§1 Zahldarstellungen und Rundungsfehler

Bekannt ist die Darstellung einer Zahl im Dezimalsystem zur Basis $d = 10$:

$$272.5 = 2 \cdot 10^2 + 7 \cdot 10^1 + 2 \cdot 10^0 + 5 \cdot 10^{-1},$$

$$\pi = 3.1415\ 92653\ 58979\ldots$$

$$1 = 1.00\ldots = 0.999\ldots \quad \text{(nicht eindeutig)}.$$

Die allgemeine Darstellung einer Zahl zur Basis $d \in \mathbb{N}$ wird wiedergegeben durch den folgenden Satz:

(1.1) **Satz**: Sei $d \in \mathbb{N}$ mit $d \geq 2$ und sei $x \in \mathbb{R}$, $x \neq 0$. Dann gibt es eine Darstellung

$$x = \pm \sum_{i=k}^{-\infty} a_i d^i$$

mit $k \in \mathbb{N}$ und $a_i \in \{0, 1, \ldots, d-1\}$. Diese Darstellung ist eindeutig, wenn zusätzlich gilt: zu jedem $n \in \mathbb{N}$ gibt es einen Index $i \leq -n$ mit $a_i \neq d-1$.

Beweis: Vgl. HÄMMERLIN / HOFFMANN (S. 2,3), WALTER (S. 105).

Schreibweise:

$$x = \pm a_k a_{k-1} \ldots a_0 \cdot a_{-1} a_{-2} \cdots,$$

$$0 \leq a_i \leq d-1, \quad a_i \in \mathbb{N}.$$

Die am häufigsten verwendeten Basen sind $d = 2, 8, 10, 16$ mit den Ziffern in der folgenden Tabelle:

Name des Systems	Basis d	Ziffern
Dual-	2	0, 1 oder 0, L
Oktal-	8	0, 1, 2, ..., 7
Dezimal-	10	0, 1, 2, ..., 8, 9
Hexadezimal-	16	0, 1, 2, ..., 8, 9, A, B, C, D, E, F

Beispiele:

$d = 2$

$18.5 = 1 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1}$

$= L\,0\,0\,L\,0.L$

$3.2 = LL \cdot \overline{00LL}$

$d = 16$:

$29 = 1D$

Problem: Auf der Rechenanlage können nur endlich viele Zahlen dargestellt werden. Die Menge der Maschinenzahlen A hat die Form

$$(1.2)\quad A = \left\{ x = \pm \sum_{k=1}^{t} a_k d^{-k} \times d^e \;\middle|\; 0 \le a_k \le d-1 \right\}.$$

wobei

$m = \sum_{k=1}^{t} a_k d^{-k}$: Mantisse, $|m| < 1$,

$t \in \mathbb{N}$ fest : Mantissenlänge

$e \in \mathbb{Z}' \subset \mathbb{Z}$: Exponent, $\mathbb{Z}'$ endlich

Es gibt mehrere spezielle Datentypen:

Integerzahlen:

$$x = \pm \sum_{k=1}^{t} a_k d^{-k} \times d^t = \pm \sum_{k=1}^{t} a_k d^{t-k}$$

Festkommazahlen:

$$x = \pm \sum_{k=1}^{t} a_k d^{-k} \times d^q, \quad q \in \mathbb{Z} \text{ fest}.$$

Diese Festkommadarstellung ist in vielen Fällen ungeeignet, da etwa physikalische Konstanten über mehrere Dekaden streuen, z.B.

Ruhemasse des Elektrons: $m_0 = 9.11 \cdot 10^{-28} g$

Lichtgeschwindigkeit: $c = 2.998 \cdot 10^{10} cm/sec.$

(Normierte) Gleitkommazahlen:

$$x = \pm \sum_{k=1}^{t} a_k d^{-k} \times d^e, \quad a_1 \neq 0 \text{ für } x \neq 0.$$

Beispiel: $5420 = 0.542 \times 10^4$.

Für den Exponenten e gilt

$$e_- \le e \le e_+ ,$$

sodaß alle darstellbaren Zahlen $x \neq 0$ im Bereich

$$d^{e_-} \leq |x| < d^{e_+}$$

liegen.

Exponentenunterlauf: $|x| < d^{e_-}$, x wird durch Null ersetzt.

Exponentenüberlauf: $|x| > d^{e_+}$.

Beispiel: Bei der IBM 360 ist $d = 16$, $e_- = -64$, $e_+ = 63$.

1.2 Rundung

Eine Abbildung

$$rd : \mathbb{R} \to A \text{ heißt Rundung, wenn}$$

(1.3)
$$|rd(x) - x| = \min_{a \in A} |a - x| .$$

Bei der Beschreibung der Rundung einer Zahl

$$x = \pm d^e \times \sum_{k=1}^{\infty} a_k d^{-k} \neq 0$$

auf t Stellen beschränken wir uns auf normierte Gleitkommazahlen und nehmen an, daß keine Bereichsüberschreitungen auftreten, d.h. $e_- \leq e \leq e_+$ gilt. Die auf t Stellen gerundete Zahl x ist

(1.4)
$$rd(x) = \begin{cases} \pm d^e \sum_{k=1}^{t} a_k d^{-k} & , a_{t+1} < \frac{d}{2} \\ \pm d^e \left(\sum_{k=1}^{t} a_k d^{-k} + d^{-t} \right) & , a_{t+1} \geq \frac{d}{2} \end{cases} .$$

Beispiele: $d = 10$, $t = 4$

$\pi = 3.14159..$, $rd(\pi) = 0.3142 \times 10^1$

$\sqrt{57} = 7.5498$, $rd(\sqrt{57}) = 0.7550 \times 10^1$

$x = 0.1253499..$, $rd(x) = 0.1253 \times 10^0$

$d = 2$, $t = 3$

$.x = L.00LL$, $rd(x) = 0.L0L \times 2^1$.

Die Zahl

$$eps = \frac{1}{2} d^{-t+1}$$

heißt die Maschinengenauigkeit einer normierten Gleitkommamaschine mit

Mantissenlänge t zur Basis d.

(1.5) Definition: Sei $x \in \mathbb{R}$ und sei $\tilde{x} \in \mathbb{R}$ eine Näherung (Approximation) für x.

(i) $|\tilde{x} - x|$ heißt der absolute Fehler von $\tilde{x}$.

(ii) Für $x \neq 0$ heißt $\left|\frac{\tilde{x}-x}{x}\right|$ der relative Fehler von $\tilde{x}$.

(1.6) Satz: Sei rd eine Rundung mit $t \geq 1$. Dann gilt für $x \neq 0$:

(i) $\left|\frac{rd(x)-x}{x}\right| \leq eps$

(ii) $\left|\frac{rd(x)-x}{rd(x)}\right| \leq eps$.

Beweis: zu (i): Sei

$$x = \pm 0.a_1 .. a_t a_{t+1} .. \times d^e, \quad a_1 \neq 0 .$$

Die Rundungsvorschrift (1.4) zeigt

$$|rd(x) - x| \leq \frac{d}{2} d^{-(t+1)} d^e = \frac{1}{2} d^{-t} d^e ,$$

$$|x| \geq d^{-1} d^e \quad \text{wegen } a_1 \neq 0$$

$$\Rightarrow \left|\frac{rd(x)-x}{x}\right| \leq \frac{1}{2} d^{-t} d^e \;/\; d^{-1} d^e$$

$$= \frac{1}{2} d^{-t+1} = eps .$$

Die Beh. (ii) folgt wie in (i). ■

Aus dem Satz folgt die Darstellung

(1.7) $$\boxed{rd(x) = x(1+\varepsilon),\ |\varepsilon| \leq eps = \frac{1}{2} d^{-t+1}}$$

1.3 Gleitkommaarithmetik

Das Symbol $\square$ bezeichne eine der Rechenoperationen $+, -, \cdot, /$.
Wenn $x, y \in A$ zwei Gleitkommazahlen mit t-stelliger Mantisse sind, dann ist im allgemeinen $x \square y$ nicht mit t-stelliger Mantisse darstellbar, d.h. die Operation $\square$ ist nicht abgeschlossen in A.

Beispiel: $d = 10,\ t = 2$

$$0.11 + 0.0011 = 0.1111 \notin A$$
$$1.1 * 1.1 = 1.21 \notin A$$
$$1 / 0.9 = 1.\overline{11} \notin A$$

Nach einer Operation $\square$ muß also im allgemeinen gerundet werden. Die Operation $\square$ wird auf der Maschine folgendermaßen ausgeführt:

(a) Möglichst genaue Berechnung von $z = x \square y$ (etwa auf $2t$ Ziffern)

(b) Normalisierung von z und Rundung auf t Stellen.

Das Ergebnis ist die Gleitpunktoperation

$$gl(x \square y).$$

Die Arithmetik auf dem Rechner kann so organisiert werden, daß gilt

$$gl(x \square y) = rd(x \square y) \text{ für } x, y \in A.$$

Aus (1.7) ersieht man die Darstellung

$$(1.8) \quad \boxed{gl(x \square y) = (x \square y)(1+\varepsilon), \quad |\varepsilon| \le eps}$$

Die Gleitpunktoperationen sind weder assoziativ noch distributiv.

Beispiel: $d = 10,\ t = 2$

$$0.75 + 0.055 - 0.80 = 0.005$$
$$gl(0.75 + 0.055) = rd(0.805) = 0.81$$
$$gl(0.81 - 0.80) = rd(0.01) = 0.01$$

Andererseits ist

$$gl(0.75 + gl(0.055 - 0.80))$$
$$= gl(0.75 + rd(-0.745))$$
$$= gl(0.75 - 0.75) = 0.$$

§2 Fehleranalyse

2.1 Fehlertypen

Gegeben sei $D \subset \mathbb{R}^n$ und $f: D \to \mathbb{R}^m$.
Das Problem bestehe in der Berechnung des Ausdrucks

$$y = f(x), \qquad x \in D, \tag{2.1}$$

d.h.

$$y_i = f_i(x_1, .., x_n); \; i = 1, .., m,$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \to \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = y$$

Eingabedaten — Resultatdaten

Man unterscheidet die folgenden Fehlertypen, welche die Genauigkeit der Berechnung von $y = f(x)$ begrenzen:

(1) Fehler in den Eingabedaten x,
(2) Abbrechfehler oder Diskretisierungsfehler
(3) Rundungsfehler während der Rechnungen.

