal{E}(\Phi^t x) = \nabla \mathcal{E}(\Phi^t x) \cdot \frac{d}{dt}\Phi^t x = \nabla \mathcal{E}(\Phi^t x) \cdot f(\Phi^t x)$. □

Beispiel. Wir zeigen Energieerhaltung des Fadenpendels. Für dieses Modell gilt:

$$f(x) = f(p,q) = \begin{pmatrix} -mgl \sin q \\ \frac{p}{ml^2} \end{pmatrix}$$
$$\mathcal{E}(x) = \mathcal{E}(p,q) = \frac{1}{2}\frac{p^2}{ml^2} - mgl \cos q$$

Der Gradient der Energie ist

$$\nabla \mathcal{E}(p,q) = \begin{pmatrix} \frac{p}{ml^2} \\ mgl \sin q \end{pmatrix}$$

Damit erhält man

$$\nabla \mathcal{E}(p,q) \cdot f(p,q) = \frac{p}{ml^2}(-mgl \sin q) + mgl \sin q \frac{p}{ml^2} = 0$$

Satz 1.18 ([HLW16, Thm. IV.1.5]). Alle Runge-Kutta-Verfahren erhalten lineare Invarianten.

Beweis. Sei $\mathcal{E}$ lineare Invariante, also $\mathcal{E}(x) = d^T x$ mit festem Vektor d. Nach dem vorigen Satz ist dann $d^T f(x) = 0$ für alle $x \in \Omega_0$. Für eine Stufe k_i eines beliebigen RK-Verfahrens ist dann

$$d^T k_i = d^T f\Big(x + \tau \sum_{j=1}^{s} a_{ij} k_j\Big) = 0.$$

Also ist

$$\mathcal{E}(x_{k+1}) = d^T x_{k+1} = d^T \Big(x_k + \tau \sum_{i=1}^{s} b_i k_i\Big) = d^T x_k = \mathcal{E}(x_k)$$ □

Für die quadratischen Invarianten betrachten wir zunächst einen wichtigen Spezialfall:

Frage: Für welche linearen autonomen Differentialgleichungen $x' = Ax$ erhält der Phasenfluss Φ^t die Euklidische Norm $\|\Phi^t x\|_2 = \|x\|_2$ für alle t? **Antwort:** Genau dann, wenn $\Phi^t = \exp(tA)$ eine orthogonale Matrix ist.

Satz 1.19 ([DB08, 6.18]). Sei $A \in \mathbb{R}^{d \times d}$. Die Matrix $\exp(tA)$ ist genau dann orthogonal, wenn A schiefsymmetrisch ist.

Beweis. **(⇒)** Sei $\exp(tA) \in O(d)$ für alle t. Dann ist

$$I = \exp(tA)^T \exp(tA) = \exp(tA^T) \exp(tA).$$

Differenziere nach t und betrachte $t = 0$

$$0 = \Big(A^T \exp(tA^T) \exp(tA) + \exp(tA^T) A \exp(tA) \Big) \Big|_{t=0} = A^T + A$$

(⇐)

$$\begin{aligned} I &= \exp(tA - tA) = \exp(tA) \cdot \exp(-tA) \quad \text{(da } A \text{ mit } A \text{ kommutiert)} \\ &= \exp(tA) \cdot \exp(tA^T) \\ &= \exp(tA) \cdot \exp(tA)^T \end{aligned}$$

□

Zentral ist anscheinend die Eigenschaft

$$\exp(z) \cdot \exp(-z) = 1 \qquad \forall z \in \mathbb{C}.$$

Das nennt man **Reversibilität**.

Man hätte diese Eigenschaft gerne auch für diskrete Verfahren.

Definition. Eine diskrete Evolution Ψ heißt reversibel, wenn

$$\Psi^{t,t+\tau} \Psi^{t+\tau,t} x = x$$

für alle $(t, x) \in \Omega$ und hinreichend kleine τ.

Beispiel. Das explizite Euler-Verfahren ist nicht reversibel.

Reversible rationale Approximationen der Exponentialfunktion erzeugen normerhaltende diskrete Flüsse.

Satz 1.20 ([DB08, 6.21]). Sei R eine rationale, konsistente, reversible Approximation der Exponentialfunktion. Dann gilt für eine Matrix $A \in \mathbb{R}^{d \times d}$

$$R(\tau A) \in O(d) \qquad \forall \tau > 0$$

genau dann, wenn $A = -A^T$.

Beweis. Weitestgehend wie bei Satz 1.19. □

Beispiel.

$$R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}} = 1 + z + \frac{z^2}{2} + \frac{z^3}{4} + \mathcal{O}(z^4) = e^z + \mathcal{O}(z^3)$$

- Die entsprechende Matrix-Abbildung heißt **Cayley-Transformation**.
- Stabilitätsfunktion insbesondere der impliziten Mittelpunktsregel → dem einfachsten Gauß-Verfahren.

Gauß-Verfahren erhalten sogar beliebige quadratische Invarianten!

