Praxisorientierte Einführung in die Numerik im Sommersemester 2020
Lernskript
Stand 29. Juni 2020, 10:08:21

Frank Wübbeling

29. Juni 2020

Inhaltsverzeichnis

Vorwort

Dieses Skript entsteht zur Online–Vorlesung Praxisorientierte Einführung in die Numerik. Es stellt ausgewählte Kapitel der Vorlesungen Numerische Lineare Algebra und Numerische Analysis im Anwendungszusammenhang vor.

Hauptthema der Vorlesung ist die numerische Behandlung gewöhnlicher Differentialgleichungen. Auf dem Weg zu diesem Ziel führen wir zunächst grundlegende Techniken zur Numerik ein: Fehlerrechnung, (direkte und iterative) Lösung von linearen und nichtlinearen Gleichungssystemen, Berechnung von Eigenwerten, Interpolation, numerische Integration und Differentiation.

Achtung: Dies ist ein Lernskript und fasst die wesentlichen Ideen (Definitionen und Sätze) der Vorlesung pro Veranstaltung zusammen. Beweise, Motivationen und Hintergründe finden Sie jeweils in der angegebenen Hintergrundliteratur, am einfachsten in meinen Skripten zur Numerischen Linearen Algebra und zur Numerischen Analysis.

Kapitel 1

Hauptbeispiele der Vorlesung und einfache numerische Verfahren

1.1 Differentialgleichungen: Grundbegriffe

Für die Hauptbeispiele benötigen wir Differentialgleichungen. Wir erinnern an die Definitionen.

EIne Differentialgleichung ist eine Gleichung, in der eine Ableitung einer (unbekannten) Funktion y auftaucht. Lösung der Differentialgleichung ist eine konkrete Funktion y, die die Gleichung erfüllt. **Lösung einer Differentialgleichung ist immer eine Funktion.**

y kann eine Funktion in einer Variablen (gewöhnliche Differentialgleichung) oder in mehreren Variablen (partielle Differentialgleichung) sein. Wir werden im Rahmen dieser Vorlesung ausschließlich gewöhnliche Differentialgleichungen betrachten.

Ausdrücklich zugelassen ist, dass y eine Funktion in den $\mathbb{R}^n$ bzw. $\mathbb{C}^n$ ist. Falls $n = 1$, so heißt die Gleichung skalar.

Für $n > 1$ haben wir n Koordinatenfunktionen, entsprechend müssen n Gleichungen gegeben sein. Die Gleichungen heißen System von Differentialgleichungen.

Es ist zugelassen, dass höhere Ableitungen von y in der Gleichung auftauchen. Die Ordnung der höchsten Ableitung heißt Grad der Differentialgleichung.

Einige Beispiele:

1.
$$y : [a, b] \mapsto \mathbb{R},\ y'(t) = cy(t)$$
(Lineare) skalare gewöhnliche Differentialgleichung 1. Ordnung. Eine Lösung ist
$$y(t) = e^{ct}.$$

2.
$$y : [a, b] \mapsto \mathbb{R},\ y'(t) = c(t)y(t),\ c : [a, b] \mapsto \mathbb{R}$$
(Allgemeine lineare) skalare gewöhnliche Differentialgleichung 1. Ordnung. Eine Lösung ist
$$y(t) = e^{\int_a^t c(s)\,ds}.$$

3.
$$y : [a, b] \mapsto \mathbb{R},\ y''(t) = -y(t)$$
Skalare Differentialgleichung 2. Ordnung. Lösungen sind $\cos t$ und $\sin t$.

4.
$$y : [a, b] \mapsto \mathbb{R},\ y'(t) = 1 + y(t)^2$$
Skalare Differentialgleichung 1. Ordnung. Eine Lösung ist $\tan x$.

5.
$$y = (H(t), F(t)),\ H, F : [a, b] \mapsto \mathbb{R},\ y : [a, b] \mapsto \mathbb{R}^2$$
$$\begin{aligned} H'(t) &= aH(t) - cH(t)F(t) \\ F'(t) &= -bF(t) + cH(t)F(t) \end{aligned}$$
oder
$$y'(t) = (H'(t), F'(t)) = (aH(t) - cH(t)F(t), -bF(t) + cH(t)F(t)) =: f(y(t))$$
System von 2 Differentialgleichungen in 2 Funktionen $(F(t), H(t))$ oder in der 2–dimensionalen Funktion y.
(Räuber–Beute–Modell, Lotka–Volterra–Differentialgleichung)

6.
$$y : [a, b] \times \mathbb{R}^2 \mapsto \mathbb{R},\ \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)(t, x, y) = \frac{\partial u}{\partial t}(t, x, y)$$
Partielle Differentialgleichung 2. Ordnung in 2 Dimensionen (Wärmeleitungsgleichung), siehe 2. Hauptbeispiel.

7.

$$y : [a, b] \mapsto \mathbb{R},\ u''(x) = -s(x)$$

Skalare stationäre Wärmeleitungsgleichung in einer Dimension, siehe 2. Hauptbeispiel.

Die Lösungen von Differentialgleichungen sind im Allgemeinen nicht eindeutig. Wir brauchen daher noch zusätzliche Bedingungen. Wir betrachten zwei Bedingungen, die zu völlig unterschiedlichen numerischen Algorithmen führen.

1. Anfangswertaufgaben: Hier ist ein y_0 gegeben mit $y(a) = y_0$. Für Systeme von Differentialgleichungenn muss also für jede Koordinatenfunktion ein Anfangswert vorgegeben werden, im Räuber–Beute–Modell müssen also $F(a)$ und $H(a)$ vorgegeben sein.
2. Randwertprobleme: Für gewöhnliche Differentialgleichungen höherer Ordnung oder Systeme von gewöhnlichen Differentialgleichungen sind häufig Bedingungen bei a und b, also auf dem gesamten Rand, vorgegeben. In diesem Fall sprechen wir von Randwertproblemen.
   Beispiel: Skalare stationäre Wärmeleitungsgleichung mit $y(a) = y(b) = 0$ (siehe 2. Hauptbeispiel).

Für Anfangswertprobleme kann man die Aufgabe so formulieren: Bei $t = a$ ist die Funktion bekannt. Setze die Funktion fort unter Berücksichtigung der Differentialgleichung. Mit dem Satz von Picard–Lindelöf gilt: Diese Fortsetzung ist zumindest auf einem Teilintervall von $[a, b]$ möglich, falls f Lipischitz–stetig ist in der zweiten Variablen.

Bei Randwertproblemen ist dies nicht möglich, weil man keinen Anfangspunkt hat, sondern die Bedingungen von links und rechts beachten muss. Man erhält ein (lineares) Gleichungssystem, in dem der linke und rechte Randwert vorkommen.

Wir betrachten bei Anfangswertproblemen nur (Systeme von) Gleichungen erster Ordnung und nehmen an, dass die einzelnen Gleichungen jeweils nach y'_k aufgelöst sind (Normalform). Die einzelnen Gleichungen sind also von der Form

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

mit

$$y(t) = (y_1(t), \ldots, y_n(t)),\ y'(t) = (y'_1(t), \ldots, y'_n(t)),\ f : [a, b] \times \mathbb{R}^n \mapsto \mathbb{R}^n.$$

Alle Beispiele oben sind bereits in dieser Form gegeben, mit Ausnahme der stationären Wärmeleitungsgleichung, denn diese ist von der Ordnung 2. Tatsächlich

lässt sich mit einem einfach Trick jede Gleichung höherer Ordnung in ein System von Gleichungen erster Ordnung umwandeln. Am Beispiel der stationären Wärmeleitungsgleichung:

Wir führen die zusätzliche Funktion $v(x) = u'(x)$ ein. Dann erhalten wir das System von Gleichungen:

$$\begin{aligned} v'(x) = u''(x) &= -s(x) \\ u'(x) &= v(x) \end{aligned}$$

und dieses ist in Normalform, denn alle Gleichungen sind von erster Ordnung und jeweils nach der ersten Ableitung aufgelöst. Wir können uns also tatsächlich bei der Betrachtung auf die Normalform beschränken.

1.2 Euler–Verfahren

Das Euler–Verfahren ist das einfachste numerische Verfahren zur Lösung von Anfangswertaufgaben. Wir wollen es auf zwei Arten motivieren. Sei dazu das skalare Anfangswertproblem

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

gegeben und zu lösen auf dem Intervall $[a, b]$.

Graphische Lösung: Wir kennen also den Wert der Funktion $y(t)$ an der Stelle $t = a$ und wollen sie nach rechts fortsetzen. Da y die Differentialgleichung erfüllllt, gilt auch

$$y'(a) = f(a, y(a)).$$

Wir kennen also auch die Ableitung der Funktion an der Stelle $t = a$. Die Ableitung ist die Tangentensteigung, also kennen wir die Tangente an $y(t)$ im Punkt $(a, y(a))$. Die Geradengleichung der Tangente ist

$$z(t) = y_0 + (t - a) \cdot y'(a).$$

In einer sehr kleinen Umgebung von a stimmen die Tangente z und die Kurve y fast überein, d.h. dort ist die Tangente eine gute Näherung für die Funktion (siehe Abbildung 1.1).

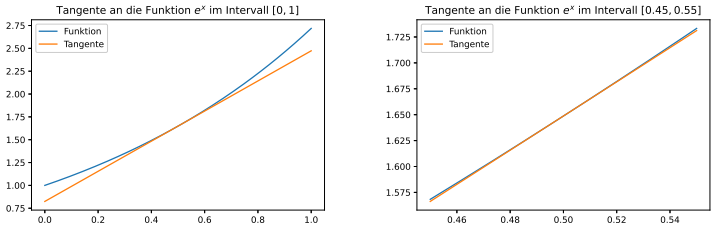


Abbildung 1.1: Tangente an e^x im Punkt $x = 0.5$.
Auf dem kleinen Intervall ist die Tangente eine gute Approximation für e^x.

Das Eulerverfahren geht nun so vor: Im Intervall werden $(N-1)$ Punkte $t_1, \dots t_{N-1}$ zwischen $t_0 = a$ und $t_N = b$ mit gleichem Abstand $h = (b-a)/N$ (äquidistant) eingefügt. Es gilt also $t_k = a + kh$.

An den Punkten t_k, $k = 0 \dots N$, wollen wir Näherungen y_k für $y(t_k)$ berechnen.

Wir gehen wie geplant von links nach rechts vor. Für $k = 0$ wissen wir $y(t_0) = y(a) = y_0$. Um eine Näherung für y an der Stelle t_1 auszurechnen, stellen wir die Tangentengleichung auf und setzen den Punkt t_1 ein, verfolgen also graphisch die Tangente ein kurzes Stück. Nach unserer Rechnung vorhin gilt

$$y_1 = z(t_1) = z(a+h) = y_0 + hy'(a) = y_0 + hf(t_0, y_0).$$

Sei nun y_k bereits ausgerechnet. Zur Berechnung von y_{k+1} berechnen wir die Tangente an die y, falls sie wirklich durch den Punkt (t_k, y_k) geht, also:

$$z(t) = y_k + (t - t_k)f(t_k, y_k)$$

und setzen t_{k+1} ein. Wir erhalten

$$y_{k+1} = z(t_{k+1}) = y_k + (t_{k+1} - t_k)f(t_k, y_k) = y_k + hf(t_k, y_k).$$

Dies ist das Eulerverfahren. Wir erhalten als Ergebnis einen Vektor mit Zahlen $y_0, \dots, y_N$. Die Zahlen sind nach unserer Motivation Approximationen für die Funktion y an den Stellen $t_0, \dots, t_N$.

Dies ist natürlich kein Beweis, sondern nur eine Motivation. Im Kapitel über Differentialgleichungen werden wir zeigen, dass dieses Verfahren tatsächlich eine Approximation liefert und Abschätzungen für den Fehler angeben.

Alternative Herleitung: Angenommen, wir wissen, dass $y_k = y(t_k)$. Dann ist mit Taylorentwicklung

$$y(t_{k+1}) = y(t_k + h) = y_k + y'(t_k)h + R,$$

wobei wir die Taylorreihe nach dem zweiten Glied abgebrochen und durch ein Restglied R ersetzt haben. Unter vernünftigen Annahmen ist dieses Restglied klein für $h \mapsto 0$ und gut abschätzbar (Restgliedabschätzungen) mit Formeln der Analysis 1. Daher ist $y_k + y'(t_k)h = y_k + hf(t_k, y_k)$ eine gute Abschätzung für $y(t_{k+1})$, und es macht Sinn, zu definieren

$$y_{k+1} = y_k + hf(t_k, y_k).$$

1.3 Federbeispiel

Es sei $s(t)$ die Auslenkung einer Feder mit Federkonstante c, an die ein Gewicht der Masse m gehängt wird. Dann erfüllt s mit der Gravitationskonstanten g die Differentialgleichung

$$s''(t) = -g - \frac{c}{m}s(t).$$

Diese Gleichung berücksichtigt keine Dämpfung. Mit einer Dämpfungsfunktion d:

$$s''(t) = -g - \frac{c}{m}s(t) - \frac{d(s'(t))}{m}.$$

d kann linear, aber auch nichtlinear sein.

Für das 3–Feder–Beispiel von Bollhöfer und Mehrmann erhalten wir ein System aus drei Differentialgleichungen zweiter Ordnung, also nach Umstellung auf die Normalform ein System aus sechs Differentialgleichungen erster Ordnung.

Wir nehmen jeweils an, dass wir die Situation zum Anfangszeitpunkt komplett kennen (Positionen und Geschwindigkeit). Dies sind Beispiele für Systeme von Anfangswertaufgaben.

1.4 Stationäre Wärmeleitungsgleichung

Es sei $U(t, x)$ die Wärmeverteilung in einem Draht, der an den Enden a und b auf 0 Grad gekühlt und in der Mitte zeitkonstant erwärmt wird mit der Hitzezufuhr $s(x)$. Dann stellt sich nach einiger Zeit eine zeitkonstante Wärmeverteilung $u(x)$ ein.

u erfüllt das Randwertproblem $-u''(x) = s(x)$, $u(a) = u(b) = 0$.

Numerische Lösungsidee: Wie bei den Anfangswertaufgaben fügen wir wieder äquidistante Zwischenpunkte $x_1, \ldots, x_{N-1}$ zwischen a und b ein. Die Punkte haben den

Abstand $h = (b-a)/N$. Es gilt

$$\lim_{\epsilon \mapsto 0} \frac{u(x+\epsilon) - 2u(x) + u(x-\epsilon)}{\epsilon^2} = u''(x).$$

Ansatz über l'Hospital:

$$\frac{u(x+\epsilon) - 2u(x) + u(x-\epsilon)}{\epsilon^2} \mapsto \frac{u'(x+\epsilon) - u'(x-\epsilon)}{2\epsilon} \mapsto u''(x).$$

Alternativ über Taylorentwicklung.

Für kleine h gilt also

$$\frac{-u(x+h) + 2u(x) - u(x-h)}{h^2} \sim -u''(x) = s(x).$$

Wir wollen nun Näherungen u_k für $u(x_k)$ berechnen. Dazu fordern wir, dass diese Gleichung exakt erfüllt ist für die Näherungen. Wegen $x_k + h = x_{k+1}$ und $x_k - h = x_{k-1}$ erhalten wir damit für $k = 1 \dots N-1$ die Gleichungen

$$s(x_k) = \frac{-u_{k+1} + 2u_k - u_{k-1}}{h^2}.$$

Wir multiplizieren jetzt noch diese Gleichung mit h^2 und erhalten

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

Da die Temperatur an den Rändern Null ist, können wir setzen $u_0 = u_N = 0$. Wir erhalten also ein lineares Gleichungssystem in den Variablen $u_1, \dots, u_{N-1}$.

Dieses Gleichungssystem wird beschrieben über die Matrixgleichung $Au = b$ mit der Matrix

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

Zu zeigen ist jetzt natürlich, dass diese Matrix invertierbar ist usw. Diese Matrix, die das Anwendungsproblem der Verteilung der Wärme in einem Draht repräsentiert, wird uns als Beispiel bei der Lösung linearer Gleichungssysteme dienen.

1.5 Google–Matrix

Google bestimmt mit Hilfe dieser Matrix die Wichtigkeit von Webseiten. Die Größe ist (Anzahl der Webseiten)$\times$(Anzahl der Webseiten). Die Matrix enthält nur nichtnegative Einträge, die Spaltensumme ist 1 für jede Spalte (stochastische Matrix). Bei der eigentlichen Google–Matrix sind die Hauptdiagonalelemente 0, bei der modifizierten $d > 0$, etwa $d \sim 0.6$.

Die Matrix hat den betragsmaximalen Eigenwert 1, die Vielfachheit des Eigenwerts im charakteristischen Polynom der Matrix ist ebenfalls 1. Zur Lösung des Problems muss ein Eigenvektor zun diesem Eigenwert 1 ausgerechnet werden.

Diese Matrix wird uns als Beispiel für Eigenwertprobleme dienen.

Kapitel 2

Referenz zur linearen Algebra

Wir stellen hier einmal die Grundbegriffe der linearen Algebra als Referenz kompakt zusammen, so wie wir sie nutzen werden. Wir beschränken uns grundsätzlich auf die Betrachtung von Vektorräumen über $K = \mathbb{R}$ oder $K = \mathbb{C}$, häufig der Einfachheit halber nur auf $\mathbb{R}$. Seien also im Folgenden immer U und V Vektorräume über K.

Hier werden keine Beweise oder ähnliches gegeben, sie sollten alles in Ihrem Lineare Algebra–Skript finden (oder zum Beispiel in dem von Siegfried Echterhoff unter https://ivv5hpp.uni-muenster.de/u/echters/Lineare-AlgebraI/Skript/).

2.1 Normierte Vektorräume

Grundlegend für alle numerischen Überlegungen ist der Begriff der Norm, denn nur so lassen sich Fehler messen.

Definition 2.1 *(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 2.2 *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).

Beispiel 2.3 *Sei $V = C^0(I)$ der Raum der stetigen Funktionen auf einer kompakten Teilmenge $I \subset \mathbb{R}^n$, $f \in V$.*

$$||f||_p = \left(\int_I |f(x)|^p dx \right)^{1/p} (p < \infty),\ ||f||_\infty = \sup_{x \in I} |f(x)|$$

heißt p–Norm (und ist eine Norm).

Definition 2.4 *(Banachraum) Sei $(V, ||\cdot||)$ normierter Vektorraum. V heißt vollständig oder Banachraum, falls jede Cauchyfolge in V einen Grenzwert in V besitzt (bzgl. $||\cdot||$).*

Beispiel 2.5
$(C^0(I), ||\cdot||_2)$ ist nicht vollständig.
$(C^0(I), ||\cdot||_\infty)$ ist vollständig.

Definition 2.6 *(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 induzierte Norm

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

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

Beispiel 2.7

1) Sei $V = \mathbb{C}^n$. Dann ist

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

ein Skalarprodukt.

2) Sei $V = C^0(I)$. Dann ist

$$(f, g) = \int_I f(x)\overline{g(x)}dx,\ f, g \in V$$

ein Skalarprodukt.

Wir werden beide stillschweigend als Standard–Skalarprodukte auf den jeweiligen Räumen verwenden. Die induzierte Norm ist jeweils $||\cdot||_2$.

Satz 2.8 *(Cauchy–Schwarz)*
Sei V ein Vektorraum mit Skalarprodukt. 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.

Die wichtigste Folgerung ist

Satz 2.9 *Sei V ein Vektorraum mit Skalarprodukt. Dann ist*

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

eine Norm.

Satz 2.10 *Sei V endlichdimensional und seien $||\cdot||$ und $|||\cdot|||$ zwei Normen auf V. Dann sind $||\cdot||$ und $|||\cdot|||$ äquivalent, d.h. $\exists\, C_1, C_2 > 0$:*

$$C_1|||v||| \leq ||v|| \leq C_2|||v||| \,\forall v \in V.$$

Eine wichtige Folgerung dieses Satzes ist: Wenn eine Folge in endlichdimensionalen Räumen konvergiert bezüglich einer Norm, so konvergiert sie gegen den gleichen Grenzwert bezüglich aller Normen.

2.2 Lineare Operatoren

Wir werden in dieser Vorlesung im wesentlichen Matrizen als Spezialfall linearer Operatoren untersuchen. Im folgenden steht T immer für eine allgemeine lineare Abbildung zwischen Vektorräumen, A für eine Abbildung zwischen endlichdimensionalen Räumen (die wir immer sofort mit einer Matrix identifizieren).

Definition 2.11 *(lineare Operatoren) Seien U, V Vektorräume. $T : U \mapsto V$ heißt linear genau dann, wenn*

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

Sind U und V endlichdimensional, so kann T durch eine Matrix A bezüglich vorgegebener Basen dargestellt werden. Falls $U = V$ und T in zwei verschiedenen Basen durch die Matrizen A und B dargestellt wird, so heißen A und B ähnlich, und es gibt eine Matrix X mit

$$A = XBX^{-1}.$$

Die Menge aller linearen Operatoren $L(U, V)$ bildet auf natürliche Weise selbst wieder einen Vektorraum.

Definition 2.12 *(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.

Es gilt: Jede stetige Abbildung besitzt eine Adjungierte (ohne Beweis).

Beispiel 2.13 *Sei $U = \mathbb{C}^n$, $V = \mathbb{C}^m$, $A \in L(U, V)$ (also A $(n \times m)$–Matrix, wobei wir immer unzulässigerweise die Matrizen mit den Abbildungen identifizieren, die sie darstellen). U und V seien versehen mit dem Standardskalarprodukt. Dann gilt für $u \in U$, $v \in V$*

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

und damit $A^* = \overline{A^t}$, über $\mathbb{R}$ natürlich $A^* = A^t$. Matrizen mit der Eigenschaft

$$A = A^* = \overline{A^t}$$

heißen hermitesch, reelle Matrizen mit der Eigenschaft

$$A = A^* = A^t$$

heißen symmetrisch.

Satz 2.14 *(Rechenregeln für adjungierte Operatoren)*

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

Definition 2.15 *(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 2.16 *(Diagonalisierbarkeit)*
Sei $T \in L(U,U)$, $\dim U < \infty$. T heißt diagonalisierbar, falls U eine Basis aus Eigenvektoren v_k von T besitzt. Es gilt

$$D = W^{-1}TW,\ W = (v_1 v_2 \cdots v_n), T = \mathit{diag}(\lambda_k).$$

Satz 2.17 *Selbstadjungierte Operatoren haben reelle Eigenwerte. Eigenvektoren zu unterschiedlichen Eigenwerten stehen senkrecht aufeinander.*

Definition 2.18 *(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 2.19 *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 2.20 *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 I. Bitte machen Sie sich klar, dass Ihre Formulierung der folgenden entspricht.

Satz 2.21 *(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*

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

Dann ist J (fast) eine Diagonalmatrix, möglicherweise mit einigen Einsen oberhalb der Hauptdiagonalen, auf der die Eigenwerte von A stehen.

Satz 2.22 *Sei A hermitesche $(n \times n)$–Matrix. Dann ist A diagonalisierbar. U besitzt eine Orthonormalbasis aus Eigenvektoren von A.*

Korollar 2.23 *Die Matrix A sei hermitesch. A ist positiv definit (semidefinit) genau dann, wenn alle Eigenwerte von A positiv (nichtnegativ) sind.*

Satz 2.24 *Eine hermitesche Matrix ist genau dann positiv definit (semidefinit), wenn alle ihre Hauptminoren positiv (nichtnegativ) sind.*

Kapitel 3

Fehler beim numerischen Rechnen

Fehler können beim numerischen Rechnen an mindestens vier Stellen entstehen:

1. Der **Modellierungsfehler** entsteht dadurch, dass wir ein (womöglich vereinfachtes) mathematisches Modell zugrunde legen, das nicht die gesamte Anwendung umsetzt. Beispiel: Im CT–Beispiel haben wir keine Streuung berücksichtigt.
2. Der **Diskretisierungsfehler** entsteht dadurch, dass wir nicht die exakte mathematische Formel implementieren. Beispiel: Approximation des Differentialquotienten durch einen Differenzenquotienten.
3. Der **Messfehler** bewirkt, dass unsere Eingangsdaten nur eine endliche Genauigkeit haben.
4. Der **Rechenfehler** entsteht durch Rundung bei der Durchführung der Rechnung.

Wir werden uns mit den Punkten drei und vier beschäftigen. Der Messfehler dominiert dabei üblicherweise den Rechenfehler.

3.1 Fehlerdefinitionen

Definition 3.1 *(absoluter und relativer Fehler)*
Sei $x \in V$, $(V, ||\cdot||)$ normierter Vektorraum, $\tilde{x}$ eine Näherung für x. Dann heißt

$$||dx||,\ dx = x - \tilde{x}$$

absoluter Fehler *von* $\tilde{x}$. *Falls* $x \neq 0$, *so heißt*

$$\epsilon := \frac{||dx||}{||x||}$$

relativer Fehler *von* $\tilde{x}$.

Bemerkung: Es sei

$$\widetilde{\epsilon} := \frac{||dx||}{||\widetilde{x}||}.$$

Dann gilt sofort

$$\widetilde{\epsilon}(1-\epsilon) \leq \epsilon \leq \widetilde{\epsilon}(1+\epsilon)$$

Für kleines ϵ gilt also $\epsilon \sim \widetilde{\epsilon}$.

Für $x = 0$ wird kein relativer Fehler definiert.

Natürlich hängen alle diese Definitionen von der verwendeten Norm ab. Im Allgemeinen gibt die Anwendung eine Norm vor. Wir werden immer den relativen Fehler betrachten.

3.2 Fehlerverstärkung

Auszuwerten sei die stetig differenzierbare Funktion $f : \mathbb{R} \mapsto \mathbb{R}$ an der Stelle $x \in \mathbb{R}$, zu berechnen ist also $f(x)$. Statt x sei nur eine Näherung $\widetilde{x} = x + dx$ bekannt, wir können also nur $f(x + dx)$ berechnen.

