Analysis und Numerik von gewöhnlichen Differentialgleichungen

Frank Wübbeling

Stand: 7. August 2022

Inhaltsverzeichnis

Kapitel -1

Vorwort

Der vorliegende Text entsteht als Begleitmaterial zur Vorlesung Analysis und Numerik gewöhnlicher Differentialgleichungen im Sommersemester 2022. Die Vorlesung richtet sich an Studierende des Bachelorstudiengangs Mathematik im dritten Semester sowie Studierende in den Lehramtsstudiengängen Mathematik. Für die Korrektheit des Textes wird keinerlei Garantie übernommen, vermutlich sind noch reichlich Schreibfehler enthalten. Für Bemerkungen und Korrekturen bin ich dankbar.

Da der Großteil der Studierenden heute kaum noch einen physikalischen Hintergrund hat, habe ich auf die Darstellung der Beziehungen zwischen Angewandter Mathematik und Physik, wie sie in den klassischen Lehrbüchern und Vorlesungen üblich war, größtenteils verzichtet. Übungen zu den einzelnen Kapiteln finden sich im Netz, ebenso eine (subjektive) Literaturliste. Klassische Hintergrundliteratur für den analytischen Teil ist Walter [2013], für den numerischen Teil Bulirsch and Stoer [1966] (alle Bücher der Reihe).

Ich habe mich bemüht, zu den vorgestellten Algorithmen eine Beispiel–Implementation in Python zu liefern. Einige Programme sind dabei in der PDF–Datei enthalten. Klick auf die jeweilige Textstelle öffnet die Beispielimplementation. Ebenso sind alle Bilder beigefügt, Klick liefert jeweils das zugehörige Bild. Dies funktioniert in

Acrobat (Reader) und in einigen anderen PDF–Readern, in vielen PublicDomain–Readern aber nicht.

Billerbeck, im Sommer 2022

Frank Wübbeling

Kapitel 0

Angewandte Mathematik

Dieses Kapitel dient der Motivation - es ist nicht Teil des Prüfungsstoffs.

Angewandte Mathematik überträgt mathematische Konzepte auf die Realität und macht sie so für praktische Probleme nutzbar. Das Zusammenspiel zwischen Theorie (Mathematik) und Praxis (Anwendung) ist der Reiz dieser Disziplin und bildet die Grundlage für die modernen Naturwissenschaften (Physik, Biologie, Chemie, Geophysik, Medizin, ...), aber auch für andere Gebiete wie die analytischen Wirtschaftswissenschaften.

Wichtige Aufgaben sind dabei

Simulation: Ein Prozess wird auf dem Rechner nachgebildet, z.B. in der Wettervorhersage oder der Klimaforschung.

Optimierung: Es werden optimale Parameter für einen beeinflussbaren Prozess gesucht, z.B. in der Produktionsplanung oder der Strahlentherapie.

Ursachenforschung: Ermittlung von Größen, die nur indirekt gemessen werden können, z.B. in der Tomographie oder beim Scharfrechnen von geglätteten Bildern.

Die Vorgehensweise bei der Lösung eines Problems mit Hilfe der Mathematik ist:

1. Genaue Formulierung des Problems
2. Übertragung in die Mathematik (Modellierung)
3. Vereinfachung des Modells, so dass es lösbar wird (Diskretisierung)
4. Design eines Lösungswegs (Algorithmus)

5. Implementation des Algorithmus
6. Interpretation
7. Anwendung

Die numerische Mathematik ist dabei für die Schritte 3-5 zuständig. Jeder der mathematischen Schritte unterliegt dabei einer Untersuchung mit analytischen Methoden:

- Wie genau ist das Modell?
- Wie genau ist das vereinfachte Modell?
- Liefert der Algorithmus das richtige Ergebnis?
- Was passiert bei Messfehlern?
- Wie effizient (schnell) ist der Algorithmus?

Wir betrachten einige Beispiele.

Beispiel 0.1 *(Computertomographie)*
Ein Röntgenbild zeigt immer zweidimensionale Schattenbilder. Eine dreidimensionale Lagebeziehung (etwa: liegt der Tumor vor oder hinter dem Knochen?) kann man den Bildern nicht entnehmen. Mitte des letzten Jahrhunderts kam die Idee auf, viele Röntgenbilder aufzunehmen und daraus eine dreidimensional Darstellung zu berechnen. Ein einfaches mathematisches Modell: Sei R die Position der Röntgenquelle, P eine Position auf der Fotoplatte. Sei weiter $g(x)$, $g : \mathbb{R}^3 \mapsto \mathbb{R}$ die Stärke, mit der ein Röntgenstrahl am Punkt x geschwächt wird.
Die Helligkeit der Fotoplatte am Punkt P ist umso größer, je weniger der Röntgenstrahl auf seinem (geraden) Weg von R nach P geschwächt wurde: Ging der Strahl durch Knochen (dort ist g groß), so bleibt die Fotoplatte schwarz, ging er durch Luft, so wird die Platte weiß. Auf diese Weise bekommen wir eine zweidimensionale Projektion von g auf die Fotoplatte. Wir hätten aber gern nicht die Projektion, sondern g selbst - die Fragestellung ist daher: Wie berechnet man g aus seinen zweidimensionalen Projektionen?
Mathematisch ist die Schwärzung proportional zum Linienintegral von g über die Linie zwischen R und P. Die mathematische Fragestellung lautet daher: Kann man eine Funktion von $\mathbb{R}^n$ nach $\mathbb{R}$ aus Linienintegralen über die Funktion berechnen? Diese pure mathematische Fragestellung wurde weitgehend schon 1905 von Radon bei der Untersuchung der später nach ihm benannten Radon–Transformation beantwortet, der sogar eine Inversionsformel angeben konnte. Leider kann man zeigen, dass diese Inversionsformel nicht praktikabel ist (ihre direkte Implementation ist langsam und liefert große Fehler), siehe hierzu die Diskussionen in Natterer [2001]

und Natterer and Wübbeling [2001].
Sehr erfolgreich war dagegen eine viel einfachere Vorgehensweise: Man teilt den gesamten Raum in Würfel (Voxel) auf und nimmt an, dass g auf jedem Voxel konstant ist. Man macht sich schnell klar (am einfachsten in 2D), dass die Werte in jedem Voxel dann Lösung eines linearen Gleichungssystems sind, das nur noch invertiert werden muss. Effiziente Verfahren zur Lösung dieses Gleichungssystems (das ca. 512^3 Unbekannte hat) bilden heute den Kern der meisten Computertomographie–Geräte. Eine genauere Diskussion finden Sie zuhauf in der Literatur, z.B. in Natterer and Wübbeling [2001].

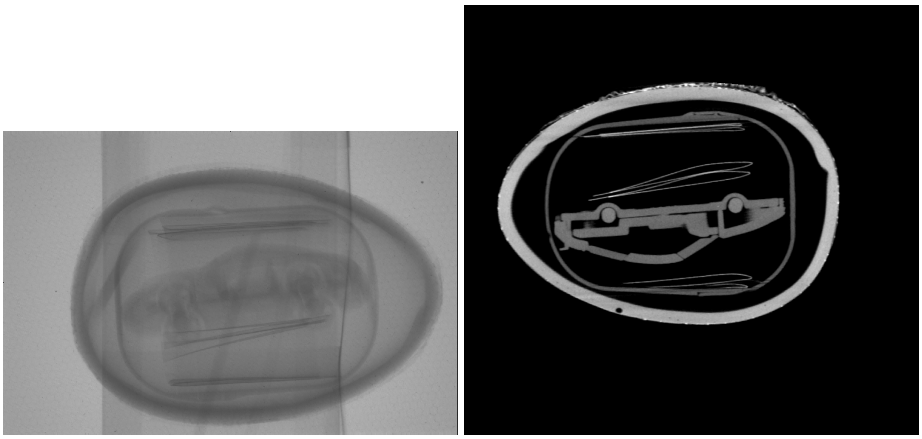


Abbildung 1: Röntgenbild/Tomographie eines Überraschungseis. Nur in der Tomographie sind Details erkennbar.

Klick für Bild autoxray
Klick für Bild autotomo

Beispiel 0.2 *Berechnung der Ableitung einer Funktion*
Die Funktion f sei auf dem Intervall I differenzierbar. Ihre Ableitung an der Stelle $x \in I$ ist definiert als

$$f'(x) = \lim_{h \mapsto 0,\, h \neq 0} \frac{f(x+h) - f(x)}{h}.$$

Diese Definition ist für die Praxis, in der die Ableitung etwa als Geschwindigkeit als Ableitung der zurückgelegten Wegstrecke auftritt, nutzlos, wenn nur diskrete (endlich viele) Funktionsauswertungen vorliegen. Es liegt in diesem Fall nahe, die Ableitung durch die Approximation

$$f'(x) \sim \frac{f(x+h) - f(x)}{h}$$

für ein kleines h zu ersetzen (Modellvereinfachung, Diskretisierung). Der Fehler dieses Modells kann für zweimal stetig differenzierbare Funktionen einfach angegeben werden. Mit der Taylorentwicklung und dem Lagrange-Restglied gilt

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi)$$

mit einem ξ zwischen x und $x+h$. Der maximale Fehler kann also abgeschätzt werden durch

$$\left| f'(x) - \frac{f(x+h) - f(x)}{h} \right| \leq \frac{|h|}{2} ||f''||_\infty \tag{1}$$

Entsprechend zeigt man für viermal stetig differenzierbare Funktionen eine Approximationsformel für die zweite Ableitung:

$$\left| f''(x) - \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \right| \leq \frac{h^2}{12} ||f''''||_\infty \tag{2}$$

Hierbei steht natürlich jeweils die Unendlichnorm für das Betragsmaximum auf I.

Nach dieser Analyse scheint klar: Je kleiner das h, desto besser das Ergebnis. Dies berücksichtigt aber natürlich die Messfehler nicht. Ist h sehr klein, so führt offensichtlich schon ein kleiner Fehler im Zähler zu riesigen Fehlern. Ist h zu groß, ist der Modellfehler, den wir angegeben haben, zu groß. Diesen Zusammenhang werden wir genauer untersuchen.

Beispiel 0.3 *Wärmeleitung in einem isolierten Stab*
Als Beispiel für ein komplexeres Anwendungsproblem betrachten wir einen wärmeisolierten Stab, der an beiden Enden auf eine feste Temperatur gebracht wird. Der Stab befinde sich im Intervall $[0, \pi]$ auf der x-Achse und sei homogen. Die Anfangstemperatur zum Zeitpunkt $t = 0$ sei bekannt. Es bezeichne $T(x,t)$ die Temperatur zum Zeitpunkt t an der Stelle x, $t \geq 0$, $x \in [0, \pi]$. Mögliche Fragestellungen:

1. *Bestimme den Temperaturverlauf T unter Berücksichtung einer externen Wärmequelle $q(x,t)$.*
2. *Nach längerer Zeit stellt sich ein fester Endzustand $T_0(x)$ ein. Bestimme T_0.*

Zunächst benötigen wir eine Mathematisierung (Modellierung). Unter Vernachlässigung vieler physikalischer und mathematischer Gesichtspunkte und aller Konstanten können wir diese leicht motivieren. Sei dazu $[a,b]$ ein beliebiges Teilintervall von $[0,\pi]$ und $t_2 > t_1 \geq 0$. Sei weiter

$$Q(t) = \int_a^b T(x,t)\, dx$$

die Wärmeenergie im Intervall $[a, b]$ zum Zeitpunkt t. Es ist anschaulich, dass die Wärmeenergie, die zu einem Zeitpunkt t durch einen Punkt x läuft, proportional zur Ableitung von T nach x ist: Ist die Ableitung 0, so ändert sich nichts, und es wird auch keine Wärme verschoben. Ist die Ableitung groß, hat man einen großen Temperaturunterschied links und rechts des Punkts, und Wärmeenergie fließt durch diesen Punkt in einer Richtung, die vom Vorzeichen des Unterschieds abhängt (Fouriersches Gesetz).

Zunächst gilt mal

$$Q(t_2) - Q(t_1) = \int_a^b T(x, t_2) - T(x, t_1) = \int_a^b \int_{t_1}^{t_2} T_t(x, t)\, dt\, dx$$

Da sich Q nur durch Zufluss oder Abfluss von Energie am linken oder rechten Rand oder durch externe Wärmezufuhr ändert, gilt

$$\begin{aligned} Q(t_2) - Q(t_1) &= \int_{t_1}^{t_2} T_x(b, t) - T_x(a, t)dt + \int_{t_1}^{t_2} \int_a^b q(x, t)\, dx\, dt \\ &= \int_{t_1}^{t_2} \int_a^b T_{xx}(x, t) + q(x, t)\, dx\, dt \end{aligned}$$

wobei T_x für die partielle Ableitung von T nach x steht und $q(x, t)$ für die Stärke einer externen Wärmequelle zum Zeitpunkt t an der Stelle x.

Für $b \mapsto a$ und $t_2 \mapsto t_1$ konvergiert nach dem Mittelwertsatz der Integralrechnung

$$\frac{1}{(b-a)(t_2-t_1)} \int_a^b \int_{t_1}^{t_2} T_t(x, t)\, dt\, dx \mapsto T_t(a, t_1)$$

und

$$\frac{1}{(b-a)(t_2-t_1)} \int_{t_1}^{t_2} \int_a^b T_{xx}(x, t) + q(x, t)\, dx\, dt \mapsto T_{xx}(a, t_1) + q(a, t_1)$$

und damit gilt für alle $x \in [0, \pi]$, $t \geq 0$

$$T_t(x, t) = T_{xx}(x, t) + q(x, t). \tag{3}$$

Eine genauere Herleitung bekommen Sie zum Beispiel in der Vorlesung Modellierung.

Wir haben damit den kompletten physikalischen Vorgang in einer mathematischen Beschreibung verstecken können, in einer Gleichung, die für alle x und t erfüllt sein muss und die die Ableitungen der gesuchten Funktion T enthält (partielle Differentialgleichung, Wärmeleitungsgleichung). Streng analytisch kann man nun zeigen:

Die Wärmeleitungsgleichung mit bekanntem Temperaturverlauf $T(x,0) = t_0(x)$ und festen Temperaturen am Rand ($T(0,t) = C_1$, $T(\pi,t) = C_2$) hat eine eindeutige Lösung (mit Bedingungen an q).

Sätze dieser Art sind Inhalt der Vorlesung "Partielle Differentialgleichungen".

Für den Endzustand $T_0(x)$ gilt, dass der Temperaturverlauf von der Zeit nicht mehr abhängt, also $(T_0)_t = 0$, er erfüllt damit

$$-(T_0)'' = q \tag{4}$$

und die Randbedingung (stationäre Wärmeleitungsgleichung).

Für sehr einfache Funktionen q lässt sich der Endzustand T_0 direkt angeben. Wir setzen der Einfachheit halber $C_1 = C_2 = 0$, am linken und rechten Ende des Stabes wird also auf 0 Grad gekühlt.

Eine in der Physik beliebte Methode, Aufgaben dieser Art zu lösen, bestimmt zunächst einmal die Eigenvektoren (Eigenfunktionen) der Abbildung auf der linken Seite der Differentialgleichung, also der zweiten Ableitung. Wir suchen also nicht–verschwindende Funktionen u_k mit $u_k(0) = u_k(\pi) = 0$ und $(u_k)_{xx} = \lambda_k u_k$. Man zeigt leicht, dass

$$u_k(x) = \sin kx,\ \lambda_k = -k^2$$

für $k \in \mathbb{N}$ dies leisten. Sei nun

$$T_0(x) = \sum_k a_k u_k(x)$$

eine Lösung der Differentialgleichung. Dann gilt

$$q(x) = -(T_0)_{xx}(x) = -\sum_k a_k \lambda_k u_k(x) = \sum_k a_k k^2 \sin kx.$$

Umgekehrt gilt: Hat q eine solche Reihenentwicklung, so ist $\sum_k a_k u_k(x)$ eine (die) Lösung der stationären Wärmeleitungsgleichung. Die Koeffizienten lassen sich mit Hilfe der Fouriertransformation berechnen.

Damit ist das Problem mathematisch eigentlich komplett gelöst: Wir haben gezeigt, dass es eine eindeutige Lösung gibt, und können diese sogar aus der Fourierreihendarstellung von q direkt berechnen. Sei etwa q die charakteristische Funktion des Intervalls $[\pi/2 - \epsilon, \pi/2 + \epsilon]$ für ein $1 > \epsilon > 0$. Dann gilt

$$q(x) = \sum_k A_k \sin(kx),\ A_k = \frac{2}{\pi} \int_{\pi/2-\epsilon}^{\pi/2+\epsilon} \sin(ky) dy = -\frac{2}{k\pi} \cos(ky) \Big|_{\pi/2-\epsilon}^{\pi/2+\epsilon}.$$

und die Lösung des stationären Wärmeleitungsproblems ist

$$T_0(x) = \sum_{k=1}^{\infty} -\frac{A_k}{k^2} \sin(kx).$$

Wir können die Lösung also exakt angeben. Dies geht aber offensichtlich nur, weil wir die Wärmequelle unrealistisch vereinfacht haben. Möglicherweise ist diese nur gemessen und besitzt keine geschlossene Darstellung - in diesem Fall können wir auch die Fourierreihe nicht berechnen und die analytische Lösung wird wertlos.

Bemerkung: Die so erzielte Lösung ist zwar physikalisch absolut sinnvoll, aber nicht differenzierbar und damit keine Lösung der Differentialgleichung. Dies zeigt, dass unsere mathematische Modellierung nicht vollständig ist.

Es gibt aber eine sehr einfache numerische Lösung für unser Problem durch Diskretisierung. Hierzu verteilen wir zunächst $N+1$ Gitterpunkte x_k gleichmäßig im Intervall $[0, \pi]$, also $x_k = kh$, $h = \pi/N$, $k = 0, \ldots, N$. Wir beschränken uns darauf, Näherungen u_k für $T_0(x_k)$ zu bestimmen. Mit (4) gilt an jedem Gitterpunkt

$$-T_0''(x_k) = q(x_k).$$

Wir approximieren die Differentialgleichung mit (2), also

$$-u_{k-1} + 2u_k - u_{k+1} = h^2 q(x_k),\ k = 1, \ldots, N-1.$$

Zusätzlich wissen wir wegen der Randbedingung $u_0 = T_0(0) = 0$ und $u_N = T_0(\pi) = 0$. Insgesamt erhalten wir damit $N-1$ lineare Gleichungen für die $N-1$ Unbekannten u_1 bis u_{N-1}:

$$\begin{aligned}
-0 + 2u_1 - u_2 &= h^2 q(x_1) \\
-u_1 + 2u_2 - u_3 &= h^2 q(x_2) \\
-u_2 + 2u_3 - u_4 &= h^2 q(x_3) \\
&\vdots \\
-u_{N-3} + 2u_{N-2} - u_{N-1} &= h^2 q(x_{N-2}) \\
-u_{N-2} + 2u_{N-1} - 0 &= h^2 q(x_{N-1})
\end{aligned} \tag{5}$$

oder in Matrixschreibweise

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} q(x_1) \\ q(x_2) \\ q(x_3) \\ \vdots \\ q(x_{N-2}) \\ q(x_{N-1}) \end{pmatrix} \tag{6}$$

Zur numerischen Lösung müssen wir also nur die Matrix invertieren und das Gleichungssystem lösen. Dies ist in Python schnell getan, Abbildung 3 zeigt den Vergleich zweier Lösungen, die jeweils mit der analytischen und diskreten Methode erzielt wurden. Bei großem N (hoher Auflösung) sind die Kurven praktisch gleich.

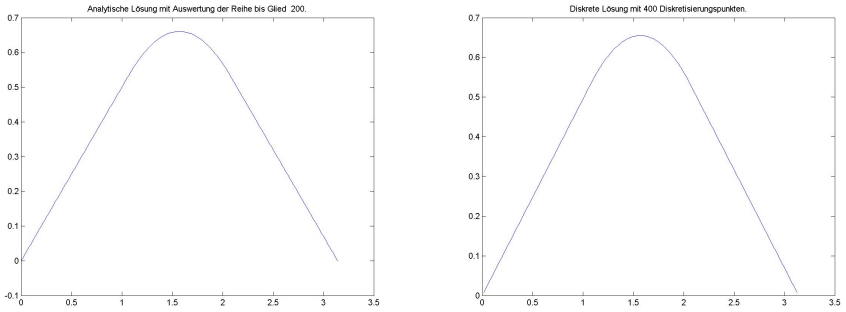


Abbildung 2: Analytische/Diskrete Lösung der stationären Wärmeleitungsgleichung

Klick für Bild heatanalytic

Klick für Matlab Figure heatanalytic

Klick für Bild heatdiscrete

Klick für Matlab Figure heatdiscrete

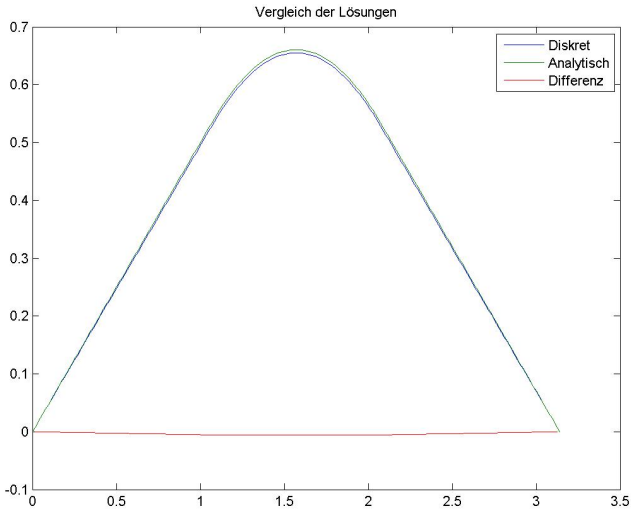


Abbildung 3: Vergleich der diskreten/analytischen Lösung der stationären Wärmeleitungsgleichung

Wir halten also fest: Zur Diskretisierung praktischer Probleme müssen am Ende meist (große) lineare Gleichungssysteme gelöst werden. Dies möglichst genau und effizient zu tun, wird den Großteil dieser Vorlesung einnehmen.

Bemerkung: *Die Cramersche Regel wäre aus mathematischer Sicht hierzu bereits absolut ausreichend, leider ist sie weder effizient noch liefert ihre direkte Implementation genaue Werte.*

Warnung: *Dies ist eine reine Motivation. Das analytische Modell durch ein diskretes zu ersetzen, scheint hier eine gute Idee zu sein. Tatsächlich kann aber natürlich erst eine genaue mathematische Analyse zeigen, ob das Ergebnis brauchbar ist bzw. mit welchem Fehler die diskrete Lösung behaftet ist. Zur Warnung schauen wir uns daher auch noch die zeitabhängige Wärmeleitungsgleichung an. Wir diskretisieren die Zeit an den Zeitpunkten $t_k = k$dt, $k \in \mathbb{N}_0$. Sei $u(t_k, x)$ die gesuchte Näherung für die Temperatur am Punkt x zum Zeitpunkt t_k. Mit Hilfe der Formel für die Diskretisierung der ersten Ableitung erhalten wir also*

$$u(t_{k+1}, x) = u(t_k + dt, x) = u(t_k, x) + dt(u_{xx} + q).$$

Wir können also eine Approximation für die Temperatur zum Zeitpunkt t_{k+1} angeben, wenn wir die Temperatur zum Zeitpunkt t_k kennen. Für $t = 0$ ist die Temperatur be-

kannt (hier konstant 0), zur Berechnung der zweiten Ableitung verwenden wir wieder unsere Formel für die Diskretisierung der zweiten Ableitung, und wir erhalten sofort die im Programm implementierte Formel. Um unser Programm zu testen, lassen wir es für einige Zeit laufen und vergleichen den Endzustand mit dem vorher berechneten aus der stationären Wärmeleitungsgleichung.

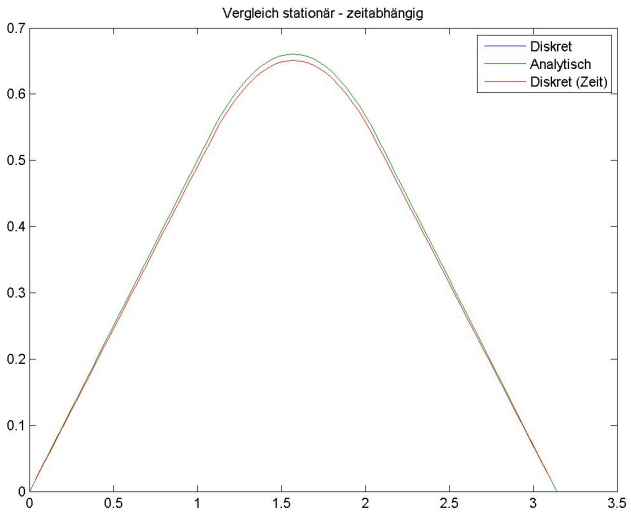


Abbildung 4: Vergleich der stationären Lösung mit der zeitabhängigen Lösung

Alles ist so, wie wir es erwarten: Der Endzustand liegt nah an der Lösung der stationären Gleichung. Wir wollen nun etwas genauer werden und erhöhen die Diskretisierung auf der Raumachse wenig, statt 80 wählen wir nun 100 Diskretisierungspunkte und lassen unsere Simulation wieder laufen.

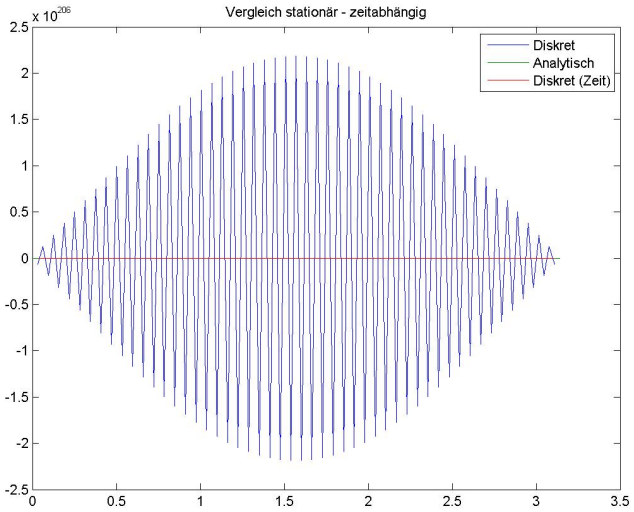


Abbildung 5: Vergleich der stationären Lösung mit der zeitabhängigen Lösung, instabil

Das ist definitiv nicht, was wir erwarten. Eine Verbesserung der Approximation führt zu einem völlig chaotischen (in der Numerik: instabilen) Verhalten. Dies zeigt deutlich, dass blinde Diskretisierung ohne zusätzliche mathematische Analyse zu unsinnigen Ergebnissen führen kann.

Kapitel 1

Modellierung einiger elementarer Differentialgeichungen

In diesem Kapitel wollen wir einige einfache Differentialgleichungen motivieren. Dies ist rein heuristisch, im dritten Kapitel werden wir die Anfangswertaufgaben fundiert behandeln.

1.1 Steinwurf: Die Bewegungsgleichung

Ein Stein wird senkrecht nach oben geworfen. Zum Zeitpunkt t_0 hat er die Höhe h_0 und die vertikale Geschwindigkeit v_0. Auf den Stein wirkt die Erdbeschleunigung $g \sim 9.81 \frac{m}{s^2}$ nach unten. Zu bestimmen ist die Höhe zum Zeitpunkt $t > t_0$.

Wir bezeichnen mit $h(t)$ die Höhe zum Zeitpunkt t, mit $v(t)$ die Geschwindigkeit zum Zeitpunkt t, mit $a(t)$ die Beschleunigung zum Zeitpunkt t (hier: konstant $-g$). Wir nehmen an, dass $a(t)$ bekannt ist.

Aufgabe: Bestimme $h(t)$ und $v(t)$.

Geschwindigkeit ist der Höhenunterschied pro Zeit im Limes für Zeit gegen 0 oder

$$v(t) = \lim_{dt \to 0} \frac{h(t+dt) - h(t)}{dt} = h'(t).$$

Entsprechend ist die Beschleunigung der Geschwindigkeitsunterschied pro Zeit, also

$$a(t) = \lim_{dt \to 0} \frac{v(t+dt) - v(t)}{dt} = v'(t).$$

Interpretation: Die Ableitung einer Funktion ist die Geschwindigkeit, mit der sich die Werte der Funktion ändern.

Zur zweiten Gleichung: Gesucht ist also eine Funktion $v(t)$ mit den Eigenschaften

$$v'(t) = a(t),\ v(t_0) = v_0.$$

Die eindeutige Lösung ist nach Analysis I für stetiges a gegeben durch

$$v(t) = \int_{t_0}^{t} a(s)\, ds + v_0.$$

und entsprechend für $h(t)$

$$h(t) = \int_{t_0}^{t} v(s)\, ds + h_0.$$

Für die allgemeine Form einer Anfangswertaufgabe lassen wir zu, dass die rechte Seite der Gleichung für $v'(t)$ auch von $v(t)$ abhängen darf, also

$$v'(t) = f(t, v(t)),\ v(t_0) = v_0.$$

Zusätzlich lassen wir auch zu, dass wir gleichzeitig mehrere Funktionen $v_1(t), \ldots, v_n(t)$ suchen mit $n > 1$. Für $(h(t), v(t))$:

$$(h'(t), v'(t)) = (v(t), a(t)),\ h(t_0) = h_0, v(t_0) = v_0$$

(und hier kommt jetzt die unbekannte Funktion v auch auf der rechten Seite vor). Wir sprechen von einem Anfangswertproblem für ein System von Differentialgleichungen.

Für den Steinwurf gilt $a(t) = -g$ und damit

$$v(t) = -\int_{t_0}^{t} g\, ds + v_0 = -(t - t_0)\, g + v_0$$

sowie

$$h(t) = \int_{t_0}^{t} v(s)\, ds + h_0 = -\frac{(t - t_0)^2}{2}\, g + (t - t_0)v_0 + h_0.$$

Man kann sich die Frage stellen, ob es auch Sinn macht, höhere Ableitungen zu betrachten. So können wir das System von oben auch beschreiben mit der Differentialgleichung

$$h''(t) = v'(t) = a(t)$$

und dann das v komplett weglassen. Dies ist jetzt keine Differentialgleichung in unserem Sinne (denn links steht die zweite Ableitung). Durch den Trick von oben (Einführung der Ableitung von v als zusätzliche Funktion) kann man aber solche Differentialgleichungen höherer Ordnung immer in Systeme von Differentialgleichungen der Ordnung 1 umwandeln (Übungen). Wir beschränken uns daher auf die Definition oben.

Wir halten fest:

- Die Lösung einer Differentialgleichung ist eine Funktion (oder mehrere Funktionen bei einem System).
- Die allgemeine Form einer Anfangswertaufgabe ist

  $$u'(t) = f(t, u(t)),\ u(t_0) = u_0$$

  bzw. als System

  $$(u_1'(t), \ldots, u_n'(t)) = (f_1(t, u_1(t), \ldots, u_n(t)), \ldots, f_n(t, u_1(t), \ldots, u_n(t)))$$

  $$u_1(t_0) = u^1, \ldots, u_n(t_0) = u^n.$$

- Differentialgleichungen mit höherer Ableitung lassen sich in Systeme von Differentialgleichungen mit erster Ableitung umwandeln.

1.2 Populationsdynamik: Entwicklung der Bevölkerung

Es sei $u(t)$ die Größe einer Population zum Zeitpunkt t, $u(t_0) = u_0 > 0$.

Aufgabe: Bestimme $u(t)$.

Die Änderung der Population im Zeitintervall $[t, t + dt]$ ist gegeben durch

$$u(t + dt) - u(t)$$

und entspricht der Differenz aus Geburten und Sterbefällen in diesem Zeitraum. Diese wird für kleine dt im Wesentlichen proportional zu dt und zur Größe der Bevölkerung zum Zeitpunkt t sein, d.h.

$$u(t + dt) - u(t) \sim dt\, C(t, u(t))\, u(t).$$

Der Proportionalitätsfaktor $C(t, u(t))$ hängt natürlich von der Zeit ab (etwa ist die Geburtenrate im Sommer höher) und von $u(t)$ (etwa bei Überbevölkerung, ist $u(t)$ groß, so wird der Proportionalitätsfaktor klein).

Im Limes für $dt \to 0$ gilt (für C stetig)

$$u'(t) = C(t, u(t))\, u(t),\ u(t_0) = u_0.$$

Wir erhalten also eine Anfangswertaufgabe für u.

Einfachstes (exponentielles) Modell: Wir nehmen an, dass $C(t, u(t))$ konstant ist, also $C(t, u(t)) = \lambda$ oder

$$u'(t) = \lambda u(t) \Longleftrightarrow \frac{u'(t)}{u(t)} = \lambda,\ u(t) \neq 0.$$

Dann gilt mit einer Integrationskonstanten c

$$\begin{aligned}(\log|u(t)|)' = \lambda &\Rightarrow \log|u(t)| = \lambda t + c \\ &\Rightarrow u(t) = e^{\lambda t} \cdot e^c.\end{aligned}$$

Wir setzen $c = \log u_0 - \lambda t_0$ und erhalten:

$$u(t) = u_0 e^{\lambda(t-t_0)}$$

ist Lösung der Anfangswertaufgabe.

Es gibt also zwei Möglichkeiten für das Langzeitverhalten $t \mapsto \infty$ von u: Für $\lambda < 0$ stirbt die Bevölkerung aus, für $\lambda > 0$ wächst die Bevölkerungszahl ohne Grenze exponentiell.

Insbesondere das zweite ist allein schon wegen der Nahrungsbeschränkungen offensichtlich unrealistisch: Wir müssen berücksichtigen, dass für $u(t)$ groß der Proportionalitätsfaktor kleiner wird.

Im zweiten Modell nehmen wir daher an, dass dauerhaft Nahrung für M Individuen zur Verfügung steht. Also: Ist $u(t) = M$, so sollte $C(t, u(t)) = 0$ sein (die Änderung wird 0), für $u(t) > M$ negativ, für $u(t) < M$ positiv. Wir nehmen daher an

$$C(t, u(t)) = \lambda\,(M - u(t)),\ \lambda > 0.$$

Der Einfachheit halber setzen wir in der folgenden Rechnung $M = 1$.

Wir suchen also eine Funktion u mit

$$u'(t) = \lambda(1 - u(t))\, u(t) \Leftrightarrow \frac{u'(t)}{u(t)\,(1 - u(t))} = \lambda,\ 1 \neq u(t) \neq 0.$$

Leider steht hier jetzt nicht mehr die Ableitung des Nenners im Zähler, aber mit Partialbruchzerlegung (Forster I 19.14, $\frac{1}{z(1-z)} = \frac{1}{z} + \frac{1}{1-z}$) gilt für $u(t) \in (0,1)$

$$\lambda = u'(t)\frac{1}{u(t)(1-u(t))} = \frac{u'(t)}{u(t)} + \frac{u'(t)}{1-u(t)} = (\log u(t) - \log(1-u(t)))'$$

und damit mit einer Integrationskonstanten c

$$\log \frac{u(t)}{1-u(t)} = \lambda t + c \Leftrightarrow \frac{u(t)}{1-u(t)} = e^{\lambda t + c}.$$

Auflösen nach $u(t)$ liefert

$$u(t) = \frac{1}{1+e^{-\lambda t - c}}.$$

c kann aus der Anfangsbedigung $u(t_0) = u_0$ bestimmt werden.

Wie nach der Herleitung erwartet, gilt hier im Langzeitverhalten

$$u(t) \xrightarrow{t\to\infty} 1 = M.$$

Wir halten fest:

- $Ce^{\lambda t}$ ist Lösung der Differentialgleichung

$$y'(t) = \lambda y(t).$$

- Die Lösungsmethode für den zweiten Fall werden wir im nächsten Kapitel verallgemeinern (autonome Differentialgleichungen).

1.3 Federmodell: Schwingungsgleichung

Im Buch von Bollhöfer und Mehrmann finden Sie die Modellierung des schwingenden Sitzes in einer Fahrerkabine, der auf mehreren Federn steht (Reifen, Autofederung, Sitzfederung) durch ein System von Differentialgleichungen.

Wir beschränken uns auf einen gefederten Sitz, der Sitz sei der Einfachheit halber massenlos. Auf dem Sitz nimmt eine Person mit Masse m langsam Platz, die Feder wird nach unten gedrückt. Es sei $h(t)$ die Auslenkung der Feder.

Aufgabe: Bestimme $h(t)$.

Auf den Sitz wirken zwei Kräfte: Die Gravitation $-m\,g$ und die Federkraft $-c\,h(t)$ (Hookesches Gesetz). Hier ist c die Federkonstante. Insgesamt also

$$F(t) = -m\,g - c\,h(t).$$

In der Ruhelage, bei sitzendem Fahrer und Federauslenkung H, ist die Summe F der Kräfte gleich Null, d.h.

$$F = -m\,g - c\,H = 0 \Longleftrightarrow H = -\frac{m\,g}{c}.$$

Durch ein Schlagloch werde nun der Sitz aus der Ruhelage geworfen. Die Auslenkung der Feder sei h_0 (also $h(0) = h_0$), und die vertikale Geschwindigkeit sei v_0 (also $h'(0) = v_0$).

Die auf den Sitz wirkende Beschleunigung ist $F(t)/m$, also

$$a(t) = -g - \frac{c\,h(t)}{m}.$$

Wir nutzen diesmal zur Modellierung die Gleichung zweiter Ordnung aus Beispiel 1, also

$$h''(t) = a(t) = -g - \frac{c\,h(t)}{m}.$$

Man wird vermuten, dass der Sitz anfängt, um die Ruhelage herum zu schwingen, d.h. wir erwarten etwas wie

$$h(t) = H + \alpha\sin(\lambda t) + \beta\cos(\lambda t)$$

mit zu bestimmenden Konstanten α, β, λ. Unter Berücksichtigung dieser Vermutung machen wir den Ansatz $u(t) = h(t) - H$. Dann gilt

$$u''(t) = h''(t) = -g - \frac{c\,h(t)}{m} = -g - \frac{c\,(u(t)+H)}{m} = -\frac{c}{m}u(t).$$

Wir machen den Ansatz $u(t) = e^{\lambda t}$. Dann gilt $u''(t) = \lambda^2 u(t)$, und damit ist

$$u(t) = e^{\lambda t},\ \lambda^2 = -\frac{c}{m}$$

Lösung der Differentialgleichung.

Wir haben also zwei Fälle:

1. $-\frac{c}{m} > 0$: In diesem Fall ist λ reell. Die Lösungen sind Exponentialfunktionen. Das macht für das Federbeispiel keinen Sinn - aber tatsächlich macht ja auch in unserer Modellierung $m < 0$ oder $c < 0$ keinen Sinn.

2. $-\frac{c}{m} < 0$: Sei $\lambda^2 = \frac{c}{m}$. Dann ist

$$(i\lambda)^2 = (-i\lambda)^2 = -\frac{c}{m}.$$

Mit den Eulerschen Formeln sind $u(t) = \sin \lambda t$ und $u(t) = \cos \lambda t$ Lösungen der Differentialgleichung für u, und auch alle ihre Linearkombinationen. Die Funktionen

$$h(t) = H + u(t) = H + \alpha \sin \lambda t + \beta \cos \lambda t$$

sind also tatsächlich, wie vermutet, Lösungen der Differentialgleichung für h.

.

Wir halten fest:

- $e^{\lambda t}$, $e^{-\lambda t}$ und alle ihre Linearkombinationen sind Lösungen der Gleichung

$$u''(t) = \lambda^2 u(t).$$

- $\sin \lambda t$, $\cos \lambda t$ und alle ihre Linearkombinationen sind Lösungen der Differentialgleichung

$$u''(t) = -\lambda^2 u(t).$$

- Änderung eines Vorzeichens in der Differentialgleichung ändert das Lösungsverhalten komplett (exponentiell, schwingend).

1.4 Stationäre Wärmeleitungsgleichung: Randwertprobleme

Abschließend schauen wir auf ein Problem, das nicht in unser bisheriges Framework passt. Wir werden dies erst in einem späteren Kapitel behandeln. Im Buch von Bollhöfer und Mehrmann finden Sie ein komplexeres und viel praktischeres Beispiel (Kühlrippe).

Ein homogener Draht sei zwischen zwei Punkten (0 und 1) gespannt. Es sei $T(t, x)$ die Temperatur im Punkt x zum Zeitpunkt t. Der Draht wird von unten erhitzt, die Energiezufuhr im Punkt x sei $q(x)$. Der Draht wird an den Rändern auf 0 Grad gekühlt, d.h. $T(t, 0) = 0 = T(t, 1)$. Nach einiger Zeit ändert sich die Temperatur nicht mehr, d.h. T hängt nicht mehr von t ab, es gilt $T(t, x) = u(x)$.

Aufgabe: Bestimme $u(x)$.

Im Folgenden nehmen wir an, dass u zweimal stetig differenzierbar und q stetig ist.

Auf jeden Punkt $s \in [0,1]$ wirken hier zwei Effekte: Einmal wird von unten erhitzt, andererseits fließt Energie, wenn rechts oder links von s die Temperatur größer oder kleiner ist.

Nach dem Fourierschen Gesetz ist der Energiefluss durch den Punkt s proportional zu $-u'(s)$. Dies lässt sich leicht verstehen: Gilt $u'(s) > 0$, so ist die Temperatur rechts höher, und es fließt Energie von rechts nach links (negativer Abfluss) und umgekehrt. Der Proportionalitätsfaktor ist der (nichtnegative) Wärmeleitkoeffizient κ des Drahts.

Wir schauen nun auf die Energie im Intervall $[s-h, s+h]$. Die gesamte in diesem Intervall zugeführte Energie ist

$$\int_{s-h}^{s+h} q(x)\,dx.$$

Da sich die Temperatur nicht mehr ändert, fließt durch den rechten und linken Rand genau so viel Energie im Intervall ab, wie von unten zugeführt wird.

Wir erhalten mit dem Mittelwertsatz der Integralrechnung

$$\begin{aligned}
&\int_{s-h}^{s+h} q(x)\,dx = \kappa(u'(s-h) - u'(s+h)) \\
\iff \qquad & 2hq(\xi) = -\kappa h \left(\frac{u'(s-h) - u'(s)}{-h} + \frac{u'(s+h) - u'(s)}{h} \right) \\
\iff \qquad & q(\xi) = -\frac{\kappa}{2} \left(\frac{u'(s-h) - u'(s)}{-h} + \frac{u'(s+h) - u'(s)}{h} \right)
\end{aligned}$$

mit einem $\xi \in [s-h, s+h]$. Für $h \mapsto 0$ konvergiert die linke Seite daher gegen $q(s)$. Die rechte Seite ist eine Summe aus zwei Differenzenquotienten für u' am Punkt s, für $h \mapsto 0$ konvergiert sie also gegen $-\kappa u''(s)$. u muss daher erfüllen:

$$-u''(s) = \frac{q(s)}{\kappa},\ u(0) = u(1) = 0.$$

Die Differentialgleichung kennen wir schon aus dem ersten Beispiel.

Aber: In diesem Fall sind die Werte nicht am Anfang des Intervalls vorgegeben (Anfangswertaufgabe), sondern auf den beiden Rändern (Randwertaufgabe). Wir werden in Kapitel X sehen, dass dies die numerische und analytische Behandlung komplett ändert.

Wir halten fest:

- Bei Randwertprobleme sind Werte am linken und rechten Rand eines Intervalls vorgeschrieben.

Kapitel 2

Analytische Lösung von Differentialgleichungen

Wir behandeln einige typische Beispiele für analytisch lösbare Differentialgleichungen und Anfangswertaufgaben.

2.1 Elementare Differentialgleichung

Hier hängt die rechte Seite der Differentialgleichung nur von t, nicht von $y(t)$, ab. Wir behandeln den eindimensionalen Fall

$$y'(t) = f(t)$$

mit einer stetigen Funktion $f : [a, b] \mapsto \mathbb{R}$. Dann ist nach dem Fundamentalsatz der Differential– und Integralrechnung

$$y(t) = \int_a^t f(s)\, ds + C$$

für jede Integrationskonstante C eine Lösung der Differentialgleichung.

Hinweis: Zur Vereinfachung betrachten wir hier (fast) alles über $\mathbb{R}$. Natürlich kann man auch komplexwertige Funktionen y und f betrachten.

Beispiel 2.1 *Eine Lösung der Anfangswertaufgabe*

$$y'(t) = t + 1 =: f(t),\ y(a) = y_0$$

ist die Funktion

$$y(t) = \int_a^t f(s)\, ds + y_0 = \int_a^t s + 1\, ds + y_0 = \frac{1}{2}(t^2 - a^2) + (t - a) + y_0.$$

2.2 Autonome Differentialgleichung

Hier hängt die rechte Seite der Differentialgleichung nur von $y(t)$, nicht von t, ab. Wir behandeln den eindimensionalen Fall für das Anfangswertproblem

$$y'(t) = f(y(t)),\, y(a) = y_0.$$

mit einer stetigen Funktion $f : J \mapsto \mathbb{R}$ mit y_0 im Inneren von $J \subset \mathbb{R}$. Weiter sei $f(y_0) \neq 0$.

Da f stetig, y_0 im Inneren von J und $f(y_0) \neq 0$ gibt es eine Umgebung $J_0 \subset J$ von y_0 mit $\operatorname{sgn}(f(y)) = \operatorname{sgn}(f(y_0))\, \forall y \in J_0$. Hierbei ist sgn die Vorzeichenfunktion, d.h. f wechselt auf J_0 sein Vorzeichen nicht. Insbesondere hat $f(y)$ keine Nullstelle in J_0.

Wir definieren für $y \in J_0$

$$G(y) := \int_{y_0}^{y} \frac{1}{f(z)}\, dz \Rightarrow G'(y) = \frac{1}{f(y)}$$

und diese Funktion ist nach den Voraussetzungen wohldefiniert und streng monoton.

Nach dem Satz über implizite Funktionen (lokale Umkehrfunktion, z.B. Forster I Kapitel 15, Satz 3) gilt:

Sei $G(y_0) = a + C$. Da $G'(y_0) \neq 0$, gibt es eine in einer Umgebung $I \subset \mathbb{R}$ von a definierte differenzierbare Funktion Y mit den Eigenschaften

$$G(Y(t)) = t + C,\, Y'(t) = \frac{1}{G'(Y(t))},\, Y(a) = y_0.$$

Beweis der Formel für die Ableitung mit der Kettenregel für die Ableitung:

$$1 = G'(Y(t))\, Y'(t) = \frac{Y'(t)}{f(Y(t))}.$$

Es gilt also:

Satz 2.2 *Y ist Lösung des Anfangswertproblems.*

Beispiel 2.3 *Wir suchen eine Lösung der Differentialgleichung*

$$y'(t) = y(t)^2 =: f(y(t)).$$

Sei G eine Stammfunktion von $\frac{1}{f(z)}$, also z.B.

$$G(y) = \int^y \frac{1}{f(z)}\, dz = \int^y \frac{1}{z^2}\, dz = -\frac{1}{y}.$$

Wir suchen wie im Satz eine Funktion Y mit der Eigenschaft

$$t + C = G(Y(t)) = -\frac{1}{Y(t)} \Rightarrow Y(t) = -\frac{1}{t + C}.$$

Tatsächlich gilt

$$Y'(t) = \frac{1}{(t + C)^2} = Y(t)^2.$$

Beispiel 2.4 *Wir suchen eine Lösung der Differentialgleichung*

$$y'(t) = 1 + y(t)^2.$$

Wir setzen

$$G(y) = \int^y \frac{1}{1 + z^2}\, dz = \arctan y.$$

Wie im Satz

$$\arctan(Y(t)) = t + C \Rightarrow Y(t) = \tan(t + C)$$

und $Y(t)$ ist Lösung der Differentialgleichung.

Bemerkung: Die rechte Seite der Differentialgleichung in diesem Beispiel ist auf ganz $\mathbb{R}$ definiert. Die hier berechneten Lösungen der Differentialgleichung sind aber nur definiert in Intervallen der Länge π - der Tangens hat jeweils bei $k\pi + \pi/2$ einen Pol. Wir bemerken: Der Satz garantiert nur eine Lösbarkeit in einer Umgebung I von a, die Gleichung besitzt nicht notwendig globale Lösungen.

2.3 Getrennte Variable

Satz 2.5 *Es seien $f : J \mapsto \mathbb{R}$ und $g : [a, b] \mapsto \mathbb{R}$ stetig, $f(y) \neq 0$. Weiter sei F eine Stammfunktion von $\frac{1}{f}$ und G eine Stammfunktion von g.*

Es sei $y : [a, b] \mapsto \mathbb{R}$ eine differenzierbare Funktion mit $F(y(t)) - G(t) + C = 0$. Dann ist y Lösung der Differentialgleichung

$$y'(t) = f(y(t))\, g(t).$$

Beweis und Beispiel: Übungen.

Bemerkung: Autonome und elementare Differentialgleichungen sind Spezialfälle für getrennte Variable.

2.4 Exakte Differentialgleichung, integrierender Faktor

Satz 2.6 *Es seien $g, h : [a, b] \times J \mapsto \mathbb{R}$. Es sei $F : [a, b] \times \mathbb{R} \mapsto \mathbb{R}$ zweimal stetig differenzierbar. Weiter gelte*

$$F_t(t, y) = g(t, y),\ F_y(t, y) = h(t, y).$$

Dann gilt $g_y(t, y) = h_t(t, y)$. Die Differentialgleichung

$$y'(t) = -\frac{g(t, y(t))}{h(t, y(t))}$$

heißt exakt. Hierbei steht F_t für die Ableitung von F nach der ersten Variablen (t), F_y für die Ableitung nach der zweiten Variablen (y) usw.

Sei weiter $Y : [a, b] \mapsto \mathbb{R}$ eine Funktion mit $F(t, Y(t)) = C$ und $h(t, Y(t)) \neq 0$.

Dann ist Y eine Lösung der exakten Differentialgleichung.

Beweis und Beispiel: Übungen.

Leider ist dieser Satz nur sehr selten anwendbar, denn nur wenige Differentialgleichungen erfüllen die Voraussetzung des Satzes. Hier hilft die folgende Umformulierung:

Satz 2.7 *In der obigen Situation sei $M(t, y)$ eine Funktion mit*

$$g_y(t, y)\, M(t, y) + g(t, y) M_y(t, y) = h_t(t, y) M(t, y) + h(t, y) M_t(t, y).$$

Dann ist die Differentialgleichung

$$y'(t) = -\frac{M(t, y(t)) g(t, y(t))}{M(t, y(t)) h(t, y(t))} = -\frac{g(t, y(t))}{h(t, y(t))}$$

exakt. M heißt integrierender Faktor.