Satz 1.21 ([DB08, 6.58]). Die Differentialgleichung $x' = f(x)$ mit lokal Lipschitzstetigem f besitze das quadratische erste Integral $\mathcal{E}$, d.h.

$$\mathcal{E}(x) = x^T E x + e^T x + \eta$$

mit $E \in \mathbb{R}^{d\times d},\ e \in \mathbb{R}^d,\ \eta \in \mathbb{R}$. Jedes Gaus-Verfahren erzeugt einen Phasenfluss Ψ, der $\mathcal{E}$ erhält, d.h.

$$\mathcal{E}(\Psi^\tau x) = \mathcal{E}(x)$$

für alle $x \in \Omega_0$ und zulässige τ.

Beweis. Ganz ähnlich wie der Beweis der B-Stabilität. Sei $x \in \Omega_0$, und τ so klein, dass das Kollokationspolynom

$$u \in P_s, \quad u(0) = x, \quad u(\tau) = \Psi^\tau x$$

existiert. Da $\mathcal{E}$ quadratisch ist, ist $q(\theta) := \mathcal{E}(u(\theta\tau))$ ein Polynom in P_{2s}. Der Hauptsatz der Integralrechnung liefert

$$\mathcal{E}(\Psi^\tau x) = q(1) = q(0) + \int_0^1 q'(\theta)\,d\theta = \mathcal{E}(x) + \int_0^1 q'(\theta)\,d\theta.$$

Zu zeigen ist nun also $\int_0^1 q'(\theta)\,d\theta = 0$. Nutze die Quadraturformel des Gauß-Verfahrens. Diese ist für Polynome in P_{2s-1} exakt:

$$\int_0^1 q'(\theta)\,d\theta = \sum_{j=1}^{s} b_j q'(c_j).$$

Es sind aber alle $q'(c_j) = 0$, denn

$$\begin{aligned}
q'(c_j) &= \big(\mathcal{E}(u(c_j\tau))\big)' \\
&= \tau\nabla\mathcal{E}(u(c_j\tau)) \cdot u'(c_j\tau) && \text{(Kettenregel)} \\
&= \tau\nabla\mathcal{E}(u(c_j\tau)) \cdot f(u(c_j\tau)) && \text{(Kollokationseigenschaft)} \\
&= 0 && \text{(da } \mathcal{E} \text{ eine Invariante ist)}
\end{aligned}$$

□

Was ist mit der Energieerhaltung des Fadenpendels? → Das behandeln wir später mit der Theorie der Hamiltonschen Systeme.

2 Numerik von Hamilton-Systemen

2.1 Hamilton-Systeme

Extrem wichtige Klasse von Differentialgleichungen entstammen der klassischen Mechanik, Quantenmechanik und relativistische Mechanik. Dazu gehören auch spezielle numerische Verfahren — eine „schöne Mathematik“.

„Vereinigendes Prinzip“: Bringt ganz unterschiedliche Gleichungen auf eine gemeinsame Form.

Beispiel. Mathematisches Pendel – Fadenpendel

- Koordinate: Winkel α
- Masse m, Fadenlänge l, Erdbeschleunigung g

Bewegungsgleichungen: $\ddot{\alpha} + \frac{g}{l} \sin \alpha = 0$

Beispiel. Teilchen in einem Kraftfeld $F(x)$

$$m\ddot{x} = F(x) \qquad \text{(Newtons Gesetz)}$$

Beispiel. 1d-Wellengleichung — Longitudinale Auslenkung einer elastischen Schnur

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 u}{\partial t^2} && x \in [a,b], t \geq 0 \\ u(a,t) = u(b,t) &= 0 && \forall t \geq 0 \end{aligned}$$

2.1.1 Die Lagrange-Gleichungen

Wir betrachten ein mechanisches System mit d Freiheitsgraden $q = (q_1, \ldots, q_d)$.

- Kinetische Energie: $T = T(q, \dot{q})$ (häufig: $T(q, \dot{q}) = \frac{1}{2}\dot{q}^T M(q)\dot{q}$ mit $M(q)$ s.p.d.)
- Potentielle Energie: $U = U(q)$

Definition. Die Lagrange-Funktion eines mechanischen Systems ist $L = T - U$.

Das mechanische System löst die Lagrange-Gleichungen

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial L}{\partial q}$$

Warum? Es gilt das Prinzip der stationären Wirkung.

Definition (Prinzip der stationären Wirkung / Hamilton'sches Prinzip). Sei $q : [t_0, t_1] \to \mathbb{R}^d$ eine Trajektorie eines mechanischen Systems. Für die in der Natur vorkommenden Trajektorien ist die *Wirkung*

$$W := \int_{t_0}^{t_1} L(q(t)\ \dot{q}(t))\ \mathrm{d}t$$

stationär.