$\widetilde{x}$ hat also einen relativen Fehler von $\epsilon = \frac{|dx|}{|x|}$. Wie groß ist der dadurch verursachte erwartete relative Fehler

$$\frac{|f(x+dx) - f(x)|}{|f(x)|} \text{ ?}$$

Mit dem Mittelwertsatz gilt sofort

$$\exists \xi \in [x - |dx|, x + |dx|] : \frac{f(x+dx) - f(x)}{dx} = f'(\xi)$$

und damit

$$\frac{|f(x+dx)-f(x)|}{|f(x)|} = \frac{|f'(\xi)|\,|dx|}{|f(x)|}\frac{|x|}{|x|} = \frac{|f'(\xi)|\,|x|}{|f(x)|}\frac{|dx|}{|x|} \leq \widetilde{M}\epsilon,$$
$$\widetilde{M} := \sup_{y\in[x-|dx|,x+|dx|]} |f'(y)| \left|\frac{x}{f(x)}\right|$$

Der Eingangsfehler ϵ wird also durch den Faktor M verstärkt.

Wir betrachten sehr kleine Fehler ϵ. Im Limes für $\epsilon \mapsto 0$ gilt $M = |f'(x)|$. Wir schätzen daher die Fehlerverstärkung ab mit dem Faktor

$$M := \frac{|x|\,|f'(x)|}{|f(x)|}.$$

Diese Rechnung kann man genau so auch für Funktionen $f : \mathbb{R}^n \mapsto \mathbb{R}$ ausführen. Sei hierzu gegeben ein $x + dx$ eine Näherung für $x \in \mathbb{R}^n$, und sei $\epsilon_i := \frac{|dx_i|}{|x_i|}$ der relative Fehler der i. Komponente.

Satz 3.2
Sei $f : \mathbb{R}^n \mapsto \mathbb{R}$ stetig differenzierbar. Weiter seien $x,\ dx \in \mathbb{R}^n$. Dann gilt

$$\frac{|f(x+dx)-f(x)|}{|f(x)|} \leq \sum_{k=1}^{n} \widetilde{M_i}\epsilon_i,$$
$$\widetilde{M_i} := \sup_{y\in[x-|dx|,x+|dx|]} |\frac{\partial f}{\partial x_i}(y)| \left|\frac{x_i}{f(x)}\right|.$$

Weiter gilt

$$\lim_{dx\mapsto 0} \widetilde{M_i} = \left|\frac{\partial f}{\partial x_i}(x)\right| \left|\frac{x_i}{f(x)}\right| =: M_i.$$

Die M_i heißen Verstärkungsfaktoren.

Liegen die M_i in der Größenordnung von 1, so liegt der maximale Fehler in der Größenordnung der Eingangsfehler. Gilt $M_i >> 1$, so ist der maximale Fehler viel größer als die Eingangsfehler. Dieses Verhalten nennen wir gut bzw. schlecht gestellt.

Definition 3.3 *Ein Problem heißt gut gestellt, falls kleine Eingangsfehler zu kleinen Fehlern im Ergebnis führen. Ein Problem heißt schlecht gestellt, falls kleine Eingangsfehler zu großen Fehlern im Ergebnis führen können.*

Wir schauen nun als Beispiel auf die Grundrechenarten.

Beispiel 3.4 *1. Multiplikation:*

$$f(x,y) := xy$$

Es gilt

$$M_x = |y| \left| \frac{x}{xy} \right| = 1,\ M_y = 1.$$

Die Multiplikation ist immer gut gestellt.

2. Addition:

$$f(x,y) := x + y$$

Es gilt

$$M_x = 1 \left| \frac{x}{x+y} \right| = \left| \frac{x}{x+y} \right|,\ M_y = \left| \frac{y}{x+y} \right|.$$

Mit $M_x := \left| \frac{1}{1+y/x} \right|$ *wird* M_x *sehr groß, falls* $x \sim -y$*. Die Addition ist also ein schlecht gestelltes Problem, falls* x *und* y *in der gleichen Größenordnung liegen und unterschiedliches Vorzeichen haben. Diese Situation nennen wir* **Auslöschung**.

Beispiel 3.5 *zur Auslöschung: Es sei* $x = 1.01,\ y = -1$ *und* $\widetilde{x} = 1, \widetilde{y} = -1$*. Dann hat* $\widetilde{x}$ *einen Fehler von ca. 0.01=1 %. Es gilt*

$$f(x,y) = x + y = 0.01,\ f(\widetilde{x},\widetilde{y}) = \widetilde{x} + \widetilde{y} = 0$$

und damit für den relativen Fehler des Ergebnisses

$$\epsilon = \frac{|f(\widetilde{x},\widetilde{y}) - f(x,y)|}{|f(x,y)|} = 1 = 100\%$$

Der Fehler hat sich also erheblich erhöht. Der Verstärkungsfaktor ist

$$M_x = \frac{1.01}{0.01} = 101$$

und das erklärt das schlechte Ergebnis.

Bemerkung: Der verstärkte Anfangsfehler ist ein **unvermeidlicher Fehler**. Dieser Fehler entsteht nicht durch einen Algorithmus oder durch falsches Rechnen auf dem Computer, auch bei exakter Rechnung macht man diesen Fehler.

3.3 Maschinendarstellung reeller Zahlen

Ein Rechner kann nicht jede Zahl exakt darstellen, sondern wird nur eine endliche Anzahl von Stellen bereitstellen. Naheliegend wäre es vielleicht, eine Zahl auf dem Rechner in einer Basis b (auf dem Rechner $b = 2$, hier der Einfachheit halber $b = 10$) darzustellen als

$$m_p m_{p-1} \dots m_1 . m_{-1} m_{-2} \dots m_{-r}.$$

Zahlen mit mehr als $n+1$ Stellen ließen sich dann sofort nicht mehr darstellen. Aus diesem Grund multipliziert man zunächst die Größenordnung heraus und repräsentiert die Zahlen in der normalisierten Darstellung

$$\pm 0.m_1 m_2 \dots m_p \, b^e$$

mit $m_1 \neq 0$ (die Null passt dabei nicht und erhält eine Sonderdarstellung). Einige Beispiele für normalisierte Darstellung im Zehnersystem:

$$\begin{aligned} 10 &= 0.1 \cdot 10^2 \\ 0.23 &= 0.23 \cdot 10^0 \\ 0.0000001 &= 0.1 \cdot 10^{-6} \end{aligned}$$

Definition 3.6 *(Gleitkommazahlendarstellung nach IEEE 754, 1985)*
Seien $b >= 2$ (Basis), $p >= 1$ (Mantissenlänge) für ein Format fest gewählte ganze Zahlen. Dann ist die Menge M der **Maschinenzahlen** *definiert durch*

$$M := \left\{ \pm \left(\sum_{k=1}^{p} m_k b^{-k} \right) b^e,\ m_k, e \in \mathbb{Z},\ 0 \leq m_k \leq b-1,\ m_1 \neq 0 \right\} \cup \{0\}.$$

Auf Computern gebräuchlich sind die Werte $b = 2$ und $p = 23$ (single precision float) und $p = 52$ (double precision float).

Bemerkung: Eigentlich hat man noch eine Bedingung an e, aber die sei so großzügig gewählt, dass sie für uns nicht ins Gewicht fällt.

Will man nun eine Zahl auf den Rechner bringen, ist das im Allgemeinen nicht exakt möglich, sie muss auf die nächstgelegene darstellbare Zahl gerundet werden. Die dazu verwendete Funktion nennen wir Rundungsfunktion.

Definition 3.7 *Eine Rundungsfunktion zu einer Menge von Maschinenzahlen M ist eine Funktion $rd : \mathbb{R} \mapsto M$ mit der Eigenschaft*

$$|rd(x) - x| \leq \min_{y \in M} |y - x|.$$

Diese Funktion ist nicht eindeutig definiert. Im Zehnersystem kann man etwa die 5 nach oben oder unten runden, beide Möglichkeiten sind in den Standards zugelassen und während der Rechnung wählbar.

Für uns ist natürlich entscheidend: Wie groß ist der maximale relative Fehler, wenn wir eine Zahl auf den Rechner bringen? Dazu betrachten wir ein kurzes Beispiel mit $b = 10$ und $p = 2$ mit einer zugehörigen Rundungsfunktion rd.

Die unbekannte Zahl x sei zu $m = 0.44 \cdot 10^e$ gerundet worden. Dann muss x zwischen $0.435 \cdot 10^e$ und $0.445 \cdot 10^e$ liegen. Der maximale Fehler ist also

$$0.005 \cdot 10^e = 10^{-3} \frac{10}{2} 10^e = b^{-p-1} \frac{b}{2} b^e.$$

Diese Rechnung gilt für alle Zahlen. Nach Definition der Maschinenzahlen ist $m_1 \neq 0$, d.h. für alle Zahlen ist $m \geq b^{-1} b^e$. Also gilt für den relativen Rundungsfehler

$$\left| \frac{rd(x) - x}{rd(x)} \right| \leq \text{eps}, \text{eps} = \frac{b^{-p+1}}{2}.$$

Satz 3.8 *(Abschätzung des Rundungsfehlers)*

$$|\frac{\mathrm{rd}(x) - x}{\mathrm{rd}(x)}| \leq \mathrm{eps}$$

und

$$|\frac{\mathrm{rd}(x) - x}{x}| \leq \mathrm{eps}\,.$$

Beweis: Teil 1 mit der obigen Rechnung, Teil 2 mit derselben Idee. □

Es ist $eps \sim 10^{-7}$ für einfach und $eps \sim 10^{-16}$ für doppelt genaue Zahlen.

Wenn wir eine reelle Zahl auf den Rechner bringen, entsteht also ein zusätzlicher Fehler. Weiter sind die Maschinenzahlen nicht abgeschlossen. Es gilt etwa für $x = 0.11 \in M$ in dem obigen Format:

$$x \cdot x = 0.121 \notin M,$$

d.h. das Ergebnis muss wieder gerundet werden, mit zusätzlichem Fehler. Dieser Fehler ist sehr klein, viel kleiner als der Messfehler. Trotzdem kann er wichtig sein, nämlich dann, wenn in einer Rechnung eine künstliche Auslöschung auftritt.

3.4 Stabilität

Wir betrachten als Beispiel die Funktion $f(x, y) = x+y$, die wir ausrechnen mit dem Rechenweg $f(x, y) = (x+Z)+(y-Z)$ für eine große Zahl Z. Es seien die Vorzeichen von x und y gleich, d.h. wir wissen bereits, die Aufgabe ist gut konditioniert, der unvermeidbare Fehler ist in der Größenordnung der Eingangsfehler.

Wir setzen nun $x = y = 1$ und $Z = 1000$ und bringen diese Zahlen auf den Rechner von oben, also $p = 2$ und $b = 10$. Dann ist $eps = 0.05$. Es ist $x + Z = 1001$, dieses Ergebnis wird gerundet zu $0.1 \cdot 10^4$. $y - Z = -999$ wird gerundet zu $-0.1 \cdot 10^4$. Auf dem Rechner ist dann das Gesamtergebnis $1000 - 1000 = 0$, der Fehler beträgt also 100%, und das ist viel größer als der Eingangsfehler.

Definition 3.9 *Algorithmen, die einen Fehler liefern, der in der Größenordnung des unvermeidlichen Fehlers liegt, heißen stabil. Ansonsten heißen sie instabil.*

Es sieht vielleicht so aus, als sei das ja auch unsinnig, so etwas zu tun (eine Zahl erst addieren und dann subtrahieren). Manchmal tut man das aber implizit, und tatsächlich ist das bei der pq–Formel zur Lösung quadratischer Gleichungssysteme der Fall. Wir betrachten

$$x_1 = -\frac{p}{2} + \sqrt{\left(\frac{p}{2}\right)^2 - q}$$

Hier gibt es zwei gefährliche Stellen (Auslöschung): Die Subtraktion unter der Wurzel und die äußere Addition der Wurzel. Wir betrachten den zweiten Fall. Falls $|q| << p$, so ist der Wert der Wurzel in der Größenordnung von $\frac{p}{2}$. Daher haben daher die beiden addierten Zahlen unterschiedliches Vorzeichen und liegen in derselben Größenordnung, wir erwarten Auslöschung. Im Jupyter Notebook wird gezeigt, dass dies tatsächlich der Fall ist, und dass der auftretende Fehler viel größer ist als der unvermeidliche Fehler. Dies lässt sich korrigieren durch Anwendung des Satzes von Vieta.

Korollar 3.10 *Die pq–Formel zur Berechnung der Lösung quadratischer Gleichungssysteme ist instabil.*

3.5 Induzierte Matrixnorm

Im Folgenden wollen wir den unvermeidlichen Fehler für Systeme linearer Gleichungen der Form $Ax = b$ untersuchen. Dazu wollen wir zulassen, dass die Matrizen A fehlerhaft sind. Um diesen Fehler zu messen, benötigen wir sinnvolle Matrixnormen. Die wollen wir zunächst definieren und untersuchen.

Naheliegend wäre es, die Norm einer Matrix einfach auf den einzelnen Koeffizienten zu definieren, wie bei den Vektoren auch, also

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

für $p < \infty$. Für $p = 2$ heißt diese Norm z.B. Frobenius–Norm.

Der Vorteil dieser Normen ist, dass sie schnell auszurechnen sind. Der Nachteil ist, dass sie nicht notwendig verträglich sind mit der Vektorraumnorm (d.h. es gilt nicht $||Av|| \leq ||A||\,||v||$). Für die Zwecke dieser Vorlesung sind sie damit im Allgemeinen unbrauchbar.

Zur Motivation: Sei in unserem Problem $dA = 0$, also $A = \widetilde{A}$. Dann gilt $\widetilde{x} = A^{-1}\widetilde{b}$. Der relative Fehler von $\widetilde{x}$ gegen die wahre Lösung x ist dann

$$\frac{||A^{-1}db||}{||x||}.$$

Wir wollen nun den Zähler abschätzen. Für die Matrizen haben wir noch keine Norm festgelegt. Naheliegend wäre eine Definition für $||A^{-1}||$, so dass gilt

$$||A^{-1}db|| \leq ||A^{-1}||\,||db|| \,\forall db \in \mathbb{R}^n,$$

und natürlich sollte diese Grenze möglichst klein sein. Teilen durch $||db||$ legt nahe

$$||A^{-1}|| := \sup_{db \neq 0} \frac{||A^{-1}\,db||}{||db||}.$$

Wir definieren daher

Definition 3.11 *(induzierte Matrixnorm)*
Sei A in $\mathbb{R}^{n\times m}$, und seien $||\cdot||_{\mathbb{R}^m}$ und $\cdot||_{\mathbb{R}^n}$ Normen im $\mathbb{R}^m$ bzw. $\mathbb{R}^n$. Dann heißt

$$||A|| := \sup_{x\in\mathbb{R}^m, x\neq 0} \frac{||Ax||_{\mathbb{R}^n}}{||x||_{\mathbb{R}^m}} = \sup_{x\in\mathbb{R}^m, x\neq 0} ||A\frac{x}{||x||_{\mathbb{R}^m}}||_{\mathbb{R}^n} = \sup_{x\in\mathbb{R}^m, ||x||=1} ||Ax||$$

(induzierte, verträgliche) Matrixnorm von A.

Die erste Frage ist natürlich: Ist das wohldefiniert, d.h. ist das $\sup < \infty$, und ist dies eine Norm?

Satz 3.12 *Die Matrixnorm ist wohldefiniert und ist eine Norm.*

Beweis: Mit den Bezeichnungen aus 3.11: Wir werden gleich zeigen, dass in der Supremumsnorm $||\cdot||_\infty$ gilt

$$\forall A \; \exists M \in \mathbb{R} : \frac{||Ax||_\infty}{||x||_\infty} \leq M \, \forall x \in \mathbb{R}^m, \, x \neq 0.$$

Wegen der Normäquivalenz 2.10 gibt es Zahlen c_1 und c_2 mit

$$||Ax||_{\mathbb{R}^n} \leq c_1 ||Ax||_\infty, \; ||x||_{\mathbb{R}^n} \geq c_2 ||x||_\infty$$

und damit

$$\frac{||Ax||_{\mathbb{R}^n}}{||x||_{\mathbb{R}^m}} \leq \frac{c_1}{c_2} \frac{||Ax||_\infty}{||x||_\infty} \leq \frac{c_1}{c_2} M \, \forall x \in \mathbb{R}^m, \, x \neq 0.$$

Also ist der Quotient nach oben durch eine Zahl beschränkt und das Supremum ist endlich. Die Normeigenschaften zeigt man durch einfaches Nachrechnen. □
Im Folgenden werden wir die Indizes an den Normen weglassen.

Satz 3.13 *(Eigenschaften der induzierten Norm)*
Sei $A \in \mathbb{R}^{n \times m}$, $x \in \mathbb{R}^m$. Dann gilt

$$||Ax|| \leq ||A|| \, ||x||.$$

Sei $B \in \mathbb{R}^{m \times r}$. Dann gilt

$$||AB|| \leq ||A|| \, ||B||.$$

Beweis:

$$||Ax|| = ||A \frac{x}{||x||}|| \, ||x|| \leq ||A|| \, ||x|| \, (x \neq 0).$$

$$||AB|| = \sup_{||x||=1} ||ABx|| \leq \sup_{||x||=1} ||A|| \, ||Bx|| \leq \sup_{||x||=1} ||A|| \, ||B|| \, ||x|| = ||A|| \, ||B||.$$

□

Korollar 3.14 *Lineare Abbildungen auf endlichdimensionalen Vektorräumen sind stetig.*

Beweis: Sei alles wie in 3.12. Sei x_n eine Folge in $\mathbb{R}^m$, die gegen x konvergiert. Dann gilt

$$||Ax_n - Ax|| \leq ||A|| \, ||x_n - x|| \mapsto 0,$$

also konvergiert Ax_n gegen Ax und A ist stetig. □

In endlichdimensionalen Banachräumen wird das Infimum angenommen. Um zu zeigen, dass $||\cdot||$ eine induzierte Matrixnorm ist, ist also zu zeigen:

1. $||Ax|| \leq ||A||\,||x|| \forall x \in \mathbb{R}^m$

2. $\exists \overline{x} \in \mathbb{R}^m : ||A\overline{x}|| = ||A||\,||\overline{x}||$.

Wir berechnen die Matrixnorm an zwei Beispielen, $||\cdot||_\infty$ und $||\cdot||_2$.

Beispiel 3.15 *Sei $A = (A_{k,j}) \in \mathbb{R}^{n\times m}$ nicht die Nullmatrix. Wir bestimmen $||A||_\infty$. Sei dazu $x = (x_k) \in \mathbb{R}^m$ beliebig. Nach Definition der Supremumsnorm gilt*

$$|x_j| \leq ||x||_\infty \, \forall j.$$

Dann gilt:

$$||Ax||_\infty = ||\sum_j A_{k,j}x_j||_\infty = \max_k |\sum_j A_{k,j}x_j| \leq \max_k \sum_j |A_{k,j}x_j| \leq \max_k \sum_j |A_{k,j}|\,||x||_\infty$$

und damit $||A|| \leq \max_k \sum |A_{k,j}|$.

Zu zeigen ist noch, dass diese Grenze angenommen wird. Sei $\tilde{k}$ der Index, an dem das Zeilenmaximum angenommen wird, also

$$\max_k \sum_j |A_{k,j}| = \sum_j |A_{\tilde{k},j}|.$$

Sei $\overline{x} \in \mathbb{R}^n$ mit $\overline{x}_j = \operatorname{sgn}(A_{\tilde{k},j})$ mit der Definition

$$\operatorname{sgn} : \mathbb{R} \mapsto \mathbb{R},\ \operatorname{sgn}(y) = \begin{cases} 1 & y > 0 \\ 0 & y = 0 \\ -1 & y < 0 \end{cases}$$

Mit dieser Definition gilt $y\operatorname{sgn}(y) = |y|$, also für unser $\overline{x}$ mit $||\overline{x}|| = 1$:

$$||A\overline{x}||_\infty \geq (A\overline{x})_{\tilde{k}} = \sum_j A_{\tilde{k},j}\overline{x}_j = \sum_j |A_{\tilde{k},j}| = \max_k \sum_j |A_{k,j}|.$$

und damit

$$||A||_\infty = \max_k \sum_j |A_{k,j}|.$$

Wichtiger als die ∞–Norm ist die 2–Norm. Wir erinnern zunächst an einige Definitionen und Sätze der Linearen Algebra.

Seien $x, y \in \mathbb{C}^m$. Dann sind das euklidische Standardskalarprodukt und die euklidische Norm definiert durch

$$(x, y) = x^t \overline{y},\ ||x||_2^2 = (x, x).$$

Wir werden häufig die Adjungierte einer Matrix benutzen, definiert durch

$$A^* := \overline{A^t} \quad (2.12).$$

Dann ist A^*A positiv semidefinit, und der $\mathbb{R}^m$ besitzt eine Orthonormalbasis aus Eigenvektoren v_k von A^*A zu Eigenwerten $\lambda_k \geq 0$. Eigenwerte seien immer dem Betrag nach geordnet, d.h.

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_m|.$$

Seien $x, y \in \mathbb{C}^m$. Da die v_k eine Basis bilden, lassen sich x und y in dieser Basis darstellen:

$$x = \sum_k \alpha_k v_k$$
$$y = \sum_j \beta_j v_j$$

Dann gilt

$$(x, y) = (\sum_k \alpha_k v_k, \sum_j \beta_j v_j) = \sum_{k,j} \alpha_k \overline{\beta_j} (v_k, v_j) = \sum_k \alpha_k \overline{\beta_k}$$

denn $(v_k, v_j) = \delta_{k,j}$ mit dem Kronecker-δ.

Insbesondere gilt daher

$$||x||^2 = (x, x) = \sum_k |\alpha_k|^2$$

und

$$||Ax||^2 = (Ax, Ax) = (A^*Ax, x) = \sum_k \lambda_k |\alpha_k|^2.$$

Definition 3.16 *Sei $B \in \mathbb{C}^{m\times m}$. Dann heißt*

$$\rho(B) = \max\{|\lambda| : \lambda \textit{ Eigenwert von } B\}$$

Spektralradius von B.

Mit unseren Definitionen von oben gilt also

$$\rho(A^*A) = |\lambda_1| = \lambda_1.$$

Satz 3.17 *Sei $A \in \mathbb{C}^{m\times n}$. Dann gilt*

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

Beweis: Wir müssen nur noch einsetzen. Mit unseren Bezeichnungen von oben gilt für jedes $x \in \mathbb{R}^m$

$$||Ax||_2^2 = \sum_k \lambda_k |\alpha_k|^2 \leq \lambda_1 \sum_k |\alpha_k|^2 = \rho(A^*A)||x||_2^2.$$

Außerdem gilt

$$||Av_1||_2^2 = (A^*Av_1, v_1) = \lambda_1(v_1, v_1) = \lambda_1||v_1||_2^2.$$

Nach Ziehen der Wurzel gilt also für alle $x \in \mathbb{R}^m$

$$||Ax||_2 \leq \rho(A^*A)^{1/2}||x||_2$$

und Gleichheit für $x = v_1$, also ist $\rho(A^*A)^{1/2} = ||A||_2$. □

3.6 Fehler bei linearen Gleichungssystemen

Wir wollen lösen ein Gleichungssystem $Ax = b$, A invertierbar. Für A steht uns nur eine Näherung $\widetilde{A} = A + dA$ zur Verfügung, für b eine Näherung $\widetilde{b} = b + db$. Wir können also nur das Gleichungssystem

$$\widetilde{b} = \widetilde{A}\widetilde{x} = (A + dA)(x + dx),\ \widetilde{x} = x + dx.$$

lösen. Wie groß ist der relative Fehler von $\widetilde{x}$, also $\frac{||dx||}{||x||}$?

Zunächst wollen wir klären, unter welchen Voraussetzungen $\widetilde{A}$ überhaupt invertierbar ist. Dies ist natürlich nicht immer der Fall, wenn etwa gilt $A = I_2$ und wir uns vermessen im Element $(1, 1)$, so entsteht

$$\widetilde{A} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

und die Matrix ist natürlich nicht invertierbar. Wir fragen daher zunächst: Wie stark darf man eine Einheitsmatrix verändern, damit sie invertierbar bleibt? Die Antwort gibt

Satz 3.18 *(Neumannsche Reihe)*
Sei $A \in \mathbb{R}^{n\times n}$ mit $||A|| < 1$ für eine induzierte Norm. Dann ist $(I - A)$ invertierbar, und

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

Beweis: Dies ist natürlich nichts anderes als eine geometrische Reihe, nur dass man statt einer Zahl eine Matrix einsetzt. Der Beweis ist exakt der gleiche.
Wegen $||A|| < 1$ bilden die Partialsummen von $\sum_{k=0}^{\infty} A^k$ eine Cauchyfolge, also konvergiert die Summe. Es gilt

$$(I - A)\sum_{k=0}^{n} A^k = I - A^{n+1} \mapsto I,$$

und damit

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

□

Korollar 3.19 *Sei $A \in \mathbb{R}^{n\times n}$ invertierbar, $dA \in \mathbb{R}^{n\times n}$. Weiter sei*

$$q := ||A^{-1}||\,||dA|| < 1.$$

Dann ist $(A + dA)$ invertierbar und

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

Beweis:

$$(A + dA) = A(I - (-A^{-1}dA))$$

ist invertierbar nach 3.18.

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

□

Dieser Satz lässt sich so interpretieren: Die Matrix A sei invertierbar. Bekannt ist eine Approximation $\widetilde{A}$. Falls $||A-\widetilde{A}||$ klein genug ist, so ist auch $\widetilde{A}$ invertierbar.

Korollar 3.20 *Die Menge der invertierbaren $(n \times n)$–Matrizen ist offen.*

Als Anwendung berechnen wir jetzt den unvermeidbaren Fehler bei der Lösung eines linearen Gleichungssystems

$$Ax = b$$

mit einer invertierbaren $n \times n$–Matrix A und $b \in \mathbb{R}^n$. Wir bestimmen also eine Abschätzung für den Fehler, der entsteht, wenn die Koeffizienten der invertierbaren Matrix A oder des Vektors b nicht genau bekannt sind, sondern statt dessen nur Näherungen $\widetilde{A} = A + dA$ und $\widetilde{b} = b + db$ zur Verfügung stehen und wir ersatzweise die Lösung des Gleichungssystems

$$(A + dA)\tilde{x} = b + db$$

berechnen.