Beispiel: Mit einer Mantissenlänge $t = 3$ sei die unendliche Reihe

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

zu berechnen für $x = 1.2345$.

(1) Führe die Rechnungen mit der Approximation $\tilde{x} = rd(x) = 1.23$ durch.

(2) Bei der Approximation von e^x durch eine endliche Summe

$$S_N = \sum_{k=0}^{N} \frac{x^k}{k!}$$

entsteht der Abbrechfehler

$$e^x - S_N = \sum_{k=N+1}^{\infty} \frac{x^k}{k!}.$$

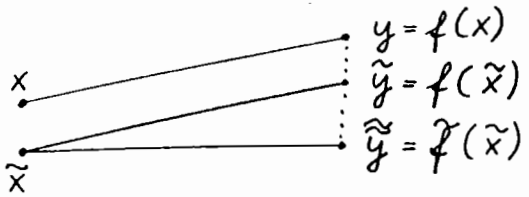
(3) Berechne die gerundeten Ausdrücke

$$gl(\tilde{x}^k/k!), \quad k = 0, .., N.$$

Anstatt der exakten Lösung $y = f(x)$ wird man daher eine Approximation

$\tilde{f}(\tilde{x})$, $\tilde{x}$ Approximation für x, $\tilde{f}$ " für f,

berechnen:



2.2 Kondition und Gutartigkeit

Wir befassen uns zunächst mit dem Fehlertyp (1) und studieren, wie sich Fehler in x auf das Ergebnis $y = f(x)$ auswirken. Sei $\tilde{x}$ eine Näherung von x und sei

$\Delta x = \tilde{x} - x$ der absolute Fehler,

$\frac{\tilde{x}_i - x_i}{x_i}$ der relative Fehler $(i = 1, .., n)$

$\tilde{y} = f(\tilde{x})$ der Näherungswert für $y = f(x)$.

Die Taylor-Entwicklung erster Ordnung liefert für den absoluten Fehler $\Delta y = \tilde{y} - y$ die Approximation (f sei C^1-Funktion)

$$(2.2)\quad \Delta y_i = f_i(x + \Delta x) - f_i(x) \approx \sum_{j=1}^{n} \frac{\partial f_i(x)}{\partial x_j} \Delta x_j \qquad (i = 1, .., m).$$

Für den relativen Fehler erhält man dann

$$(2.3)\quad \frac{\Delta y_i}{y_i} \approx \sum_{j=1}^{n} \left(\frac{\partial f_i(x)}{\partial x_j} \cdot \frac{x_j}{f_i(x)} \right) \cdot \left(\frac{\Delta x_j}{x_j} \right), \quad x_j, y_i \neq 0.$$

(2.4) Definition:

(a) Die Zahlen

$$k_{ij}(x) = \left| \frac{\partial f_i(x)}{\partial x_j} \cdot \frac{x_j}{f_i(x)} \right|$$

heißen Verstärkungsfaktoren bzw. (relative) Konditionszahlen.

(b) Das Problem "Berechne $y = f(x)$" heißt gut konditioniert, falls alle $k_{ij}(x)$ die Größenordnung 1 haben. Andernfalls heißt das Problem schlecht konditioniert.

Zuerst untersuchen wir damit die arithmetischen Operationen $+, -, *, /$.

(1) Multiplikation: $y = f(x_1, x_2) = x_1 \cdot x_2$

$k_{11}(x) = k_{12}(x) \equiv 1$: gutartig

(2) Division: $y = f(x_1, x_2) = x_1/x_2$

$k_{11}(x) = k_{12}(x) \equiv 1$: gutartig

(3) Addition, Subtraktion:

$y = f(x_1, x_2) = x_1 + x_2$,

$k_{11}(x) = \left| \frac{x_1}{x_1+x_2} \right|$, $k_{12}(x) = \left| \frac{x_2}{x_1+x_2} \right|$.

Das Problem ist schlecht konditioniert, falls $x_1 \sim -x_2$. Daher ist die Subtraktion nahezu gleichgroßer Zahlen mit gleichen Vorzeichen schlecht konditioniert.
Dieses Phänomen heißt Auslöschung.

Beispiel:

$1.31 - 1.25 = 0.06$
$1.32 - 1.24 = 0.08$

rel. Fehler: 0.8%, 30%

Genauer gilt:

$x = (1.31, -1.25)$, $y = x_1 + x_2 = 0.06$

$\Delta x = (0.01, 0.01)$,

$\left| \frac{\Delta x_i}{x_i} \right| \le 0.008$, $k_{1i}(x) \le 22$ $(i = 1, 2)$.

Der relative Fehler im Ergebnis ist ca. 40 mal größer als der relative Fehler in den Daten.

(4) Wurzel: $y = f(x) = \sqrt{x}$, $x > 0$,

$k(x) = \frac{1}{2} \frac{1}{\sqrt{x}} \frac{x}{\sqrt{x}} = \frac{1}{2}$: gutartig.

Bem.: Bei einigen Problemen kann die Auslöschung durch geeignete Umformulierung vermieden werden; vgl. Beispiele (2.5), (2.6).

(2.5) Beispiel:

Es sei eine quadratische Gleichung in der Form

$y^2 + 2py - q = 0$, $p, q > 0$, $p \gg q$

gegeben, und die Aufgabe bestehe darin, die größte Wurzel

$$y = f(p, q) = -p + \sqrt{p^2 + q} = \frac{q}{p + \sqrt{p^2 + q}}$$

zu berechnen. Die Konditionszahlen in (2.4) (a) sind

$$k_p(p,q) = \left| \frac{p}{f(p,q)} \frac{\partial f}{\partial p} \right| = \frac{p}{\sqrt{p^2+q}} < 1$$

$$k_q(p,q) = \left| \frac{q}{f(p,q)} \frac{\partial f}{\partial q} \right| = \frac{p+\sqrt{p^2+q}}{2\sqrt{p^2+q}} < 1 \,,$$

also ist die Aufgabe gut konditioniert.

2.3 Algorithmen

Ein Algorithmus zur Berechnung der Lösung $y = f(x)$ eines Problems ist eine Sequenz von endlich vielen "elementaren Operationen" (z.B. $+, -, *, /, \sqrt{x}, \cos(x)..$). Es gibt i.a. mehrere Anordnungen der Rechenschritte, welche zum gleichen Ergebnis $y = f(x)$ führen. In jedem Rechenschritt fallen Rundungsfehler an. Dabei kann der Fall eintreten, daß bei der Lösung eines an sich gut konditionierten Problems eine ungünstige Anordnung der Rechenschritte zum Aufschaukeln der Rundungsfehler führt. Der zugehörige Algorithmus ist numerisch instabil.

(2.6) Beispiel: Fortsetzung Beispiel (2.5).

Die Berechnung der größten Wurzel der quadratischen Gleichung

$$y^2 + 2py - q = 0, \quad p, q > 0, \quad p \gg q$$

kann mit zwei Methoden erfolgen:

Methode 1: $y = -p + \sqrt{p^2+q}$

(Auslöschung wegen $p \gg q > 0$)

Methode 2: $y = \frac{q}{p+\sqrt{p^2+q}}$

Die Berechnung beider Ausdrücke erfordert zunächst die Größen

$$s = p^2 \,,$$
$$t = s + q \,,$$
$$u = \sqrt{t} \,.$$

Im weiteren unterscheidet man die beiden Algorithmen:

Algorithmus 1: $y = f_1(u) = -p + u$

Die hierbei auftretende Auslöschung findet sich in der Konditionszahl wieder:

$$k_u = \left| \frac{u}{f_1(u)} \frac{df_1(u)}{du} \right| = \frac{u}{-p+u}$$

$$= \frac{\sqrt{p^2+q}}{-p+\sqrt{p^2+q}} = \frac{1}{q}\left(p\sqrt{p^2+q} + p^2+q\right)$$

$$\geq \frac{2p^2}{q} \gg 1$$

Algorithmus 2:

$$v = p+u, \quad y = f_2(u) = \frac{q}{v} = \frac{q}{p+u}$$

Die Konditionszahl

$$k_u = \left| \frac{u}{f_2(u)} \frac{df_2(u)}{du} \right| = \frac{\sqrt{p^2+q}}{p+\sqrt{p^2+q}} < 1$$

zeigt, daß dieser Algorithmus numerisch stabil ist.

Zahlenbeispiel: (Mantisse $t = 12$)

$p = 1000$, $q = 0.018\,000\,000\,081$

exakt : $0.900\,000\,000\,000 \times 10^{-5}$

Algorithmus 1 : $0.900\,030\,136\,108 \times 10^{-5}$

Algorithmus 2 : $0.899\,999\,999\,99\underline{9} \times 10^{-5}$

Der Verstärkungsfaktor bei Algorithmus 1

ist

$$\frac{2p^2}{q^2} \approx 10^8,$$

d.h. man kann 8 falsche Stellen bei Algorithmus 1 erwarten.

Auch das nachfolgende Beispiel zeigt, daß man durch ungeschicktes Anordnen der einzelnen Rechenschritte zu völlig unbrauchbaren Resultaten geführt werden kann.