Sei q eine Trajektorie, und δq eine Variation davon, die die Endpunkte fest lässt, also $\delta q(t_0) = \delta q(t_1) = 0$. Stationarität von q heißt dann, dass für alle solche δq

$$\frac{d}{d\epsilon} S(q + \epsilon\delta q)|_{\epsilon=0} = 0.$$

gilt. Ausrechnen:

$$\begin{aligned}
\frac{d}{d\epsilon}\int_{t_0}^{t_1} L(q+\epsilon\delta q, \dot{q}+\epsilon\delta\dot{q})\ \mathrm{d}t|_{\epsilon=0} &= \int_{t_0}^{t_1} \frac{\partial L}{\partial q}\delta q + \frac{\partial L}{\partial \dot{q}}\delta\dot{q}\, dt \\
&= \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial q}\delta q - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\delta q\right)\ \mathrm{d}t \\
&\qquad\qquad \text{(partielle Integration)} \\
&= \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\right)\delta q\ \mathrm{d}t
\end{aligned}$$

Da dieser Ausdruck für alle hinreichend glatten Funktionen δq gleich Null sein muss, erhält man die Lagrange-Gleichung

$$\delta W = 0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q}$$

Beispiel (Pendel).

- Kinetische Energie

$$T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) = \frac{1}{2}ml^2\dot{\alpha}^2$$

- Potentielle Energie

$$U = mgy = -mgl\cos\alpha$$

- Lagrange-Funktion

$$L(\alpha, \dot{\alpha}) = \frac{1}{2}ml^2\dot{\alpha}^2 + mgl\cos\alpha$$

- Lagrange-Gleichung

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = \frac{d}{dt}(ml^2\dot{\alpha}) + mgl\sin\alpha = ml^2\ddot{\alpha} + mgl\sin\alpha.$$

Beispiel (Teilchen in einem Kraftfeld). Angenommen das Kraftfeld ist **konservativ**, d.h. es gibt ein $U : \mathbb{R}^3 \to \mathbb{R}$, so dass $F(x) = -\nabla U(x)$.

- Kinetische Energie

$$T(x, \dot{x}) = \frac{1}{2}m\langle\dot{x}, \dot{x}\rangle$$

- Potentielle Energie

$$U$$

- Lagrange-Gleichung

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = \frac{d}{dt}(m\dot{x}) + \nabla U(x) = m\ddot{x} - F(x)$$

Beispiel (Eindimensionale Wellengleichung). Ein unendlich-dimensionales System wird nicht beschrieben durch d Freiheitsgrade $(q_1, \ldots, q_d)$, sondern durch die Funktion $u : [a, b] \to \mathbb{R}$. Diese beschreibt die transversale Auslenkung einer Saite.

- Kinetische Energie

$$T(u, \dot{u}) = \frac{1}{2}\int_a^b m\dot{u}(x)^2 dx$$

  Dabei ist m die Massendichte.

- Potentielle Energie

$$U(u) = \int_a^b S\Big[\sqrt{1 + u'(x)^2} - 1\Big]dx \approx \int_a^b S\frac{u'(x)^2}{2}dx$$

  Dabei ist S die Zugsteifigkeit.

- Lagrange-Funktion

$$L(u, \dot{u}) = T(u, \dot{u}) - U(u)$$

- Lagrange-Gleichung

$$\frac{\partial L}{\partial u} = \frac{\partial}{\partial x}\frac{\partial L}{\partial u'} + \frac{\partial}{\partial t}\frac{\partial L}{\partial \dot{u}}$$

  Einsetzen:

$$0 = \frac{\partial}{\partial x}\Big(-Su'(x)\Big) + \frac{\partial}{\partial t}m\dot{u}$$

  Umstellen:

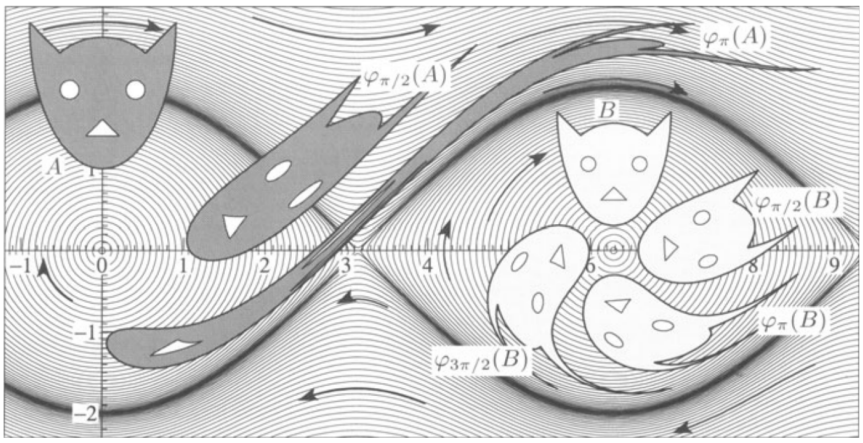
$$\frac{\partial^2 u}{\partial t^2} = \frac{S}{m}\frac{\partial^2 u}{\partial x^2}$$

  Das ist die eindimensionale Wellengleichung.

2.2 Symplektizität

Flüsse von Hamiltonschen Systemen haben eine weitere wichtige Erhaltungseigenschaft, die sogenannte **Symplektizität**, die ähnlich wie Volumenerhaltung im Phasenraum vorstellbar ist.