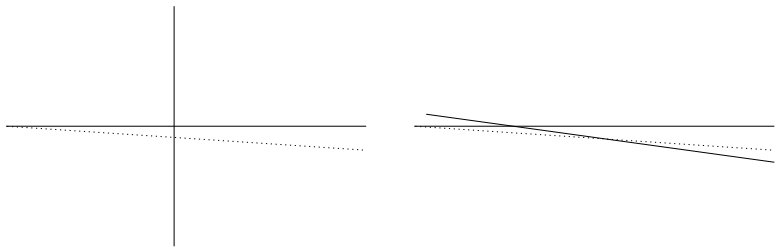


Abbildung 3.1: Graphische Lösung von Gleichungssystemen: Links gut gestellt, rechts schlecht gestellt, kleine Änderungen (gestrichelte Linie) in den Koeffizienten führen zu großer Änderung des Schnittpunkts.

Sei zunächst $n = 2$. 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 3.21 *Sei $A \in \mathbb{R}^{n\times n}$ invertierbar. Sei $x \in \mathbb{R}^n$ und $Ax = b$. Sei weiter $dA \in \mathbb{R}^{n\times n}$ und $db \in \mathbb{R}^n$. Es sei*

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

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

$$q = k(A)\frac{||dA||}{||A||} < 1.$$

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

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

Dann gilt für den relativen Fehler in der Lösung

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

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

$$M = k(A)/(1-q)$$

verstärkt.

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

Beweis: Nach 3.19 ist $A + dA$ invertierbar, und es gilt

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

Es gilt

$$(A + dA)(x + dx) = (b + db)$$

und damit wegen $Ax = b$

$$(A + dA)dx = db - dA\,x$$

und

$$dx = (A + dA)^{-1}(db - dA\,x),$$

also insbesondere

$$||dx|| \leq ||(A + dA)^{-1}||(||db|| + ||dA||\,||x||).$$

Für den relativen Fehler für $x \neq 0$

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

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

Kapitel 4

Direkte Lösung linearer Gleichungssysteme

Fast jedes praktische Problem führt am Ende nach langer Modellierung auf ein lineares Gleichungssystem. Deshalb ist ihre Lösung von fundamentaler Bedeutung für die Angewandte Mathematik. Wir betrachten zunächst direkte Verfahren, die in endlicher Zeit eine Lösung liefern, gegenüber iterativen Verfahren, bei denen eine Folge ausgerechnet wird, die gegen die Lösung konvergiert. Direkte Verfahren sind dabei typischerweise langsam für große Matrizen und spielen heute eine untergeordnete Rolle.

Eine gute Quelle für klassische Algorithmen und Analysen zu diesem Bereich ist das Buch von Golub und van Loan, Matrix Computations.

4.1 Gauß–Elimination und LR–Zerlegung

Die **Gauß–Elimination** sollte bereits aus der Schule bekannt sein. Wir rechnen trotzdem zur Einführung ein Mikro–Beispiel.

Zeilenoperation

$$
\begin{array}{rcrcrcrcrlll}
3 & x_1 & + & 2 & x_2 & + & & x_3 & = & 8 & & & \\
6 & x_1 & + & 5 & x_2 & - & 4 & x_3 & = & 12 & l_{21}=2 & II - l_{21} I & \equiv \mathbf{A^{(1)}x = b^{(1)}} \\
-3 & x_1 & + & & x_2 & - & 2 & x_3 & = & -3 & l_{31}=-1 & III - l_{31} I &
\end{array}
$$

$$
\begin{array}{rcrcrcrcrlll}
3 & x_1 & + & 2 & x_2 & + & & x_3 & = & 8 & & & \\
 & & & & x_2 & - & 6 & x_3 & = & -4 & & & \equiv \mathbf{A^{(2)}x = b^{(2)}} \\
 & & & 3 & x_2 & - & & x_3 & = & 5 & l_{32}=3 & III - l_{32} II &
\end{array}
$$

$$
\begin{array}{rcrcrcrcrlll}
3 & x_1 & + & 2 & x_2 & + & & x_3 & = & 8 & & & \\
 & & & & x_2 & - & 6 & x_3 & = & -4 & & & \equiv \mathbf{A^{(3)}x = b^{(3)}} \\
 & & & & & & 17 & x_3 & = & 17 & & &
\end{array}
$$

Durch **Rückwärtseinsetzen** ergibt sich damit

$$x_3 = 17/17 = 1,\ x_2 = (-4+6)/1 = 2,\ x_3 = (8-1-2\cdot 2)/3 = 1.$$

Wir werden Algorithmen immer in einem Pseudocode formulieren.

Zu lösen sei $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$. Setze $A^{(1)} = A$ und $b^{(1)} = b$. Es sei $A^{(k)} = (a_{jl}^{(k)})$ usw.

Für $i = 1 \ldots n-1$

Zur Konstruktion des Gleichungssystems $A^{(i+1)}x = b^{(i+1)}$

Übernehme die ersten i Gleichungen, d.h. die ersten i Zeilen.

Für $j = i+1 \ldots n$

$l_{ji} = \frac{a_{ji}^{(i)}}{a_{ii}^{(i)}}$ falls $a_{ii}^{(i)} \neq 0$.

Für $k = i+1 \ldots n$

$$a_{jk}^{(i+1)} = a_{jk}^{(i)} - l_{ji} \cdot a_{ik}^{(i)}$$

$$b_j^{(i+1)} = b_j^{(i)} - l_{ji} b_i^{(i)}$$

Setze die restlichen Einträge auf 0.

Für $i = n \ldots 1$

$$x_i = \left(b_i^{(n)} - \textstyle\sum_{j=i+1}^{n} a_{ij}^{(n)} x_j \right) / a_{ii}^{(i)}.$$

Hierbei benötigen wir die Matrizen $A^{(k)}$ zur Berechnung der Lösung nicht, es liegt also nahe, jeweils $A^{(k)}$ mit $A^{(k+1)}$ zu überschreiben. Es wird also im Laufe des Algorithmus kein zusätzlicher Speicherplatz benötigt.

Wir bestimmen den Aufwand zur Lösung des Systems. Wir vereinbaren zunächst: Da Addition und Multiplikation fast immer zusammen auftreten, zählen wir sie als eine **Rechenoperation**. Tatsächlich sind moderne Rechnerarchitekturen in der Lage, diese beiden Operationen gleichzeitig durchzuführen (fused multiply add).

Für das Auflösen des Gleichungssystems werden dann

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(2 + \sum_{k=i+1}^{n} 1 \right) = \frac{1}{6}(2n^3 + 3n^2 - 5n) = \frac{1}{6} n(n-1)(2n+5)$$

Rechenoperationen und n Divisionen benötigt, wobei wir für die einzelnen Divisionen jeweils einmal den Kehrwert der $a_{ii}^{(i)}$ ausrechnen und dann mit ihm multiplizieren. Die Division ist nämlich tatsächlich recht aufwändig, einen Algorithmus zu ihrer schnellen Berechnung (mit einigen Rechenoperationen) werden wir im Kapitel über die Newton–Iteration herleiten.

Zur Durchführung des Rückwärtseinsetzens erhalten wir

$$\sum_{i=1}^{n} (2 + \sum_{j=i+1}^{n} 1) = n^2/2 + 7/2\, n.$$

Alle Berechnungen dieser Art interessieren uns immer nur für große n. Dann dominieren aber sofort die Terme mit hoher Potenz die mit kleiner, und nur der Leitterm mit der höchsten Potenz ist interessant. Es würde also reichen, den größten Term (mit einer Abschätzung für den Rest) zu kennen.

Für h nah bei 0 ist es genau umgekehrt: Hier ist etwa $h^2 << h$, d.h. hier sind die kleinen Potenzen wichtig. Wir definieren daher das Landau-Symbol:

Definition 4.1 *(**Landau–Symbole**)*

1. *Sei* $f : \mathbb{N} \mapsto \mathbb{N}$.

$$f(n) = O(n^p) \textit{ für große } n \Leftrightarrow \exists C > 0 : \ |f(n)| \leq C n^p \, \forall n > 0.$$

2. *Sei* $f : \mathbb{R} \mapsto \mathbb{R}$.

$$f(h) = O(h^p) \textit{ für kleine } h \Leftrightarrow \exists C > 0 : \ |f(h)| \leq C |h|^p \, \forall |h| \leq 1.$$

Beispiel 4.2

1. $O(n^\alpha) = O(n^\beta)$ *für* $0 < \alpha \leq \beta$.
   $\sin(n) = O(1)$.

2. $O(h^\alpha) = O(h^\beta)$ *für* $\alpha \geq \beta > 0$.
   $\sin(h) = O(1)$
   $\sin(h) = O(h)$

Mit dieser Konvention gilt

Satz 4.3 *Die Auflösung einer Gleichung mit n Unbekannten mit dem Gauß–Algorithmus benötigt $n^3/3 + O(n^2)$ Rechenoperationen und n Divisionen.*

Bemerkung:

1. Die Cramersche Regel rechnet die Determinanten der Matrix aus, was bei direkter Berechnung die Komplexität $n!$ hat (und damit völlig unbrauchbar ist).
2. Die Gauss–Elimination ist durchführbar genau dann, wenn alle $a_{ii}^{(i)} \neq 0$ (denn dann können wir die Divisionen durchführen).
3. Falls $a_{ii}^{(i)} = 0$, aber $a_{ki}^{(i)} \neq 0$ für ein $k > i$, so vertausche die $k.$ und die $i.$ Zeile des Gleichungssystems (was die Lösung natürlich nicht ändert).
4. Falls $a_{ki}^{(i)} = 0$ für alle $k \geq i$, so ist x_i aus den Gleichungen i bis n bereits eliminiert, und der $i.$ Schritt muss gar nicht erst durchgeführt werden. In diesem Fall hat $A^{(i)}$ die Form

$$\begin{pmatrix} * & & & & & \\ 0 & * & & & & \\ \vdots & \ddots & \ddots & & & \\ 0 & \cdots & 0 & * & & \\ 0 & \cdots & 0 & 0 & * & * \\ \vdots & & \vdots & \vdots & \vdots & * \\ 0 & \cdots & 0 & 0 & * & * \end{pmatrix}$$

   Entwicklung der Determinante nach der ersten Spalte zeigt sofort: Dann ist $A^{(i)}$ singulär, und damit auch A. Falls A invertierbar ist, kann dieser Fall also nicht auftreten.

   Die Gauss-Elimination ist auf einer Permutation des Systems immer ausführbar.

5. Fehleranalyse: Eine genaue Fehleranalyse ist für den Gauß–Algorithmus schwierig. Wir betrachten nur die Berechnung von x_1. $\tilde{x_1}$ wird berechnet durch

$$\tilde{x_1} = \frac{1}{a_{11}} \underbrace{(b_1 - \sum_{k=2}^{n} a_{1k}\tilde{x_k})}_{a_{11}x_1}.$$

Ist nun $|a_{11}x_1|$ klein gegenüber b_1, so kann dies nur dadurch entstanden sein, dass in der Differenz Auslöschung aufgetreten ist. Fehler werden also stark verstärkt, wenn $|a_{11}|$ klein ist. Wir ordnen deshalb im $i.$ Schritt die Gleichungen so an, dass das Diagonalelement in der $i.$ Spalte unterhalb der Hauptdiagonalen betragsmaximal ist, dass also gilt

$$|a_{ii}^{(i)}| = \max_{k \geq i} |a_{ki}^{(i)}|.$$

Diese Strategie heißt **Spaltenpivotsuche** und macht den Gaußalgorithmus bereits zu einem stabilen (in praktischen Fällen).

Wir rechnen dazu ein kurzes Beispiel, zunächst ohne Pivotsuche. Wir benutzen zur Rechnung einen Rechner mit $b = 10$ und $p = 2$, also mit zwei Stellen.

$$\begin{array}{rcl} 10^{-4}\ x_1 + & x_2 &= 1 + 10^{-4} \\ x_1 + & x_2 &= 2 \end{array}$$

$$\begin{array}{rcl} 10^{-4}\ x_1 + & x_2 &= 1 \\ \underbrace{(-10^4 + 1)}_{=-10^4} & x_2 &= \underbrace{-10^4 + 2}_{=-10^4} \end{array}$$

und damit $\tilde{x_2} = -10^4/-10^4 = 1$ und $\tilde{x_1} = (1-1)/10^{-4} = 0$. Da $x = (1,1)$ die korrekte Lösung ist, hat x_1 einen Fehler von 100% (genau wie oben vorausgesagt sorgt die Auslöschung beim Rückwärtseinsetzen für einen großen Fehler). Die Kondition der Matrix ist kleiner als drei, dieser Fehler ist also nicht unvermeidbar. Nun dasselbe mit Pivotsuche:

$$\begin{array}{rcl} x_1 + & x_2 &= 2 \\ 10^{-4}\ x_1 + & x_2 &= 1 + 10^{-4} \end{array}$$

$$\begin{array}{rcl} x_1 + & x_2 &= 2 \\ \underbrace{(1 - 10^{-4})}_{=1} & x_2 &= \underbrace{1 - 2 \cdot 10^{-4}}_{=1} \end{array}$$

und wir erhalten die korrekte Lösung $x_2 = 1$, $x_1 = 1$.

Die Matrix $A^{(n)}$ hat Zeilenstufenform, d.h. unterhalb der Hauptdiagonalen stehen nur Nullen.

Definition 4.4 *Sei $A \in \mathbb{R}^{n \times n}$. Falls alle Elemente unterhalb der Hauptdiagonalen Null sind, so heißt A* **rechte obere Dreiecksmatrix**.
Entsprechend heißt A **linke untere Dreiecksmatrix**, *falls alle Elemente oberhalb der Hauptdiagonalen Null sind.*
Falls auf der Hauptdiagonalen von L nur Einsen stehen, so heißt L normiert.

Definition 4.5

1. *Eine normierte linke untere Dreiecksmatrix $L = (l_{kj})$ heißt* **Elementarmatrix,** *wenn nur in einer Spalte i unterhalb der Hauptdiagonalen Einträge ungleich 0 stehen.*
2. *Eine Matrix $P \in \mathbb{R}^{n\times n}$ heißt* **Permutationsmatrix,** *falls in jeder Zeile und Spalte genau eine 1 auftaucht und alle anderen Einträge 0 sind, d.h.*

$$\exists\, \sigma \in \{1 \dots n\}^n : \ \sigma_k \neq \sigma_j \text{ für } k \neq j,\ a_{i,k} = \begin{cases} 1, & k = \sigma_i \\ 0, & k \neq \sigma_i \end{cases} .$$

Beispiel 4.6

$$L^{(i)} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{i+1,i} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{n,i} & & & 1 \end{pmatrix}$$

ist Elementarmatrix.

$$P = \begin{pmatrix} & 1 & & \\ & & 1 & \\ & & & 1 \\ 1 & & & \end{pmatrix}, \ P^t = \begin{pmatrix} & & & 1 \\ 1 & & & \\ & 1 & & \\ & & 1 & \end{pmatrix}$$

sind Permutationsmatrizen zu $\sigma = (2, 3, 4, 1)$ bzw. $(4, 1, 2, 3)$. Offensichtlich gilt

$$PP^t = P^tP = I.$$

Bemerkung:

1. LA lässt die ersten i Zeilen von A konstant und subtrahiert jeweils von den Zeilen $j = i+1, \dots, n$ das $l_{j,i}$–fache der i. Zeile. Dies ist genau die Operation, die den i. Eliminationsschritt durchführt.
2. Diese Operation kann man rückgängig machen, indem man auf die Zeilen $j = i+1, \dots, n$ das $l_{j,i}$–fache der i. Zeile addiert. Die Inverse von L erhält man also einfach durch Streichen des Minuszeichens.
3. Das Produkt zweier Elementarmatrizen $L^{(i)}L^{(i+1)}$ zu den Spalten i und $i+1$ ist eine normierte linke untere Dreiecksmatrix, die unterhalb der Hauptdiagonalen die Elemente $l_{j,i}$ und $l_{j,i+1}$ enthält.

Begründung: Statt erst ein Vielfaches der $(i+1)$. und dann ein Vielfaches der i. Zeile zu addieren, kann man beides gleichzeitig tun.

4. PA bringt die Zeilen von A in die Reihenfolge $\sigma_1 \dots \sigma_n$.

Offensichtlich sind das genau die Matrizen, die wir zur exakten Beschreibung des Gauß–Algorithmus benötigen. Wir formulieren dies als

Satz 4.7 *(**LR-Zerlegung**)*
Sei A eine $n \times n$–Matrix. Dann gibt es eine Permutationsmatrix P zur Permutation σ, eine normierte linke untere Dreiecksmatrix L und eine rechte obere Dreiecksmatrix R (alles $(n \times n)$), so dass

$$PA = LR.$$

L und R heißen LR–Zerlegung von PA.

Beweis: Wir wissen bereits, dass die Gausselimination auf einer Permutation der Matrix durchführbar ist. Sei PA diese Permutation.
Es sei $A^{(1)} = PA$. Dann gilt $A^{(2)} = L^{(1)}A^{(1)} = L^{(1)}PA$, $A^{(3)} = L^{(2)}A^{(2)} = L^{(2)}L^{(1)}PA$ usw. Insbesondere

$$R := A^{(n)} = L^{(n-1)} \cdots L^{(1)}PA.$$

Damit gilt

$$PA = \underbrace{(L^{(1)})^{-1} \cdots (L^{(n-1)})^{-1}}_{L} R.$$

Nach Vorbemerkung 2 und 3 gilt

$$L = \begin{pmatrix} 1 & & & \\ l_{2,1} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n,1} & \cdots & l_{n,n-1} & 1 \end{pmatrix},$$

insbesondere ist R rechte obere und L normierte linke untere Dreiecksmatrix. □

Für unser Beispiel vom Anfang, das wir ohne Zeilenpermutationen durchgeführt hatten, gilt

$$P = I,\ L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{pmatrix},\ R = \begin{pmatrix} 3 & 2 & 1 \\ 0 & 1 & -6 \\ 0 & 0 & 17 \end{pmatrix}$$

Im zweiten Beispiel zur Spaltenpivotsuche gilt

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Bemerkung:

1. Sei $PA = LR$. Dann kann $Ax = b$ gelöst werden durch

$$\begin{aligned} Ax = b &\iff PAx = Pb \\ &\iff LRx = Pb \\ &\iff Ly = Pb,\ Rx = y \end{aligned}$$

   Dabei wird y bestimmt durch **Vorwärtseinsetzen** (Auflösung der Gleichungen von vorn nach hinten) in $Ly = Pb$ und x durch **Rückwärtseinsetzen** (Auflösung der Gleichungen von hinten nach vorn) in $Rx = y$.

2. Nicht jede invertierbare Matrix besitzt eine LR–Zerlegung. Sei etwa

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \cdot \begin{pmatrix} b & c \\ 0 & d \end{pmatrix} = \begin{pmatrix} b & c \\ ab & ac+d \end{pmatrix}.$$

   Dann gilt offensichtlich $b = 0$, also $ab = 0$ ↯.

3. Es gibt singuläre Matrizen, die eine LR–Zerlegung besitzen.

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

4. Die LR–Zerlegung mit Spaltenpivotsuche und Vorwärts–Rückwärtseinsetzen ist für praktische Zwecke ein gutartiger Algorithmus zur Bestimmung der Lösung eines linearen Gleichungssystems.

5. Der Aufwand zur Berechnung der LR–Zerlegung beträgt $n^3/3 + O(n^2)$ Rechenoperationen ($+n$ Divisionen). Der Aufwand für das Einsetzen beträgt $n^2 + O(n)$.

Satz 4.8 *(**Eindeutigkeit** der LR–Zerlegung) Sei A eine invertierbare $n \times n$–Matrix. Falls A eine LR–Zerlegung besitzt, so ist diese Zerlegung eindeutig.*

Beweis: Wir benutzen ohne Beweis das Lemma: Die linken unteren und rechten oberen Dreiecksmatrizen bilden einen Ring, d.h. Summe, Produkt und Inverse sind wieder linke untere bzw. rechte obere Dreiecksmatrix. Produkte von normierten Dreiecksmatrizen sind wieder normiert.

A besitze die LR–Zerlegungen (L, R) und (L', R'). Da A invertierbar ist, sind auch die Dreiecksmatrizen invertierbar. Es gilt

$$A = LR = L'R' \Longrightarrow (L')^{-1}L = R'R^{-1} =: Z.$$

Z ist Produkt linker unterer normierter Dreiecksmatrizen, also selbst wieder normierte linke untere Dreiecksmatrix. Andererseits ist Z Produkt rechter oberer Dreiecksmatrizen, also selbst wieder rechte obere Dreiecksmatrix. Damit ist Z Diagonalmatrix und hat, weil sie normiert ist, 1 auf der Hauptdiagonalen, ist also die Einheitsmatrix. Damit gilt

$$L = L' \text{ und } R = R'.$$

□

Kapitel 5

Über- und unterbestimmte Gleichungssysteme

Wir haben bereits in der Einleitung gesehen, dass in Anwendungen nicht notwendig die Anzahl der Gleichungen und Variablen in einem linearen Gleichungssystem übereinstimmen. Sollen etwa bei einer Landvermessung Positionen bestimmt werden, so macht man typischerweise mehr Messungen als notwendig, um Messfehler ausgleichen zu können. Das bekannteste Beispiel findet sich bei Gauss, der die Theorie dazu in seinem Buch "Theoria combinationis observationum erroribus minimis obnoxiae" veröffentlichte (die Society for Industrial and Applied Mathematics hat freundlicherweise für die Nicht–Lateiner eine englische Übersetzung veröffentlicht). Der alte 10 DM–Schein erinnerte an diese Arbeit von Gauss. Eine andere Motivation ist die Betrachtung von Ausgleichsgeraden (für beide Beispiele siehe Vorlesung).

Mögliche Probleme bei der Lösung sind:

1. Es gibt keine Lösung (überbestimmtes System), im allgemeinen mehr Gleichungen als Unbekannte
2. Es gibt unendlich viele Lösungen (unterbestimmtes System), im allgemeinen weniger Gleichungen als Unbekannte

5.1 Kleinste Quadrate–Lösung

Falls es keine Lösung gibt, suchen wir Vektoren x^*, für die das Residuum $Ax^* - b$ zwar nicht 0, aber doch möglichst klein ist. In diesem gesamten Kapitel verwenden wir dazu die euklidische Norm.

Definition 5.1 *(**kleinste Quadrate–Lösung, least squares solution**) Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. $x^* \in \mathbb{R}^n$ heißt kleinste Quadrate–Lösung (kQL) von $Ax = b$ genau dann, wenn*

$$||Ax^* - b||_2^2 \leq ||Ax - b||_2^2 \, \forall x \in \mathbb{R}^n.$$

Bemerkung: Wir machen keine Voraussetzungen an m, n oder den Rang von A.

Bemerkung: Falls $Ax = b$ Lösungen besitzt, so sind genau diese auch die kleinste Quadrate–Lösungen.

Wir halten mal direkt fest: Falls A die Nullmatrix ist, so gilt für jedes $x^* \in \mathbb{R}^n$

$$||Ax^* - b||_2^2 = ||b||_2^2 = ||Ax - b|| \, \forall x \in \mathbb{R}^n.$$

Damit sind alle Vektoren aus dem $\mathbb{R}^n$ kleinste Quadrate–Lösungen, insbesondere ist die kQL also nicht notwendig eindeutig.

Leider hilft uns diese Definition nicht bei der Berechnung der kleinsten Quadrate–Lösung. Wir geben daher eine zur Definition äquivalente Bedingung an. Sei im Folgenden immer $A \in \mathbb{R}^{m \times n}$. Wir beschränken uns hier der Einfachheit halber auf reelle Matrizen, obwohl alles sofort auch ins Komplexe übertragbar ist. Wir beginnen mit

Lemma 5.2 *Sei $A \in \mathbb{R}^{m \times n}$.*

1. $$\mathit{Bild}(A)^\perp = \mathit{Ker}(A^t).$$

2. $$\mathbb{R}^m = \mathit{Bild}(A) \oplus \mathit{Bild}(A)^\perp = \mathit{Bild}(A) \oplus \mathit{Ker}(A^t)$$

3. $$\mathit{Ker}(A^t A) = \mathit{Ker}(A).$$

Hierbei ist $\oplus$ die orthogonale Summe.

Beweis:

1. Sei $y \in \text{Ker}(A^t) \subset \mathbb{R}^m$, $x \in \mathbb{R}^n$ beliebig.

   $$y \in \text{Ker}(A^t) \Rightarrow A^t y = 0 \Rightarrow 0 = (A^t y, x) = (y, Ax) \Rightarrow y \in \text{Bild}(A)^\perp,$$

   also $\text{Ker}(A^t) \subset \text{Bild}(A)^\perp$. Sei nun $y \in \text{Bild}(A)^\perp$. Dann gilt

   $$0 = (y, AA^t y) = (A^t y, A^t y) = ||A^t y||_2^2$$

und damit $y \in \text{Ker}(A^t)$, insgesamt also $\text{Ker}(A^t) = \text{Bild}(A)^\perp$.

2. Klar nach 1.
3. Übungen.

□

Satz 5.3 *(**Gauss–Normalgleichung**)*
Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. $x^* \in \mathbb{R}^n$ ist genau dann kleinste Quadrate–Lösung von $Ax = b$, falls

$$A^t A x^* = A^t b.$$

Diese Gleichung heißt Gausssche Normalgleichung. Die Menge der kleinste Quadrate–Lösungen ist nicht leer.

Beweis: Nach Lemma 5.2 gibt es b_1, b_2, x^* mit

$$b = b_1 + b_2,\ b_1 \in \text{Bild}(A) \Rightarrow b_1 = Ax^*\,, b_2 \in \text{Ker}(A^t),\ b_1 \perp b_2.$$

Sei $x \in \mathbb{R}^m$. Dann gilt

$$||Ax - b||_2^2 = ||\underbrace{Ax - b_1}_{\in \text{Bild}(A)} - \underbrace{b_2}_{\in \text{Ker}(A^t)}||_2^2 = ||Ax - b_1||_2^2 + ||b_2||_2^2 \geq \underbrace{||Ax^* - b_1||_2^2}_{=0} + ||b_2||_2^2.$$

x^* ist also kleinste Quadrate–Lösung.
Weiter ist ein $x \in \mathbb{R}^n$ genau dann kleinste Quadrate–Lösung, wenn $||Ax - b_1||_2^2 = 0$.
Damit gilt $Ax = b_1 = Ax^*$, also

$$A\,(x - x^*) = 0 \overset{\text{Lemma 5.2}}{\Longleftrightarrow} 0 = A^t A\,(x - x^*) = A^t Ax - A^t Ax^* = A^t Ax - A^t(b_1 + \underbrace{b_2}_{\in \text{Ker}(A^t)}),$$

also $A^t Ax = A^t b$.
□
Bemerkung: Seien x_1^*, x_2^* zwei kleinste Quadrate–Lösungen. Dann gilt

$$x_1^* - x_2^* \in \text{Ker}(A^t A) = \text{Ker}(A).$$

Die kleinste Quadrate–Lösung ist also genau dann **eindeutig**, wenn A den Rang n hat. Im Allgemeinen ist sie es **nicht**.