Beweis und Beispiel: Übungen.

2.5 Lineare Differentialgleichungen und Variation der Konstanten

Differentialgleichungen der Form

$$y'(t) = \alpha(t) y(t) + \beta(t)$$

heißen lineare Differentialgleichungen. Falls $\beta = 0$, so heißt die Gleichung homogen, ansonsten inhomogen.

Hier lassen wir ausdrücklich Systeme zu, d.h. dass eine Funktion $y : I \mapsto \mathbb{R}^n$ oder $y : I \mapsto \mathbb{C}^n$ gesucht wird, und ebenso $\alpha : I \mapsto \mathbb{C}^{n\times n}$, $\beta : I \mapsto \mathbb{C}^n$. Diese werden wir in Kapitel 4 untersuchen.

Hier beschränken wir uns auf den reellen eindimensionalen (skalaren) Fall mit α, β stetig. Für die homogene Gleichung hatten wir in Kapitel 1 hergeleitet: Sei

$$Y(t) = e^{\int_a^t \alpha(s)\,ds}.$$

Dann ist $CY(t)$ Lösung der AWA mit $Y(a) = C$. Sei nun $\beta \neq 0$. In diesem Fall machen wir den Ansatz (Variation der Konstanten)

$$y(t) = C(t)\,Y(t).$$

Damit y Lösung der Differentialgleichung ist, muss gelten

$$C(t)Y'(t) + C'(t)Y(t) = y'(t) = \alpha(t)y(t) + \beta(t) = \alpha(t)C(t)Y(t) + \beta(t)$$

und damit, da $Y'(t) = \alpha(t)Y(t)$,

$$C'(t)Y(t) = \beta(t) \Rightarrow C(t) = \int_a^t \frac{\beta(s)}{Y(s)}\,ds + C_1.$$

Nachrechnen zeigt sofort:

Satz 2.8 *(Variation der Konstanten)*
Für $C_1 \in \mathbb{R}$ ist eine Lösung der skalaren Differentialgleichung gegeben durch

$$y(t) = \left(\int^t \frac{\beta(s)}{Y(s)}\,ds + C_1\right) Y(t),\ Y(t) = e^{\int^t \alpha(s)\,ds}$$

mit $Y(s) \neq 0$. Lösung der Anfangswertaufgabe mit $y(a) = y_0$ ist

$$y(t) = \left(\int_a^t \frac{\beta(s)}{Y(s)}\,ds + y_0\right) Y(t),\ Y(t) = e^{\int_a^t \alpha(s)\,ds}$$

.

Beispiel 2.9 *Wir suchen eine Lösung der AWA*

$$y'(t) = y(t) + 1,\ y(0) = 1.$$

Wir gehen vor wie oben (mit einer allgemeinen Stammfunktion)

$$Y(t) = e^{\int^t 1\,ds} = e^t$$

und erhalten

$$C(t) = \int^t \frac{1}{e^s}\,ds + C_0 = -e^{-t} + C_0$$

oder

$$y(t) = \left(-e^{-t} + C_0\right) e^t.$$

Zur Lösung unserer Anfangswertaufgabe wählen wir $C_0 = 2$ *und erhalten*

$$y(t) = 2e^t - 1.$$

2.6 Zusammenfassung

2.6.1 Kompetenzen

- Analytische Lösung von Anfangswertaufgaben der Typen Elementar/Autonom/Getrennte Variable.
- Analytische Lösung von Anfangswertaufgaben für exakte DGL (mit integrierenden Faktoren).
- Analytische Lösung von linearen skalaren Differentialgleichungen durch Variation der Konstanten.
- Anfangswertaufgaben sind im Allgemeinen nicht global, sondern nur lokal lösbar. Beispiel: $y'(t) = 1 + y(t)^2$.

2.6.2 Mini–Aufgaben

Geben Sie bei diesen Differentialgleichungen jeweils den Typ an und skizzieren Sie den Lösungsweg. Einige lassen sich mehreren Typen zuordnen.

- $y'(t) = t\,y(t)$
- $y'(t) = t\,y(t) + t^2$
- $y'(t) = t\,y(t)^2$
- $y'(t) = -\frac{ty(t)^2}{y(t)t^2}$
- $y'(t) = \frac{y(t)}{t}$

Geben Sie für die letzte Gleichung einen integrierenden Faktor an (ohne Rechnung - schauen Sie in die vierte Gleichung).

Kapitel 3

Existenz und Eindeutigkeit der Lösung von Anfangswertaufgaben

Für den Beweis des Satzes von Picard–Lindelöf beweisen wir zunächst den Banachschen Fixpunktsatz. Dieser wird später eine zentrale Rolle bei der numerischen Lösung von Gleichungen spielen.

3.1 Der Banachsche Fixpunktsatz

Definition 3.1 *(kontrahierend, Fixpunkt)*
Seien X, Y normierte Räume, $D \subset X$.

1. *Eine Funktion*
$$g : D \mapsto Y$$
heißt kontrahierend genau dann, wenn eine Konstante $0 \leq q < 1$ existiert mit
$$||g(x) - g(y)|| \leq q||x - y|| \ \forall x, y \in D.$$
q heißt Kontraktionskonstante.

2. *Sei $g : D \mapsto X$. $\overline{x} \in D$ heißt Fixpunkt von g genau dann, wenn*
$$g(\overline{x}) = \overline{x}.$$

Lemma 3.2 *Sei g kontrahierend. Dann ist g stetig.*

Beweis: Sei x_n eine gegen x konvergente Folge, q die Kontraktionskonstante von g. Dann gilt
$$||g(x_n) - g(x)|| \leq q||x_n - x|| \to 0.$$

□

Satz 3.3 *(**Banachscher Fixpunktsatz**)*
Es gelte:

1. X *ist ein vollständiger, normierter Raum (Banachraum), d.h. jede Cauchyfolge in* X *konvergiert gegen einen Grenzwert in* X.
2. $\emptyset \neq D \subset X$ *ist abgeschlossen. Jede Cauchyfolge mit Folgegliedern in* D *hat einen Grenzwert in* D.
3. $g : D \mapsto D$ *ist eine Funktion auf* D *mit Werten in* D.
4. g *ist kontrahierend.*

Dann gilt:

1. g *hat genau einen Fixpunkt* $\overline{x}$.
2. *Sei* $x^{(0)} \in D$. *Dann konvergiert die rekursiv definierte Folge*

$$x^{(k+1)} = g(x^{(k)})$$

   gegen $\overline{x}$. $x^{(k)}$ *heißt Fixpunktfolge oder Fixpunktiteration.*

3. *Für die Fixpunktiteration gelten die Fehlerabschätzungen*

$$||\overline{x} - x^{(k)}|| \leq \frac{q^k}{1-q}||x^{(1)} - x^{(0)}|| \text{ (a priori)}$$

   und

$$||\overline{x} - x^{(k)}|| \leq \frac{q}{1-q}||x^{(k)} - x^{(k-1)}|| \text{ (a posteriori)}$$

Beweis:
Zu Aussage 1.: Sei $0 \leq q < 1$ Kontraktionskonstante von g.
Seien x und y zwei Fixpunkte von g. Dann gilt

$$||x - y|| = ||g(x) - g(y)|| \leq q||x - y|| \Longleftrightarrow (1-q)||x - y|| \leq 0.$$

Da $q < 1$, ist dies nur möglich für $x = y$. Also ist der Fixpunkt eindeutig.
Die Existenz zeigen wir konstruktiv und geben eine konvergente Folge an, deren Grenzwert ein Fixpunkt ist. Sei $x^{(0)} \in D$ beliebig und $x^{(k)}$ die zugehörige Fixpunktiteration.

g ist kontrahierend, also gilt mit der Definition von $x^{(k)}$

$$\begin{aligned} ||x^{(k+1)} - x^{(k)}|| &= ||g(x^{(k)}) - g(x^{(k-1)})|| \\ &\leq q||x^{(k)} - x^{(k-1)}|| \\ &\leq q^2||x^{(k-1)} - x^{(k-2)}|| \\ &\vdots \\ &\leq q^k||x^{(1)} - x^{(0)}||. \end{aligned}$$

Sei $\epsilon > 0$ beliebig und M so groß, dass

$$\frac{q^M}{1-q}||x^{(1)} - x^{(0)}|| \leq \epsilon.$$

Seien $l, k > M$ und ohne Einschränkung $l \geq k$. Dann gilt

$$\begin{aligned} ||x^{(l)} - x^{(k)}|| &\leq \underbrace{||x^{(l)} - x^{(l-1)}||}_{\leq q^{l-1}||x_1 - x_0||} + \underbrace{||x^{(l-1)} - x^{(l-2)}||}_{\leq q^{l-2}||x_1 - x_0||} + \ldots + \underbrace{||x^{(k+1)} - x^{(k)}||}_{\leq q^k||x_1 - x_0||} \\ &\leq q^k \sum_{j=0}^{l-k-1} q^j ||x^{(1)} - x^{(0)}|| \\ &\leq \frac{q^k}{1-q}||x^{(1)} - x^{(0)}|| \qquad (3.1) \\ &\leq \epsilon \end{aligned}$$

nach Wahl von k und M. Also ist $x^{(k)}$ eine Cauchyfolge in D und hat einen Grenzwert $\overline{x} \in D$. Es gilt, da g stetig ist,

$$x_{k+1} = g(x_k) \Longrightarrow_{k \mapsto \infty} \overline{x} = g(\overline{x}),$$

also ist $\overline{x}$ Fixpunkt und wegen der Vorbemerkung der einzige Fixpunkt von g. Dies beweist die Aussagen 1. und 2.
Zu den Abschätzungen: Nach 3.1 gilt

$$||\overline{x} - x^{(k)}|| = \lim_{l \to \infty, l > k} ||x^{(l)} - x^{(k)}|| \leq \frac{q^k}{1-q}||x^{(1)} - x^{(0)}||.$$

Sei nun y die Fixpunktfolge mit Startwert $x^{(k)}$, also $y^{(0)} = x^{(k-1)}$ und $y^{(1)} = g(x^{(k-1)}) = x^{(k)}$. Wir wenden die gerade bewiesene Abschätzung auf die Folge y an und erhalten

$$||\overline{x} - x^{(k)}|| = ||\overline{x} - y^{(1)}|| \leq \frac{q}{1-q}||y^{(1)} - y^{(0)}|| = \frac{q}{1-q}||x^{(k)} - x^{(k-1)}||.$$

□

Bemerkung: Mit Hilfe der ersten Abschätzung können wir im Vorhinein (a priori), nach Berechnung nur von $x^{(1)}$, eine obere Schranke für den Fehler von $x^{(k)}$ angeben, basierend auf $||x^{(1)} - x^{(0)}||$.
Mit Hilfe der zweiten Abschätzung können wir im Nachhinein (a posteriori), wenn wir also $x^{(k)}$ berechnet haben, ebenfalls eine obere Schranke für den Fehler angeben, basierend auf $||x^{(k)} - x^{(k-1)}||$.

3.2 Der Existenz– und Eindeutigkeitssatz von Picard–Lindelöf

Diesen Satz haben Sie vermutlich bereits in der Analysis kennengelernt. Wenn ja, schauen Sie nochmal in Ihre Aufzeichnungen. Da er für uns zentral ist, werden wir ihn noch einmal beweisen.

Definition 3.4 *(Anfangswertaufgabe, AWA)*
Es sei $D \subset \mathbb{R} \times \mathbb{R}^n$. Sei $(a, y_0) \in D$.
Im Allgemeinen werden wir im Folgenden wählen $D = I \times J$, $I \subset \mathbb{R}$, $J \subset \mathbb{R}^n$, $a \in I$, $y_0 \in \mathring{J}$. Weiter sei $f : D \mapsto \mathbb{R}^n$.
Die Aufgabe: Bestimme eine auf einem Intervall I_0, $a \in I_0$, definierte Funktion y mit

$$y : I_0 \mapsto \mathbb{R}^n,\ y'(t) = f(t, y(t)),\ y(a) = y_0,\ (t, y(t)) \in D$$

heißt Anfangswertaufgabe für die gewöhnliche Differentialgleichung.

Lemma 3.5 *(Integraldarstellung der Differentialgleichung)*
Es seien y und f stetig, und es sei $(s, y(s)) \in D \forall s \in I_0$. Dann ist y genau dann Lösung der Anfangswertaufgabe 3.4, wenn

$$y(t) = y_0 + \int_a^t f(s, y(s))\, ds \,\forall t \in I_0.$$

Beweis: Hauptsatz der Differential– und Integralrechnung. □

Definition 3.6 *(Lipschitzstetigkeit)*
Es seien X und Y normierte Vektorräume, $D \subset X$, $f : D \mapsto Y$. f heißt lipschitzstetig genau dann, wenn es eine (Lipschitz–) Konstante $L \geq 0$ gibt mit

$$||f(x) - f(y)|| \leq L||x - y|| \,\forall x, y \in D.$$

Für D offen heißt f lokal lipschitzstetig, falls $\forall x \in D$ eine Umgebung U von x und eine Konstante L existieren, so dass

$$||f(z) - f(y)|| \leq L||z - y|| \, \forall y, z \in U.$$

Korollar 3.7 *(Kriterium zur Lipschitzstetigkeit)*

1. *Falls f lipschitzstetig ist, so ist f stetig.*
2. *Sei f lipschitzstetig mit Lipschitzkonstante L. Falls $L < 1$, so ist f kontrahierend. Kontrahierende Funktionen sind lipschitzstetig, aber nicht notwendig umgekehrt.*
3. *Es seien $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$. Es sei D konvex, und f stetig differenzierbar. Sei*

$$L := \sup_{s \in D} ||f'(s)|| = ||f'||_\infty.$$

   (a) *Falls $L < \infty$, so ist f lipschitzstetig mit der Lipschitzkonstanten L. Dies ist insbesondere der Fall für D kompakt.*

   (b) *Falls $L < 1$, so ist f kontrahierend mit der Kontraktionskonstanten $q = L$.*

   (c) *Falls $L \geq 1$, so ist f nicht kontrahierend.*

   (d) *Falls $L = \infty$, so ist f nicht lipschitzstetig.*

Beweis: zu 3a für den Spezialfall $n = m = 1$. Seien $x, y \in D$. Da D konvex ist, gibt es ein $\xi \in [x, y]$ (Mittelwertsatz) mit

$$f(x) - f(y) = f'(\xi)\,(x - y).$$

Also gilt

$$|f(x) - f(y)| = |f'(\xi)|\,|x - y| \leq L|x - y|.$$

Rest in den Übungen.
Für $n > 1$ ist f' die Jakobimatrix von f. In diesem Fall ist $||f'(s)||$ die induzierte Matrixnorm (siehe Kapitel über Matrixnormen). □
Der dritte Teil dieses Korollars gibt uns ein sehr nützliches Kriterium dafür, ob eine Funktion kontrahierend ist, ohne über die Definition gehen zu müssen.

3.3 Picard–Lindelöf bei globaler Lipschitzstetigkeit

Satz 3.8 *(Picard–Lindelöf, globale Version)*
In der Differentialgleichung 3.4 sei I ein abgeschlossenes Intervall und $J = \mathbb{R}^n$.

Weiter sei f stetig und lipschitzstetig in der zweiten Variablen mit Lipschitzkonstante L, d.h.

$$||f(t,y) - f(t,z)|| \leq L||y - z|| \, \forall t \in I, y, z \in \mathbb{R}^n.$$

Sei weiter $a \in I$ und $y_0 \in \mathbb{R}^n$. Dann gibt es genau eine Funktion

$$y : I \mapsto \mathbb{R}^n : \, y'(t) = f(t, y(t)) \, \forall y \in I, \, y(a) = y_0.$$

Ihr erstes Gefühl sollte sein: Das kann nicht stimmen. Denn die Lösung existiert auf ganz I – wie kann das sein, angesichts von Beispiel 2.4?

Tatsächlich sind dort die Voraussetzungen nicht erfüllt, f ist dort nicht lipschitzstetig. Tatsächlich ist der Satz in dieser Form praktisch unbrauchbar, weil nur wenige Funktionen lipschitzstetig im $\mathbb{R}^n$ sind. Üblicherweise sind sie es nur auf kleinen Teilmengen.

Diesen Fall wird der dann folgende Satz (lokale Version von Picard–Lindelöf) abdecken.

Beweis: Es sei

$$\epsilon = \frac{1}{2L}, \, I_1 := \{t \in I : |t - a| \leq \epsilon\}.$$

I_1 ist also ein abgeschlossenes Intervall.
Es sei nun X der Raum $C_0(I_1)$ der stetigen Funktionen auf dem Intervall I_1, versehen mit der Supremumsnorm

$$||y||_\infty = \sup_{s \in I_1} ||y(s)||.$$

Nach Analysis II (z.B. Forster II, Paragraph 2, Satz 9) ist X vollständig. Wir definieren die Funktion

$$g : X \mapsto X, \, (gy)(t) := y_0 + \int_a^t f(s, y(s)) \, ds.$$

g ist also eine Funktion, die eine stetige Funktion auf eine stetige Funktion abbildet.
g ist eine Selbstabbildung auf X.
Wir zeigen: g ist kontrahierend bzgl. der Supremumsnorm. Seien also $y, z \in X$.

Dann gilt

$$\begin{aligned}
||g(y) - g(z)||_\infty &= \sup_{t\in I_1} ||(gy)(t) - (gz)(t)|| \\
&= \sup_{t\in I_1} \int_a^t ||f(s, y(s)) - f(s, z(s))||\, ds \\
&\leq \sup_{t\in I_1} \int_a^t ||f(s, y(s)) - f(s, z(s))||\, ds \\
&\leq \sup_{t\in I_1} \int_a^t L\, \|y(s) - z(s)\|\, ds \\
&\leq \sup_{t\in I_1} \int_a^t L||y - z||_\infty\, ds \\
&= \sup_{t\in I_1} |t - a|L||y - z||_\infty \\
&\leq \frac{1}{2}||y - z||_\infty
\end{aligned}$$

und damit ist g kontrahierend mit Kontraktionskonstante $\frac{1}{2}$. Wir haben alles zusammen für den Fixpunktsatz von Banach. Es gibt genau eine Funktion $y \in X$ mit $g(y) = y$, und diese ist nach 3.5 Lösung der AWA.
Diese Lösung y ist natürlich nur auf dem Intervall I_1 definiert. Angenommen, I geht rechts noch weiter, also etwa $[a, a + 2\epsilon] \in I$. Dann betrachten wir das gleiche Anfangswertproblem mit Anfangswerten $(a + \epsilon, y(a + \epsilon))$, d.h.

$$z'(t) = f(t, z(t)),\ z(a + \epsilon) = y(a + \epsilon).$$

Wie wir gerade gezeigt haben, gibt es eine Lösung z dieses Problems auf dem Intervall $[a, a+2\epsilon]$, und y und z stimmen auf $[a, a+\epsilon]$ überein (denn der Banachsche Fixpunktsatz garantiert dort eindeutige Lösbarkeit).
Die Lösung lässt sich also auf $[a - \epsilon, a + 2\epsilon]$ fortsetzen, und dies kann man auf dem gesamten Intervall I tun. □

Bemerkung: (Endwertaufgaben)
Picard–Lindelöf garantiert, dass eindeutige Lösungen der Anfangswertaufgabe zu beiden Seiten existieren, d.h. wir können für ein Intervall $[a, b]$ die Anfangsbedingung auch bei b vorgeben.

3.4 Picard–Lindelöf bei Lipschitzstetigkeit auf einem Streifen

Im lokalen Satz von Picard–Lindelöf schwächen wir die Voraussetzung ab, erhalten aber entsprechend auch nur eine schwächere Aussage.

Satz 3.9 *(Picard–Lindelöf, lokale Version; Streifensatz)*
Es sei alles wie in 3.8, aber jetzt sei $J \subset \mathbb{R}^n$ kompakt. y_0 liege im Inneren von J. Sei ϵ so klein, dass die abgeschlossene ϵ–Umgebung U von y_0 ganz in J liegt. Dann besitzt das Anfangswertproblem 3.4 genau eine Lösung auf dem Intervall

$$I_0 := \{t \in I : |t-a| \leq \delta\},\ \delta = \frac{\epsilon}{M},\ M = \sup_{I\times J} |f(t,y)|.$$

$I_0 \times U$ heißt Streifen.

Zur Verdeutlichung wenden wir zunächst 3.8 auf Beispiel 2.4 an. Dort war

$$f(t,y) = 1 + y^2.$$

Angenommen, f ist lipschitzstetig im zweiten Argument mit Lipschitzkonstante L. Es gilt aber

$$|f(t,L+1) - f(t,L)| = |1 + (L+1)^2 - (1+L^2)| = |2L+1| > L = L|(L+1) - L|.$$

Also ist f nicht (global) lipschitzstetig, und 3.8 ist nicht anwendbar.

Sei nun J ein abgeschlossenes endliches Intervall. Es gilt $f_y(t,y) = 2y$ und

$$L := \sup_{y\in J} |2y| < \infty.$$

Mit dem Mittelwertsatz gibt es für alle $x, y \in J$ ein $\xi \in J$ mit

$$|f(t,y) - f(t,z)| = |f_y(t,\xi)|\,|y-z| \leq L|y-z|.$$

f ist also lipschitzstetig im Sinne von 3.9, und der Satz ist anwendbar.

Beweis: Wir führen den Beweis mit Hilfe von 3.8. Hierzu setzen wir zunächst f stetig auf $I_0 \times \mathbb{R}^n$ fort. Für $n = 1$ und J Intervall gelingt dies mit der folgenden Konstruktion. Sei

$$c_{\max} = \max_{y\in J} y,\ c_{\min} = \min_{y\in J} y.$$

Sei nun $y > c_{\max}$, $t \in I_0$. Dann definieren wir

$$\widetilde{f}(t,y) := \begin{cases} f(t,c_{\max}) & y > c_{\max} \\ f(t,y) & c_{\min} \leq y \leq c_{\max} \\ f(t,c_{\min}) & y < c_{\min} \end{cases}.$$

d.h. $\widetilde{f}$ ist eine Fortsetzung von f, bei dem wir den obersten Wert, auf dem f noch definiert ist, nach oben fortsetzen. Man sieht sofort: Das so definierte f ist global stetig und lipschitzstetig mit Lipschitzkonstante L im zweiten Argument.
Die Voraussetzungen von 3.8 für die Funktion $\widetilde{f}$ sind nun erfüllt, d.h. es gibt eine Lösung der AWA für $\widetilde{y}$ auf I_0. Nach der Definition von $\widetilde{f}$ gilt für $y(t) \in J$ also

$$y'(t) = f(t, y(t)),\ y(a) = y_0.$$

Sei nun $t \in I_0$, d.h. $|t - a| \leq \delta = \frac{\epsilon}{M}$.

$$|y(t)) - y(a)| = |\int_a^t f(s, y(s))\, ds| \leq |t - a|M \leq \epsilon$$

und damit gilt $y(t) \in J$, und insbesondere ist y Lösung der AWA auf dem Intervall I_0. □
Dieser Satz lässt sich auch geometrisch deuten.

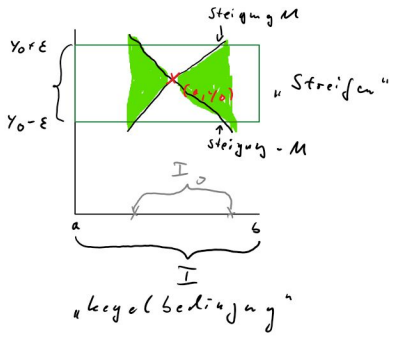


Abbildung 3.1: Kegelbedingung

Auf dem Streifen stimmen die AWA und das Hilfsproblem überein. Da $|y'(t)| = |f(t, y(t))| \leq M$, liegt die im Hilfsproblem berechnete Lösung im grünen Kegel. Solange dieser Kegel also ganz in D liegt, ist die Lösung des Hilfsproblems auch eine Lösung der AWA (Kegelbedingung).

3.5 Picard–Lindelöf bei lokaler Lipschitzstetigkeit: Maximales Existenzintervall

Es sei nun f nur noch lokal lipschitzstetig auf einer offenen Menge D. Sei $(a, y_0) \in D$. Da f lokal lipschitzstetig ist, ist es lipschitzstetig auf einer kleinen Umgebung

von (a, y_0). Diese Umgebung enthält einen Streifen wie in 3.9, es existiert also eine Lösung der zugehörigen Anfangswertaufgabe auf einem Intervall I_0, das a enthält. Wir untersuchen, welche Fälle hier auftreten können.

Der Zeichnung kann man sehr schön entnehmen, warum für einen Streifen die Lösung u.U. nicht global ist: Sobald der Graph $(t, y(t))$ die Menge D verlässt, also den Rand schneidet, ist die Lösung des Hilfsproblems für $\widetilde{f}$ keine Lösung der AWA mehr. Tatsächlich zeigt man den Satz:

Satz 3.10 *(Fortsetzbarkeit von Lösungen)*
Es sei $D \subset \mathbb{R} \times \mathbb{R}^n$ offen. f sei stetig und genüge einer lokalen Lipschitzbedingung auf D. Weiter sei $(a, y_0) \in D$. Dann hat nach 3.9 die Anfangswertaufgabe 3.4 genau eine auf einer Umgebung U von a definierte Lösung y. U kann so gewählt werden, dass y über U hinaus nicht fortgesetzt werden kann, U heißt dann maximales Existenzintervall.
Der Abschluss G des Graphen $\{(t, y(t)) : t \in U\}$ kommt dabei dem Rand beliebig nahe, d.h. es gibt keine kompakte Teilmenge $K \subset D$, die G enthält.

(Beweis: Walter, Paragraph 6, Satz 7).

Ein typischer Fall ist Beispiel 2.4: f ist lokal lipschitzstetig auf $D = \mathbb{R} \times \mathbb{R}$. Die Lösung des Anfangswertproblems mit $y(0) = 0$ ist $y(t) = \tan t$. Das maximale Existenzintervall ist das offene Intervall $(-\pi/2, \pi/2)$. An den Enden des Intervalls geht $|y(t)|$ gegen ∞. Der Graph $(t, y(t))$ ist unbeschränkt in der zweiten Variablen, also liegt der Graph in keiner kompakten Teilmenge von D.

Es sei $b = \sup_{t \in U} t$. Dann können für den rechten Rand des Intervalls nur drei Varianten auftreten:

1. $b < \infty$ und G schneidet den Rand von D (siehe Zeichnung). Insbesondere kommt G dem Rand beliebig nahe.
2. $b = \infty$, d.h. die Lösung existiert für alle $t \geq a$.
3. Falls weder 1 noch 2: Dann gilt $\limsup_{t \to b^-} |y(t)| = \infty$. In diesem Fall ist das Intervall rechts offen.

3.6 Konstruktive Lösung der AWA mit Banach

Der Banachsche Fixpunktsatz ist konstruktiv – er gibt einen Algorithmus an, mit dem sich der Fixpunkt ausrechnen lässt. Es gilt:

Korollar 3.11 *(Berechnung der Lösung von AWA mit Banach)*
Es sei alles wie in 3.8, und g definiert wie im Beweis dort, also

$$g : C_0(I_1) \mapsto C_0(I_1),\ (gy)(t) = y_0 + \int_a^t f(s, y(s))\, ds.$$

Insbesondere sei f stetig und lipschitzstetig im zweiten Argument, und I_1 sei so klein, dass g kontrahierend ist bezüglich $||\cdot||_\infty$. Sei

$$y^{(0)} \in C_0(I_1),\ y^{(k+1)} = g(y^{(k)}).$$

Dann konvergiert die Funktionenfolge $y^{(k)}$ gegen die Lösung der Anfangswertaufgabe bzgl. $||\cdot||_\infty$.

Beweis: Fixpunktsatz von Banach. □

Beispiel 3.12
Gegeben sei die AWA

$$y'(t) = y(t),\ y(0) = 1.$$

Das zugehörige f ist global lipschitzstetig in y, denn die Ableitung nach y ist 1. Wir wählen $y^{(0)} = 1$. Dann gilt

$$\begin{aligned} y^{(0)}(t) &= 1 \\ y^{(1)}(t) = 1 + \int_0^t 1\, ds &= 1 + t \\ y^{(2)}(t) = 1 + \int_0^t 1 + s\, ds &= 1 + t + \frac{t^2}{2}. \end{aligned}$$

Per Induktion:

$$y^{(k)}(t) = \sum_{j=0}^{k} \frac{t^j}{j!}$$

und diese Funktionenfolge konvergiert gegen die Lösung e^t der Differentialgleichung, auf jedem kompakten Intervall I_1 sogar gleichmäßig.

3.7 Existenzsatz von Peano

Man kann sich fragen, ob man auf die Lipschitzstetigkeit im Satz von Picard–Lindelöf verzichten kann. Wir betrachten das (bei $(0, 0)$ nicht lokal lipschitzstetige)

Beispiel 3.13

$$y'(t) = \sqrt{|y(t)|},\ y(0) = 0.$$

Dies hat offensichtlich die Lösung $y(t) = 0$, aber auch $y(t) = \frac{1}{4}t\,|t|$ ist eine Lösung. Wir erhalten also eine Lösung, aber die Lösung ist nicht eindeutig. Dies ist die Aussage des Existenzsatzes von Peano.

Satz 3.14 *(Existenzsatz von Peano)*
Sei f stetig in einer Umgebung von (a, y_0). Dann gibt es ein Intervall I mit $a \in I$ und eine stetig differenzierbare Funktion $y : I \mapsto \mathbb{R}^n$ mit

$$y'(t) = f(t, y(t)),\ y(a) = y_0.$$

Beweis: z.B. Walter, Paragraph 7. Idee: Der Beweis zeigt wieder, dass die Funktion g aus 3.8 einen Fixpunkt besitzt, benutzt aber statt dessen den Fixpunktsatz von Schauder/Brouwer für relativ kompakte Mengen, und den Satz von Arzela–Ascoli, der die relativ kompakten Teilmengen der stetigen Funktionen auf einem Intervall charakterisiert. Die einzelnen Sätze findet man auch bei Harro Heuser, Analysis I/II.

3.8 Zusammenfassung

3.8.1 Kompetenzen

- Fixpunktsatz von Banach mit allen seinen Voraussetzungen kennen.
- Kontraktionseigenschaft mit Hilfe der Ableitung nachrechnen.
- Existenz– und Eindeutigkeitssätze kennen und Voraussetzungen nachrechnen (insb. Lipschitzstetigkeit).
- Kegelbedingung kennen und interpretieren können

3.8.2 Mini–Aufgaben

- Welche der folgenden Funktionen sind (global oder lokal) lipschitzstetig im zweiten Argument?

$$f_1(t, y) = |t|^{3/2},\ f_2(t, y) = t^n,\ f_3(t, y) = t^{1/2} \text{ für } t > 1.$$

- Ist die Funktion $x + e^{-x}$, $x \geq 0$, kontrahierend?
- Lösen Sie die AWA $y'(t) = 2y(t),\ y(0) = 1$ konstruktiv mit dem Fixpunktsatz von Banach.

Kapitel 4

Iterative Lösung von Gleichungen mit Fixpunktiterationen

Wir haben in den Vorbemerkungen gesehen, dass zur Lösung von Randwertproblemen große Gleichungssysteme gelöst werden müssen. In diesem Kapitel tun wir dies iterativ mit dem Fixpunktsatz von Banach. Wir beginnen mit Beispielen für nichtlineare Gleichungen und dem Newton–Verfahren und gehen dann zu linearen Gleichungssystemen über.

4.1 Fixpunkte von nichtlinearen Gleichungen

Wir betrachten einige Beispiele, der Einfachheit halber über $\mathbb{R}$. Zu berechnen sei jeweils ein Fixpunkt von g.

Beispiel 4.1

1. $$g : \mathbb{R} \mapsto \mathbb{R},\ g(x) := 0.9\ \cos(x)$$

   erfüllt die Voraussetzungen des Banachschen Fixpunktsatzes. Seien $x, y \in \mathbb{R}$. *Dann ist nach dem Mittelwertsatz*

   $$|g(x) - g(y)| = |g'(\xi)(x - y)| = 0.9\,|-\sin(\xi)|\,|x - y| \leq 0.9\,|x - y|.$$

   Also ist g *kontrahierend mit der Kontraktionskonstanten* 0.9*, Selbstabbildung,* $\mathbb{R}$ *ist vollständig und abgeschlossen, also sind alle Voraussetzungen erfüllt.*

2.

$$g : \mathbb{R} \mapsto \mathbb{R}\, g(x) := \cos(x)$$

ist nicht kontrahierend, denn

$$\lim_{\frac{\pi}{2} \neq x \mapsto \frac{\pi}{2}} \frac{|\cos(x) - \cos(\frac{\pi}{2})|}{|x - \frac{\pi}{2}|} = |-\sin(\frac{\pi}{2})| = 1$$

und damit gibt es kein $q < 1$, das diesen Ausdruck nach oben begrenzt.

3.

$$g : [0, 0.1] \mapsto [0.99, 1],\ g(x) = \cos(x)$$

ist kontrahierend, aber keine Selbstabbildung. Beweis der Abbildungseigenschaft: $\cos(x)$ ist monoton fallend auf $[0, 0.1]$, nimmt also seinen größten Wert in 0 an (1) und seinen kleinsten in 0.1 (~ 0.995).

4.

$$D := [0.6, 0.9],\ g : D \mapsto D,\ g(x) = \cos(x)$$

erfüllt die Voraussetzungen des Banachschen Fixpunktsatzes.
Beweis: $\cos(x)$ ist in diesem Bereich monoton fallend. Es gilt $\cos(0.6) < 0.9$, aber $\cos(0.9) > 0.6$, also wird D auf D abgebildet.
Seien nun $x, y \in D$. Dann gilt wie oben

$$g(x) - g(y) = g'(\xi)(x - y) = -\sin(\xi)(x - y).$$

Da D konvex ist und ξ zwischen x und y liegt, gilt auch $\xi \in D$. Der Sinus ist monoton steigend auf D, nimmt also seinen größten Wert in 0.9 an. $\sin(0.9) < 0.8 =: q$, und damit ist g kontrahierend mit der Kontraktionskonstante q. Da $\mathbb{R}$ Banachraum, D abgeschlossen, sind alle Voraussetzungen des Banachschen Fixpunktsatzes erfüllt.

Beispiel 4.2 *Häufig ist die zielführende Formulierung der Fixpunktgleichung nicht sofort klar. Wir suchen einen Fixpunkt von $\tan x$ in $[\pi/2, 3\pi/2]$. Die Wahl*

$$g(x) = \tan x,\ g'(x) = \frac{1}{\cos^2 x} \geq 1$$

führt offensichtlich nicht zum Ziel, denn g ist dann nicht kontrahierend, am Rand sogar nicht einmal definiert. Wir formen daher die Fixpunktgleichung um. Hier wählen wir

$$\overline{x} = \tan(\overline{x}) \Longleftrightarrow \arctan \overline{x} = \overline{x} - \pi \Longleftrightarrow \overline{x} = \pi + \arctan \overline{x}.$$

Mit

$$D = [\frac{\pi}{2}, \frac{3\pi}{2}],\ g(x) = \pi + \arctan x,\ g'(x) = \frac{1}{1 + x^2}$$

bildet g D auf D ab und ist damit kontrahierend mit Kontraktionskonstante

$$q = \frac{1}{1+\pi^2/4} \sim 0.29.$$

D enthält den gesuchten Fixpunkt. Wir führen die Fixpunktiteration durch mit dem Startwert $x^{(0)} = \pi$ und erhalten

k	$x^{(k)}$	*a priori*	*a posteriori*	$\lvert\overline{x} - x^{(k)}\rvert$	$\lvert\overline{x} - x^{(k)}\rvert / \lvert\overline{x} - x^{k-1}\rvert$
1	3.1416			$1.3518e+000$	
2	4.4042	$1.4758e-001$	$1.7744e+000$	$8.9190e-002$	$6.5978e-002$
3	4.4891	$4.2563e-002$	$1.1931e-001$	$4.2900e-003$	$4.8100e-002$.
4	4.4932	$1.2275e-002$	$5.7439e-003$	$2.0266e-004$	$4.7239e-002$
5	4.4934	$3.5401e-003$	$2.7132e-004$	$9.5871e-006$	$4.7307e-002$
6	4.4934	$1.0210e-003$	$1.2804e-005$	$4.7571e-007$	$4.9619e-002$

Offensichtlich ist die a priori–Abschätzung viel zu pessimistisch.

Satz 4.3 *(**Lokaler Konvergenzsatz**)*
Sei X vollständig, $g : U \mapsto X$, $U \subset X$, kontrahierend. Sei $\overline{x}$ ein Fixpunkt von g im Inneren von U. Dann gibt es eine Umgebung D von $\overline{x}$, so dass die Fixpunktiteration mit Startwerten in D gegen $\overline{x}$ konvergiert.

Beweis: Sei q die Kontraktionskonstante von g. Sei D eine abgeschlossene Kugel um $\overline{x}$ mit Radius $\epsilon > 0$, die ganz in U liegt. Wir zeigen: g bildet D in sich selbst ab. Sei $x \in D$.

$$\begin{aligned} ||g(x) - \overline{x}|| &= ||g(x) - g(\overline{x})|| \\ &\leq q||x - \overline{x}|| \\ &\leq q \cdot \epsilon < \epsilon. \end{aligned}$$

Also gilt $g(x) \in D$, und die Aussage folgt mit 3.3. □

Korollar 4.4 *Sei $g : U \mapsto \mathbb{R}^n$ stetig differenzierbar, $U \subset \mathbb{R}^n$, $\overline{x}$ ein Fixpunkt von g, und sei*

$$||g'(\overline{x})|| < 1.$$

Dann gibt es eine Umgebung D von $\overline{x}$, so dass die Fixpunktiteration mit Startwerten in D gegen $\overline{x}$ konvergiert.

Beweis: g' ist stetig, also gibt es eine Umgebung U' von $\overline{x}$ mit

$$||g'(x)|| \leq \frac{1 + ||g'(\overline{x})||}{2} < 1 \, \forall x \in U'.$$

□

4.2 Newton–Verfahren

Die kurze Behandlung in diesem Abschnitt wird der Bedeutung der Newton–Verfahren nicht gerecht. Tatsächlich ist das Newton–Verfahren eins der am häufigsten genutzten numerischen Verfahren, die Konvergenzanalyse ist aber recht übersichtlich.

Sei zunächst $f : \mathbb{R} \mapsto \mathbb{R}$ stetig differenzierbar. Wir suchen eine Nullstelle $\overline{x}$ von f. Dazu müssen wir zunächst

$$f(\overline{x}) = 0$$

in eine Fixpunkgleichung umwandeln. Es bietet sich an eine Formulierung wie

$$x = x - f(x) =: g(x).$$

Damit die zugehörige Fixpunktiteration konvergiert, muss gelten

$$|g'(\overline{x})| < 1 \Longleftrightarrow f'(\overline{x}) \in (0, 2).$$

Diese Bedingung legt nahe, f mit $1/f'(x)$ zu multiplizieren. Dazu gibt es eine geometrische Motivation.
Sei $x^{(0)}$ eine Näherung für $\overline{x}$. Wir approximieren die Funktion f in der Nähe des Punktes $(x^{(0)}, f(x^{(0)}))$ durch ihre Tangente, und suchen statt einer Nullstelle der Funktion die Nullstelle der Tangente. Falls $x^{(0)}$ nah an $\overline{x}$ liegt, so ist diese Approximation gut. Die Tangentenfunktion hat die Darstellung

$$T(x) = f(x^{(0)}) + f'(x^{(0)}) \cdot (x - x^{(0)})$$

mit der Nullstelle

$$x^{(0)} - f'(x^{(0)})^{-1} f(x^{(0)}).$$

Wir setzen also für eine gegebene Näherung

$$x^{(k+1)} = g(x^{(k)}),\ g(x) := x - (f'(x))^{-1} f(x),$$

und erhalten die Fixpunktiteration zu g, das Newton–Verfahren zur Bestimmung einer Nullstelle von f. So, wie wir es aufgeschrieben haben, ist das Verfahren auch in höheren Dimensionen definiert (hier ist dann $f'(x)$ die Jakobimatrix).

Wir formulieren die folgenden Sätze für den $\mathbb{R}^n$, betrachten sie dann aber aus Zeitgründen nur für $n = 1$. Alle Sätze bleiben in höheren Dimenstionen korrekt und natürlich auf Teilmengen des $\mathbb{R}^n$ anwendbar. Die Beweise sind immer übertragbar. Sollte dies nicht der Fall sein, steht dies immer dabei.

Definition 4.5 *(**Newton–Verfahren**, auch **Newton–Raphson–Verfahren**)*
Sei $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ differenzierbar. Sei $x^{(0)} \in \mathbb{R}^n$. Falls die auftretenden Ableitungen $f'(x^{(k)})$ invertierbar sind für alle $k \in \mathbb{N}$, so heißt die Folge mit

$$x^{(k+1)} = x^{(k)} - (f'(x^{(k)}))^{-1} f(x^{(k)})$$

Newton–Verfahren zur Bestimmung einer Nullstelle von f. Hierbei ist $f'(x)$ für $n > 1$ die Jakobimatrix von f an der Stelle x. Für $n = 1$ ist natürlich einfach

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}.$$

Wir führen zunächst ein Maß für die Geschwindigkeit der Konvergenz einer Folge ein.

Definition 4.6 *(Konvergenzordnung, lineare, quadratische Konvergenz, Landau–Symbole)*
Es sei $x^{(k)} \in \mathbb{R}^n$ eine konvergente Folge mit Grenzwert $\overline{x}$.

1. *Falls*
$$\exists\, 0 \leq q < 1,\, k_0 > 0 :\ ||x^{(k+1)} - \overline{x}|| \leq q||x^{(k)} - \overline{x}|| \forall\, k > k_0,$$
so heißt $x^{(k)}$ (mindestens) konvergent von der Ordnung 1 oder linear konvergent.

2. *Sei $p > 1$. Falls*
$$\exists\, C > 0,\, k_0 > 0 :\ ||x^{(k+1)} - \overline{x}|| \leq C\, ||x^{(k)} - \overline{x}||^p \,\forall\, k > k_0,$$
so heißt $x^{(k)}$ (mindestens) konvergent von der Ordnung p. Für $p = 2$ heißt $x^{(k)}$ auch quadratisch konvergent.

3. *Sei nun $\epsilon^{(k)}$ eine reelle Nullfolge, die konvergent von der Ordnung p ist. Falls*
$$\exists\, C > 0 : ||x^{(k)} - \overline{x}|| \leq C\epsilon^{(k)},$$
so heißt auch $x^{(k)}$ konvergent von der Ordnung p.

4. *Es seien f, g zwei reellwertige Funktionen. Falls*
$$\exists\, C > 0,\, h_0 > 0 : |f(h)| \leq C|g(h)| \,\forall\, h : |h| < h_0,$$
so schreiben wir kurz
$$f(h) = O(g(h)) \text{ für } h \to 0.$$

Entsprechend: Falls

$$\exists C > 0,\, n_0 > 0 : |f(n)| \leq |g(n)| \,\forall\, n > n_0,$$

so schreiben wir kurz

$$f(n) = O(g(n)) \textit{ für } n \to \infty.$$

Bemerkung:

1. Für eine kontrahierende Funktion g ist die zugehörige Fixpunktfolge linear konvergent:

$$||x^{(k+1)} - \overline{x}|| \leq ||g(x^{(k)}) - g(\overline{x})|| \leq q||x^{(k)} - \overline{x}||.$$

2. Für linear konvergente Folgen mit $q = 1/10$ gewinnt man in jedem Schritt eine Dezimalstelle.
3. Wenn eine Folge quadratisch konvergent ist, so verdoppeln sich in jedem Schritt die Anzahl der gültigen Dezimalstellen.
4. Je höher die Ordnung, desto schneller konvergiert die Folge und desto weniger aufwändig kann man den Grenzwert berechnen.
   Für lineare Verfahren: Je kleiner die Kontraktionskonstante, desto besser.
5. Sei $0 \leq q < 1$. Dann ist
$$x^{(k)} = q^{(p^k)}$$
   konvergent von der Ordnung p. Beweis:
$$|x^{(k+1)} - 0| = q^{p^{k+1}} = q^{p^k\, p} = (q^{p^k})^p = |x^{(k)} - 0|^p.$$
6. Unter der Voraussetzung von Teil (3) der Definition gilt
$$||\overline{x} - x^{(k)}|| = O(\epsilon^{(k)}).$$
7. In der Übung zum Sekantenverfahren 4.10 wird deutlich werden, warum man Teil (3) der Definition benötigt, der in Büchern häufig fehlt.

Lemma 4.7 *Es sei $g : \mathbb{R} \mapsto \mathbb{R}$, $g \in C^p$. Weiter sei $\overline{x}$ ein Fixpunkt von g und $g^{(j)}(\overline{x}) = 0$, $j = 1 \ldots p-1$, $p > 1$. Dann gibt es eine abgeschlossene ϵ–Umgebung U von $\overline{x}$, so dass die Fixpunktfolge $x^{(k)}$ konvergent von der Ordnung p ist für Startwerte $x^{(0)} \in U$.*

Beweis: Da $g'(\overline{x}) = 0$, gibt es nach 4.3 gibt es eine abg. ϵ–Umgebung U von $\overline{x}$, so dass die Fixpunktfolge mit Startwerten in U gegen $\overline{x}$ konvergiert. Sei $x^{(k)}$ eine solche Folge, und

$$C = \max_{\xi \in U} |g^{(p)}(\xi)|.$$

Dann gilt mit Taylorentwicklung von g für ein $\xi \in U$ (Restglied von Lagrange)

$$\begin{aligned}
|x^{(k+1)} - \overline{x}| &= |g(x^{(k)}) - g(\overline{x})| \\
&= |\sum_{j=0}^{p-1} g^{(j)}(\overline{x}) \frac{(x^{(k)} - \overline{x})^j}{j!} + g^{(p)}(\xi) \frac{(x^{(k)} - \overline{x})^p}{p!} - g(\overline{x})| \\
&= |g^{(p)}(\xi) \frac{(x^{(k)} - \overline{x})^p}{p!}| \\
&\leq \frac{C}{p!} |x^{(k)} - \overline{x}|^p.
\end{aligned}$$

□

Satz 4.8 *(**Konvergenz des Newtonverfahrens**)*
Sei $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ zweimal stetig differenzierbar. Sei $\overline{x}$ eine Nullstelle von f.

1. *Sei $f'(\overline{x})$ invertierbar. Dann gibt es eine Umgebung U von $\overline{x}$, so dass das Newtonverfahren*
$$x^{(k+1)} = g(x^{(k)}),\ g(x) := x - f'(x)^{-1} f(x)$$
für $x^{(0)} \in U$ gegen $\overline{x}$ konvergiert (lokale Konvergenz). Die Ordnung der Konvergenz ist quadratisch.

2. *Falls $f'(\overline{x})$ nicht invertierbar ist, aber $f'(x)$ invertierbar ist in einer kleinen Umgebung von $\overline{x}$ für $x \neq \overline{x}$, ist das Newtonverfahren immer noch lokal konvergent, aber die Konvergenz ist nur noch linear.*

Beweis: Sei also $n = 1$.

1. $f'(\overline{x})$ ist invertierbar, also gibt es auch eine kleine Umgebung U' von $\overline{x}$, so dass $f'(x)$ invertierbar ist für $x \in U'$ und damit g auf U' wohldefiniert ist. Nach Lemma 4.4 ist für die Konvergenz nur zu zeigen: $|g'(\overline{x})| < 1$. Es gilt
$$g'(\overline{x}) = 1 - \frac{f'(\overline{x}) \cdot f'(\overline{x}) - f(\overline{x}) f''(\overline{x})}{f'(\overline{x})^2} = \frac{f(\overline{x}) f''(\overline{x})}{f'(\overline{x})^2} = 0.$$
Für $g \in C^2$ folgt die quadratische Konvergenz mit 4.7. Ansonsten muss man etwas mehr arbeiten, siehe z.B. Wübbeling [2022].

2. Sei nun
$$f'(\overline{x}) = f(\overline{x}) = 0.$$
Für einen einfachen Beweis nehmen wir an $f''(\overline{x}) \neq 0$. Dann gilt mit l'Hospital
$$g(x) = x - \frac{f(x)}{f'(x)} \xrightarrow[x\to\overline{x}]{} \overline{x} - \frac{f'(\overline{x})}{f''(\overline{x})} = \overline{x}$$
und
$$\begin{aligned} g'(\overline{x}) &= \lim_{x\to\overline{x}} \frac{g(x) - g(\overline{x})}{x - \overline{x}} \\ &= \lim_{x\to\overline{x}} g'(x) \\ &= \lim_{x\to\overline{x}} \frac{f(x)}{f'(x)^2} f''(x) \\ &= \left(\lim_{x\to\overline{x}} \frac{f'(x)}{2f'(x)f''(x)}\right) f''(\overline{x}) \\ &= \frac{1}{2} \end{aligned}$$
und damit haben wir auch in diesem Fall (allerdings nur lineare) Konvergenz nach dem lokalen Konvergenzsatz mit der Kontraktionskonstante $q = \frac{1}{2}$.

□

Beispiel 4.9

1. *Sei $f(x) = x^n - a$, $a > 0$, $n \in \mathbb{N}$. Gesucht wird die Nullstelle $\overline{x} = a^{1/n}$ von f. Für das Newtonverfahren gilt*
$$x^{(k+1)} = x^{(k)} - \frac{(x^{(k)})^n - a}{n(x^{(k)})^{n-1}} = \frac{n-1}{n} x^{(k)} + \frac{a}{n} \frac{1}{(x^{(k)})^{n-1}} =: g(x^{(k)})$$
und
$$g'(x) = \frac{n-1}{n} - \frac{a(n-1)}{n} \frac{1}{x^n} = \frac{n-1}{n} \left(1 - \frac{a}{x^n}\right).$$
Die einzige positive Nullstelle von g' ist $\overline{x}$, und offensichtlich nimmt g für $x > 0$ dort sein Minimum $\overline{x}$ an. Es gilt also
$$g(x) \geq g(\overline{x}) = \overline{x} \, \forall x > 0.$$

Sei nun $x^{(0)} > 0$. Dann ist

$$x^{(k+1)} = g(x^{(k)}) \geq \overline{x} \, \forall k \geq 0.$$

Sei nun $x^{(k)} > 0$. Dann ist

$$x^{(k+1)} = g(x^{(k)}) = x^{(k)} - \underbrace{\frac{(x^{(k)})^n - a}{n(x^{(k)})^{n-1}}}_{\geq 0} \leq x^{(k)}.$$

$x^{(k)}$ ist also monoton fallend und beschränkt und konvergiert gegen einen Grenzwert $\widetilde{x}$ (ab Folgeglied 1). Da g stetig ist, gilt

$$x^{(k+1)} = g(x^{(k)}) \Rightarrow \widetilde{x} = g(\widetilde{x}) \Rightarrow \widetilde{x} = \overline{x}.$$

Das Verfahren ist eine Möglichkeit, die n. Wurzel einer Zahl näherumgsweise zu berechnen und bekannt unter dem Namen Verfahren von Heron.