Beispiel. Volumenerhaltung beim mathematischen Pendel (Bild aus [HLW16]):



Betrachte die Hamiltonschen Gleichungen

$$\dot{p} = -\frac{\partial H}{\partial q}(p,q), \quad \dot{q} = \frac{\partial H}{\partial p}(p,q)$$

Umschreiben:

$$\begin{pmatrix} \dot{p} \\ \dot{q} \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial H}{\partial p} \\ \frac{\partial H}{\partial q} \end{pmatrix}$$

Diese Beziehung wollen wir jetzt abstrakter betrachten.

Bemerkung. *Die harmlos aussehende Matrix* $\left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$ *hat eine besondere Eigenschaft. Es gilt nämlich*

$$\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}^2 = -\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

Sie verhält sich also wie die imaginäre Einheit i und erzeugt damit eine komplexe Struktur auf $\mathbb{R}^{2d}$.

Wir betrachten 2-dimensionale Parallelogramme in $\mathbb{R}^{2d}$ aufgespannt durch Vektoren

$$\xi = \begin{pmatrix} \xi^p \\ \xi^q \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}.$$

Hier und im Folgenden bezeichnen $\xi^p \in \mathbb{R}^d$ und $\xi^q \in \mathbb{R}^d$ die Impuls- bzw. Ortskomponenten von ξ.

Falls $d = 1$, so ist die orientierte Fläche des Parallelogramms gerade

$$\det \begin{pmatrix} \xi^p & \eta^p \\ \xi^q & \eta^q \end{pmatrix} = \xi^p \eta^q - \xi^q \eta^p = \begin{pmatrix} \xi^p & \xi^q \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}.$$

Das verallgemeinern wir jetzt für höhere Dimensionen.

Definition (Symplektische Form). Die symplektische Form $\omega : \mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}$ ist

$$\omega(\xi,\eta) := \sum_{i=1}^{d} \det \begin{pmatrix} \xi_i^p & \eta_i^p \\ \xi_i^q & \eta_i^q \end{pmatrix} = \sum_{i=1}^{d} (\xi_i^p \eta_i^q - \xi_i^q \eta_i^p).$$

- Bilineare Form
- Interpretation: Summe der orientierten Flächen der Projektionen auf die Koordinatenebenen (p_i, q_i).
- Matrixdarstellung

$$\omega(\xi,\eta) = \begin{pmatrix} \xi^{pT} & \xi^{qT} \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}.$$

Da die Matrix $\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$ wichtig zu sein scheint geben wir ihr den Namen J.

Eine wichtige Eigenschaft von Hamiltonschen Systemen ist nun, dass ihre Flüsse $\Phi^t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ die symplektische Form erhalten. Das muss man natürlich erklären.

Definition (Lineare symplektische Abbildung). Eine lineare Abbildung $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ heißt **symplektisch**, wenn

$$\omega(A\xi, A\eta) = \omega(\xi, \eta) \qquad \forall \xi, \eta \in \mathbb{R}^{2d}.$$

Alternativ: wenn $A^T J A = J$.

Bemerkung. *Für $d = 1$ bedeutet das gerade, dass A flächenerhaltend ist.*

Definition (Differenzierbare symplektische Abbildung). Sei U eine offene Teilmenge von $\mathbb{R}^{2d}$. Eine differenzierbare Abbildung $g : U \to \mathbb{R}^{2d}$ heißt **symplektisch**, falls die Jacobi-Matrix $\nabla g(p, q)$ für alle $(p, q) \in U$ symplektisch ist.

Jetzt kommt der zentrale Satz: die Flüsse Φ^t von Hamiltonschen Systemen erhalten die symplektische Form:

Satz 2.1 (Poincaré, 1899). Sei $H(p, q)$ zweimal stetig differenzierbar auf $U \subset \mathbb{R}^{2d}$. Sei Φ^t der Phasenfluss der Differentialgleichung

$$\dot{y} = J^{-1} \nabla H(y)$$

mit $y = (p, q)$.

Für jedes feste t ist Φ^t eine symplektische Abbildung.

Beweis. Der Beweis erfolgt in zwei Schritten:

1. Φ^0 ist symplektisch.
2. Die „Abweichung von der Symplektizität“ hängt nicht von t ab.

(zu 1) Φ^0 ist symplektisch, wenn seine erste Ableitung an jedem Punkt $y_0 = (p_0, q_0)$ symplektisch ist. Da $\Phi^0 y_0 = y_0$ gilt

$$\left(\frac{\partial \Phi^0 y_0}{\partial y_0}\right)^T J \left(\frac{\partial \Phi^0 y_0}{\partial y_0}\right) = I^T J I = J$$

Also ist Φ^0 symplektisch.