Beispiel 5.4

1. *Eine feste Länge L wird m–mal gemessen mit Ergebnissen l_1 bis l_m. Das zugehörige überbestimmte Gleichungssystem lautet*

$$\begin{array}{rcl} L & = & l_1 \\ L & = & l_2 \\ & \vdots & \\ L & = & l_m \end{array} \Rightarrow \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}_{A} L = \underbrace{\begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{pmatrix}}_{b}.$$

Es gilt

$$A^t A = (1, \ldots, 1) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = m$$

und

$$A^t b = (1, \ldots, 1) \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_m \end{pmatrix} = l_1 + l_2 + \ldots + l_m.$$

Für die kleinste Quadrate–Lösung L^* erhalten wir also

$$mL^* = A^t A L^* = A^t b = l_1 + l_2 + \ldots + l_m$$

und damit

$$L^* = \frac{\sum_{i=1}^{m} l_i}{m},$$

also, nicht sehr überraschend, den Mittelwert der l_i.

2. *Zu Zeitpunkten t_i werden die Messwerte y_i gemessen, $i = 1 \ldots 4$.*

t_i	-2	0	1	1
y_i	-2	-4	4	6

Es wird ein linearer Zusammenhang der Form $y(t) = at + b$ vermutet. Wir bestimmen die Ausgleichsgerade. Das überbestimmte Gleichungssystem lautet

$$\begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} -2 \\ -4 \\ 4 \\ 6 \end{pmatrix}.$$

Die Normalgleichung lautet

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ -2 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ -2 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} -2 \\ -4 \\ 4 \\ 6 \end{pmatrix}$$

also

$$\begin{pmatrix} 4 & 0 \\ 0 & 6 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 4 \\ 14 \end{pmatrix}$$

und damit erhalten wir die Ausgleichsgerade $(7/3)x + 1$.

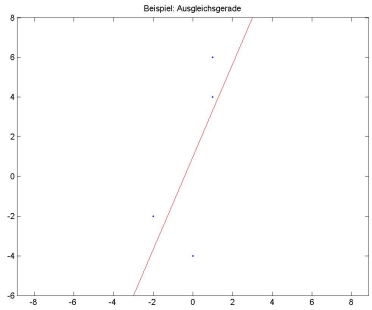


Abbildung 5.1: Beispiel zur Ausgleichsgeraden

Sei A *die* $m \times n$*–Nullmatrix. Dann ist jedes* $x \in \mathbb{R}^n$ *kleinste–Quadrate–Lösung von* $Ax = b$*, denn*

$$A^t Ax = 0 = A^t b.$$

5.2 Die Minimum Norm–Lösung

Im allgemeinen ist die kleinste Quadrate–Lösung nicht eindeutig. Wir wählen in diesen Fällen eine spezielle aus, nämlich die mit kleinster Norm. Dies führt zur Definition der Minimum–Norm–Lösung.

Definition 5.5 *(***Minimum Norm–Lösung, verallgemeinerte Lösung, Moore–Penrose–Lösung***)*
Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. $x^+ \in \mathbb{R}^n$ *heißt Minimum Norm–Lösung von* $Ax = b$ *genau dann, wenn gilt*

1. x^+ *ist kleinste Quadrate–Lösung von* $Ax = b$.

2. x^+ *hat unter allen kleinste Quadrate–Lösungen von* $Ax = b$ *die kleinste Norm, d.h.*

$$||x^+|| \leq ||x^*|| \forall x^* : x^* \textit{ ist kleinste Quadrate–Lösung von } Ax = b$$

.

Bemerkung: Wir machen keine Voraussetzungen an m, n oder den Rang von A.

Diese Definition erlaubt uns noch nicht die direkte Berechnung von x^+. Wir geben wieder eine zum Optimierungsproblem äquivalente nachrechenbare Bedingung an.

Satz 5.6 *(Berechnung und Eindeutigkeit der Minimum Norm–Lösung)*
Sei $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. x^+ *ist genau dann Minimum Norm–Lösung von* $Ax = b$, *falls gilt*

1. $A^t A x^+ = A^t b$.
2. $x^+ \in \textit{Bild}(A^t)$.

Die Minimum Norm–Lösung ist eindeutig bestimmt.

Beweis: Sei x^* eine kleinste Quadrate–Lösung von $Ax = b$. Nach Nach Lemma 5.2, angewandt auf A^t, können wir zerlegen

$$x^* = x^+ \oplus x_2,\ x^+ \in \text{Bild}(A^t),\ x_2 \in \text{Ker}(A).$$

Damit gilt

$$A(x^* - x^+) = Ax_2 = 0$$

und damit

$$A^t A x^+ = A^t A(x^+ + (x^* - x^+)) = A^t A x^* = A^t b,$$

also ist auch x^+ kleinste Quadrate–Lösung. Sei nun $\overline{x}$ eine weitere kleinste Quadrate–Lösung. Mit der Bemerkung zu Satz 5.3 gilt

$$\overline{x} = x^+ + w,\ w \in \text{Ker}(A)$$

und wieder mit Pythagoras

$$||\overline{x}||_2^2 = ||x^+ + w||_2^2 = || \underbrace{x^+}_{\in \text{Bild}(A^t)} + \underbrace{w}_{\in \text{Ker}(A)} ||_2^2 = ||x^+||_2^2 + ||w||_2^2 \geq ||x^+||_2^2.$$

Also ist x^+ Minimum Norm–Lösung. Gleichheit bekommen wir genau dann, wenn $w = 0$, also $\overline{x} = x^+ + w = x^+$, die Minimum Norm–Lösung ist also eindeutig. □

Beispiel 5.7

1. *Sei A die $m \times n$–Nullmatrix, $b \in \mathbb{R}^m$ beliebig. Dann erfüllt jedes $x^* \in \mathbb{R}^n$ die Normalengleichung*

$$A^t A x^* = A^t b$$

und ist damit kleinste Quadrate–Lösung. Die Minimum Norm–Lösung ist unter diesen die mit kleinster Norm, also $x^+ = 0$. Offensichtlich ist x^+ auch eindeutig bestimmt durch die Bedingung

$$x^+ \in \mathit{Bild}(A^t) = \{0\}\,.$$

2. *Wir suchen die Minimum Norm–Lösung des Ausgleichsproblems für eine einzelne Messung (t_1, y_1) und den linearen Ansatz $at + b$. (a, b) ist kleinste Quadrate–Lösung, also muss gelten*

$$\begin{pmatrix} 1 \\ t_1 \end{pmatrix} \begin{pmatrix} 1 & t_1 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 1 \\ t_1 \end{pmatrix} \begin{pmatrix} y_1 \end{pmatrix}$$

also

$$\begin{pmatrix} 1 & t_1 \\ t_1 & t_1^2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} y_1 \\ t y_1 \end{pmatrix}$$

oder

$$b + a t_1 = y_1.$$

Da in diesem Fall Lösungen von $Ax = b$ existieren, sind genau diese natürlich auch die kleinste Quadrate–Lösungen, den Ansatz über die Normalengleichung hätten wir uns also sparen können.
Wegen

$$x^+ \in \mathit{Bild}(A^t) = \left\{A^t u | u \in \mathbb{R}^m\right\} = \left\{(u, t_1 u)^t | u \in \mathbb{R}\right\}$$

gilt für ein u: $b = u$, $a = t_1 u$ und

$$y_1 = b + a t_1 = u + t_1^2 u$$

und damit

$$u = \frac{y_1}{1 + t_1^2}$$

und die gesuchte Gerade ist

$$y(t) = \frac{y_1 t_1}{1 + t_1^2} t + \frac{y_1}{1 + t_1^2}.$$

Insbesondere erfüllt diese Gerade natürlich $y(t_1) = y_1$. Tatsächlich hat diese Gerade eine kleinere 2–Norm auf den Koeffizienten als die eigentlich viel naheliegendere Lösung $y(t) = y_1$.

5.3 Die Pseudoinverse

Falls $A \in \mathbb{R}^{m \times n}$ maximalen Rang hat (also Rang$(A) = \min(n, m)$), so lässt sich die Minimum Norm–Lösung durch Matrixinversion berechnen.

Satz 5.8 Pseudoinverse, Moore–Penrose–Inverse, verallgemeinerte Inverse
Die Abbildung $A^+ \in \mathbb{R}^{n \times m} : \mathbb{R}^m \mapsto \mathbb{R}^n$, $A^+ b = x^+$, x^+ Minimum Norm–Lösung von $Ax = b$, ist linear. A^+ heißt Pseudoinverse (Moore–Penrose–Inverse, verallgemeinerte Inverse) von A.

1. *Falls $n = m$ und A invertierbar, so gilt $A^+ = A^{-1}$.*
2. *Falls $m > n$ und A injektiv (Rang$(A) = n$), so ist $A^t A$ invertierbar und*
$$A^+ = (A^t A)^{-1} A^t.$$
3. *Falls $m < n$ und A surjektiv (Rang$(A) = m$), so ist AA^t invertierbar und*
$$A^+ = A^t (AA^t)^{-1}.$$

Beweis: Sei $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$.

1. Falls A invertierbar ist (Rang$(A) = m = n$), so ist die einzige kleinste Quadrate–Lösung die eindeutige Lösung von $Ax = b$, also gilt
$$A^+ = A^{-1}.$$
2. Für $m > n$ ist der Zielraum in der Dimension größer als der Urbildraum. A kann also nicht surjektiv sein, aber injektiv. Sei A injektiv, d.h. Rang$(A) = n$. Wegen Ker$(A) =$ Ker$(A^t A)$ ist auch $A^t A$ injektiv, also invertierbar. x^+ erfüllt die Normalengleichung
$$A^t A x^+ = A^t b$$
also
$$x^+ = (A^t A)^{-1} A^t b.$$
3. Für $m < n$ ist der Urbildraum in der Dimension größer als der Zielraum. A kann also nicht injektiv sein, aber surjektiv. Sei A surjektiv, d.h. Rang$(A) = m$. Dann gibt es Lösungen von $Az = b$, und genau diese sind die kleinste Quadrate–Lösungen.
Wegen
$$\text{Rang}(A^t) = \text{Rang}(A) = \dim \text{Bild}(A) = m$$

ist A^t injektiv, also ist AA^t invertierbar. Wegen $x^+ \in \text{Bild}(A^t)$ gilt $x^+ = A^t y$ für ein $y \in \mathbb{R}^m$, also

$$b = Ax^+ = AA^t y$$

und damit

$$x^+ = A^t y = A^t (AA^t)^{-1} b.$$

□

Kapitel 6

Iterative Lösung linearer Gleichungssysteme

Wir werden als Grundlage den aus der Analysis bekannten Banachschen Fixpunktsatz beweisen, und daraus iterative Methoden für lineare und nichtlineare Probleme herleiten. Im ganzen Kapitel sind die Matrizen A immer quadratisch und invertierbar, wir betrachten zunächst keine über– oder unterbestimmten Gleichungssysteme.

6.1 Der Banachsche Fixpunktsatz

Definition 6.1 *(**kontrahierend, Fixpunkt**)*
Seien X, Y normierte Räume, $D \subset X$.

1. *Eine Funktion*
$$g : D \mapsto Y$$
heißt kontrahierend in D genau dann, wenn eine Konstante $0 < q < 1$ 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*
$$g(\overline{x}) = \overline{x}.$$

Bemerkung: Sei g kontrahierend. Dann ist g stetig.

Beweis: Sei x_n eine gegen x konvergente Folge, dann gilt

$$||g(x_n) - g(x)|| \leq q||x_n - x|| \mapsto 0.$$

□

Satz 6.2 *(**Banachscher Fixpunktsatz**)*
Sei X ein vollständiger normierter Raum (Banachraum). Sei $\emptyset \neq D \subset X$ abgeschlossen, d.h. jede Cauchyfolge in D konvergiert in D.
Sei $g : D \mapsto D$ kontrahierend. Dann hat g genau einen Fixpunkt.

Beweis: Sei $q < 1$ Kontraktionskonstante von g. Seien zunächst x und y zwei Fixpunkte von g. Dann gilt

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

also $x = y$ wegen $q < 1$. Damit ist der Fixpunkt eindeutig.
Die Existenz zeigen wir konstruktiv und geben eine konvergente Folge an, deren Grenzwert der Fixpunkt ist. Sei $x^{(0)} \in D$ beliebig. Wir definieren in D die Folge $x^{(k)}$ durch

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

$x^{(k)}$ heißt **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

$$\begin{aligned} ||x^{(l)} - x^{(k)}|| &\leq \underbrace{||x^{(l)} - x^{(l-1)}||}_{\leq q^{l-1}||x_1 - x_0||} + \underbrace{||x^{(l-1)} - x^{(l-2)}||}_{\leq q^{l-2}||x_1 - x_0||} + \ldots + \underbrace{||x^{(k+1)} - x^{(k)}||}_{\leq q^k||x_1 - x_0||} \\ &\leq q^k \sum_{j=0}^{l-k-1} q^j ||x^{(1)} - x^{(0)}|| \\ &\leq \frac{q^k}{1-q}||x^{(1)} - x^{(0)}|| \qquad (6.1) \\ &\leq \epsilon \end{aligned}$$

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. □

Korollar 6.3 *(Konvergenz der Fixpunktiteration)*
Seien für g die Voraussetzungen aus 6.2 erfüllt, insbesondere g kontrahierend. Dann konvergiert die Fixpunktiteration

$$x^{(k+1)} = g(x^{(k)}),\ x^{(0)} \in D$$

gegen einen Fixpunkt von g.

Wenn wir die Fixpunktiteration zur approximativen Berechnung des Fixpunkts nutzen wollen, brauchen wir Abschätzungen, wie nah ein Folgeglied bereits am Grenzwert liegt.

Korollar 6.4 Fehlerabschätzung
Seien für g die Voraussetzungen aus 6.2 erfüllt, und q sei die Kontraktionskonstante von g. Es gilt

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

mit (6.1).
Sei $y^{(j)}$ eine zweite Fixpunktiteration mit Startwert $x^{(k)}$. Dann lautet die Abschätzung, angewandt auf $y^{(0)}$

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

oder angewandt auf $y^{(1)}$

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

Mit Hilfe der ersten Abschätzung können wir im Vorhinein (a priori) eine obere Schranke für den Konvergenzfehler angeben. Mit Hilfe der zweiten Abschätzung können wir im Nachhinein (a posteriori), wenn wir also bereits das $k+1$. Folgeglied berechnet haben, ebenfalls eine obere Schranke angeben. Notwendigerweise ist die a priori–Abschätzung eine worst case–Abschätzung, während die a posteriori–Abschätzung auf dem tatsächlichen Folgenverlauf basiert. Deshalb ist normalerweise die a posteriori–Abschätzung deutlich schärfer.

Bevor wir uns einige Beispiele anschauen, bemerken wir ein nützliches Kriterium zur Kontraktionseigenschaft.

Satz 6.5 *(Abschätzung der Kontraktionskonstante für differenzierbare Funktionen) Seien $g : D \mapsto \mathbb{R}^m$, $D \subset \mathbb{R}^n$ abgeschlossen, und g sei stetig differenzierbar. Es gelte*

$$||g'(x)|| \leq q < 1 \, \forall x \in D,$$

wobei $g'(x)$ die Jakobimatrix von g an der Stelle x ist und $||\cdot||$ die induzierte Matrixnorm. Weiter sei D konvex. Dann ist g kontrahierend mit der Kontraktionskonstante q.
Falls $||g'(x)|| \geq 1$ für ein x im Inneren von D, so ist g in einer Umgebung von x nicht kontrahierend.

Beweis: Seien $x, y \in D$. Da D konvex ist, liegt die Strecke von x nach y ganz in D, und nach dem Mittelwertsatz gibt es ein $\xi \in D$ zwischen x und y mit

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

Insbesondere ist damit

$$||g(x) - g(y)|| = ||g'(\xi)(x - y)|| \leq ||g'(\xi)|| \, ||x - y|| \leq \underbrace{\sup_{\xi \in D} ||g'(\xi)||}_{=:q<1} \, ||x - y||.$$

Für die zweite Bemerkung wähle $y = x + \epsilon u$ für ein beliebiges $u \in \mathbb{R}^n$ mit Norm 1 und betrachte $\epsilon \mapsto 0$. Dann konvergiert

$$\frac{||g(x) - g(y)||}{||x - y||} \mapsto ||g'(x)u||.$$

Da $||g'(x)|| \geq 1$ in der induzierten Matrixnorm, ist die rechte Seite nicht durch ein $q < 1$ nach oben beschränkt. Am einfachsten macht man sich dieses Argument für $n = 1$ klar. □

Beispiel 6.6 *1. Es sei*

$$B \in \mathbb{R}^{n \times n}, \, c \in \mathbb{R}^n, \, g : \mathbb{R}^n \mapsto \mathbb{R}^n, \, g(x) := Bx + c.$$

Dann gilt $g'(x) = B$ und g ist kontrahierend genau dann, wenn $||B|| < 1$ in der induzierten Matrixnorm.
Alle Voraussetzungen des Banachschen Fixpunktsatzes sind dann erfüllt, und alle Fixpunktfolgen konvergieren gegen den eindeutig bestimmten Fixpunkt

$$\overline{x} = (I - B)^{-1} c.$$

2. *Es sei*
$$g : \mathbb{R} \mapsto \mathbb{R},\ g(x) := 0.9 \cos x.$$
Es gilt $\forall x \in \mathbb{R}$
$$|g'(x)| = |-0.9 \sin x| \leq 0.9 =: q < 1.$$
Also ist g *kontrahierend mit der Kontraktionskonstanten* $q = 0.9$.
Alle Voraussetzungen des Banachschen Fixpunktsatzes sind dann erfüllt.

3. *Es sei*
$$g : \mathbb{R} \mapsto \mathbb{R},\ g(x) := \cos x.$$
Da $|g'(\pi/2)| = |-\sin(\pi/2)| = 1$*, ist* g *nicht kontrahierend.*

4. *Es sei*
$$g : [-0.1, 0.1] \mapsto \mathbb{R},\ g(x) := \cos x.$$
Es gilt für $x \in [-0.1, 0.1]$
$$|g'(x)| \leq 0.5 < 1,$$
also ist g *kontrahierend.*
g *ist aber keine Selbstabbildung, daher sind die Voraussetzungen von Banach nicht erfüllt.*

5. *Es sei* $D := [0.6, 0.9]$ *und*
$$g : D \mapsto D,\ g(x) := \cos x.$$
g *ist wieder kontrahierend, denn* $|g'(x)| \leq 0.9 < 1$ *für* $x \in D$.
Da g *monoton ist und* $g(0.6) \sim 0.82 \in D$, $g(0.9) \sim 0.62 \in D$ *gilt* $g(D) \subset D$ *und* g *ist korrekt definiert, also Selbstabbildung.*
Also sind alle Voraussetzungen des Fixpunktsatzes von Banach erfüllt.

6. *Bemerkung:* g *kann in einer induzierten Norm kontrahierend sein, in einer anderen nicht. Für die Matrix*
$$B = \frac{1}{10} \begin{pmatrix} 6 & 6 \\ 0 & 6 \end{pmatrix}$$
gilt $||B||_\infty = 1.2$*, aber* $||B||_2 = 0.97$*, es ist also die Fixpunktiteration zu* $g(x) := Bx + c$ *kontrahierend in der euklidischen Norm, aber nicht kontrahierend in der Supremumsnorm.*

7. *Bemerkung: Die Formulierung einer Fixpunktgleichung ist wichtig bei der Betrachtung von Fixpunktfolgen.*
Gesucht sei der Fixpunkt von
$$g(x) := \tan x$$

im Intervall $[\frac{\pi}{2}, \frac{3\pi}{2}]$. *Es gilt*

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

und g *ist nicht kontrahierend. Wir betrachten die Umkehrfunktion. Es gilt*

$$x = \arctan x + \pi =: g(x),\ g'(x) = \frac{1}{1+x^2} < \frac{1}{2} =: q < 1.$$

in unserem Intervall, und g *ist kontrahierend.*

6.2 Iterative Fixpunktverfahren für lineare Gleichungen

Gesucht sei eine Approximation an die Lösung $\overline{x}$ des linearen Gleichungssystems $Ax = b$ mit Genauigkeit ϵ. $A \in \mathbb{R}^{n\times n}$ sei invertierbar und dünn besetzt (sparse), d.h. A habe viele Nullen. Dann vermutet man, dass sich das Gleichungssystem schnell lösen lässt. Manchmal ist dies der Fall (etwa für Bandmatrizen wie der Matrix der Wärmeleitungsgleichung), aber nicht immer. In diesem Fall kommen iterative Methoden zum Einsatz.

Dazu definieren wir jeweils eine Matrix $B \in \mathbb{R}^{n\times n}$ und einen Vektor $c \in \mathbb{R}^n$ so, dass $\overline{x}$ Fixpunkt der Funktion

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

ist.

Falls g die Voraussetzungen des Banachschen Fixpunktsatzes erfüllt, so wählen wir $x^{(0)} \in \mathbb{R}^n$ beliebig und berechnen Folgeglieder $x^{(k)}$ der zugehörigen Fixpunktfolge. Dabei kontrollieren wir in jedem Schritt die Genauigkeit mit Hilfe der a posteriori– und a priori–Abschätzungen. Sobald die Genauigkeit besser ist als ϵ, akzeptieren wir das aktuelle Folgeglied als Approximation.

Da $Ax = b$ nicht in Fixpunktform $g(x) = x$ gegeben ist, müssen wir es umformen. Wir betrachten einige Beispiele.

Beispiel 6.7 *1. Wir lösen die linke Seite nach* A *auf.*

$$Ax = b \Longleftrightarrow x = \underbrace{A^{-1}b}_{c} =: g(x)$$

Das ist zwar korrekt, macht aber keinen Sinn, denn wir müssten A^{-1} *ausrechnen, um* g *zu berechnen. Wenn wir das können, brauchen wir kein iteratives Verfahren.*

2. *Wir bringen das Ax auf die rechte Seite und addieren auf beiden Seiten ein x.*

$$Ax = b \Longleftrightarrow x = x + b - Ax = \underbrace{(I - A)}_{=:B} x + \underbrace{b}_{=:c} =: g(x).$$

Das ist ebenfalls korrekt, und im Grunde die Neumannsche Reihe, aber bei praktischen Problemen ist leider dieses g selten kontrahierend, daher nutzt es uns ebenfalls nichts.

Eine Kombination dieser beiden Verfahren führt zum Erfolg. Wir setzen

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

Hierbei enthält L die Einträge unterhalb der Hauptdiagonalen, D die Diagonaleinträge und R die Einträge oberhalb der Hauptdiagonalen. Zur Berechnung von L,D und R muss man natürlich nicht rechnen, dies ist nur eine Aufteilung.

Wir bringen nun einen Teil von Ax auf die rechte Seite und invertieren. Dies führt zu

Definition 6.8 *Sei $A = L + D + R$ wie oben, und D sei invertierbar.*

1. *Es gilt*

$$Ax = (L+D+R)x = b \Longleftrightarrow Dx = -(L+R)x+b \Longleftrightarrow x = \underbrace{-D^{-1}(L + R)}_{=:B} x+\underbrace{D^{-1}b}_{=:c}.$$

Das zugehörige Fixpunktverfahren

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

heißt Gesamtschritt– oder Jakobi–Verfahren.

2. *Es gilt*

$$Ax = (L+D+R)x = b \Longleftrightarrow (D+L)x = -Rx+b \Longleftrightarrow x = \underbrace{-(D + L)^{-1}R}_{=:B} x+\underbrace{(L + D)^{-1}b}_{=:c}.$$

Das zugehörige Fixpunktverfahren

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

heißt Einzelschritt– oder Gauss–Seidel–Verfahren.
Man berechnet natürlich nicht wirklich die Inverse $(D + L)^{-1}$, sondern löst zur Berechnung von $x = (D + L)^{-1}y$ das Gleichungssystem $(D + L)x = y$ durch Vorwärtseinsetzen (wie im Beispiel).

Die historische Idee von Gauss ist: Wähle einen beliebigen Startvektor $x \in \mathbb{R}^n$.
Ändere x_1 so ab, dass die erste Gleichung erfüllt ist.
Ändere dann x_2 so ab, dass die zweite Gleichung erfüllt ist.
usw., wenn man bei x_n angekommen ist, fängt man wieder bei x_1 an.

Man macht sich sofort klar, dass dies äquivalent zum Einzelschrittverfahren ist.

Beispiel 6.9 *Gegeben sei das Gleichungssystem*

$$\begin{pmatrix} 3 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}.$$

Dann gilt

$$D = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}, \; L = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \; R = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Im Gesamtschrittverfahren gilt

$$B = -D^{-1}(L+R) = -\begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -\begin{pmatrix} 0 & \frac{1}{3} \\ \frac{1}{4} & 0 \end{pmatrix}$$

und

$$c = D^{-1}b = \begin{pmatrix} \frac{4}{3} \\ \frac{5}{4} \end{pmatrix}.$$

Es gilt

$$||B||_\infty = \frac{1}{3} =: q < 1$$

und damit ist die Funktion

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

kontrahierend mit der Kontraktionskonstante q.
Wir wählen $x^{(0)} = 0$, damit gilt $x^{(1)} = c$. Es ist

$$||x^{(1)} - x^{(0)}||_\infty = \frac{4}{3} \Longrightarrow ||x^{(k)} - \overline{x}||_\infty \leq \frac{q^k}{1-q}\frac{4}{3} = 2\left(\frac{1}{3}\right)^k$$

für die a priori–Fehlerabschätzung. Weiter gilt

$$x^{(2)} = Bx^{(1)} + c = \begin{pmatrix} \frac{11}{12} \\ \frac{11}{12} \end{pmatrix} \sim \begin{pmatrix} 0.92 \\ 0.92 \end{pmatrix}$$

Die a posteriori–Abschätzung liefert hier

$$||x^{(2)} - \overline{x}||_\infty \leq \frac{q}{1-q}||x^{(2)} - x^{(1)}||_\infty = \frac{5}{24}.$$

Für das Einzelschrittverfahren gilt

$$D + L = \begin{pmatrix} 3 & 0 \\ 1 & 4 \end{pmatrix}.$$

Wegen $c = (D+L)^{-1}b$ gilt $(D+L)c = b$. Wir berechnen c durch Vorwärtseinsetzen:

$$\begin{pmatrix} 3 & 0 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix} \Longrightarrow c_1 = \frac{4}{3},\ c_2 = \frac{1}{4}(5 - \frac{4}{3}) = \frac{11}{12}.$$

Weiter erhält man

$$q = \frac{1}{3},\ x^{(2)} = \begin{pmatrix} 1.03 \\ 0.99 \end{pmatrix}$$

und damit eine deutlich bessere Approximation als beim Gesamtschrittverfahren.