2. *Formal können wir in Beispiel 1 auch $n = -1$ setzen, also*

$$f(x) = 1/x - a, \; g(x) = x - \frac{1/x - a}{-\frac{1}{x^2}} = x + (x - ax^2) = x(2 - ax).$$

f hat den einzigen Nullpunkt $\overline{x} = 1/a$. Die Argumente aus Beispiel 1 kehren sich gerade um, $\overline{x}$ ist ein Maximum von g und für $0 < x < \overline{x}$ monoton steigend. Das Newtonverfahren konvergiert für $x \in (0, 2/a)$.
Das sieht nicht besonders sinnvoll aus – wir erhalten eine Iteration, die gegen $1/a$ konvergiert. Tatsächlich ist diese Formel schon seit einigen Jahren die wohl mit riesigem Abstand am häufigsten benutzte Anwendung des Newton–Verfahrens.
$g(x)$ benutzt nur Multiplikationen und Additionen. Wir erhalten also ein Verfahren, um den Kehrwert einer Zahl nur mit Multiplikationen und Additionen zu realisieren. Dies ist interessant für CPU–Designer, die sich damit die komplizierte Realisierung der Division in Hardware sparen können. Intel hat die genutzten Algorithmen für seine IA64–Prozessoren offengelegt in Harrison [2000], der Newtonschritt steht in 3.2. Leider ist die dort angesprochene Beispielimplementation nicht mehr verfügbar. In groben Zügen wird zunächst eine Approximation z.B. durch einen Lookup–Table gefunden, die dann durch wenige Schritte des Newton–Verfahrens verbessert wird.

3. *Für $n = 2$ betrachten wir die Funktion*

$$f : \mathbb{R}^2 \mapsto \mathbb{R}^2, \; f(x, y) = \begin{pmatrix} x - \frac{1}{4}(\cos x - \sin y) \\ y - \frac{1}{4}(\cos x - 2 \sin y) \end{pmatrix}.$$

Die Jakobimatrix lautet

$$f'(x,y) = \begin{pmatrix} 1 + 1/4 \sin x & 1/4 \cos y \\ 1/4 \sin x & 1 + 1/2 \cos y \end{pmatrix}.$$

Das g aus der Newton–Iteration ist in der Nähe von $(0.16, 0.2)$ kontrahierende Selbstabbildung, also gibt es dort eine Nullstelle von f. Die Determinante von f' ist positiv (s.u.), daher ist $f'(x,y)$ invertierbar. Also konvergiert das Newtonverfahren bei geeignet gewählten Startwerten quadratisch. Die Iteration lautet

$$\begin{pmatrix} x^{(k+1)} \\ y^{(k+1)} \end{pmatrix} = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} - \frac{1}{\det f'(x^{(k)}, y^{(k)})} \cdot \begin{pmatrix} 1 + 1/2 \cos y^{(k)} & -1/4 \cos x^{(k)} \\ -1/4 \sin x^{(k)} & 1 + 1/4 \sin x^{(k)} \end{pmatrix} \cdot \begin{pmatrix} x^{(k)} - 1/4(\cos x^{(k)} - \sin y^{(k)}) \\ y^{(k)} - 1/4(\cos x^{(k)} - 2 \sin y^{(k)}) \end{pmatrix}$$

mit

$$\det f'(x,y) = (1 + 1/4 \sin x)(1 + 1/2 \cos y) - (1/4 \sin x)(1/4 \cos y).$$

Bemerkung:

1. Im $\mathbb{R}^n$ invertiert man die Jakobi–Matrizen im Newtonverfahren nicht explizit, sondern nutzt statt dessen

$$f'(x^{(k)})(x^{(k)} - x^{(k+1)}) = f(x^{(k)}).$$

2. In jedem Schritt des Newton–Verfahrens muss einmal die Funktion und einmal ihre Ableitung ausgewertet werden. Im $\mathbb{R}^n$ muss zusätzlich ein Gleichungssystem gelöst werden.
   Falls f' nicht explizit zur Verfügung steht, muss es durch Differenzen approximiert werden, wir berechnen also

$$\frac{df}{dx_i}(x^{(k)}) \sim \frac{f(x^{(k)}) - f(x^{(k)} + he_i)}{h}$$

   mit den Einheitsvektoren e_i und berechnen daraus eine Approximation der Jakobimatrix. In diesem Fall werden $n + 1$ Funktionsauswertungen benötigt.

3. Ausdrücklich: Das Newtonverfahren ist im Allgemeinen **nicht** global konvergent. Falls die zugrundeliegende Funktion einen Nullpunkt $\overline{x}$ besitzt, so konvergiert das Newtonverfahren gegen $\overline{x}$, falls der Anfangspunkt nah genug an $\overline{x}$ liegt.

Definition 4.10 *(Vereinfachtes Newtonverfahren, Sekantenverfahren)*
f erfülle die Voraussetzungen des Newtonverfahrens.

1. *Für großes n ersetzt man im Newtonverfahren die Jakobimatrix an der Stelle $x^{(k)}$ durch die Matrix an der Stelle $x^{(0)}$ und spart sich damit die Berechnung der Ableitung. Wir erhalten*

$$f'(x^{(0)})(x^{(k)} - x^{(k+1)}) = f(x^{(k)}).$$

   Diese Iteration heißt **vereinfachtes Newtonverfahren**. *Das vereinfachte Newtonverfahren ist lokal linear konvergent.*

2. *Statt durch die Tangente kann man in einer Dimension die Funktion auch durch die Verbindungsgerade (Sekante) zweier Punkte auf der Kurve approximieren. Hierzu wählt man zwei Startwerte $x^{(0)}$, $x^{(1)}$. Die Verbindungsgerade der zugehörigen Punkte auf der Kurve hat die Gleichung*

$$S(x) = f(x^{(0)}) + (f(x^{(1)}) - f(x^{(0)}))\frac{x - x^{(0)}}{x^{(1)} - x^{(0)}}.$$

   Die Nullstelle dieser Geraden ist

$$x^{(0)} - \frac{x^{(1)} - x^{(0)}}{f(x^{(1)}) - f(x^{(0)})} f(x^{(0)})$$

   und wir erhalten das **Sekantenverfahren**

$$x^{(k+2)} = x^{(k)} - \frac{x^{(k+1)} - x^{(k)}}{f(x^{(k+1)}) - f(x^{(k)})} f(x^{(k)}).$$

   Das Sekantenverfahren ist lokal konvergent mit der Konvergenzordnung $\frac{1+\sqrt{5}}{2} \sim 1.62$ (Übungen).

3. *Durch Berücksichtigung von weiteren Termen in der Taylorentwicklung (neben der Linearisierung) kann man Verfahren höherer Ordnung herleiten.*

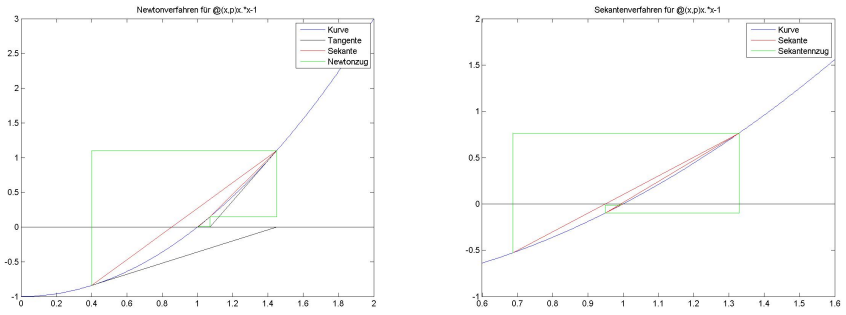


Abbildung 4.1: Newtonverfahren und Sekantenverfahren für $x^2 - 1$ und Startwert 0.7

Klick für Bild Newton
Klick für Matlab Figure Newton
Klick für Bild Sekante
Klick für Matlab Figure Sekante

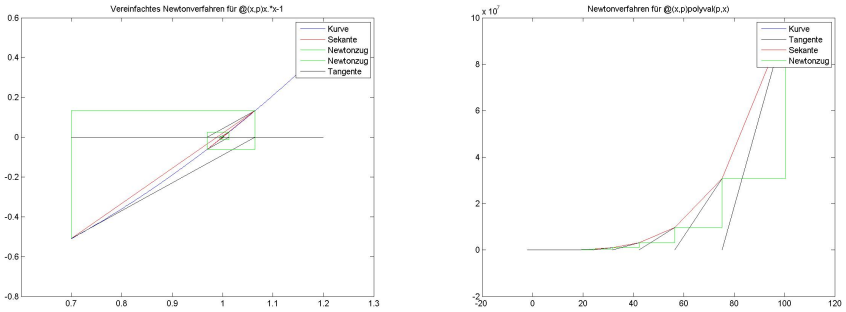


Abbildung 4.2: Vereinfachtes Newtonverfahren und typisches Verhalten bei Nicht–Konvergenz

Klick für Bild Vereinfacht
Klick für Matlab Figure Vereinfacht
Klick für Bild Newtonnoconv
Klick für Matlab Figure Newtonnoconv

4.3 Homotopiemethoden und der Satz von Gerschgorin

Ein Problem beim Newtonverfahren ist das Finden einer geeigneten Anfangsnäherung. Hat man eine solche, konvergiert das Newton–Verfahren meist mit wenigen Schritten.

Geeignete Anfangsnäherungen und global konvergente Verfahren lassen sich mit **Homotopiemethoden** gewinnen.

Gesucht sei die Nullstelle von $f(x)$. Wir definieren eine Funktion $f(x,t)$, $t \in [0,1]$, mit den Eigenschaften:

1. $f(x,1) = f(x)$.
2. $f(x,t)$ ist zweimal stetig differenzierbar in x.
3. Die Nullstelle $x(t)$ von $f(x,t)$ hängt stetig von t ab.
4. Die Nullstelle $x(0)$ lässt sich einfach bestimmen.

Damit können wir die Nullstellen $x(t)$ verfolgen. Sei $h = 1/N$ und N fest. Ausgehend von der Nullstelle $x(0)$ bestimmen wir mit einigen Schritten des Newton–Verfahrens eine Näherung für $x(h)$. Da die Nullstellen stetig von t abhängen, wird das Newton–Verfahren schnell konvergieren, falls h klein genug ist. Ausgehend von dieser Näherung an $x(h)$ bestimmen wir dann eine Näherung an $x(2h)$ usw. bis zur Nullstelle $x(1)$ von $f(x)$.

Im Matlab–Beispiel wird eine Homotopiemethode gerechnet zur Bestimmung der Nullstellen von

$$p(x) = x^4 - 3x^3 + 5x^2 + x - 2$$

mit der Homotopiefunktion

$$f(x,t) = (1-t)(x^4 - 1) + tp(x).$$

Es illustriert das große Problem der Homotopiemethoden: Will man alle Nullstellen einer Funktion bestimmen, muss man, sobald zwei Nullstellen zusammen– und wieder auseinanderlaufen (Bifurkation), sicherstellen, dass man alle Zweige weiterverfolgt (dies ist im Programm nicht der Fall, deshalb erhält man am Ende nur drei der vier Nullstellen).

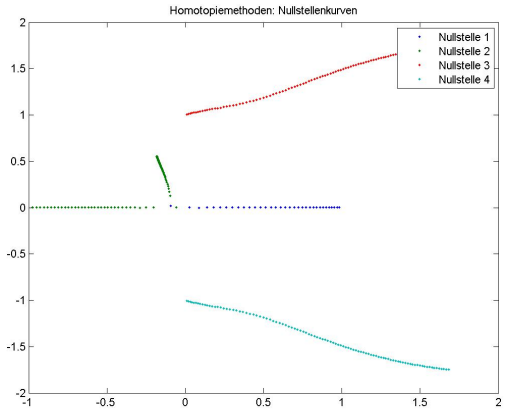


Abbildung 4.3: Nullstellen $x_k(t)$ für $f(x,t)$.

Hier wurde stillschweigend vorausgesetzt, dass die Nullstellen eines Polynoms stetig von den Koeffizienten abhängen. Unter dieser Voraussetzung kann man mit Hilfe der Homotopie–Methoden einen interessanten Satz beweisen.

Satz 4.11 *(**Satz von Gerschgorin**)*
Sei $A = (a_{ij}) \in \mathbb{C}^{n\times n}$. Sei $K_i \in \mathbb{C}$ (also in der komplexen Ebene) der Kreis um das Diagonalelement $a_{i,i}$ mit dem Radius der Summe der Beträge der Außerdiagonalelemente in Zeile i, also

$$r_i = \sum_{j\neq i} |a_{i,j}|,\ K_i = \{z : |z - a_{i,i}| \leq r_i\}\,.$$

Dann liegen alle Eigenwerte von A in der Vereinigung der Kreise K_i.
Falls die Vereinigung V von m Kreisen disjunkt ist zum Rest der Kreise, so liegen in V genau m Eigenwerte von A.

Also: Sei $M \subset \{1 \ldots n\}$, $m = |M|$. Weiter sei

$$\bigcup_{i\in M} K_i \cap \bigcup_{i\notin M} K_i = \emptyset,$$

dann ist

$$|\{\lambda_k \in \bigcup_{i\in M} K_i : \lambda_k \text{ Eigenwert von } A\}| = m,$$

wobei die Eigenwerte mit ihrer Vielfachheit im charakteristischen Polynom gezählt werden.

Zunächst ein kurzes Beispiel. Wir betrachten

$$A = \begin{pmatrix} 4 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1/2 \end{pmatrix}.$$

Die Gerschgorinkreise sind der Kreis K_1 um 4 mit Radius 1, der Kreis K_2 um 1 mit Radius 1 und der Kreis K_3 um $1/2$ mit Radius 1 (alles in der komplexen Ebene, natürlich). Dann garantiert der Satz von Gerschgorin, dass in K_1 genau ein Eigenwert von A liegt, in $K_2 \cup K_3$ liegen zwei.
Ausdrücklich: Der Satz von Gerschgorin garantiert in diesem Fall **nicht**, dass in K_2 bzw. K_3 ein Eigenwert liegt (nur in der Vereinigung liegen zwei).

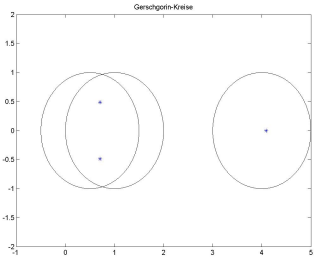


Abbildung 4.4: Gerschgorin–Kreise von A

Beweis:

1. Sei λ ein Eigenwert von A. Sei x Eigenvektor von A zum Eigenwert λ mit $||x||_\infty = 1$. Es gibt also ein m mit $|x_m| = 1$.

$$\begin{aligned}(A - \lambda I)x = 0 \quad &\Longrightarrow \quad (a_{m,m} - \lambda)x_m = -\sum_{j \neq m} a_{m,j}x_j \\ &\Longrightarrow \quad |a_{m,m} - \lambda| \leq \sum_{j \neq m} |a_{m,j}| \cdot |x_j| \\ &\Longrightarrow \quad |a_{m,m} - \lambda| \leq \sum_{j \neq m} |a_{m,j}| = r_m\end{aligned}$$

2. Es sei

$$A = D + L + R, D,\ L,\ R \in \mathbb{R}^{n \times n}.$$

Hierbei enthalte D die Elemente auf der Hauptdiagonalen (HD) von A, L die Elemente unterhalb der HD, R die Elemente oberhalb der HD, also

$$L_{i,k} = \begin{cases} a_{i,k} & i > k \\ 0 & sonst \end{cases},\ D_{i,k} = \begin{cases} a_{i,k} & i = k \\ 0 & sonst \end{cases},\ R_{i,k} = \begin{cases} a_{i,k} & i < k \\ 0 & sonst \end{cases}.$$

Wir betrachten die Matrizen

$$A(t) = D + t(L + R), t \in [0, 1],\ A(0) = D,\ A(1) = A.$$

Wir zitieren den Satz: Die Nullstellen eines Polynoms hängen stetig von seinen Koeffizienten ab. Insbesondere gibt es stetige Funktionen $\lambda_1, \ldots, \lambda_n$, so dass $\lambda_1(t) \ldots \lambda_n(t)$ die Eigenwerte von $A(t)$ sind (der arithmetischen Vfh nach gezählt).

Falls die Eigenwerte keine mehrfachen Nullstellen des charakteristischen Polynoms sind, so folgt das einfach mit dem Satz über implizite Funktionen, andernfalls muss man etwas mehr arbeiten. Ein vollständiger Beweis mit der Ordnung der Abhängigkeit findet sich in Kato [1995], Satz II.1.7.
Wir betrachten nun $\lambda_i(t)$ als Kurve in der komplexen Ebene.
$A(0) = D$ ist Diagonalmatrix, d.h. die Eigenwerte stehen auf der Hauptdiagonalen. Es gilt also $\lambda_i(0) = a_{i,i}$ (bei geeigneter Numerierung der Kurven). Sei

$$V = \bigcup_{i \in M} K_i,\ W = \bigcup_{i \notin M} K_i,\ V \cap W = \emptyset.$$

Die Gerschgorinkreise $K_i(t)$ von $A(t)$ sind Kreise um $a_{i,i}$ mit Radius $tr_i \leq r_i$, also gilt $K_i(t) \subset K_i$. Damit gilt nach Teil 1 des Satzes

$$\lambda_i(t) \in V \cup W \,\forall i = 1 \ldots n,\ t \in [0,1].$$

Da aber $V \cap W = \emptyset$, können die Kurven λ_i die Mengen V und W nicht verlassen, d.h. sie liegen ganz entweder in V oder W.
Es gilt $a_{i,i} \in K_i$ und damit

$$i \in M \Rightarrow \lambda_i(0) = a_{i,i} \in V,\ i \notin M \Rightarrow \lambda_i(0) = a_{i,i} \in W.$$

Also beginnen (und enden) genau m Kurven in V. Insbesondere liegen genau $m = |M|$ Eigenwerte von $A(1) = A$ in V.

□

Bemerkung: Da die Eigenwerte von A und A^t dieselben sind, kann man den Satz statt auf die Zeilensumme auch auf die Spaltensumme anwenden.
Häufig kann man die Abschätzung verschärfen, indem man das Kriterium statt auf A auf $\mathcal{D}A\mathcal{D}^{-1}$ mit einer Diagonalmatrix $\mathcal{D}$ anwendet.

4.4 Zusammenfassung

4.4.1 Kompetenzen

- Fixpunkte von Gleichungen im $\mathbb{R}^n$ mit Banach bestimmen können.
- Definition des Newton–Verfahrens kennen
- Definition quadratischer/linearer Konvergenz kennen.
- Lokalen Konvergenzsatz für das Newton–Verfahren anwenden können.

- Konvergenz des Newton–Verfahrens für konkrete Funktionen nachweisen können.
- Satz von Gerschgorin kennen und auf Matrizen anwenden können.

4.4.2 Mini–Aufgaben

- Zeigen Sie, dass im ersten Beispiel zum Newton–Verfahren die Funktion $g|_D$, $D := \{x : x \geq \overline{x}\}$, die Voraussetzungen des Fixpunktsatzes von Banach erfüllt. Achtung: Benutzen Sie das, was dort schon gezeigt wurde, dann ist dies ein Einzeiler.
- Wenden Sie den Satz von Gerschgorin auf die Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

  an.
- Gesucht sei eine Nullstelle von $f(x) = x - \cos x$. Stellen Sie das Newton–Verfahren auf und zeigen Sie, dass es lokal quadratisch konvergiert.
- Warum ist es sinnlos, das Newton–Verfahren auf lineare Gleichungen der Form $Ax - b = 0$ anzuwenden?

Kapitel 5

Lösung Linearer Gleichungssysteme mit Fixpunktiterationen

Insbesondere die Lösung von Randwertproblemen führt uns auf das Problem der Lösung sehr großer linearer Gleichungssysteme. Ziel in diesem Kapitel ist es, dazu wieder den Fixpunktsatz von Banach und Fixpunktiterationen zu nutzen. Deshalb (und im Vorgriff auf das Kapitel über lineare Differentialgleichungen) wiederholen wir hier einige Grundbegriffe der linearen Algebra.

Dabei werden wir uns der Einfachheit halber häufig auf reelle, endlichdimensionale Vektorräume oder sogar auf den $\mathbb{R}^n$ zurückziehen. Dort, wo eine Übertragung auf (komplexe) unendlichdimensionale Vektorräume nicht möglich ist, werden wir dies bemerken.

5.1 Grundbegriffe der linearen Algebra

5.1.1 Normierte Vektorräume

Definition 5.1 *(normierte Vektorräume)*
Sei V ein Vektorraum. $|| \cdot || : V \mapsto \mathbb{R}^{\geq 0}$ *heißt Norm, falls*

1) $||\alpha x|| = |\alpha| ||x|| \, \forall \alpha \in K, \, x \in V$.

2) $||x|| = 0 \Leftrightarrow x = 0$.

3) $||x + y|| \leq ||x|| + ||y|| \, \forall x, y \in V$.

$(V, || \cdot ||)$ *heißt normierter Vektorraum.*

Beispiel 5.2 *Sei $V = \mathbb{R}^n$, $p \in [1, \infty]$, $v = (v_1, \ldots, v_n) \in V$.*

$$||v||_p := \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p} (p < \infty),\ ||v||_\infty = \max_i |v_i|$$

heißt p–Norm (und ist eine Norm).

Der folgende topologische Satz spielt für uns eine wichtige Rolle.

Satz 5.3 *(Normäquivalenz in endlichdimensionalen Räumen)*
Es seien $||\cdot||$ und $|||\cdot|||$ zwei Normen im $\mathbb{R}^n$. Dann sind $||\cdot||$ und $|||\cdot|||$ äquivalent, d.h. $\exists\, 0 < c \leq C < \infty$:

$$c|||v||| \leq ||v|| \leq C|||v|||\, \forall v \in V.$$

Beweis: Wir zeigen, dass jede Norm $||\cdot||$ äquivalent zur $||\cdot||_\infty$–Norm ist, und damit sind alle Normen zueinander äquivalent.

1. $||\cdot||$ ist eine stetige Funktion bzgl. $||\cdot||_\infty$
   Sei $x = (x_k) \in \mathbb{R}^n$. Seien e_k die Einheitsvektoren, dann gilt $x = \sum_k x_k e_k$ und damit

   $$||x|| = ||\sum_k x_k e_k|| \leq \sum_k |x_k|\, ||e_k|| \leq \underbrace{\left(\sum_k ||e_k|| \right)}_{=:L} ||x||_\infty.$$

   Seien nun $x, y \in \mathbb{R}^n$. Dann gilt

   $$||x|| = ||x - y + y|| \leq ||x - y|| + ||y||,\ ||y|| = ||y - x + x|| \leq ||x - y|| + ||x||$$

   und damit

   $$|\, ||x|| - ||y||\, | \leq ||x - y|| \leq L||x - y||_\infty.$$

   Also ist $||\cdot||$ lipschitzstetig mit Lipschitzkonstante L.

2. Sei nun $x \in \mathbb{R}^n$, $x \neq 0$. Dann gilt

   $$||x|| = ||\frac{x}{||x||_\infty}||\, ||x||_\infty$$
   $$\begin{cases} \leq \underbrace{\sup_{||u||_\infty = 1} ||u||}_{=:C}\ ||x||_\infty \\ \geq \underbrace{\inf_{||u||_\infty = 1} ||u||}_{=:c}\ ||x||_\infty \end{cases}.$$

Die Einheitskugel im $\mathbb{R}^n$ ist kompakt, $||\cdot||$ ist stetig, also werden Supremum und Infimum angenommen, und es gilt

$$c||x||_\infty \leq ||x|| \leq C||x||_\infty,\ 0 < c \leq C < \infty.$$

□

In unendlichdimensionalen Räumen ist dieser Satz falsch.

Korollar 5.4 *(Normkonvergenz in endlichdimensionalen Räumen)*
Sei $x^{(k)}$ eine Folge im endlichdimensionalen Raum X und konvergent gegen $\overline{x}$ bezüglich der Norm $||\cdot||$. Dann ist $x^{(k)}$ konvergent gegen $\overline{x}$ bezüglich jeder Norm auf X.

Beweis: Sei $|||\cdot|||$ eine Norm auf X. Dann gilt

$$|||x^{(k)} - \overline{x}||| \leq C\,||x^{(k)} - \overline{x}|| \to 0.$$

□

Definition 5.5 *(Vektorräume mit Skalarprodukt, euklidische Vektorräume)*
$(\cdot,\cdot) : V \times V \mapsto K$ heißt Skalarprodukt, falls

1) $(v,v) \geq 0$ *und* $(v,v) = 0 \Leftrightarrow v = 0\ \forall v \in V$.

2) $(u,v) = \overline{(v,u)}\ \forall u,v \in V$.

3) $(\cdot,v)$ *ist linear für alle festen* $v \in V$.

Üblicherweise wird auf euklidischen Räumen die Norm

$$||v||_2 = (v,v)^{1/2},\ v \in V$$

benutzt. V heißt dann Prä–Hilbertraum. Ist V mit dieser Norm vollständig, so heißt V Hilbertraum.

Beispiel 5.6
Sei $V = \mathbb{C}^n$. Dann ist

$$(u,v) = u^t\overline{v},\ u,v \in V$$

ein Skalarprodukt. Wir werden dies stillschweigend als Standard–Skalarprodukte verwenden. Die induzierte Norm ist jeweils $||\cdot||_2$.

Satz 5.7 *(Cauchy–Schwarz)*
Sei V ein (Prä–) Hilbertraum. Dann gilt

$$|(u,v)|^2 \leq ||u||^2||v||^2 \, \forall u,v \in V$$

und Gleichheit genau dann, wenn u und v linear abhängig sind.

Beweis: Der Einfachheit halber zeigen wir den Satz für reelle Vektorräume. Falls $v = 0$, so ist der Satz richtig. Sei also $v \neq 0$. Es gilt

$$0 \leq || \, ||v||^2 u - (u,v)v||^2 = ||v||^4(u,u) - 2|(u,v)|^2||v||^2 + |(u,v)|^2(v,v)$$

und damit

$$|(u,v)|^2 \leq ||u||^2||v||^2$$

und Gleichheit genau dann, wenn $u = \lambda v$. □

Satz 5.8 *Sei V ein Vektorraum mit Skalarprodukt. Dann ist*

$$||v|| = (v,v)^{1/2}, \; v \in V$$

eine Norm.

Beweis:

$$||u+v||^2 = ||u||^2 + 2\mathrm{Re}(u,v) + ||v||^2 \leq ||u||^2 + 2||u|| \, ||v|| + ||v||^2 = (||u|| + ||v||)^2.$$

□

5.1.2 Lineare Operatoren

Definition 5.9 *(lineare Operatoren) Seien U, V Vektorräume. $T : U \mapsto V$ heißt lineare Abbildung genau dann, wenn*

$$T(\alpha x + y) = \alpha Tx + Ty, \; \forall \alpha \in K, \; x,y \in U.$$

Für $U = \mathbb{R}^n$, $V = \mathbb{R}^m$ identifizieren wir die Abbildung T immer direkt mit der darstellenden Matrix aus dem $\mathbb{R}^{m \times n}$.
Die Menge aller linearen Operatoren $L(U,V)$ bildet auf natürliche Weise selbst wieder einen Vektorraum. Auf diesem definieren wir eine durch die Normen in U und V induzierte Norm.

Definition 5.10 *(induzierte Operatornorm)*
Seien $(U, ||\cdot||_U)$ *und* $(V, ||\cdot||_V)$ *normierte Vektorräume. Sei* $T \in L(U, V)$*. Dann heißt*

$$||T|| := \sup_{u\in U, u\neq 0} \frac{||Tu||_V}{||u||_U} = \sup_{u\in U, u\neq 0} ||T\frac{u}{||u||_U}||_V = \sup_{u\in U, ||u||_U=1} ||Tu||_V$$

(induzierte) Operatornorm von T*.*

Beispiel 5.11 *(Induzierte Matrixnorm bzgl.* $||\cdot||_\infty$*)*
Sei $A = (A_{k,j}) \in \mathbb{R}^{n\times m}$ *nicht die Nullmatrix. Wir betrachten* A *als Operator von* $(\mathbb{R}^m, ||\cdot||_\infty)$ *nach* $(\mathbb{R}^n, ||\cdot||_\infty)$ *und bestimmen die zugehörige Norm* $||A||_\infty$*.*
$\forall\, u = (u_k) \in \mathbb{R}^m$ *gilt*

$$||Au||_\infty = ||(\sum_j A_{k,j}u_j)||_\infty = \max_k |\sum_k A_{k,j}u_j| \leq \underbrace{\left(\max_k \sum_j |A_{k,j}|\right)}_{=:C} ||u||_\infty$$

$$\Longrightarrow ||A||_\infty = \sup_{u\neq 0} \frac{||Au||_\infty}{||u||_\infty} \leq C.$$

Sei nun l *ein Index, an dem das Maximum angenommen wird, also*

$$\sum_j |A_{l,j}| = C.$$

Sei $u \in \mathbb{R}^n$ *mit* $u_j = \operatorname{sgn}(A_{l,j})$ *mit der Vorzeichenfunktion* sgn*. Insbesondere ist* $||u||_\infty = 1$*. Es gilt* $A_{l,j}\operatorname{sgn}(A_{l,j}) = |A_{l,j}|$*, also:*

$$||A||_\infty \geq ||Au||_\infty \geq (Au)_l = \sum_j A_{l,j}u_j = \sum_j |A_{l,j}| = C.$$

und damit $||A||_\infty = C$*.*

Es ist noch ungeklärt, ob das Supremum unendlich sein kann.

Satz 5.12 *(Normeigenschaft der induzierten Matrixnorm)*
Es sei $A \in \mathbb{R}^{n\times m}$ *lineare Abbildung von* $U := (\mathbb{R}^m, ||\cdot||_U)$ *nach* $V : \mathbb{R}^n, ||\cdot||_V$*. Dann ist* $||A|| < \infty$*. Die induzierte Matrixnorm ist eine Norm auf den linearen Abbildungen von* U *nach* V*.*

Beweis: Nach 5.3 gibt es $c > 0$ und $C < \infty$ mit

$$||v||_V \leq C||v||_\infty \,\forall v \in V,\ ||u||_U \geq c||u|| \,\forall u \in U.$$

Also gilt

$$||A|| = \sup_{u \in U, u \neq 0} \frac{||Tu||_V}{||u||_U} \leq \frac{C}{c} \sup_{u \in U, u \neq 0} \frac{||Tu||_\infty}{||u||_\infty} = \frac{C}{c} ||T||_\infty < \infty.$$

Damit ist jetzt die induzierte Matrixnorm eine wohldefinierte Abbildung in die reellen Zahlen. Die Eigenschaften der Norm zeigt man durch einfaches Nachrechnen. □

Satz 5.13 *(Eigenschaften der induzierten Norm)*
Seien $(U, ||\cdot||_U)$, $(V, ||\cdot||_V)$, $(W, ||\cdot||_W)$ *normierte Vektorräume. Sei* $T \in L(U, V)$, $S \in L(V, W)$. *Dann gilt*

$$||Tu||_V \leq ||T||\, ||u||_U \, \forall u \in U$$

und

$$||ST|| \leq ||S||\, ||T||.$$

Beweis: Teil 1: Definition der induzierten Operatornorm.
Teil 2: Nach Teil 1 ist

$$||STu||_W \leq ||S||\, ||Tu||_V \leq ||S||\, ||T||\, ||u||_U.$$

□

Korollar 5.14 *Es sei* $B \in \mathbb{R}^{n \times n}$ *und* $||\cdot||$ *eine Norm. Genau dann wenn* $||B|| < 1$ *in der induzierten Matrixnorm ist die Funktion*

$$g : \mathbb{R}^n \mapsto \mathbb{R}^n,\ g(x) := Bx - c$$

kontrahierend (bezüglich der $||\cdot||$*–Norm).*

Beweis: Seien $x, y \in \mathbb{R}^n$ und sei $||B|| < 1$. Es gilt

$$||g(x) - g(y)|| = ||(Bx + c) - (By + c)|| = ||(B(x - y)|| \leq ||B||\, ||x - y||.$$

Umgekehrt: Sei g kontrahierend mit Konstante $q < 1$. Dann gilt

$$||Bx|| = ||g(x) - g(0)|| \leq q\, ||x - 0|| = ||x||$$

und damit

$$||B|| = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{||Bx||}{||x||} \leq q < 1.$$

□

Bemerkung: Alternativ kann man Normen auf dem Vektorraum der Matrizen definieren durch

$$||A|| = (\sum_{k,j} |A_{k,j}|^p)^{\frac{1}{p}} \text{ bzw. } ||A|| = \max_{k,j} |A_{k,j}|$$

für $p < \infty$. Für $p = 2$ heißt diese Norm Frobenius–Norm.
Der Vorteil dieser Normen ist, dass sie schnell auszurechnen sind. Der Nachteil ist, dass sie nicht notwendig verträglich sind mit der Vektorraumnorm für $p \neq 2$ (d.h. es gilt nicht $||Av|| \leq ||A||\,||v||$). Für die Zwecke dieser Vorlesung sind sie damit im Allgemeinen unbrauchbar.

Korollar 5.15 *Sei A_k eine Folge von Matrizen. A_k konvergiert gegen A in einer beliebigen Norm $||\cdot||$ auf dem Vektorraum der Matrizen genau dann, wenn alle Matrixelemente gegeneinander konvergieren.*

Beweis: Äquivalenz zur Unendlichnorm der Koeffizienten. □

5.1.3 Adjungierte Abbildungen und Eigenwerte

Definition 5.16 *(Adjungierte Abbildung)*
Seien $(U, (\cdot,\cdot)_U)$ und $(V, (\cdot,\cdot)_V)$ Vektorräume mit Skalarprodukt. Sei $T \in L(U, V)$, $T^* \in L(V, U)$. Falls

$$(Tu, v)_V = (u, T^*v)_U \forall u \in U,\ v \in V,$$

so heißt T^* die zu T adjungierte Abbildung.
Falls $U = V$ und $T = T^*$, so heißt T selbstadjungiert.

Beispiel 5.17 *Sei $U = \mathbb{R}^n$, $V = \mathbb{R}^m$, $A \in \mathbb{R}^{n\times m}$. U und V seien versehen mit dem Standardskalarprodukt. Dann gilt für $u \in U$, $v \in V$*

$$(Au, v) = u^t A^t \overline{v} = u^t (A^t v) = (u, A^t v)$$

und damit $A^* = A^t \in \mathbb{R}^{m\times n}$. Über $\mathbb{C}$ gilt entsprechend $A^* = \overline{A^t}$.

Definition 5.18 *(hermitesch, symmetrisch)*
Matrizen mit der Eigenschaft $A = A^* = \overline{A^t}$ heißen hermitesch, reelle Matrizen mit der Eigenschaft $A = A^* = A^t$ heißen symmetrisch.

Satz 5.19 *(Rechenregeln für adjungierte Operatoren)*

1. $(T_1T_2)^* = T_2^*T_1^*$.
2. $(T^*)^* = T$.
3. *TT^* und T^*T sind selbstadjungiert.*

Beweis: Durch einfaches Nachrechnen. □

Definition 5.20 *(Eigenwerte und Eigenvektoren)*
Sei $T \in L(U,U)$, $v \in U$, $v \neq 0$. v heißt Eigenvektor zum Eigenwert $\lambda \in \mathbb{C}$, falls $Tv = \lambda v$.

Definition 5.21 *(Diagonalisierbarkeit)*
Sei $T \in L(U,U)$, $\dim U < \infty$. T heißt diagonalisierbar, falls U eine Basis aus Eigenvektoren v_k zu Eigenwerten λ_k von T besitzt. Es gilt

$$D = W^{-1}TW,\ W = (v_1v_2\cdots v_n), T = \mathit{diag}(\lambda_k).$$

Satz 5.22 *Selbstadjungierte Operatoren haben reelle Eigenwerte. Eigenvektoren zu unterschiedlichen Eigenwerten stehen senkrecht aufeinander.*

Beweis: Sei T selbstadjungiert. Sei $Tx = \lambda x$, $x \neq 0$. Dann gilt

$$\lambda(x,x) = (\lambda x, x) = (Tx, x) = (x, Tx) = (x, \lambda x) = \overline{\lambda}(x,x)$$

und wegen $(x,x) \neq 0$ gilt $\lambda = \overline{\lambda}$.
Sei $Tx = \lambda_1 x$, $Ty = \lambda_2 y$, $\lambda_1 \neq \lambda_2$, $x \neq 0$, $y \neq 0$. Dann gilt

$$\lambda_1(x,y) = (Tx,y) = (x,Ty) = \overline{\lambda_2}(x,y) = \lambda_2(x,y)$$

und damit wegen $\lambda_1 \neq \lambda_2$: $(x,y) = 0$. □

Definition 5.23 *(Positiv definite Operatoren)*
Sei U Vektorraum mit Skalarprodukt, $T \in L(U,U)$. T heißt (symmetrisch) positiv definit, wenn T selbstadjungiert ist und

$$(Tu,u) > 0\, \forall u \in U,\ u \neq 0.$$

Gilt nur $\geq$, so heißt T positiv semidefinit.

Satz 5.24 *Sei U Vektorraum mit Skalarprodukt, $T \in L(U,U)$ symmetrisch positiv definit. Dann ist*

$$(u,v)_T := (Tu,v),\ u \in U,\ v \in U$$

ein Skalarprodukt auf U.

Satz 5.25 *Sei $T \in L(U,V)$. T^*T ist positiv semidefinit. Falls T injektiv ist, so ist T positiv definit.*

Beweis: T^*T ist selbstadjungiert, und $(T^*Tx, x) = (Tx, Tx) \geq 0$. □

Den Satz über die Jordan–Normalform kennen Sie aus der Linearen Algebra. Bitte machen Sie sich klar, dass Ihre Formulierung der folgenden entspricht.

Satz 5.26 *(Jordan–Normalform)*
Sei A eine $(n \times n)$–Matrix. v heißt Hauptvektor k. Stufe zum Eigenwert λ von A, falls

$$(A - \lambda I)^k v = 0, (A - \lambda I)^{k-1} v \neq 0.$$

Hauptvektoren erster Stufe sind Eigenvektoren.

1. *Jede Matrix besitzt eine Basis aus Hauptvektoren v_j.*
2. *Sei J die Darstellung von A in dieser Basis, also*

$$J = X^{-1}AX,\ X = (v_1 v_2 \cdots v_n).$$

Dann ist J eine Jordan–Matrix, d.h.

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{pmatrix},\ J_k \in \mathbb{C}^{p \times p},\ J_k = \lambda_k I + N,\ N_{il} = \begin{cases} 1 & i+1 = l \\ 0 & \textit{sonst.} \end{cases}$$

Satz 5.27 *Sei A hermitesche $(n \times n)$–Matrix. Dann ist A diagonalisierbar. $\mathbb{R}^n$ besitzt eine Orthonormalbasis aus Eigenvektoren von A.*

Beweis: Zu zeigen ist: Alle Hauptvektoren sind Eigenvektoren, also Hauptvektoren erster Stufe. Sei $(A - \lambda I)^2 v = 0$. Dann gilt

$$0 = ((A - \lambda I)^2 v, v) = ((A - \lambda I)v, (A - \lambda I)v) = ||(A - \lambda I)v||^2$$

und damit schon $(A - \lambda I)v = 0$, es gibt also keine Hauptvektoren höherer Stufe, und die Jordan–Normalform ist eine Diagonalmatrix. Es gibt also eine Basis aus Eigenvektoren. Die Eigenvektoren zu unterschiedlichen Eigenwerten stehen bereits senkrecht aufeinander nach 5.22. In den Eigenräumen zum gleichen Eigenwert wählt man eine ONB als Basis. □
Dies ist ein Hauptsatz für hermitesche Matrizen. Viele ihrer Eigenschaften können einfach durch Entwicklung in diese ONB bewiesen werden.

Korollar 5.28 *Die Matrix A sei hermitesch. A ist positiv definit (semidefinit) genau dann, wenn alle Eigenwerte von A positiv (nichtnegativ) sind.*

Beweis: Sei A hermitesch mit positiven Eigenwerten λ_k und einer zugehörigen ONB aus Eigenvektoren v_k. Sei $x \in \mathbb{C}^n$. Dann lässt sich x in dieser Basis darstellen, also $x = \sum_k c_k v_k.$, und es gilt

$$(Ax, x) = (\sum_j c_j A v_j, \sum_k c_k v_k) = \sum_k \lambda_k |c_k|^2 (v_k, v_k) > 0.$$

Falls A positiv definit ist, so gilt

$$\lambda_k = (\lambda_k v_k, v_k) = (Av_k, v_k) > 0.$$

Für semidefinit entsprechend. □

Satz 5.29 *(entfernt)*

Mit diesen Vorbemerkungen können wir nun leicht die 2–Norm einer Matrix berechnen. Sei $A \in \mathbb{C}^{m \times n}$ und $B = A^*A$.

Definition 5.30 *Sei $A \in \mathbb{C}^{n \times n}$. Dann heißt*

$$\rho(A) = \max\{|\lambda_k| : \lambda_k \textit{ Eigenwert von } A\}$$

Spektralradius von A.

Satz 5.31 *Sei $A \in \mathbb{C}^{m \times n}$. Dann gilt*

$$||A||_2 = \rho(A^*A)^{1/2}.$$

Falls $m = n$ und $A = A^*$, so gilt

$$||A||_2 = \rho(A).$$

Beweis: $B = A^*A$ ist symmetrisch positiv semidefinit, also besitzt $\mathbb{C}^n$ eine Orthonormalbasis aus Eigenvektoren v_k zu nichtnegativen Eigenwerten λ_k von B. Sei $v \in \mathbb{C}^m$. v_k ist Basis, also gibt es μ_k mit $v = \sum_k \mu_k v_k$. Es gilt

$$||v||^2 = (v, v) = (\sum_k \mu_k v_k, \sum_j \mu_j v_j) = \sum_k |\mu_k|^2.$$

Weiter gilt

$$\begin{aligned} ||Av||_2^2 = (Av, Av) &= (A^*Av, v) \\ &= (\sum_k \mu_k \lambda_k v_k, \sum_j \mu_j v_j) \\ &= \sum_k \lambda_k |\mu_k|^2 \\ &\leq \rho(B) \sum_k |\mu_k|^2 \\ &= \rho(B)\, ||v||_2^2 \end{aligned}$$

$$\Longrightarrow ||A||_2^2 = \sup_{v \neq 0} \frac{||Av||_2^2}{||v||_2^2} \leq \rho(B).$$

Da $\lambda_k \geq 0$, ist $\rho(B)$ der größte Eigenwert von B. Sei w ein zugehöriger Eigenvektor. Dann gilt

$$||Aw||_2^2 = (Aw, Aw) = (A^*Aw, w) = \rho(B)\,||w||_2^2 \Rightarrow ||A||_2^2 \geq \frac{||Aw||_2^2}{||w||_2^2} = \rho(B).$$

Also gilt

$$||A||_2 = \sup_{v \neq 0} \frac{||Av||_2}{||v||_2} = \rho(A^*A)^{1/2}.$$

□

5.2 Fehlerabschätzung für lineare Gleichungssysteme

Als Anwendung für diese Sätze beweisen wir eine Fehlerabschätzung.
Zu lösen sei das Gleichungssystem $Ax = b$ mit A invertierbar. Statt A und b stehen aber nur Näherungen $\widetilde{A} = A + \Delta A$ und $\widetilde{b} = b + \Delta b$ zur Verfügung. Man kann also nur das Gleichungssystem $\widetilde{A}\widetilde{x} = \widetilde{b}$ lösen. Dazu stellen sich sofort zwei Fragen:

1. Unter welchen Voraussetzungen ist $\widetilde{A}$ invertierbar, d.h. das Gleichungssystem für jede rechte Seite eindeutig lösbar?
2. Wie groß ist der Fehler von $\widetilde{x} = x + \Delta x$?

Wenn wir hier von Fehler sprechen, meinen wir immer den relativen Fehler, also z.B. $||\Delta x||/||x||$, typischerweise angegeben in Prozent.

Satz 5.32 *(Neumannsche Reihe)*
Sei $(V, ||\cdot||)$ ein Banachraum, $T : V \mapsto V$ linear mit $||T|| < 1$ (induzierte Norm). Dann ist $(I - T)$ invertierbar, und

$$(I - T)^{-1} = \sum_{k=0}^{\infty} T^k.$$

Beweis: Es sei $v \in V$ und $v^{(k)} = \sum_{j=0}^{k} T^j v$. Da $||T^j v|| \leq ||T||^j\,||v||$, ist $v^{(k)}$ eine Cauchyfolge. V ist vollständig, also $v^{(k)} \to w \in V$. Es gilt

$$(I - T)w = \lim_{k \mapsto \infty} (I - T) \sum_{j=0}^{k} T^j v = \lim_{k \mapsto \infty} (v - T^{k+1} v) = v \Longrightarrow w = (I - T)^{-1} v.$$

□

Korollar 5.33 *Seien V Banachraum, $T \in L(V,V)$ invertierbar, $\Delta T \in L(V,V)$ und T^{-1} sei stetig. Weiter sei $q = ||T^{-1}||\,||\Delta T|| < 1$. Dann ist $(T + \Delta T)$ invertierbar und*

$$||(T + \Delta T)^{-1}|| \leq \frac{||T^{-1}||}{1 - q}.$$

Beweis:

$$(T + \Delta T) = T(I - (-T^{-1}\Delta T))$$

ist invertierbar nach 5.32.

$$\begin{aligned} ||(T + \Delta T)^{-1}|| &= ||\sum_{k=0}^{\infty}(-T^{-1}\Delta T)^k T^{-1}|| \\ &\leq ||T^{-1}|| \sum_{k=0}^{\infty} q^k \\ &= ||T^{-1}|| \frac{1}{1 - q} \end{aligned}$$

□

Dieser Satz lässt sich so interpretieren: Die Matrix A sei invertierbar, für die Näherung $\widetilde{A}$ gelte $||A - \widetilde{A}|| < \frac{1}{||A^{-1}||}$. Dann ist auch $\widetilde{A}$ invertierbar.

Korollar 5.34 *Die Menge der invertierbaren $(n \times n)$–Matrizen ist offen.*

Was passiert nun, wenn wir die Matrix A oder die rechte Seite b in einem linearen Gleichungssystem nicht genau kennen? Zunächst ein Beispiel.

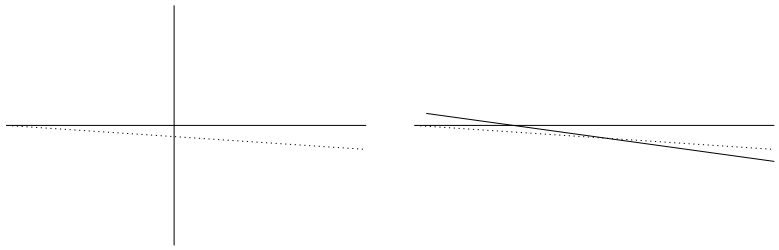


Abbildung 5.1: Graphische Lösung von Gleichungssystemen: Links gut gestellt, rechts schlecht gestellt, kleine Änderungen (gestrichelte Linie) in den Koeffizienten führen zu großer Änderung des Schnittpunkts.

Sei $n = 2$. Dann können wir die Lösung des Gleichungssystems als Schnittpunkt zweier Geraden im $\mathbb{R}^2$ graphisch bestimmen. Kleine Änderungen in den Koeffizienten führen zu kleinen Änderungen in der Lage der Linien. **Aber:** Falls die Linien fast parallel liegen, führt eine kleine Änderung in der Lage der Linien zu großen Änderungen beim Schnittpunkt. Die Verstärkung des Eingangsfehlers muss also von der Richtung der Linien, also von A, abhängen.

Satz 5.35 *Sei $A \in \mathbb{R}^{n\times n}$ invertierbar. Sei $x \in \mathbb{R}^n$ und $Ax = b$. Sei weiter $\Delta A \in \mathbb{R}^{n\times n}$ und $\Delta b \in \mathbb{R}^n$. Es sei*

$$k(A) = ||A||\,||A^{-1}||$$

die **Kondition** *von A und es gelte*

$$q = k(A)\frac{||\Delta A||}{||A||} = ||A^{-1}||\,||\Delta A|| < 1.$$

Dann ist $A + \Delta A$ invertierbar. Sei $\tilde{x} = x + \Delta x$ die Lösung von

$$(A + \Delta A)\tilde{x} = (b + \Delta b).$$

Dann gilt für den relativen Fehler in der Lösung

$$\frac{||\Delta x||}{||x||} \leq \frac{k(A)}{1-q}\left(\underbrace{\frac{||\Delta b||}{||b||}}_{\text{rel.Fehler in b}} + \underbrace{\frac{||\Delta A||}{||A||}}_{\text{rel. Fehler in A}}\right)$$

Die relativen Fehler in A und b werden also (höchstens) um den Faktor

$$M = \frac{k(A)}{1-q}$$

verstärkt.

Für sinnvolle Anwendungen ist $||\Delta A||$ klein gegen $||A||$, also $q \sim 0$ und damit $M \sim k(A)$.