(2.7) Beispiel: Es soll das Integral

$$I_n = \int_0^1 \frac{x^n}{x+5}\, dx$$

für $n = 0, 1, .., 20$ berechnet werden. Für die Zahlen I_n kann sofort eine Rekursion angegeben werden:

$$I_n + 5\, I_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} = \int_0^1 x^{n-1}\, dx = \frac{1}{n}.$$

Ausgehend von dem Wert

$I_0 = \ln \frac{6}{5} = 0.1823215..$ können

theoretisch alle Werte

$$I_n = \frac{1}{n} - 5\, I_{n-1}$$

berechnet werden. Die Rechnung liefert jedoch schon für $n \geq 15$ unbrauchbare Ergebnisse.

schon I_{13} muß verkehrt sein, denn $I_{n+1} < I_n$!

vielleicht: I_{14} offensichtlich verkehrt, denn $I_n > 0$!

n	$I_n = -5I_{n-1} + \frac{1}{n}$ $I_0 = \ln\frac{6}{5}$	$I_{n-1} = \frac{1}{5}(-I_n + \frac{1}{n})$ $I_{50} = 0.$
1	0.088 392 216	0.088 392 216
2	0.058 038 919	0.058 038 919
3	0.043 138 734	0.043 138 734
4	0.034 306 327	0.034 306 329
5	0.028 468 364	0.028 468 352
6	0.024 324 844	0.024 324 905
7	0.021 232 922	0.021 232 615
8	0.018 835 389	0.018 836 924
9	0.016 934 162	0.016 926 489
10	0.015 329 188	0.015 367 550
11	0.014 263 149	0.014 071 338
12	0.012 017 583	0.012 976 639
13	0.016 835 157	0.012 039 876
14	-0.012 747 213	0.011 229 186
15	0.130 402 734	0.010 520 733
16	-0.589 513 672	$9.896\,332\,328 \cdot 10^{-3}$
17	3.006 391 892	$9.341\,867\,769 \cdot 10^{-3}$
18	$-1.497\,640\,391 \cdot 10^{1}$	$8.846\,216\,741 \cdot 10^{-3}$
19	$7.493\,465\,113 \cdot 10^{1}$	$8.400\,495\,394 \cdot 10^{-3}$
20	$-3.746\,232\,556 \cdot 10^{1}$	$7.997\,523\,028 \cdot 10^{-3}$

Betrachtet man nämlich die Akkumulation der Rundungsfehler, so wird in jedem Schritt der Rundungsfehler mit dem Faktor 5 multipliziert. Der Rundungsfehler in I_0 führt dann zu einem Fehler

$$|\varepsilon_n| \leq 5^n \cdot \frac{1}{2} \cdot 10^{-t+1}$$

$$\approx 5 \cdot 10^5 \quad \text{für } n = 20,\ t = 9.$$

Wird hingegen die Rekursion in der Form

$$I_{n-1} = \frac{1}{5n} - \frac{1}{5} I_n \qquad \text{(rückwärts)}$$

ausgewertet, so reduziert sich der Fehler bei der Berechnung von I_{n-1} gegenüber dem Fehler in I_n um den Faktor $\frac{1}{5}$. Man überlegt leicht, daß $I_n \to 0$ für $n \to \infty$. Beginnend mit dem Näherungswert $I_{50} = 0$ (exakt: $I_{50} = 0.00327\,851\,462$) erweist sich die Berechnung der Zahlen $I_{20}, I_{19}, \ldots, I_1, I_0$ als äußerst stabil. Die Ergebnisse sind auf 10 Stellen genau.

Kap. II: Lineare Gleichungssysteme

Sei A eine (m,n)-Matrix und sei $b \in \mathbb{R}^m$. Gesucht ist ein Vektor $x \in \mathbb{R}^n$, welcher das lineare Gleichungssystem (LGS)

$$Ax = b$$

löst. In diesem Kapitel werden direkte Methoden zur Lösung von $Ax = b$ vorgestellt, welche eine Lösung x in endlich vielen Schritten berechnen.

Es sind folgende drei Fälle möglich:

(1) $m = n$: $\operatorname{rang}(A) = n \Rightarrow Ax = b$ ist eindeutig lösbar; vgl. § 3.

(2) $m > n$: Das LGS $Ax = b$ heißt überbestimmt und hat im allgemeinen keine Lösung. Stattdessen wird ein Ersatzproblem gelöst: vgl. Lineare Ausgleichsprobleme (§ 20).

(3) $m < n$: Das LGS $Ax = b$ heißt unterbestimmt. Wenn eine Lösung existiert, dann hat der Lösungsraum die Dimension $n - \operatorname{rang}(A)$.
Anwendungen: Lineare Optimierung (§ 7).

§3 Die LR-Zerlegung einer Matrix und Gauß-Elimination

Sei $A = (a_{ik})$ eine (n,n)-Matrix und $b \in \mathbb{R}^n$.
Zu lösen sei das LGS

$$(3.1) \qquad Ax = b.$$

Motivationsbeispiel: (Gauß-Elimination mit $n = 3$)

$$\begin{aligned} 2x_1 + 4x_2 + 6x_3 &= 2 \\ x_1 + x_2 &= 1 \quad \Big| \; -\tfrac{1}{2} \times 1.\text{Zeile} \\ x_1 + x_3 &= 2 \quad \Big| \; -\tfrac{1}{2} \times 1.\text{Zeile} \end{aligned}$$

Nach dem 1. Schritt:

$$\begin{aligned} 2x_1 + 4x_2 + 6x_3 &= 2 \\ -x_2 - 3x_3 &= 0 \\ -2x_2 - 2x_3 &= 1 \quad \Big| \; -2 \times 2.\text{Zeile} \end{aligned}$$

Nach dem 2. Schritt

$$\begin{aligned} 2x_1 + 4x_2 + 6x_3 &= 2 \\ -x_2 - 3x_3 &= 0 \\ 4x_3 &= 1 \end{aligned}$$

d.h.

$$\begin{pmatrix} 2 & 4 & 6 \\ 0 & -1 & -3 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} \Rightarrow \quad \begin{aligned} x_3 &= \tfrac{1}{4} \\ x_2 &= -3x_3 = -\tfrac{3}{4} \\ x_1 &= \tfrac{1}{2}(2 - 4x_2 - 6x_3) = \tfrac{7}{4} \end{aligned}$$

$$R \qquad x \qquad = \qquad c$$

Idee des Gauß-Elimination: Führe das LGS $Ax = b$ durch geeignete Linearkombinationen von Gleichungen in ein gestaffeltes Gleichungssystem über:

$$Rx = c \in \mathbb{R}^n,$$

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ 0 & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}$$ rechte Dreiecks-Matrix

Für $r_{ii} \neq 0$ wird $Rx = c$ rekursiv gelöst durch

(3.2)
$$\boxed{x_i = \frac{1}{r_{ii}} \left(c_i - \sum_{k=i+1}^{n} r_{ik} x_k \right), \quad i = n, n-1, \dots, 1.}$$

Zur Durchführung der Gauß-Elimination werden Elementarmatrizen L_j benötigt:

$$L_j = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ 0 & & -l_{j+1,j} & \ddots \\ & & \vdots & \\ & & -l_{nj} & & 1 \end{pmatrix} \leftarrow j\text{-te Zeile}$$

$\uparrow$
j-te Spalte

Rechenregeln für Elementarmatrizen: Sei

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \quad \text{mit } a_i = (a_{i1}, \dots, a_{in}).$$

(3.3)
$$L_j A = \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ a_{j+1} - \ell_{j+1,j}\, a_j \\ \vdots \\ a_n - \ell_{nj}\, a_j \end{pmatrix},$$

(3.4)
$$L_j^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{j+1,j} & \ddots & \\ & & \vdots & & 1 \\ & & \ell_{nj} & & \end{pmatrix}$$

(3.5)
$$L_j^{-1} \cdot L_k^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{j+1,j} & \ddots & \\ & & \vdots & & 1 \\ & & \vdots & & \ell_{k+1,k} & \ddots \\ & & \ell_{nj} & & \ell_{nk} & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow j \\ \\ \leftarrow k \\ \\ \end{matrix}$$

$\uparrow$ j $\qquad$ $\uparrow$ k

LR-Zerlegung und Gauß-Elimination ohne Pivotsuche: Sei

$$A := \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ \vdots & & \vdots \\ a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

1. Schritt: Sei $a_{11}^{(1)} \neq 0$

$$L_1 A = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} \cdots & a_{nn}^{(2)} \end{pmatrix} \qquad (\text{vgl. } (3.3) \text{ mit } j=1)$$

mit

$$l_{i1} := \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, \ldots, n,$$

$$a_{ik}^{(2)} = a_{ik}^{(1)} - l_{i1} \cdot a_{1k}^{(1)}, \quad i,k = 2, \ldots, n.$$

(In Worten: subtrahiere von der i-ten Zeile der Matrix A das l_{i1}-fache der 1. Zeile, $i = 2, \ldots, n$.)

Ausgangsmatrix vor dem j-ten Schritt ($j \geq 2$):

$$L_{j-1} \cdots L_1 A = \begin{pmatrix} a_{11}^{(1)} & & & a_{1n}^{(1)} \\ & \ddots & & \vdots \\ & & a_{jj}^{(j)} \cdots a_{jn}^{(j)} \\ 0 & & \vdots \quad\quad \vdots \\ & & a_{nj}^{(j)} \cdots a_{nn}^{(j)} \end{pmatrix}.$$

<u>j-ter Schritt ($j \geq 2$)</u>: Sei $a_{jj}^{(j)} \neq 0$

$$L_j L_{j-1} \cdots L_1 A = \begin{pmatrix} a_{11}^{(1)} & & & & a_{1n}^{(1)} \\ & \ddots & & & \vdots \\ & & a_{jj}^{(j)} & \cdots & a_{jn}^{(j)} \\ & & 0 & a_{j+1,j+1}^{(j+1)} \cdots & a_{j+1,n}^{(j+1)} \\ & & \vdots & \vdots \quad\quad \vdots & \\ & & 0 & a_{n,j+1}^{(j+1)} \cdots & a_{nn}^{(j+1)} \end{pmatrix}$$

mit

$$l_{ij} = a_{ij}^{(j)} / a_{jj}^{(j)}, \quad i = j+1, \ldots, n$$

$$a_{ik}^{(j+1)} = a_{ik}^{(j)} - l_{ij} a_{jk}^{(j)}, \quad i,k = j+1, \ldots, n.$$

Nach $n-1$ Schritten erhält man

$$(3.6)\quad L_{n-1} \cdots L_1 A = \begin{pmatrix} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} \\ & \ddots & \vdots \\ 0 & & a_{nn}^{(n)} \end{pmatrix} =: R = (r_{ik})$$

mit $r_{ii} = a_{ii}^{(i)} \neq 0$.