6.3 Infimum der induzierten Matrixnormen

Wir haben gesehen: Es ist wichtig, die richtige Norm zu wählen, damit $||B|| < 1$ und damit die Fixpunktfolge konvergiert. Die Frage ist natürlich: Wie klein kann man die induzierte Norm einer Matrix machen? Dies beantwortet der folgende Satz.

Satz 6.10 *Sei $B \in \mathbb{R}^{n\times n}$, und $||\cdot||$ eine induzierte Matrixnorm. Dann gilt mit dem Spektralradius $\rho(B)$*

$$||B|| \geq \rho(B).$$

Sei weiter $\epsilon > 0$. Dann gibt es eine induzierte Matrixnorm $||\cdot||_\epsilon$, so dass

$$||B||_\epsilon \leq \rho(B) + \epsilon.$$

Beweis: Sei zunächst λ ein Eigenwert von B zum Eigenvektor y. Dann gilt

$$||B|| = \sup_{x\neq 0} \frac{Bx}{x} \geq \frac{By}{y} = |\lambda|,$$

also $||B|| \geq |\lambda|$.
Den zweiten Teil zeigen wir nur für den (trivialen) Fall dass B symmetrisch positiv definit ist. Dann gilt für die induzierte euklidische Norm

$$||B||_2^2 = \rho(B^t B) = \rho(B^2) = \rho(B)^2$$

und damit

$$||B||_2 = \rho(b) < \rho(B) + \epsilon.$$

□

Korollar 6.11 *Sei $B \in \mathbb{R}^{n\times n}$. Falls $\rho(B) < 1$, so gibt es eine induzierte Norm, so dass die Funktion*

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

kontrahierend ist. Damit konvergiert die zugehörige Fixpunktfolge für alle Startwerte $x^{(0)}$ und alle c gegen den eindeutigen Fixpunkt.
Falls $\rho(B) \geq 1$, so gibt es Startwerte und c, so dass die Fixpunktfolge nicht gegen $(I - B)^{-1}c$ konvergiert.

Bemerkung: Eigentlich bekommt man die Konvergenz zunächst nur in dieser speziellen Norm, da aber in endlichdimensionalen Räumen alle Normen äquivalent sind, konvergiert die Folge für alle Normen. Dies bedeutet natürlich insbesondere: Die Fixpunktfolgen *können* konvergieren, obwohl die Funktion g nicht kontrahierend ist. Kontraktion ist hinreichend, aber nicht notwendig.

Beweis: Zu zeigen ist nur noch der zweite Teil des Korollars. Sei λ Eigenwert von B zum Eigenvektor y und $|\lambda| \geq 1$. Wir setzen $c = 0$, dann ist 0 Fixpunkt von g, und es gilt

$$x^{(k)}y = B^k y = \lambda^k y$$

und dies ist wegen $|\lambda| \geq 1$ keine Nullfolge. □

6.4 Satz von Gerschgorin

Der Satz von Gerschgorin liefert eine grobe Abschätzung über die Lage der Eigenwerte einer Matrix.

Satz 6.12 *Sei $A \in \mathbb{C}^{n\times n}$. Weiter sei*

$$r_i := \sum_{k\neq i} |A_{i,k}|$$

und

$$K_i := \{x : |x - A_{i,i}| \leq r_i\}$$

der Kreis um $A_{i,i}$ mit Radius r_i in der komplexen Ebene.
Dann sind alle Eigenwerte von A in der Vereinigung der K_i enthalten.
Falls l der K_i disjunkt sind vom Rest, so sind in der Vereinigung dieser Kreise genau l Eigenwerte enthalten, der Vielfachheit nach gezählt.

Beweis: Sei λ ein Eigenwert von A zum Eigenvektor x mit $||x||_\infty = 1$. Dann gibt es ein m mit $|x_m| = 1$, und $|x_k| \leq 1 \forall k$. Dann gilt

$$0 = (A - \lambda I)_m = \sum_k A_{m,k}x_k - \lambda x_m = (A_{m,m} - \lambda)x_m - \sum_{k\neq m} A_{m,k}x_k$$

und damit

$$|A_{m,m} - \lambda| \leq \sum_{k \neq m} |A_{m,k}| \, |x_k| \leq r_m,$$

also $\lambda \in K_m$.
Zum Beweis des zweiten Teils sei $B_t \in \mathbb{R}^{n \times n}$,

$$(B_t)_{i,k} = \begin{cases} B_{i,k}, & i = k \\ tB_{i,k}, & i \neq k \end{cases}$$

Also ist B_0 eine Diagonalmatrix mit den Diagonaleinträgen von B, und $B_1 = B$.
Es gilt der Satz: Die Nullstellen eines Polynoms hängen stetig von den Koeffizienten ab.
Seien $\lambda_k(t)$ die Nullstellen von B_t. Dann ist λ_k ein stetiger Weg in den komplexen Zahlen. Nach Teil 1 muss dieser ganz in der Vereinigung aller Kreise liegen. Da alle λ_k auf den Diagonalelementen beginnen, können sie die Vereinigung der l disjunkten Kreise nicht verlassen, und der Satz gilt. □

Beispiel 6.13 *1. Wärmeleitungsmatrix: Es sind K_1 und K_n Kreise um 2 mit Radius 1, alle anderen sind Kreise um 2 mit Radius 2.*

2.

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

K_1 ist der Kreis um 4 mit Radius 1, K_2 der Kreis um 0 mit Radius 1, K_3 der Kreis um 1 mit Radius 1.
Die Vereinigung von K_2 und K_3 ist disjunkt von K_1, also liegen in der Vereinigung genau zwei Eigenwerte und in K_1 liegt genau ein Eigenwert, jeweils der arithmetischen Vielfachheit nach gezählt.

6.5 Zeilensummenkriterien

Definition 6.14 *(**strikte Diagonaldominanz, starkes Zeilensummenkriterium**)*
Sei $A \in \mathbb{R}^{n \times n}$. Seien

$$r_i = \sum_{i \neq k} |a_{i,k}|$$

die Radien der Gerschgorinkreise und

$$r_i < |a_{i,i}| \, \forall i \in \{1 \dots n\}.$$

Dann heißt A strikt diagonaldominant, oder auch A erfüllt das starke Zeilensummenkriterium.

Korollar 6.15 *Sei A strikt diagonaldominant. Dann ist A invertierbar, Einzel– und Gesamtschrittverfahren konvergieren gegen die Lösung von $Ax = b$.*

Beweis: Da $|a_{ii} - 0| > r_i$, ist 0 nicht im i–ten Gerschgorinkreis enthalten, also ist 0 kein Eigenwert von A und damit A invertierbar.
Wir zeigen nur die Konvergenz des GSV. Es ist zu zeigen, dass

$$\rho(B) = \rho(D^{-1}(L + R)) < 1.$$

$B_{i,i} = 0$, also sind alle Gerschgorinkreise Kreise um 0 mit Radius $r_i/|a_{i,i}| < 1$, also sind alle Eigenwerte kleiner als 1. □

Dieses Kriterium ist nicht für unser Standardbeispiel, die Wärmeleitungsmatrix, anwendbar. Wir definieren daher

Definition 6.16 *(schwache Diagonaldominanz, schwaches Zeilensummenkriterium)*
Es sei $A \in \mathbb{R}^{n\times n}$. Seien wieder $r_i := \sum_k |a_{i,k}|$ die Radien der Gerschgorinkreise. A heißt schwach diagonaldominant, falls:

1. $|a_{i,i}| \geq r_i$
2. $\exists m : |a_{m,m}| > r_m$
3. *A ist unzerlegbar, d.h.*

$$\forall I \subset \{1 \dots n\}, \emptyset \neq I \neq \{1 \dots n\}\, \exists i \in I,\, k \notin I : |a_{i,k}| \neq 0.$$

Unzerlegbarkeit bedeutet, dass man die Matrix nicht in zwei Teile zerlegen kann, die nichts miteinander zu tun haben. Beispiel für eine zerlegbare Matrix wäre

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

(wähle $I = \{2\}$). Die Matrix der Wärmeleitung ist unzerlegbar und damit schwach diagonaldominant.

Satz 6.17 *$A \in \mathbb{R}^{n\times n}$ erfülle das schwache Zeilensummenkriterium 6.16. Dann ist A invertierbar, Gesamt– und Einzelschrittverfahren konvergieren gegen die Lösung von $Ax = b$.*

Beweis: Wir zeigen wieder nur das Gesamtschrittverfahren. Sei

$$Ax = \lambda x,\ ||x||_\infty = 1.$$

Wir gehen exakt so vor wie beim Beweis zu Gerschgorin und betrachten zwei Fälle.

$|x_m| = 1$: Dann gilt wie bei Gerschgorin

$$|\lambda| = |\lambda x_m| = |(Bx)_m| \leq \frac{1}{|a_{m,m}|} \sum_{k \neq m} |a_{m,k} x_k| \leq \frac{r_m}{|a_{m,m}|} < 1.$$

$|x_m| < 1$: Sei dann i mit $|x_i| = 1$ (mindestens eins gibt es, denn $||x||_\infty = 1$). Wie gerade argumentieren wir

$$|\lambda| = |\lambda x_i| = |(Bx)_m| \leq \frac{1}{|a_{i,i}|} \sum_{k \neq i} |a_{i,k} x_k| \underbrace{\leq}_{!!!} \frac{r_i}{|a_{i,i}|} \leq 1.$$

Das ist blöd – für Zeile i haben wir nun $r_i \leq |a_{i,i}|$, und damit $|\lambda| \leq 1$, und das hilft uns nicht, wir brauchen ein $<$.
Wann steht bei !!! ein $<$-Zeichen? Sicherlich dann, wenn mindestens ein $|x_k| < 1$ ist und das zugehörige $a_{i,k}$ nicht Null ist. Wir setzen nun in Teil 3 von 6.16

$$I = \{i : |x_i| = 1\}.$$

Dann gibt es i, k mit $i \in I$, $k \notin I$ (also $|x_i| = 1$, $|x_k| < 1$) mit $a_{i,k} \neq 0$, und der Satz ist bewiesen. □

Kapitel 7

Das Newton–Verfahren

Sei $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ differenzierbar und nichtlinear. Gesucht sei eine Nullstelle $\overline{x}$ von f.

Das bekannteste Verfahren zur Bestimmung einer Approximation an diese Lösung ist das Newton–Verfahren. Wir betrachten es hier der Einfachheit halber für $n = 1$.

Sei x eine Näherung für die Nullstelle $\overline{x}$ von f. Wir approximieren die Funktion f durch ihre Tangente im Punkt $(x, f(x))$, also

$$y(t) = (t - x)f'(x) + f(x).$$

Die Tangente schneidet die x-Achse im Punkt

$$\widetilde{x} = x - \frac{f(x)}{f'(x)}.$$

Falls die Tangente eine gute Approximation für die Funktion ist, so erwarten wir, dass $\widetilde{x}$ eine bessere Approximation für $\overline{x}$ ist als x. Dies wird nun iteriert.

Definition 7.1 *Sei $f : \mathbb{R} \mapsto \mathbb{R}$ differenzierbar, $x^{(0)} \in \mathbb{R}$, und $\overline{x}$ eine Nullstelle von f. Sei*

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

Falls die Folge $(x^{(k)})$ wohldefiniert ist und gegen $\overline{x}$ konvergiert, so heißt sie Newtonverfahren zur Bestimmung von $\overline{x}$.

Satz 7.2 *Sei alles wie in 7.1. Dann gibt es eine Umgebung D von $\overline{x}$, so dass das Newtonverfahren für $x^{(0)} \in D$ gegen $\overline{x}$ konvergiert.*

Beweis: Zunächst mal ist $\overline{x}$ genau dann Nullstelle von f, wenn $\overline{x}$ Fixpunkt von g ist. Es reicht also, die Voraussetzungen des Fixpunktsatzes nachzurechnen. Es gilt

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$$

für $f'(x) \neq 0$.
Sei zunächst $f'(\overline{x}) \neq 0$. Wegen $f(\overline{x}) = 0$ gilt dann $g'(\overline{x}) = 0$.
f' und g' sind stetig. Also gibt es mit dem ϵ–δ–Kriterium ein $\epsilon > 0$ mit

$$|g'(x)| \leq \frac{1}{2} =: q \text{ und } |f'(x)| \geq \frac{|f'(\overline{x})|}{2} =: \mu > 0 \, \forall x \in [\overline{x} - \epsilon, \overline{x} + \epsilon] =: D.$$

Insbesondere ist damit wegen $|f'(x)| \neq 0$ auf D die Folge wohldefiniert. Damit ist schon mal g kontrahierend auf D.
Sei nun $x \in D$, also $|x - \overline{x}| \leq \epsilon$. Dann gilt

$$|g(x) - \overline{x}| = |g(x) - g(\overline{x})| \leq q|x - \overline{x}| \leq \frac{\epsilon}{2}$$

und damit auch $g(x) \in D$. Also ist g eine Selbstabbildung von D nach D, kontrahierend auf D, D ist abgeschlossen und konvex, und $\mathbb{R}$ ist vollständig. Also konvergiert die Fixpunktfolge $x^{(k)}$ gegen den eindeutigen Fixpunkt $\overline{x}$ von g, und das Newtonverfahren ist konvergent.
Zur Konvergenzgeschwindigkeit: Wir bemerken zunächst

$$0 = f(\overline{x}) = f(x) + f'(x)(\overline{x} - x) + f''(\xi)\frac{1}{2}(\overline{x} - x)^2$$

mit einem ξ zwischen x und $\overline{x}$, also $\xi \in D$. Wir setzen darin für x das $x^{(k)}$ ein, lösen nach $f(x^{(k)})$ auf und erhalten

$$\begin{aligned}
|\overline{x} - x^{(k+1)}| &= |\overline{x} - x^{(k)} + \frac{f(x^{(k)})}{f'(x^{(k)})}| \\
&= |\overline{x} - x^{(k)} + \frac{-f'(x^{(k)})(\overline{x} - x^{(k)}) - f''(\xi)\frac{1}{2}(\overline{x} - x^{(k)})^2}{f'(x^{(k)})} \\
&\leq \frac{1}{\mu} \, ||f''||_\infty \, \frac{1}{2}(\overline{x} - x^{(k)})^2, \; ||f''||_\infty = \sup_{x \in D} |f''(x)|.
\end{aligned}$$

Bis auf eine Konstante quadriert sich der Fehler in jedem Schritt, dies nennen wir auch quadratisch konvergent.
Wir wissen bereits, dass die Folge konvergiert. Sei der Abstand $|\overline{x} - x^{(k)}| < 10^{-3}$. Dann erhalten wir in den nächsten Schritten die Genauigkeiten 10^{-6}, 10^{-12}, 10^{-24} (bis auf Konstante). Das Newtonverfahren konvergiert für $f'(\overline{x}) \neq 0$ extrem schnell.

Sei nun $f'(\overline{x}) = 0$. Wir betrachten ohne Einschränkung den Spezialfall $f'(x) \neq 0$ in einer kleinen Umgebung außerhalb von $\overline{x}$, $f''(\overline{x}) \neq 0$, $f \in C^4$.
In diesem Fall ist g im Punkt $\overline{x}$ zwar nicht definiert, aber stetig fortsetzbar. Mit L'Hospital gilt:

$$\frac{f(x)}{f'(x)} \longrightarrow \frac{f'(x)}{f''(x)} \longrightarrow_{x \longrightarrow \overline{x}} 0$$

und damit ist $g(\overline{x}) = \overline{x}$ stetige Fortsetzung von g, und $\overline{x}$ ist wieder Fixpunkt von g. Auch g' ist stetig fortsetzbar und damit Ableitung von g:

$$\begin{aligned} g'(x) &= \frac{f(x)f''(x)}{f'(x)^2} \\ &\longrightarrow \frac{f(x)f'''(x) + f'(x)f''(x)}{2f'(x)f''(x)} \\ &\longrightarrow \frac{f(x)f''''(x) + f'(x)f'''(x) + f'(x)f'''(x) + f''(x)f''(x)}{2\left(f''(x)^2 + f'(x)f'''(x)\right)} \\ &\longrightarrow_{x \longrightarrow \overline{x}} \frac{1}{2}. \end{aligned}$$

Damit ist g auch in diesem Fall kontrahierend und wie oben Selbstabbildung auf einer kleinen Umgebung von $\overline{x}$. Leider klappt der Quadrat–Trick in diesem Fall nicht, die Konvergenz ist erheblich langsamer. □
Nochmal: Dieser Satz sagt nicht aus, dass das Newton–Verfahren grundsätzlich konvergiert. Die Fixpunktfunktion g des Newton–Verfahrens ist Selbstabbildung und kontrahierend **in einer kleinen Umgebung D von** $\overline{x}$. Also muss auch $x^{(0)}$ in dieser Umgebung liegen, damit der Satz anwendbar ist, d.h. der Startwert darf nicht zu weit von der gesuchten Stelle entfernt liegen. Dies heißt auch lokal konvergent (im Gegensatz zu den global konvergenten Einzel– und Gesamtschrittverfahren, wenn sie die Voraussetzungen erfüllen).

Beispiel 7.3 *(Verfahren von Heron)*
Es sei $f(x) = x^2 - a$, $a > 0$. Gesucht sei die (einzige) positive Nullstelle $\overline{x}$ von f, also $\overline{x} = \sqrt{a}$. Es gilt

$$f'(x) = 2x,\ g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2 - a}{2x} = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

und damit für das Newton–Verfahren

$$x^{(k+1)} = g(x^{(k)}) = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right).$$

Behauptung: Sei $x^{(0)} > 0$. Dann konvergiert das Newton–Verfahren.

1. *Mit $x^{(0)} > 0$ sind alle $x^{(k)} > 0$, insbesondere wohldefiniert.*
2. *$g(x)$ hat ein globales Minimum bei $\sqrt{a}$ für $x > 0$, denn*

$$0 = g'(\widetilde{x}) = \frac{1}{2}\left(1 - \frac{a}{\widetilde{x}^2}\right) \Longrightarrow \widetilde{x} = \sqrt{a},\ g''(\sqrt{a}) = \frac{1}{\sqrt{a}} > 0 \Longrightarrow g(x) \geq g(\sqrt{a}) = \sqrt{a}.$$

   Also ist g Selbstabbildung auf $D := [\sqrt{a}, \infty)$.

3. *Sei $x \in D$. Dann gilt*

$$|g'(x)| = \frac{1}{2}|1 - \frac{a}{x^2}| \leq \frac{1}{2}.$$

   D ist konvex, damit ist g kontrahierend auf D, und das Newtonverfahren konvergiert für $x^{(0)} \in D$ gegen den einzigen Fixpunkt $\overline{x} = \sqrt{a}$ von g.

Sei nun $0 < x^{(0)} \leq \sqrt{a}$. Dann gilt $x^{(1)} \geq \sqrt{a}$, und die Folge ist wieder konvergent.

Beispiel 7.4 *Es sei $f(x) = a - \frac{1}{x}$, $a > 0$. Gesucht sei die einzige positive Nullstelle von f, also $\overline{x} = \frac{1}{a}$. Es gilt*

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

Behauptung: Sei $x^{(0)} \in (0, \frac{1}{a}] =: D$. Dann konvergiert das Newtonverfahren

$$x^{(k+1)} = x^{(k)}(2 - ax^{(k)})$$

gegen $\overline{x} = \frac{1}{a}$. (Bemerkung: Hier können wir den Fixpunktsatz nicht direkt benutzen, denn D ist nicht abgeschlossen!)

1. $$x \in D \Longrightarrow 2 - ax \geq 1 \Longrightarrow g(x) \geq x \Longrightarrow x^{(k)} \text{ ist monoton steigend}.$$

2. $$0 = g'(\widetilde{x}) = 2(1 - a\widetilde{x}) \Longrightarrow \widetilde{x} = \frac{1}{a},\ g''(\widetilde{x}) = -2a < 0 \Longrightarrow g(x) \leq \frac{1}{a}.$$

   Damit ist die Folge monoton und beschränkt, also konvergent.

3. *Es gilt $x^{k+1} = g(x^{(k)})$. Wir betrachten auf beiden Seiten $k \longrightarrow \infty$ und erhalten, da g stetig ist,*

$$\lim x^{(k)} = g(\lim x^{(k)}).$$

   Also ist der Grenzwert Fixpunkt von g, also konvergiert die Folge gegen $\overline{x} = \frac{1}{a}$.

Kapitel 8

Eigenwerte

Bei der Berechnung der 2–Norm einer Matrix und bei der Konvergenz des Gesamt– und Einzelschrittverfahrens tauchte der betragsgrößte Eigenwert einer Matrix auf. Wir betrachten die Potenzmethode, eine Möglichkeit, diesen zu berechnen.

Definition 8.1 *(Potenzmethode)*
Sei $A \in \mathbb{R}^{n\times n}$. Sei

$$x^{(0)} \in \mathbb{R}^n,\ x^{(k+1)} = Ax^{(k)},\ \textit{also}\ x^{(k)} = A^k x^{(0)}.$$

Sei weiter

$$d \in \mathbb{R}^n,\ \alpha^{(k)} = \frac{(x^{(k+1)}, d)}{(x^{(k)}, d)}.$$

Dann heißt $\alpha^{(k)}$ Potenzmethode zur Bestimmung des betragsmaximalen Eigenwerts von A.

Satz 8.2 *(Konvergenz der Potenzmethode)*
Sei alles wie in 8.1. Seien λ_k die Eigenwerte von A, der arithmetischen Vielfachheit nach gezählt und nach dem Betrag geordnet, d.h.

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Falls

1. *Falls $\lambda_1 = \ldots = \lambda_r$ und $|\lambda_r| > |\lambda_{r+1}|$, so konvergiert $\alpha^{(k)}$ gegen λ_1 für fast alle d und $x^{(0)}$. Falls die arithmetische und geometrische Vielfachheit von λ_1 übereinstimmen, so ist die Konvergenz schnell (wie $\left(\frac{\lambda_{r+1}}{\lambda_1}\right)^k$). Andernfalls ist die Konvergenz langsam (wie $\frac{1}{k}$).*

2. *Falls es ein λ_r gibt mit $|\lambda_1| = |\lambda_r|$, $\lambda_1 \neq \lambda_r$, so konvergiert $\alpha^{(k)}$ für fast alle d, $x^{(0)}$ nicht.*

Beweis: Wir betrachten zunächst den Fall, dass A diagonalisierbar ist.
Sei dann x_i eine Basis aus Eigenvektoren von A mit zugehörigen Eigenwerten λ_i. Dann gibt es Koeffizienten μ_i mit

$$x^{(0)} = \sum_{i=1}^{n} \mu_i x_i \Longrightarrow x^{(k)} = A^k x^{(0)} = \sum_{i=1}^{n} \mu_i A^k x_i = \sum_{i=1}^{n} \mu_i \lambda_i^k x_i.$$

Es sei nun die erste Bedingung 1 erfüllt. Dann gilt

$$\begin{aligned} x^{(k)} &= \sum_{i=1}^{r} \mu_i \lambda_1^k x_i + \sum_{i=r+1}^{n} \mu_i \lambda_i^k x_i \\ &= \lambda_1^k \underbrace{\left(\underbrace{\sum_{i=1}^{r} \mu_i x_i}_{=:y} + \underbrace{\sum_{i=r+1}^{n} \mu_i \frac{\lambda_i^k}{\lambda_1^k} x_i}_{\mapsto_{k\mapsto\infty} 0} \right)}_{=:r^{(k)}} \end{aligned}$$

Damit konvergiert $r^{(k)}$ gegen y. y ist Linearkombination der Eigenvektoren zum Eigenwert λ_1. Falls $y \neq 0$, so ist y ebenfalls ein Eigenvektor zu λ_1. Damit y Null ist, muss gelten $\mu_i = 0$, $i = 1 \dots r$. Dies ist nur für eine Nullmenge im $\mathbb{R}^n$ der Fall, also ist $y \neq 0$ für fast alle Startwerte $x^{(0)}$.
Weiter gilt

$$\alpha^{(k)} = \frac{(x^{(k+1)}, d)}{(x^{(k)}, d)} = \frac{\lambda_1^{k+1}(r^{(k+1)}, d)}{\lambda_1^k (r^{(k)}, d)} \mapsto \lambda_1 \frac{(y, d)}{(y, d)} = \lambda_1,$$

falls $(y, d) \neq 0$. $(y, d) = 0$ nur für $d \in y^\perp$, also wieder eine Nullmenge.
Die Konvergenzgeschwindigkeit ist gegeben durch die von $r^{(k)}$, also gerade

$$\left(\frac{\lambda_{r+1}}{\lambda_1} \right)^k.$$

Sei nun Bedingung 2 erfüllt, und ohne Einschränkung

$$|\lambda_1| = |\lambda_2|,\ \lambda_1 \neq \lambda_2,\ |\lambda_1| > |\lambda_3|.$$

Dann gilt wie oben

$$x^{(k)} = \lambda_1^k \left(\underbrace{\mu_1 x_1}_{=:y} + \underbrace{\mu_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k x_2}_{\text{divergiert}} + \underbrace{\sum_{i=3}^{n} \mu_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i}_{\mapsto 0} \right).$$

Der erste und dritte Term ist derselbe wie oben. Der zweite Term konvergiert aber nicht wegen $|\lambda_1/\lambda_2| = 1$, $\lambda_1/\lambda_2 \neq 1$, für fast alle $x^{(0)}$, und damit auch nicht das $\alpha^{(k)}$.