Beweis: Nach 5.33 ist $A + \Delta A$ invertierbar, und es gilt

$$||(A + \Delta A)^{-1}|| \leq \frac{||A^{-1}||}{1-q}.$$

Es gilt

$$(A + \Delta A)(x + \Delta x) = (b + \Delta b)$$

und damit wegen $Ax = b$

$$(A + \Delta A)\Delta x = \Delta b - \Delta A x$$

und

$$\Delta x = (A + \Delta A)^{-1}(\Delta b - \Delta A x),$$

also insbesondere

$$||\Delta x|| \leq ||(A + \Delta A)^{-1}||(||\Delta b|| + ||\Delta A||\,||x||).$$

Für den relativen Fehler für $x \neq 0$

$$\begin{aligned}
\frac{||\Delta x||}{||x||} &\leq \frac{||A^{-1}||}{1-q}\left(\frac{||\Delta b||}{||x||} + ||\Delta A||\right) \\
&= \frac{k(A)}{1-q}\left(\frac{||\Delta b||}{||A||\,||x||} + \frac{||\Delta A||}{||A||}\right) \\
&\leq \frac{k(A)}{1-q}\left(\frac{||\Delta b||}{||b||} + \frac{||\Delta A||}{||A||}\right)
\end{aligned}$$

wegen $||b|| = ||Ax|| \leq ||A||\,||x||$. □

5.3 Fixpunktverfahren für lineare Gleichungssysteme

Zu lösen sei im Folgenden immer das lineare Gleichungssystem $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $b, x \in \mathbb{R}^n$. Wir definieren dazu eine Matrix $B \in \mathbb{R}^{n\times n}$ und einen Vektor $c \in \mathbb{R}^n$, so dass die Funktion

$$g(x) = Bx + c$$

die Voraussetzungen des Banachschen Fixpunktsatzes erfüllt, und der dann eindeutig bestimmte Fixpunkt $\overline{x}$ das Gleichungssystem löst. Den Fixpunkt bestimmen wir dann, wie beim Newtonverfahren, mit Hilfe der Fixpunktiteration.

Definition 5.36 *(Einzelschritt– und Gesamtschrittverfahren)*
Es sei $A = D + L + R$ wie im Beweis zu 4.11. D sei invertierbar, d.h. auf der Hauptdiagonalen von A stehen keine Nullen.

1. *Es sei*

   $$B = -D^{-1}(L + R),\ c = D^{-1}b.$$

   Dann heißt die Fixpunktiteration zu der Funktion

   $$g(x) := Bx + c = D^{-1}(-(L + R)x + b)$$

   Gesamtschritt– oder Jacobi–verfahren.

2. *Es sei*

$$B = -(D+L)^{-1}R,\ c = (D+L)^{-1}b.$$

Dann heißt die Fixpunktiteration zu der Funktion

$$g(x) := Bx + c = (D+L)^{-1}(-Rx + b)$$

Einzelschritt– oder Gauss–Seidel–Verfahren.

Lemma 5.37
Sei $||B|| < 1$ in der induzierten Matrixnorm zur Norm $||\cdot||$ auf dem $\mathbb{R}^n$. Dann konvergieren Einzelschritt– und Gesamtschritt–Verfahren bezüglich jeder Norm gegen eine Lösung des Gleichungssystems $Ax = b$.

Beweis: Wir betrachten $g : (\mathbb{R}^n, ||\cdot||) \mapsto (\mathbb{R}^n, ||\cdot||)$. Nach 5.14 ist g kontrahierende Selbstabbildung. $\mathbb{R}^n$ ist vollständig und abgeschlossen. Also sind alle Voraussetzungen für Banach erfüllt und die Fixpunktfolge konvergiert bezüglich $||\cdot||$ gegen ein $\overline{x}$. Also konvergiert sie nach 5.3 bezüglich jeder Norm gegen $\overline{x}$. $\overline{x}$ ist Fixpunkt, also gilt für das Gesamtrschrittverfahren

$$\overline{x} = g(\overline{x}) = D^{-1}(-(L+R)\overline{x} + b) \Longrightarrow b = (D+L+R)\overline{x} = A\overline{x}.$$

Entsprechend für das Einzelschrittverfahren

$$\overline{x} = g(\overline{x}) = (D+L)^{-1}(-R\overline{x} + b) \Longrightarrow b = (D+L+R)\overline{x} = A\overline{x}.$$

□

Wir müssen also jeweils sicherstellen, dass es eine Norm gibt mit $||B|| < 1$.

Satz 5.38 *Es sei $B \in \mathbb{R}^{n\times n}$, $\rho(B) < 1$. Dann gibt es eine Norm $||\cdot||$ auf dem $\mathbb{R}^n$ so dass für die zugehörige induzierte Matrixnorm gilt $||B|| < 1$.*

Beweis: Wir zeigen den Satz nur für zwei Spezialfälle, einen genaueren Beweis finden Sie in Wübbeling [2022].

1. Es sei B selbstadjungiert. Dann gilt nach 5.31

$$||B||_2 = \rho(B) < 1.$$

2. Es sei B diagonalisierbar und invertierbar, d.h. $B = XJX^{-1}$ für eine Matrix X aus Eigenvektoren und eine invertierbare Diagonalmatrix $J \in \mathbb{R}^{n\times n}$. Auf der Hauptdiagonalen von J stehen die Eigenwerte von B, und damit gilt nach 5.11 $||J||_\infty = \rho(B)$.

$$||v|| := ||J^{-1}X^{-1}v||_\infty$$

ist eine Norm im $\mathbb{R}^n$. Für die induzierte Norm ist

$$\begin{aligned}
||B|| &= \sup_{v\neq 0} \frac{||Bv||}{||v||} \\
&= \sup_{v\neq 0} \frac{||J^{-1}X^{-1}XJX^{-1}v||_\infty}{||J^{-1}X^{-1}v||_\infty} \quad u := J^{-1}X^{-1}v \\
&= \sup_{u\neq 0} \frac{||Ju||_\infty}{||u||_\infty} \\
&= ||J||_\infty = \rho(B) < 1.
\end{aligned}$$

□

Korollar 5.39 *(starke Diagonaldominanz)*
Falls für alle $i = 1 \dots n$ gilt

$$\sum_{j\neq i} |a_{i,j}| < |a_{i,i}|,$$

so konvergieren Gesamt– und Einzelschrittverfahren. Matrizen mit dieser Eigenschaft heißen stark diagonaldominant.

Beweis: ´

1. Im Gesamtschrittverfahren ist $B = D^{-1}(L + R)$, also

$$B = \begin{pmatrix} 0 & \frac{a_{1,2}}{a_{1,1}} & \frac{a_{1,3}}{a_{1,1}} & \dots & \frac{a_{1,n}}{a_{1,1}} \\ \frac{a_{2,1}}{a_{2,2}} & 0 & \frac{a_{2,3}}{a_{2,2}} & \dots & \frac{a_{2,n}}{a_{2,2}} \\ & \ddots & \ddots & \ddots & \end{pmatrix}$$

   Nach 4.11 sind alle Gerschgorinkreise Kreise in der komplexen Ebene um 0 mit Radius

$$\frac{1}{|a_{i,i}|} \sum_{j\neq i} |a_{i,j}| < 1.$$

   Damit gilt für alle Eigenwerte $|\lambda| < 1$, also $\rho(B) < 1$, und damit konvergiert das Gesamtschrittverfahren gegen eine Lösung von $Ax = b$. Insbesondere ist das Gleichungssystem immer lösbar, d.h. stark diagonaldominante Matrizen sind invertierbar.

2. Im Einzelschrittverfahren ist $B = -(D + L)^{-1}R$. Sei $|\lambda| \geq 1$. Dann ist auch $R + \lambda(D + L)$ stark diagonaldominant und damit nach Teil 1 invertierbar. Es gilt für das charakteristische Polynom χ_B von B

$$\chi_B(\lambda) = \det(-(D+L)^{-1}R - \lambda I) = \det(-(D+L)^{-1}) \det(R+\lambda(D+L)) \neq 0$$

und damit ist λ kein Eigenwert und $\rho(B) < 1$.

□

Bemerkung:
Die Bedingung der starken Diagonaldominanz lässt sich erheblich abschwächen zur schwachen Diagonaldominanz (Wübbeling [2022]).

5.4 Zusammenfassung

5.4.1 Kompetenzen

- Die zitierten Grundlagen beherrschen (ggf. wiederholen).
  Hinweis: Wir nutzen die *Sätze*, nicht die dortigen Beweise.
- Definition(en) der induzierten Matrixnorm kennen. Aus der Definition induzierte Matrixnorm ausrechnen können. Formeln für $||A||_\infty$ und $||A||_2$ kennen.
- Fehlerabschätzung zur Lösung linearer Gleichungen kennen und anwenden können.
- Gesamtschritt– und Einzelschrittverfahren kennen und anwenden können. Konvergenzbedingung kennen, insb. starkes Zeilensummenkriterium.

5.4.2 Mini–Aufgaben

- Geben Sie für den $\mathbb{R}^n$ c und C an so dass

$$c||x||_\infty \leq ||x||_2 \leq C||x||_\infty.$$

- – Es sei

$$A \in \mathbb{R}^{n\times n},\ A_{ik} = \begin{cases} 1 & i = k \\ \frac{1}{2n} & \text{sonst} \end{cases}.$$

  Geben Sie mit dem Satz von Gerschgorin Abschätzungen für $||A||_2$, $||A^{-1}||_2$ und die Kondition $k_2(A)$ an. Begründen Sie, dass A positiv definit ist.

  – Nehmen Sie nun an, dass für A nur eine Näherung $\widetilde{A}$ mit einem Fehler von 10% bekannt ist. Können Sie garantieren, dass $\widetilde{A}$ invertierbar ist?

  – Zu lösen sei nun das Gleichungssystem $Ax = b$. Statt b ist nur eine Näherung $\widetilde{b}$ mit einem Fehler von 10% bekannt. Welche Fehlerabschätzung für die Lösung $\widetilde{x}$ des Gleichungssystems $\widetilde{A}\widetilde{x} = \widetilde{b}$ können Sie garantieren?

– Zeigen Sie, dass Einzel– und Gesamtschrittverfahren für diese Matrix konvergieren.

Kapitel 6

Systeme von Linearen Differentialgleichungen

Nach diesen Vorbemerkungen zur linearen Algebra können wir uns wieder den Differentialgleichungen zuwenden.

In 2.5 wurden bereits lineare Differentialgleichungen der Form

$$y'(t) = \alpha(t)\, y(t) + \beta(t)$$

untersucht und eine explizite Lösungsformel wurde angegeben. Wir wollen dies nun auf Systeme ausweiten. Im Folgenden sei immer I konvex (d.h. ein Intervall), möglicherweise unbeschränkt.

Definition 6.1 *(Lineare Systeme von Differentialgleichungen)*
Es sei $I \subset \mathbb{R}$ konvex und $A : I \mapsto \mathbb{R}^{n\times n}$, $b : I \mapsto \mathbb{R}^n$. Das System von Differentialgleichungen

$$y'(t) = A(t)\, y(t) + b(t),\ y : I \mapsto \mathbb{R}^n,\ y \in C^1$$

heißt linear. Es heißt homogen, falls $b \equiv 0$, ansonsten inhomogen.

Es gilt

Satz 6.2 *(Existenz und Eindeutigkeit der Lösung von AWA für lineare DGL)*
Es seien A und b stetig in I. Dann hat das Anfangswertproblem

$$y'(t) = A(t)\, y(t) + b(t),\ y(a) = y_0,\ a \in I$$

eine eindeutige Lösung.

Beweis: Es sind nur die Voraussetzungen von 3.8 zu zeigen. Für unsere Differentialgleichung gilt $f(t, y) = A(t)\, y + b(t)$. Sei J ein abgeschlossenes Intervall in I,

$a \in J$. $A(t)$ ist stetig, also gibt es für jede induzierte Matrixnorm ein $L \in \mathbb{R}$ mit $||A(t)|| \leq L \, \forall t \in J$. Es gilt $\forall y_1, \, y_2 \in \mathbb{R}^n$

$$||f(t, y_1) - f(t, y_2)|| = ||(A(t)y_1 + b(t)) - (A(t)y_2 + b(t)|| \leq ||A(t))|| \, ||y_1 - y_2|| \leq L||y_1 - y_2||.$$

f ist also global lipschitzstetig. Auf jedem abgeschlossenen Teilintervall von I gibt es eine eindeutige Lösung, und damit auf ganz I. □

6.1 Homogene Systeme

Es sei immer A stetig auf I.

Satz 6.3 *(Lösungsisomorphismus)*

1. *Die Lösungen eines homogenen Systems von linearen Differentialgleichungen bilden einen n–dimensionalen Untervektorraum von C^1.*
2. *Sei $z(t, y_0)$ die Lösung der homogenen Anfangswertaufgabe mit $z(a, y_0) = y_0$. Dann ist die Abbildung*

$$\Phi : \mathbb{R}^n \mapsto C^1(I), \; \Phi(y_0)(t) := z(t, y_0)$$

ein Vektorraumisomorphismus von $\mathbb{R}^n$ auf den Lösungsraum.

Beweis: Wir zeigen den zweiten Teil des Satzes.
Die Existenz und Eindeutigkeit einer Lösung und damit die Wohldefiniertheit von Φ ist nach 6.2 sichergestellt.
Es seien nun $z_1, \, z_2 \in \mathbb{R}^n$, $\alpha, \, \beta \in \mathbb{R}$. $y_1(t) = \Phi(t, z_1)$, $y_2(t) = \Phi(t, z_2)$ und $y = \alpha y_1 + \beta y_2$. Dann gilt

$$y' = \alpha y_1' + \beta y_2' = \alpha A y_1 + \beta A y_2 = A(\alpha y_1 + \beta y_2) = Ay$$

und $y(a) = \alpha y_1(a) + \beta y_2(a) = \alpha a_1 + \beta a_2$. Es gilt also

$$\Phi(\alpha z_1 + \beta z_2) = y = \alpha y_1 + \beta y_2 = \alpha \Phi(z_1) + \beta \Phi(z_2).$$

Φ ist also linear.
Sei y irgendeine Lösung, dann löst y genau ein Anfangswertproblem.
Also ist Φ bijektiv und damit ein Vektorraumisomorphismus. Insbesondere haben $\mathbb{R}^n$ und der Lösungsraum die gleiche Dimension n. □

Korollar 6.4 *(Fundamentalsystem)*
Es seien $y_1 \dots y_m$ Lösungen der homogenen Differentialgleichung.

- *$y_1, \dots, y_m$ sind genau dann linear unabhängig in C^1, wenn $y_1(a), \dots, y_m(a)$ linear unabhängig sind in $\mathbb{R}^n$.*
- *Sei $t \in I$. $y_1(a), \dots, y_m(a)$ sind genau dann linear unabhängig, wenn $y_1(t), \dots, y_m(t)$ linear unabhängig sind.*
- *Falls $n = m$ und $y_1, \dots y_n$ linear unabhängig, so ist jede Lösung y der homogenen Differentialgleichung eindeutige Linearkombination von $y_1, \dots, y_n$, d.h.*

$$y = \sum_{k=1}^{n} \alpha_k \, y_k, \; \alpha_k \in \mathbb{R}.$$

$y_1, \dots, y_n$ sind Basis des Lösungsraums.

Definition 6.5 *(Fundamentalsystem, Fundamentalmatrix)*
Es seien $y_1, \dots, y_n$ linear unabhängige Lösungen der homogenen Differentialgleichung 6.1. Dann heißt $y_1, \dots, y_n$ ein Fundamentalsystem (von Lösungen der homogenen Differentialgleichung).
Die Matrix

$$Y(t) = \big(y_1(t), \dots, y_n(t)\big)$$

heißt Fundamentalmatrix.

Korollar 6.6 *Sei $Y(t)$ Fundamentalmatrix zu 6.1.*

1. *Es gilt $Y'(t) = A(t)Y(t)$.*
2. *Sei $C \in \mathbb{R}^{n \times n}$ invertierbar. Dann ist $Z(t) = Y(t)\,C$ eine Fundamentalmatrix, d.h. ihre Spalten bilden ein Fundamentalsystem.*

Beweis: Durch Einsetzen. □
Da $y_1, \dots, y_n$ linear unabhängig sind, ist $Y(t)$ für alle t invertierbar. Wählt man im Korollar $C = Y(a)^{-1}$, so gilt $Z(a) = I$. Die Lösung der homogenen AWA mit $y(a) = y_0$ ist dann gegeben durch $Z(t)\,y_0$.

Definition 6.7 *(Wronski–Determinante)*
Es seien $y_1, \dots, y_m$ Lösungen der homogenen Differentialgleichung 6.1 und

$$Y(t) = \big(y_1(t), \dots, y_n(t)\big)\,.$$

Dann heißt $W(t) = \det Y(t)$ Wronski–Determinante.

Korollar 6.8 *Es gilt entweder $W \equiv 0$ oder $W(t) \neq 0 \, \forall t \in I$.*

Beweis: 6.4 □

Satz 6.9 *(Berechnung der Wronski–Determinante)*
Seien (y_k), $k = 1 \dots n$, *Lösungen der homogenen Differentialgleichung und* $W(t) = \det(y_1(t), \dots, y_n(t))$ *die zugehörige Wronski–Determinante. Dann gilt*

$$W'(t) = \operatorname{Spur}(A(t))W(t)$$

mit $\operatorname{Spur}(A) = A_{1,1} + \dots + A_{n,n}$.

Beweis: Es sei $s \in I$, $z_i(t)$ Lösung der Anfangswertaufgabe mit $z_i(s) = e_i$ und $Z(t)$ die zugehörige Fundamentalmatrix.
Sei $M(t) = (M_1(t), \dots, M_n(t))$ eine differenzierbare Matrixfunktion. Die Determinante ist definiert durch (Bosch [2014], Kapitel 4.29)

$$\det(M(t)) = \sum_{\pi \in \Pi_n} \operatorname{sgn} \pi \, M_{1,\pi(1)}(t) \cdots M_{n,\pi(n)(t)}$$

mit den Permutationen Π_n der Zahlen $1, \dots, n$ und damit

$$\begin{aligned}
(\det M)'(t) &= \sum_{\pi \in \Pi_n} \sum_{k=1}^{n} \operatorname{sgn} \pi \, M_{1,\pi(1)}(t) \cdots M_{k-1,\pi(k-1)}(t) M'_{k,\pi(k)}(t) M_{k+1,\pi(k+1)}(t) \cdots M_{n,\pi(n)}(t) \\
&= \sum_{k=1}^{n} \sum_{\pi \in \Pi_n} \operatorname{sgn} \pi \, M_{1,\pi(1)}(t) \cdots M_{k-1,\pi(k-1)}(t) M'_{k,\pi(k)}(t) M_{k+1,\pi(k+1)}(t) \cdots M_{n,\pi(n)}(t) \\
&= \sum_{k=1}^{n} \det(M_1(t), \dots, M_{k-1}(t), M'_k(t), M_{k+1}(t), \dots, M_n(t)).
\end{aligned}$$

Es gilt $z_k(s) = e_k$, $z'_k(s) = A(s)z_k(s) = A(s)e_k$. Eingesetzt:

$$\begin{aligned}
(\det Z)'(s) &= \sum_{k=1}^{n} \det(z_1(s), \dots, z_{k-1}(s), z'_k(s), z_{k+1}(s), \dots, z_n(s)) \\
&= \sum_{k=1}^{n} \det(e_1, \dots, e_{k-1}, A(s)e_k, e_{k+1}, \dots, e_n) \\
&= \sum_{k=1}^{n} A_{k,k}(s) \\
&= \operatorname{Spur}(A(s)).
\end{aligned}$$

Es gilt $Y(t) = Z(t)\,Y(s)$. Also gilt mit dem Determinanten–Produktsatz für die Wronski–Determinante W

$$W'(s) = (\det(Z\,Y(s))'(s) = (\det Z)'(s))\det Y(s) = \operatorname{Spur}(A(s))W(s).$$

□

Korollar 6.10 *Es gilt*

$$W(t) = W(a)e^{\int_a^t \operatorname{Spur}(A(s))\,ds}.$$

Insbesondere kann man die Wronski–Determinante ausrechnen, ohne die Lösungen des Systems zu kennen.

6.2 Inhomogene Lineare Systeme

Wir betrachten 6.1 im inhomogenen Fall, d.h. diesmal ist nicht notwendig $b \equiv 0$. Es seien wieder immer A und b stetig.

Satz 6.11 *(Lösungsraum inhomogener Gleichungen)*

1. *Die Lösungen von 6.1 bilden einen affinen Unterraum von C^1.*
2. *Sei y_p eine Lösung von 6.1 (das p steht für parikuläre (spezielle) Lösung). z ist genau dann eine weitere Lösung von 6.1, wenn $z = y_p + y$, y Lösung des homogenen Systems $y' = A\,y$.*
3. *Sei y_k ein Fundamentalsystem des homogenen Systems $y' = A\,y$ und Y die zugehörige Fundamentalmatrix. Der Lösungsraum von 6.1 ist dann gegeben durch*

$$y_p + Yc,\ c \in \mathbb{R}^n.$$

Beweis: Sei y_p also eine Lösung der inhomogenen Gleichung. Eine solche existiert nach 6.2.
Sei z eine weitere Lösung, dann ist

$$(z - y_p)' = (A\,z + b) - (A,\ y_p + b) = A\,(z - y_p)$$

und damit ist $y = z - y_p$ Lösung der homogenen Gleichung. Andererseits: Ist y Lösung des homogenen Systems, so ist $y = Y(t)\,c$ für ein $c \in \mathbb{R}^n$, und

$$(y_p + y)' = Ay_p + b + A\,y = A(y_p + y) + b.$$

□

Es reicht also, ein Fundamentalsystem für das homogene Problem und eine spezielle Lösung anzugeben, um den Lösungsraum des inhomogenen Problems zu bestimmen. In 2.8 hatten wir für skalare Probleme bereits mit Variation der Konstanten spezielle Lösungen berechnet. Im Fall von Systemen geht dies genauso.

Sei $Y(t)$ eine Fundamentalmatrix des homogenen Problems. Wir machen wieder den Ansatz

$$y_p(t) = Y(t)\, c(t),\ c \in C^1.$$

Wir wollen $c(t)$ so bestimmen, dass $y_p' = Ay_p + b$, also gilt wegen $Y' = AY$

$$Y'\, c + Y\, c' = A\, Y\, c + b \Longrightarrow Yc' = b \Longrightarrow c' = Y^{-1}b.$$

Die Lösung dieses Systems mit $c(a) = 0$ ist gegeben durch

$$c(t) = \int_a^t Y(s)^{-1} b(s)\, ds.$$

Satz 6.12 *(Variation der Konstanten für Systeme)*
Für das inhomogenen System 6.1 sei $Y(t)$ eine Fundamentalmatrix aus Lösungen des homogenen Problems $y' = Ay$, $a \in I$. Dann ist

$$y_p(t) = Y(t)\, c(t),\ c(t) = \int_a^t Y(s)^{-1} b(s)\, ds$$

Lösung des inhomogenen Systems mit $c(a) = 0$.

$$z(t) = Y(t)\,(c(t) + Y(a)^{-1}\, y_0)$$

ist Lösung der Anfangswertaufgabe zu 6.1 mit $z(a) = y_0$.

Beweis: y_p ist Lösung nach den Vorbemerkungen. Es gilt $z(t) = y_p(t) + Y(t)\,(Y(a)^{-1}\, y_0)$, also ist z Lösung des inhomogenen Problems nach 6.11. Außerdem ist

$$z(a) = Y(a)\,(0 + Y^{-1}(a)\, y_0) = y_0.$$

□

6.3 Lineare Systeme mit konstanten Koeffizienten

Leider sind die homogenen Gleichungen häufig nicht einfach zu lösen. Für Beispiele ziehen wir uns daher auf einen scheinbar trivialen Fall zurück: Wir nehmen an, dass $A(t) \equiv A$ konstant ist.

Beemerkung: In 6.1 darf ruhig weiterhin $b(t) \not\equiv 0$ sein, es geht nur um die Matrixfunktion A. Im Folgenden sei also immer $A(t) \in \mathbb{C}^{n\times n}$ konstant.

Definition 6.13 *(System von linearen DGL mit konstanten Koeffizienten)*
6.1 heißt System mit konstanten Koeffizienten, wenn $A(t) = A \in \mathbb{C}^{n\times n}$ konstant.

Nach 5.26 gibt es zu jeder Matrix A eine invertierbare Matrix X und eine Jordanmatrix J, jeweils im $\mathbb{C}^{n\times n}$, so dass $A = XJX^{-1}$.

Sei nun y eine Lösung der Differentialgleichung $y'(t) = Ay(t)$. Dann gilt mit $z(t) = X^{-1}y(t)$

$$\begin{aligned} z'(t) &= X^{-1}y'(t) \\ &= X^{-1}Ay(t) \\ &= X^{-1}XJX^{-1}y(t) \\ &= Jz(t) \end{aligned}$$

und entsprechend: Ist z eine Lösung von $z' = Jz$, so ist $y' = Ay$. Also:

Lemma 6.14 *(Diagonalisierung von linearen DGL mit konstanten Koeffizienten)*
In 6.13 sei $A = XJX^{-1}$ eine Jordanzerlegung. Es seien

$$y, z : I \mapsto \mathbb{C}^n, y, z \in C^1,\ y(t) = Xz(t).$$

y ist Lösung der Differentialgleichung $y' = Ay$ genau dann, wenn $z' = Jz$.

Zur Struktur der Matrix X: Seien x_k die Spalten von X. Dann gilt

$$X = (x_1, \ldots, x_n) \Rightarrow (Ax_1, \ldots, Ax_n) = AX = XJX^{-1}X = XJ.$$

Insbesondere: Falls A diagonalisierbar ist, also J eine Diagonalmatrix ist mit $\lambda_1 \ldots, \lambda_n$ auf der Hauptdiagonalen, so gilt

$$(Ax_1, \ldots, Ax_n) = (x_1, \ldots, x_n)J = (\lambda_1 x_1, \ldots, \lambda_n x_n),$$

d.h. auf der Hauptdiagonalen stehen die Eigenwerte λ_k, die Spalten von X sind die zugehörigen Eigenvektoren.

Korollar 6.15 *(6.13 für diagonalisierbare Matrizen)*
Es sei in 6.13 A diagonalisierbar, d.h. es gibt eine Basis aus Eigenvektoren x_k zu Eigenwerten λ_k, $k = 1 \ldots n$. Dann ist

$$y_k(t) = e^{\lambda_k t}x_k,\ k = 1 \ldots n$$

ein Fundamentalsystem.

Beweis: In 6.14 ist $J = D$. Auf der Hauptdiagonalen von D stehen die Eigenwerte λ_k und in X stehen spaltenweise die Eigenvektoren x_k. Es sei $z(t) = (z_1(t), \ldots, z_n(t))$

eine Lösung von $z' = Dz$. Dann gilt $z'_k = \lambda_k z_k$ und damit ist nach Beispiel 1.2 $z_k(t) = c_k\, e^{\lambda_k t}$.
Nach 6.14 lassen sich alle Lösungen von $y' = Ay$ in der Form

$$y(t) = X\, z(t) = (x_1, \ldots, x_n) \begin{pmatrix} c_1\, e^{\lambda_1 t} \\ \vdots \\ c_n\, e^{\lambda_n t} \end{pmatrix} = \sum_{k=1}^{n} c_k e^{\lambda_k t} x_k = \sum_{k=1}^{n} c_k y_k(t)$$

darstellen, also ist y_k, $k = 1 \ldots n$, eine Basis des Lösungsraums, und damit Fundamentalsystem. □

Für diagonalisierbare Matrizen können wir also konstruktiv Fundamentalsysteme für das homogene Problem in 6.13 angeben und dann durch Variation der Konstanten 6.12 die inhomogenen Probleme lösen.

Ein Problem dabei: Häufig sind unsere Differentialgleichungen rein reell, d.h. alle Koeffizienten sind reell. Die charakteristischen Polynome zerfallen aber sicher nur über $\mathbb{C}$. Entsprechend sind unsere konstruierten Lösungen y_k möglicherweise komplex.

Korollar 6.16 *(reelle diagonalisierbare Matrizen)*
Es sei in 6.13 A reell und (über $\mathbb{C}$) diagonalisierbar. Dann gibt es ein reelles Fundamentalsystem.

Beweis: Es sei x ein Eigenvektor zum (komplexen, nicht reellen) Eigenwert λ von A. Wegen $A = \overline{A}$ gilt

$$A\overline{x} = \overline{Ax} = \overline{\lambda x} = \overline{\lambda}\,\overline{x}.$$

$\overline{x}$ ist also Eigenvektor von A zum Eigenwert $\overline{\lambda}$. Also sind nach 6.15

$$y_1(t) = e^{\lambda t}x,\ y_2(t) = e^{\overline{\lambda} t}\overline{x} = \overline{y_1(t)}$$

l.u. Lösungen des linearen Systems. Diese sind nach Voraussetzung nicht reell und ebenfalls linear unabhängig. Dann sind aber auch die Funktionen

$$\frac{1}{2}(y_1(t) + y_2(t)) = \Re y_1(t),\ \frac{1}{2i}(y_1(t) - y_2(t)) = \Im y_1(t)$$

l.u. Lösungen der homogenen Differentialgleichung, und sie sind reell. □

Beispiel 6.17 *Wir betrachten das homogene System*

$$\begin{aligned} f'(t) &= 3f(t) + 5g(t) \\ g'(t) &= -5f(t) - 3g(t). \end{aligned}$$

Die zugehörige Matrix A ist

$$A = \begin{pmatrix} 3 & 5 \\ -5 & -3 \end{pmatrix} \Rightarrow \chi_A(\lambda) = (3-\lambda)\cdot(-3-\lambda) - 5\cdot(-5) = \lambda^2 + 16.$$

Die beiden Eigenwerte sind $\pm 4i$. Ein zugehöriger Eigenvektor $x = (x_1, x_2)$ liegt im Kern von $A - \lambda I$, d.h.

$$\begin{pmatrix} 3-\lambda & 5 \\ -5 & -3-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 \Rightarrow x = \begin{pmatrix} -5 \\ 3-\lambda \end{pmatrix} \text{ ist Eigenvektor.}$$

Wir haben also

$$\lambda_1 = 4i,\ x_1 = \begin{pmatrix} -5 \\ 3-4i \end{pmatrix},\ \lambda_2 = -4i,\ x_2 = \begin{pmatrix} -5 \\ 3+4i \end{pmatrix}.$$

Ein (komplexes) Fundamentalsystem ist nun gegeben durch

$$y_1(t) = e^{4it}\begin{pmatrix} -5 \\ 3-4i \end{pmatrix},\ y_2(t) = e^{-4it}\begin{pmatrix} -5 \\ 3+4i \end{pmatrix}.$$

Die Eigenwerte, Eigenvektoren und das Fundamentalsystem sind konjugiert, wie zuvor bewiesen. Reelle Systeme konstruieren wir nun aus dem Real– und Imaginärteil von y_1. Es gilt

$$z(t) := \Re y_1(t) = \Re((\cos 4t + i\sin 4t)\begin{pmatrix} -5 \\ 3-4i \end{pmatrix}) = \begin{pmatrix} -5\cos 4t \\ 3\cos 4t + 4\sin 4t \end{pmatrix}.$$

Tatsächlich gilt für die Funktion z

$$z'(t) = \begin{pmatrix} 20\sin 4t \\ -12\sin 4t + 16\cos 4t \end{pmatrix} = \begin{pmatrix} 3 & 5 \\ -5 & -3 \end{pmatrix}\begin{pmatrix} -5\cos 4t \\ 3\cos 4t + 4\sin 4t \end{pmatrix} = Az(t).$$

Im allgemeinen Fall, also J ist eine Jordanmatrix, muss man etwas mehr arbeiten. In diesem Fall ist

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & j_r \end{pmatrix},\ J_r \in \mathbb{R}^{n_k \times n_k},\ J_k = \begin{pmatrix} \lambda_k & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_k & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \lambda_k & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & \lambda_k \end{pmatrix}.$$

Wir betrachten exemplarisch die Spalten für das erste Eigenkästchen. Es gilt wie oben

$$(Ax_1, \dots, Ax_{n_r}) = (x_1, \dots, x_{n_r})J_1 = (\lambda_1 x_1, x_1 + \lambda_1 x_2, \dots, x_{n_r-1} + \lambda_1 x_{n_r})$$

und damit ist x_1 Eigenvektor, für die restlichen Vektoren gilt $(A - \lambda_1)x_k = x_{k-1}$. Entsprechend für die restlichen Jordankästchen.

Wir betrachten in 6.14 exemplarisch die Gleichungen für das erste Jordankästische J_1. Dann gilt

$$\begin{pmatrix} z_1' \\ z_2' \\ \vdots \\ z_{n_1}' \end{pmatrix} = \begin{pmatrix} \lambda_1 z_1 + z_2 \\ \lambda_1 z_2 + z_3 \\ \dots \\ \lambda_1 z_{n_1} \end{pmatrix}.$$

Der Ansatz $z_2, \dots z_{n_1} \equiv 0$ führt auf die skalare Differentialgleichung $z_1' = \lambda z_1$, also ist

$$z_1 = c_1 e^{\lambda_1 t} \Longrightarrow y_1(t) = c_1 e^{\lambda_1 t} x_1$$

und y_1 ist Lösung der Differentialgleichung $y' = Ay$. Für jedes Jordankästchen bekommen wir also eine Lösung wie bei diagonalisierbaren Matrizen.
Wir machen nun den Ansatz $z_3, \dots, z_{n_1} \equiv 0$. Dann ist $z_2' = \lambda z_2$, also $z_2(t) = c_2 e^{\lambda_1 t}$. z_1 erfüllt dann die Gleichung

$$z_1'(t) = \lambda_1 z_1(t) + c_2 e^{\lambda_1 t}.$$

Wir lösen diese lineare skalare Gleichung mit Variation der Konstanten nach 2.8 und erhalten

$$z_2(t) = (c_2 t + \widetilde{c_2})\, e^{\lambda_1 t}.$$

Die zugehörige Lösung y_2 der Differentialgleichung $y' = Ay$ ist

$$y_2(t) = z_1(t)x_1 + z_2(t)x_2 = ((c_2 t + \widetilde{c_2})x_1 + c_2 x_2)\, e^{\lambda_1 t} = \begin{pmatrix} p_{2,1}(t) \\ \vdots \\ p_{2,n}(t) \end{pmatrix} e^{\lambda_1 t}.$$

Hierbei sind $p_{2,1}, \dots, p_{2,n}$ Polynome vom Grad ≤ 1. Diese Konstruktion kann man für $z_3, \dots, z_{n_r}$ fortsetzen und erhält per Induktion (Übungen) für $k = 1, \dots, n_{r-1}$

$$y_k(t) = \begin{pmatrix} p_{k,1}(t) \\ \vdots \\ p_{k,n}(t) \end{pmatrix} e^{\lambda_1 t},$$

wobei $p_{k,1}, \dots p_{k,n}$ Polynome vom Grad $\leq k - 1$ sind. Zu J_r erhält man also mit dieser Konstruktion n_r linear unabhängige Lösungen der Differentialgleichung $y' = Ay$.

Satz 6.18 *(Allgemeines Fundamentalsystem für 6.13)*

1. *Zu einer k–fachen Nullstelle λ des charakteristischen Polynoms gibt es k linear unabhängige Lösungen der Form*

$$y_1(t) = P_0(t)e^{\lambda t}, \ldots, y_k(t) = P_{k-1}(t)e^{\lambda t}.$$

   Die P_j sind Vektoren des $\mathbb{R}^n$, dessen Einträge Polynome vom Grad $\leq q-1$ sind, wobei q die Größe des größten Jordankästchens zum Eigenwert λ ist.

2. *Die Vereinigung der Funktionen aus (1) für alle (möglicherweise komplexen) Nullstellen ist ein Fundamentalsystem.*

Beweis: Teil 1 nach den Vorbemerkungen. Zu Teil 2: Das charakteristische Polynom zerfällt vollständig in Linearfaktoren, d.h. die Summe der Vielfachheiten der Nullstellen ist n, und damit erhalten wir in (1) n linear unabhängige Lösungen, und diese bilden ein Fundamentalsystem nach Definition 6.5. □

Beispiel 6.19 *Wir betrachten das homogene System*

$$\begin{aligned} f'(t) &= f(t) - g(t) \\ g'(t) &= 4f(t) - 3g(t) \end{aligned}$$

mit der zugehörigen Matrix

$$A = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix} \Rightarrow \chi_A(\lambda) = (1-\lambda)(-3-\lambda) + 4 = (\lambda+1)^2.$$

-1 ist also doppelte Nullstelle des charakteristischen Polynoms. Ein zugehöriger Eigenvektor x muss im Kern von $A+I$ liegen, also

$$\begin{pmatrix} 2 & -1 \\ 4 & -2 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ ist Eigenvektor.}$$

Die zugehörige Lösung der homogenen Differentialgleichung ist also

$$y_1(t) = e^{-t}\begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Weitere linear unabhängige Eigenvektoren gibt es nicht. Die Matrix A ist also nicht diagonalisierbar, sie ist ähnlich zu der Jordanmatrix

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Wir wissen, dass sich y_2 schreiben lässt als

$$y_2(x) = (c_0 + tc_1)e^{-t}$$

mit zwei unbekannten Vektoren in $\mathbb{R}^2$. y_2 ist Lösung, also gilt

$$(c_1 - c_0 - tc_1)e^{-t} = y_2'(t) = Ay_2(t) = (Ac_0 + tAc_1)e^{-t}.$$

Die Polynome links und rechts müssen gleich sein, also gilt

$$Ac_1 = -c_1,\ (A + I)c_0 = c_1.$$

Also ist c_1 ein Eigenvektor von A zum Eigenwert -1, also z.B.

$$c_1 = \begin{pmatrix}1\\2\end{pmatrix} \Rightarrow \begin{pmatrix}2 & -1\\4 & -2\end{pmatrix} c_0 = \begin{pmatrix}1\\2\end{pmatrix} \Rightarrow c_0 = \begin{pmatrix}0\\-1\end{pmatrix} \text{ ist Lösung.}$$

Eine zweite Lösung des homogenen Systems ist also gegeben durch

$$y_2(t) = (c_0 + tc_1)e^{-t} = \begin{pmatrix}t\\-1 + 2t\end{pmatrix} e^{-t}.$$

6.4 Stabilität und reelle Systeme der Ordnung $n = 2$

Als Anwendung der Überlegungen im letzten Kapitel schauen wir nun konkret auf Systeme der Ordnung $n = 2$. Wir untersuchen das homogene Problem

$$y'(t) = Ay(t),\ A = \begin{pmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{pmatrix} \in \mathbb{R}^{2\times 2},\ t \in \mathbb{R}.$$

Wir gehen vor wie in 6.3 vorgegeben. Die Eigenwerte der Matrix A sind die Nullstellen des charakteristischen Polynoms χ_A

$$\begin{aligned}\chi_A(\lambda) &= \det(A - \lambda I)\\ &= \det(\begin{pmatrix}a_{11} - \lambda & a_{12}\\a_{21} & a_{22} - \lambda\end{pmatrix})\\ &= \lambda^2 - \underbrace{(a_{11} + a_{22})}_{\text{Spur } A =: S}\lambda + \underbrace{(a_{11}a_{22} - a_{21}a_{12})}_{\det A =: D}\end{aligned}$$

Mit der pq–Formel sind die Nullstellen gegeben durch

$$\lambda_{1,2} = \frac{1}{2}(S \pm \sqrt{S^2 - 4D}).$$

Für $S^2 - 4D \neq 0$ erhalten wir zwei unterschiedliche (möglicherweise komplexe) Eigenwerte λ_1 und λ_2 mit linear unabhängigen Eigenvektoren x_1 und x_2, in diesem Fall ist ein Fundamentalsystem nach 6.3 gegeben durch

$$y_1(t) = e^{\lambda_1 t} x_1,\ y_2(t) = e^{\lambda_2 t} x_2.$$

Falls $S^2 = 4D$, so ist $\lambda_1 = \lambda_2 = \lambda = \frac{S}{2}$. Die Jordannormalform könnte die Gestalt

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \text{ oder } \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

annehmen. Bei Diagonalisierbarkeit ist

$$A = XJX^{-1} = X(\lambda I)X^{-1} = \lambda I,$$

und wie im ersten Fall erhalten wir ein Fundamentalsystem für die Basis aus Eigenvektoren $x_1 = e_1$, $x_2 = e_2$.

Andernfalls ist ein Fundamentalsystem gegeben durch

$$y_1(t) = e^{\lambda t} x_1,\ y_2(t) = e^{\lambda t}(x_2 + t x_1)$$

mit dem (einzigen linear unabhängigen) Eigenvektor x_1 und dem Hauptvektor x_2 mit $(A - \lambda I)x_2 = x_1$ (Übungen).

Wir wollen nun das qualitative Verhalten dieser Lösungen untersuchen. Hierzu nutzt man im $\mathbb{R}^2$ die Phasenportraits: Sei (f, g) eine Lösung der Differentialgleichung. Dann heißt die Kurve $(f(t), g(t))$ ihre Trajektorie. Zeichnet man viele dieser Trajektorien in ein Diagramm, so heißt das Diagramm auch Phasenportrait. Da man den Kurven nicht ansieht, in welcher Richtung sie durchlaufen wird, versieht man sie mit Pfeilen.

Bemerkung: Wenn die Trajektorien zweier Lösungen einer autonomen Differentialgleichung einen Schnittpunkt haben, so sind sie gleich (Übungen).

Wir beschränken uns auf die Betrachtung der Jordanmatrix, d.h. $x_1 = e_1$, $x_2 = e_2$. Im allgemeinen Fall werden die Kurven affin verzerrt (Drehung, Streckung, ...), aber wesentliche Eigenschaften (etwa das Verhalten für $t \to \infty$) bleiben erhalten.

Beispiel 6.20 *(Trajektorien für $n = 2$)*

1. *Sei J invertierbare Diagonalmatrix und λ_1, λ_2 reell mit $\operatorname{sgn} \lambda_1 = \operatorname{sgn} \lambda_2$. Nach unserer Vorüberlegung heißt dies $S^2 \geq 4D$, $D > 0$. Die zugehörigen Trajektorien sind*

$$y(t) = (ae^{\lambda_1 t}, be^{\lambda_2 t}),\ a, b \in \mathbb{R}$$

und für $a, \tau > 0$ mit der Parametrisierung $t = \frac{1}{\lambda_1} \log \frac{\tau}{a}$

$$y(\tau) = (\tau, b\left(\frac{\tau}{a}\right)^{\lambda_2/\lambda_1}) = (\tau, C\,\tau^{\alpha}),\ \alpha = \frac{\lambda_2}{\lambda_1} > 0,\ C = \frac{b}{a^{\alpha}}.$$

Für $\alpha = 1$ sind die Trajektorien Halbgeraden, für $\alpha = 2$ Parabeln usw. Falls die Eigenwerte das Vorzeichen -1 haben (also $S < 0$), so gehen die Lösungen gegen Null für $t \mapsto \infty$, und die Pfeile im Phasenporträt zeigen zum Nullpunkt, für $S > 0$ zeigen sie nach außen.
In gewisser Weise "konvergieren" die Lösungen der Differentialgleichung für $t \to \infty$ im ersten Fall gegen die Lösung $y \equiv 0$.

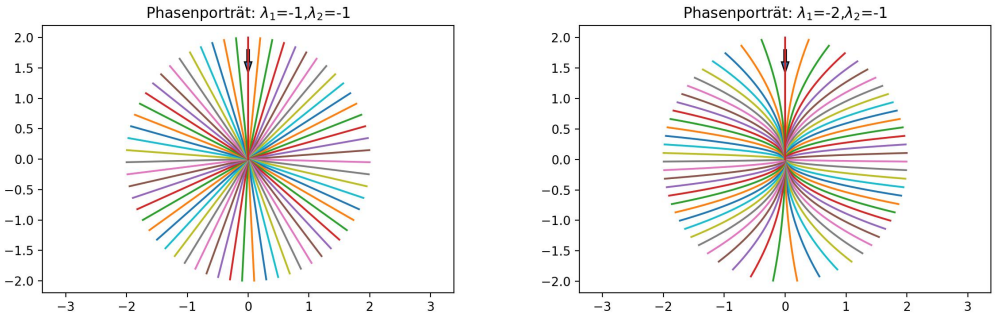


Abbildung 6.1: Phasenporträts im Fall reeller Ew. mit gleichem Vorzeichen
Klick für Bild port1
Klick für Bild port2

2. *Wie 1), aber $\lambda_1 < 0 < \lambda_2$. Dies entspricht $S^2 \geq 4D$, $D < 0$. Wir können vorgehen wie oben, aber in diesem Fall ist $\alpha < 0$. Die Kurven sind also Hyperbeln. Außer in trivialen Fällen ($b = 0$) streben die Lösungen im Betrag gegen ∞.*

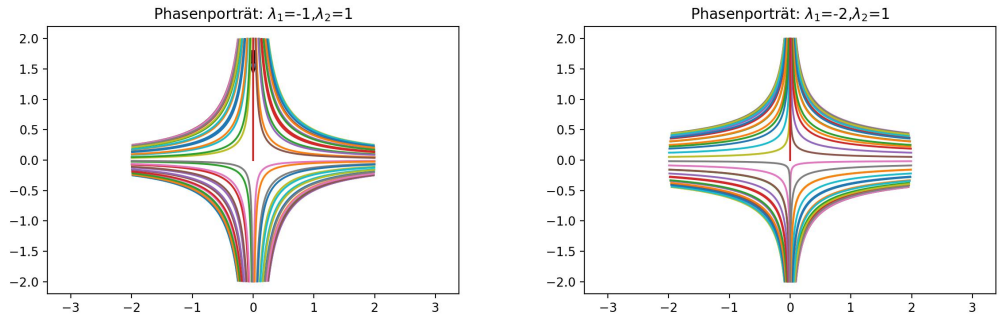


Abbildung 6.2: Phasenporträts im Fall reeller Ew. mit unterschiedlichem Vorzeichen
Klick für Bild port3
Klick für Bild port4

3. *Zwei nicht–reelle, komplex konjugierte Eigenwerte: In diesem Fall ist das oben angegebene Fundamentalsystem mit $e^{\lambda_k t}$ nicht reell. Wir beschaffen uns wie in 6.16 ein reelles Fundamentalsystem. Es sei $\lambda_1 = \alpha + i\omega$. Dann sind*

$$y_1(t) = e^{\alpha t}(\cos \omega t, -\sin \omega t),\ y_2(t) = e^{\alpha t}(\sin \omega t, \cos \omega t)$$

Lösungen der homogenen Differentialgleichung, und die Wronski–Determinante ist $e^{n\alpha t} \neq 0$, also ist y_1, y_2 ein Fundamentalsystem.
Die Trajektorien von y_1 und y_2 sind für $\alpha = 0$ Kreise um den Nullpunkt, ansonsten Spiralen um den Nullpunkt. Für $\alpha > 0$ zeigen die Pfeile nach außen, für $\alpha < 0$ nach innen.

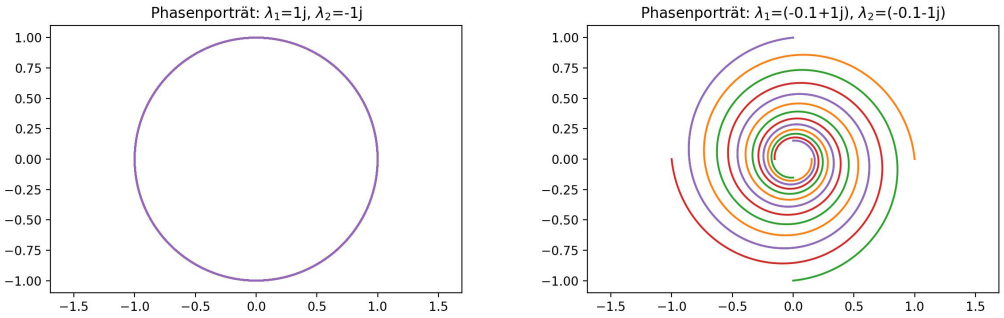


Abbildung 6.3: Phasenporträts im Fall nichtreeller Eigenwerte
Klick für Bild port5
Klick für Bild port6

Wir interessieren uns insbesondere für das Verhalten für $t \to \infty$. Den Fundamentalsystemen sieht man sofort an, dass das Verhalten vom Vorzeichen von $\Re\lambda_k$ abhängt. Es gilt

$$||y_k(t)|| \begin{cases} \to 0, & \Re\lambda_k < 0 \\ \to \infty, & \Re\lambda_k > 0 \\ \to \infty, & \Re\lambda_k = 0, A \text{ nicht diagonalisierbar}, k = 2 \\ \text{konstant}, & \text{sonst} \end{cases}.$$

Korollar 6.21 *(Verhalten homogener Lösungen von 6.13)*
Es seien λ_k, $k = 1 \dots n$, die Eigenwerte der Matrix $A \in \mathbb{C}^{n\times n}$.

1. *Falls $\Re\lambda_k < 0$, $k = 1 \dots n$, so konvergieren alle Lösungen des homogenen Problems 6.13 gegen 0 für $t \to \infty$.*
2. *Falls $\Re\lambda_k \leq 0$, und für die Eigenwerte mit $\Re\lambda_k = 0$ arithmetische und geometrische Vielfachheit übereinstimmen (d.h. die Jordankästchen haben die Größe 1), so sind alle Lösungen des homogenen Problems 6.13 beschränkt für $t > t_0$.*
3. *Andernfalls gibt es eine Lösung y des homogenen Problems 6.13 mit*

$$||y(t)|| \xrightarrow{t\to\infty} \infty.$$

Wir untersuchen das Verhalten der Trajektorien.

Definition 6.22 *(kritischer Punkt, Gleichgewichtspunkt)*
Sei y Lösung einer autonomen Differentialgleichung ($y'(t) = f(y(t))$, siehe 2.2, insbesondere ist 6.13 autonom). Falls die Trajektorie zu y nur aus einem Punkt besteht, d.h. $y(t) = \overline{y}$, so gilt

$$f(\overline{y}) = y'(t) = 0.$$

$y(t) = \overline{y}$ heißt stationäre Lösung. $\overline{y}$ heißt dann kritischer oder Gleichgewichtspunkt.

Korollar 6.23 *Für das homogene System 6.13 ist der Nullraum von A die Menge der kritischen Punkte. Falls A invertierbar, so ist A der einzige kritische Punkt.*

Dies gilt unabhängig vom Verhalten des Fundamentalsystems.

Wir betrachten nun wieder $n = 2$. Zu lösen sei nun in 6.13 die Anfangwertaufgabe mit $y(0) = y_0$. Statt y_0 stehe nur eine Näherung $\overline{y}$ zur Verfügung mit $||\overline{y} - y_0||_\infty < \epsilon$.

Sei $\widetilde{y}$ Lösung des Anfangswertproblems mit $y(0) = \overline{y}$. Dann ist $y - \widetilde{y}$ Lösung des homogenen Problems und lässt sich darstellen in der Form

$$y - \overline{y} = ay_1 + by_2.$$

Falls $Y(0) = I$, so sind $|a|, |b| < \epsilon$.

Definition 6.24 *(Stabilität, Instabilität)*

1. *Falls alle homogenen Lösungen von 6.1 gegen 0 gehen für $t \to \infty$, so gilt*

   $$||y(t) - \widetilde{y}(t)|| \xrightarrow{t\to\infty} 0,$$

   y und $\widetilde{y}$ haben also unabhängig vom Fehler das gleiche Langzeitverhalten. Dies ist zwar sehr angenehm, ist in der Anwendung aber langweilig: Alle Lösungen haben das gleiche Verhalten. Wir nennen diesen Fall asymptotisch stabil.