Wendet man die Matrizen L_j direkt auf die erweiterte Matrix (A, b) an, so ergibt sich

$$L_{n-1} \cdots L_1 (A, b) = (R, c).$$

Das LGS $Ax = b$ ist dann äquivalent zu $Rx = c$ und kann gemäß (3.2) gelöst werden (Gauß-Elimination).

Aus (3.6) folgt mittels der Formeln (3.4)(3.5) die LR-Zerlegung der Matrix A:

$$(3.7)\quad A = L_1^{-1} \cdots L_{n-1}^{-1} R =: LR$$

$$L = \begin{pmatrix} 1 & & 0 \\ l_{21} & \ddots & \\ \vdots & \ddots & \ddots \\ l_{n1} & \cdots & l_{n,n-1} \quad 1 \end{pmatrix}$$ linke Dreiecksmatrix

Bei gegebener LR-Zerlegung $A = LR$ ist das LGS $Ax = b$ äquivalent zu den beiden leicht auflösbaren LGS

$$Lc = b, \quad Rx = c.$$

Insbesondere folgt noch aus (3.7)

$$\det(A) = \det(L)\det(R) = \prod_{j=1}^{n} r_{jj}.$$

Problem: Wann gilt $a_{jj}^{(j)} \neq 0$?

(3.8) Satz: Sei A eine (n,n)-Matrix, deren Hauptabschnittsmatrizen A_j regulär sind. Dann gibt es eine eindeutige Zerlegung

$A = LR$,
L linke Dreiecksmatrix mit $l_{jj} = 1$, $j = 1, \ldots, n$
R reguläre rechte Dreiecksmatrix.

Beweis: Vgl. Satz (4.1).

Zur Behandlung des Falles $a_{jj}^{(j)} = 0$ für ein j benötigen wir Permutationsmatrizen. Hierzu sei

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow i$$

der i-te kanonische Einheitsvektor

Eine Matrix P heißt <u>Permutationsmatrix</u>, wenn eine Permutation $(i_1, \dots i_n)$ von $(1, \dots, n)$ existiert mit

$$P = \begin{pmatrix} e_{i_1}^T \\ \vdots \\ e_{i_n}^T \end{pmatrix}.$$

Insbesondere gilt $P^2 = I$, also $P^{-1} = P$.

<u>LR-Zerlegung und Gauß-Elimination mit Pivotsuche:</u>

<u>j-ter Schritt ($j \geq 1$)</u>: Die Ausgangsmatrix sei

$$A^{(j)} := \begin{pmatrix} a_{11}^{(1)} & & & a_{1n}^{(1)} \\ & \ddots & & \vdots \\ & & a_{jj}^{(j)} \dots a_{jn}^{(j)} & \vdots \\ 0 & & \vdots & \vdots \\ & & a_{nj}^{(j)} \dots a_{nn}^{(j)} & \end{pmatrix}, \qquad A^{(1)} := A \tag{3.9}$$

<u>Spaltenpivot-Suche</u>: Wähle r mit

$$\left| a_{rj}^{(j)} \right| = \max_{i \geq j} \left| a_{ij}^{(j)} \right|.$$

<u>1. Fall</u>: $a_{rj}^{(j)} = 0$: A ist singulär (Beweis!), setze $L_j = I$.

2. Fall: $a_{rj}^{(j)} \neq 0$: Vertausche die j-te Zeile mit der r-ten Zeile in $A^{(j)}$. Dies entspricht einer Multiplikation von links mit der Permutationsmatrix

$$P_j = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & 0 & & 1 & \\ & & & & \ddots & & \\ & & & 1 & & 0 & \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow j \\ \\ \leftarrow r \\ \\ \\ \\ \end{matrix} .$$

Damit erhalten wir im ungünstigen Fall nach $n-1$ Schritten

$$(3.10) \quad L_{n-1} P_{n-1} \dots L_2 P_2 L_1 P_1 A = R .$$

Dies führt zu einer LR-Zerlegung mittels des folgenden Hilfssatzes (Beweis als Übung).

(3.11) Hilfssatz: Sei $j < k$. Die Permutationsmatrix P_k vertausche die Zeilen k und $r \geq k$. Dann gilt

$$P_k L_j = L_j' P_k ,$$

wobei L_j' aus L_j dadurch hervorgeht, daß man in der j-ten Spalte das k-te und r-te Element vertauscht.

Beweis:

$$L_j = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\ell_{kj} & 1 & \\ & & \vdots & & \ddots \\ & & -\ell_{rj} & 0 & 1 \\ & & \vdots & & \end{pmatrix} \begin{matrix} \\ \\ \leftarrow j \\ \leftarrow k \\ \\ \leftarrow r \\ \\ \end{matrix}$$

$$P_k L_j = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \vdots & & \ddots \\ & & -\ell_{rj} & 0 \cdots & 1 \\ & & \vdots & \vdots & \vdots \\ & & -\ell_{kj} & 1 \cdots & 0 \\ & & \vdots & & \ddots \end{pmatrix} \begin{matrix} \\ \\ \leftarrow j \\ \\ \leftarrow k \\ \\ \leftarrow r \\ \\ \end{matrix}$$

$\uparrow k \qquad \uparrow r$

$$\Rightarrow P_k L_j P_k = L_j' , \qquad P_k^2 = I$$

$$\Rightarrow P_k L_j = L_j' P_k .$$

∎

Die Anwendung des Hilfssatzes auf (3.10) zeigen wir der Einfachheit halber für $n = 4$:

$$L_3 P_3 L_2 P_2 L_1 P_1 A = R$$

$$\Leftrightarrow \quad L_3 P_3 L_2 L_1' P_2 P_1 A = R$$

$$\Leftrightarrow \quad L_3 L_2' L_1'' \underbrace{P_3 P_2 P_1}_{=:P} A = R$$

P Permutationsmatrix

$$\Leftrightarrow \quad PA = LR, \quad L := L_1''^{-1} L_2'^{-1} L_3^{-1}.$$

Die Anwendung der obigen Operationen auf die erweiterte Matrix (A, b) führt auf die Matrix (R, c). R ist regulär, wenn A regulär ist, und das LGS $Rx = c$ kann gemäß (3.2) gelöst werden.

Zusammenfassend erhalten wir

(3.12) Satz: (LR-Zerlegung und Gauß-Elimination)
Zu jeder (n,n)-Matrix A gibt es eine Permutationsmatrix P, eine linke Dreiecksmatrix L und eine rechte Dreiecksmatrix R, so daß

$$PA = LR, \quad l_{jj} = 1 \text{ für } j = 1, \dots n.$$

Ist A regulär, so ist auch R regulär, und die Gauß-Elimination liefert die eindeutige Lösung von $Ax = b$.

Bei der praktischen Durchführung der Gauß-Elimination kann man die wesentlichen Elemente von L, d.h. l_{ik}, $i \geq k+1$, $k \leq j-1$, auf den Null-Elementen der Matrix $A^{(j)}$ in (3.9) abspeichern.

Beispiel:

$$\begin{pmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 7 \\ 4 \end{pmatrix}$$

Die Pivot-Elemente in den erweiterten Matrizen werden durch Kreise markiert.

1. Schritt:

$$\left(\begin{array}{ccc|c} \textcircled{3} & 1 & 6 & 2 \\ 2 & 1 & 3 & 7 \\ 1 & 1 & 1 & 4 \end{array}\right)$$

Anwendung von L_1:

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ \hline 2/3 & 1/3 & -1 & 17/3 \\ 1/3 & \textcircled{2/3} & -1 & 10/3 \end{array}\right)$$

$\uparrow$
l_{i1}

2. Schritt:

Vertausche Zeile 2 und 3

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & \textcircled{2/3} & -1 & 10/3 \\ 2/3 & 1/3 & -1 & 17/3 \end{array}\right)$$

Anwendung von L_2:

$$\left(\begin{array}{ccc|c} 3 & 1 & 6 & 2 \\ 1/3 & 2/3 & -1 & 10/3 \\ 2/3 & 1/2 & -1/2 & 4 \end{array}\right)$$

$$\Rightarrow \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{pmatrix}$$

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad PA = \begin{pmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix}$$

$$PA = LR$$

Gestaffeltes Gleichungssystem:

$$\begin{pmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 10/3 \\ 4 \end{pmatrix}, \qquad \begin{aligned} x_1 &= 19 \\ x_2 &= -7 \\ x_3 &= -8 \end{aligned}$$

$$R \qquad\qquad x \quad = \quad c$$

Gauß-Elimination zur Lösung von

$$Ax = b$$

(1) $PA = LR$

$p = (p_1, \dots, p_n)$ Permutationsvektor

(2) $Lc = Pb$

Vorwärtseinsetzen: $i = 1, \dots, n$:

$$c_i = b_{p_i} - \sum_{k=1}^{i-1} l_{ik} c_k$$

(3) $Rx = c$

Rückwärtseinsetzen: $i = n, n-1, \dots, 1$:

$$x_i = \frac{1}{r_{ii}} \left(c_i - \sum_{k=i+1}^{n} r_{ik} x_k \right)$$

Anzahl der Operationen:

(1) $PA = LR$

$$\sum_{j=1}^{n-1} \left[(n-j) + (n-j)^2 \right]$$

$$= \frac{1}{2} n(n-1) + \frac{1}{6} n(n-1)(2n-1)$$

$$= \frac{1}{3} (n^3 - n)$$

(2) $Lc = Pb$

$$1 + 2 + \dots + (n-1) = \frac{1}{2} (n^2 - n)$$

(3) $Rx = c$

$$1 + 2 + \dots + n = \frac{1}{2} (n^2 + n)$$

Gesamt: $\frac{1}{3} n^3 + n^2 - \frac{1}{3} n$

Programm: $PA = LR$

für $j = 1, .., n$:

$p_j = j$

für $j = 1, .., n-1$:

Pivotsuche:

$max = |a_{jj}|$, $r = j$

für $i = j+1, .., n$:

falls $|a_{ij}| > max$:

$max = |a_{ij}|$, $r = i$

falls $max = 0$: STOP A singulär

Zeilentausch:

falls $r > j$:

für $k = 1, .., n$:

$hr = a_{jk}$, $a_{jk} = a_{rk}$, $a_{rk} = hr$

$hi = p_j$, $p_j = p_r$, $p_r = hi$

Transformation:

für $i = j+1, .., n$:

$a_{ij} = a_{ij} / a_{jj}$

für $k = j+1, .., n$:

$a_{ik} = a_{ik} - a_{ij} a_{jk}$

§4 Matrizen mit speziellen Eigenschaften

4.1 Diagonalstrategie

Zunächst geben wir Bedingungen an, die die Durchführung der Gauß-Elimination ohne Pivotsuche ermöglichen (Diagonalstrategie).