Zu zeigen ist noch, was bei nicht diagonalisierbaren Matrizen passiert. Wir betrachten nur den Spezialfall

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{=:N} = \lambda I + N.$$

Die Matrix hat nur den Eigenwert λ (doppelte Nullstelle des charakteristischen Polynoms), aber nur einen Eigenvektor, d.h. hier ist die geometrische Vielfachheit kleiner als die arithmetische.
N ist nilpotent, d.h. $N^2 = 0$. Mit der binomischen Reihe gilt

$$A^k = (\lambda I + N)^k = \lambda^k I + k\lambda^{k-1} N.$$

Alle anderen Summanden fallen weg wegen $N^2 = 0$. Wir setzen wieder an wie oben und erhalten

$$x^{(k)} = A^k x^{(0)} = k\lambda^k \underbrace{\left(\underbrace{\frac{x^{(0)}}{k}}_{\mapsto 0} + \frac{1}{\lambda} N x^{(0)} \right)}_{=:r^{(k)}}$$

(und für diese Matrix ist $Nx^{(0)}$ tatsächlich Eigenvektor). Damit gilt

$$\alpha^{(k)} = \frac{(x^{(k+1)}, d)}{(x^{(k)}, d)} = \underbrace{\frac{k+1}{k}}_{\mapsto 1} \lambda \underbrace{\frac{(r^{(k+1)}, d)}{(r^{(k)}, d)}}_{\mapsto 1} \mapsto \lambda$$

aber beide Terme konvergieren nur wie $\frac{1}{k}$. Dieses Argument lässt sich auf jede Matrix übertragen (ist aber sehr technisch). □

Um die Potenzmethode tatsächlich anzuwenden, würden wir einige Folgeglieder ausrechnen und stoppen, wenn die Näherung gut genug ist. Dazu brauchen wir einen Satz, der sagt, wie weit unsere Näherung noch von einem Eigenwert entfernt liegt. Wir betrachten dies nur für symmetrische Matrizen.

Satz 8.3 *(Fehlerabschätzung für Eigenwerte)*
Es sei $A \in \mathbb{R}^{n \times n}$ symmetrisch. Seien $\widetilde{\lambda} \in \mathbb{R}$ und $\widetilde{x} \in \mathbb{R}^n$ Näherungen für Eigenwert und Eigenvektor von A.
Sei $d := A\widetilde{x} - \widetilde{\lambda}\widetilde{x}$. Dann gibt es einen Eigenwert λ von A mit

$$|\widetilde{\lambda} - \lambda| \leq \frac{||d||_2}{||\widetilde{x}||_2}.$$

Beweis: Sei x_i eine ONB aus Eigenvektoren von A zu Eigenwerten λ_i. Dann gilt

$$\widetilde{x} = \sum_{i=1}^{n} \mu_i x_i,\ \mu_i = (\widetilde{x}, x_i),\ ||x||_2^2 = \sum_{i=1}^{n} \mu_i^2$$

(nach Beispiel 2 in 3.15). Weiter gilt

$$d = A\widetilde{x} - \widetilde{\lambda}\widetilde{x} = \sum_{i=1}^{n}(\mu_i A x_i - \widetilde{\lambda}\mu_i x_i) = \sum_{i=1}^{n} \mu_i(\lambda_i - \widetilde{\lambda})x_i$$

und damit

$$||d||_2^2 = \sum_{i=1}^{n} \mu_i^2(\lambda_i - \widetilde{\lambda})^2 \geq \sum_{i=1}^{n} \mu_i^2(\min_j(\lambda_j - \widetilde{\lambda})^2) = ||\widetilde{x}||_2^2 \min_j(\lambda_j - \widetilde{\lambda})^2.$$

Insgesamt

$$\min_j |\lambda_j - \widetilde{\lambda}| \leq \frac{||d||_2}{||\widetilde{x}||_2}$$

und dies war zu zeigen. □

Mit Hilfe dieses Satzes können wir also abschätzen, wie gut die Approximation, die wir aktuell haben, ist, und geeignet stoppen. Der Satz hat eine zweite interessante Folgerung.

Korollar 8.4 *(Kondition des Eigenwertproblems)*
Statt A sei nur eine Näherung $\widetilde{A} = A + dA$ bekannt, und es seien $A,\ dA,\ \widetilde{A} \in \mathbb{R}^{n\times n}$ symmetrisch. Sei $\widetilde{\lambda}$ Eigenwert von $\widetilde{A}$. Dann gibt es einen Eigenwert λ von A mit

$$|\lambda - \widetilde{\lambda}| \leq ||dA||_2.$$

Beweis: Sei $\widetilde{x}$ ein zugehöriger Eigenvektor von $\widetilde{A}$. Wir betrachten $\widetilde{\lambda}$ und $\widetilde{x}$ als Näherungen für Eigenwert und Eigenvektor von A und wenden 8.3 an.

$$||d||_2 = ||A\widetilde{x} - \widetilde{\lambda}\widetilde{x}||_2 = ||(\widetilde{A} - dA)\widetilde{x} - \widetilde{\lambda}\widetilde{x}||_2 = ||dA\,\widetilde{x}||_2$$

und damit gibt es einen Eigenwert λ von A mit

$$|\widetilde{\lambda} - \lambda| \leq \frac{||dA\,\widetilde{x}||_2}{||\widetilde{x}||_2} \leq ||dA||_2.$$

□

Der Satz sagt also: Wenn wir eine Matrix mit dem Fehler dA stören, verschieben sich die Eigenwerte höchstens um $||dA||_2$, insbesondere führen kleine Fehler in der Matrix zu kleinen Fehlern bei der Eigenwertberechnung. Das Eigenwertproblem für symmetrische Matrizen ist gut konditioniert.

Beispiel 8.5

1. *Sei A ähnlich zu*
$$\begin{pmatrix} -3 & & \\ & 2 & \\ & & 1 \end{pmatrix}.$$
Dann konvergiert die Potenzmethode schnell gegen -3.

2. *Sei A ähnlich zu*
$$\begin{pmatrix} 3 & & \\ & 3 & \\ & & 1 \end{pmatrix}.$$
Dann konvergiert die Potenzmethode schnell gegen 3.

3. *Sei A ähnlich zu*
$$\begin{pmatrix} 3 & 1 & \\ & 3 & \\ & & 1 \end{pmatrix}.$$
Dann konvergiert die Potenzmethode langsam gegen 3.

4. *Sei A ähnlich zu*
$$\begin{pmatrix} 3 & & \\ & -3 & \\ & & 1 \end{pmatrix}.$$
Dann konvergiert die Potenzmethode nicht.

Im letzten Fall kann man mit Shifts die Konvergenz erzwingen. Wir betrachten statt A die Matrix $\widetilde{A} = A + I$. Die Addition verschiebt die Eigenwerte um 1 nach oben, $\widetilde{A}$ hat die Eigenwerte $4, -2, 2$, und die Potenzmethode konvergiert gegen 4. Wir ziehen die 1 wieder ab und erhalten einen betragsmaximalen Eigenwert, 3.

Manchmal sucht man auch den betragskleinsten Eigenwert. Setze dann hier $\widetilde{A} = A^{-1}$. Die Eigenwerte von $\widetilde{A}$ sind die Kehrwerte der Eigenwerte von A, insbesondere ist der betragsgrößte Eigenwert von $\widetilde{A}$ der Kehrwert des betragskleinsten Eigenwerts von A. Falls $\widetilde{A}$ die Bedingung zur Konvergenz erfüllt, so konvergiert die Potenzmethode also gegen den Kehrwert des betragskleinsten Eigenwerts von A.

Korollar 8.6 *(Inverse Iteration nach Wieland)*
Sei $A \in \mathbb{R}^{n\times n}$, $\mu \in \mathbb{R}$, und
$$\widetilde{A} = (A - \mu I)^{-1}.$$

Die Potenzmethode zu $\widetilde{A}$ konvergiere gegen $\widetilde{\lambda}$. Dann ist der zu μ nächstgelegene Eigenwert λ gegeben durch

$$\lambda = \frac{1}{\widetilde{\lambda}} + \mu.$$

Beweis: Seien λ_k die Eigenwerte von A, dann sind die Eigenwerte von $\widetilde{A}$ gegeben durch

$$\frac{1}{\lambda_k - \mu}.$$

□

Kapitel 9

Interpolation

Bei der allgemeinen Interpolationsaufgabe wird eine Funktion f gesucht, die an vorgegebenen Stützstellen vorgegebene Stützwerte annimmt. Also:

Definition 9.1 *(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.

9.1 Polynominterpolation

Definition 9.2 (Aufgabe der Polynominterpolation, Polynomraum)
Sei $N \geq 0$. Dann ist $\mathcal{P}_N$ der Raum der Polynome vom Grad kleiner oder gleich N.
Seien $x_0, \dots, x_N$ paarweise verschieden, $y_0, \dots, y_N$ gegeben. Dann ist die Aufgabe der **Polynominterpolation***:*
Finde ein $p \in \mathcal{P}_N$ mit $p(x_i) = y_i \, \forall i = 0 \dots N$.

Damit gilt:

Satz 9.3 *Die Aufgabe der Polynominterpolation 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 9.4 (Vandermondematrizen)
Es seien x_i, $i = 0 \ldots N$ paarweise verschieden. Die Matrix $V \in \mathbb{C}^{(n+1)\times(n+1)}$, $V_{ik} = (x_i)^k$, $i, k = 0 \ldots N$, heißt Vandermondematrix zu $x_0, \ldots, x_N$.

Also:

$$V(x_0, \ldots, x_N) = \begin{pmatrix} x_0^0 & \ldots & x_0^N \\ \vdots & \ddots & \vdots \\ x_N^0 & \ldots & x_N^N \end{pmatrix}$$

Satz 9.5 (Invertierbarkeit der Vandermondematrizen)
Seien $x_0, \ldots, x_N$ paarweise verschiedene Zahlen, $y_0, \ldots, y_N$ in $\mathbb{R}$ oder $\mathbb{C}$. Sei $p(x) = \sum_{k=0}^{N} a_k x^k$. Sei $y = (y_0, \ldots, y_N)^t$, $a = (a_0, \ldots, a_N)^t$, $V = V(x_0, \ldots, x_N)$ Vandermonde–Matrix zu $x_0, \ldots, x_N$. Dann gilt:

1. *p ist genau dann Lösung des Polynominterpolationsproblems, 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 9.3, 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 9.6 (Abschätzung des Interpolationsfehlers)

Sei $f \in C^{(N+1)}([a,b])$, $f : [a,b] \mapsto \mathbb{R}$. Seien x_k paarweise verschieden in $[a,b]$, $k = 0 \ldots N$, 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

$$F(x) := (f(x) - p(x)) - Kw(x),\ K = \frac{f(\overline{x}) - p(\overline{x})}{w(\overline{x})}.$$

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

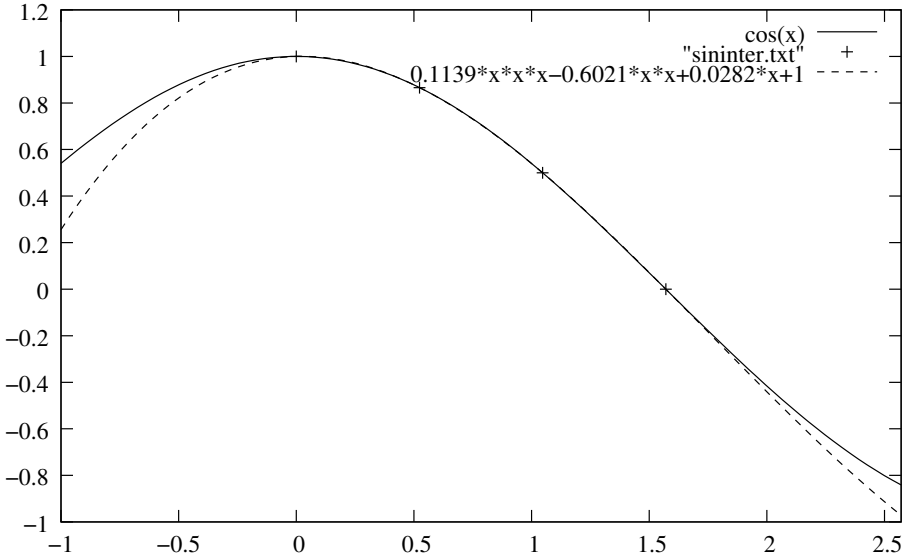
$$0 = F^{(N+1)}(\xi) = f^{(N+1)}(\xi) - K(N+1)! \Rightarrow K = \frac{f^{(N+1)}(\xi)}{(N+1)!}$$

und damit

$$0 = F(\overline{x}) = f(\overline{x}) - p(\overline{x}) - \frac{f^{(N+1)}(\xi)}{(N+1)!} w(\overline{x}).$$

□

Beispiel 9.7 (Interpolation des Cosinus)

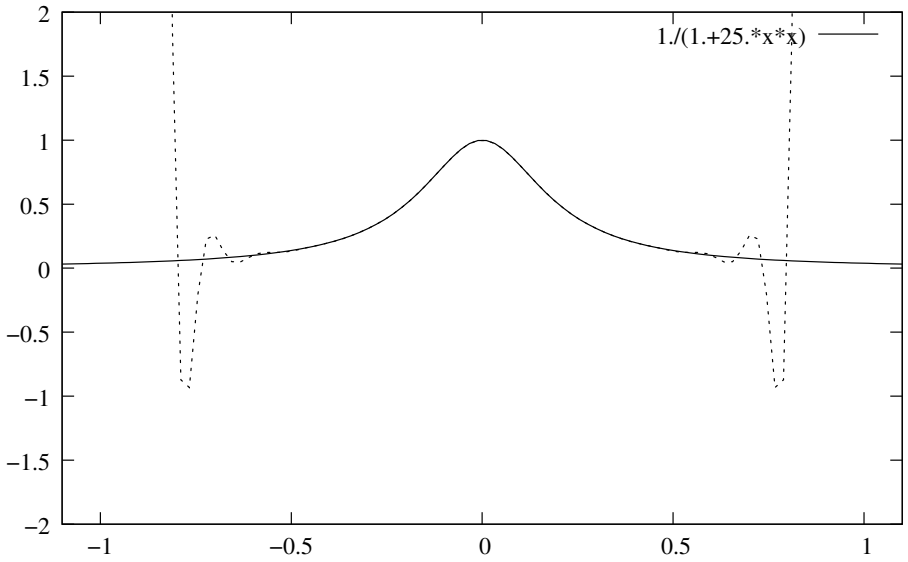


Interpolation des Cosinus auf $[0, \pi/2]$ mit vier Stützpunkten. Die Approximation ist bereits so exakt, dass innerhalb des von den Stützstellen abgedeckten Intervalls kaum ein Unterschied zwischen dem Cosinus und dem Interpolationspolynom vom Grade 3 sichtbar ist. Außerhalb steigt dagegen der Fehler schnell dramatisch an.

Beispiel 9.8 (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 9.6 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 9.9 *(***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 9.10 Eigenschaften der Tschebyscheff–Polynome
Für die Tschebyscheff–Polynome T_n gilt:

1. $T_n \in \mathcal{P}_n$. *Für $n > 0$ hat $T_n(x)$ den Höchstkoeffizienten 2^{n-1}.*
2. *Die Nullstellen von T_{n+1} sind*
$$x_k^n = \cos\left(\frac{2k+1}{2(n+1)}\pi\right), k = 0 \dots n.$$
3. *Wählt man für eine Polynominterpolation vom Grad n die Stützstellen x_k^n, $k = 0 \dots n$, 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 9.10 und 9.6 die Abschätzung

$$||f - p||_\infty \leq \frac{||f^{(N+1)}||_\infty}{2^N (N+1)!}.$$

9.2 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 9.11 (Splines)
Seien $s_0 < s_1 < \ldots < s_n$ reelle Zahlen. Eine Funktion

$$s : [s_0, s_n] \mapsto \mathbb{R}$$

heißt Spline der Ordnung k (zu den Knoten $s_0 \ldots s_n$), falls

1. $s \in C^{(k-2)}([s_0, s_n])$ *für* $k > 1$.
2. $s|_{[s_i, s_{i+1})} \in \mathcal{P}_{k-1}$, $i = 0 \ldots n-1$.

Üblicherweise wird der Spline über sein eigentliches Definitionsgebiet hinaus fortgesetzt, z.B. linear oder periodisch.

Beispiel 9.12

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. Ohne Beweis, Sie finden diesen z.B. in meinem Skript zur Numerischen Analysis, Satz 2.48. Im Wesentlichen muss man dazu zweimal partiell integrieren.

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 9.13 *(Konvergenz von Splines der Ordnung 1)*
Sei $f : [a, b] \mapsto \mathbb{R} \in C^1([a, b])$, und x_k, $k = 0 \ldots n$, seien äquidistant verteilt in $[a, b]$, also

$$x_k = a + kh,\ h = (b-a)/n.$$

Weiter sei

$$s_0 = a,\ s_k = \frac{x_{k-1} + x_k}{2},\ k = 1 \ldots n, s_{n+1} = b.$$

Es sei $s^{(n)}$ der Spline der Ordnung 1 mit $s^{(n)}(x_k) = f(x_k)$ zu den Knoten $s_0, \ldots, s_{n+1}$. Dann gilt

$$||s^{(n)} - f||_\infty = O(h) \mapsto_{n \mapsto \infty} 0.$$

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)

$$||s^{(n)} - f||_\infty = O(h^2).$$

Kapitel 10

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 zugrudeliegende Idee: Wenn eine Operation (Integration, Differentiation) nicht direkt möglich ist, berechne das Interpolationspolynom und führe die Operation auf dem Polynom durch.

10.1 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 10.1

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 10.2

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 nur (2), der Rest in den Übungen. 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:

$$D_h f(x) = f'(x) + \frac{h^2}{12}(f'''(\xi_1) - f'''(\xi_2)) = f'(x) + O(h^2).$$

□

10.2 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

$$\begin{aligned}\int_a^b f(x)\,dx &\sim \int_a^b p(x)\,dx \\ &= \int_a^b \sum_{k=0}^{N} f(x_k) \prod_{j\neq k} \frac{x-x_j}{x_k-x_j}\,dx \qquad \text{Lagrange-Form} \\ &= \sum_{k=0}^{N} \underbrace{\int_a^b \prod_{j\neq k} \frac{x-x_j}{x_k-x_j}\,f(x_k)\,dx}_{=:A_k} \\ &= \sum_{k=0}^{N} A_k f(x_k).\end{aligned}$$

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 $N = 2$ erhält man $x_0 = a$, $x_1 = (a+b)/2$, $x_2 = b$, $h = (b-a)/2$ und

$$A_0 = \frac{1}{3}h,\ A_1 = \frac{4}{3}h, A_2 = \frac{1}{3}h,$$

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 10.3 *(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

$$C_N = \frac{1}{(N+1)!}\int_a^b |w(x)|\,dx \leq \frac{||w||_\infty}{(N+1)!}(b-a) \leq \frac{(b-a)^{N+2}}{(N+1)!} = O(h^{N+2}).$$

2. *Sei N gerade, $f \in C^{(N+2)}$, und die Stützstellen seien nach Newton–Cotes gewählt, also*

$$x_k = a + kh,\ h = \frac{b-a}{N},\ k = 0 \ldots N.$$

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

$$p \in \mathcal{P}_N,\ p(x_k) = f(x_k),\ k = 0 \ldots N.$$

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: Da die Stützstellen alle symmetrisch zur Mitte liegen bei Newton–Cotes und N gerade ist, gilt

$$w(a+x) = -w(b-x) \Longrightarrow \int_a^b w(x)\,dx = 0.$$

Wir entwickeln das $f^{(N+1)}(\xi(x))$ mit Taylor um die Intervallmitte $(a+b)/2$ und erhalten

$$f^{(N+1)}(\xi(x)) = f^{(N+1)}(\frac{a+b}{2}) + (\xi(x) - \frac{a+b}{2}) f^{(N+2)}(\mu(x))$$

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 (Freund and Hoppe [2007], 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 10.4 *(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

$$|\int_a^b f(x)\,dx - \widetilde{I_N}(f)| = O(h^{N+1}).$$

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 I gilt

$$|\int_I f(x)\,dx - I_N(f)| \leq ch^{N+2}$$

und damit

$$\begin{aligned}|\int_a^b f(x)\,dx - \widetilde{I_N}(f)| &\leq \sum_I |\int_I f(x)\,dx - I_N(f)| \\ &\leq p\,c\,h^{N+2} \\ &= \frac{p(b-a)}{Np} ch^{N+1} \\ &= O(h^{N+1})\end{aligned}$$

und entsprechend für den zweiten Teil. □

10.3 Richardson–Extrapolation

Zu bestimmen sei der Grenzwert $F(0) := \lim_{h\mapsto 0} F(h)$ 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 zunächst 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 zusammengesetzte Trapezregel. Das sieht zunächst unsinnig aus, wir haben oben gesehen, dass (im einfachsten Fall) das Richardson–Verfahren die Genauigkeit von $O(h)$ auf $O(h^2)$ erhöht, und die zusammengesetzte Trapezregel ist ja schon von der Ordnung h^2.

Hier hilft ein Trick. Wir setzen $\widetilde{h} := h^2$ und

$$F(\widetilde{h}) := I_h.$$

I_h ist die zusammengesetzte Trapezregel zur Schrittweite h. Dann gilt:

$$F(\widetilde{h}) = I_h = \int_a^b f(x)\,dx + O(h^2) = \int_a^b f(x)\,dx + O(\widetilde{h}).$$

Richardson würde hier die Genauigkeit von $\widetilde{h}$ auf $\widetilde{h}^2$ erhöhen, d.h. auf $O(h^4)$.

Es seien wieder bekannt die Näherungen I_h mit der Trapezregel für die Schrittweite h und $I_{h/2}$ für die Schrittweite $h/2$. Nach Definition von F ist

$$F(h^2) = I_h,\ F(h^2/4) = I_{h/2}.$$

Diesmal gilt also für das Interpolationspolynom p

$$p(x) = I_h\frac{x - h^2/4}{h^2 - h^2/4} + I_{h/2}\frac{x - h^2}{h^2/4 - h^2}$$

und

$$p(0) = -\frac{1}{3}T_h + \frac{4}{3}T_{h/2}.$$

Kleine Aufgabe: Genaues Hinschauen zeigt – dies ist die Simpson–Regel, und die ist tatsächlich, wie erwartet, von der Ordnung $O(h^4)$.

Kapitel 11

Anfangswertprobleme gewöhnlicher Differentialgleichungen

Wir wiederholen zunächst einige Grundbegriffe aus den Vorbemerkungen und der Analysis II.

Definition 11.1 *(Anfangswertprobleme gewöhnlicher Differentialgleichungen, AWA) Sei*

$$f : [a, b] \times G \mapsto \mathbb{R}^n,\ G \subset \mathbb{R}^n \textit{ offen und zusammenhängend}, y_0 \in G.$$

Die Aufgabe: Bestimme

$$y : [a, b] \mapsto G :\ y'(t) = f(t, y(t))\ \forall t \in [a, b],\ y(a) = y_0$$

heißt Anfangswertproblem.

In Kapitel I haben wir bereits gesehen, dass sich Aufgaben mit höheren Ableitungen in diese Form (mit $n > 1$) zurückführen lassen.

Lemma 11.2 *(Integraldarstellung der Anfangswertaufgabe)*
Sei y stetig. y ist genau dann Lösung der Anfangswertaufgabe 11.1, wenn

$$y(s) = y_0 + \int_a^s f(t, y(t))\, dt\, \forall s \in [a, b].$$

Beweis: Sei y Lösung der Anfangswertaufgabe. Dann gilt

$$y(s) - y(a) = \int_a^s y'(t)\, dt = \int_a^s f(t, y(t))\, dt$$

und die Rückwärtsrichtung durch Einsetzen.

Beispiel 11.3

1. *Skalare lineare Differentialgleichung mit konstantem Koeffizienten*

$$y'(t) = \beta\, y(t),\ y(0) = y_0,\ \beta \in \mathbb{R}.$$

Lösung ist

$$y(s) = y_0 e^{\beta(s-a)}.$$

2. *Allgemeine lineare homogene Differentialgleichung*

$$y'(t) = \beta(t)\, y(t),\ y(0) = y_0,\ \beta \textit{ stetig}.$$

Lösung ist

$$y(s) = y_0 e^{\int_a^s \beta(t)\, dt}.$$

3. *Allgemeine lineare inhomogene Differentialgleichung*

$$y'(t) = \alpha(t) + \beta(t)\, y(t),\ y(0) = y_0,\ \alpha, \beta \textit{ stetig}.$$

Lösung in den Übungen mit Variation der Konstanten.

4. $$y'(t) = 1 + y(t)^2,\ y(0) = 0.$$

Lösung ist $y(s) = \tan(s)$*. Insbesondere hat die Aufgabe keine globale Lösung auf ganz* $\mathbb{R}$ *(denn der Tangens hat einen Pol bei* $\frac{\pi}{2}$*).*

5. $$y'(t) = y(t)^{\frac{1}{3}},\ y(0) = 0.$$

Diese Aufgabe hat die Lösungen

$$y_1(t) = 0,\ y_2(t) = \left(\frac{2}{3}t\right)^{3/2}.$$

Insbesondere ist die Lösung der Aufgabe nicht eindeutig.

Wir beweisen den Satz von Picard–Lindelöf zur eindeutigen Lösbarkeit von Anfangswertaufgaben in einer einfachen Version. Wir bemerken zunächst:

Lemma 11.4 *(Fixpunkteigenschaft der Lösung von AWA)*
Sei alles wie in 11.1. Es sei $X = C^0([a,b], \mathbb{R}^n)$ *der Raum der stetigen Funktionen auf dem Intervall* $[a,b]$ *mit Werten im* $\mathbb{R}^n$*, versehen mit der Supremumsnorm. Dann ist*

X vollständig (Analysis 2).
Weiter sei

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

Dann ist $\overline{y} \in X$ genau dann Lösung der AWA, wenn $\overline{y} = g(\overline{y})$, also wenn $\overline{y}$ ein Fixpunkt von g ist.