(zu 2) Wir müssen die Ableitung $\frac{\partial \Phi^t y_0}{\partial y_0}$ untersuchen, d.h. die linearisierte Störung der Lösung bei einer Störung des Startwerts – also gerade die Wronski-Matrix Ξ. Diese löst die Gleichung

$$\dot{\Xi} = J^{-1} \underbrace{\nabla^2 H(\Phi^t(y_0))}_{\text{Hesse-Matrix von } H} \Xi$$

Konkret heißt das hier

$$\frac{d}{dt} \frac{\partial \Phi^t}{\partial y_0} = J^{-1} \nabla^2 H(\Phi^t y_0) \frac{\partial \Phi^t}{\partial y_0} \tag{2.1}$$

Die Produktregel liefert uns

$$\frac{d}{dt}\left[\left(\frac{\partial \Phi^t}{\partial y_0}\right)^T J \left(\frac{\partial \Phi^t}{\partial y_0}\right)\right] = \left(\frac{d}{dt}\frac{\partial \Phi^t}{\partial y_0}\right)^T J \left(\frac{\partial \Phi^t}{\partial y_0}\right) + \left(\frac{\partial \Phi^t}{\partial y_0}\right)^T J \left(\frac{d}{dt}\frac{\partial \Phi^t}{\partial y_0}\right)$$

Dort wird jetzt (2.1) eingesetzt:

$$\frac{d}{dt}\left[\left(\frac{\partial\Phi^t}{\partial y_0}\right)^T J\left(\frac{\partial\Phi^t}{\partial y_0}\right)\right] = \left(\frac{\partial\Phi^t}{\partial y_0}\right)^T \nabla^2 H(\Phi^t y_0)^T J^{-T} J\left(\frac{\partial\Phi^t}{\partial y_0}\right) + \left(\frac{\partial\Phi^t}{\partial y_0}\right)^T J J^{-1} \nabla^2 H(\Phi^t y_0)\left(\frac{\partial\Phi^t}{\partial y_0}\right)$$

Aber $J^T = -J$, also $J^{-T}J = -I$, und $\nabla^2 H$ ist symmetrisch. Deshalb ist

$$\frac{d}{dt}\left[\left(\frac{\partial\Phi^t}{\partial y_0}\right)^T J\left(\frac{\partial\Phi^t}{\partial y_0}\right)\right] = 0 \qquad \square$$

Es gilt sogar die Umkehrung des Satzes: *nur* Hamiltonsche Systeme haben symplektische Flüsse!

Definition (lokal Hamiltonsch). Eine Differentialgleichung $x' = f(x)$ heißt **lokal Hamiltonsch**, wenn für jedes $x_0 \in U$ eine Umgebung existiert, in der

$$f(x) = J^{-1}\nabla H(x)$$

für eine Funktion H.

Satz 2.2 ([HLW16, Satz VI.2.6]). Sei $f\colon U \to \mathbb{R}^{2d}$ stetig differenzierbar. Dann ist $x' = f(x)$ genau dann lokal Hamiltonsch, wenn der Fluss $\Phi^t x$ für alle $x \in U$ und alle t hinreichend klein symplektisch ist.

2.3 Symplektische Verfahren

Wir wollen Verfahren entwickeln, die die Symplektizität von Hamiltonschen Flüssen erben.

Definition. Ein Einschrittverfahren heißt symplektisch, falls der diskrete Fluss

$$\Psi^t\colon \mathbb{R}^{2d} \to \mathbb{R}^{2d}$$

symplektisch ist, wenn das Verfahren auf ein Hamiltonsches System angewendet wird.

Die einfachsten symplektischen Verfahren sind die symplektischen Euler-Verfahren

$$\begin{aligned} p_{k+1} &= p_k - \tau H_q(p_{k+1}, q_k) \\ q_{k+1} &= q_k + \tau H_p(p_{k+1}, q_k) \end{aligned}$$

und

$$\begin{aligned} p_{k+1} &= p_k - \tau H_q(p_k, q_{k+1}) \\ q_{k+1} &= q_k + \tau H_p(p_k, q_{k+1}). \end{aligned}$$

Satz 2.3 ([HLW16, Satz VI.3.3]). Die symplektischen Euler-Verfahren sind symplektisch.

Beweis. Beweis für die erste Methode:

- Methode ist symplektisch, wenn

$$\frac{\partial \Psi^\tau y}{\partial y} \in \mathbb{R}^{2d \times 2d}$$

für alle $y = (p, q)$ die symplektisch Form erhält, wenn also

$$\Big(\frac{\partial \Psi^\tau y}{\partial y}\Big)^T J \Big(\frac{\partial \Psi^\tau y}{\partial y}\Big) = J. \tag{2.2}$$

- Wir bestimmen die vier Komponenten von $\frac{\partial \Psi^\tau y}{\partial y}$:

1) Erste Gleichung des Verfahrens:

$$p_{k+1} = p_k - \tau H_q(p_{k+1}, q_k)$$

Ableiten nach p_k:

$$\frac{\partial p_{k+1}}{\partial p_k} = I - \tau H_{qp}(p_{k+1}, q_k) \cdot \frac{\partial p_{k+1}}{\partial p_k}$$

$\Leftrightarrow$

$$\frac{\partial p_{k+1}}{\partial p_k}\Big(I + \tau H_{qp}\Big) = I$$

Ebenso z.B.

$$\frac{\partial p_{k+1}}{\partial q_k}\Big(I + \tau H_{qp}\Big) = -\tau H_{qq}$$

etc.