2. *Falls alle homogenen Lösungen von 6.1 für $t > 0$ durch C nach oben beschränkt sind, so gilt*

   $$||y(t) - \widetilde{y}(t)|| < C\epsilon.$$

   In diesem Fall haben y und $\widetilde{y}$ nicht das gleiche Langzeitverhalten, aber der Fehler ist durch den Fehler im Anfangswert beschränkt. Wir nennen diesen Fall stabil.

3. *Falls es unbeschränkte Lösungen des homogenen Problems 6.1 gibt, so gibt es Anfangswerte $\overline{y}$, so dass*

   $$||y(t) - \widetilde{y}(t)|| \xrightarrow{t\to\infty} \infty.$$

   In diesem Fall kann schon ein beliebig kleiner Fehler in den Anfangswerten dazu führen, dass der Fehler von $\widetilde{y}$ beliebig groß wird. Wir nennen diesen Fall instabil.

Korollar 6.25 *(Stabilität für konstante Koeffizienten)*
Die Differentialgleichung in Fall 1. von 6.21 ist asymptotisch stabil, in Fall 2. stabil, in Fall 3. instabil.

6.5 Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten

Zum Abschluss betrachten wir nun noch eine Anwendung auf homogene Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten.

In Beispiel 1.1 haben wir bereits gesehen, wie man diese Differentialgleichung in ein System von Differentialgleichungen erster Ordnung umwandeln kann. Wir zeigen dies noch einmal an einem etwas komplexeren Beispiel.

Gesucht seien Lösungen der homogenen Differentialgleichung

$$u^{(5)} + 4u^{(4)} + 2u^{(3)} - 4u^{(2)} + 8u^{(1)} + 16u = 0.$$

Wir setzen $y_1 = u$, $y_2 = u' = y_1'$, $y_3 = u'' = y_2'$, $y_4 = u''' = y_3'$, $y_5 = u'''' = y_4'$. Dann gilt

$$\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}' &= \begin{pmatrix} y_2 \\ y_3 \\ y_4 \\ y_5 \\ -16y_1 - 8y_2 + 4y_3 - 2y_4 - 4y_5 \end{pmatrix} \\
&= \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -16 & -8 & 4 & -2 & -4 \end{pmatrix}}_{=:A} \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}}_{=:y} \\
&= Ay.
\end{aligned}$$

Es gilt also $y' = Ay$. Im allgemeinen Fall

$$\sum_{j=0}^{n} a_j u^{(j)} = 0,\ a_j \text{ konstant},\ a_n = 1$$

erhält man entsprechend die Matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ & \cdots & & & & \\ 0 & \cdots & & & 0 & 1 \\ -a_0 & -a_1 & & \cdots & & -a_{n-1} \end{pmatrix}.$$

Laut Übungen sind die Eigenwerte dieser Matrix gerade die Nullstellen des Polynoms

$$p(\lambda) = a_0 + a_1\lambda + \ldots + a_{n-1}\lambda^{n-1} + \lambda^n,$$

und zu jeder Nullstelle λ_j mit Vielfachheit q_j gibt es nur einen linear unabhängigen Eigenvektor und ein Jordankästchen der Größe $q_j \times q_j$.

Das System besitzt ein Fundamentalsystem in der Form 6.18. Insbesondere ist die erste Komponente der Funktionen des Fundamentalsystems, also u, von der Form $p_{k,j}(t)e^{\lambda_j t}$ mit linear unabhängigen Polynomen $p_{k,j}$. Die $p_{k,j}$ bilden eine Basis des Raums der Polynome vom Grad $\leq q_j - 1$, denn: Wären sie linear abhängig, so wären $u_{k,j}$ und alle seine Ableitungen für festes j linear abhängig. Aus dieser kann man die Monome linear kombinieren. Es gilt:

Satz 6.26 *(Fundamentalsystem für homogene Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten)*
Zu lösen sei die Differentialgleichung

$$\sum_{j=0}^{n} a_j u^{(j)} = 0,\ a_j \textit{ konstant},\ a_n = 1.$$

Es sei

$$p(\lambda) = a_0 + a_1\lambda + \ldots + a_{n-1}\lambda^{n-1} + \lambda^n$$

mit den Nullstellen $\lambda_1 \ldots \lambda_m$ und den Vielfachheiten $q_1 \ldots q_m$. Dann bilden die n Funktionen

$$u_{kl}(t) = t^l e^{\lambda_k t},\ k = 1 \ldots m,\ l = 0 \ldots q_k - 1,$$

ein Fundamentalsystem von Lösungen der Differentialgleichung.

Beweis: Einen eleganteren Beweis für diesen Satz finden Sie im Walter [2013]. □

Falls λ_k nichtreell ist, so kann man sich wie in 6.16 reelle Lösungen verschaffen. In unserem Beispiel gilt

$$p(\lambda) = (\lambda + 2)^3(\lambda - 1 + i)(\lambda - 1 - i).$$

Ein reelles Fundamentalsystem ist gegeben durch die Funktionen

$$e^{-2t},\ te^{-2t},\ t^2e^{-2t},\ e^t \sin t,\ e^t \cos t.$$

6.6 Zusammenfassung

6.6.1 Kompetenzen

- Definition linearer Differentialgleichungen. Wissen, dass diese immer eine globale Lösung haben bei stetigen Funktionen $A(t)$, $b(t)$.
- Definitionen Fundamentalsystem, Fundamentalmatrix, Wronski–Determinante, 6.4 kennen.

- Lösungen inhomogener Gleichungen mit Variation der Konstanten berechnen können.
- Fundamentalsystem von Lösungen bei Gleichungen mit konstanten Koeffizienten ausrechnen können mit 6.18 und den Beispielen danach.
- Klassifikation asymptotisch stabil/stabil/instabil für eine Differentialgleichung angeben können, insbesondere im Fall $n = 2$.
- Differentialgleichung höherer Ordnung in ein System erster Ordnung umwandeln können. Fundamentalsystem angeben und berechnen können.

6.6.2 Mini–Aufgaben

- Beispiele zu 6.18 nachrechnen und das Lösungsrezept verinnerlichen.
- in Beispiel 6.20 fehlt der Fall für nicht diagonalisierbare Matrizen. Wie sehen die Trajektorien in diesem Fall aus?
- Zeigen Sie in Satz 6.26 direkt: Falls $y(t) = e^{\lambda t}$ Lösung der Differentialgleichung ist, so gilt $p(\lambda) = 0$.
- Geben Sie ein reelles Fundamentalsystem an für die Differentialgleichung $u'' - a\,u = 0$ für $a \in \{\pm 1, 0\}$. Geben Sie alle Lösungen der Differentialgleichung $u'' - a\,u = 1$ an. Lösen Sie jeweils die AWA mit $u(0) = 2$, $u'(0) = 2$.
- Wie verändert sich der Begriff der Stabilität, wenn man statt des Verhaltens für $t \to \infty$ das Verhalten auf einem kompakten Intervall betrachtet? (Tipp: alle Gleichungen werden stabil)

Kapitel 7

Stabilität für AWA: Gronwallsche Ungleichung

Wir untersuchen nun eine allgemeine AWA der Form

$$y'(t) = f(t, y(t)),\ y(a) = y_0$$

auf ihre Stabilität, d.h. auf ihr Verhalten, falls f und y_0 nicht genau bekannt sind. Wir setzen im Folgenden immer voraus, dass f und y_0 auf dem gesamten Intervall $I = [a, b]$ eine Kegelbedingung im Sinne von 3.9 erfüllt, so dass die Existenz einer eindeutigen Lösung auf dem gesamten Intervall sichergestellt ist. y_0 liegt laut Voraussetzung im Inneren des Streifens, der die Kegelbedingung erfüllt, d.h. auch bei kleinen Fehlern ist die Bedingung weiterhin erfüllt.

Wir betrachten die skalare lineare Anfangswertaufgabe 2.5

$$u'(t) = \alpha(t)\, u(t) + \beta(t),\ u(a) = u_0.$$

Es sei

$$v(t) = e^{\int_a^t \alpha(\xi) d\xi},$$

also Lösung des homogenen Problems. Die Differentialgleichung besitzt nach 2.8 (Variation der Konstanten) die Lösung

$$u(t) = v(t) \left(u(a) + \int_a^t \frac{1}{v(s)} \beta(s)\, ds \right).$$

Die Aussage des Lemmas von Gronwall ist nun: Man kann in Differentialgleichung und Lösung "$=$" durch "$\leq$" ersetzen.

7.1 Lemma von Gronwall

Satz 7.1 (Lemma von Gronwall)
Seien $I = [a, b]$ *und*

$$\alpha : I \mapsto \mathbb{R},\ \beta : I \mapsto \mathbb{R},\ u : I \mapsto \mathbb{R}$$

stetig. Sei

$$v(t) = e^{\int_a^t \alpha(\xi) d\xi}.$$

1. *Differentielle Form: Falls* u *differenzierbar ist und*

$$u'(t) \leq \alpha(t)\, u(t) + \beta(t) \,\forall\, t \in I,$$

so gilt

$$u(t) \leq v(t) \left(u(a) + \int_a^t \beta(s) \frac{1}{v(s)}\, ds \right) \,\forall\, t \in I.$$

2. *Integralform: Falls* $\alpha \geq 0$ *und*

$$u(t) \leq \beta(t) + \int_a^t \alpha(s) u(s) ds \,\forall\, t \in I,$$

so gilt

$$u(t) \leq \beta(t) + \int_a^t \beta(s)\alpha(s) \frac{v(t)}{v(s)}\, ds \,\forall\, t \in I.$$

Beweis:

1. Lemma: Sei $\varphi \in C^1(I, \mathbb{R})$, $\psi \in C^0(I, \mathbb{R})$, und

$$\varphi'(t) \leq \psi(t),\ t \in I.$$

Dann gilt

$$\varphi(t) = \varphi(a) + \int_a^t \varphi'(s)\, ds \leq \varphi(a) + \int_a^t \psi(s)\, ds.$$

2. Beweis der differentiellen Form: Zunächst gilt

$$\frac{1}{v(x)} = e^{\int_a^t -\alpha(\xi)\, d\xi} \Rightarrow \left(\frac{1}{v}\right)' = -\alpha \frac{1}{v}.$$

Da $v(t) > 0$, gilt mit der vorgegebenen Ungleichung

$$(u\frac{1}{v})' = u'\frac{1}{v} + u\left(\frac{1}{v}\right)' \leq (\alpha\, u + \beta)\frac{1}{v} - \alpha\frac{1}{v}u = \frac{\beta}{v}.$$

Nach 1. gilt damit

$$\frac{u(t)}{v(t)} \leq \frac{u(a)}{v(a)} + \int_a^t \frac{\beta(s)}{v(s)} \mathsf{ds},\ t \in I.$$

Multiplikation mit $v(t)$ liefert das Gewünschte.

3. Beweis der Integralform: Sei nun

$$w(t) := \int_a^t \alpha(s)u(s) \mathsf{ds}.$$

Mit der gegebenen Ungleichung gilt

$$u(t) \leq \beta(t) + w(t)$$

und da $\alpha(t) \geq 0$

$$w'(t) = \alpha(t)u(t) \leq \alpha(t)\beta(t) + \alpha(t)w(t),\ w(a) = 0.$$

Einsetzen in die differentielle Form liefert

$$w(t) \leq \int_a^t \beta(s)\alpha(s)\frac{v(t)}{v(s)}\, ds$$

und damit

$$u(t) \leq \beta(t) + w(t) \leq \beta(t) + \int_a^t \beta(s)\alpha(s)\frac{v(t)}{v(s)}\, ds.$$

□

Korollar 7.2 *(Gronwall für konstantes α)*
Es seien $\beta, u : I \mapsto \mathbb{R}$ stetig, $\alpha \in \mathbb{R}$, $\alpha \geq 0$, und es gelte

$$u(t) \leq \beta(t) + \alpha \int_a^t u(s)\, ds\, \forall t \in I.$$

Dann gilt

$$u(t) \leq \beta(t) + \alpha \int_a^t \beta(s)e^{\alpha(t-s)}\, ds.$$

Wir betrachten nun das Anfangswertproblem

$$y'(t) = f(t, y(t)),\ y(a) = y_0.$$

Statt f stehe nur eine Näherung $\widetilde{f}$ zur Verfügung, und statt y_0 nur eine Näherung $\widetilde{y_0}$. Wir können also nur die Lösung des Anfangswertproblems

$$\widetilde{y}'(t) = \widetilde{f}(t, \widetilde{y}(t)),\ \widetilde{y}(a) = \widetilde{y_0}$$

berechnen. Wie groß ist der Fehler, d.h. der Unterschied zwischen y und $\widetilde{y}$? Dies beantwortet

7.2 Stabilität von Anfangswertaufgaben

Satz 7.3 (Stetigkeit der Lösung von Anfangswertaufgaben)
Sei

$$y'(t) = f(t, y(t)),\ y(a) = y_0$$

eine Anfangswertaufgabe, die die Voraussetzungen von 3.9 erfüllt, insbesondere sei f Lipschitz–stetig im zweiten Argument mit der Lipschitz–Konstanten L.
Statt f und y_0 seien nur Näherungen $\widetilde{f}$ und $\widetilde{y}_0$ bekannt mit

$$||f - \widetilde{f}||_\infty \leq \epsilon,\ ||y_0 - \widetilde{y}_0|| \leq \widetilde{\epsilon}.$$

Falls die ungestörte Gleichung und die gestörte Gleichung

$$\widetilde{y}'(t) = \widetilde{f}(t, \widetilde{y}(t)),\ \widetilde{y}(a) = \widetilde{y}_0$$

Lösungen y bzw. $\widetilde{y}$ im Intervall $[a, b]$ besitzen, so gilt

$$||\widetilde{y}(t) - y(t)|| \leq (\tilde{\epsilon} + \epsilon(t-a))e^{L(t-a)}\,\forall t \in [a, b].$$

Diesen Satz kann man so interpretieren: Falls $\widetilde{y}_0 \to y_0$ und $\widetilde{f} \to f$, so gilt $\widetilde{y} \to y$. Die Lösung hängt also stetig von den Anfangswerten und der Funktion f ab (bezüglich der Supremumsnorm). Die Lösung einer Differentialgleichung ist in diesem Sinne stabil.

Bemerkung: Wir setzen hier die Existenz einer Lösung voraus. Falls das Anfangswertproblem die Kegelbedingung erfüllt, so ist sichergestellt, dass es eine Lösung auf dem gesamten Intervall $[a, b]$ besitzt. Falls ϵ und $\widetilde{\epsilon}$ klein genug sind, so erfüllt auch das gestörte Problem die Kegelbedingung und besitzt ebenfalls eine Lösung auf $[a, b]$.

Beweis: Mit $u(t) := ||\widetilde{y}(t) - y(t)||$ gilt mit der Integraldarstellung der Anfangswertaufgabe 3.5

$$\begin{aligned}
u(t) &= ||\widetilde{y}_0 - y_0 + \int_a^t \widetilde{f}(s, \widetilde{y}(s)) - f(s, y(s))\,ds|| \\
&\leq ||\widetilde{y}_0 - y_0|| + \int_a^t (||\widetilde{f}(s, \widetilde{y}(s)) - f(s, \widetilde{y}(s)) + f(s, \widetilde{y}(s)) - f(s, y(s)))||\,ds \\
&\leq ||\widetilde{y}_0 - y_0|| + \int_a^t (||\widetilde{f}(s, \widetilde{y}(s)) - f(s, \widetilde{y}(s))|| + ||f(s, \widetilde{y}(s)) - f(s, y(s)))||\,ds \\
&\leq \widetilde{\epsilon} + \int_a^t (\epsilon + L||\widetilde{y}(s) - y(s)||)\,ds \\
&\leq \underbrace{\widetilde{\epsilon} + \epsilon(t-a)}_{\beta(t)} + \underbrace{L}_{\alpha} \int_a^t u(s)\,ds.
\end{aligned}$$

Anwendung von 7.2 liefert das Gewünschte:

$$\begin{aligned} u(t) &\leq (\widetilde{\epsilon} + \epsilon(t-a)) + L \int_a^t \underbrace{(\widetilde{\epsilon} + \epsilon(s-a))}_{\leq \widetilde{\epsilon} + \epsilon(t-a)} e^{L(t-s)}\, ds \\ &\leq (\widetilde{\epsilon} + \epsilon(t-a)) \left(1 + L \int_a^t e^{L(t-s)}\, ds\right) \\ &\leq (\widetilde{\epsilon} + \epsilon(t-a)) \left(1 + [-e^{L(t-s)}]_a^t\right) \\ &= (\widetilde{\epsilon} + \epsilon(t-a)) e^{L(t-a)}. \end{aligned}$$

□

7.3 Diskretes Lemma von Gronwall

Beim Nachweis der Stabilität von Einschrittverfahren werden wir noch eine weitere Variante des Lemmas von Gronwall benutzen, dies sei der Vollständigkeit halber hier schon einmal zitiert.

Lemma 7.4 (Diskretes Lemma von Gronwall)
Seien (β_k), (α_k), (e_k) *reelle nichtnegative Folgen und*

$$e_{k+1} \leq \beta_k + (1 + \alpha_k) e_k,\ k \geq 0.$$

Dann gilt

$$e_k \leq (e_0 + \sum_{j=0}^{k-1} \beta_j) e^{\sum_{j=0}^{k-1} \alpha_j}.$$

Beweis: Durch vollständige Induktion (Übungen). □

Zum Zusammenhang der diskreten und kontinuierlichen Formulierung: Mit den Definitionen aus Kapitel 10 kann man e_k als die Auswertung einer Gitterfunktion e auf dem äquidistanten Gitter mit Gitterweite h interpretieren. Bringt man in der Formulierung oben nun e_k auf die linke Seite und teilt durch h, so ergibt sich

$$\frac{e_{k+1} - e_k}{h} \leq \frac{1}{h}(\alpha_k e_k + \beta_k).$$

Auf der linken Seite steht nun ein Differenzenquotient, also eine Approximation der Ableitung. Dies ist aber gerade die Voraussetzung der differentiellen Form des kontinuierlichen Lemmas. Approximiert man nun in der Folgerung das Integral durch eine Summe, erhält man das diskrete Lemma. (Dies ist eine reine formale Motivation.)

7.4 Zusammenfassung

7.4.1 Kompetenzen

- Aussage kennen: Lösungen einer Differentialgleichung auf einem Intervall hängen stetig von den Anfangsdaten und der Funktion f ab.

Kapitel 8

Interpolation, Numerische Integration und Differentiation

Wir suchen im Folgenden numerische Lösungen des Anfangswertproblems. Dies bedeutet: Wir suchen nicht direkt die Lösung y, sondern wir versuchen, Näherungen für $y(t_k)$ an einigen Stellen t_k anzugeben. Hierzu werden wir im Allgemeinen die Integraldarstellung 3.5 nutzen. Wir erhalten also

$$y(t_k) = y(a) + \int_a^{t_k} f(t, y(t))\, dt.$$

Problem dabei: Unter dem Integral taucht die Funktion y auf, von der wir aber nur Approximationen an den Stellen t_k kennen, d.h. wir können den Integranden nur an den Stellen t_k auswerten. Wir müssen also das Problem lösen:

Approximiere das Integral $I(g) = \int_a^b g(x)\, dx$, wobei ausschließlich Funktionsauswertungen $g(x_k)$ genutzt werden dürfen.

Naheliegend ist die folgende Idee: Bestimme eine (möglichst einfache) Funktion G, die durch die Punkte $(t_k, g(t_k))$ verläuft. Integriere statt der Funktion g die Funktion G.

Definition 8.1 *(allgemeine Interpolationsaufgabe)*
Gegeben seien paarweise verschiedene Stützstellen x_i und Stützwerte y_i, $i = 0 \dots N$. Bestimme eine Funktion p aus einem Funktionenraum X mit

$$p(x_i) = y_i,\ i = 0 \dots N.$$

Im Folgenden seien immer die Stützstellen x_i paarweise verschieden.

8.1 Polynominterpolation

Definition 8.2 (Aufgabe der Polynominterpolation, Polynomraum)
Sei $N \geq 0$. Dann ist $\mathcal{P}_N$ der Raum der Polynome vom Grad kleiner oder gleich N. Seien $x_0, \ldots, x_N$ paarweise verschieden, $y_0, \ldots, y_N$ gegeben. Dann ist die Aufgabe der **Polynominterpolation***:*
Finde ein $p \in \mathcal{P}_N$ mit $p(x_i) = y_i \, \forall i = 0 \ldots N$.

Damit gilt:

Satz 8.3 *8.2 ist* **eindeutig** *lösbar.*

Beweis:

1. Formel von Lagrange, **Existenz** einer Lösung: Sei

$$w_j(x) := \prod_{\substack{k=0 \\ k \neq j}}^{N} \frac{x - x_k}{x_j - x_k}, \; j = 0 \ldots N.$$

   Dann ist $w_j \in \mathcal{P}_N$, und

$$w_j(x_k) = \delta_{j,k} := \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}$$

   für $j, k = 0 \ldots N$ mit dem Kronecker–δ. Sei

$$p(x) := \sum_{j=0}^{N} y_j w_j(x).$$

   Dann ist $p \in P_N$, und es gilt

$$p(x_k) = \sum_{j=0}^{N} y_j w_j(x_k) = \sum_{j=0}^{N} y_j \delta_{j,k} = y_k$$

   für alle $k = 0 \ldots N$.

2. **Eindeutigkeit** der Lösung: Seien p_1 und p_2 Lösungen der Polynominterpolationsaufgabe. Sei $p = p_1 - p_2$. Dann ist $p \in \mathcal{P}_N$, und es gilt

$$p(x_k) = p_1(x_k) - p_2(x_k) = y_k - y_k = 0$$

   für alle $k = 0 \ldots N$. Also ist p ein Polynom vom Grad kleiner oder gleich N mit $N + 1$ Nullstellen, also ist nach dem Fundamentalsatz der Algebra $p = 0$, und damit $p_1 = p_2$.

□

Die Formel von Lagrange sichert die Existenz einer Lösung und gibt sie konstruktiv an. Alternativ kann man die Koeffizienten des Interpolationspolynoms mit Hilfe der Vandermondematrizen bestimmen.

Definition 8.4 (Vandermondematrizen)
Es seien x_i, $i = 0 \dots N$ paarweise verschieden. Die Matrix $V \in \mathbb{C}^{(N+1)\times(N+1)}$, $V_{ik} = (x_i)^k$, $i, k = 0 \dots N$, heißt Vandermondematrix zu $x_0, \dots, x_N$.

Also:

$$V(x_0, \dots, x_N) = \begin{pmatrix} x_0^0 & \dots & x_0^N \\ \vdots & \ddots & \vdots \\ x_N^0 & \dots & x_N^N \end{pmatrix}$$

Satz 8.5 (Invertierbarkeit der Vandermondematrizen)
Seien $x_0, \dots, x_N$ paarweise verschiedene Zahlen, $y_0, \dots, y_N$ in $\mathbb{R}$ oder $\mathbb{C}$. Sei $p(x) = \sum_{k=0}^{N} a_k x^k$. Sei $y = (y_0, \dots, y_N)^t$, $a = (a_0, \dots, a_N)^t$, $V = V(x_0, \dots, x_N)$ Vandermonde–Matrix zu $x_0, \dots, x_N$. Dann gilt:

1. *p ist genau dann Lösung des Polynominterpolationsproblems 8.2, wenn $Va = y$.*
2. *V ist invertierbar.*

Beweis:

1. Es gilt $(Va)_j = p(x_j)$ und $p \in \mathcal{P}_N$.
2. Die Interpolationsaufgabe besitzt eine eindeutige Lösung nach 8.3, also ist V injektiv und surjektiv, also invertierbar.

□

Damit lassen sich die Koeffizienten eines Interpolationspolynoms durch Lösen eines linearen Gleichungssystems der Ordnung $(N + 1)$ bestimmen.

Satz 8.6 (Abschätzung des Interpolationsfehlers)

Sei $f \in C^{(N+1)}([a, b])$, $f : [a, b] \mapsto \mathbb{R}$. Seien x_k paarweise verschieden in $[a, b]$, $k = 0 \dots N$, und sei $p \in \mathcal{P}_N$ das zugehörige Interpolationspolynom mit $p(x_k) = f(x_k)$.

Dann gilt:

$$\forall\, \overline{x} \in [a,b]\, \exists \xi \in [a,b] \text{ mit } f(\overline{x}) - p(\overline{x}) = w(\overline{x}) \frac{f^{(N+1)}(\xi)}{(N+1)!},\ w(x) := \prod_{k=0}^{N} (x - x_k).$$

Insbesondere gilt

$$\forall \overline{x} \in [a,b] :\ |f(\overline{x}) - p(\overline{x})| \leq |w(\overline{x})| \frac{||f^{(N+1)}||_\infty}{(N+1)!}$$

und

$$||f - p||_\infty \leq ||w||_\infty \frac{||f^{(N+1)}||_\infty}{(N+1)!}$$

mit der Maximumnorm $||f||_\infty = \max_{x \in [a,b]} |f(x)|$.

Beweis:

1. Sei $\overline{x} = x_k$ für ein k. Dann ist $f(\overline{x}) = p(\overline{x})$, $w(\overline{x}) = 0 \Rightarrow$ Behauptung.
2. Sei $\overline{x} \neq x_k$ für alle $k = 0 \dots N$, also $w(\overline{x}) \neq 0$. Wir betrachten den Interpolationsfehler. Dieser hat bereits $(N+1)$ Nullstellen an den interpolierenden Punkten. Wir modifizieren die Fehlerfunktion nun leicht so, dass sie noch eine zusätzliche Nullstelle bei $\overline{x}$ hat. Sei also

   $$F(x) := (f(x) - p(x)) - K w(x),\ K = \frac{f(\overline{x}) - p(\overline{x})}{w(\overline{x})}.$$

   F hat mindestens die $(N+2)$ verschiedenen Nullstellen $\overline{x}$ und x_k, $k = 0 \dots N$. Nach dem Satz von Rolle hat F' mindestens $(N+1)$ verschiedene Nullstellen, F'' mindestens N Nullstellen und $F^{(N+1)}$ hat mindestens eine Nullstelle ξ im Intervall $[a,b]$. $p \in \mathcal{P}_N$, also verschwindet $p^{(N+1)}$. Der Höchstkoeffizient von $x^{(N+1)}$ in w ist 1, also gilt

   $$w^{(N+1)}(x) = (N+1)!$$

   und damit insgesamt

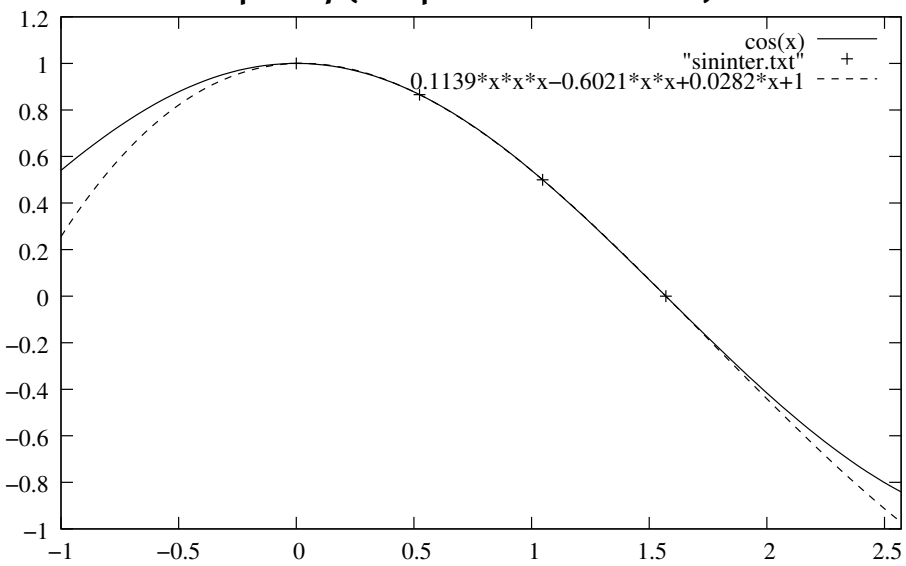
   $$0 = F^{(N+1)}(\xi) = f^{(N+1)}(\xi) - K(N+1)! \Rightarrow K = \frac{f^{(N+1)}(\xi)}{(N+1)!}$$

   und damit

   $$0 = F(\overline{x}) = f(\overline{x}) - p(\overline{x}) - \frac{f^{(N+1)}(\xi)}{(N+1)!} w(\overline{x}).$$

□

Beispiel 8.7 (Interpolation des Cosinus)

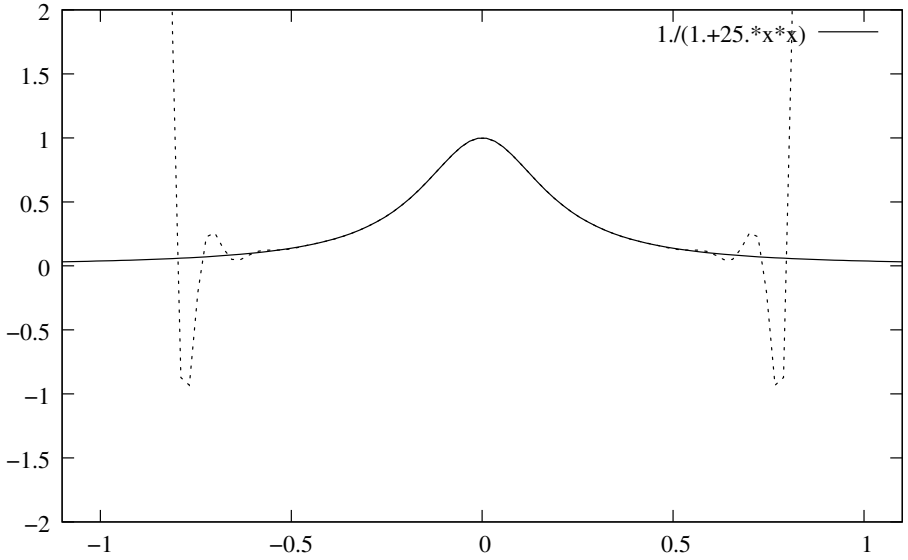


Interpolation des Cosinus auf $[0, \pi/2]$ mit vier Stützpunkten. Die Approximation ist bereits so exakt, dass innerhalb des von den Stützstellen abgedeckten Intervalls kaum ein Unterschied zwischen dem Cosinus und dem Interpolationspolynom vom Grade 3 sichtbar ist. Außerhalb steigt dagegen der Fehler schnell dramatisch an.

Beispiel 8.8 (Runge–Beispiel) *Runge and König [1925]*
Leider sind die Verhältnisse nicht immer so gut. Von Carl Runge stammt das Beispiel der Funktion

$$f(x) = \frac{1}{1 + 25x^2}$$

auf dem Intervall $[-1, 1]$: Für steigende Zahl der Stützstellen nimmt der maximale Fehler schnell zu.



Interpolation von $f(x) = 1/(1 + 25x^2)$ auf dem Einheitsintervall mit 30 äquidistanten Stützstellen. Die Approximation in der Nähe der o ist gut, am Rand beliebig schlecht.

Im Licht von Satz 8.6 stellt sich die Frage: Falls wir frei sind in der Wahl der Stützstellen, welche Wahl liefert die beste Fehlerabschätzung, also den kleinsten Wert für $||w||_\infty$?

Definition 8.9 *(**Tschebyscheff–Polynome**)*

$$T_n : [-1, 1] \mapsto \mathbb{R},\ T_n(x) := \cos(n \arccos x),\ n \in \mathbb{N}$$

heißt **Tschebyscheff–Polynom** *der Ordnung n.*

Satz 8.10 Eigenschaften der Tschebyscheff–Polynome
Für die Tschebyscheff–Polynome T_n gilt:

1. $T_n \in \mathcal{P}_n$.
2. *Für $n > 0$ hat $T_n(x)$ den Höchstkoeffizienten 2^{n-1}.*
3. *Die Nullstellen von T_{n+1} sind*

$$x_k^n = \cos\left(\frac{2k+1}{2(n+1)}\pi\right), k = 0 \ldots n.$$

4. *Wählt man für eine Polynominterpolation vom Grad n die Stützstellen x_k^n, $k = 0 \ldots n$, so ist*

$$w(x) = \prod_{k=0}^{n}(x - x_k^n) = \frac{1}{2^n}T_{n+1}(x).$$

Beweis: Übungen. □

Die Polynominterpolation, bei der wir die Stützstellen $x_0 \dots x_N$ als Nullstellen des Tschebyscheff–Polynoms T_{N+1} wählen, nennen wir **Tschebyscheff–Interpolation**. Wir erhalten für die Tschebyscheff–Interpolation nach 8.10 und 8.6 die Abschätzung

$$||f - p||_\infty \leq \frac{||f^{(N+1)}||_\infty}{2^N (N+1)!}.$$

8.2 Splines

Bei der Polynominterpolation gibt es ein riesiges Problem: Falls N groß ist, so können wir die zugehörigen Polynome nicht mehr vernünftig auswerten.

Splines beheben diesen Mangel: Sie teilen zunächst das Intervall $[a, b]$ an Knotenpunkten s_i auf. Auf jedem Einzelintervall $[s_i, s_{i+1}]$ sind die Splines (der Ordnung k) Polynome p_i vom Grad $k - 1$, mit der zusätzlichen Forderung, dass an den Knoten die zusammengesetzte Funktion $(k-2)$–mal differenzierbar ist, die Polynome von links und rechts also bis zur $(k-2)$–ten Ableitung übereinstimmen.

Definition 8.11 (Splines)
Seien $s_0 < s_1 < \dots < s_n$ reelle Zahlen. Eine Funktion

$$s : [s_0, s_n] \mapsto \mathbb{R}$$

heißt Spline der Ordnung k (zu den Knoten $s_0 \dots s_n$), falls

1. $s \in C^{(k-2)}([s_0, s_n])$ *für* $k > 1$.
2. $s|_{[s_i, s_{i+1})} \in \mathcal{P}_{k-1}$, $i = 0 \dots n-1$.

Üblicherweise wird der Spline über sein eigentliches Definitionsgebiet hinaus fortgesetzt, z.B. linear oder periodisch.

Beispiel 8.12

Die stückweise konstanten Funktionen sind Splines der Ordnung 1.

Polygonzüge (stückweise lineare stetige Funktionen) sind Splines der Ordnung 2.

Die Splines der Ordnung 4 (kubische Funktionen auf jedem Intervall, die an den Intervallenden zweimal stetig differenzierbar sind) entsprechen der Straklatteninterpolation aus dem Schiffsbau (Übungen).

Wir notieren, dass in diesem viel simpleren Fall die Konvergenz der Interpolation gegen die gegebene Funktion f (für $n \mapsto \infty$) trivial ist, ganz anders als bei den Polynomen.

Satz 8.13 *(Konvergenz von Splines der Ordnung 1)*
Sei $f : [a,b] \mapsto \mathbb{R} \in C^1([a,b])$, und x_k, $k = 0 \dots n$, seien äquidistant verteilt in $[a,b]$, also

$$x_k = a + kh,\ h = (b-a)/n.$$

Weiter sei

$$s_0 = a,\ s_k = \frac{x_{k-1} + x_k}{2},\ k = 1 \dots n, s_{n+1} = b.$$

Es sei $s^{(n)}$ der Spline der Ordnung 1 mit $s^{(n)}(x_k) = f(x_k)$ zu den Knoten $s_0, \dots, s_{n+1}$. Dann gilt

$$||s^{(n)} - f||_\infty = O(h) \xrightarrow[n\to\infty]{} 0.$$

Beweis: Sei $x \in [a,b]$, und x liege im Intervall $I = [s_k, s_{k+1}]$. I hat höchstens die Länge h. In I liegt genau ein Interpolationspunkt x_j. Nach Definition der Splines der Ordnung 1 gilt $s^{(n)}|_I \in \mathcal{P}_0$, und $s^{(n)}(x_j) = f(x_j)$. Also ist $s^{(n)}$ in diesem Intervall Interpolationspolynom der Ordnung $N = 0$. Mit unserer Formel für den Interpolationsfehler gilt

$$|s^{(n)}(x) - f(x)| \leq ||f'|_I||_\infty |x - x_j| \leq h||f'||_\infty.$$

□

Bemerkung: Für Splines der Ordnung 0 kann man die s_k dem linken oder rechten Intervall zuschlagen, der Beweis bleibt gleich.
Bemerkung: Für Splines der Ordnung 2 (Polygonzüge) gilt sogar (Übungen)

$$||s^{(n)} - f||_\infty = O(h^2).$$

8.3 Zusammenfassung

8.3.1 Kompetenzen

- Grundaufgabe der Polynominterpolation 8.2 kennen.
- Interpolationspolynom ausrechnen können mit der Lagrange–Interpolation und Vandermonde–Matrizen
- Abschätzung für den Fehler des Interpolationspolynoms 8.6 kennen und interpretieren könnnen.

- Begründen, warum die Interpolation mit hohen Polynomgraden problematisch ist (Beispiel von Runge) und warum eine günstige Auswahl von Interpolationspunkten eine bessere Abschätzung liefert als die äquidistante.
- Idee der Spline–Interpolation kennen.
- Fehlerabschätzung für die lineare Spline–Interpolation herleiten können

8.3.2 Aufgaben

In den Übungen.

Kapitel 9

Anwendungen der Polynominterpolation

Aus dem vergangenen Kapitel nehmen wir mit, dass man sich bei der Polynominterpolation auf Polynome kleinen Grades beschränken sollte. Bei vielen Interpolationspunkten (und hoher erwarteter Genauigkeit) sollte man sich auf kleine Intervalle beschränken.

Bei allen Anwendungen ist die zugrundeliegende Idee: Wenn eine Operation (Integration, Differentiation) nicht direkt möglich ist, berechne das Interpolationspolynom und führe die Operation auf dem Polynom durch.

9.1 Numerische Differentiation

Gegeben seien Funktionsauswertungen $f(x_k)$ einer differenzierbaren Funktion f, $k = 0 \dots N$. Zu berechnen sei daraus eine Approximation an eine Ableitung von f an der Stelle x. Hierzu berechnen wir das Interpolationspolynom und leiten es an der Stelle x ab.

Beispiel 9.1

1. *Gegeben seien $f(x)$ und $f(x+h)$. Das Interpolationspolynom $p \in \mathcal{P}_1$ ist*

$$p(t) = f(x) + \frac{f(x+h) - f(x)}{h}(t - x).$$

Die Approximation für die erste Ableitung ist

$$p'(x) = \frac{f(x+h) - f(x)}{h} =: D_h^+(f)(x)$$

(rechtsseitiger Differenzenquotient).

2. *Gegeben seien $f(x-h)$ und $f(x)$. Das Interpolationspolynom $p \in \mathcal{P}_1$ ist*

$$p(t) = f(x) + \frac{f(x-h) - f(x)}{h}(x-t).$$

Die Approximation für die erste Ableitung ist

$$p'(x) = \frac{f(x) - f(x-h)}{h} =: D_h^-(f)(x)$$

(linksseitiger Differenzenquotient).

3. *Gegeben seien $f(x-h)$ und $f(x+h)$. Das Interpolationspolynom $p \in \mathcal{P}_1$ ist*

$$p(t) = f(x+h) + \frac{f(x-h) - f(x+h)}{2h}(x+h-t).$$

Die Approximation für die erste Ableitung ist

$$p'(x) = \frac{f(x+h) - f(x-h)}{2h} =: D_h(f)(x)$$

(zentraler Differenzenquotient).

4. *Gegeben seien $f(x-h)$, $f(x)$ und $f(x+h)$. Das Interpolationspolynom $p \in \mathcal{P}_2$ ist (in Lagrange–Form)*

$$\begin{aligned} p(t) = f(x+h)&\frac{(t-x)(t-(x-h))}{h(2h)} \\ + f(x)&\frac{(t-(x-h))(t-(x+h))}{h(-h)} \\ + f(x-h)&\frac{(t-(x+h))(t-x)}{(-h)(-2h)} \end{aligned}$$

Die Approximation für die zweite Ableitung ist

$$p''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} =: D_h^2(f)(x)$$

(zentraler Differenzenquotient der zweiten Ableitung).

Satz 9.2

1. *Sei $f \in C^2([a,b])$. Dann gilt $\forall x \in (a,b)$*

$$|f'(x) - D_h^+(f)(x)| = O(h),\ |f'(x) - D_h^-(f)(x)| = O(h).$$

2. *Sei* $f \in C^3([a,b])$. *Dann gilt* $\forall x \in (a,b)$

$$|f'(x) - D_h(f)(x)| = O(h^2).$$

3. *Sei* $f \in C^4([a,b])$. *Dann gilt* $\forall x \in (a,b)$

$$|f''(x) - D_h^2(f)(x)| = O(h^2).$$

Beweis: Wir beweisen exemplarisch (2). Taylorreihe mit Restglied liefert

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(\xi_1)$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(\xi_2)$$

Einsetzen:

$$D_h f(x) = f'(x) + \frac{h^2}{12}(f'''(\xi_1) - f'''(\xi_2)) = f'(x) + O(h^2).$$

□

9.2 Numerische Integration: Newton–Cotes–Formeln

Aufgabe: Zu berechnen sei das Integral

$$\int_a^b f(x)\,dx$$

aus den Auswertungen der Funktion f an den Stützstellen $x_k \in [a,b]$. Wir approximieren das Integral durch das Integral des Interpolationspolynoms, sei also

$$p \in \mathcal{P}_N,\ p(x_k) = f(x_k),\ k = 0 \ldots N \ \Rightarrow \int_a^b f(x)\,dx \sim \int_a^b p(x)\,dx =: I_N(f).$$

Dann gilt

$$\begin{aligned}\int_a^b f(x)\,dx &\sim \int_a^b p(x)\,dx \\ &= \int_a^b \sum_{k=0}^{N} f(x_k) w_k(x)\,dx \qquad \text{Lagrange-Form} \\ &= \sum_{k=0}^{N} \underbrace{\int_a^b w_k(x)\,dx}_{=:A_k} f(x_k)\,dx \\ &= \sum_{k=0}^{N} A_k f(x_k).\end{aligned}$$

Wir betrachten den Spezialfall der Newton–Cotes–Formeln. Hier werden die Stützstellen aquidistant verteilt, also

$$x_k = a + kh,\ h = \frac{b-a}{N},\ k = 0 \ldots N.$$

Für $N = 1$ gilt $x_0 = a$, $x_1 = b$, $h = b - a$ und

$$A_0 = \int_a^b \frac{x-b}{a-b}\,dx = \frac{1}{a-b}\frac{(a-b)^2}{2} = \frac{b-a}{2} = \frac{h}{2}$$

und $A_0 = A_1$, also

$$\int_a^b f(x)\,dx \sim I_1(f) = \frac{h}{2}(f(a) + f(b)).$$

Dies ist die *Trapezregel*.

Für höhere Ordnungen halten wir zunächst fest:

$$\int_a^b f(x)\,dx = \frac{b-a}{2} \int_{-1}^{1} f(\frac{a+b}{2} + t\frac{b-a}{2})\,dt$$

Wir können uns bei der Integration also immer auf das Referenzintervall $[-1, 1]$ zurückziehen.

Für $N = 2$ gilt dann $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ und man erhält die Lagrange–Polynome

$$w_0(x) = \frac{(x-0)(x-1)}{(-1-0)(-1-1)},\ w_1(x) = \frac{(x+1)(x-1)}{(0+1)(0-1)},\ w_2(x) = \frac{(x+1)(x-0)}{(1+1)(1-0)}.$$

und die Koeffizienten

$$\int_{-1}^{1} w_0(x)\,dx = \int_{-1}^{1} \frac{1}{2}(x^2 - x) = \frac{1}{3}$$
$$\int_{-1}^{1} w_1(x)\,dx = \int_{-1}^{1} -x^2 + 1\,dx = -\frac{2}{3} + 2 = \frac{4}{3}$$
$$\int_{-1}^{1} w_2(x)\,dx = \int_{-1}^{1} \frac{1}{2}(x^2 + x) = \frac{1}{3}.$$

Mit $h = \frac{b-a}{2}$ gilt

$$A_0 = \frac{1}{3}h,\ A_1 = \frac{4}{3}h, A_2 = \frac{1}{3}h,$$

also

$$\int_a^b f(x)\,dx \sim I_2(f) = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Diese Formel heißt Simpsonregel oder Keplersche Faßregel.

Satz 9.3 *(Fehlerabschätzung für die Numerische Integration)*

1. *Sei*

$$f \in C^{N+1}([a,b]),\ w(x) := \prod_{j=0}^{N}(x - x_j).$$

Dann gilt

$$|I_N(f) - \int_a^b f(x)\,dx| \leq C_N\,||f^{(N+1)}||_\infty$$

mit

$$\begin{aligned} C_N &= \frac{1}{(N+1)!}\int_a^b |w(x)|\,dx \\ &\leq \frac{||w||_\infty}{(N+1)!}(b-a) \\ &\leq \frac{(b-a)^{N+2}}{(N+1)!} \end{aligned}$$

Für äquidistante Stützstellen gilt mit $h = \frac{b-a}{N}$

$$C_N \leq \left(\frac{b-a}{N}\right)^{N+2}\frac{N^{N+2}}{(N+1)!} = O(h^{N+2}).$$

2. *Sei N gerade, $f \in C^{(N+2)}$, und die Stützstellen seien nach Newton–Cotes gewählt, also*

$$x_k = a + kh,\ h = \frac{b-a}{N},\ k = 0 \dots N.$$

Dann gilt sogar

$$|I_N(f) - \int_a^b f(x)\,dx| \leq \frac{b-a}{2} C_N ||f^{(N+2)}||_\infty = O(h^{N+3}).$$

Beweis: Sei p das Interpolationspolynom zu f an den Stützstellen x_k, also

$$p \in \mathcal{P}_N,\ p(x_k) = f(x_k),\ k = 0 \dots N.$$

Nach Definition von I_N gilt

$$\begin{aligned}
|I_N(f) - \int_a^b f(x)\,dx| &= |\int_a^b p(x) - f(x)\,dx| \\
&= |\int_a^b f^{(N+1)}(\xi(x)) \frac{w(x)}{(N+1)!}\,dx| \quad \text{Interpolationsfehler} \\
&\leq \underbrace{\int_a^b \frac{|w(x)|}{(N+1)!}\,dx}_{=:C_N} ||f^{(N+1)}||_\infty.
\end{aligned}$$

Zum zweiten Teil: Sei $M = N/2$. x_M ist der Mittelpunkt $\frac{a+b}{2}$ des Intervalls. Alle x_k liegen symmetrisch links und rechts von x_M, also $x_j - a = b - x_{N-j}$. Also gilt

$$\begin{aligned}
w(a+x) &= (a + x - x_M) \prod_{j=0}^{M-1} (a + x - x_j) \prod_{j=M+1}^{N} (a + x - x_j) \\
&= (-b + x + x_M) \prod_{j=0}^{M-1} (-b + x + x_{N-j}) \prod_{j=M+1}^{N} (-b + x + x_{N-j}) \\
&= -(b - x - x_M)\,(-1)^M \prod_{j=M+1}^{N} (b - x - x_j)\,(-1)^M \prod_{j=0}^{M-1} (b - x - x_j) \\
&= -w(b-x).
\end{aligned}$$

$$w(a+x) = -w(b-x) \Longrightarrow \int_a^b w(x)\,dx = 0.$$

Wir entwickeln das $f^{(N+1)}(\xi(x))$ mit Taylor um die Intervallmitte x_M und erhalten

$$f^{(N+1)}(\xi(x)) = f^{(N+1)}(x_M) + (\xi(x) - x_M) f^{(N+2)}(\mu(x))$$

und damit

$$|I_N(f) - \int_a^b f(x)\,dx| \leq ||f^{(N+2)}||_\infty \frac{b-a}{2} C_N.$$

□

Bemerkung: Eine etwas genauere Rechnung zeigt, dass man die Konstanten noch verbessern kann (Bulirsch and Stoer [1966], p. 175). Damit erhält man die endgültigen Fehlerformeln

Fehler	Integrationsformel
$\frac{h^3}{12}\|\|f^{(2)}\|\|_\infty$	Trapezregel
$\frac{h^5}{90}\|\|f^{(4)}\|\|_\infty$	Simpson–Regel

Diese Formeln sind für großes N natürlich unbrauchbar, weil dann der Interpolationsfehler schnell wächst. Daher arbeiten wir hier mit einer erweiterten Idee, den zusammengesetzten Formeln.
Dazu teilen wir das Intervall $[a, b]$ in p Teilintervalle gleicher Größe, schreiben das Integral $\int_a^b$ als Summe der Integrale über die Teilintervalle, und verwenden auf den einzelnen Intervallen Newton–Cotes–Formeln kleiner Ordnung.
Für die Simpson–Regel ($N = 2$) erhält man so etwa für drei Teilintervalle und $x_k = a + kh$, $h = \frac{b-a}{Np} = \frac{b-a}{6}$

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_{x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \int_{x_4}^{x_6} f(x)\,dx \\ &\sim \frac{h}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + 4f(x_5) + f(x_6)).\end{aligned}$$

Satz 9.4 *(Fehlerabschätzung für zusammengesetzte Formeln)*
Sei $f \in C^{(N+1)}$. Das Integral $\int_a^b f(x)\,dx$ werde als Summe von p Teilintegralen gleicher Länge geschrieben, und auf jedem Intervall werde Newton–Cotes der Ordnung N verwendet. Die Summe liefert eine Approximation $\widetilde{I_N}(f)$. Dann gilt mit $h = \frac{b-a}{Np}$

$$|\int_a^b f(x)\,dx - \widetilde{I_N}(f)| = O(h^{N+1}).$$

Falls N gerade, $f \in C^{N+2}$ so gilt sogar

$$|\int_a^b f(x)\,dx - \widetilde{I_N}(f)| = O(h^{N+2}).$$

Beweis: Auf jedem Teilintervall J gilt

$$|\int_J f(x)\,dx - I_N(f)| \leq ch^{N+2}$$

und damit

$$\begin{aligned} |\int_a^b f(x)\,dx - \widetilde{I_N}(f)| &\leq \sum_J |\int_J f(x)\,dx - I_N(f)| \\ &\leq p\,c\,h^{N+2} \\ &= \frac{p(b-a)}{Np} c h^{N+1} \\ &= O(h^{N+1}) \end{aligned}$$

und entsprechend für den zweiten Teil. □

9.3 Richardson–Extrapolation

Zu bestimmen sei der Grenzwert

$$F(0) := \lim_{h \to 0} F(h)$$

einer Funktion F. Zur Verfügung stehen die Auswertungen der Funktion F an den Stützstellen h_k. Berechne eine Approximation für $F(0)$.