(4.1) Satz: Sei A eine (n,n)-Matrix, deren Hauptabschnittsmatrizen A_j regulär sind. Dann gibt es eine eindeutige Zerlegung

$$A = LR$$

L: linke Dreiecksmatrix mit $l_{jj} = 1$, $j = 1,..,n$,

R: reguläre rechte Dreiecksmatrix.

Beweis: Der Beweis wird durch Induktion über n geführt.

Für $n=1$ ist die Beh. trivial. Die Beh. sei richtig für $n-1$. Für eine (n,n)-Matrix ist die folgende Zerlegung zu zeigen:

$$A = \left(\begin{array}{c|c} A_{n-1} & c \\ \hline a^T & a_{nn} \end{array}\right) = \left(\begin{array}{c|c} L_{n-1} & 0 \\ \hline l^T & 1 \end{array}\right)\left(\begin{array}{c|c} R_{n-1} & r \\ \hline 0 & r_{nn} \end{array}\right).$$

Nach Induktionsvor. gibt es eine Zerlegung

$$A_{n-1} = L_{n-1} R_{n-1}.$$

Für die gesuchten $l, r \in \mathbb{R}^{n-1}$, $r_{nn} \in \mathbb{R}$ erhält man die Gleichungen

(1) $c = L_{n-1} r$

(2) $l^T R_{n-1} = a^T \Rightarrow R_{n-1}^T l = a$

(3) $l^T r + r_{nn} = a_{nn}$.

Die Gleichungen sind eindeutig auflösbar, da nach Vor. L_{n-1}, R_{n-1} regulär sind. ∎

Mit

$$D = \operatorname{diag}(r_{jj}) = \operatorname{diag}(a_{jj}^{(j)})$$

erhält man somit die Zerlegung

$$\boxed{A = LDR}, \quad l_{jj} = 1, \quad r_{jj} = 1.$$

Die Regularität der Hauptabschnittsmatrizen von A kann mit einer einfachen Bedingung für die Elemente a_{ij} von A nachgeprüft werden.

(4.2) <u>Definition:</u> Die Matrix A heißt <u>diagonaldominant</u>, wenn

$$|a_{ii}| > \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}| \quad (i = 1, \ldots, n).$$

(4.3) <u>Satz:</u> Bei einer diagonaldominanten Matrix A sind alle Hauptschnittsmatrizen regulär, also existiert die LR-Zerlegung $A = LR$.

<u>Beweis:</u> Für die j-te Abschnittsmatrix A_j gelte

$$A_j x = 0 \qquad \text{für ein } x \in \mathbb{R}^j.$$

Zu zeigen ist dann $x = 0$. Wäre

$$|x_\tau| = \max_{1 \le i \le j} |x_i| > 0,$$

so betrachten wir die τ-te Gleichung

$$\sum_{i=1}^{j} a_{\tau i} x_i = 0.$$

Zusammen mit $|a_{\tau\tau}| > \sum_{\substack{k=1 \\ k \neq \tau}}^{j} |a_{\tau k}|$ ergibt sich hieraus ein Widerspruch:

$$|a_{rr}||x_r| = \Big|\sum_{\substack{k=1\\k\neq r}}^{j} a_{rk}x_k\Big|$$

$$\leq \sum_{k\neq r} |a_{rk}||x_k|$$

$$\leq \sum_{k\neq r} |a_{rk}||x_r| < |a_{rr}||x_r|. \quad \blacksquare$$

Beispielsweise ist die bei der Berechnung von Spline-Funktionen (§17) auftretende tridiagonale Matrix

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

diagonal dominant und damit LR-zerlegbar.

Spezielle Matrizen, die das Kriterium in Satz (4.1) erfüllen, sind die positiv definiten Matrizen.

4.2 Das CHOLESKY-Verfahren für positiv definite Matrizen

Eine (n,n)-Matrix A heißt positiv definit, falls

(1) $A = A^T$, d.h. A ist symmetrisch,

(2) $x^T A x > 0$ für alle $x \in \mathbb{R}^n$, $x \neq 0$.

Für positiv definite Matrizen A kann die LR-Zerlegung ohne Pivotsuche durchgeführt werden.

(4.4) Satz: Sei A positiv definit.

(i) Alle Hauptabschnitt-Matrizen von A sind positiv definit und regulär. Insbesondere ist A regulär.

(ii) Es gibt genau eine linke Dreiecksmatrix L mit $l_{ii} > 0$, $i = 1, \dots, n$, so daß gilt

$$\boxed{A = L L^T}$$

(Beachte: $l_{ii} = 1$ wird nicht gefordert)

Beweis: zu (i): Übung

zu (ii): Nach Satz (4.1) gibt es genau eine Zerlegung

$$A = U V,$$

$U = (u_{ik})$: linke Dreiecksmatrix, $u_{ii} = 1$,

$V = (v_{ik})$: reguläre rechte Dreiecksmatrix

Sei

$$D = \begin{pmatrix} v_{11} & & 0 \\ & \ddots & \\ 0 & & v_{nn} \end{pmatrix}, \quad v_{ii} \neq 0.$$

Setze

$R = D^{-1} V$: rechte Dreiecksmatrix, $r_{ii} = 1$.

$\Rightarrow A = U D R$, $A = A^T = R^T D^T U^T = R^T D U^T$.

Wegen der Eindeutigkeit der Zerlegung folgt

$$R^T = U, \quad \text{d.h.} \quad A = U D U^T = R^T D R.$$

Behauptung: D ist positiv definit, d.h. $v_{ii} > 0$.

Für alle $x \neq 0$ gilt

$$0 < x^T A x = x^T R^T D R x = (Rx)^T D R x.$$

$\Rightarrow 0 < y^T D y$ für alle $y \neq 0$, da R regulär,

$\Rightarrow D$ positiv definit.

Mit

$$D^{1/2} := \begin{pmatrix} \sqrt{v_{11}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{v_{nn}} \end{pmatrix}, \qquad L := U D^{1/2}$$

gilt

$$\boxed{A = L^T L}. \quad \blacksquare$$

Berechnung der Elemente l_{ik}:

Man geht induktiv spaltenweise vor:

$$l_{11} = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}} \quad (i = 2, \dots, n).$$

Sei l_{ij} für $j \leq k-1$ bekannt.

Die Bestimmungsgleichungen für l_{ik}, $i \geq k$, lauten

$$a_{kk} = l_{k1}^2 + \ldots + l_{kk}^2$$

$$a_{ik} = \sum_{j=1}^{k-1} l_{ij}\, l_{kj} + l_{ik}\, l_{kk}\,.$$

Nach (4.4)(ii) ist die Auflösung möglich

$$l_{kk} = \Big(a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2\Big)^{1/2}$$

(4.5)

$$l_{ik} = \frac{1}{l_{kk}}\Big(a_{ik} - \sum_{j=1}^{k-1} l_{ij}\, l_{kj}\Big), \qquad i \geq k+1.$$

Insbesondere erhält man die Abschätzung

$$|l_{kj}| \leq \sqrt{a_{kk}}, \qquad j = 1, \ldots, k, \quad k = 1, 2, \ldots, n.$$

<u>Rechenaufwand</u>:

n Wurzeln, $\boxed{\frac{n^3}{6}} + \frac{n^2}{2} - \frac{2}{3}n$ Operationen.

Die Lösung des LGS $Ax = b$ nach der Methode von CHOLESKY erfolgt in den drei Schritten

(4.6)

1.	$A = LL^T$: CHOLESKY-Zerlegung
2.	$Lc = b$: Vorwärtseinsetzen
3.	$L^T x = c$: Rückwärtseinsetzen

Bei positiv definiten Matrizen A sind die Hauptdiagonalelemente $a_{ii} = e_i^T A e_i > 0$ positiv. Darüberhinaus kann man leicht zeigen, daß diagonal-dominante Matrizen (vgl. Def (4.2)) mit $a_{ii} > 0$, d.h.

$$a_{ii} > \sum_{k \neq i} |a_{ik}| \quad (i = 1, \dots, n),$$

positiv definit sind.