Zusammen erhält man

$$\begin{pmatrix} I + \tau H_{qp}^T & 0 \\ -\tau H_{pp} & I \end{pmatrix} \begin{pmatrix} \frac{\partial p_{k+1}}{\partial p_k} & \frac{\partial p_{k+1}}{\partial q_k} \\ \frac{\partial q_{k+1}}{\partial p_k} & \frac{\partial q_{k+1}}{\partial q_k} \end{pmatrix} = \begin{pmatrix} I & -\tau H_{qq} \\ 0 & I + \tau H_{qp} \end{pmatrix},$$

also

$$\frac{\partial \Psi^\tau y}{\partial y} = \begin{pmatrix} I + \tau H_{qp}^T & 0 \\ -\tau H_{pp} & I \end{pmatrix}^{-1} \begin{pmatrix} I & -\tau H_{qq} \\ 0 & I + \tau H_{qp} \end{pmatrix}.$$

Damit kann man die Erhaltungseigenschaft (2.2) direkt nachrechnen. □

Die symplektischen Euler-Verfahren sind **keine** RK-Verfahren. Stattdessen gehören sie zu den sog. **partitionierten** RK-Verfahren. Betrachte Differentialgleichungen der Form

$$y' = f(y, z), \qquad z' = g(y, z),$$

wobei $y \in \mathbb{R}^{n_1}$ und $z \in \mathbb{R}^{n_2}$.

Idee: Nimm für y und z zwei verschiedene RK-Verfahren. Details bei [HLW16, Kapitel II.2]

Es gibt auch ein „einfaches“ Verfahren zweiter Ordnung, das symplektisch ist.

Satz 2.4 ([HLW16, Satz VI.3.5]). Die implizite Mittelpunktsregel

$$y_{k+1} = y_k + \tau J^{-1} \nabla H\Big(\frac{y_{k+1} + y_k}{2}\Big)$$

ist symplektisch.

Beweis. Wir leiten wieder ab

$$\begin{aligned}\frac{\partial \Psi^\tau y_k}{\partial y_k} &= \frac{\partial y_{k+1}}{\partial y_k} \\ &= I + \tau J^{-1}\nabla^2 H\Big(\frac{y_{k+1}+y_k}{2}\Big)\cdot\Big(\frac{1}{2}\frac{\partial y_{k+1}}{y_k}+\frac{1}{2}\Big).\end{aligned}$$

Umformen ergibt

$$\frac{\partial y_{k+1}}{\partial y_k} = \Big(I - \frac{\tau}{2}J^{-1}\nabla^2 H\Big)^{-1}\Big(I+\frac{\tau}{2}J^{-1}\nabla^2 H\Big).$$

Dann kann man direkt nachrechnen dass $\left(\frac{\partial y_{k+1}}{\partial y_k}\right)^T J \frac{\partial y_{k+1}}{\partial y_k} = J$.

Ein paar Details zu dieser Rechnung einfügen!

2.3.1 Symplektische RK-Verfahren

Dabei handelt es sich um relativ neue Verfahren, die erst Ende der 1980er Jahre systematisch untersucht wurden.

Wir interessieren uns wieder für die Ableitung

$$\Xi(t) = \frac{\partial \Phi^t y_0}{\partial y_0}.$$

Diese löst bekanntlich eine lineare Differentialgleichung.

Lemma 2.1 ([HLW16, Lemma VI.4.1]). Das folgende Diagramm kommutiert für alle Runge-Kutta-Verfahren und alle partitionierten Runge-Kutta-Verfahren:

$$\begin{array}{ccc} \dot y = f(y),\ y(0)=y_0 & \xrightarrow{\frac{\partial}{\partial y_0}} & \begin{array}{l}\dot y = f(y), \quad y(0)=y_0 \\ \dot\Xi = f'(y)\Xi, \quad \Xi(0)=I\end{array} \\ \text{RK-Verfahren}\downarrow & & \downarrow\text{RK-Verfahren} \\ \{y_k\} & \xrightarrow{\frac{\partial}{\partial y_0}} & \{y_k,\Xi_k\} \end{array}$$

Beweis (Beweisidee:). Betrachte exemplarisch das explizite Euler-Verfahren

$$y_{k+1} = y_k + \tau f(y_k).$$

Ableiten nach y_0 ergibt

$$\Xi_{k+1} = \Xi_k + \tau f'(y_k)\Xi_k.$$

Das ist gerade das explizite Euler-Verfahren für die Gleichung

$$\dot\Xi = f'(y)\Xi.$$

Startwert passt auch, denn $I = \frac{\partial y_0}{\partial y_0} = \Xi_0$. □

Idee. *Die Symplektizitätsbedingung ist eine quadratische Invariante des erweiterten Systems für die Variablen y und Ξ.*

Satz 2.5. Alle Verfahren die quadratische Invarianten erhalten sind symplektisch.