Wir gehen vor wie bei den anderen Anwendungen: Wir approximieren den Wert $F(0)$ durch $p(0)$ mit dem Interpolationspolynom p, also:

$$p \in \mathcal{P}_N,\, p(h_k) = F(h_k),\, k = 0 \ldots N \implies F(0) \sim p(0).$$

$p(0)$ heißt Richardson–Extrapolation für den Wert $F(0)$.

Wir schauen auf eine einfache Anwendung, die Berechnung der ersten Ableitung mit dem rechtsseitigen Differenzenquotienten.

Es sei nun $F(h) := D_h^+(f)(x)$. Zusätzlich stehe die Näherung $F(h/2) = D_{h/2}^+(f)(x)$ zur Verfügung. Wir wissen bereits:

$$F(h) = f'(x) + O(h).$$

Mit Lagrange erhalten wir für das Interpolationspolynom p, das F an den Stellen h und $h/2$ interpoliert,

$$p(x) = F(h)\frac{x - h/2}{h - h/2} + F(h/2)\frac{x - h}{h/2 - h}$$

und damit

$$p(0) = -F(h) + 2F(h/2).$$

Wir untersuchen die Genauigkeit dieser Formel mit Taylor. Es gilt

$$F(h) = F(0) + hF'(0) + \frac{h^2}{2}F''(\xi_1),\ F(h/2) = F(0) + \frac{h}{2}F'(0) + \frac{h^2}{8}F''(\xi_2)$$

und damit

$$p(0) = -F(h) + 2F(h/2) = F(0) + O(h^2)$$

und wir haben die Abschätzung für den Fehler von $O(h)$ auf $O(h^2)$ erhöht.

Eine weitere gängige Anwendung ist das Romberg–Verfahren. Es wendet Richardson an auf die zusammengesetzten Formeln zur numerischen Integration (Übungen).

9.4 Integration nach Gauss

Auch bei der Integration einer Funktion kann man sich natürlich wieder fragen: Angenommen, wir sind frei in der Wahl der Stellen, an denen wir auswerten. Was wären die optimalen Auswertepunkte, die die kleinsten Fehler liefern?

Die Antwort liefert die Integration nach Gauss. Wir halten erstmal fest:

Lemma 9.5 *(Exaktheit der Newton–Cotes–Formeln für Polynome)*
Sei $f \in \mathcal{P}_N$. Dann wird f durch Newton–Cotes der Ordnung N exakt integriert, d.h.

$$I(f) = I_N(f).$$

Beweis: $p := f$ ist Interpolationspolynom, denn $p(x_k) = f(x_k)$ und $p \in \mathcal{P}_N$. Also ist

$$I_N(f) = I(p) = I(f).$$

□

Wir versuchen nun, die Stützstellen x_k so zu wählen, dass sogar Polynome vom Grad $2N - 1$ exakt integriert werden. Wir definieren die orthogonalen Polynome.

Definition 9.6 *(orthogonale Polynome)*
Es sei $(,)$ ein Skalarprodukt auf dem Vektorraum $C^0([a,b])$ der stetigen Funktionen auf dem Intervall $[a,b]$. Durch Anwendung des Schmidtschen Orthogonalisierungsverfahrens (Bosch [2014], Satz 7.5) auf die Monome x^0, x^1, ... erhält man eine Folge

von Polynomen $q_n \in \mathcal{P}_n$ mit $q_n \perp \mathcal{P}_{n-1}$.
Die Polynome heißen orthogonale Polynome.

Die Polynome und ihre Nullstellen sind in Tabellenwerken vertafelt (etwa in Abramowitz and Stegun [1965]). Für das Standardskalarprodukt heißen die Polynome Legendre–Polynome q_n. Wir betrachten im Folgenden nur dieses.

Lemma 9.7 *Das orthogonale Polynom q_n hat den Grad n und besitzt genau n Nullstellen im Intervall $[a, b]$.*

Beweis: Ohne Beweis, siehe z.B. Bulirsch and Stoer [1966], Satz 3.6.10. □

Satz 9.8 *(Gauss–Integration)*
Es sei

$$(f, g) := \int_a^b f(x)\, g(x)\, dx.$$

Weiter seien $x_0, \ldots, x_N$ die Nullstellen des zugehörigen orthogonalen Polynoms q_{N+1}. Dann werden durch die zugehörigen Newton–Cotes–Formeln Polynome $p \in \mathcal{P}_{2N+1}$ exakt integriert.
Die zugehörigen Gewichte A_k sind positiv.

Beweis: Nach 9.5 gilt

$$I_N(p) = I(p) \,\forall\, p \in \mathcal{P}_N.$$

Sei nun $p \in \mathcal{P}_{2N+1}$. Mit Polynomdivision gilt

$$p = s\, q_{N+1} + r,\ r, s \in \mathcal{P}_N.$$

Nach Voraussetzung steht q_{N+1} senkrecht auf s, also gilt

$$I(p) = \int_a^b s(x)\, q_{N+1}(x) + r(x)\, dx = \int_a^b r(x)\, dx = I(r).$$

Da auch $r \in \mathcal{P}_N$, gilt

$$\begin{aligned} I(p) &= I(r) \\ &= I_N(r) \\ &= \sum_{k=0}^{N} A_k r(x_k) \\ &= \sum_{k=0}^{N} A_k (s(x_k) q_{N+1}(x_k) + r(x_k)) \\ &= \sum_{k=0}^{N} A_k p(x_k) \\ &= I_N(p). \end{aligned}$$

Sei nun $w_k \in \mathcal{P}_N$ das Lagrange–Polynom aus 8.3. Wir haben gerade gezeigt: $w_k^2 \in P_{2N}$ wird exakt integriert. Also gilt

$$A_k = \sum_{j=0}^{N} A_j (w_k(x_j))^2 = \int_a^b w_j(x)^2 \, dx > 0.$$

□

Korollar 9.9 *(Fehlerabschätzung für die Gauss–Integration)*
Es sei $f \in C^{2N+2}$. Für die Gauss–Integration gilt

$$|I(f) - I_N(f)| \leq 2(b-a) \left(\frac{b-a}{2} \right)^{2N+2} \frac{1}{(2N+2)!} ||f^{(2N+2)}||_\infty.$$

Beweis: Sei $\overline{x} = \frac{a+b}{2}$ und $x \in [a,b]$. Mit Taylorentwicklung gilt

$$f(x) = \underbrace{\sum_{k=0}^{2N+1} f^{(k)}(x) \frac{(x-\overline{x})^k}{k!}}_{=:p(x) \in \mathcal{P}_{2N+1}} + \underbrace{f^{2N+2}(\xi(x)) \frac{(x-\overline{x})^{(2N+2)}}{(2N+2)!}}_{=:r(x)}.$$

Nach 9.8 gilt $I(p) = I_N(p)$. Da insbesondere das konstante Polynom 1 korrekt integriert wird, gilt

$$(b-a) = \int_a^b 1 \, dx = \sum_{k=0}^{N} A_k \, 1 = \sum_{k=0}^{N} |A_k|$$

mit 9.8. Damit gilt

$$\begin{aligned}|I(f) - I_N(f)| &= |I(r) - I_N(r)| \\ &\leq \int_a^b |r(x)|\,dx + \sum_{k=0}^{N} |A_k|\,|r(x_k)|.\end{aligned}$$

Der Satz folgt nun mit

$$|r(x)| \leq ||f^{2N+2}||_\infty \left(\frac{b-a}{2}\right)^{2N+2} \frac{1}{(2N+2)!}.$$

□

Die Gauss–Formeln sind optimal, denn

Satz 9.10 *Keine Newton–Cotes–Formel ist exakt für alle Polynome in $\mathcal{P}_{2N}$.*

Beweis: Übungen. □

Bemerkung: Die Sätze bleiben richtig für Skalarprodukte der Form

$$(f, g) := \int_a^b w(x)\, f(x)\, g(x)\, dx$$

mit einer positiven Gewichtsfunktion w. In diesen Fällen gelten die Abschätzungen für das Integral

$$I(f) = \int_a^b w(x)\, f(x)\, dx.$$

Die Beweise oben gehen ohne Änderung durch.

9.5 Zusammenfassung

9.5.1 Kompetenzen

- Grundidee der Anwendungen der Polynominterpolation kennen: Eine Operation wird statt auf einer Funktion f auf einem zugehörigen Interpolationspolynom p durchgeführt.
- Anwendung auf Funktionsauswertung, Approximation der Ableitung, Approximation des Integrals kennen.
- Newton–Cotes–Formeln mit dieser Idee ausrechnen können.

- Fehlerabschätzung kennen oder aus dem Fehler der Polynominterpolation herleiten können.
- Zusammengesetzte Formeln kennen und wissen, warum man sie einführt (kleine Polynomgrade bei der Interpolation).

9.5.2 Mini–Aufgaben

Siehe Übungen.

Kapitel 10

Diskrete Lösung von Anfangswertaufgaben

In den folgenden Kapiteln betrachten wir Anfangswertaufgaben der Form:

Definition 10.1 *(Allgemeine Anfangswertaufgabe)*
Gesucht sei eine Funktion $y : [a, b] \mapsto \mathbb{R}^n$ *mit*

$$y'(t) = f(t, y(t)),\ y(a) = y_0.$$

Hierbei seien stets die Voraussetzungen des lokalen Satzes von Picard–Lindelöf (3.9) erfüllt, d.h. f *stetig,* f *lipschitzstetig im 2. Argument mit Lipschitzkonstante* L*, und die Kegelbedingung sei auf* $[a, b]$ *erfüllt, d.h. auf einem Streifen gilt* $|f(t, y)| \leq M$ *und der von* y_0 *ausgehende Kegel* K_M *mit Steigung* M *liege ganz in diesem Streifen.*

Die Lösung dieser Aufgabe ist eindeutig bestimmt, und jede Anfangswertaufgabe zu dieser Differentialgleichung mit Anfangswerten im Kegel besitzt eine eindeutige Lösung. Wir werden im Folgenden die Lösbarkeit nicht genauer betrachten, dies ist immer bereits durch die starken Voraussetzungen gesichert.

Gelegentlich werden wir uns bei Beweisen und Betrachtungen auf skalare Differentialgleichungen ($n = 1$) zurückziehen. Grundsätzlich gelten die Sätze auch für Systeme und sind auch so formuliert. Sollte dies nicht der Fall sein, ist dies ausdrücklich vermerkt.

10.1 Numerische Verfahren

Ein numerisches Verfahren bestimmt Näherungen an die Lösung der Differentialgleichung auf einer endlichen Folge von Zahlen in $[a, b]$.

Definition 10.2 *(Gitter)*
Es sei

$$I_h = \{t_0 = a, t_1, \ldots, t_{N-1}, t_N = b\}$$

mit $t_0 < t_1 < \ldots < t_{N-1} < t_N$*. Dann heißt* I_h *(zulässiges) Gitter auf dem Intervall* $[a, b]$*.*

$$h = \max_k (t_{k+1} - t_k)$$

heißt Feinheit des Gitters.

Definition 10.3 *(numerisches Verfahren, Gitterfunktion)*
Ein numerisches Verfahren bestimmt zu einer gegebenen Anfangswertaufgabe 10.1 ein Gitter $I_h = \{t_k\}$ *auf dem Intervall* $[a, b]$ *und eine Funktion* $y_h : I_h \mapsto \mathbb{R}^n$ *mit* $y_k := y_h(t_k) \sim y(t_k)$*,* $k = 0 \ldots N$*.* y_h *heißt Gitterfunktion (und ist nur auf dem Gitter definiert).*

Im Folgenden werden wir annehmen, dass die t_k äquidistant verteilt sind auf dem Intervall $[a, b]$, also $t_k = a + kh$ mit $h = \frac{b-a}{N}$.

Im Programmierdokument zur Vorlesung motiviere ich das Eulersche Polygonzugverfahren graphisch. Die zentrale Idee: es gilt

$$y_{k+1} := y_k + hf(t_k, y_k).$$

Die wesentliche Idee: Wir wissen $y(t_0) = y_0$. Also gilt auch $y'(t_0) = f(t_0, y_0)$, denn y ist Lösung der Differentialgleichung. Wir können die Tangente an die Funktion im Punkt (t_0, y_0) also angeben, die Tangentengleichung ist

$$T(t) = y_0 + (t - t_0)\, f(t_0, y_0).$$

Die Idee von Euler ist: Approximiere die Funktion y durch ihre Tangente, also

$$y(t_1) \sim T(t_1) = y_0 + (t_1 - t_0)\, f(t_0, y_0) = y_0 + h\, f(t_0, y_0) =: y_1.$$

Damit haben wir eine Näherung für $y(t_1)$. Ausgehend von dieser Näherung bilden wir nun wieder die Tangentengleichung usw., insgesamt erhalten wir

$$y(t_{k+1}) \sim \underbrace{y_k + h\, f(t_k, y_k) =: y_{k+1}}_{\text{Numerische Approximation}}, k = 0 \ldots N - 1.$$

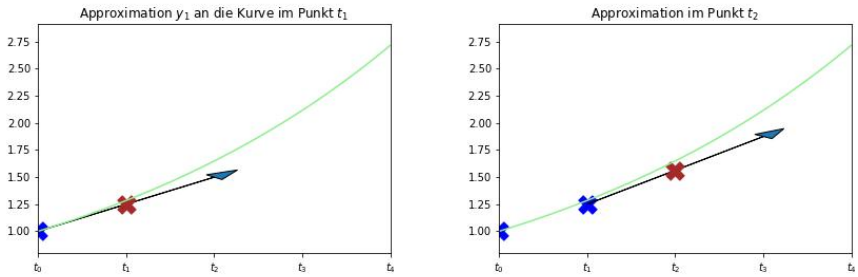


Abbildung 10.1: Motivation des Eulerverfahrens
Klick für Bild euler1
Klick für Bild euler2

Hier ist eine Bemerkung ganz wichtig: Im ersten Schritt benutzen wir $y_0 = y(t_0)$, und wir haben eine echte Tangente an die Kurve. Der Fehler im ersten Schritt ist also, dass wir die Funktion durch ihre Tangente approximieren.

Schon im nächsten Schritt ist aber nur noch $y_1 \sim y(t_1)$. Die Tangente selbst hat also einen Fehler, und dazu kommt noch der Fehler den wir durch die Approximation machen. Der Fehler $|y(t_2) - y_2|$ setzt sich also zusammmen aus einem mitgeschleppten Fehler und dem Fehler, der an diesem Punkt durch das Ersetzen der Kurve entsteht. Letzteren bezeichnen wir auch als *Lokalen Diskretisierungsfehler* (wird später noch genau definiert).

Wir wollen diese Formel noch dreimal analytisch motivieren. Diese drei Zugänge werden später zu unterschiedlichen numerischen Verfahren führen.

Taylorentwicklung: Es gilt

$$\begin{aligned} y(t_{k+1}) &= y(t_k + h) \\ &\sim y(t_k) + hy'(t_k) \\ &= y(t_k) + hf(t_k, y(t_k)) \end{aligned}$$

Also

$$y(t_{k+1}) \sim y(t_k) + hf(t_k, y(t_k)) \sim y_k + hf(t_k, y_k) =: y_{k+1}.$$

Hier sehen wir ganz klar, dass tatsächlich für $k > 0$ zwei Approximationen durchgeführt werden.

Numerische Differentiation: Wir ersetzen die Ableitung durch den Differenzenquotienten und erhalten

$$\frac{y(t_k + h) - y(t_k)}{h} \sim y'(t_k) = f(t_k, y(t_k))$$

Einsetzen von $y(t_k) \sim y_k$ und $y(t_k + h) \sim y_{k+1}$ liefert wieder das Gewünschte.

Numerische Integration: In der Integraldarstellung der Differentialgleichung gilt

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt \tag{10.1}$$

Wir verwenden den einzigen Stützpunkt $t = t_k$ für die numerische Integration und erhalten

$$\int_{t_k}^{t_{k+1}} f(t, y(t))\, dt \sim h f(t_k, y(t_k)) \sim h f(t_k, y_k).$$

Dies ist natürlich alles reine Motivation, wir müssen beweisen, dass diese Ideen gute Approximationen liefern, insbesondere, dass die Approximation immer besser wird, wenn $h \to 0$. Wir messen daher den Unterschied zwischen der Lösung und der Approximation mit dem globalen Diskretisierungsfehler.

Definition 10.4 (globaler Diskretisierungsfehler)
Sei y_h diskrete Näherung für die Lösung y der Anfangswertaufgabe 10.1 auf dem Gitter I_h mit Feinheit h, also y_h Gitterfunktion auf I_h. Dann heißt

$$e_h : I_h \mapsto \mathbb{R}^n,\ e_h := y|_{I_h} - y_h$$

die Fehlerfunktion der Näherung.

$$||e_h||_\infty = \max_{t \in I_h} ||e_h(t)||$$

heißt globaler Diskretisierungsfehler.

Der globale Diskretisierungsfehler ist das Betragsmaximum des Unterschieds von y und y_h auf dem Gitter. Es liegt nahe, zu definieren: Ein Verfahren ist konvergent, wenn diese Differenz gegen 0 geht, d.h. wenn man die Feinheit der Gitter immer kleiner wählt.

Definition 10.5 (Konvergenz von numerischen Verfahren)
Gegeben sei ein numerisches Verfahren, dass zur Feinheit h ein Gitter I_h und eine Approximation (Gitterfunktion) y_h liefert. Das Verfahren heißt konvergent, falls der globale Diskretisierungsfehler von y_h mit h gegen 0 geht, also

$$||e_h||_\infty \xrightarrow{h \to 0} 0.$$

Das Verfahren heißt konvergent von der Ordnung p, falls

$$||e_h||_\infty = O(h^p).$$

Hierbei bedeutet eine hohe Ordnung (ein großes p) wieder, dass der Fehler schnell mit h gegen 0 geht.

Wir wollen nun zunächst numerische Verfahren klassifizieren. Wir betrachten alle Verfahren in der an das Euler–Verfahren angelehnten Form

$$y_{k+1} = y_k + h\varphi. \tag{10.2}$$

Hierbei ist φ ein Ausdruck, in dem Auswertungen der Funktion f, die Gitterpunkte t_k und die Näherungen y_k vorkommen können φ heißt Verfahrensfunktion. Für das Eulerverfahren etwa gilt

$$\varphi = f(t_k, y_k).$$

Definition 10.6 *Klassifizierung von Numerischen Verfahren*
Ein Verfahren sei gegeben in der Form 10.2. Dann heißt das Verfahren

Explizites Einschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_k, y_k).$$

In diesem Fall wird nur der letzte berechnete Wert genutzt, um den nächsten auszurechnen.

Implizites Einschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_k, y_k, y_{k+1}).$$

In diesem Fall muss in jedem Schritt eine Gleichung gelöst werden.

Explizites Mehrschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_{k-r}, \ldots, t_k, y_{k-r}, \ldots, y_k).$$

In diesem Fall werden die letzten $r+1$ Näherungen genutzt, um die nächste auszurechnen.

Implizites Mehrschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_{k-r}, \ldots, t_k, y_{k-r}, \ldots, y_k, y_{k+1}).$$

In diesem Fall werden die letzten $r+1$ Näherungen genutzt, um die nächste auszurechnen, und es muss in jedem Schritt eine Gleichung gelöst werden.

10.2 Beispiele, Konsistenz und Konvergenz für explizite Einschrittverfahren

Wir behandeln zunächst nur die expliziten Einschrittverfahren. Unser numerisches Verfahren definiert also ein φ, und es gilt

$$\underbrace{y_{k+1}}_{\sim y(t_{k+1})} = \underbrace{y_k}_{\sim y(t_k)} + h\varphi(t_k, \underbrace{y_k}_{\sim y(t_k)}).$$

Dies macht schon klar, wie wir das φ wählen sollten, nämlich

$$\varphi(t_k, y(t_k)) \sim \frac{y(t_k + h) - y(t_k)}{h}.$$

Den Unterschied dieser beiden Terme bezeichnen wir als Konsistenzfehler. Dies ist genau der lokale Diskretisierungsfehler, den wir schon oben erwähnt haben.

Definition 10.7 *(Konsistenz von expliziten Einschrittverfahren)*
Sei φ die Verfahrensfunktion eines expliziten Einschrittverfahrens. Sei y (irgend-) eine Lösung der Differentialgleichung, $(t, y(t))$ im Kegel K_M.

$$\tau_h(t, y(t)) := \frac{y(t + h) - y(t)}{h} - \varphi(t, y(t))$$

heißt Konsistenzfehler oder lokaler Diskretisierungsfehler. Das Verfahren heißt konsistent, falls

$$\sup_{y,t} |\tau_h(t, y(t))| = ||\tau_h||_\infty \xrightarrow{h\to 0} 0.$$

Das Verfahren heißt konsistent von der Ordnung p, falls

$$\sup_{y,t} |\tau_h(t, y(t))| = ||\tau_h||_\infty = O(h^p).$$

Schreibt man den Konsistenzfehler als

$$\tau_h(t, y(t)) := \frac{1}{h}(y(t + h) - (y(t) + h\varphi(t, y(t))),$$

so sieht man: Der Konsistenzfehler ist der Unterschied zwischen der Lösung $y(t+h)$ und der durch das diskrete Verfahren vorhergesagten Näherung, wenn man in das diskrete Verfahren die Lösung $y(t)$ einsetzt, und ist damit der Fehler, der lokal an der Stelle t entsteht.

Zum Nachweis der Konsistenz eines Verfahrens ist das folgende Lemma nützlich.

Lemma 10.8 *Sei f stetig differenzierbar. Dann $\exists\, C \in \mathbb{R}$ so dass*

$$||y''||_\infty \leq C$$

für alle Lösungen y der Differentialgleichung, deren Graph im Kegel K_M liegt. Insbesondere ist y zweimal stetig differenzierbar.

Beweis: Es gilt

$$y'(t) = f(t, y(t)) \Longrightarrow y''(t) = f_t(t, y(t)) + f_y(t, y(t)) f(t, y(t)).$$

(Hierbei sei immer f_t die Ableitung von f nach der ersten Variablen usw.)
Also

$$||y''||_\infty \leq ||f_t||_\infty + ||f_y||_\infty \, ||f||_\infty =: C.$$

Es gilt $C < \infty$, denn $(t, y(t))$ liegt in der kompakten Menge K_M und alle Funktionen sind stetig. □

Korollar 10.9 *Es sei f r–mal stetig differenzierbar. Dann $\exists\, C \in \mathbb{R}$ so dass*

$$||y^{(r+1)}||_\infty \leq C$$

für alle Lösungen y der Differentialgleichung, deren Graph im Kegel K_M liegt. Insbesondere ist y $(r+1)$–mal stetig differenzierbar.

Beispiel 10.10 *(Beispiele für explizite Einschrittverfahren und Konsistenz) Sei im Folgenden immer y irgendeine Lösung der Differentialgleichung.*

1. **Eulersches Polygonzugverfahren:**
   Das Eulersche Polygonzugverfahren ist ein explizites Einschrittverfahren mit der Verfahrensfunktion $\varphi(t_k, y_k) = f(t_k, y_k)$, also

   $$y_{k+1} = y_k + h f(t_k, y_k).$$

   Sei f stetig differenzierbar. Dann ist nach 10.8 y zweimal stetig differenzierbar auf I. Es gilt mit Taylorentwicklung und der Differentialgleichung

   $$\begin{aligned} \tau_h(t, y(t)) &= \frac{y(t+h) - y(t)}{h} - f(t, y(t)) \\ &= \frac{y(t) + h y'(t) + \frac{h^2}{2} y''(\xi) - y(t)}{h} - y'(t) \\ &= \frac{h}{2} y''(\xi) \\ &\leq \frac{||y''||_\infty}{2} h = O(h). \end{aligned}$$

Das Eulerverfahren ist also konsistent, und zwar von der Ordnung 1. Das Eulerverfahren benötigt eine Auswertung von f in jedem Schritt. Bei der Abschätzung der zweiten Ableitung haben wir natürlich das Lemma 10.8 benutzt.

2. **Verbessertes Eulerverfahren:**
Ein verbessertes Verfahren ergibt sich, wenn wir zur Approximation des Integrals in 10.1 in der Mitte des Intervalls auswerten statt am linken Rand. Mit Taylorentwicklung gilt

$$\begin{aligned}\int_{t_k}^{t_k+h} f(t,y(t))dt &\sim h(f(t_k+h/2, y(t_k+h/2)))\\ &\sim h(f(t_k+\frac{h}{2}, y(t_k)+\frac{h}{2}y'(t_k))\\ &= h(f(t_k+\frac{h}{2}, y(t_k)+\frac{h}{2}f(t_k,y(t_k)))).\end{aligned}$$

Als Verfahrensfunktion wählen wir also

$$\varphi(t_k,y_k) = f(t_k+\frac{h}{2}, y_k+\frac{h}{2}f(t_k,y_k)).$$

Die Konsistenzordnung weisen wir wieder durch Taylorentwicklung nach. Wir benötigen diesmal, dass f zweimal stetig differenzierbar ist. Damit existiert nach 10.9 die dritte Ableitung von y auf I. Zusätzlich beachten wir wieder, dass

$$y''(t) = (f_t + f f_y)(t,y(t)).$$

Mit ein– bzw. zweidimensionaler Taylorentwicklung gilt

$$\begin{aligned}\tau_h(t,y(t)) &= \frac{y(t+h)-y(t)}{h} - f(t+\frac{h}{2}, y(t)+\frac{h}{2}f(t,y(t)))\\ &= \frac{y(t)+hy'(t)+\frac{h^2}{2}y''(t)+\frac{h^3}{6}y'''(\xi(h))-y(t)}{h} -\\ &\quad (f(t,y(t))+\frac{h}{2}f_t(t,y(t))+\frac{h}{2}f(t,y(t))f_y(t,y(t))+O(h^2))\\ &= O(h^2).\end{aligned}$$

Das Verfahren ist konsistent von zweiter Ordnung und benötigt zwei Auswertungen von f pro Schritt.

3. **Verfahren von Heun:**
Wir können auch mit der Trapezregel integrieren, hierdurch ergibt sich das Verfahren von Heun. Wir nehmen an, dass f zweimal stetig differenzierbar ist.

$$\begin{aligned}\int_{t_k}^{t_{k+1}} f(t,y(t))dt &\sim \frac{h}{2}(f(t_k,y(t_k))+f(t_k+h,y(t_k+h)))\\ &\sim \frac{h}{2}(f(t_k,y(t_k))+f(t_k+h,y(t_k)+hf(t_k,y(t_k)))).\end{aligned}$$

Die Verfahrensfunktion ist

$$\varphi(t,y)=\frac{1}{2}(f(t,y)+f(t+h,y+hf(t,y))).$$

Wieder gilt mit Taylorentwicklung (für f in zwei Dimensionen)

$$\begin{aligned}\tau_h(t,y(t)) &= \underbrace{\frac{y(t+h)-y(t)}{h}}_{y'+\frac{h}{2}y''+O(h^2)} - \frac{1}{2}\underbrace{(f(t,y(t))+f(t+h,y(t)+hf(t,y(t))))}_{y'+(y'+h(f_t+ff_y)+O(h^2))=2y'+hy''+O(h^2)}\\ &= O(h^2).\end{aligned}$$

Das Verfahren ist also ebenfalls konsistent von der Ordnung 2 und benötigt ebenfalls zwei Auswertungen von f pro Schritt.

Der Konsistenzfehler kann also in unseren Beispielen sehr einfach abgeschätzt werden, viel einfacher als etwa der globale Diskretisierungsfehler. Aber es ist noch völlig unklar, warum wir die Konsistenz überhaupt betrachten. Uns interessiert eigentlich die Genauigkeit unserer Abschätzung, und das ist der globale Diskretisierungsfehler. Der folgende Satz klärt das.

Satz 10.11 (Konvergenz von expliziten Einschrittverfahren)
Ein explizites numerisches Einschrittverfahren zur Lösung der Anfangswertaufgabe 10.1 mit Verfahrensfunktion φ sei lipschitzstetig in der zweiten Variable y mit Lipschitzkonstanten L' und konsistent (von der Ordnung p). Dann ist das Verfahren auch konvergent (von der Ordnung p).

Bemerkung: φ enthält in allen unseren Beispielen nur Auswertungen von f. Die Lipschitzstetigkeit von φ folgt daher sofort aus der Lipschitzstetigkeit von f, mit derselben Lipschitzkonstanten L.

Beweis: Sei y die Lösung der Anfangswertaufgabe 10.1. Sei $I_h = (t_k)$ das äquidistante Gitter mit Feinheit h mit zugehöriger numerischer Approximation y_h. Wir setzen zunächst

$$e_k := ||y(t_k)-y_k||$$

(globaler Diskretisierungsfehler an der Stelle t_k). Insbesondere ist $e_0 = 0$.

$$\begin{aligned} e_{k+1} &= ||y(t_{k+1}) - y_{k+1}|| \\ &= ||y(t_{k+1}) - (y_k + h\varphi(t_k, y_k))|| \\ &= ||y(t_{k+1}) - (y(t_k) + h\varphi(t_k, y(t_k))) + y(t_k) - y_k \\ &\quad + h(\varphi(t_k, y(t_k)) - \varphi(t_k, y_k))|| \\ &\leq h|\tau_h(t_k, y(t_k))| + e_k + hL'||y(t_k) - y_k|| \\ &= \underbrace{h|\tau_h(t_k, y(t_k))|}_{\beta_k} + (1 + \underbrace{hL'}_{\alpha_k})e_k. \end{aligned}$$

Mit dem diskreten Lemma von Gronwall (7.4) gilt also $(k = 0 \dots N)$

$$\begin{aligned} e_k &\leq \left(e_0 + \sum_{j=0}^{k-1} h|\tau_h(t_j, y(t_j))| \right) e^{L' \sum_{j=0}^{k-1} h} \\ &\leq (e_0 + kh||\tau_h||_\infty)\, e^{L'kh} \\ &\leq (e_0 + (b-a)||\tau_h||_\infty) e^{L'(b-a)} \\ &\xrightarrow{h \to 0} 0. \end{aligned}$$

Die Schranke hängt nicht von k ab, die Konvergenz ist gleichmäßig, also konvergiert die Supremumsnorm des globalen Diskretisierungsfehlers gegen 0.
Falls ein Verfahren konsistent ist (von der Ordnung p), so ist es auch konvergent (von der Ordnung p). □
Dies lässt sich in dem Merksatz zusammenfassen:

Für Einschrittverfahren gilt: Aus Konsistenz folgt Konvergenz.

Korollar 10.12 (Konvergenz der Referenzverfahren)
Das Eulerverfahren ist konvergent von der Ordnung 1. *Das Verfahren von Heun und das verbesserte Eulerverfahren sind konvergent von der Ordnung* 2.

Wenn man hier genau hinschaut, haben wir das Korollar eigentlich nur für skalare Differentialgleichungen bewiesen. Für Runge–Kutta–Verfahren (siehe Abschnitt 10.16), und alle bisher betrachteten Verfahren sind Vertreter dieser Klasse, folgt daraus aber auch die Konvergenz für Systeme.

10.3 Implizite Einschrittverfahren

Alle Verfahren zur Lösung von Anfangswertaufgaben unterscheiden sich nur durch die Wahl des Ausdrucks φ in 10.2. Bei den expliziten Einschrittverfahren war φ eine Funktion in t_k und y_k. Für die impliziten Verfahren lassen wir jetzt zu, dass φ auch noch von y_{k+1} abhängt. Wir betrachten zunächst zwei Beispiele.

Beispiel 10.13 (Beispiele für implizite Verfahren)

1. *Implizites Eulerverfahren: Gemäß unserer Herleitung des Eulerverfahrens über Numerische Integration 10.1 verwenden wir nun den Stützpunkt t_{k+1} und erhalten*

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt \sim y(t_k) + hf(t_{k+1}, y(t_{k+1}))$$

oder durch Einsetzen unserer Approximationen

$$y_{k+1} = y_k + hf(t_{k+1}, y_{k+1})$$

2. *Implizite Trapezregel: Verwenden wir zur Approximation des Integrals die Trapezregel, so gilt*

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t))\, dt \sim y(t_k) + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y(t_{k+1})))$$

oder

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})).$$

Beispiel 10.14 Durchführung impliziter Verfahren
In der Durchführung wird die definierende Gleichung, wann immer möglich, aufgelöst. Wir schauen auf unser Standardbeispiel $y'(t) = \lambda y(t)$. In diesem Fall ergibt sich für das implizite Eulerverfahren

$$y_{k+1} = y_k + h\, f(t_{k+1}, y_{k+1}) = y_k + \lambda\, h\, y_{k+1} \Rightarrow y_{k+1} = \frac{1}{1 - \lambda\, h} y_k.$$

Für das Trapezverfahren erhält man entsprechend

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})) = y_k + \frac{\lambda\, h}{2}(y_k + y_{k+1})$$

und damit

$$y_{k+1} = \frac{1 + \lambda\, h/2}{1 - \lambda\, h/2}\, y_k.$$

Wir werden sehen, dass implizite Verfahren nützlich sind. Es stellt sich aber die Frage, ob sie überhaupt wohldefiniert sind (d.h. ob die Gleichungen, die sie definieren, eindeutige Lösungen haben).

Satz 10.15 *(**Wohldefiniertheit für implizite Einschrittverfahren**)*
Sei $\varphi(t_k, y_k, y_{k+1})$ die Schrittfunktion eines impliziten Einschrittverfahrens zur Lösung von 10.1. Sei φ stetig, und lipschitzstetig bzgl. y_{k+1} mit der Lipschitzkonstanten L'. Dann gibt es ein h_0, so dass die Gleichung

$$y_{k+1} = y_k + h\varphi(t_k, y_k, y_{k+1})$$

für $h \leq h_0$ für alle t_k und y_k lokal (in einer kleinen Umgebung von y_k) eindeutig nach y_{k+1} auflösbar ist (d.h. das Verfahren ist überhaupt durchführbar).

Beweis: Wir zeigen, dass die rechte Seite bei der Definition der impliziten Verfahren eine Selbstabbildung und kontrahierend ist, dann folgt die Wohldefiniertheit aus dem Banachschen Fixpunktsatz.
K_M ist kompakt, φ ist stetig, also gilt

$$M' = \max_{(t,y),(t,\overline{y})\in K_M} \varphi(t, y, \overline{y}) < \infty.$$

Sei δ so klein, dass $(t_k, [y_k - \delta, y_k + \delta])$ noch ganz in K_M liegt. Sei nun h_0 so klein, dass

$$q := L'h_0 < \frac{1}{2} \text{ und } M'h_0 \leq \frac{\delta}{2},$$

Seien t_k und y_k fest. Wir setzen

$$g : [y_k - \delta, y_k + \delta] \to [y_k - \delta, y_k + \delta],\ g(z) := y_k + h\varphi(t_k, y_k, z).$$

Der Grundraum $\mathbb{R}^n$ ist Banachraum, die Teilmenge, auf der g definiert ist, ist abgeschlossen. Zu zeigen für den Fixpunktsatz von Banach ist noch: g ist wohldefiniert (Selbstabbildung) und kontrahierend. Sei $z \in [y_k - \delta, y_k + \delta]$.

$$\begin{aligned} ||g(z) - y_k|| &= ||h\varphi(t_k, y_k, z)|| \\ &\leq M'h_0 \leq \frac{\delta}{2} \end{aligned}$$

und damit $g(z) \in [y_k - \delta, y_k + \delta]$. Weiter gilt für $z,\ z' \in [y_k - \delta, y_k + \delta]$ mit der Lipschitzkonstanten L'

$$\begin{aligned} ||g(z) - g(z')|| &= ||h(\varphi(t_k, y_k, z) - \varphi(t_k, y_k, z'))|| \\ &\leq h_0 L'||z - z'|| \leq q||z - z'||. \end{aligned}$$

Also ist g auch kontrahierend und besitzt mit dem Banachschen Fixpunktsatz 3.3 einen eindeutigen Fixpunkt

$$y_{k+1} = v(t_k, y_k).$$

Die Fixpunktiteration für g konvergiert insbesondere für den Startwert y_k gegen y_{k+1}.
□

Im Lichte dieses Satzes ist es jetzt auch klar, was wir tun, falls die entstehenden impliziten Gleichungen nicht auflösbar sind: Wir nutzen einfach eine Fixpunktfolge mit Startwert y_k.

Nach Auflösung der definierenden Gleichung sind die impliziten einfach nur spezielle explizite Verfahren, d.h. der Konvergenzsatz 10.11 gilt, aber Konsistenz ist schwieriger zu zeigen. Für das implizite Eulerverfahren (Ordnung 1) und die implizite Trapezregel (Ordnung 2) folgt die Konsistenz aus dem allgemeinen Satz über die Konsistenzordnung der Runge–Kutta–Verfahren 10.19, siehe 10.21.

10.4 Runge–Kutta–Verfahren

Wir wollen die oben hergeleiteten Verfahren nun noch verallgemeinern, um Verfahren beliebiger Ordnung konstruieren zu können, die nur Auswertungen von f benötigen. Wir betrachten zunächst noch einmal das Verfahren von Heun. Die Schrittfunktion φ war hier definiert durch

$$\varphi(t_k, y_k) = \frac{1}{2}(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k))).$$

Dies schreiben wir in der Form:

$$\begin{array}{rcl} f_1 & = & f(t_k, y_k) \\ f_2 & = & f(t_k + h, y_k + hf_1) \\ \hline \varphi(t_k, y_k) & = & \frac{1}{2}f_1 + \frac{1}{2}f_2 \end{array}$$

Diese Schreibweise legt die folgende Definition nahe.

Definition 10.16 (Definition der Runge–Kutta–Verfahren)
Seien α_j, γ_j, β_{jl} fest gewählt, $j, l = 1 \ldots m$. Die Schrittfunktion φ sei definiert durch

$$\varphi(t_k, y_k) = \gamma_1 f_1 + \gamma_2 f_2 + \ldots + \gamma_m f_m = \sum_{j=1}^{m} \gamma_j f_j$$

mit

$$f_j = f(t_k + \alpha_j h, y_k + h \sum_{l=1}^{m} \beta_{jl} f_l),\ j = 1 \ldots m.$$

Dann heißt das zugehörige numerische Verfahren m–stufiges Runge–Kutta–Verfahren. Falls

$$\beta_{jl} = 0 \text{ für } l \geq j,$$

so íst das Verfahren explizit, ansonsten implizit.

Beim Nachweis der Konvergenz eines Runge–Kutta–Verfahrens haben wir zwei Probleme: Zunächst haben wir uns immer auf skalare Differentialgleichungen zurückgezogen, es ist also unklar, ob unsere Verfahren auch für Systeme funktionieren. Aber selbst in diesem Fall ist eine Schwierigkeit die mehrdimensionale Entwicklung der Funktion f, das wird sehr schnell sehr aufwändig und unübersichtlich. Der folgende Satz ist deshalb **sehr** nützlich.

Satz 10.17 *Gegeben sei ein Runge–Kutta–Verfahren wie in 10.16. Dann gilt:*

1. *Jede Differentialgleichung lässt sich in ein autonomes System von Differentialgleichungen umwandeln.*
2. *Die Normierungsbedingungen*

$$\sum_{j=1}^{m} \gamma_j = 1, \sum_{l} \beta_{jl} = \alpha_j \forall j = 1 \dots m$$

seien für ein Runge–Kutta–Verfahren erfüllt.
Dann gilt: Falls das Verfahren für skalare Anfangswertaufgaben konvergent von der Ordnung p ist, so ist es auch für Systeme von Differentialgleichungen konvergent von der Ordnung p.

Beweis: Zu 1. in den Übungen. Zu 2.: Hier ohne Beweis, siehe z.B. **??**. □

Korollar 10.18 *Beim Nachweis der Konvergenzordnung für ein Runge–Kutta–Verfahren, das die Normierungsbedingungen erfüllt, kann man sich auf skalare autonome Differentialgeichungen zurückziehen.*

Satz 10.19 (Ordnung der Runge–Kutta–Verfahren)

1. *Sei $f \in C^1$ und*

$$\sum_{j=1}^{m} \gamma_j = 1, \sum_{l} \beta_{jl} = \alpha_j \forall j = 1 \dots m.$$

Dann ist das zugehörige Runge–Kutta–Verfahren mindestens konvergent von der Ordnung 1.

2. *Sei sogar $f \in C^2$ und zusätzlich*

$$\sum_{j=1}^{n} \alpha_j \gamma_j = \frac{1}{2}.$$

Dann ist das zugehörige Runge–Kutta–Verfahren mindestens konvergent von der Ordnung 2.

Beweis: Für Lipschitz–stetiges f ist auch φ Lipschitz–stetig für alle Runge–Kutta–Verfahren. Zum Beweis der Konsistenz nutzen wir 10.18. Sei also f skalar und $f = f(y)$. Dann gilt

$$y'(t) = f(y(t)) \Rightarrow y''(t) = f'(y(t))\, y'(t) = f'(y(t))\, f(y(t)).$$

Sei also wieder y irgendeine Lösung der Differentialgleichung. Für den Konsistenzfehler setzen wir wieder die Lösung y in die Diskretisierung ein, also

$$\tau_h(t, y(t)) = \frac{y(t+h) - y(t)}{h} - \varphi(t, y(t)).$$

Zunächst berechnen wir die Taylorentwicklung der Zwischenstufen f_j.

$$\begin{aligned}
f_j &= f(y(t) + h \sum_{l=1}^{m} \beta_{jl} f_l) \\
&= f(y(t)) + h\, f'(y(t)) \left(\sum_{l=1}^{m} \beta_{jl} \underbrace{f_l}_{f(y(t))+O(h)} \right) + O(h^2) \\
&= y'(t) + h\alpha_j (f'(y(t))\, f(y(t))) + O(h^2) \\
&= y'(t) + h\alpha_j y''(t) + O(h^2).
\end{aligned}$$

Eingesetzt in die Definition des lokalen Diskretisierungsfehlers

$$\begin{aligned}
\tau_h(t, y(t)) &= \frac{1}{h}(y(t+h) - y(t)) - \varphi(t, y(t)) \\
&= y'(t) + \frac{h}{2} y''(t) - \sum_{j=1}^{m} \gamma_j f_j + O(h^2) \\
&= y'(t) + \frac{h}{2} y''(t) - \sum_{j=1}^{m} \gamma_j (y'(t) + h\alpha_j y''(t)) + O(h^2) \\
&= (1 - \sum_{j=1}^{m} \gamma_j) y'(t) + h(\frac{1}{2} - \sum_{j=1}^{m} \gamma_j \alpha_j) y''(t) + O(h^2).
\end{aligned}$$

□

Beispiel 10.20 *Wir weisen noch einmal, mit Hilfe des Korollars 10.18, die Konsistenzordnung des Verfahrens von Heun nach, um zu zeigen, dass die Rechnung viel einfacher wird. Es sei also $y'(t) = f(y(t))$ und $f \in C^2$. Wie in 10.10 gilt für die Verfahrensfunktion*

$$\varphi(y) = \frac{1}{2}(f(y) + f(y + h\, f(y)))$$

$$\begin{aligned} \tau_h(t, y(t)) &= \underbrace{\frac{y(t+h) - y(t)}{h}}_{y'(t)+\frac{h}{2}y''(t)+O(h^2)} - \frac{1}{2}(f(y(t)) + \underbrace{f(y(t) + h\, f(y(t)))}_{f(y(t))+h\, f(y(t))\, f'(y(t))+O(h^2)}) \\ &= O(h^2). \end{aligned}$$

Korollar 10.21 *(Konsistenordnung impliziter Euler und implizite Trapezregel)*

1. *Für den impliziten Euler gilt*

$$\begin{aligned} y_{k+1} &= y_k + h\, f_1 \\ f_1 &= f(t_k + h, \underbrace{y_k + h\, f_1}_{=y_{k+1}}). \end{aligned}$$

   Dies ist ein einstufiges Runge–Kutta–Verfahren mit

$$\gamma_1 = 1,\ \alpha_1 = 1,\ \beta_{1,1} = 1.$$

   Die Normierungsbedingung ist erfüllt, das Verfahren ist von der Ordnung 1.

2. *Für die implizite Trapezregel gilt*

$$\begin{aligned} y_{k+1} &= y_k + \frac{h}{2}(f_1 + f_2) \\ f_1 &= f(t_k, y_k) \\ f_2 &= f(t_k + h, \underbrace{y_k + \frac{h}{2}(f_1 + f_2)}_{=y_{k+1}}). \end{aligned}$$

   Dies ist ein zweistufiges Runge–Kutta–Verfahren mit

$$\gamma_1 = \gamma_2 = \frac{1}{2},\ \alpha_1 = 0,\ \beta_{1,1} = \beta_{1,2} = 0,\ \alpha_2 = 1,\ \beta_{2,1} = \beta_{2,2} = \frac{1}{2}.$$

   Die Normierungsbedingungen sind erfüllt, und zusätzlich gilt

$$\alpha_1\gamma_1 + \alpha_2\gamma_2 = \frac{1}{2},$$

   also ist das Verfahren (mindestens) von der Ordnung 2.

10.5 Energieerhaltung

Durch die Runge–Kutta–Verfahren haben wir jetzt einen ganzen Strauß unterschiedlicher Verfahren mit unterschiedlichen Konsistenzordnungen. Die Frage ist: Gibt es die eine beste? Wir wollen uns in diesem Abschnitt klarmachen, dass die Verfahren unterschiedliche Eigenschaften haben, auch jenseits der Konsistenzordnung.

Unsere Verfahren berechnen diskrete Approximationen an eine Anwendungsaufgabe. Wir erwarten natürlich, dass unsere Approximationen die wesentlichen Eigenschaften der Lösungen der Aufgabe reproduzieren. Leider ist dies nicht immer der Fall.

Wir kehren nochmal zurück zum Steinwurfbeispiel 1.1. Die zugehörige Anfangswertaufgabe für die vertikale Geschwindigkeit $V(t)$ und Höhe $H(t)$ eines Steins mit Masse m war

$$H'(t) = V(t),\ V'(t) = -g,\ H(0) = 0,\ V(0) = V_0.$$

Für die potentielle und kinetische Energie gilt

$$E_{pot}(t) = m\,g\,H(t),\ E_{kin} = m\frac{V(t)^2}{2}.$$

Für die Gesamtenergie $E(t) = E_{pot}(t) + E_{kin}(t)$ gilt

$$E'(t) = m\,g\,H'(t) + m\frac{2V(t)V'(t)}{2} = m(g\,V(t) - g\,V(t)) = 0$$

und das heißt, dass die Gesamtenergie konstant ist über die Zeit (Energieerhaltung).

Natürlich erwarten wir das auch für unsere Näherungen. Wir wählen das Eulerverfahren auf einem äquidistanten Gitter auf dem Interval $[0, 1]$ mit der Schrittweite $h = 1/N$. Für unsere Differentialgleichung ist

$$f(t, \begin{pmatrix} H \\ V \end{pmatrix}) = \begin{pmatrix} V \\ -g \end{pmatrix}.$$

Dann gilt

$$\begin{pmatrix} H_{k+1} \\ V_{k+1} \end{pmatrix} = \begin{pmatrix} H_k \\ V_k \end{pmatrix} + hf(t_k, \begin{pmatrix} H_k \\ V_k \end{pmatrix}) = \begin{pmatrix} H_k + hV_k \\ V_k - hg \end{pmatrix}.$$

Die Energie für unsere Approximation zum Zeitpunkt t_{k+1} ist also

$$\begin{aligned} E_h(t_{k+1}) &= m\,g\,H_{k+1} + m\frac{V_{k+1}^2}{2} \\ &= m\,g(H_k + hV_k) + m\frac{(V_k - h\,g)^2}{2} \\ &= m\,gH_k + m\frac{V_k^2}{2} + m\frac{h^2g^2}{2} \\ &= E_h(t_k) + \underbrace{m\frac{h^2g^2}{2}}_{>0}. \end{aligned}$$

Die Energie nimmt in unserer Approximation also bei jedem Zeitschritt zu.

Entsprechend gilt für das implizite Eulerverfahren

$$\begin{pmatrix} H_{k+1} \\ V_{k+1} \end{pmatrix} = \begin{pmatrix} H_k \\ V_k \end{pmatrix} + hf(t_{k+1}, \begin{pmatrix} H_{k+1} \\ V_{k+1} \end{pmatrix}) = \begin{pmatrix} H_k + h(V_k - hg) \\ V_k - hg \end{pmatrix}.$$

und für die Energie

$$E_h(t_{k+1}) = m\,g(H_k + h(V_k - hg)) + m\,\frac{(V_k - hg)^2}{2} = E_h(t_k) - m\frac{h^2g^2}{2}$$

Die Energie nimm also bei jedem Zeitschritt zu.

Dies ist jetzt noch nicht überraschend, denn alle unsere Werte sind ja eben nur Näherungen. Eine wesentliche physikalische Grundlage ist damit aber verletzt – und wir müssen damit rechnen, dass unsere Näherung dann auch physikalischen Unfug liefert. Dies passiert, wenn wir versuchen, eine der Diskretisierungen auf das Federbeispiel 1.3 anzuwenden (siehe Rechnerbeispiel). Hier kommt bei der Berechnung der Energie noch die in der Feder gespeicherte Energie hinzu, aber ansonsten bleibt unsere Betrachtung gleich: Beim Eulerverfahren nimmt die Energie zu, d.h. die Feder schwingt immer stärker, beim impliziten Eulerverfahren nimmt die Energie ab, d.h. die Schwingung wird immer kleiner. Das Langzeitverhalten der Schwingung wird also falsch vorhergesagt.

Hier sieht man nun auch die Nützlichkeit der Phasenportraits: Wir tragen die Trajektorien der drei Lösungen (analytisch, Euler, impliziter Euler) in einer Zeichnung auf und sehen direkt die Unterschiede im Langzeitverhalten.