Für eine positiv definite Matrix A ist die Reduktion der quadratischen Form $x^T A x$ auf eine Summe von Quadraten (im Körper der reellen Zahlen) möglich:

$$x^T A x = x^T L L^T x = (L^T x)^T (L^T x)$$

$$= \sum_{j=1}^{n} \left(\sum_{k=j}^{n} l_{kj} x_k \right)^2 .$$

Zusätzlich ergibt sich für die Hauptabschnittsmatrizen (Hauptminoren):

$$\det A = \prod_{j=1}^{n} l_{jj}^2 = \prod_{j=1}^{n} a_{jj}^{(j)} > 0 ,$$

$$\det \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix} = \prod_{j=1}^{k} l_{jj}^2 > 0 .$$

(4.7) <u>Folgerung</u>: Eine symmetrische Matrix A ist genau dann positiv definit, wenn

$$\det \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix} > 0 \quad \text{für } k = 1, \dots, n .$$

Zum Beispiel ist die bei der Diskretisierung von Randwertproblemen für Differentialgleichungen auftretende Matrix

$$A_n = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \Bigg\} n$$

positiv definit, denn mittels der Rekursion

$$\det A_{n+1} = 2 \det A_n - \det A_{n-1}$$

erkennt man $\det A_n = n+1 > 0$.

4.3 Bandmatrizen

In vielen Anwendungen spielen Bandmatrizen eine wichtige Rolle

(4.8) Definition: Unter der Bandbreite einer Matrix A versteht man die kleinste natürliche Zahl $m < n$, so daß gilt

$$a_{ik} = 0 \quad \text{für alle } i \text{ und } k \text{ mit } |i-k| > m.$$

Beispielsweise hat

$$A = \begin{pmatrix} * & * & * & & & & \\ * & * & * & * & & & \\ * & * & * & * & * & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & * & * & * & * & * \\ & & & * & * & * & * \\ & & & & * & * & * \end{pmatrix}$$

die Bandbreite $m = 2$. Ein Blick auf den Gaußalgorithmus zeigt leicht die folgende Aussage:

(4.9) Satz: Besitzt die Matrix A mit der Bandbreite m eine LR-Zerlegung $A = LR$, so haben L und R ebenfalls die Bandbreite m, denn es gilt

$$l_{ik} = 0 \quad \text{für } i,k \text{ mit } i-k > m,$$
$$r_{ik} = 0 \quad \text{für } i,k \text{ mit } k-i > m.$$

(4.10) Korollar: Die Linksdreiecksmatrix L der Cholesky-Zerlegung $A = LL^T$ einer positiv definiten Bandmatrix mit der Bandbreite m besitzt ebenfalls Bandbreite m.

Besonders einfach zu behandelnde LGS erhält man für tridiagonale Matrizen A der Bandbreite $m=1$. Zu lösen sei das LGS $Ax=d$ mit

$$A=\begin{pmatrix} a_1 & b_1 & & & \\ c_2 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{pmatrix}, \quad d=\begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}$$

Es existiere die LR-Zerlegung $A=LR$. Nach (4.9) sind L und R bidiagonal und können in der Form angesetzt werden

$$L=\begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & l_3 & 1 & & \\ & & \ddots & \ddots & \\ & & & l_n & 1 \end{pmatrix}, \quad R=\begin{pmatrix} m_1 & r_1 & & & \\ & m_2 & r_2 & & \\ & & \ddots & \ddots & \\ & & & m_{n-1} & r_{n-1} \\ & & & & m_n \end{pmatrix}.$$

Die Ausmultiplikation $A=LR$ führt auf die Beziehung $r_i=b_i$ $(i=1,..,n)$

und den folgenden Algorithmus zur Lösung von $Ax=d$:

(4.11)

$A=LR$

$m_1=a_1$

für $i=2,..,n$:

$l_i=c_i/m_{i-1}$

$m_i=a_i-l_i\cdot b_{i-1}$

$Ly=d$:

$y_1=d_1$

für $i=2,..,n$:

$y_i=d_i-l_i\cdot y_{i-1}$

$Rx=y$:

$x_n=y_n/m_n$

für $i=n-1,\ n-2,..,1$:

$x_i=(y_i-b_i\cdot x_{i+1})/m_i$

§ 5 Fehlerabschätzungen

A sei eine reguläre (n,n)-Matrix und $x \in \mathbb{R}^n$ sei die eindeutige Lösung des LGS $Ax = b$. Wir untersuchen die Frage, wie sich Fehler in A, b, d.h. Änderungen der Form

(1) $b \to b + \Delta b$,

(2) $A \to A + \Delta A = A(I + F)$, $F := A^{-1}\Delta A$

auf die Lösung x auswirken. Als Fehlermaße verwenden wir Normen.

(5.1) Definition: Eine (Vektor-)Norm in $\mathbb{R}^n$ ist eine Abbildung $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ mit folgenden Eigenschaften:

(a) $\|x\| > 0$ für alle $x \in \mathbb{R}^n$, $x \neq 0$,

(b) $\|\alpha x\| = |\alpha| \, \|x\|$ für alle $\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$ (Homogenität)

(c) $\|x + y\| \leq \|x\| + \|y\|$ für alle $x, y \in \mathbb{R}^n$ (Dreiecksungleichung).

Beispiele: Im folgenden benutzen wir nur die Normen:

$$\|x\|_2 := \sqrt{x^T x} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$ (euklidische Norm)

$$\|x\|_\infty := \max_i |x_i|$$ (Maximumnorm)

Als bekannt setzen wir voraus (Grundvorlesung)

(a) Jede Norm $\|\cdot\|$ in $\mathbb{R}^n$ ist eine gleichmäßig stetige Funktion bezüglich der üblichen Topologie des $\mathbb{R}^n$,

(b) je zwei Normen $\|\cdot\|$, $\|\cdot\|'$ in $\mathbb{R}^n$ sind äquivalent, d.h. es gibt Konstanten m, M mit

$$m\,\|x\|' \le \|x\| \le M\,\|x\|' \quad \text{für alle } x \in \mathbb{R}^n.$$

Eine <u>Matrixnorm</u> $\|A\|$ für eine (n,n)-Matrix A ist eine Vektornorm in $\mathbb{R}^{n^2}$, d.h. analog zu (5.1) gilt

$$\|A\| > 0 \quad \text{für alle } A \neq 0,$$

$$\|\alpha A\| = |\alpha|\,\|A\|, \quad \alpha \in \mathbb{R},$$

$$\|A+B\| \le \|A\| + \|B\|.$$

Beispielsweise ist die FROBENIUS-Norm

$$\|A\|_F = \Big(\sum_{i,k=1}^{n} a_{ik}^2 \Big)^{1/2}$$

eine Matrix-Norm. Diese ist <u>verträglich</u> mit der euklidischen Norm $\|\cdot\|$ in $\mathbb{R}^n$:

$$\|Ax\|_2 \le \|A\|_F\,\|x\|_2 \quad \text{für } x \in \mathbb{R}^n.$$

Im folgenden beschränken wir uns auf die sog. zugeordnete Matrixnorm.

(5.2) Definition: Sei A eine (n,n)-Matrix und $\|\cdot\|$ eine Vektornorm in $\mathbb{R}^n$. Die Zahl

$$\|A\| := \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

heißt die der Vektor-Norm $\|\cdot\|$ zugeordnete Matrix-Norm (Operator-Norm).

Beispiele: Sei $A = (a_{ik})$

(a) Die $\|\cdot\|_\infty$ zugeordnete Matrix-Norm ist die Zeilensummen-Norm

$$\|A\|_\infty := \max_i \sum_{k=1}^{n} |a_{ik}| .$$

(b) Die $\|\cdot\|_2$ zugeordnete Matrix-Norm ist die Spektral-Norm

$$\|A\|_2 := \left(\rho(A^T A)\right)^{1/2},$$

wobei $\rho(B)$ den Spektralradius von B bedeutet:

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

Dies ersieht man aus

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} (x^T A^T A x)^{1/2}$$

$$= \max\{|\lambda|^{1/2} \mid \lambda \text{ Eigenwert von } A^T A\}.$$

(5.3) Satz: Zugeordnete Matrix-Normen haben folgende Eigenschaften:

(i) $A \to \|A\|$ ist eine Vektornorm in $\mathbb{R}^{n^2}$.

(ii) $\|AB\| \le \|A\| \, \|B\|$, also auch $\|A^k\| \le \|A\|^k$, $k \ge 1$

(iii) $\|I\| = 1$ für die Einheitsmatrix I.

(iV) $\|A\| = \min \{ K : \|Ax\| \le K\|x\| \text{ für alle } x \in \mathbb{R}^n \}$.

(5.4) Satz: Ist F eine (n,n)-Matrix mit $\|F\| < 1$, so existiert $(I+F)^{-1}$ und es gilt

(i) $(I+F)^{-1} = \sum_{k=0}^{\infty} (-F)^k$ (Neumann'sche Reihe).

(ii) $\|(I+F)^{-1}\| \le \frac{1}{1-\|F\|}$.

Beweis: Es gilt

$$\left\| \sum_{k=0}^{\infty} (-F)^k \right\| \le \sum_{k=0}^{\infty} \|F^k\| \le \sum_{k=0}^{\infty} \|F\|^k = \frac{1}{1-\|F\|} < \infty .$$

Also konvergiert die Reihe komponentenweise in $\mathbb{R}^{n^2}$, und es folgt durch gliedweise Multiplikation

$$(I+F) \sum_{k=0}^{\infty} (-F)^k = I .$$

Damit ergibt sich die Darstellung (i), und aus der obigen Abschätzung folgt (ii). ∎

(5.5) **Definition:** Als **Kondition von A** (bzgl. der gewählten Matrixnorm) bezeichnen wir die Zahl

$$\boxed{\text{cond}(A) = \|A\| \, \|A^{-1}\|}$$

(5.6) **Hilfssatz:** Sei ΔA eine Matrix mit

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

Dann ist auch $A + \Delta A$ invertierbar, und es gilt

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

Beweis: Es ist

$$A + \Delta A = A(I + A^{-1}\Delta A) = A(I + F), \qquad F := A^{-1}\Delta A$$

$$\Rightarrow \|F\| = \|A^{-1}\Delta A\| \leq \|A^{-1}\| \, \|\Delta A\| = q < 1.$$

Also existiert nach Satz (5.4) die Inverse von $I + F$ und man erhält

$$\|(A + \Delta A)^{-1}\| = \|(I + F)^{-1} A^{-1}\| \leq \|A^{-1}\| / (1-q). \quad \blacksquare$$