Beweis. Der quadratische Ausdruck $\Xi^T J \Xi$ is Invariante der Gleichung

$$\dot{\Xi} = J^{-1}\nabla^2 H(y)\Xi,$$

denn

$$\begin{aligned}\frac{d}{dt}\left(\Xi^T J \Xi\right) &= \dot{\Xi}^T J \Xi + \Xi^T J \dot{\Xi} \\ &= (J^{-1}\nabla^2 H \Xi)^T J \Xi + \Xi^T J J^{-1} \nabla^2 H \Xi \\ &= \Xi^T \nabla^2 H \underbrace{J^{-T} J}_{=-I} + \Xi^T \underbrace{J J^{-1}}_{=I} \nabla^2 H \Xi \\ &= 0\end{aligned}$$

□

Korollar. Gauß-Verfahren sind symplektisch.

2.3.2 Reversibilität vs. Symplektizität

Es gibt reversible Verfahren, die nicht symplektisch sind. Es gibt symplektische Verfahren, die nicht reversibel sind.

Für quadratische Hamilton-Funktionen ist das aber anders.

Satz 2.6 ([HLW16, Satz VI.4.9.]). Für RK-Verfahren sind die folgenden Aussagen äquivalent:

(i) Die Methode ist reversibel für lineare Probleme

$$\dot{y} = Ly$$

(ii) Die Methode ist symplektisch für Hamilton-Gleichungen mit quadratischer Hamilton-Funktion

$$H(y) = \frac{1}{2} y^T C y. \qquad C \text{ s.p.d.}$$

Beweis. $ii) \to i)$

- Die Hamilton-Gleichungen haben die Form

  $$\dot{y} = J^{-1}\nabla H(y) = J^{-1} C y,$$

  sind also linear.
- Das Runge-Kutta-Verfahren dafür hat also die Form

  $$\Psi^\tau y = R(\tau J^{-1} C) y$$

  wobei R die Stabilitätsfunktion ist.
- Da das Verfahren symplektisch ist, gilt

  $$R(\tau J^{-1} C)^T J R(\tau J^{-1} C) = J.$$

- Da $R = PQ^{-1}$ für Polynome P, Q erhält man

$$P(\tau J^{-1}C)^T J P(\tau J^{-1}C) = Q(\tau J^{-1}C)^T J Q(\tau J^{-1}C). \tag{2.3}$$

- Betrachte das Produkt „Polynom in $J^{-1}C$“ mit J.
- Für jedes Monom $(J^{-1}C)^k$, $k \in \mathbb{N}$ gilt (C ist symmetrisch, und $J^T = -J$)

$$\begin{aligned}
((J^{-1}C)^k)^T J &= (C^T J^{-T})^k J \\
&= \underbrace{C^T J^{-T} \dots C^T J^{-T}}_{k \text{ mal}} J \\
&= -C^T \underbrace{J^{-T} C^T \dots J^{-T} C^T}_{k-1 \text{ mal}} \\
&= -C \underbrace{J^{-T} C \dots J^{-T} C}_{k-1 \text{ mal}} \\
&= -J^T J^{-T} C \underbrace{J^{-T} C \dots J^{-T} C}_{k-1 \text{ mal}} \\
&= -J^T (J^{-T} C)^k \\
&= J(-J^{-1}C)^k.
\end{aligned}$$

- Also folgt aus (2.3)

$$P(-\tau J^{-1}C) \cdot P(\tau J^{-1}C) = Q(-\tau J^{-1}C) \cdot Q(\tau J^{-1}C)$$

  bzw.

$$R(-\tau J^{-1}C) \cdot R(\tau J^{-1}C) = I.$$

- Das ist gerade die Reversibilität des Verfahrens. □

2.4 Energieerhaltung

Wir haben einige Mühe in Verständnis und Erhaltung der Symplektizität gesteckt. Aber Symplektizität ist eine sehr abstrakte Eigenschaft. Wozu soll die gut sein? Hier komme eine etwas konkretere Rechtfertigung.

Betrachte das mathematische Pendel.

- Kinetische Energie:

$$T(q, \dot{q}) = \frac{ml^2}{2}\dot{q}$$

  (q ist der Winkel)

- Potentielle Energie:

$$U(q) = -mgl \cos q$$

- Bewegungsgleichungen:

$$\ddot{q} + \frac{g}{l} \sin q = 0$$

- Gesamtenergie:
$$E = \frac{ml^2}{2}\dot{q}^2 - mgl\cos q$$
Dies entspricht der Hamilton-Funktion
$$H(p,q) = \frac{1}{2ml^2}p^2 - mgl\cos q$$
Damit ist die Gesamtenergie eine Erhaltungsgröße!
- Aber: weder linear noch quadratisch. Wird daher nicht automatisch von z.B. Gauß-Verfahren erhalten.

Wird die Energie von symplektischen Verfahren erhalten?

Nein! Aber fast...

Satz 2.7 ([BG94]; [HLW16, Thm. IX.8.1]). Betrachte ein Hamilton-System mit analytischer Hamilton-Funktion $H : D \to R$ ($D \subset \mathbb{R}^{2d}$) und wende ein symplektisches Verfahren Ψ^τ mit Schrittweite τ an. Wenn die numerische Lösung in einer kompakten Menge $K \subset D$ bleibt, dann existiert ein τ_0, so dass
$$H(y_n) = H(y_0) + O(\tau^p)$$
für exponentiell lange Zeitintervalle $n\tau \leq e^{\frac{\tau_0}{2\tau}}$.