Beweis: Integraldarstellung der Differentialgleichung. □

Damit ist schon klar, was wir tun müssen, um die eindeutige Lösbarkeit von AWA zu beweisen: Wir müssen zeigen, dass g die Voraussetzungen des Fixpunktsatzes von Banach erfüllt. Wir beginnen mit

Definition 11.5 *(Lipschitzstetigkeit)*
Es sei $f : D \mapsto X$. f heißt lipschitzstetig mit Lipschitzkonstante L genau dann, wenn

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

Bemerkung:

1. Falls $L < 1$, so ist f kontrahierend.
2. f ist stetig.
3. Falls D konvex und kompakt ist, und f stetig differenzierbar, so ist f lipschitzstetig mit der Lipschitzkonstanten $L = ||f'||_\infty$.
4. Die Funktion
   $$f(y) = y^2,\ f : D \subset \mathbb{R} \mapsto \mathbb{R}$$
   ist lipschitzstetig, falls D beschränkt ist, ansonsten ist sie nicht lipschitzstetig.

Beweis: Genau wie bei den entsprechenden Bemerkungen zu kontrahierenden Funktionen. □

Satz 11.6 *(Picard–Lindelöf)*
Es sei alles wie in 11.1, und f sei stetig. Zusätzlich sei f lipschitzstetig im zweiten Argument mit der Lipschitzkonstanten L, d.h.

$$||f(t, y_1) - f(t, y_2)|| \leq L||y_1 - y_2|| \,\forall t \in [a, b], y_1, y_2 \in G.$$

Dann $\exists \epsilon > 0$: Das AWA besitzt eine eindeutige Lösung auf dem Intervall

$$I_\epsilon := [a, a + \epsilon].$$

Beweis: G ist offen, $y_0 \in G$. Also gibt es eine abgeschlossene Kugel B um y_0 mit Radius $\delta > 0$, die ganz in G liegt.
$I \times B$ ist abgeschlossen und beschränkt, f ist stetig, d.h.

$$M := \sup_{t \in [a,b],\, y \in B} ||f(t,y)|| < \infty.$$

Sei $0 < q < 1$ beliebig, und

$$\epsilon := \min(b-a, \frac{q}{L}, \frac{\delta}{M}).$$

Dann gilt $I_\epsilon \subset I$.
Sei nun g wie in 11.4 für $X = C^0(I_\epsilon, \mathbb{R}^n)$ mit der Supremumsnorm. Wir rechnen die Bedingungen des Fixpunktsatzes nach.

1. Bereits bemerkt: X ist vollständig.
2. Setze $D := C^0(I_\epsilon, B) \subset X$, D ist also der Raum der stetigen Funktionen auf I_ϵ mit Werten in B.
   Sei y_n Folge in D, die gegen $y \in X$ konvergiert. Dann gilt

   $$||y_n(s) - y(s)|| \leq ||y_n - y||_\infty \mapsto 0,$$

   also konvergiert $y_n(s)$ gegen $y(s)$ für alle s. B ist abgeschlossen, $y_n(s) \in B$, also auch $y(s) \in B$, also $y \in D$. Also ist auch D abgeschlossen.
3. Sei $y \in D$, $s \in I_\epsilon$. Nach Definition von g, M, ϵ gilt

   $$\begin{aligned} ||g(y)(s) - y_0|| &= || \int_a^s f(t, y(t))\, dt|| \\ &\leq \int_a^s ||f(t,y(t))||\, dt \\ &\leq (s-a)\, M \leq \epsilon M \leq \delta. \end{aligned} \tag{11.1}$$

   Also gilt $g(y)(s) \in B$, also $g(y) \in D$. Also ist g eine Selbstabbildung von D nach D.

4. Seien $u, v \in D$. Es gilt

$$\begin{aligned}
||g(u) - g(v)||_\infty &= \sup_{s \in I_\epsilon} || \int_a^s f(t, u(t)) - f(t, v(t))\, dt|| \\
&\leq \sup_{s \in I_\epsilon} \int_a^s ||f(t, u(t)) - f(t, v(t))||\, dt \\
&\leq \int_a^{a+\epsilon} L||u(t) - v(t)||\, dt \\
&\leq \int_a^{a+\epsilon} L||u - v||_\infty \\
&\leq \epsilon L||u - v||_\infty \leq q||u - v||_\infty
\end{aligned}$$

   und damit ist g kontrahierend.

Der Fixpunktsatz von Banach liefert das gewünschte Ergebnis. □
Diesen Satz kann man noch etwas verbessern.

Satz 11.7 *(Globaler Satz von Picard-Lindelöf)*
Sei alles wie in 11.6, und es gelte

$$M(b - a) \leq \delta.$$

Dann besitzt die AWA eine Lösung auf dem gesamten Intervall $[a, b]$.

Beweis: In meinem Skript zur Numerischen Analysis.
Beweisidee: Wähle eine gewichtete Supremumsnorm auf X. □

Bemerkung: Sei $\overline{y}$ Lösung der AWA aus 11.6. Dann gilt mit den Bezeichnungen dort $\overline{y} = g(\overline{y})$ und nach 11.1

$$||\overline{y}(s) - y_0|| \leq (s - a)M.$$

Für $n = 1$ bedeutet dies: Der Graph von $\overline{y}$ liegt in einem Dreieck $K_M(a, y_0)$ mit Spitze bei (a, y_0), begrenzt durch Geraden mit Steigung M bzw. $-M$ mit $s \in [a, b]$. Die AWA besitzt eine Lösung auf $[a, b]$, wenn dieses Dreieck ganz im Definitionsgebiet $[a, b] \times G$ von f enthalten ist (Kegelbedingung).

Wir werden im Folgenden immer (stillschweigend) annehmen, dass dies der Fall ist, und damit die Existenz einer eindeutigen Lösung der AWA auf dem gesamten Intervall $[a, b]$ gesichert ist.

Wir bemerken auch noch: Da G offen ist, können wir y_0 etwas nach oben oder unten verschieben, so dass die Kegelbedingung erfüllt bleibt. Außerdem besitzen alle Anfangswertprobleme $y(a') = y'_0$ mit $(a', y'_0) \in K_M(a, y_0)$ eine eindeutige Lösung auf dem Intervall $[a', b]$, denn $K_M(a', y'_0) \subset K_M(a, y)$.

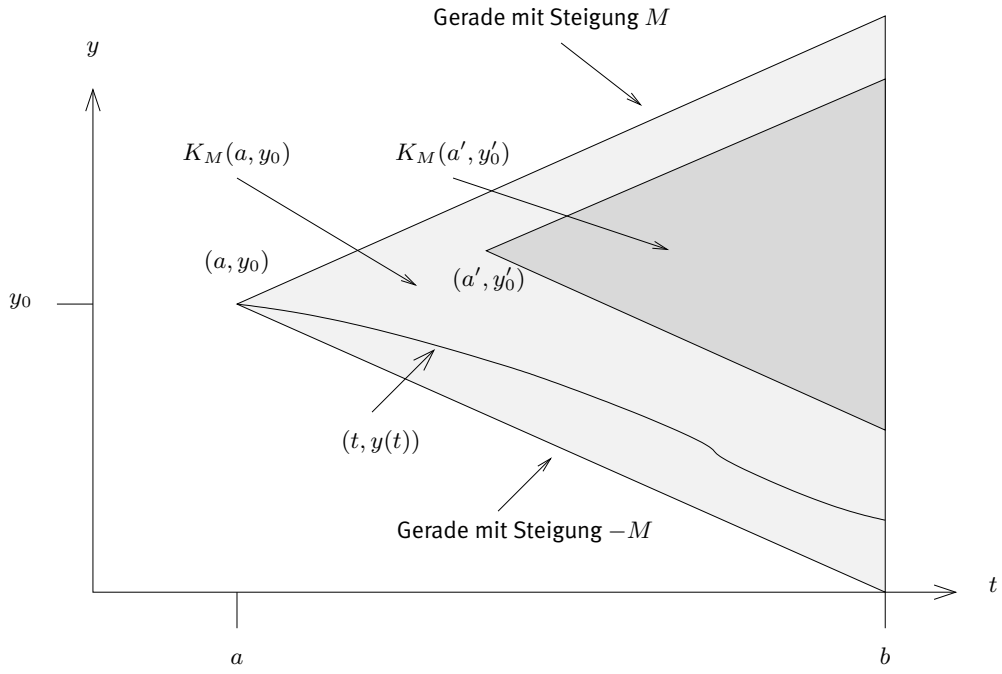


Abbildung 11.1: Kegel $K_M(a, y_0)$

Satz 11.8 (Lemma von Gronwall)

Seien $I = [a, b]$ *und*

$$\alpha : I \mapsto \mathbb{R},\ \beta : I \mapsto \mathbb{R},\ u : I \mapsto \mathbb{R}$$

stetig.

1. *Differentielle Form: Falls* u *differenzierbar ist und*

$$u'(t) \leq \alpha(t) + \beta(t)u(t)\,\forall\, t \in I,$$

so gilt

$$u(t) \leq u(a)e^{\int_a^t \beta(s)ds} + \int_a^t \alpha(s)e^{\int_s^t \beta(\xi)d\xi}ds\,\forall\, t \in I.$$

2. *Integralform: Falls* $\beta \geq 0$ *und*

$$u(t) \leq \alpha(t) + \int_a^t \beta(s)u(s)ds\,\forall\, t \in I,$$

so gilt

$$u(t) \leq \alpha(t) + \int_a^t \alpha(s)\beta(s)e^{\int_s^t \beta(\xi)d\xi}ds\,\forall\, t \in I.$$

Beweis:

1. 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. Sei v die Lösung von

$$v'(t) = -\beta(t)v(t),\ v(a) = 1,$$

also

$$v(t) := e^{-\int_a^t \beta(\xi) d\xi}.$$

Dann gilt, da $v(t) > 0$, mit der Ungleichung aus der differentiellen Form

$$(uv)' = u'v + v'u = u'v - \beta vu \leq \alpha v + \beta uv - \beta vu = \alpha v.$$

Nach 1. gilt damit

$$u(t)v(t) \leq u(a) + \int_a^t \alpha(s)v(s)\mathrm{ds},\ t \in I.$$

Multiplikation mit $1/v(t)$ liefert Teil 1 des Satzes.

3. Sei nun

$$v(t) := \int_a^t \beta(s)u(s)\mathrm{ds}.$$

Mit den Voraussetzungen von Teil 2 gilt dann

$$u(t) \leq \alpha(t) + v(t)$$

und

$$v'(t) = \beta(t)u(t) \leq \beta(t)\alpha(t) + \beta(t)v(t),\ v(a) = 0.$$

Einsetzen in Teil 1 des Satzes liefert

$$v(t) \leq \int_a^t \alpha(s)\beta(s)e^{\int_s^t \beta(\xi)\, d\xi}\, ds$$

und Teil 2 folgt wegen

$$u(t) \leq \alpha(t) + v(t).$$

□

Korollar 11.9 *(Gronwall für konstantes β)*
Es seien $\alpha, u : I \mapsto \mathbb{R}$ stetig, $\beta \in \mathbb{R}$, $\beta \geq 0$, und es gelte

$$u(t) \leq \alpha(t) + \beta \int_a^t u(s)\, ds \, \forall t \in I.$$

Dann gilt

$$u(t) \leq \alpha(t) + \beta \int_a^t \alpha(s) e^{\beta(t-s)}\, ds.$$

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

Satz 11.10 (Stetigkeit des Anfangswertproblems für Anfangswertaufgaben)

Sei

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

ein Anfangswertproblem, das die Voraussetzungen von 11.6 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

$$||\widetilde{y}(t) - y(t)|| \leq (\widetilde{\epsilon} + \epsilon(t-a)) \exp(L(t-a)) \forall t \in [a, b].$$

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

$$\begin{aligned}
u(t) &= ||\widetilde{y}_0 - y_0 + \int_a^t \widetilde{f}(s, \widetilde{y}(s)) - f(s, y(s))\mathsf{ds}|| \\
&\leq ||\widetilde{y}_0 - y_0|| + \int_a^t (||\widetilde{f}(s, \widetilde{y}(s)) - f(s, \widetilde{y}(s)) + f(s, \widetilde{y}(s)) - f(s, y(s)))||\mathsf{ds} \\
&\leq ||\widetilde{y}_0 - y_0|| + \int_a^t (||\widetilde{f}(s, \widetilde{y}(s)) - f(s, \widetilde{y}(s))|| + ||f(s, \widetilde{y}(s)) - f(s, y(s)))||\mathsf{ds} \\
&\leq \widetilde{\epsilon} + \int_a^t (\epsilon + L||\widetilde{y}(s) - y(s)||)\mathsf{ds} \\
&\leq \underbrace{\widetilde{\epsilon} + \epsilon(t-a)}_{\alpha(t)} + \underbrace{L}_{\beta} \int_a^t u(s)\mathsf{ds}.
\end{aligned}$$

Anwendung von 11.9 liefert das Gewünschte:

$$\begin{aligned}
u(t) &\leq (\widetilde{\epsilon} + \epsilon(t-a)) + L \int_a^t \underbrace{(\widetilde{\epsilon} + \epsilon(s-a))}_{\leq \widetilde{\epsilon} + \epsilon(t-a)} \exp(L(t-s))\mathsf{ds} \\
&\leq (\widetilde{\epsilon} + \epsilon(t-a))(1 - [\exp(L(t-s))]_a^t) \\
&= (\widetilde{\epsilon} + \epsilon(t-a)) \exp(L(t-a)).
\end{aligned}$$

□

Kapitel 12

Diskrete Lösung von Anfangswertaufgaben

In den folgenden Kapiteln betrachten wir Anfangswertaufgaben der Form:

Definition 12.1 *(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 globalen Satzes von Picard–Lindelöf erfüllt, d.h. f *stetig,* f *lipschitzstetig im 2. Argument mit Lipschitzkonstante* L*, und die Kegelbedingung sei erfüllt, d.h. der Kegel* K_M *liege ganz im Definitionsgebiet von* f*.*

Daher ist die Lösung eindeutig bestimmt, und jede Anfangswertaufgabe zu dieser Differentialgleichung mit Anfangswerten im Kegel besitzt eine eindeutige Lösung. Wir werden daher im Folgenden die Lösbarkeit nicht genauer betrachten, dies ist immer bereits durch die starken Vorausetzungen gesichert.

Ein numerisches Verfahren bestimmt Näherungen an die Lösung der Differentialgleichung auf einer Teilmenge I_h von $[a, b]$.

Definition 12.2 *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.

Im Folgenden werden wir zunächst immer annehmen, dass die t_k äquidistant verteilt sind auf dem Intervall $[a, b]$, also $t_k = a + kh$ mit $h = (b - a)/N$. Weiter setzen wir zur Motivation $n = 1$.
Es sei $\overline{y}$ die Lösung unserer Anfangswertaufgabe. Wir wollen eine Funktion

$$y_h : I_h \mapsto \mathbb{R}^n$$

bestimmen mit

$$y_k := y_h(t_k) \sim \overline{y}(t_k).$$

In 1.2 haben wir bereits das Eulersche Polygonzugverfahren graphisch motiviert und erhalten die rekursive Definition

$$y_{k+1} := y_k + hf(t_k, y_k).$$

Wir wollen diese Formel noch dreimal analytisch motivieren. Diese drei Zugänge werden später zu unterschiedlichen numerischen Verfahren führen.

Taylorentwicklung: Es gilt

$$\begin{aligned}\overline{y}(t_{k+1}) &= \overline{y}(t_k + h) \\ &\sim \overline{y}(t_k) + h\overline{y}'(t_k) \\ &= \overline{y}(t_k) + hf(t_k, \overline{y}(t_k))\end{aligned}$$

Also

$$y_{k+1} = y_h(t_{k+1}) \sim \overline{y}(t_{k+1}) \sim \overline{y}(t_k) + hf(t_k, \overline{y}(t_k)) \sim y_k + hf(t_k, y_k).$$

Numerische Differentiation: Wir wenden auf die Differentialgleichung den Vorwärts–Differenzenquotienten D_h^+ an und erhalten

$$\frac{\overline{y}(t_k + h) - \overline{y}(t_k)}{h} \sim f(t_k, \overline{y}(t_k))$$

Einsetzen von $\overline{y}(t_k) \sim y_k$ und $\overline{y}(t_k + h) \sim y_{k+1}$ liefert wieder das Gewünschte.

Numerische Integration: In der Integraldarstellung der Differentialgleichung gilt

$$\overline{y}(t_{k+1}) = \overline{y}(t_k) + \int_{t_k}^{t_{k+1}} f(t, \overline{y}(t))\, dt \tag{12.1}$$

Wir verwenden den einzigen Stützpunkt $t = t_k$ für die numerische Integration und erhalten

$$\int_{t_k}^{t_{k+1}} f(t, \overline{y}(t))\, dt \sim hf(t_k, \overline{y}(t_k)) \sim hf(t_k, y_k).$$

Dies ist natürlich alles reine Motivation, wir müssen beweisen, dass diese Ideen gute Approximationen liefern.

Definition 12.3 (globaler Diskretisierungsfehler)

Sei y_h diskrete Näherung für die Lösung $\overline{y}$ der Anfangswertaufgabe 12.1 auf dem Gitter I_h. Dann heißt

$$e_h : I_h \mapsto \mathbb{R}^n,\ e_h := \overline{y}|_{I_h} - y_h$$

die Fehlerfunktion der Näherung.

$$||e_h||_\infty = \max_{t \in I_h} ||e_h(t)||$$

heißt globaler Diskretisierungsfehler.

Der globale Diskretisierungsfehler ist das Maximum 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.

Definition 12.4 (Konvergenz von numerischen Verfahren)

Gegeben sei eine Folge von Gittern I_h, deren Feinheit gegen 0 geht, und ein numerisches Verfahren für 12.1, das jedem Gitter I_h die Näherung y_h zuordnet. Das Verfahren heißt konvergent, falls der globale Diskretisierungsfehler von y_h mit h gegen 0 geht, also

$$||e_h||_\infty \to_{h \to 0} 0.$$

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

$$y_{k+1} = y_k + h\varphi \tag{12.2}$$

Hierbei ist φ ein Ausdruck, in dem Auswertungen der Funktion f, die Gitterpunkte t_k, die Näherungen y_k, und h vorkommen, und heißt Verfahrensfunktion. Für das Eulerverfahren etwa gilt

$$\varphi = f(t_k, y_k).$$

Definition 12.5 *Klassifizierung von Numerischen Verfahren*
Ein Verfahren sei gegeben in der Form 12.2. Dann heißt das Verfahren

Explizites Einschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_k, y_k, h).$$

In diesem Fall wird nur der letzte berechnete Wert genutzt, um den nächsten auszurechnen.

Implizites Einschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_k, y_k, y_{k+1}, h).$$

In diesem Fall muss in jedem Schritt eine Gleichung gelöst werden.

Explizites Mehrschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_{k-r}, \ldots, t_k, y_{k-r}, \ldots, y_k, h).$$

In diesem Fall werden die letzten $r+1$ Näherungen genutzt, um die nächste auszurechnen.

Implizites Mehrschrittverfahren *falls*

$$y_{k+1} = y_k + h\varphi(t_{k-r}, \ldots, t_k, y_{k-r}, \ldots, y_k, y_{k+1}, h).$$

In diesem Fall werden die letzten $r+1$ Näherungen genutzt, um die nächste auszurechnen, und es muss in jedem Schritt eine Gleichung gelöst werden.

Wir behandeln zunächst nur die expliziten Einschrittverfahren. Unser numerisches Verfahren definiert also ein φ, und es gilt

$$\underbrace{y_{k+1}}_{\sim y(t_{k+1})} = \underbrace{y_k}_{\sim y(t_k)} + h\varphi(t_k, \underbrace{y_k}_{\sim y(t_k)}, h).$$

Dies macht schon klar, wie wir das φ wählen sollten, nämlich

$$\varphi(t_k, y(t_k), h) \sim \frac{y(t_k + h) - y(t_k)}{h}.$$

Den Unterschied dieser beiden Terme bezeichnen wir als Konsistenzfehler.

Definition 12.6 *(Konsistenz von expliziten Einschrittverfahren)*
Sei φ die Verfahrensfunktion eines expliziten Einschrittverfahrens. Sei y (irgend-) eine Lösung der Differentialgleichung mit Graph im Kegel K_M.

$$\tau_h(t, y(t)) := \frac{y(t+h) - y(t)}{h} - \varphi(t, y(t), h)$$

heißt Konsistenzfehler oder lokaler Diskretisierungsfehler. Das Verfahren heißt konsistent, falls

$$\sup_{y,t} |\tau_h(t, y(t))| = ||\tau_h||_\infty \to_{h\to 0} 0.$$

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

$$\tau_h(t, y(t)) := \frac{1}{h}(y(t+h) - (y(t) + h\varphi(t, y(t), h)),$$

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 12.7 *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 12.8 *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 12.9 *(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 $\varphi(t_k, y_k, h) = f(t_k, y_k)$, also

   $$y_{k+1} = y_k + hf(t_k, y_k).$$

   Sei f stetig differenzierbar. Dann ist nach 12.7 y zweimal stetig differenzierbar auf I. Es gilt mit Taylorentwicklung und der Differentialgleichung

   $$\begin{aligned}
   \tau_h(t, y(t)) &= \frac{1}{h}(y(t+h) - (y(t) + h\varphi(t, y(t), h))) \\
   &= \frac{y(t+h) - y(t)}{h} - f(t, y(t)) \\
   &= \frac{y(t) + hy'(t) + \frac{h^2}{2}y''(\xi(h)) - y(t)}{h} - y'(t) \\
   &= \frac{h}{2}y''(\xi(h)) \\
   &\leq \frac{||y''||_\infty}{2}h = O(h).
   \end{aligned}$$

   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 12.7 benutzt.

2. **Verbessertes Eulerverfahren:**

   Ein verbessertes Verfahren ergibt sich, wenn wir zur Approximation des Integrals in 12.1 die Mittelpunktregel anwenden und Taylorentwicklung nutzen:

   $$\begin{aligned}
   \int_{t_k}^{t_k+h} f(t, y(t))dt &\sim h(f(t_k + h/2, y(t_k + h/2))) \\
   &\sim h(f(t_k + \frac{h}{2}, y(t_k) + \frac{h}{2}y'(t_k)) \\
   &= h(f(t_k + \frac{h}{2}, y(t_k) + \frac{h}{2}f(t_k, y(t_k)))).
   \end{aligned}$$

   Als Verfahrensfunktion wählen wir also

   $$\varphi(t_k, y_k, h) = f(t_k + \frac{h}{2}, y_k + \frac{h}{2}f(t_k, y_k)).$$

Die Konsistenzordnung weisen wir wieder durch Taylorentwicklung nach. Wir benötigen diesmal, dass f zweimal stetig differenzierbar ist. Damit existiert nach 12.8 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

$$\begin{aligned}
\tau_h(t, y(t)) &= \frac{1}{h}(y(t+h) - (y(t) + h\varphi(t, y(t), h))) \\
&= \frac{y(t+h) - y(t)}{h} - f(t + \frac{h}{2}, y(t) + \frac{h}{2} f(t, y(t))) \\
&= \frac{y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y'''(\xi(h)) - y(t)}{h} - \\
&\quad (f(t, y(t)) + \frac{h}{2} f_t(t, y(t)) + \frac{h}{2} f(t, y(t)) f_y(t, y(t)) + O(h^2)) \\
&= O(h^2).
\end{aligned}$$

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

$$\varphi(t, y, h) = \frac{1}{2}(f(t, y) + f(t + h, y + hf(t, y))).$$

Wieder gilt mit Taylorentwicklung (für f in zwei Dimensionen)

$$\begin{aligned}
\tau_h(t, y(t)) &= \underbrace{\frac{y(t+h) - y(t)}{h}}_{y' + \frac{h}{2}y'' + O(h^2)} - \frac{1}{2} \underbrace{(f(t, y(t)) + f(t + h, y(t) + hf(t, y(t))))}_{y' + (y' + h(f_t + f f_y) + O(h^2)) = 2y' + hy'' + O(h^2)} \\
&= O(h^2).
\end{aligned}$$

Das Verfahren ist also ebenfalls konsistent von der Ordnung 2 und benötigt ebenfalls zwei Auswertungen von f pro Schritt.

Natürlich 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.

Lemma 12.10 (Diskretes Lemma von Gronwall)

Seien (α_k), (β_k), (e_k) reelle nichtnegative Folgen und

$$e_{k+1} \leq \alpha_k + (1+\beta_k)e_k,\ k \geq 0.$$

Dann gilt

$$e_k \leq (e_0 + \sum_{j=0}^{k-1} \alpha_j) \exp(\sum_{j=0}^{k-1} \beta_j).$$

Beweis: Durch vollständige Induktion (Übungen). □

Satz 12.11 (Konvergenz von expliziten Einschrittverfahren)

Ein explizites numerisches Einschrittverfahren zur Lösung der Anfangswertaufgabe 12.1 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 $\overline{y}$ die Lösung der Anfangswertaufgabe 12.1. Sei $I_h = (t_k)$ das äquidistante Gitter mit Feinheit h mit zugehöriger numerischer Approximation y_h. Wir setzen zunächst

$$e_k := ||\overline{y}(t_k) - y_k||$$

(globaler Diskretisierungsfehler an der Stelle t_k).

$$\begin{aligned}
e_{k+1} &= ||\overline{y}(t_{k+1}) - y_{k+1}|| \\
&= ||\overline{y}(t_{k+1}) - (y_k + h\varphi(t_k, y_k, h))|| \\
&= ||\overline{y}(t_{k+1}) - (\overline{y}(t_k) + h\varphi(t_k, \overline{y}(t_k), h)) + \overline{y}(t_k) - y_k \\
&\quad + h(\varphi(t_k, \overline{y}(t_k), h) - \varphi(t_k, y_k, h))|| \\
&\leq h|\tau_h(t_k, \overline{y}(t_k))| + e_k + hL'||\overline{y}(t_k) - y_k|| \\
&= \underbrace{h|\tau_h(t_k, \overline{y}(t_k))|}_{\alpha_k} + \underbrace{(1+hL')}_{\beta_k} e_k.
\end{aligned}$$

Mit dem diskreten Lemma von Gronwall gilt also

$$\begin{aligned} e_k &\leq \left(||e_h(t_0)|| + \sum_{j=0}^{k-1} h|\tau_h(t_j, \overline{y}(t_j))| \right) \exp(L' \sum_{j=0}^{k-1} h) \\ &\leq (e_0 + (t_k - a) \max_j |\tau_h(t_j, \overline{y}(t_j)))| \exp\left(L'(t_k - a)\right) \\ &\leq (e_0 + (b-a)||\tau_h||_\infty \exp\left(L'(b-a)\right) \\ &\to_{h\to 0} 0. \end{aligned}$$

Die Schranke hängt nicht von t_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). Wir dürfen sogar noch zulassen, dass die Anfangswerte falsch sind (bis auf einen Fehler $O(h^p)$). □
Dies lässt sich (für lipschitzstetige Verfahrensfunktionen) in dem Merksatz zusammenfassen:

Für Einschrittverfahren gilt: Aus Konsistenz folgt Konvergenz.