Beide Verfahren sind also ungeeignet, denn sie erhalten die Energie nicht. Die implizite Trapezregel dagegen erhält die Energie und ist deshalb hier das richtige Verfahren. Es gilt

$$\begin{pmatrix} H_{k+1} \\ V_{k+1} \end{pmatrix} = \begin{pmatrix} H_k \\ V_k \end{pmatrix} + \frac{h}{2}(f(t_k, \begin{pmatrix} H_k \\ V_k \end{pmatrix}) + f(t_{k+1}, \begin{pmatrix} H_{k+1} \\ V_{k+1} \end{pmatrix})) = \begin{pmatrix} H_k + h(V_k - \frac{hg}{2}) \\ V_k - hg \end{pmatrix}.$$

und für die Energie

$$E_h(t_{k+1}) = m\,g(H_k + h(V_k - \frac{h\,g}{2})) + m\,\frac{(V_k - hg)^2}{2} = E_h(t_k).$$

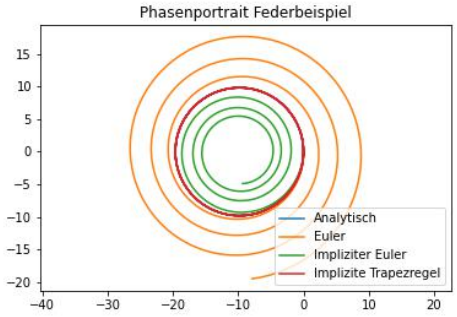


Abbildung 10.2: Trajektorien für das Federbeispiel aus 1.3

10.6 Fehlerabschätzung und Schrittweitensteuerung

Mit dem Konvergenzsatz 10.11 kann man den globalen Diskretisierungsfehler abschätzen durch die Summe der Konsistenzfehler. Es gilt

$$e_k \leq (b-a)||\tau_h||_\infty e^{L'(b-a)}.$$

Unsere Idee ist jetzt: Wir wählen kein äquidistantes Gitter mehr, sondern bestimmen in jedem Schritt h so, dass der lokale Diskretisierungsfehler unter einer Schranke bleibt.

Wir schätzen diese Fehler wie folgt:

In jedem Schritt des Verfahrens werden zwei Runge–Kutta–Verfahren mit unterschiedlicher Konsistenzordnung benutzt. Hierbei sei Verfahren 1 mit Verfahrensfunktion $\varphi^{(1)}$ das genauere (höherer Konsistenzordnung). Wir berechnen

$$y_{k+1}^{(1)} = y_k + h\,\varphi^{(1)}(t_k, y_k)$$
$$y_{k+1}^{(2)} = y_k + h, \varphi^{(2)}(t_k, y_k)$$

Da das erste Verfahren eine höhere Konsistenzordnung hat, gilt für die Lösung y der Differentialgleichung mit $y(t_k) = y_k$

$$|y_{k+1}^{(2)} - y(t_{k+1})| \leq |y_{k+1}^{(2)} - y_{k+1}^{(1)}| + |y_{k+1}^{(1)} - y(t_{k+1})| \sim |y_{k+1}^{(2)} - y_{k+1}^{(1)}|.$$

Wir können also eine Approximation des lokalen Diskretisierungsfehlers für das zweite Verfahren nur mit Hilfe der berechneten Werte angeben.

Ein einfaches Beispiel erhalten wir mit der Kombination Euler und Heun: Wir berechnen unsere Approximationen mit Heun (Ordnung 2) und schätzen die Fehler durch den Unterschied zum Eulerverfahren (Ordnung 1). Hierbei entsteht keine zusätzlicher Aufwand, denn Euler benötigt nur die Auswertung an der Stelle $f(t_k, y_k)$, und die hat man bei Heun sowieso ausgerechnet.

In der Praxis genutzt werden eher Verfahren höherer Ordnung, etwa Dormand–Prince, Referenz 1.

10.7 A–Stabilität für Einschrittverfahren

Zu guten numerischen Verfahren zur Lösung von gewöhnlichen Differentialgleichungen gehört noch etwas mehr als nur Stabilität und Konvergenz.

Wir versuchen, die Verfahren so zu wählen, dass möglichst viele Eigenschaften der Lösungen der Differentialgleichung für die Lösungen der Differenzengleichungen erhalten bleiben. Wir betrachten dazu

Definition 10.22 (Modellproblem von Dahlquist)

$$y'(t) = \lambda y(t),\ y(0) = 1,\ \lambda \in \mathbb{C},\ Re\lambda < 0.$$

Die analytische Lösung ist $y(t) = \exp(\lambda t)$. Offensichtlich ist der Betrag der Lösung monoton fallend, und das erwarten wir auch von der numerischen Approximation, unabhängig von h.

Definition 10.23 (A-Stabilität)
Ein Verfahren heißt A–stabil, wenn für das Dahlquist–Problem und äquidistante Gitter mit Schrittweite h die numerische Approximation im Betrag monoton fallend ist für alle h und alle λ mit Realteil kleiner 0, dass also gilt

$$|y_h(t_{k+1})| \leq |y_h(t_k)| \,\forall k.$$

Für Runge–Kutta–Verfahren lässt sich die A–Stabilität einfach überprüfen.

Satz 10.24 (Stabilitätsfunktion und Stabilitätsbereich von Runge–Kutta–Verfahren)
Für die durch Runge–Kutta–Verfahren gelieferten Näherungen für die Lösung des Dahlquist–Problems gilt

$$y_h(t_{k+1}) = y_{k+1} = R(\lambda h) y_k = R(\lambda h) y_h(t_k).$$

Für explizite Verfahren ist R ein Polynom, für implizite Verfahren eine rationale Funktion.

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

heißt Stabilitätsbereich. Es gilt

$$|y_h(t_{k+1})| \leq |y_h(t_k)|$$

genau dann, wenn $\lambda h \in S$.
Das Verfahren ist A–stabil genau dann, wenn der Stabilitätsbereich die komplette linke komplexe Halbebene umfasst.

Beweis: Übungen.

Beispiel 10.25 (Beispiele zur Stabilitätsfunktion)

1. *Euler explizit:*

$$y_{k+1} = y_k + hf(t_k, y_k) = y_k + h\lambda y_k = (1 + h\lambda) y_k.$$

   Es ist also $R(z) = 1 + z = z - (-1)$. Das Verfahren ist nicht A–stabil. Das Stabilitätsgebiet ist ein Kreis um -1 mit Radius 1.

2. *Euler implizit:*

$$y_{k+1} = y_k + hf(t_k, y_{k+1}) = y_k + h\lambda y_{k+1}$$

   und damit

$$y_{k+1} = \frac{1}{1 - h\lambda} y_k.$$

   Die Stabilitätsfunktion ist

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

   Es gilt $|R(z)| \leq 1$ außerhalb eines Kreises mit Radius 1 um die 1. Die linke komplexe Halbebene ist ganz im Stabilitätsbereich enthalten. Das Verfahren ist A–stabil.

3. *implizite Trapzeregel: Siehe 10.14 für die rationale Funktion R.*

Satz 10.26 (A–Stabilität expliziter Runge–Kutta–Verfahren)
Kein explizites, konsistentes Runge–Kutta–Verfahren ist A–stabil.

Beweis: Für explizite Verfahren ist R ein Polynom. Da das Verfahren konsistent ist, muss der Grad n von $R > 0$ sein. Sei

$$R(x) = a_n x^n + \sum_{k=0}^{n-1} a_k x^k = x^n \underbrace{\left(a_n + \sum_{k=0}^{n-1} a_k x^{k-n} \right)}_{\to a_n, |x| \to \infty}, \; a_n \neq 0.$$

Insbesondere gilt $|R(x)| \to \infty$ für $x \to -\infty$, also ist $|R(z)|$ größer als 1 für ein z in der linken Halbebene, und das Verfahren ist nicht A–stabil. □

Warum ist dies nun problematisch? Hierzu schauen wir auf eine lineare Differentialgleichung $y'(t) = A\,y(t)$ im $\mathbb{R}^2$. A habe die Eigenwerte $\lambda_1 << \lambda_2 < 0$. Die allgemeine Lösung dieser Differentialgleichung ist gegeben durch

$$y(t) = \alpha e^{\lambda_1 t} x_1 + \beta e^{\lambda_2 t} x_2.$$

Da $\lambda_1 << \lambda_2 < 0$, geht $e^{\lambda_1 t}$ sehr viel schneller gegen 0 als $e^{\lambda_2 t}$. Der erste Summand spielt also eigentlich gar keine Rolle.

Aber: Wenn ein Verfahren nicht insgesamt A–stabil ist, müssen wir h klein wählen, damit $\lambda_1 h \in S$. Es entscheidet also über die Wahl von h der Eigenwert, der eigentlich in der Lösung überhaupt keine Rolle spielt. Dies kann dazu führen, dass wir h sehr viel kleiner wählen müssen als in A–stabilen Verfahren.

Gleichungen mit dieser Eigenschaft heißen steife Differentialgleichungen.

10.8 Zusammenfassung

10.8.1 Kompetenzen

- Definition numerischer Verfahren kennen, insbesondere die Begriffe Gitter, globaler Diskretisierungsfehler, Konvergenz, Ein– und Mehrschrittverfahren.
- Wichtig: Interpretation der Approximation kennen: Die Funktion y_h approximiert die Lösung y an den Punkten des Gitters I_h.
- Anhand einer vorgegebenen Differentialgleichung und eines vorgegebenen Verfahrens die Approximation ausrechnen können.

- Definitionen von Konsistenz und Konvergenz kennen. Abschätzung des Konsistenzfehlers mit Taylor ausrechnen können.
- Definition der Runge–Kutta–Verfahren und die Konsistenzsätze für Runge–Kutta kennen.
- Konvergenzsatz kennen: Aus Konsistenz folgt Konvergenz.
- Definition der A–Stabilität und ihre Bedeutung für steife Differentialgleichungen kennen. A–Stabilität für konkrete Verfahren ausrechnen können.

10.8.2 Mini–Aufgaben

Siehe Übungen.

Kapitel 11

Lineare Mehrschrittverfahren

Bevor wir uns nun den Mehrschrittverfahren zuwenden, wiederholen wir noch einmal die zentralen Begriffe der Einschrittverfahren:

- Konsistenz: Das Verfahren ist Diskretisierung der Differentialgleichung. Der lokale Diskretisierungsfehler, bei dem man die echte Lösung y von 10.1 in die diskretisierte Gleichung einsetzt, geht für kleine Schrittweiten h gegen 0.
- Konvergenz: Das Verfahren liefert für eine Familie von Gittern, deren Feinheit gegen 0 geht, die exakte Lösung (dies ist das eigentliche Ziel).
- Stabilität: Kleine Fehler im einzelnen Schritt führen zu kleinen Gesamtfehlern, begründet den Satz: Aus Konsistenz folgt Konvergenz (für Einschrittverfahren), und ist eine Folgerung der Gronwallschen Ungleichung.

Die Idee bei den Mehrschrittverfahren ist, die vergangenen Funktionsauswertungen bei der Berechnung der nächsten Approximation mitzunehmen. Wir erhoffen uns dadurch eine deutliche Verringerung des notwendigen Aufwands oder eine deutliche Erhöhung der möglichen Konsistenzordnung. In der Vorgehensweise ähneln die Einschrittverfahren den vielleicht aus der Stochastik bekannten gedächtnislosen Markovprozessen: Der nächste Wert hängt ausdrücklich nur vom aktuellen Zustand ab, nicht von einer Historie. Im Gegensatz dazu stehen die Mehrschrittverfahren.

Mehrschrittverfahren haben hohe Konsistenzordnungen bei nur einer Evaluationen von f. Auf der anderen Seite sind sie nicht notwendig stabil, wie es die Einschrittverfahren sind. Eine Schrittweitensteuerung ist schwierig. Daher sind sie häufig nicht die Standardsolver. In Matlab steht der Adams–Bashforth–Solver als Standard–Mehrschrittverfahren unter dem Namen ode113 zur Verfügung.

Wir werden in diesem Kapitel zunächst einige Verfahren herleiten, die Bezeichnungen der Einschrittverfahren auf Mehrschrittverfahren übertragen und überprüfen, welche Sätze erhalten bleiben. Zunächst schränken wir die komplette Betrachtung auf lineare Verfahren auf äquidistanten Gittern ein.

11.1 Definition und Beispiele

Definition 11.1 (lineare Mehrschrittverfahren)
Ein numerisches Verfahren, das auf einem **äquidistanten** *Gitter I_h mit Schrittweite h die Näherung $y_h(t_k) = y_k$ der Differentialgleichung berechnet mit*

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = h \sum_{j=0}^{m} \beta_j f_{k+j}, f_{k+j} := f(t_{k+j}, y_{k+j}),\ k = 0 \dots N - m$$

$(\alpha_m \neq 0)$ heißt lineares m–Schritt–Mehrschrittverfahren oder lineares Mehrschrittverfahren der Stufe m. Das Verfahren heißt explizit, falls $\beta_m = 0$, ansonsten implizit.

Wir dürfen also bei einem m–Mehrschrittverfahren zur Berechnung von y_{k+m} die letzen m Funktionsauswertungen f_k bis f_{k+m-1} mitbenutzen. Nur f_{k+m} muss neu ausgerechnet werden.

Dies bedeutet natürlich, dass wir im ersten Schritt des Verfahrens das y_m berechnen und dafür y_0 bis y_{m-1} benötigen, wobei wir eigentlich nur y_0 haben. Die fehlenden Terme werden üblicherweise mit hochgenauen Einschrittverfahren berechnet, hat man alle zusammen, macht man ab hier mit den Mehrschrittverfahren weiter.

Beispiel 11.2 (Mittelpunktregel)
Mehrschrittverfahren lassen sich wie die Einschrittverfahren durch Numerische Integration herleiten (siehe 10.2). Wir approximieren das Integral mit einer Funktionsauswertung in der Mitte des Intervalls und erhalten

$$y(t_{k+2}) - y(t_k) = \int_{t_k}^{t_{k+2}} y'(t)\,dt = \int_{t_k}^{t_{k+2}} f(t, y(t))\,dt \sim 2\,h f(t_{k+1}, y(t_{k+1})).$$

Wir erhalten das numerische Verfahren

$$y_{k+2} - y_k = 2h\, f(t_{k+1}, y_{k+1}) = 2h\, f_{k+1}.$$

Dieses Beispiel lässt sich vielfach variieren.

Beispiel 11.3
Es gilt

$$y(t_{k+2}) - y(t_{k+1}) = \int_{t_{k+1}}^{t_{k+2}} f(t, y(t))\,dt.$$

Wir wollen zur Approximation des Integrals die Näherungen $f_k = f(t_k, y_k) \sim f(t_k, y(t_k))$ und $f_{k+1} = f(t_{k+1}, y_{k+1}) \sim f(t_{k+1}, y(t_{k+1}))$ nutzen. Sei p das Interpolationspolynom vom Grad 1 mit $p(t_k) = f_k$ und $p(t_{k+1}) = f_{k+1}$. Wir gehen zur Approximation über und erhalten

$$y_{k+1} - y_k = \int_{t_{k+1}}^{t_{k+2}} p(s)\, ds.$$

Das Interpolationspolynom rechnen wir explizit aus mit Lagrange:

$$p(s) = \frac{s - t_{k+1}}{h} f_k + \frac{t_k - s}{h} f_{k+1},$$

also

$$\int_{t_{k+1}}^{t_{k+2}} p(s)\, ds = h(\frac{3}{2} f_{k+1} - \frac{1}{2} f_k).$$

Insgesamt erhalten wir das Verfahren

$$y_{k+2} = y_{k+1} + h(\frac{3}{2} f_{k+1} - \frac{1}{2} f_k).$$

Dieses Verfahren ist explizit, denn y_{k+2} kommt auf der rechten Seite nicht vor.

Wir fassen dies nun allgemeiner.

Für die Lösung unserer Anfangswertaufgabe 10.1 gilt

$$y(t_{k+m}) - y(t_{k+m-r}) = \int_{t_k+m-r}^{t_{k+m}} f(t, y(t))\mathrm{dt}.$$

Für die Approximation ersetzen wir die Funktion unter dem Integralzeichen durch sein Interpolationspolynom p_k mit den Stützstellen $t_k + jh$ und den Stützwerten

$$f_{k+j} = f(t_{k+j}, y_{k+j}).$$

Wir erhalten damit z.B. für ein explizites Verfahren, das die Stützwerte f_k bis f_{k+m-1} benutzt

$$y_{k+m} - y_{k+m-r} \quad = \quad \int_{t_{k+m-r}}^{t_{k+m}} p_k(t)\mathrm{dt}$$

Für das Interpolationspolynom haben wir dann noch die Wahl zwischen den Interpolationsstellen t_k bis t_{k+m} (implizit) und t_k bis t_{k+m-1} (explizit). Für $r = 1$ erhalten

wir die Verfahren von Adams–Bashforth und Adams–Moulton, für $r = 2$ die Verfahren von Nyström und Milne–Simpson. Wir fassen das Ergebnis in folgender Tabelle zusammen:

Interpolation benutzt	$r = 1$	$r = 2$	
$f_k \dots f_{k+m-1}$	Adams–Bashforth	Nyström	explizit
$f_k \dots f_{k+m}$	Adams–Moulton	Milne–Simpson	implizit

11.2 Konsistenz von Mehrschrittverfahren

Der Konsistenzfehler für Mehrschrittverfahren wird wie bei den Einschrittverfahren 10.7 definiert.

Definition 11.4 *Gegeben sei ein durch die Konstanten α_j und β_j beschriebenes lineares Mehrschrittverfahren. Sei y irgendeine Lösung der Differentialgleichung. Dann ist der lokale Diskretisierungsfehler gegeben durch*

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}\sum_{j=0}^{m} \alpha_j y(t + jh) - \sum_{j=0}^{m} \beta_j f(t + jh, y(t + jh)) \\ &= \frac{1}{h}\sum_{j=0}^{m} \alpha_j y(t + jh) - \sum_{j=0}^{m} \beta_j y'(t + jh).\end{aligned}$$

Wieder bekommen wir den lokalen Diskretisierungsfehler, indem wir die Lösungen y der Differentialgleichung in die diskrete Gleichung einsetzen.

Korollar 11.5 (Konsistenzordnung der Mittelpunktregel)
Sei $f \in C^2$. Dann hat die Mittelpunktregel die Konsistenzordnung 2.

Beweis: Wir benutzen für

$$y(t + 2h) - y(t) = y((t + h) + h) - y((t + h) - h)$$

jeweils eine Taylorentwicklung um $y(t+h)$. Dann fallen alle Terme in h^j mit geradem j weg, und es gilt

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}(y((t + h) + h) - y((t + h) - h)) - 2f(t + h, y(t + h)) \\ &= 2y'(t + h) + O(h^2) - 2y'(t + h)) = O(h^2).\end{aligned}$$

□

Satz 11.6 *(Konsistenzordnung der durch Integration hergeleiteten MSV)*
Ein numerisches Verfahren mit m Schritten sei durch Integration des Interpolationspolynoms an m (explizit) bzw. $(m+1)$ (implizit) Stützstellen hergeleitet worden. Dann hat es mindestens die Konsistenzordnung m (explizit) bzw. $(m+1)$ (implizit).

Beweis: Für explizite Verfahren: Sei y eine Lösung der Differentialgleichung. Weiter sei p das Integrationspolynom, das an den Stellen $t+jh$ den Wert $f(t+jh, y(t+jh))$ annimmt, $j = 0 \dots m-1$. Dann gilt nach Konstruktion der Integrationsformeln

$$\begin{aligned}
\tau_h(t, y(t)) &= \frac{1}{h}(y(t+mh) - y(t+(m-r)h)) - \sum_{j=0}^{m-1} \beta_j f(t+jh, y(t+jh)) \\
&= \frac{1}{h} \int_{t+(m-r)h}^{t+mh} y'(t)\mathrm{dt} - \frac{1}{h} \int_{t+(m-r)h}^{t+mh} p(t)\mathrm{dt} \\
&= \frac{1}{h} \int_{t+(m-r)h}^{t+mh} f(t, y(t)) - p(t)\mathrm{dt} \\
&= O(h^m)
\end{aligned}$$

nach 9.3. □

Eine schöne Übersicht über all diese Verfahren mit Rechnungen und Beispielen findet sich (mit leicht anderen Bezeichnungen) in **?**, Kapitel III.1.

Bemerkung: Dies ist ein phantastisches Ergebnis: Mit nur einer zusätzlichen Auswertung von f können beliebig hohe Konsistenzordnungen erreicht werden!

Wir erwarten nun den Satz: Aus Konsistenz folgt Konvergenz. Leider ist für Mehrschrittverfahren die Situation komplexer.

Beispiel 11.7 *Es werde ein Verfahren möglichst hoher Ordnung der Form*

$$y_{k+2} - (1+\alpha)y_{k+1} + \alpha y_k = h\left(\frac{3-\alpha}{2} f_{k+1} - \frac{1+\alpha}{2} f_k\right)$$

gesucht. Sei $f \in C^3$. Wir bestimmen α durch Taylorentwicklung:

$$
\begin{aligned}
\tau_h(t) &= \frac{1}{h}(y(t+2h) - (1+\alpha)y(t+h) + \alpha y(t)) - \left(\frac{3-\alpha}{2}y'(t+h) - \frac{1+\alpha}{2}y'(t)\right) \\
&= y'(t)\underbrace{\left(2 - (1+\alpha) - \frac{3-\alpha}{2} + \frac{1+\alpha}{2}\right)}_{0} \\
&\quad + y''(t)\underbrace{\left(\frac{4h}{2} - (1+\alpha)\frac{h}{2} - \frac{3-\alpha}{2}h\right)}_{0} \\
&\quad + y'''(t)\underbrace{\left(\frac{8}{6}h^2 - (1+\alpha)\frac{h^2}{6} - \frac{3-\alpha}{2}\frac{h^2}{2}\right)}_{\frac{h^2}{12}(5+\alpha)} \\
&\quad + O(h^3)
\end{aligned}
$$

Also insgesamt eine Konsistenzordnung 3 für $\alpha = -5$ und 2 sonst. Nach unseren Erfahrungen mit den Einschrittverfahren erwarten wir eine entsprechende Konvergenzordnung.

Leider zeigt das numerische Experiment: Das geht gewaltig schief. Für $\alpha > 1$ und $\alpha < -1.5$ ist der Algorithmus nicht einmal konvergent.

11.3 Stabilität und Konvergenz von Mehrschrittverfahren

Aufgrund des letzten Beispiels vermuten wir bereits, dass der Satz "Aus Konsistenz folgt Konvergenz" für die Mehrschrittverfahren nicht ohne Weiteres korrekt ist.

Wir überlegen noch einmal, warum dies für die Einschrittverfahren galt. Das diskrete Lemma von Gronwall garantierte uns, dass die kleinen Einzelfehler, die wir in jedem Schritt des Verfahrens machen, nicht katastrophal verstärkt werden und damit möglicherweise die Konvergenz verhindern.

Für die Mehrschrittverfahren brauchen wir einen entsprechenden Satz. Dies erreicht man über die analytische Betrachtung von Differenzengleichungen. Zur Vereinfachung nutzen wir einen Satz ohne Beweis, Sie finden ihn z.B. in meinem Skript zur Vorlesung Numerische Analysis.

Die Idee: Wenn ein konsistentes Verfahren konvergent ist, dann muss es insbesondere konvergent sein für die einfachste aller Anfangswertaufgaben, das Modellpro-

blem

$$y'(t) = 0,\ y(0) = 0,$$

mit der Lösung $y = 0$. Die durch das numerische Verfahren gelieferte Lösung muss also für $h \to 0$ gleichmäßig gegen 0 konvergieren. Der Satz sagt: Dies reicht bereits aus, damit das Verfahren für alle Anfangswertaufgaben konvergent ist.

Satz 11.8 *(Reduktion auf das Modellproblem)*
Gegeben sei ein lineares Mehrschrittverfahren der Stufe m. Sei I_h eine Folge von äquidistanten Gittern mit Feinheit h. Falls für jede Wahl der Anlaufwerte $(y_h)_j$ mit $(y_h)_j \to_{h\to 0} 0$, $j = 0, \dots, m-1$, gilt: Die Folge (y_h) der Näherungen für das Modellproblem konvergiert gegen 0, so ist das Mehrschrittverfahren stabil für alle Anfangswertaufgaben.
Für stabile Verfahren folgt aus Konsistenz des Mehrschrittverfahrens (der Ordnung p) Konvergenz (der Ordnung p).

Der Satz sagt zweierlei: Erstens, wir können uns auf die einfachste aller Anfangswertaufgaben (das Modellproblem) beschränken. Zweitens, für die Betrachtung der Fehler reicht es, sich die Auswirkung der Fehler am Anfang anzuschauen.

Wir schauen nun, wann die durch das Mehrschrittverfahren für das Modellproblem gelieferten Näherungen gegen Null konvergieren. Nach Definition des Verfahrens gilt

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = 0,\ \alpha_m \neq 0.$$

Insbesondere hängt die Folge nicht vom Gitter I_h ab. Eine Gleichung dieser Form nennen wir eine (homogene) Differenzengleichung, eine Folge mit dieser Eigenschaft eine Lösung der Differenzengleichung.

Als Beispiel betrachten wir die Fibonaccifolge, die Sie schon einmal beim Sekantenverfahren 4.10 in den Übungen analysiert haben. Sie ist definiert durch

$$-y_k - y_{k+1} + y_{k+2} = 0$$

(ohne Einschränkung schreiben wir die Gleichungen immer mit $\alpha_m = 1$, die Differenzengleichung kann man immer mit einer Konstanten multiplizieren, und die Lösungen bleiben dieselben).

Eine Lösung der Differenzengleichung ist durch die Anlaufwerte y_0 und y_1 festgelegt. Sei $y^{(0)}$ die Lösung mit den Anlaufwerten $(1, 0)$, $y^{(1)}$ die Lösung mit den Anlaufwerten $(0, 1)$. Die Lösungen sind unabhängig von h usw. Die Lösungen bilden einen

Unterraum, daher ist jede Linearkombination von Lösungen automatisch auch wieder eine Lösung. Der Unterraum hat die Dimension $m = 2$.

Sei nun y_h die Lösung für die Anlaufwerte $((y_h)_0, (y_h)_1)$. Offensichtlich ist

$$(y_h)_0 y^{(0)} + (y_h)_1 y^{(1)}$$

eine Lösung der Differenzengleichung mit denselben Anlaufwerten wie y_h, also gilt

$$y_h = (y_h)_0 y^{(0)} + (y_h)_1 y^{(1)}.$$

Es gelte nun, dass $(y_h)_j \to_{h\to 0} 0$, $j = 0 \dots m-1$. Falls $y^{(j)}$ beschränkt ist, so gilt

$$||y_h||_\infty \leq |(y_h)_0| \, ||y^{(0)}||_\infty + |(y_h)_1| \, ||y^{(1)}||_\infty \to 0.$$

Korollar 11.9 *Der globale Diskretisierungsfehler (für das Modellproblem, und damit für alle Anfangswertaufgaben, bei konsistenten Anlaufwerten) geht mit der Gitterfeinheit gegen Null, wenn die Lösungen der Differenzengleichung*

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = 0$$

beschränkt sind für alle Anlaufwerte.

Wir müssen also untersuchen: Wann sind die Lösungen einer homogenen Differenzengleichung beschränkt? Wir geben zunächst eine alternative, nicht–rekursive Basis für die Lösungen der homogenen Differenzengleichung an.

Satz 11.10 *Es sei*

$$\rho(x) = \sum_{j=0}^{m} \alpha_j x^j$$

das charakteristische Polynom der Differenzengleichung

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = 0.$$

Seien x_l die (komplexen) Nullstellen von ρ mit Vielfachheiten σ_l. Dann ist eine Basis für den Unterraum U der Lösungen der Differenzengleichung im Raum aller Folgen gegeben durch

$$(y^{(l,r)})_j = j^r \, x_l^j, \; r = 0 \dots \sigma_l - 1.$$

Bemerkung: Wenn Sie sich diesen Satz genauer anschauen, sehen Sie eine direkte Beziehung der Lösungen der diskreten Differenzengleichungen und den Lösungen der linearen Differentialgleichungen 6.26.

Beweis: Die angegebenen Folgen sind linear unabhängig. Die Anzahl der Folgen ist $\sum_l \sigma_l = m$. Wenn wir zeigen können, dass die Folgen Lösungen der Differenzengleichung sind, sind wir fertig.
Zunächst gilt

$$\begin{aligned} 0 &= \rho(x_l) \\ &= x_l^k \rho(x_l) \\ &= \sum_{j=0}^{m} \alpha_j x_l^{k+j} \\ &= \sum_{j=0}^{m} \alpha_j (y^{(l,0)})_{k+j} \end{aligned}$$

und damit ist $y^{(l,0)}$ Lösung der Differenzengleichung.
Sei nun x_l eine doppelte Nullstelle von ρ, also $\sigma_l \geq 2$. Dann ist x_l auch eine Nullstelle von ρ', und es gilt

$$\begin{aligned} 0 &= x_l^{k+1} \rho'(x_l) + k x_l^k \rho(x_l) \\ &= x_l^{k+1} \sum_{j=0}^{m} j \alpha_j x_l^{j-1} + k\, x_l^k \sum_{j=0}^{m} \alpha_j x_l^j \\ &= \sum_{j=0}^{m} \alpha_j (k+j) x_l^{k+j} \\ &= \sum_{j=0}^{m} \alpha_j (y^{(l,1)})_{k+j}. \end{aligned}$$

Also ist für $\sigma_l \geq 2$ auch $y^{(l,1)}$ eine Lösung der Differenzengleichung, usw. □

Beispiel 11.11 *(Fibonacci)*
Wir betrachten wieder die Fibonaccifolge. Das charakteristische Polynom ist hier

$$\rho(x) = x^2 - x - 1$$

mit den Lösungen

$$x_{0,1} = \frac{1 \pm \sqrt{5}}{2}.$$

Diese Zahlen sind wohlbekannt aus dem goldenen Schnitt.

Also sind die Folgen

$$(y^{(0,0)})_k = \left(\frac{1+\sqrt{5}}{2}\right)^k, (y^{(1,0)})_k = \left(\frac{1-\sqrt{5}}{2}\right)^k$$

eine Basis des Raums aller Fibonaccifolgen. Die Standard–Fibonaccifolge y, die mit $(0,1)$ beginnt, lässt sich schreiben als

$$y_k = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k\right).$$

Wir gewinnen also eine nicht–rekursive Darstellung der Fibonacci–Zahlen.

Beispiel 11.12 *Wir betrachten die Differenzengleichung*

$$y_{k+2} - 2y_{k+1} + y_k = 0$$

(diese gehört zu einem Mehrschrittverfahren, das unabhängig von der Anfangswertaufgabe konsistent ist). Das charakteristische Polynom ist

$$\rho(x) = x^2 - 2x + 1.$$

$x = 1$ *ist die einzige (doppelte) Nullstelle.*

Die Basislösungen sind entsprechend gegeben durch

$$(y^{(0,0)})_k = 1^k = 1,\ (y^{(0,1)})_k = k\,1^k = k.$$

Korollar 11.13 *Die Basislösungen der Differenzengleichung sind beschränkt genau dann, wenn*

1. *Für alle Nullstellen x_l mit zugehöriger Vielfachheit σ_l des charakteristischen Polynoms gilt*
$$|x_l| \leq 1.$$
2. *Falls $|x_l| = 1$, so gilt $\sigma_l = 1$.*

Definition 11.14 *(Wurzelbedingung von Dahlquist)*
Eine Differenzengleichung heißt stabil, wenn ihre Basislösungen beschränkt sind, d.h. wenn die Bedingungen aus 11.13 erfüllt sind.

Mit den Vorbemerkungen:

Korollar 11.15 *Falls ein Mehrschrittverfahren konsistent ist (von der Ordnung p), und das zugehörige Differenzenverfahren die Wurzelbedingung von Dahlquist erfüllt, so ist das Mehrschrittverfahren konvergent (von der Ordnung p).*

Beispiel 11.16 *(Beispiele zur Stabilität von Mehrschrittverfahren)*

1. *Einschrittverfahren: Wir können Einschrittverfahren interpretieren als Mehrschrittverfahren mit $m = 1$. Sie sind von der Form*

   $$y_{k+1} - y_k = \varphi$$

   und damit

   $$\rho(x) = x - 1.$$

   Die einzige (einfache) Nullstelle von ρ ist $x = 1$. Also sind Einschrittverfahren immer stabil (und dies stimmt mit unserem alten Satz überein).

2. *Aus Integration gewonnene Mehrschrittverfahren sind von der Form*

   $$y_{k+m} - y_{k+m-r} = \dots$$

   Es gilt

   $$\rho(x) = x^m - x^{m-r} = x^{m-r}(x^r - 1).$$

   ρ hat die $(m-r)$–fache Nullstelle 0 und die r–ten Einheitswurzeln, die alle die Vielfachheit 1 haben. Also sind alle diese Verfahren stabil.

3. *Das Mehrschrittverfahren*

   $$y_{k+2} - 2y_{k+1} + y_k = 0$$

   ist konsistent, aber das Differenzenverfahren erfüllt nicht die Wurzelbedingung (1 ist doppelte Nullstelle von ρ, siehe oben), also ist das Mehrschrittverfahren nicht stabil.

4. *In 11.7 gilt*

   $$\rho(x) = x^2 - (1 + \alpha)x + \alpha.$$

   Die Nullstellen sind 1 und α, d.h. es muss erfüllt sein

   (a) $|\alpha| \leq 1$

   (b) $\alpha \neq 1$ *(ansonsten ist 1 doppelte Nullstelle von ρ).*

11.4 Zusammenfassung

11.4.1 Kompetenzen

- Definition der Mehrschrittverfahren, Konsistenz, Stabilitätsbedingung kennen.
- Approximation ausrechnen und interpretierenn können (siehe Einschrittverfahren).
- Konvergenzsatz: Aus Konsistenz und Stabilität folgt Konvergenz.

11.4.2 Mini–Aufgaben

Siehe Übungen.

Kapitel 12

Randwertprobleme

Abschließend wollen wir uns noch der Behandlung linearer Randwertaufgaben zuwenden. Bisher haben wir ausschließlich Anfangswertaufgaben behandelt, d.h. wir suchten Funktionen $y(t)$ mit

$$y'(t) = f(t, y(t)),\ y(a) = y_0, \text{ auf } [a, b].$$

Für eine lineare Differentialgleichung zweiter Ordnung lautet das Anfangswertproblem

$$y''(t) + p(t)y'(t) + q(t)y(t) = f(t),\ y(a) = y_0,\ y'(a) = y'_0.$$

Wir müssen also im Punkt a Werte für y und seine Ableitung vorgeben. Diesen Fall haben wir sehr gut untersucht, die Lösbarkeitsbedingungen sind einfach analytisch angebbar, und wir haben konvergente numerische Methoden für diesen Fall angegeben.

Will man aber etwa die Auslenkung einer eingespannten Saite oder eines an beiden Seiten festgehaltenen Seils beschreiben, so kann man das zwar durch eine lineare Differentialgleichung tun. Randbedingung ist aber offensichtlich, dass die Auslenkung am linken und rechten Rand verschwindet, d.h.

$$y(a) = y(b) = 0.$$

Gleiches passiert, wenn man die Temperaturverteilung eines Stabs berechnen möchte, der an beiden Enden erhitzt wird (1.4).

Wir landen also bei Randwertproblemen, bei denen Bedingungen an die Lösung nicht nur am Randpunkt a, sondern auch am Randpunkt b vorgegeben sind.

Definition 12.1 (Lineare) Randwertprobleme
Ein Randwertproblem sucht Funktionen

$$y(t) \in C^1([a,b] \mapsto \mathbb{R}^n)$$

mit

$$y'(t) = f(t, y(t)),\ g(y(a), y'(a), y(b), y'(b)) = 0$$

für eine Funktion

$$g : \mathbb{R}^{4n} \mapsto \mathbb{R}^n.$$

Typischerweise haben wir Dirichlet–Randbedingungen (bei denen der Wert der Funktion vorgeschrieben ist), Neumann–Randbedingungen (bei denen der Wert der Ableitung vorgeschrieben ist) oder gemischte Bedingungen (bei denen eine Kombination aus Ableitung und Wert der Funktion vorgeschrieben ist, jeweils in einem Randpunkt). Sind g und f (affin) linear, so nennen wir das Randwertproblem auch linear.

Für $n = 2$ ergibt sich damit

1. $y(a) = z_0$, $y(b) = z_1$*: Dirichlet–Randbedingungen.*
   Falls $z_0 = z_1 = 0$: Homogene Dirichlet–Randbedingungen.
2. $y'(a) = z_0$, $y'(b) = z_1$*: Neumann–Randbedingungen.*
   Falls $z_0 = z_1 = 0$: Homogene Neumann–Randbedingungen.
3. $g_1(y(a), y'(a)) = 0$, $g_2(y(b), y'(b)) = 0$*: Gemischte Randbedingungen.*
4. $y(a) = y(b)$, $y'(a) = y'(b)$*: Periodische Randbedingungen.*

Die Lösbarkeit dieses Systems ist leider erheblich schwieriger zu zeigen als bei Anfangswertaufgaben. Wir betrachten als Beispiel eine lineare Differentialgleichung zweiter Ordnung mit konstanten Koeffizienten und Dirichlet–Randbedingungen:

$$-y''(t) + \alpha^2 y(t) = 0,\ \alpha \in \mathbb{R},\ y(a) = z_0,\ y(b) = z_1,\ b > a.$$

Ein Fundamentalsystem dieser linearen Differentialgleichung (6.5) ist gegeben durch

$$y_1(t) = \exp(\alpha t),\ y_2(t) = \exp(-\alpha t)$$

und alle Lösungen sind von der Form

$$y(t) = c_1 y_1(t) + c_2 y_2(t).$$

Zur Erfüllung der Randwerte erhalten wir die Gleichungen

$$c_1 y_1(a) + c_2 y_2(a) = z_0,\ c_1 y_1(b) + c_2 y_2(b) = z_1.$$

Das System ist also lösbar, wenn die Matrix

$$W = \begin{pmatrix} y_1(a) & y_1(b) \\ y_2(a) & y_2(b) \end{pmatrix}$$

invertierbar ist bzw. wenn ihre Determinante nicht verschwindet (die Matrix erinnert an eine Wronski–Determinante, sie ist aber anders definiert als die Wronski–Determinante für Anfangswertprobleme 6.7, und ihre Invertierbarkeit ist deshalb nicht gesichert). Setzen wir die Lösungen ein, so gilt

$$W = \begin{pmatrix} \exp(\alpha a) & \exp(\alpha b)) \\ \exp(-\alpha a) & \exp(-\alpha b) \end{pmatrix}.$$

Die Determinante dieser Matrix ist gerade

$$\det(W) = \exp(\alpha(a-b)) - \exp(-\alpha(a-b))$$

und da die Exponentialfunktion für reelle Argumente monoton ist, verschwindet dieser Ausdruck nicht. Damit ist die Matrix invertierbar und die Dirichlet–Aufgabe eindeutig lösbar. Tatsächlich sind lineare Dirichlet–Randwertaufgaben zweiter Ordnung der Form

$$-y''(t) + q(t)y(t) = f(t),\ y(a) = y_0,\ y(b) = y_1$$

für stetiges f und $q(t) \geq 0$ immer eindeutig lösbar. Wir werden später einen Beweis mit Hilfe der Variationsmethoden angeben.

Anders ist es im Fall $q(t) < 0$. Wir betrachten das fast gleiche Randwertproblem für die Helmholtzgleichung

$$-y''(t) - \alpha^2 y(t) = 0 \text{ auf } [0, \pi],\ \alpha > 0.$$

Wir können die Analyse übertragen mit dem Fundamentalsystem

$$y_1(t) = \cos(\alpha t),\ y_2(t) = \sin(\alpha t).$$

Die trigonometrischen Funktionen haben natürlich ein fundamental anderes Verhalten als die Exponentialfunktion. Insbesondere ist die Matrix

$$W = \begin{pmatrix} \sin(0) & \sin(\alpha\pi) \\ \cos(0) & \cos(\alpha\pi) \end{pmatrix}$$

genau für $\alpha \in \mathbb{Z}$ nicht invertierbar. Wir erhalten also: Die Randwertaufgabe ist eindeutig lösbar, falls $\alpha \notin \mathbb{Z}$. Eine allgemeine Lösbarkeit mit einem griffigen Kriterium (Lipschitz–Stetigkeit wie bei den Anfangswertaufgaben) ist nicht gegeben, es muss jede Gleichung einzeln untersucht werden.

An dem einfachen Beispiel der linearen Differentialgleichungen zweiter Ordnung mit homogenen Randbedingungen lasen sich bereits alle Schwierigkeiten studieren. Wir betrachten im Folgenden das **Sturm–Liouville–Problem**

$$-(p(t)y'(t))' + q(t)y(t) = f(t),\ p(t) > p_0 > 0$$

mit linearen Randbedingungen, differenzierbarer Funktion p und stetiger Funktion q. Wir betrachten zunächst den Fall $p = 1$ und homogene Dirichlet–Randbedingungen, also

$$-y''(t) + q(t)y(t) = f(t),\ y(a) = y(b) = 0.$$

Wir zitieren zunächst den Satz:

Satz 12.2 *Sei $q(t) \geq 0$, f stetig. Dann ist das Sturm–Liouville–Problem mit homogenen Dirichlet–Anfangswerten eindeutig lösbar.*

In den folgenden beiden Abschnitten untersuchen wir zwei einfache Ansätze zur numerischen Lösung dieses Problems.

Bemerkung: Wir betrachten in diesem Abschnitt nur gewöhnliche Differentialgleichungen. Tatsächlich lassen sich, im Gegensatz zu Anfangswertaufgaben, alle vorgestellten Methoden und Sätze in höhere Dimensionen, also für partielle Differentialgleichungen, übertragen.

In einer Dimension sind sie aber viel anschaulicher, oft trivial. Zum Verständnis der Verhältnisse bei partiellen Differentialgleichungen ist es also sehr nützlich, die Interpretation als gewöhnliche Differentialgleichung im Kopf zu behalten. Wir werden hier immer einige Beziehungen zu den partiellen Differentialgleichungen angeben, auch wenn Sie diese bisher nicht gehört haben oder hören werden.

12.1 Schießverfahren

Der Hintergrund dieser Verfahren ist einfach zu erraten. Eine Kanone, deren Abschusswinkel α variabel ist, stehe in einer Ebene am Punkt a. Das Zielobjekt stehe am Punkt b. Sei $y_\alpha(t)$ die Höhe der Kugel auf dem Weg von a nach b an einem Punkt t. Dann ist unser Ziel, durch Versuch und Irrtum $\alpha = \widetilde{\alpha}$ so zu wählen, dass $y_{\widetilde{\alpha}}(b) = 0$. y_α genügt einer gewöhnlichen Differentialgleichung, das gesuchte $y_{\widetilde{\alpha}}$ genügt den homogenen Dirichlet–Randbedingungen $y_{\widetilde{\alpha}}(a) = y_{\widetilde{\alpha}}(b) = 0$.

Dies ist die Beschreibung des uralten QBasic–Spiels Gorillas, das noch an einigen Stellen im Internet verfügbar ist, etwa hier.

Dies interpretieren wir noch etwas um. Gesucht sei die Lösung y des Sturm–Liouville–Problems. Für jedes α besitzt die Anfangswertaufgabe mit der vorgegebenen Differentialgleichung und $y(a) = 0$, $y'(a) = \alpha$ eine Lösung und ist leicht numerisch berechenbar mit den Algorithmen des letzten Kapitels. Wir suchen nun ein α mit $F(\alpha) := y_\alpha(b) = 0$. F ist eine nichtlineare, eindimensionale, berechenbare Funktion. Wir suchen eine Nullstelle von F. Verfahren dazu haben wir kennengelernt, etwa das Newton–Verfahren oder das Sekanten–Verfahren (Regula falsi). Voraussetzung zur Anwendung dieser Verfahren ist, dass die Funktion mindestens einmal stetig differenzierbar ist. Tatsächlich gilt der Satz:

Satz 12.3 (Differenzierbarkeit der Lösungen von Anfangswertaufgaben nach ihren Anfangswerten)
Sei $y_\alpha(t)$ die Lösung der Anfangswertaufgabe für die Differentialgleichung des Sturm–Liouville–Problems für die Anfangswerte

$$y'_\alpha(a) = \alpha,\ y_\alpha(a) = 0.$$

Sei

$$F(\alpha) := y_\alpha(b).$$

Dann ist F differenzierbar.

(ohne Beweis)

Dies liefert uns ein numerisches Verfahren zur Lösung von Randwertaufgaben.

1. Implementiere die Funktion $F(\alpha)$, die eine Approximation für $y_\alpha(b)$ berechnet mit den numerischen Verfahren des letzten Kapitels.
2. Nutze ein numerisches Verfahren zur Bestimmung der Nullstelle von F, zum Beispiel das Newton–Verfahren oder das Sekanten–Verfahren.

12.2 Diskretisierungsverfahren

Wir können auch die Diskretisierungsidee aus dem letzten Kapitel zur Lösung einsetzen. Sei wieder

$$I_h = (t_0, \ldots, t_N)$$

ein zulässiges Gitter auf $[a, b]$, ohne Einschränkung wählen wir es in diesem Kapitel immer äquidistant. Sei y die Lösung des Sturm–Liouville–Problems. Wir suchen Gitterfunktionen

$$y_h = (y_k),\ y(t_k) \sim y_h(t_k).$$

In den inneren Punkten des Gitters diskretisieren wir die Differentialgleichung konsistent, gibt $N-1$ Gleichungen, plus zwei Randbedingungen, macht insgesamt $N+1$ Gleichungen für $N+1$ Unbekannte.

Wir betrachten als Modellproblem die eindimensionale **Poisson–Gleichung**

$$-y''(t) = f(t)$$

für t in $[0,1]$ auf einem äquidistanten Gitter mit homogenen Dirichlet–Randbedingungen, $h = 1/N$. Zur Diskretisierung nutzen wir die Diskretisierung der zweiten Ableitung aus 9.1. Wir erhalten die Gleichungen

$$\begin{aligned} -\frac{y_0 - 2y_1 + y_2}{h^2} &= f_1 = f(t_1) \\ &\vdots \\ -\frac{y_{N-2} - 2y_{N-1} + y_N}{h^2} &= f_{N-1} = f(t_{N-1}) \end{aligned}$$

Setzen wir die Randbedingung $y_0 = y_N = 0$ ein, so erhalten wir für die Unbekannten $y_1, \ldots, y_{N-1}$ das lineare Gleichungssystem

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}}_{=:L_h} \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{pmatrix}}_{=:y_h} = \underbrace{\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-2} \\ f_{N-1} \end{pmatrix}}_{=:f_h}.$$

Konsistenz für ein solches Verfahren definieren wir wieder wie für die Anfangswertprobleme (10.7): Wir setzen eine Lösung y der Differentialgleichung in diese Diskretiserung ein und definieren den Unterschied zwischen linker und rechter Seite als Konsistenz. In diesem speziellen Fall wäre der Konsistenzfehler also

$$\tau_h(y) = ||L_h(y|_{I_h}) - f_h||.$$

Dann ist die Diskretisierung konsistent in der $\infty-$ und euklidischen Norm nach 9.2.

Ist das Gesamtverfahren auch konvergent, d.h. gilt

$$||y|_{I_h} - y_h|| \mapsto 0$$

? Es gilt

$$\begin{aligned} ||y|_{I_h} - y_h|| &= ||L_h^{-1}(L_h(y|_{I_h}) - L_h y_h)|| \\ &= ||L_h^{-1}(L_h(y|_{I_h}) - f_h)|| \\ &\leq \underbrace{||L_h^{-1}||}_{\text{Stabilitätsfaktor}} \cdot \underbrace{||L_h(y|_{I_h}) - f_h||}_{\text{Konsistenzfehler}} \end{aligned}$$

Also:

Definition 12.4 (Stabilität für Diskretisierungen von Randwertproblemen)
Die Diskretisierung eines Randwertproblems heißt stabil, falls $||L_h^{-1}||$ unabhängig von h nach oben beschränkt ist.

Damit gilt sofort wieder unser Satz:

Satz 12.5 (Konvergenz für Diskretisierungen von Randwertproblemen)
Aus Stabilität und Konsistenz (der Ordnung p) folgt Konvergenz (der Ordnung p).

Wir bestimmen nun $||L_h^{-1}||_2$. Nach 5.31 müssen wir, da L_h symmetrisch ist, hierzu die Eigenwerte von L_h bestimmen.

Die Eigenvektoren von L_h lassen sich leicht angeben: Es sind die Vektoren

$$y^k \in \mathbb{R}^{N-1},\ (y^k)_j = \sin(k\pi jh),\ k, j = 1, \ldots, N-1,$$

zu den Eigenwerten

$$\lambda_k = \frac{4}{h^2}\sin^2(kh\pi/2),\ k = 1, \ldots, N-1.$$

Mit den Additionstheoremen gilt nämlich

$$\begin{aligned}
&-\sin(x-y) + 2\sin(x) - \sin(x+y)\\
&= -\sin(x)\cos(y) + \cos(x)\sin(y) + 2\sin(x) - \sin(x)\cos(y) - \cos(x)\sin(y)\\
&= 2(1-\cos(y))\sin(x)\\
&= 2(1-\cos(\frac{y}{2}+\frac{y}{2}))\sin(x)\\
&= 2(1-\cos^2\frac{y}{2}+\sin^2\frac{y}{2})\sin(x)\\
&= (4\sin^2\frac{y}{2})\sin(x)
\end{aligned}$$

und dann durch Einsetzen von $x = k\pi jh$ und $y = k\pi h$.

Es gilt

$$\lambda_k \geq \lambda_1 \geq \frac{4}{h^2}\left(\frac{2}{\pi}\right)^2 \frac{h^2\pi^2}{4} = 4$$

unter Benutzung der Abschätzung

$$\sin x \geq \frac{2}{\pi}x,\ x \in [0, \pi/2].$$

L_h hat keinen Eigenwert 0. Insbesondere ist L_h invertierbar, das oben angegebene Gleichungssystem für y_h ist eindeutig lösbar, und sogar sehr effizient lösbar, denn L_h ist eine Tridiagonalmatrix (d.h., sie hat nur Einträge unter– und oberhalb der Hauptdiagonalen), für diese Matrizen vereinfacht sich das Gaussverfahren enorm. L_h ist symmetrisch positiv semidefinit. Die Eigenwerte von L_h^{-1} sind die Kehrwerte der Eigenwerte von L_h, also ist nach 5.31

$$||L_h^{-1}||_2 \leq 1/4.$$

Für unser Verfahren gilt mit der euklidischen Norm: Der Stabilitätsfaktor ist beschränkt, die Konsistenz ist ein $O(h^2)$, also ist das Verfahren konvergent von der Ordnung 2. Dies legt nahe

Leider steht hier zunächst nur die euklidische Norm. Dies ist ungünstig: Es garantiert uns keine punktweise Konvergenz und insbesondere keine punktweisen Fehlerabschätzungen.