Symplektische Verfahren erhalten also **nicht** die Hamilton-Funktion bzw. die Gesamtenergie. Aber die numerische Energie bleibt „in der Nähe“ der exakten Energie!

2.5 Variationelle Integratoren

Mit dem jetzt Gelernten können wir Zeitschrittverfahren auf eine ganz neue Art konstruieren. Siehe dazu [MW01] für eine detailliertere Übersicht.

Wir erinnern an das Prinzip der stationären Wirkung (auch Hamiltonsches Prinzip genannt): Lagrange-Funktion
$$L(q,\dot{q}) = T(q,\dot{q}) - U(q)$$

Definition. Die **Wirkung** eine Trajektorie $q\colon t \mapsto (q(t),\dot{q}(t))$ ist
$$S(q) \coloneqq \int_{t_0}^{t_1} L(q(t),\dot{q}(t))\,dt.$$

Wir betrachten nur Trajektorien mit gegebenem festen Start- und Endpunkt
$$q(t_0) = q_0, \qquad q(t_1) = q_1.$$

Definition (Hamiltonsches Prinzip). Die tatsächlich vorkommenden Trajektorien sind die, die die Wirkung stationär machen.

Sei q eine Trajektorie, und δq eine Variation davon, die die Endpunkte fest lässt, also $\delta q(t_0) = \delta(t_1) = 0$. Stationarität von q heißt dann, dass für alle solche δq

$$\frac{d}{d\epsilon} S(q + \epsilon\delta q)\big|_{\epsilon=0} = 0.$$

Wie schon in Kapitel 2.1.1 gezeigt ist dies äquivalent zur Euler–Lagrange-Gleichung

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} = \frac{\partial L}{\partial q}.$$

Wir betrachten jetzt das Wirkungsintegral S als Funktion der Start- und Endposition

$$S(q_0, q_1) = \int_{t_0}^{t_1} L(q(t), \dot{q}(t))\, dt.$$

Dabei ist q die zu q_0, q_1 gehörige Lösung der Lagrange-Gleichung.

Exkurs Anfang: Erzeugendenfunktionen

Wir brauchen ein weiteres Kriterium für Symplektizität:

- Betrachte ein gegebenes Hamiltonsches System H auf einem festen Zeitintervall $[t_0, t_1]$
- Seien $p_0 \in \mathbb{R}^d$ und $q_0 \in \mathbb{R}^d$ die Startwerte zur Zeit t_0
- Bezeichne die Werte zur Zeit t_1 mit $p_1 \in \mathbb{R}^d$ und $q_1 \in \mathbb{R}^d$
- Es gibt eine Abbildung $\Phi^{t_0,t_1}(p_0, q_0) = (p_1, q_1)$.

Wie wir wissen, ist diese symplektisch. [Achtung: der folgende Satz enthält überdurchschnittlich viel didaktische Reduktion]

Satz 2.8 ([HLW16, Satz VI.5.1]). Eine Abbildung $\varphi\colon (p_0, q_0) \mapsto (p_1, q_1)$ ist genau dann symplektisch, wenn lokal eine Funktion

$$S\colon (q_0, q_1) \mapsto S(q_0, q_1) \in \mathbb{R}$$

existiert, so dass

$$\nabla S0 \begin{pmatrix} \frac{\partial S}{\partial q_0} \\ \frac{\partial S}{\partial q_1} \end{pmatrix} = \begin{pmatrix} -p_0 \\ p_1 \end{pmatrix} \tag{2.4}$$

Wenn man eine symplektische Abbildung $(p_0, q_0) \mapsto (p_1, q_1)$ hat, dann kann sie durch (2.4) aus der Funktion S rekonstruiert werden.

Aber der obige Satz ist vom Typ „Äquivalenz“. Es gilt also auch die Umkehrung: Jede hinreichen glatte (und in einem gewissen Sinne nicht degenerierte) Funktion S **erzeugt** via (2.4) eine symplektische Abbildung $(p_0, q_0) \mapsto (p_1, q_1)$.

Man kann also auf systematische Art symplektische Abbildungen erzeugen. Die Funktion S heißt deshalb Erzeugendenfunktion.

Exkurs Ende

H ist nicht ganz konstant, aber die Abweichung ist seeeehr klein - nämlich in $\mathcal{O}(\tau^p)$

Literaturverzeichnis

[BG94] Giancarlo Benettin and Antonio Giorgilli. On the hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms. *Journal of Statistical Physics*, 74:1117–1143, March 1994.

[DB08] Peter Deuflhard and Folkmar Bornemann. *Numerische Mathematik 2 – Gewöhnliche Differentialgleichungen*. de Gruyter, 2008.

[Die60] Jean Dieudonné. *Foundations of Modern Analysis*. Academic Press, 1960.

[HLW16] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration—Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, zweite auflage edition, 2016.

[MW01] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001.