Korollar 12.12 (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.

Kapitel 13

Konvergenz und Konsistenz für implizite Einschrittverfahren

Wir werden sehen, dass implizite Verfahren nützlich sind. Es stellt sich aber die Frage, ob implizite Verfahren überhaupt wohldefiniert sind (d.h. ob die Gleichungen, die sie definieren, eindeutige Lösungen haben), und ob die so entstehenden Verfahren konvergent sind.

Satz 13.1 *(**Wohldefiniertheit für implizite Einschrittverfahren**)*

Sei $\varphi(t_k, y_k, y_{k+1}, h)$ die Schrittfunktion eines impliziten Einschrittverfahrens zur Lösung von 12.1. Sei φ stetig, und lipschitzstetig bzgl. y_k und y_{k+1} mit der Lipschitzkonstanten L'. Dann gibt es ein h_0, so dass die Gleichung

$$y_{k+1} = y_k + h\varphi(t_k, y_k, y_{k+1}, h)$$

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). Es gibt also eine Funktion $v(t_k, y_k, h)$, so dass

$$y_{k+1} = v(t_k, y_k, h).$$

Insbesondere lassen sich implizite Einschrittverfahren als explizite Verfahren der Form

$$y_{k+1} = y_k + h\varphi(t_k, y_k, v(t_k, y_k, h), h)$$

formulieren. Implizite Verfahren sind spezielle explizite Verfahren.

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.

Sei δ so klein, dass für alle Punkte aus dem Kegel K_M auch die δ–Umgebung dieser Punkte noch im Definitionsgebiet von f und damit φ liegt (siehe die Bemerkung zu 11.6). Sei K die Vereinigung aller dieser Umgebungen. K ist kompakt, f ist stetig, also gilt

$$M' = \max_{(t,y),(t,y_1)\in K} \varphi(t, y, y_1, h) < \infty.$$

Sei nun h 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

$$g : [y_k - \delta, y_k + \delta] \mapsto [y_k - \delta, y_k + \delta],\ g(z) := y_k + h\varphi(t_k, y_k, z, h).$$

Der Grundraum $\mathbb{R}$ ist Banachraum, das Intervall, auf dem g definiert ist, ist abgeschlossen. Zu zeigen für den Fixpunktsatz von Banach ist noch: g ist wohldefiniert (Selbstabbildung) und kontrahierend. Sei $z \in [y_k - \delta, y_k + \delta]$.

$$\begin{aligned} |g(z) - y_k| &= |h\varphi(t_k, y_k, z, h)| \\ &\leq M'h_0 \leq \frac{\delta}{2} \end{aligned}$$

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'

$$\begin{aligned} |g(z) - g(z')| &= |h(\varphi(t_k, y_k, z, h) - \varphi(t_k, y_k, z', h))| \\ &\leq h_0 L'|z - z'| \leq q|z - z'|. \end{aligned}$$

Also ist g auch kontrahierend und besitzt mit dem Banachschen Fixpunktsatz 6.2 einen eindeutigen Fixpunkt

$$y_{k+1} = v(t_k, y_k, h).$$

Die Fixpunktiteration für g konvergiert insbesondere für den Startwert y_k gegen y_{k+1}.

□

Implizite Einschrittverfahren sind also explizite Verfahren, nur etwas anders aufgeschrieben. Insbesondere gilt der Konvergenzsatz auch für implizite Verfahren:

Korollar 13.2 (Konvergenzsatz für implizite Einschrittverfahren)
Implizite konsistente Einschrittverfahren (von der Ordnung p) sind konvergent (von der Ordnung p).

Beweis: Implizite sind spezielle explizite Verfahren, das zugehörige φ ist lipschitzstetig (hier ohne Beweis, siehe Skript zur NumAna), also folgt alles mit 12.11 und 13.1. □

Beispiel 13.3 (Beispiele für implizite Verfahren)

1. *Implizites Eulerverfahren:*

$$y_{k+1} = y_k + hf(t_{k+1}, y_{k+1})$$

Wir berechnen die Konsistenzordnung. Sei zunächst f differenzierbar. Sei y irgendeine Lösung der Differentialgleichung. Wir setzen wie bei den expliziten Verfahren

$$y_k = y_h(t_k) = y(t_k),\ y_{k+1} = y_h(t_{k+1}) = y_h(t_k + h) = y(t_k + h),\ t_k = t$$

in die Definition des Verfahrens ein, berechnen die Differenz der linken und rechten Seite und teilen durch h.

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}(y(t+h) - (y(t) + hf(t+h, y(t+h))))\\ &= \frac{1}{h}(hy'(t) + O(h^2) - hy'(t+h))\\ &= y'(t) - y'(t) + O(h)\end{aligned}$$

und damit ist das Verfahren konsistent von der Ordnung 1, also auch konvergent von der Ordnung 1.

2. *Implizite Trapezregel:*

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_{k+1})).$$

Sei f zweimal stetig differenzierbar.

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}(y(t+h) - (y(t) + \frac{h}{2}(f(t, y(t)) + f(t+h, y(t+h)))))\\ &= y'(t) + \frac{h}{2}y''(t) + O(h^2) - \frac{1}{2}(y'(t) + y'(t+h))\\ &= y'(t) - y'(t) + \frac{h}{2}y''(t) - \frac{h}{2}y''(t) + O(h^2)\end{aligned}$$

und damit ist das Verfahren von zweiter Ordnung.

Satz 13.1 gibt auch gleich eine Anleitung zur Durchführung der impliziten Verfahren. Glücklicherweise muss das v aus 13.1 nicht explizit ausgerechnet werden. Die Fixpunktiteration für das g aus dem Satz mit Startwert y_k konvergiert gegen y_{k+1}, also führt man in jedem Schritt des impliziten Einschrittverfahrens zur Lösung der impliziten Gleichung einige Schritte der Fixpunkiteration durch.

Beispiel 13.4 Fixpunktiteration für das implizite Eulerverfahren

Wir wenden das implizite Eulerverfahren an und lösen die Definitionsgleichung durch Fixpunktiterationen. Führen wir in jedem Schritt des Eulerverfahrens einen Schritt der Fixpunktiteration durch, so erhalten wir

$$y_{k+1} = y_k + hf(t_{k+1}, y_k).$$

Dieses Verfahren hat die Ordnung 1 *und benötigt eine Auswertung von* f.

Führen wir zwei Schritte der Fixpunktiteration durch, so erhalten wir

$$y_{k+1} = y_k + hf(t_{k+1}, y_k + hf(t_{k+1}, y_k)).$$

Hier machen wir ein schlechtes Geschäft: Wir benötigen zwei Auswertungen von f *pro Schritt, aber das Verfahren ist trotzdem nur von der Ordnung 1.*

Wir machen also in jedem Schritt jetzt zwei Fehler: Einmal den lokalen Diskretisierungsfehler, und zusätzlich den Abbruchfehler, der dadurch entsteht, dass wir die Fixpunktiteration nach dem $p.$ Schritt abbrechen.

Die Kontraktionskonstante im Fixpunktsatz von Banach war Lh. Wir gewinnen also in jedem Schritt der Fixpunktiteration einen Faktor $O(h)$, nach dem $p.$ Schritt ist unsere Approximation von der Ordnung $O(h^p)$. Der lokale Diskretisierungsfehler erhöht sich also, wenn man die implizite Gleichung durch p Schritte der Fixpunktiteration ersetzt, um $O(h^p)$. Es liegt nahe, das p gerade so zu wählen, dass der lokale Diskretisierungsfehler ebenso groß ist wie der Fehler, der durch den Abbruch der Fixpunktiteration entsteht (mehr hierzu im begleitenden Programmierbeispiel).

Kapitel 14

Anwendungen und Implementation

14.1 Runge–Kutta–Verfahren

Der Runge–Kutta–Ansatz liefert Verfahren beliebig hoher Ordnung, die nur Auswertungen von f benutzen. Wir betrachten zunächst noch einmal das Verfahren von Heun. Die Schrittfunktion φ war hier definiert durch

$$\varphi(t_k, y_k, h_k) = \frac{1}{2}(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k))).$$

Dies schreiben wir in der Form:

$$\begin{array}{rcl} f_1 & = & f(t_k, y_k) \\ f_2 & = & f(t_k + h, y_k + hf_1) \\ \hline \varphi(t_k, y_k, h) & = & \frac{1}{2}f_1 + \frac{1}{2}f_2 \end{array}$$

Diese Schreibweise legt die folgende Definition nahe.

Definition 14.1 (Definition der Runge–Kutta–Verfahren)

1. *Seien α_j, γ_j, β_{jl} fest gewählt, $j = 1 \ldots m$, $l = 1 \ldots j-1$. Die Schrittfunktion φ sei definiert durch*

$$\varphi(t_k, y_k, h) = \gamma_1 f_1 + \gamma_2 f_2 + \ldots + \gamma_m f_m = \sum_{j=1}^{m} \gamma_j f_j$$

mit

$$\begin{aligned} f_1 &= f(t_k + \alpha_1 h, y_k) \\ f_2 &= f(t_k + \alpha_2 h, y_k + h(\beta_{2,1} f_1)) \\ &\vdots \\ f_m &= f(t_k + \alpha_m h, y_k + h \sum_{l=1}^{m-1} \beta_{ml} f_l). \end{aligned}$$

Dann heißt das zugehörige numerische Verfahren m–stufiges explizites Runge–Kutta–Verfahren.

2. *Seien α_k, γ_k, β_{kl} fest gewählt, $k = 1 \ldots m$, $l = 1 \ldots m$. Die Schrittfunktion φ sei definiert durch*

$$\varphi(t_k, y_k, h) = \gamma_1 f_1 + \gamma_2 f_2 + \ldots \gamma_m f_m$$

mit

$$\begin{aligned} f_1 &= f(t_k + \alpha_1 h, y_k + h \sum_{l=1}^{m} \beta_{1,l} f_l) \\ f_2 &= f(t_k + \alpha_2 h, y_k + h \sum_{l=1}^{m} \beta_{2,l} f_l) \\ &\vdots \\ f_m &= f(t_k + \alpha_m h, y_k + h \sum_{l=1}^{m} \beta_{m,l} f_l). \end{aligned}$$

Dann heißt das zugehörige numerische Verfahren m–stufiges implizites Runge–Kutta–Verfahren.

Natürlich sind die expliziten Verfahren implizite Verfahren, bei denen wir setzen

$$\beta_{jl} := 0 \text{ für } l \geq j.$$

Satz 14.2 (Ordnung der Runge–Kutta–Verfahren)

1. *Sei*

$$\sum_{j=1}^{m} \gamma_j = 1.$$

Dann ist das zugehörige Runge–Kutta–Verfahren mindestens konvergent von der Ordnung 1 für alle $f \in C^1$.

2. Sei $f \in C^2$ und

$$\sum_{j=1}^{m} \gamma_j = 1$$
$$\sum_{l} \beta_{jl} = \alpha_j$$
$$\sum_{j=1}^{n} \alpha_j \gamma_j = \frac{1}{2}.$$

Dann ist das zugehörige Runge–Kutta–Verfahren mindestens konvergent von der Ordnung 2.

Wegen dieses Satzes betrachten wir grundsätzlich nur Runge–Kutta–Verfahren, die die Normierungsbedingungen

$$\sum_{j=1}^{m} \gamma_j = 1, \sum_{l} \beta_{jl} = \alpha_j \forall j = 1 \dots m$$

erfüllen.

Beweis: Für Lipschitz–stetiges f ist auch φ Lipschitz–stetig für alle Runge–Kutta–Verfahren.

Zum Beweis der Konsistenz machen wir wieder die Taylorentwicklung. Sei also wieder y irgendeine Lösung der Differentialgleichung. Zunächst berechnen wir eine Taylorentwicklung der auftretenden Zwischenstufen f_j. Wir setzen die Normierung für α gleich ein.

$$\begin{aligned}
f_j &= f(t, y(t)) + \alpha_j h f_t(t, y(t)) + h \sum_{l=1}^{m} \beta_{j,l} f_l f_y(t, y(t)) + O(h^2) \\
&= f(t, y(t)) + \alpha_j h f_t(t, y(t)) + h \sum_{l=1}^{m} \beta_{j,l} (f(t, y(t)) + O(h)) f_y(t, y(t)) + O(h^2) \\
&= y'(t) + h\alpha_j (f_t(t, y(t)) + f(t, y(t)) f_y(t, y(t))) + O(h^2) \\
&= y'(t) + h\alpha_j y''(t) + O(h^2).
\end{aligned}$$

Eingesetzt in die Definition des lokalen Diskretisierungsfehlers

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}(y(t+h) - y(t)) - \varphi(t, y(t), h) \\ &= y'(t) + \frac{h}{2}y''(t) - \sum_j \gamma_j f_j + O(h^2) \\ &= y'(t) - \sum_{j=1}^{m} \gamma_j y'(t) + h(\frac{1}{2} - \sum_{j=1}^{m} \gamma_j \alpha_j) y''(t) + O(h^2).\end{aligned}$$

□

14.2 Energieerhaltung

Dieses Kapitel liegt als Online–Dokument vor (integriertes Python–Notebook).

14.3 Fehlerabschätzung und Schrittweitensteuerung

Wir begnügen uns hier mit einer einfachen Idee.

Nach dem diskreten Lemma von Gronwall kann man den globalen Diskretisierungsfehler abschätzen durch die Summe der Konsistenzfehler. Die Konsistenzfehler sind dabei die Fehler, die in einem Schritt des Verfahrens entstehen. Wir schätzen diese Fehler wie folgt:

In jedem Schritt des Verfahrens werden zwei Verfahren mit unterschiedlicher Konsistenzordnung benutzt. Hierbei sei Verfahren 1 mit Verfahrensfunktion $\varphi^{(1)}$ das genauere. Wir berechnen also

$$\begin{aligned}y_{k+1}^{(1)} &= y_k + \varphi^{(1)}(t_k, y_k) \\ y_{k+1}^{(2)} &= y_k + \varphi^{(2)}(t_k, y_k)\end{aligned}$$

Da das erste Verfahren eine höhere Konsistenzordnung hat, gilt für eine Lösung $\overline{y}$ der Differentialgleichung

$$|y_{k+1}^{(2)} - y(t_k)| \leq |y_{k+1}^{(2)} - y_{k+1}^{(1)}| + |y_{k+1}^{(1)} - y(t_k)| \sim |y_{k+1}^{(2)} - y_{k+1}^{(1)}|.$$

Wir können also eine Approximation des lokalen Diskretisierungsfehlers für das zweite Verfahren nur mit Hilfe der berechneten Werte angeben. Implementationsbeispiele finden Sie wieder als Python–Notebooks (online).

Kapitel 15

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 12.1 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.

Definition 15.1 (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*

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = h \sum_{j=0}^{m} \beta_j f(t_{k+j}, y_{k+j}) =: h \sum_{j=0}^{m} \beta_j f_{k+j},\ k = 0 \dots N - m$$

$(\alpha_m \neq 0)$ 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 y_k bis y_{k+m-1} noch mitbenutzen. Dies bedeutet natürlich auch, 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 Einschrittverfahren berechnet, hat man alle zusammen, macht man ab hier mit den Mehrschrittverfahren weiter. Um diese Anlaufrechnung werden wir uns später kümmern.

Beispiel 15.2 (Mittelpunktregel)

$$y_{k+2} - y_k = 2hf(t_{k+1}, y_{k+1})$$

Definition 15.3 *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 15.4 (Konsistenzordnung der Mittelpunktregel)

Sei $f \in C^2$. Dann hat die Mittelpunktregel die Konsistenzordnung 2.

Beweis: Wir benutzen für

$$y(t+2h) - y(t)$$

jeweils eine Taylorentwicklung um $y(t+h)$. Dann fallen alle Terme in h^j mit geradem j weg, und es gilt

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}(y(t+2h) - y(t)) - 2f(t+h, y(t+h)) \\ &= 2y'(t+h) + O(h^2) - 2y'(t+h).\end{aligned}$$

□

Unsere Idee zur Konstruktion ist: Wir wandeln unsere Differentialgleichung wieder in die äquivalente Integralgleichung um, und benutzen unsere Quadraturformeln. Diese haben wir durch Integration des Interpolationspolynoms hergeleitet.

Wir machen ein einfaches Beispiel. Wir wollen y_{k+1} ausrechnen und dabei die Näherungen y_k und y_{k-1} benutzen. Sei f eine Lösung der Differentialgleichung. Dann schreiben wir wieder

$$y(t_{k+1}) - y(t_k) = \int_{t_k}^{t_{k+1}} f(s, y(s))\, ds.$$

Uns stehen die Näherungen y_k für $y(t_k)$ und y_{k-1} für $y(t_{k-1})$ zur Verfügung. Sei $f_j = f(t_j, y_j)$. Sei dann 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 aus der Gleichung

$$y_{k+1} - y_k = \int_{t_k}^{t_{k+1}} p(s)\, ds.$$

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

$$\int_{t_k}^{t_{k+1}} p(s)\, ds = h(\frac{3}{2} f_k - \frac{1}{2} f_{k-1}).$$

Insgesamt erhalten wir das Verfahren

$$y_{k+1} = y_k + h(\frac{3}{2} f_k - \frac{1}{2} f_{k-1}).$$

Die Ordnung dieses Verfahrens ist 2, denn wir benutzen zwei Vorgänger zur Berechnung der nächsten Approximation. Im Sinne der Definition schreiben wir dieses Verfahren mit Indexshift in der Form

$$y_{k+2} = y_{k+1} + h(\frac{3}{2}f_{k+1} - \frac{1}{2}f_k).$$

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 12.1 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	$r = 1$	$r = 2$	
$f_k \ldots f_{k+m-1}$	Adams–Bashforth	Nyström	explizit
$f_k \ldots f_{k+m}$	Adams–Moulton	Milne–Simpson	implizit

Satz 15.5 *(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 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

$$\begin{aligned}\tau_h(t, y(t)) &= \frac{1}{h}(y(t+mh) - y(t+(m-r)h)) - \sum_{j=0}^{m-1} \beta_j f(t+jh, y(t+jh)) \\ &= \frac{1}{h}\int_{t+(m-r)h}^{t+mh} y(t)\mathsf{dt} - \frac{1}{h}\int_{t+(m-r)h}^{t+mh} p(t)\mathsf{dt} \\ &= O(h^m)\end{aligned}$$

nach 10.3. □

Eine schöne Übersicht über all diese Verfahren mit Rechnungen und Beispielen findet sich (mit leicht anderen Bezeichnungen) in Hairer et al. [1993], 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 15.6 *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 der Vorlesung: Das geht gewaltig schief. Für $\alpha > 1$ und $\alpha < -1.5$ ist der Algorithmus nicht einmal konvergent. Wir müssen vermuten: Der Algorithmus ist dort zwar konsistent, aber wegen mangelnder Stabilität ist er nicht mehr konvergent. Dies versteht man durch einen kleinen Ausflug in die Theorie der linearen Differenzengleichungen.

Kapitel 16

Stabilität von Mehrschrittverfahren

Aufgrund des letzten Beispiels vermuten wir bereits, dass der Satz "Aus Konsistenz folgt Konvergenz" für die Mehrschrittverfahren nicht ohne weitere Bedingungen korrekt ist. Dass er zumindest manchmal korrekt ist, folgte ebenfalls aus den Beispielen.

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, die wir sofort einführen werden. Den korrekten Beweis finden Sie in meinem Skript zur Vorlesung Numerische Analysis, und er ist sehr technisch. Wir wollen daher hier eine vereinfachte Version beweisen, die den wesentlichen Trick zeigt, aber nicht vollständig ist.

Wir zitieren ohne Beweis: Beim Beweis der Konvergenz konsistenter Verfahren reicht es, sich auf die Auswirkungen der Fehler in den Anlaufwerten $y_0, \ldots, y_{m-1}$ des Modellproblems

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

zu beschränken. Also:

Satz 16.1 *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 $(y_h)_j \to_{h\to 0} 0$, $j = 0, \ldots, m-1$, gilt: Die Folge (y_h) der Näherungen für das Modellproblem konvergiert gegen 0, so ist das Mehrschrittverfahren stabil für*

alle Anfangswertaufgaben. Es gilt: Aus Konsistenz des Mehrschrittverfahrens (der Ordnung p) folgt 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. 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 16.2 *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 16.3 *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

$$(y^{(l,r)})_j = j^r \, x_l^j, \; r = 0 \ldots \sigma_l - 1.$$

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 offensichtlich

$$\begin{aligned} 0 &= \rho(x_l) \\ &= x_l^k \rho(x_l) \\ &= \sum_{j=0}^{m} \alpha_j x_l^{k+j} \\ &= \sum_{j=0}^{m} \alpha_j (y^{(l,0)})_{k+j} \end{aligned}$$

und damit ist $y^{(l,0)}$ Lösung der Differenzengleichung.
Sei nun x_l eine doppelte Nullstelle von ρ, also $\sigma_l \geq 2$. Dann gilt

$$\rho(x_l) = 0 = \rho'(x_l) \Longrightarrow (t^{k+1}\rho(t))'(x_l) = 0.$$

Eingesetzt

$$\begin{aligned} 0 &= \sum_{j=0}^{m} \alpha_j (k+j+1) x_l^{k+j} \\ &= \sum_{j=0}^{m} \alpha_j (k+j) x_l^{k+j} + \sum_{j=0}^{m} \alpha_j x_l^{k+j} \\ &= \sum_{j=0}^{m} \alpha_j (y^{(l,1)})_{k+j}. \end{aligned}$$

Also ist für $\sigma_l \geq 2$ auch $y^{(l,1)}$ eine Lösung der Differenzengleichung, usw. □

Beispiel 16.4 *(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 16.5 *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 16.6 *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 16.7 *(Wurzelbedingung von Dahlquist)*
Eine Differenzengleichung heißt stabil, wenn ihre Basislösungen beschränkt sind, d.h. wenn die Bedingungen aus 16.6 erfüllt sind.

Mit den Vorbemerkungen:

Korollar 16.8 *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 16.9 *(Stabilität von konkreten 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 15.6 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).*

Kapitel 17

Errata

- Abschnitt 1.3, Federbeispiel: In der zweiten Gleichung fehlte bei der Dämpfung das m im Nenner.
- Abschnitt 1.2, Seite 8, Berechnung von y_1: Stand 0 statt a
- Abschnitt 3.4, Stabilität: War $\mathrm{eps} = 0.005$, ist jetzt $\mathrm{eps} = 0.05$.
- Abschnitt 3.3, Maschinendarstellung, nach Satz 3.8: War $\mathrm{eps} \sim 10^{-23}$, ist jetzt 10^{-16} (wie in der Einführung ausgerechnet)
- Rechnung vor 3.16: x, y kommen aus dem $\mathbb{C}^n$, und deshalb fehlte bei der Berechnung von (x, y) das konjugiert über dem β_k.
- Nach Satz 4.7 ergänzt: P, L, R aus den Beispielen angegeben.
- Definition 3.16: A in B geändert.
- Vor Definition 3.11: $\forall db \in \mathbb{R}^n$, da fehlte das db.
- Beispiel 5.4 Punkt 1: Es fehlte ein L bei $mL = (1, \ldots, 1)^t(1, \ldots, 1)L = \ldots$
- Beweis zu Satz 5.6: Bei der Berechnung von A^tAx^+ musste ganz am Ende der Zeile A^tb statt nur b stehen.
- Definition 5.1 der Formulierung in der Vorlesung angepasst.
- Satz 5.8: Bei der Definition von A^+ waren n und m vertauscht.
- Beispiel 5.4: * in ^{t} geändert.
- Satz 6.5, Beweis: f in g geändert.
- Definition 6.16, schwache Diagonaldominanz: $\geq$ in $\leq$ geändert.

- Beweis zu Satz 7.2: Beim Einsetzen der Taylorformel war ein $\frac{1}{2}$ und ein Quadrat verlorengegangen.
- Beweis zu Satz 9.6: In der Mitte des Beweises $w^{N+1}(x) = (N+1)!$, nicht p.
- Satz 9.6: War richtig so, aber man schreibt hier besser $\overline{x}$ statt x nach Insbesondere gilt.
- Satz 10.1: Bei Punkt 3 fehlte der Strich in $p'(x)$. Bei Punkt 4 fehlten Klammern.
- Beweis zu Satz 10.3: Hier stand am Anfang $\mathcal{P}_n$ statt $\mathcal{P}_N$.
- Vor Satz 10.4: Stand ein Gleichheitszeichen bei der Einsetzung der Simpson–Regel. Sollte ein $\sim$ sein.
- Satz 10.2: Hier fehlte die Angabe des Grundintervalls $[a, b]$ in allen Teilen.
- Beispiele zu Richardson, Kapitel 10.3: Beispiel der Trapezregel zu rechtsseitigem Differenzenquotienten geändert. Romberg/Trapezregel als neues Beispiel eingeführt.
- Beweis zu 10.2: Für $f(x-h)$ musste $-h^3/6$ stehen.
- Richardson 10.3, Untersuchung der Genauigkeit: Taylorentwicklung für $h/2$ korrigert.
- Definition 12.5, Klassifikation der numerischen Verfahren: Bei allen Verfahren fehlte ein h vor dem φ. Es ist zwar so nicht falsch, passt aber dann nicht mehr zu den weiteren Untersuchungen.

Literaturverzeichnis

R.W. Freund and R.H.W. Hoppe. *Stoer/Bulirsch: Numerische Mathematik 1*. Springer-Lehrbuch. Springer London, Limited, 2007. ISBN 9783540453901. URL http://link.springer.com/book/10.1007/978-3-540-45390-1/page/1.

E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Solving Ordinary Differential Equations. Springer, 1993. ISBN 9783540566700. URL http://books.google.de/books?id=F93u7VcSRyYC.

Carl Runge and Hermann König. *Vorlesungen über numerisches Rechnen*. Springer Göttingen, 1925. URL http://resolver.sub.uni-goettingen.de/purl?PPN373207646.

Abbildungsverzeichnis

Listings