Bisher haben wir Konvergenz immer bezüglich der Maximumnorm betrachtet. Das wollen wir auch beibehalten. Zur Untersuchung dieser Konvergenz benötigen wir einige spezielle Eigenschaften der oben angegebenen Matrix L_h.

Definition 12.6 (M-Matrizen)
Sei A eine reelle, invertierbare $N \times N$–Matrix. A heißt M–Matrix genau dann, wenn

$$A_{ii} > 0 \, \forall i, \; A_{ik} \leq 0 \, \forall i \neq k, \; A_{ik}^{-1} \geq 0 \, \forall i, k.$$

Die wesentliche Eigenschaft von Matrizen mit positiven Einträgen ist ihre Monotonie.

Lemma 12.7 (Monotonie für M–Matrizen)
Sei $A \in \mathbb{R}^{N \times N}$ eine M–Matrix. Seien

$$u, \; v \in \mathbb{R}^n, \; u \leq v.$$

Dann gilt

$$A^{-1}u \leq A^{-1}v.$$

Hier und im Folgenden sind Vektorungleichungen immer elementweise gemeint.

Beweis: Da alle Einträge von A^{-1} nichtnegativ sind, gilt

$$A^{-1}(v - u) \geq 0.$$

□

Satz 12.8 (L_h ist M–Matrix)
Sei A eine reelle, invertierbare $N \times N$–Matrix mit positiven Hauptdiagonaleelementen und nicht–positiven Außerdiagonalelementen. Es gelte

$$A_{jj} \geq -\sum_{k \neq j} A_{jk}.$$

Dann ist A eine M–Matrix. Insbesondere ist L_h eine M–Matrix.

Beweis: Alle Voraussetzungen für die M-Matrix sind erfüllt, wir müssen nur noch zeigen, dass die Einträge von L_h^{-1} nichtnegativ sind.
Sei $b < 0$. Wir zeigen: Dann ist auch $A^{-1}b < 0$.
Sei $x = A^{-1}\, b \Rightarrow Ax = b$. Angenommen,

$$x_j = \max\{x_1, \ldots, x_N\} \geq 0.$$

Da

$$\sum_{k=1}^{N} A_{j,k}x_k = (Ax)_j = b_j,$$

gilt

$$\begin{aligned} A_{jj}x_j &= b_j - \sum_{k\neq j} A_{jk}x_k \\ &< -\sum_{k\neq j} A_{jk}x_j \\ &\leq A_{jj}x_j. \end{aligned}$$

Dies ist ein Widerspruch, also gilt

$$b < 0 \Longrightarrow A^{-1}b < 0.$$

Sei $b \leq 0$, und $b^{(l)}$ eine Folge von Vektoren mit

$$b^{(l)} < 0,\ b^{(l)} \to b.$$

Damit gilt

$$A^{-1}b = \lim A^{-1}b^{(l)} \leq 0.$$

Mit der Wahl $b = -e_k \leq 0$ folgt $A^{-1}e_k \geq 0$, die k. Spalte von A^{-1} ist also ≥ 0, und damit $A^{-1} \geq 0$, also ist A eine M–Matrix. □

Satz 12.9 (Stabilität des Standardverfahrens)
Das Standard–Differenzenschema zur Berechnung der Lösung der eindimensionalen Poisson–Gleichung ist stabil bezüglich der Maximumnorm.

Beweis: Zu zeigen ist: $||L_h^{-1}||_\infty$ ist unabhängig von h beschränkt.
Da die Einträge von $(L_h)^{-1}$ nichtnegativ sind, gilt nach 5.11

$$||(L_h)^{-1}||_\infty = \max_k \sum_{j=1}^{N-1} |(L_h)^{-1}_{k,j}| = \max_k \sum_{j=1}^{N-1} (L_h)^{-1}_{k,j} = ||L_h^{-1}\mathbf{1}||_\infty.$$

Hier ist $\mathbf{1}$ der Vektor aus dem $\mathbb{R}^{N-1}$, in dem alle Einträge 1 sind.
Wir betrachten nun das Poissonproblem mit $f = 1$. Dann ist die analytische Lösung gegeben durch

$$w(t) = \frac{1}{2}t(1-t)$$

denn

$$-w''(t) = 1,\ w(0) = 0,\ w(1) = 0,\ w(t) \in [0, \frac{1}{8}] = \frac{1}{4}.$$

Sei

$$w_h = w|_{I_h}.$$

L_h ist konsistente Diskretisierung der zweiten Ableitung, d.h.

$$||L_h(w_h) - \mathbf{1}||_\infty \leq Ch^2$$

und damit

$$L_h(w_h) \geq -Ch^2 + \mathbf{1}$$

und insbesondere, falls h klein genug ist,

$$L_h(w_h) \geq \frac{1}{2}\mathbf{1}.$$

Diese Vektorungleichungen sind wieder alle jeweils elementweise zu interpretieren. L_h ist M–Matrix. Nach 12.7 gilt

$$2w_h = 2L_h^{-1}L_h(w_h) \geq L_h^{-1}\mathbf{1} \Rightarrow ||(L_h)^{-1}\mathbf{1}||_\infty \leq 2||w_h||_\infty.$$

Dies setzen wir nun noch zusammen:

$$||L_h^{-1}||_\infty = ||L_h^{-1}\mathbf{1}||_\infty \leq 2||w_h||_\infty \leq 2||w||_\infty = \frac{2}{8}.$$

Also ist die Supremumsnorm unabhängig von h beschränkt, die Diskretisierung ist stabil und damit das Verfahren konvergent. □

Bemerkung: Für den einfachen Fall der Poissongleichung, den wir hier betrachten, ist die Diskretisierung sogar exakt, d.h. $L_h w_h = 1$. Damit ist für diesen Fall sogar

$$||L_h^{-1}||_\infty \leq ||w||_\infty = \frac{1}{8}.$$

Bemerkung: Wir haben oben gesehen, dass für $p = 1$ die analytische Lösbarkeit des Sturm–Liouville–Problems vom Vorzeichen von q abhängt. Hier bekommen wir ein ähnliches Problem: Man sieht leicht, dass q weiter positiv semidefinit und eine

M–Matrix bleibt, wenn $q \geq 0$, unsere Analyse lässt sich also mit leichten Änderungen retten.
Sobald $q < 0$, also dort, wo die analytische Lösbarkeit nicht sicher war, bricht diese Argumentation aber zusammen. Unsere analytischen Schwierigkeiten für $q < 0$ übersetzen sich in numerische Schwierigkeiten. Die Analysis und die Numerik hängen also, wie hoffentlich erwartet, sehr eng zusammen und sollten immer zusammen betrachtet werden.

Korollar 12.10 (Konvergenz des Standardverfahrens)
Das Standardverfahren zur Berechnung der Lösung der eindimensionalen Poisson–Gleichung ist konvergent von der Ordnung 2 (bezüglich der Maximumnorm).

Leider ist dieses Ergebnis unbefriedigend. Anders als bei den Anfangswertproblemen ist der Beweis recht uneinsichtig und nur schwer auf andere Differentialgleichungen oder Diskretisierungen übertragbar. Wir werden daher ein neues Hilfsmittel, die Variationsrechnung, kennenlernen.

12.3 Variationsmethoden

Bei der Behandlung des Steinwurfs in 1.1 hatten wir bereits bemerkt: Differentialgleichungen in der Physik entstehen häufig aus Energiebetrachtungen. Dabei wird meist ein durch Integrale definiertes Energiefunktional minimiert. Dies führt in der Praxis auf eine Integralgleichung für die Lösung.

Unter der Annahme, dass die Lösungen differenzierbar sind, können wir diese dann in eine Differentialgleichung umschreiben. Das ist allerdings eine Einschränkung und führt dazu, dass viele Probleme der Physik als Differentialgleichung keine Lösung besitzen. Es liegt daher nahe, statt der Differentialgleichungen die zugehörige Integralgleichung bzw. das Minimierungsproblem zu untersuchen.

Definition 12.11 (Grundräume zur Lösung des Variationsproblems)
Wir betrachten alle Funktionen immer auf dem abgeschlossenen Intervall $[a, b]$.

1. *C^k ist der Raum der k–mal stetig differenzierbaren Funktionen auf $[a, b]$. C^k ist vollständig bzgl. der Norm*
$$||f||_{k,\infty} = \max_{j=0\ldots k} ||f^{(j)}||_\infty.$$

2. *$C^\infty([a, b])$ ist der Raum der unendlich oft differenzierbaren Funktionen auf $[a, b]$.*

3.
$$C_0^\infty([a, b]) = \{f \in C^\infty([a, b]) : f^{(j)}(a) = f^{(j)}(b) = 0 \,\forall j\}.$$

$C_0^\infty(\mathbb{R})$ ist der Raum der unendlich oft differenzierbaren Funktionen mit kompaktem Träger auf $\mathbb{R}$.
Entsprechend $C_0^{(k)}$.

4. *Beispiel: Sei*
$$w(x) = \begin{cases} e^{\frac{1}{x^2-1}}, & |x| < 1 \\ 0, & \text{sonst.} \end{cases}$$
Dann ist $w \in C_0^\infty(\mathbb{R})$ mit Träger in $[-1,1]$. Entsprechend ist
$$w_{\epsilon,a} := \frac{1}{\epsilon} w\left(\frac{x-a}{\epsilon}\right)$$
in $C_0^\infty(\mathbb{R})$ mit Träger in $[a-\epsilon, a+\epsilon]$. Es gilt
$$w_{\epsilon,a}(x) > 0 \, \forall x \in [a-\epsilon, a+\epsilon], \int_{\mathbb{R}} w_{\epsilon,a}(x)\, dx = C = \int_{-1}^{1} w(x)\, dx.$$
Sei weiter f stetig. Dann gilt
$$\int_{\mathbb{R}} f(x) w_{\epsilon,a}(x) = \int_{-1}^{1} f(\epsilon y + a)\, w(y)\, dy \xrightarrow{\epsilon \to 0} C\, f(a).$$

5. *$L^2([a,b])$ ist der Raum der Funktionen auf $[a,b]$, die quadratisch integrierbar sind, d.h.*
$$\int_a^b |f(t)|^2 dt < \infty.$$
Mit dem Skalarprodukt
$$(f,g) = \int_a^b f(t) g(t) dt, \; ||f||^2 := (f,f) \, \forall f, g \in L^2$$
ist L^2 ein Hilbertraum, insbesondere vollständig. Es gilt die Cauchy–Schwarz–Ungleichung
$$|(f,g)| \leq ||f|| \cdot ||g||.$$

Satz 12.12 *(Dichtheit der stetigen Funktionen)*
C^0 liegt dicht in L^2, d.h.
$$\forall\, f \in L^2 \, \exists (f_n) \in C^0 : \; f_n \to f.$$

Beweis: Ohne Beweis, Hint: es gilt $w_\epsilon * f \to f$ (siehe 12.11 (Aussage 4) und 12.24).

Satz 12.13 *Es sei $f \in L^2([a,b])$, und es sei*

$$(f,\varphi) = 0 \,\forall \varphi \in C_0^\infty([a,b]).$$

Dann ist $f = 0$.

Beweis: Wir zeigen den Satz für f stetig, dann folgt der Satz aus 12.12.
Angenommen, $f(z) \neq 0$ für ein $z \in (a,b)$. Dann gibt es eine ϵ–Umgebung von z, so dass f dort sein Vorzeichen nicht ändert. Nach Voraussetzung gilt

$$0 = (f, w_{\epsilon,z}) = \int_a^b f(x) w_{\epsilon,z}(x)\, dx = \int_{z-\epsilon}^{z+\epsilon} \underbrace{f(x)}_{<0 \vee >0} \underbrace{w_{\epsilon,z}(x)}_{>0}\, dx \neq 0.$$

Da $w_{\epsilon,z} \in C_0^\infty([a,b]$ ist dies ein Widerspruch, also $f = 0$. □

Wir wenden dies nun an auf einen Spezialfall des Sturm–Liouville–Randwertproblems, diesmal mit **natürlichen Randbedingungen**. Achtung: Im Folgenden betrachten wir immer dieses Beispiel, insbesondere seien die Voraussetzungen an p und q immer erfüllt.

Definition 12.14 *(Sturm–Liouville–Modellproblem für Randwertprobleme)*

$$-(py')'(t) + q(t)y(t) = f(t),\ y(a) = 0,\ y'(b) = 0$$

mit $p \in C^1$, $q \in C^0$, und

$$p(t) \geq p_0 > 0, q(t) \geq 0.$$

Nach den Vorbemerkungen aus dem letzten Kapitel vermuten wir, dass dieses Problem eine (eindeutige) Lösung hat. Aber zunächst wollen wir dieses Problem in eine Variationsgleichung umschreiben. Sei dazu $\varphi \in C^1([a,b])$ und y irgendeine Lösung der Differentialgleichung. Dann gilt mit Multiplikation mit φ und Integration

$$\int_a^b (-(py')'(t) + q(t)y(t))\varphi(t)\mathrm{d}t = \int_a^b f(t)\varphi(t)\mathrm{d}t$$

und damit mit partieller Integration

$$\underbrace{\int_a^b p(t)y'(t)\varphi'(t) + q(t)y(t)\varphi(t)\mathrm{d}t}_{=:B(y,\varphi)} - [p \cdot y' \cdot \varphi]_a^b = \underbrace{\int_a^b f(t)\varphi(t)\mathrm{d}t}_{=:F(\varphi)}.$$

Mit diesen Bezeichnungen gilt

Satz 12.15 *Sei* $X = \{u \in C^1([a,b]) : u(a) = 0\}$.

1. *Falls* $u \in X$ *das Randwertproblem 12.14 erfüllt, so gilt*

$$B(u,v) = F(v)\,\forall v \in X.$$

2. *Falls*

$$B(u,v) = F(v) \forall v \in X,$$

und $u \in C^2$*, so ist* u *Lösung des Randwertproblems 12.14.*

Beweis: Falls $u \in X$ Lösung von 12.14 ist, so ist $u'(b) = 0$. Sei $v \in X$, also $v(a) = 0$. Also fällt die eckige Klammer in der Vorbemerkung weg, und es gilt

$$B(u,v) = F(v).$$

Sei nun

$$B(u,v) = F(v)\,\forall v \in X.$$

Wähle ein v aus $C_0^\infty([a,b])$ beliebig. Dann gilt insbesondere $v(a) = v(b) = 0$ und $v \in X$. Wieder fällt die eckige Klammer in der Vorbemerkung weg, und es gilt

$$0 = B(u,v) - F(v) = \int_a^b \left((-pu')'(t) + q(t)u(t) - f(t)\right) v(t)\,dt = 0.$$

Nach 12.13 ist u damit Lösung der Differentialgleichung. Da $u \in X$, gilt $u(a) = 0$. Sei nun $v \in X$ mit $v(b) \neq 0$. Da u Lösung der Differentialgleichung ist, gilt nach der Vorbemerkung

$$0 = B(u,v) - F(v) = [p \cdot u' \cdot v]_a^b.$$

Da $v \in X$, gilt $v(a) = 0$. Also ist $p(b)u'(b)v(b) = 0$. Da $p(b) \geq p_0 > 0$ und $v(b) \neq 0$, muss gelten $u'(b) = 0$, also ist u Lösung des Randwertproblems. □

Wir haben also unser Randwertproblem 12.14 in eine Variationsgleichung umgeschrieben.

Bemerkenswert daran ist, dass diese einen erweiterten Lösungsbegriff hat: Falls das Randwertproblem eine Lösung besitzt, so ist es äquivalent zum Variationsproblem. Falls das Randwertproblem keine Lösung besitzt, so kann es sein, dass das Variationsproblem trotzdem eine Lösung besitzt, wir haben hier also den Lösungsbegriff etwas erweitert. Vor dem physikalischen Hintergrund macht dies absolut Sinn, dort kommen durchaus Funktionen vor, die nicht in C^2 sind (und damit keine Lösungen der eigentlichen Differentialgleichung). Man spricht hier auch von einer schwachen Formulierung der Differentialgleichung.

Wir wollen nun noch einen Zusammenhang zwischen Lösung einer Gleichung und Minimierung beweisen. Dazu bemerken wir zunächst die folgenden Eigenschaften:

1. B ist bilinear, also linear in beiden Argumenten.
2. $B(u,v) = B(v,u)$, also ist B symmetrisch.
3. $B(u,u) \geq 0$.
4. $u \in X,\ u \neq 0 \Rightarrow B(u,u) > 0$ (Hint: Falls $B(u,u) = 0$, so ist $u' = 0$, also u konstant, und der Satz folgt wegen $u(a) = 0$).
5. F ist linear und stetig.

Mit diesen Eigenschaften beweisen wir

Satz 12.16 *Sei*

$$I : X \mapsto \mathbb{R},\ I(u) := \frac{1}{2}B(u,u) - F(u).$$

I nimmt sein Mimimum in $u \in X$ an genau dann, wenn

$$B(u,v) = F(v)\,\forall v \in X.$$

Beweis: Seien $u, v \in X$. Wir definieren $g : \mathbb{R} \mapsto \mathbb{R}$,

$$\begin{aligned} g(\epsilon) &:= I(u + \epsilon v) \\ &= \frac{1}{2}B(u + \epsilon v, u + \epsilon v) - F(u + \epsilon v) \\ &= \underbrace{\frac{1}{2}B(u,u) - F(u)}_{=I(u)} + \epsilon(B(u,v) - F(v)) + \underbrace{\frac{\epsilon^2}{2}B(v,v)}_{\geq 0}. \end{aligned}$$

Falls

$$B(u,v) = F(v)\,\forall v \in X,$$

so gilt damit

$$I(u + \epsilon v) \geq I(u)\,\forall v \in X$$

und damit nimmt I sein Minimum in u an.
Falls I sein Minimum in u annimmt, so nimmt g sein Minimum an für $t = 0$. Also gilt $g'(0) = 0$, und damit

$$B(u,v) = F(v)\,\forall v \in X.$$

Falls $B(v,v) > 0$ für $v \neq 0$ (positiv definit), so ist das Minimum sogar eindeutig. □

Bei diesem Beweis haben wir nur die oben genannten Eigenschaften benutzt. Wir schauen nun noch auf ein Beispiel im $\mathbb{R}^n$.

Korollar 12.17 *Sei $X = \mathbb{R}^n$. Zu lösen sei die Aufgabe $Ax = b$ mit $A \in \mathbb{R}^{n\times n}$ symmetrisch positiv definit. Es sei*

$$B(x, y) := (Ax, y),\ F(y) := (b, y),\ x, y \in \mathbb{R}^n.$$

Dann gilt

$$B(x, y) = F(y)\, \forall y \in \mathbb{R}^n \Leftrightarrow Ax = b.$$

Beweis: Setze $y = Ax - b$. Dann gilt

$$(Ax, Ax - b) = (b, Ax - b) \Leftrightarrow (Ax - b, Ax - b) = 0 \Leftrightarrow Ax = b.$$

□

Das so definierte B ist bilinear, symmetrisch und positiv definit, F ist linear. Also gilt mit diesen Definitionen:

Korollar 12.18 *Es sei*

$$B(x, y) := (Ax, y),\ F(y) := (b, y).$$

$x \in \mathbb{R}^n$ *ist genau dann Lösung von $Ax = b$, wenn*

$$I(y) := \frac{1}{2}B(y, y) - F(y)$$

sein Minimum an der Stelle x annimmt.

Das heißt: Anstatt das Gleichungssystem zu lösen, können wir genausogut das Minimierungsproblem lösen. Dies ist der Ausgangspunkt für die Krylovraum–Verfahren, eine Klasse von iterativen Verfahren zur Lösung von linearen Gleichungssystemen.

12.4 Sobolevräume

Mit unseren Sätzen haben wir die Existenz einer Lösung des (schwachen) Randwertproblems auf die Existenz eines Minimierers von I zurückgeführt. Diese wollen wir beweisen. Üblicherweise geht man dabei so vor: Man zeigt, dass das Infimum I_0 von I endlich ist, und dass jede Minimalfolge, also eine Folge y_n mit

$$I(y_n) \to I_0,$$

eine Cauchyfolge ist. Aus der Vollständigkeit des Grundraums folgt dann die Existenz eines Minimierers. Leider ist unser Raum nicht einmal vollständig.

Dies gehen wir zunächst an. Wir wollen C^1 in L^2 vervollständigen, indem wir seinen Abschluss mit hinzunehmen. Bisher haben wir als einzige Eigenschaften der differenzierbaren Funktionen die partielle Integration genutzt. Es liegt also nahe, eine Erweiterung von C^1 als den Teilraum von L^2 zu definieren, in dem die partielle Integration erlaubt ist.

Definition 12.19 (schwache Differenzierbarkeit)
$v' \in L^2$ heißt schwache Ableitung von $v \in L^2$, falls $\forall \varphi \in C_0^\infty$

$$\int_a^b v(t)\varphi'(t)dt = -\int_a^b v'(t)\varphi(t)dt$$

und entsprechend für höhere Ableitungen.

Damit gilt natürlich insbesondere: Falls eine Funktion differenzierbar ist, so ist sie auch schwach differenzierbar und die Ableitungen sind gleich.

Beispiel 12.20 (schwache Differenzierbarkeit der Betragsfunktion)
$v(x) = |x|$ ist schwach differenzierbar auf $[-1, 1]$:

$$\begin{aligned}\int_{-1}^1 |t|\varphi'(t)dt &= \int_{-1}^0 (-t)\varphi'(t)dt + \int_0^1 t\varphi'(t)dt \\ &= \int_{-1}^0 \varphi(t)dt + \int_0^1 -\varphi(t)dt + [t\varphi]_0^1 - [t\varphi]_{-1}^0 \\ &= -\int_{-1}^1 \varphi(t)\,\mathrm{sgn}(t)dt\end{aligned}$$

Die (schwache) Ableitung der Betragsfunktion ist also die Signumfunktion. Für die Signumfunktion gilt

$$\begin{aligned}\int_{-1}^1 \mathrm{sgn}(t)\varphi'(t)dt &= -\int_{-1}^0 \varphi'(t)dt + \int_0^1 \varphi'(t)dt \\ &= -2\varphi'(0).\end{aligned}$$

Dies lässt sich nicht als Integral schreiben, also ist die Signumfunktion nicht schwach differenzierbar, obwohl sie eine L^2–Funktion ist. Es gilt also

$$H^1 \neq L^2.$$

Definition 12.21 (Sobolev–Räume)
Der Raum H^1 ist der Raum der schwach differenzierbaren Funktionen und heißt Sobolev–Raum. Auf H^1 definieren wir das Skalarprodukt

$$(f,g) = (f,g)_{L^2} + (f',g')_{L^2} = \int_a^b f(x)g(x)\mathsf{d}x + \int_a^b f'(x)g'(x)\mathsf{d}x$$

mit der zugehörigen Norm $||f||^2_{H^1} = (f,f)$.

Satz 12.22 (Vollständigkeit der Sobolev–Räume)
H^1 ist vollständig.

Beweis: Sei (f_n) eine Cauchyfolge bzgl. $||\cdot||_{H^1}$. Dann sind (f_n) und (f_n') Cauchyfolgen bzgl. L^2. L^2 ist vollständig, also gilt

$$f_n \to f,\ f_n' \to f',\ f, f' \in L^2.$$

Dann gilt für $\varphi \in C_0^\infty$

$$\begin{aligned}\int_a^b f(t)\varphi'(t)\mathsf{d}t &= \lim_{k\to\infty}\int_a^b f_k(t)\varphi'(t)\mathsf{d}t \\ &= \lim_{k\to\infty} -\int_a^b f_k'(t)\varphi(t)\mathsf{d}t \\ &= -\int_a^b f'(t)\varphi(t)\mathsf{d}t\end{aligned}$$

und damit ist f' schwache Ableitung von f.

$$||f_n - f||^2_{H^1} = ||f_n - f||^2_{L^2} + ||f_n' - f'||^2_{L^2} \to 0$$

also konvergiert f_n gegen die schwach differenzierbare Funktion u bzgl. $||\cdot||_{H^1}$, also ist H^1 vollständig. Tatsächlich ist H^1 der kleinste Raum mit dieser Eigenschaft, der C^1 enthält, also die Vervollständigung von C^1 bzgl. $||\cdot||_{H^1}$. □

Zwei der wichtigsten Sätze über Sobolevräume sind die Sobolevsche Ungleichung und der Sobolevsche Einbettungssatz. In einer Dimension sind sie trivial.

Satz 12.23 (Sobolevsche Ungleichung)
Sei $f \in C^1$. Dann gibt es ein $C > 0$ mit

$$||f||_\infty \le C||f||_{H^1}.$$

Beweis: Wir betrachten das Problem auf $[-1,1]$. Sei zunächst $s \leq 0$. Dann gilt

$$\begin{aligned} f(s) &= \int_0^1 ((t-1)f(t+s))' \mathrm{d}t \\ &= \int_0^1 f(t+s) + (t-1)f'(t+s) \mathrm{d}t \end{aligned}$$

und damit nach Cauchy–Schwarz

$$\begin{aligned} |f(s)| &\leq (\int_0^1 1 \mathrm{d}t)^{1/2} (\int_{-1}^1 f(t)^2 \mathrm{d}t)^{1/2} + (\int_0^1 (t-1)^2 \mathrm{d}t)^{1/2} (\int_{-1}^1 f'(t)^2 \mathrm{d}t)^{1/2} \\ &\leq ||f||_{H^1}. \end{aligned}$$

Für $s > 0$ betrachtet man $-t + s$ statt $t + s$ und bekommt dieselbe Ungleichung. □

Satz 12.24 (H^1-Funktionen sind stetig, Sobolevscher Einbettungssatz)
Sei $f \in H^1$. Dann gibt es eine stetige Funktion g mit $f = g$ f.ü., und

$$||g||_\infty \leq C||f||_{H^1}.$$

Beweis: Sei $f \in H^1$. Die Funktionen

$$f_n(s) = \frac{1}{M} \int_a^b w_{1/n,s}(t) f(t) \mathrm{d}t = \frac{1}{M} \int_{s-1/n}^{s+1/n} w_{1/n,s}(t) f(t) \mathrm{d}t,\ M = \int_{\mathbb{R}} w(t)\, \mathrm{d}t$$

liegen in C^∞ (Differentiation unter dem Integralzeichen). Sie konvergieren gegen f bzgl. H^1, also sind sie insbesondere eine Cauchyfolge in H^1. Dies sieht man anschaulich ein – ein einfacher, aber längerer Beweis für alle Dimensionen findet sich in Evans [2010], Anhang C.4.
Es gilt nach 12.23

$$||f_n - f_m||_\infty \leq C\, ||f_n - f_m||_{H^1} \to 0.$$

(f_n) ist also auch eine Cauchyfolge in C^0 bzgl. $||\cdot||_\infty$. C^0 ist vollständig bezüglich $||\cdot||_\infty$, also gilt

$$f_n \to g \text{ bzgl. } ||\cdot||_\infty,\ g \in C^0.$$

Damit konvergiert f_n aber auch gegen g bzgl. L^2. f_n konvergiert also gegen g und f bezüglich L^2, also gilt $f = g$ f.ü. Weiter ist

$$||g||_\infty = \lim_{n\to\infty} ||f_n||_\infty \leq C \lim_{n\to\infty} ||f_n||_{H^1} = C||f||_{H^1}.$$

□

Der Sobolevsche Einbettungssatz hat eine wichtige Folgerung. Wir haben die H^1–Funktionen als L^2–Funktionen definiert. Damit besitzen sie keine Punktauswertung: L^2–Funktionen dürfen auf Nullmengen umdefiniert werden, ohne dass sie sich ändern. Da jetzt aber ohne Einschränkung jede H^1–Funktion stetig ist, können wir sie an Punkten auswerten.

Wir kommen zu einer der wichtigsten Ungleichungen für die Variationsrechnung.

Satz 12.25 (Poincaré-Ungleichung)
Sei

$$X = \{v \in H^1 : v(a) = 0\}.$$

Dann gibt es eine Konstante $C > 0$ mit

$$||v||_{L^2}^2 \leq C||v'||_{L^2}^2 \, \forall v \in X.$$

Beweis: Mit Cauchy–Schwarz:

$$\begin{aligned}(v(t))^2 &= \left(\int_a^t v'(s)\mathrm{d}s\right)^2 \\ &\leq \int_a^t 1\mathrm{d}s \cdot \int_a^t v'(s)^2\mathrm{d}s \\ &\leq (t-a)||v'||_{L^2}^2\end{aligned}$$

und damit

$$\begin{aligned}||v||_{L^2}^2 = \int_a^b v(t)^2\mathrm{d}t &\leq \int_a^b (t-a)||v'||_{L^2}^2\mathrm{d}s\mathrm{d}t \\ &= \frac{1}{2}(b-a)^2||v'||_{L^2}^2.\end{aligned}$$

□

12.5 Existenz– und Eindeutigkeitssatz für das Sturm–Liouville–Modellproblem

Satz 12.26 (Lösbarkeit des Sturm–Liouville–Problems)
Sei

$$X := \left\{y \in H^1 : y(a) = 0\right\}$$

versehen mit der Norm

$$||\cdot||_X := ||\cdot||_{H^1}.$$

Sei wieder wie in 12.14

$$B(v,w) = \int_a^b p(t)v'(t)w'(t) + q(t)v(t)w(t)dt \,\forall v, w \in X$$

und

$$F(v) = \int_a^b f(t)v(t)dt$$

sowie

$$I(v) = \frac{1}{2}B(v,v) - F(v).$$

Dann hat I in X einen eindeutigen Minimierer, und damit das Randwertproblem eine (schwache) Lösung.

Mit dem Darstellungssatz von Riesz aus der Funktionalanalysis zeigt man leicht das Lemma von Lax–Milgram (z.B. in Alt [2007]), und hieraus folgt mit den gezeigten Sätzen die eindeutige Lösbarkeit des Variationsproblems. Wir zeigen den Satz zu Fuß.

Beweis: Wir beweisen zunächst: $B(u,v)$ ist ein Skalarprodukt auf X. Bilinearität usw. sind offensichtlich für B erfüllt. Bleibt zu zeigen:

$$||v||_B^2 := B(v,v) = 0 \Longleftrightarrow v = 0$$

und damit ist $||v||_B$ die zugehörige Norm. Nach Voraussetzung an p und q gilt

$$||v||_B^2 \geq p_0||v'||_{L^2}^2$$

also mit 12.25

$$||v||_B = 0 \Rightarrow v = 0.$$

Damit ist B ein Skalarprodukt und $||\cdot||_B$ eine Norm. Nach Definition von B gilt

$$||v||_B^2 \leq ||p||_\infty||v'||_{L^2}^2 + ||q||_\infty||v||_{L^2}^2 \leq \max(||p||_\infty, ||q||_\infty)\,||v||_{H^1}^2.$$

Mit Poincaré (12.25) gilt aber auch

$$\begin{aligned} ||v||_{H^1}^2 &= ||v||_{L^2}^2 + ||v'||_{L^2}^2 \\ &\leq (C+1)||v'||_{L^2}^2 \\ &\leq \frac{C+1}{p_0}||v||_B^2. \end{aligned}$$

Damit ist $||\cdot||_B$ eine zu $||\cdot||_{H^1}$ äquivalente Norm. Der Konvergenzbegriff bezüglich der Normen ist der gleiche, insbesondere ist X vollständig bezüglich beider Normen. Wir zeigen nun, dass $I(v)$ nach unten beschränkt ist. Mit Cauchy-Schwarz und Poincaré (12.25) gilt

$$\begin{aligned} I(v) &= \frac{1}{2}||v||_B^2 - \int_a^b f(t)v(t)\,dt \\ &\geq \frac{p_0}{2}||v'||_{L^2}^2 - ||f||_{L^2} \cdot ||v||_{L^2} \\ &\geq \frac{p_0}{2}||v'||_{L^2}^2 - C||f||_{L^2} \cdot ||v'||_{L^2}. \end{aligned}$$

Die rechte Seite ist eine quadratische Funktion in $||v'||_{L^2}$, $p_0 > 0$, also ist I nach unten beschränkt und hat ein Infimum $I_0 > -\infty$.
Noch zu zeigen: Das Infimum wird angenommen.
Für jedes $k > 0$ gibt es ein $v_k \in V$, so dass

$$I_0 \leq I(v_k) < I_0 + 1/k,$$

denn I_0 ist Infimum von I. Insbesondere gilt für diese Folge

$$I(v_k) \to I_0.$$

Wir wollen zeigen, dass v_n eine Cauchyfolge ist. Dazu gibt es einen Standardtrick aus der Linearen Algebra mit Hilfe der Parallelogrammidentität (siehe z.B. Alt [2007], S. 97)

$$||u+v||_B^2 + ||u-v||_B^2 = 2||u||_B^2 + 2||v||_B^2.$$

Es gilt

$$\begin{aligned} \frac{p_0}{1+C^2}||v_n - v_m||_{H^1}^2 &\leq ||v_n - v_m||_B^2 \\ &= 2||v_n||_B^2 + 2||v_m||_B^2 - ||v_n + v_m||_B^2 \\ &= 4(\frac{1}{2}||v_n||_B^2 - F(v_n)) + 4(\frac{1}{2}||v_m||_B^2 - F(v_m)) \\ &- 8(\frac{1}{2}\left|\left|\frac{v_n+v_m}{2}\right|\right|_B^2 - F\left(\frac{v_n+v_m}{2}\right)) \\ &= 4I(v_n) + 4I(v_m) - 8I\left(\frac{v_n+v_m}{2}\right) \\ &\leq 4I(v_n) + 4I(v_m) - 8I_0 \\ &\to_{n,m\to\infty} 0. \end{aligned}$$

Also ist v_n eine Cauchyfolge und konvergiert gegen ein $v \in X$, denn X ist vollständig.
Noch zu zeigen:

$$I(v) = I_0.$$

B und F sind stetig, denn mit Cauchy–Schwarz gilt

$$|B(u,v)| \leq ||p||_\infty ||u'||_{L^2} ||v'||_{L^2} + ||q||_\infty ||u||_{L_2} ||v||_{L_2}$$

und

$$|F(v)| \leq ||q||_{L^2} ||f||^2_{L^2} ||v||^2_{L^2}.$$

Also gilt

$$\begin{aligned} I(v) - I(v_n) &= \frac{1}{2}(B(v,v) - B(v_n,v_n)) + F(v - v_n) \\ &= \frac{1}{2}((B(v,v) - B(v_n,v)) + (B(v_n,v) - B(v_n,v_n)) + F(v - v_n) \\ &\to_{n\to\infty} 0 \end{aligned}$$

und damit $I(v) = I_0$. □

12.6 Numerische Verfahren für variationelle Probleme

Nun können wir ein numerisches Verfahren zur Lösung von Randwertproblemen für gewöhnliche Differentialgleichungen mit Variationsmethoden definieren. Dazu fordern wir die variationelle Bedingung 12.16 nicht auf ganz X, sondern suchen einen Minimierer y_h in einem endlichdimensionalen Teilraum X_h von X.

Unsere Idee ist: Je genauer die Funktionen in X_h die Funktionen in X approximieren, umso besser sollte auch y_h die gesuchte Lösung y approximieren. Beachten Sie: Wir betrachten dabei die Approximationen weiterhin als *Funktionen*, wir betrachten ausdrücklich keine Gitterfunktionen wie bei den reinen Diskretisierungsansätzen. Dadurch bleibt unser komplettes analytisches Arsenal erhalten, und etwa die Existenz eines Minimums y_h ist trivial.

Definition 12.27 (+ Satz) Ritz–Galerkin–Verfahren
Seien B, F, I, X definiert wie in 12.14, also insbesondere B ein Skalarprodukt auf X und F ein lineares Funktional auf X.

Sei X_h ein endlichdimensionaler Teilraum von X. Dann ist die Einschränkung von B auf X_h ein Skalarprodukt auf X_h, und die Einschränkung von F auf X_h ist stetig.

Die Voraussetzungen von 12.26 sind erfüllt, d.h. die Einschränkung von I auf X_h hat einen eindeutigen Minimierer $y_h \in X_h$.

$y_h \in X_h$ heißt Galerkin–Lösung und erfüllt die Gleichung

$$B(y_h, v) = F(v) \, \forall \, v \in X_h.$$

Die approximative Lösung y_h lässt sich leicht berechnen. Sei $v_0, \dots v_{n-1}$ eine Basis von X_h, damit gibt es eine eindeutige Darstellung

$$y_h = \sum_{j=0}^{n-1} \alpha_j v_j.$$

B ist linear im ersten und zweiten Argument, F ist linear, es reicht daher y_h so zu wählen dass

$$B(v_k, y_h) = \sum_{j=0}^{n-1} B(v_k, v_j)\alpha_j = F(v_k), \, \forall k = 0 \dots n-1.$$

Die α_j erfüllen also ein lineares Gleichungssystem. Wir setzen

$$B \in \mathbb{R}^{n \times n}, \, B_{k,j} = B = (v_k, v_j), \, F \in \mathbb{R}^n, F_k = F(v_k), \, \alpha \in \mathbb{R}^n, \, (\alpha)_k = \alpha_k$$

und erhalten das lineare Gleichungssystem

$$B\alpha = F.$$

Da es immer ein y_h gibt, das das Gleichungssystem löst, ist B surjektiv und damit invertierbar.

Mit Hilfe der Sätze können wir leicht die Konvergenz des Galerkin–Verfahrens in der Supremumsnorm nachweisen.

Satz 12.28 (Konvergenz des Galerkin–Verfahrens und Fehlerabschätzung)
Sei (X_h) eine Folge von Räumen mit

$$\min_{x \in X_h} ||y - x||_{H^1} = d(y, X_h) \to 0 \, \forall y \in X.$$

Dann konvergiert y_h gegen y bezüglich der Supremumsnorm.
Es gibt eine Konstante C, die nicht von h abhängt, mit

$$||y - y_h||_\infty \leq C d(y, X_h).$$

Beweis: Sei y die Lösung des Variationsproblems auf X, y_h die Lösung auf X_h. Es gilt

$$B(y_h, v) = F(v) = B(y, v) \Rightarrow B(y_h - y, v) = 0 \, \forall v \in X_h.$$

B ist ein Skalarprodukt, also gilt mit dieser Beziehung und Cauchy–Schwarz

$$\begin{aligned} ||y - y_h||_B^2 &= B(y - y_h, y - y_h) \\ &= B(y - y_h, y) && y_h \in X_h \\ &= B(y - y_h, y - v) && v \in X_h \text{ beliebig} \\ &\leq ||y - y_h||_B \cdot ||y - v||_B. \end{aligned}$$

Falls $y = y_h$, so sind wir sowieso fertig. Andernfalls gilt

$$||y - y_h||_B \leq ||y - v||_B.$$

Da die B–Norm und die H^1–Norm äquivalent sind, gilt

$$C_1||y - y_h||_{H^1} \leq ||y - y_h||_B \leq ||y - v||_B \leq C_2||y - v||_{H^1}$$

oder mit $C = C_2/C_1$

$$||y - y_h||_{H^1} \leq C||y - v||_{H^1}.$$

Da $v \in X_h$ beliebig war, folgt schon mal

$$||y - y_h||_{H^1} \leq C \inf_{v \in X_h} ||y - v||_{H^1} = Cd(y, X_h)$$

und damit

$$y_h \to_{H^1} y.$$

Dies ist noch nicht ganz das Gewünschte: Das ist nur eine Konvergenz bezüglich der H^1–Norm.
Aber mit den bewiesenen Sätzen gilt

$$\begin{aligned} ||y - y_h||_\infty &\leq C'||y - y_h||_{H^1} && \text{(Sobolev 12.23)} \\ &= C\,C'\,d(y, X_h). \end{aligned}$$

□

Überzeugend an dieser Vorgehensweise ist, dass dieser Beweis deutlich eleganter ist als die Beweise für die Diskretisierung etwa bei den M–Matrizen: Die Konvergenz gilt unabhängig von der Art der Diskretisierung für alle Sturm–Liouville–Probleme, die die Voraussetzungen erfüllen.

Der Satz sagt: Die Qualität der Näherungen y_h hängt davon ab, wie gut die Teilräume X_h den Raum X bezüglich der H^1–Norm approximieren. Wir haben die Aufgabe

der Lösung der Differentialgleichung auf die Aufgabe, die Approximationsgüte von linearen Unterräumen zu bestimmen, zurückgeführt und damit von der Differentialgleichung (bis auf eine Konstante) abgekoppelt.

Im Kapitel über Interpolation haben wir unter anderem die folgenden Möglichkeiten kennengelernt, Funktionen in X zu approximieren:

1. Polynome: Wir wählen als $X_{1/n}$ den Polynomraum $\mathcal{P}_n$ mit $p(a) = 0$.
2. Trigonometrische Polynome (Fouriertransformation): Hier wählen wir als Ansatzfunktionen $\sin(k(t-a))$ und $\cos(kt)$.
3. Splines (Kapitel 8.2): Diese haben die besten Approximationseigenschaften. Sie sind das Standardwerkzeug zur Lösung von Variationsproblemen.

Nach den Vorbemerkungen müssen wir bei der Lösung der Variationsgleichung folgende Schritte erledigen.

1. Wähle einen endlichdimensionalen Teilraum X_h von X.
2. Bestimme eine Basis $v_0 \dots v_n$ von X_h.
3. Berechne die Matrizen B (Steifigkeitsmatrix) und F (Lastvektor).
4. Löse das Gleichungssystem $B\alpha = F$.
5. Setze $y_h = \sum_{j=0}^{n} \alpha_j v_j$.

Die Schritte 3 und 4 können sehr aufwändig sein. Es wäre daher günstig, wenn B dünn besetzt wäre. Dies ist für Splines der Fall.

Wir wollen uns hier als Beispiel die Lösung der Poissongleichung auf dem Intervall $[0,1]$ als variationelles Problem mit Splines der Ordnung 2 anschauen. Dieser Splineraum besteht aus den stückweise linearen, stetigen Funktionen auf einem Intervall. Sei $x_0 \dots x_n$ ein äquidistantes Gitter auf $[0,1]$ mit der Gitterweite $h = 1/n$. Wir setzen

$$\varphi_k(x_j) = \begin{cases} 1 & j = k \\ 0 & \text{sonst} \end{cases}$$

und linear interpoliert zwischen diesen Punkten. Dann ist $\varphi_1 \dots \varphi_n$ eine Basis des Raums der Splines der Ordnung 2 mit $\varphi(0) = 0$. Der Träger der φ_k ist (x_{k-1}, x_{k+1}), und es gilt

$$\varphi_k'(t) = \begin{cases} \frac{1}{h} & t \in (x_{k-1}, x_k) \\ -\frac{1}{h} & x \in (x_k, x_{l+1}) \\ 0 & \text{sonst} \end{cases}$$

Für die Poissongleichung ist $p = 1$ und $q = 0$. Falls $|j - k| > 1$, so überschneiden sich die Träger von φ_j und φ_k nicht, d.h. es gilt

$$B_{k,j} = B(\varphi_k, \varphi_j) = 0$$

und damit besitzt B nur Einträge auf der Hauptdiagonalen und den beiden Nebendiagonalen. Durch Einsetzen erhält man

$$B_{k,k} = \frac{2}{h},\ B_{k,k+1} = B_{k+1,k} = -\frac{1}{h}.$$

Weiter gilt

$$F_k = F(v_k) = \int_{x_{k-1}}^{x_{k+1}} f(t)v_k(t)\,dt \sim hf(x_k)$$

Das zu lösende Gleichungssystem und damit die Approximation y_h ist also am Ende genau die gleiche wie in 12.2. Dies scheint zunächst ein frustrierendes Ergebnis zu sein. Aber durch die völlig andere Herleitung und Sichtweise bekommen wir hier die Konvergenz und alles andere geschenkt.

Der funktionale Zugang über die Variationsrechnung ist also in diesem Fall der mathematisch bessere. In höheren Dimensionen, d.h. bei (partiellen) Differentialgleichungen in mehreren Veränderlichen, ist er besonders praktisch. Dort werden für die Ansatzfunktionen zu Splines äquivalente Konstrukte in höheren Dimensionen genutzt, Finite Elemente (siehe z.B. Braess [2007]). Die dort genutzten Sätze und Konstruktionen sind dabei exakt dieselben wie hier für eine Dimension.

12.7 Zusammenfassung

12.7.1 Kompetenzen

- Definition eines Randwertproblems zweiter Ordnung kennen.
- Einfache Lösungsverfahren kennen (Schießverfahren, Diskretisierung führt auf lineare Gleichungssysteme).
- Definition der Konsistenz für diskrete Verfahren kennen (und wissen, dass sie nur in Spezialfällen nachgerechnet werden kann).
- Idee der Umwandlung der Sturm–Liouville–Probleme in ein Variationsproblem/Minimierungsproblem kennen.
- Erweiterung des Lösungsraums kennen: $C^2 \subset C^1 \subset H^1$.
- Schwache Differenzierbarkeit für Funktionen ausrechnen können.

- Sobolevsche Ungleichung, Sobolevschen Einbettungssatz, Poincaré–Ungleichung (in ihrer trivialen Interpretation auf den reellen Zahlen) kennen.
- Analytische Folgerung kennen: Sturm–Liouville mit natürlichen Randbedingungen hat eine eindeutige Lösung.
- Idee von Ritz–Galerkin und Fehlerabschätzung kennen.
- Ritz–Galerkin mit Spline–Ansatzfunktionen durchführen können.

12.7.2 Mini–Aufgaben

Siehe Übungen.

Kapitel 13

Errata

- Abschnitt 1.2: Proportionalitätsfaktor war mal λ und mal τ, ist jetzt immer λ.
- Bemerkung zu Satz 2.5: Autonome statt Exakte.
- Beispiel 2.4: $\arctan y$ statt $\arctan z$. $\tan(t + C)$ statt $\tan t + C$.
- Kapitel 3: Zugelassen, dass Kontraktionskonstante und Lipschitzkonstante 0 sind. Spielt eigentlich keine Rolle, aber wenn man positiv voraussetzt, gibt es u.U. keine kleinste Konstante.
- Beispiel 1.1: Integrationsvariable und Grenze war t, Integrationsvariable zu s geändert.
- Beispiel 1.3: Im Federmodell einiges klarer gefasst.
- Abschnitt 2.4: $\operatorname{sgn}(f(y_0))$ statt $\operatorname{sgn}(y_0)$.
- Beispiel 2.3: t statt T.
- Banachscher Fixpunktsatz 3.3: A posteriori–Abschätzung umformuliert, so dass a priori und a posteriori Abschätzungen für $x^{(k)}$ liefern.
- Beispiel 3.12: In der Summe falscher Index $k- > j$
- Abschnitt 3: Zusammenfassung hinzugefügt.
- Satz 4.8: Fehlte die Bemerkung, dass g stetig fortsetzbar ist.
- Satz 5.28: Beweis hinzugefügt, Satz 5.29 entfernt.
- Satz 2.8: Bemerkung zur Lösung der AWA hinzugefügt.
- Aufgaben zu Kapitel 6: Aufgabe zu höherer Ordnung erweitert.

- Beispiel 6.17: Hier fehlte an einer Stelle ein Minuszeichen, $3 \to -3$.
- Kapitel 10.2: Hier war am Anfang fälschlich $\varphi(t, y, h)$ eingesetzt. h gestrichen.
- Kapitel 10.4: Definition der Runge–Kutta–Verfahren vereinfacht.
- Kapitel 10.4: Beweise und Formulierungen vereinfacht.
- Definition 4.6, Landau–Symbole: Hier fehlte in der Definition ein C.
- Kapitel 12/12.1/12.2: Redaktionelle Änderungen zum besseren Verständnis.
- Definition 4.6: In Teil 2 stand k statt $(k + 1)$.
- Kapitel 12.6: Viele redaktionelle Änderungen zum besseren Verständnis.

13.1 Zusammenfassung

13.1.1 Kompetenzen

-

13.1.2 Mini–Aufgaben

-

Literaturverzeichnis

M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables : [is an Outgrowth of a Conference on Mathematical Tables Held at Cambridge, Mass., on 1954*. Applied mathematics series. Dover Publ., 1965. ISBN 9780486612720. URL http://people.math.sfu.ca/~cbm/aands/abramowitz_and_stegun.pdf.

H.W. Alt. *Lineare Funktionalanalysis:*. Springer London, Limited, 2007. ISBN 9783540341871. URL http://books.google.de/books?id=TzeMiSx8_l4C.

S. Bosch. *Lineare Algebra*. Springer-Lehrbuch. Springer Berlin Heidelberg, 2014. ISBN 9783642552601. URL https://books.google.de/books?id=feAoBAAAQBAJ.

D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Verlag Berlin Heidelberg, 2007. ISBN 9783540724506. URL http://books.google.de/books?id=s1M-jAqi1sYC.

Roland Bulirsch and Josef Stoer. Numerical treatment of ordinary differential equations by extrapolation methods. *Numerische Mathematik*, 8(1):1–13, 1966. doi: 10.1007/978-3-540-45390-1. URL https://link.springer.com/book/10.1007/978-3-540-45390-1.

L.C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics. American Mathematical Society, 2010. ISBN 9780821849743. URL http://books.google.de/books?id=Xnu0o_EJrCQC.

John Harrison. Formal verification of ia-64 division algorithms. In *Proceedings of the 13th International Conference on Theorem Proving in Higher Order Logics*, TPHOLs '00, pages 233–251, London, UK, UK, 2000. Springer-Verlag. ISBN 3-540-67863-8. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.7123&rep=rep1&type=pdf.

T. Kato. *Perturbation Theory of Linear Operators*. Classics in mathematics. Springer, 1995. URL http://books.google.de/books?id=zzNqMgEACAAJ.

F. Natterer. *The Mathematics of Computerized Tomography*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2001. ISBN 9780898714937. URL http://books.google.de/books?id=gjSO1hLbcD0C.

F. Natterer and F. Wübbeling. *Mathematical Methods in Image Reconstruction*. Monographs on Mathematical Modeling and Computation, No 5 Series. Society for Industrial & Applied, 2001. ISBN 9780898714722. URL http://books.google.de/books?id=u8W9I32wNRMC.

Carl Runge and Hermann König. *Vorlesungen über numerisches Rechnen*. Springer Göttingen, 1925. URL http://resolver.sub.uni-goettingen.de/purl?PPN373207646.

Wolfgang Walter. *Gewöhnliche Differentialgleichungen: Eine Einführung*. 2013. URL https://link.springer.com/book/10.1007/978-3-642-57240-1.

Frank Wübbeling. *Skript zur Vorlesung Numerische Lineare Algebra*. 2022. URL https://www.uni-muenster.de/AMM/num/Vorlesungen/wuebbeling/.

Abbildungsverzeichnis

Listings