(5.7) **Satz:** Sei x Lösung von $Ax = b$.
Seien ΔA, Δb Störungen von A, b mit

$$q = \text{cond}(A) \frac{\|\Delta A\|}{\|A\|} < 1.$$

Dann ist auch das gestörte System

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

eindeutig lösbar und es gilt

$$\boxed{\frac{\|\Delta x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1-q}\left\{\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}\right\}}.$$

Beweis: Nach Hilfssatz (5.6) ist $A+\Delta A$ invertierbar, das gestörte System

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

also eindeutig lösbar. Subtraktion von $Ax=b$ und Umordnung ergibt

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

Mit (5.6) folgt dann

$$\|\Delta x\| \leq \frac{\|A^{-1}\|}{1-q}\left(\|\Delta b\| + \|\Delta A\|\,\|x\|\right),$$

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|}{1-q}\left(\frac{\|\Delta b\|}{\|x\|} + \|\Delta A\|\right)$$

$$= \frac{\|A^{-1}\|}{1-q}\left(\frac{\|\Delta b\|}{\|b\|}\,\frac{\|b\|}{\|x\|} + \frac{\|\Delta A\|}{\|A\|}\,\|A\|\right).$$

Mit $\|b\| = \|Ax\| \leq \|A\|\,\|x\|$ erhält man die Behauptung. ∎

Die Zahl $\operatorname{cond}(A)$ hat also die Bedeutung eines Verstärkungsfaktors und mißt die Empfindlichkeit der Lösung x gegenüber

Störungen in A und b. Das LGS $Ax = b$ heißt <u>schlecht konditioniert</u>, wenn $\text{cond}(A) >> 1$.

<u>Beispiel</u>: Auswirkung schlechter Kondition:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0.99 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\Delta A = \begin{pmatrix} 0.01 & 0.01 \\ 0 & 0 \end{pmatrix}, \quad \Delta b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad x + \Delta x = \begin{pmatrix} 200/101 \\ -100/101 \end{pmatrix}.$$

Obwohl der Fehler in A bei 1% liegt, haben x, $x + \Delta x$ nichts mehr miteinander zu tun.

Erklärung:

$$\|A\|_\infty = 2, \quad A^{-1} = \begin{pmatrix} -99 & 100 \\ 100 & -100 \end{pmatrix}, \quad \|A^{-1}\|_\infty = 200,$$

$$\text{cond}_\infty(A) = 400 \quad !$$

<u>Geometrisch</u>: Die Zeilenvektoren a_1, a_2 von A haben beinahe die gleiche Richtung.

Die Kondition eines Problems kann ggf. durch <u>Skalierung</u> der Matrix A verbessert werden. Unter Skalierung versteht man den Übergang

$$A \to DA, \quad D = \begin{pmatrix} d_1 & 0 \\ 0 & d_n \end{pmatrix}, \quad d_i \neq 0,$$

d.h. die i-te Zeile von A wird mit d_i multipliziert. Die optimale Wahl einer Diagonalmatrix D, welche $\operatorname{cond}(DA)$ möglichst klein macht, erhält man durch folgenden Satz (ohne Beweis).

(5.8) Satz: (Van der Sluis)

Für $A = (a_{ik})$ sei

$$\sum_{k=1}^{n} |a_{ik}| = 1, \quad i = 1, \dots, n \quad (\text{insbesondere } \|A\|_\infty = 1).$$

Dann gilt für jede Diagonalmatrix D mit $\det D \neq 0$

$$\operatorname{cond}_\infty(DA) \geq \operatorname{cond}_\infty(A).$$

Folgerung: Für eine beliebige reguläre Matrix $A = (a_{ik})$ ist mit der Skalierung

$$D = \operatorname{diag}(d_i), \quad d_i := \left(\sum_{k=1}^{n} |a_{ik}| \right)^{-1}$$

die Kondition $\operatorname{cond}_\infty(DA)$ möglichst klein.

§6 Die QR-Zerlegung einer Matrix, das Verfahren von Householder

Sei A eine (n,n)-Matrix (reell, nicht notwendig regulär).

LR-Zerlegung (ohne Pivotsuche)

$$\boxed{A = LR}$$

L linke Dreiecksmatrix
R rechte "

QR-Zerlegung

$$\boxed{A = QR}$$

Q orthogonal, d.h. $Q^T Q = I$,
R rechte Dreiecksmatrix.

Motivation zur QR-Zerlegung:
Zur Lösung des LGS $Ax = b$ erzeugt man bei der LR-Zerlegung und Gauß-Elimination eine Sequenz

$$(A, b) = (A^{(1)}, b^{(1)}) \to \ldots \to (A^{(j)}, b^{(j)}) \to \ldots \to (A^{(n)}, b^{(n)}) = (R, c)$$

$$(A^{(j+1)}, b^{(j+1)}) = L_j (A^{(j)}, b^{(j)}).$$

Sei $\varepsilon^{(j)}$ der Rundungsfehler bei der Berechnung von $(A^{(j)}, b^{(j)})$. Für irgendeine Vektornorm $\|x\|$ gilt nach Satz (5.7) die Abschätzung

$$\frac{\|\Delta x\|}{\|x\|} \leq \sum_{j=1}^{n} \varepsilon^{(j)} \operatorname{cond}(A^{(j)})$$

Die Gauß-Elimination ist daher nicht gutartig, falls

$$\operatorname{cond}(A^{(1)}) \gg \operatorname{cond}(A^{(1)}) = \operatorname{cond}(A).$$

Idee: Wähle Matrix Q_j mit Übergang

$$(A^{(j+1)}, b^{(j+1)}) = Q_j (A^{(j)}, b^{(j)}), \quad \operatorname{cond}(A^{(j+1)}) = \operatorname{cond}(A^{(j)}).$$

Dazu beschränken wir uns auf die euklidische Norm

$$\|x\| = \|x\|_2 = (x^T x)^{1/2}, \qquad \|A\| = \|A\|_2.$$

(6.1) Hilfssatz: Sei Q orthogonal.

(i) $\|Q\|_2 = 1$

(ii) $\|QA\|_2 = \|A\|_2$ für alle A

(iii) Wenn A regulär ist, gilt

$$\operatorname{cond}_2(QA) = \operatorname{cond}_2(A).$$

Beweis: Übung

Das Verfahren von Householder

Sei $w \in \mathbb{R}^n$ mit $w^T w = 1$ und sei

$$Q := I - 2 w w^T, \qquad w w^T = (w_i w_k)$$

Q ist symmetrisch:

$$Q^T = I - 2 (w w^T)^T = I - 2 w w^T = Q$$

Q ist orthogonal wegen $w^T w = 1$:

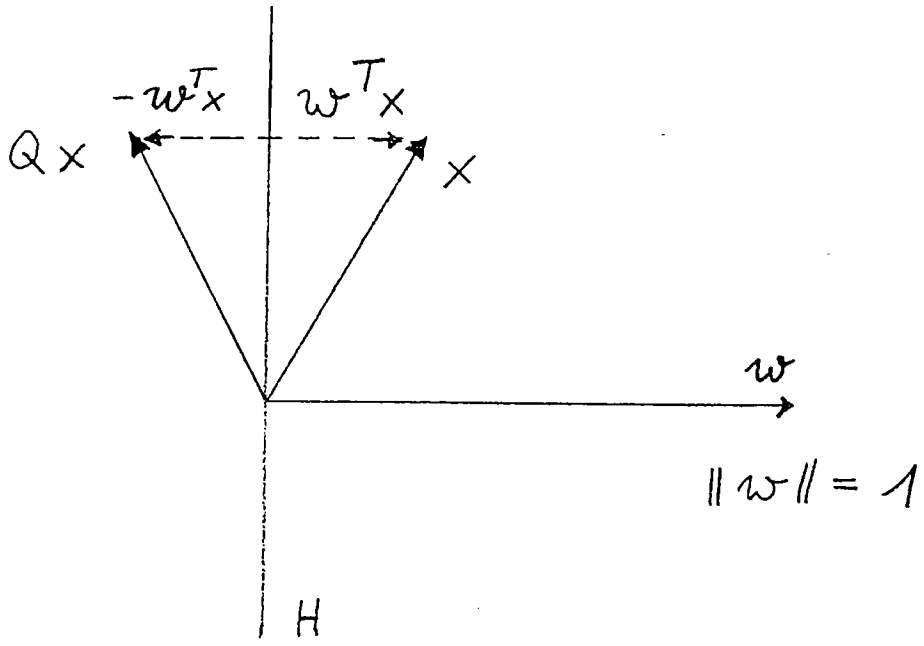
$$Q^T Q = (I - 2ww^T)(I - 2ww^T)$$
$$= I - 2ww^T - 2ww^T + 4ww^T ww^T = I$$

Für $x \in \mathbb{R}^n$ bedeutet

$$Qx = (I - 2ww^T)x = x - 2(w^T x)w$$

eine Spiegelung an der Hyperebene

$H = \{ z \in \mathbb{R}^n \mid w^T z = 0 \}$:



Q ist orthogonal:

$$Q^T Q = QQ = (I - 2ww^T)(I - 2ww^T)$$

$$= I - 2ww^T - 2ww^T + 4w\underbrace{w^T w}_{=1}w^T = I.$$

Problem: Sei $x = (x_1, \dots, x_n)^T \neq 0$ vorgegeben. Bestimme $w \in \mathbb{R}^n$, $w^T w = 1$, mit

$$Qx = k e_1, \quad k \in \mathbb{R}.$$

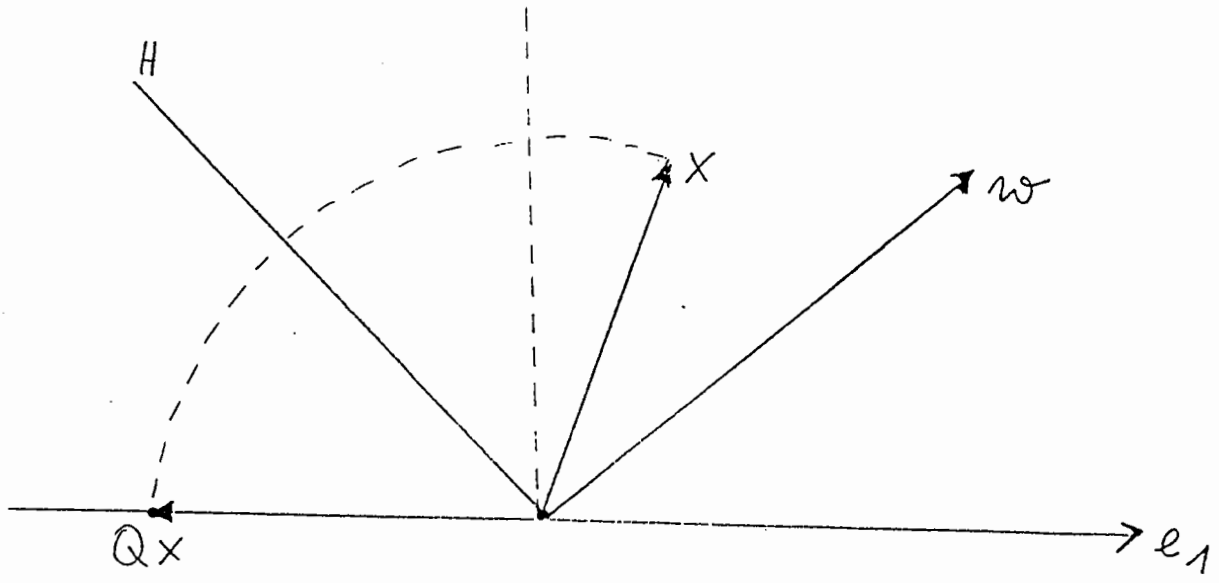
Q ist Spiegelung an einer Hyperebene: vgl. Figur

Analytische Berechnung von Q: $Qx = k e_1$

$$\Rightarrow |k| = \|Qx\| = \|x\|, \quad k = \pm \|x\|.$$

$$Qx = (I - 2ww^T)x = x - 2w(w^T x) = k e_1$$

$$\Rightarrow w = \frac{x - k e_1}{\|x - k e_1\|}, \quad \text{da } \|w\| = 1$$



$$\|x - k e_1\| = \left((x_1 - k)^2 + x_2^2 + \ldots + x_n^2 \right)^{1/2}.$$

Keine Auslöschung tritt auf für

$$k = - \operatorname{sign}(x_1) \|x\|, \quad (x_1 - k)^2 = \left(|x_1| + \|x\| \right)^2.$$

$$\Rightarrow \|x - k e_1\|^2 = \|x\|^2 + 2 \|x\| |x_1| + \|x\|^2 = 2 \|x\| \left(\|x\| + |x_1| \right)$$

Insgesamt erhalten wir

(6.2)

$$Q = I - 2 w w^T = I - 2 \frac{(x - k e_1)(x - k e_1)^T}{\|x - k e_1\|^2}$$

$$= I - \beta u u^T,$$

$$k = - \operatorname{sign}(x_1) \|x\|, \quad \beta = \frac{1}{\|x\| \left(|x_1| + \|x\| \right)}$$

$$u := x - k e_1 = \begin{pmatrix} \operatorname{sign}(x_1) \left(|x_1| + \|x\| \right) \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Householder - Transformation

QR-Zerlegung von A: Bilde Sequenz

$$A = A^{(1)} \to A^{(2)} \to \ldots \to A^{(n)} = R,$$

$$A^{(j+1)} = Q_j A^{(j)}, \quad Q_j \text{ orthogonal}.$$

j-ter Schritt ($j \geq 1$): Sei

$$A^{(j)} = \begin{pmatrix} * & \cdots & * & * & \cdots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & * & * & \cdots & * \\ \hline 0 & & & a_{jj}^{(j)} & \cdots & a_{jn}^{(j)} \\ & & & \vdots & & \vdots \\ & & & a_{nj}^{(j)} & \cdots & a_{nn}^{(j)} \end{pmatrix} \begin{matrix} \} j-1 \\ \\ \} n-j+1 \end{matrix}$$

$$x := \left(a_{jj}^{(j)}, \ldots, a_{nj}^{(j)} \right)^T \in \mathbb{R}^{n-j+1} .$$

1. Fall: $x = 0$: A ist singulär (Beweis!), setze $Q_j := I$

2. Fall: $x \neq 0$: Bestimme nach (6.2) die orthogonale $(n-j+1, n-j+1)$-Matrix $\tilde{Q}_j$ mit

$$\tilde{Q}_j \begin{pmatrix} a_{jj}^{(j)} \\ \vdots \\ a_{nj}^{(j)} \end{pmatrix} = k \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{n+1-j} .$$

Setze

$$Q_j = \begin{pmatrix} I_{j-1} & 0 \\ 0 & \tilde{Q}_j \end{pmatrix} \quad (n,n) \text{ orthogonal, symmetrisch}$$

Nach $n-1$ Schritten erhält man

$$(6.3) \qquad R := A^{(n)} = Q_{n-1} Q_{n-2} \cdots Q_1 A .$$

Definiere die orthogonale Matrix

$$Q := (Q_{n-1} \cdots Q_1)^{-1} = Q_1 \cdots Q_{n-1}, \text{ da } Q_j \text{ orthog., symm.}$$

$$\Rightarrow \quad A = QR$$

(6.4) Satz: Zu jeder (n,n)-Matrix A existiert eine orthogonale (n,n)-Matrix Q und eine rechte Dreiecksmatrix R mit

$$A = QR.$$

Ist A regulär, so ist R regulär.

Bei einer regulären Matrix A bildet man zur Lösung des LGS $Ax = b$ analog zu (6.3) den Ausdruck

$$c := b^{(n)} = Q_{n-1} \cdot \ldots \cdot Q_1 b$$

und löst dann das gestaffelte LGS $Rx = c$.

Anzahl der Operationen: $\boxed{\approx \frac{2}{3} n^3}$

Die QR-Zerlegung kann unmittelbar auf nichtquadratische (m,n)-Matrizen A $(m > n)$ erweitert werden. Hier bildet man eine Sequenz

$$A^{(j+1)} = Q_j A^{(j)} \quad (j \geq 1), \quad A^{(1)} = A$$

Q_j: orthogonale (m,m)-Matrix.

Wegen $m > n$ erhält man nach n Schritten

(6.5)

$$A^{(n+1)} = \tilde{Q} A = \begin{pmatrix} R \\ 0 \end{pmatrix} \begin{matrix} \} n \\ \} m-n \end{matrix}$$

$$\tilde{Q} = Q_n \cdot \ldots \cdot Q_1 \quad \text{orthogonal},$$

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix} \text{ obere Dreiecksmatrix}$$

Praktische Durchführung mit (6.2)

$$A^{(j)} = \underbrace{\left(\begin{array}{c|c} * & * \\ \hline 0 & \tilde{A}^{(j)} \end{array}\right)}_{n} \begin{matrix} j-1 \\ m-j+1 \end{matrix}, \quad A^{(1)} = A$$

$$Q_j = \underbrace{\begin{pmatrix} I_{j-1} & 0 \\ 0 & \tilde{Q}_j \end{pmatrix}}_{m} \begin{matrix} j-1 \\ m-j+1 \end{matrix}$$

$$\tilde{Q}_j = I - \beta_j u_j u_j^T, \quad j = 1, \ldots, n,$$

wobei nach (6.2)

$$x = \left(a_{jj}^{(j)}, \ldots, a_{mj}^{(j)} \right)^T \in \mathbb{R}^{m-j+1},$$

$$k_j = -\operatorname{sign}(x_1) \|x\|,$$

$$\beta_j = \frac{1}{\|x\| \left(|x_1| + \|x\| \right)},$$

$$u_j = x - k_j e_j.$$

$$\tilde{Q}_j \tilde{A}^{(j)} = \tilde{A}^{(j)} - u_j s_j^T$$

$$s_j^T = \beta_j u_j^T \tilde{A}^{(j)} .$$

d.h. $(u_j s_j^T)_{i,k} = a_{ij} \beta_j \sum_{\ell=j}^{m} a_{\ell j} a_{\ell k}$

Programm $QR(A, d)$

die u_j stehen spaltenweise
im linken Teil von A,
$R \setminus \mathrm{diag}(R)$ steht im rechten
Teil von A,
$\mathrm{diag}(R)$ steht auf $d = (d_1, \ldots, d_n)$.

für $j = 1, \ldots, n$:

$$xnorm = \left(\sum_{i=j}^{m} a_{ij}^2 \right)^{1/2}$$

falls $xnorm = 0$: STOP

$d_j = -\mathrm{sign}(a_{jj}) \times xnorm$

$beta = 1/(xnorm(|a_{jj}| + xnorm))$

$a_{jj} = a_{jj} - d_j$

für $k = j+1, \ldots, n$:

$$s = beta \times \sum_{\ell=j}^{m} a_{\ell j} a_{\ell k}$$

für $i = j, \ldots, m$:

$$a_{ik} = a_{ik} - a_{ij} \times s .$$

§ 7 Grundlagen der Linearen Optimierung

7.1 Ein Beispiel: Optimale Produktionsplanung

Ein Unternehmer produziert n Produkte $P_1, \dots, P_n$, zu deren Herstellung m Aktivitäten $A_1, \dots, A_m$ (Rohstoffe, Arbeitskräfte, Arbeitsstunden o.ä.) benötigt werden. Das Produkt P_i enthalte a_{ji} Anteile der Aktivität A_j und möge beim Verkauf pro Einheit einen Reingewinn von c_i Zahlungseinheiten erzielen. Von der Aktivität A_j sei die Menge b_j verfügbar. Die Produktionsmenge x_i des Produktes P_i soll nun so bestimmt werden, daß der Gewinn maximal wird. Es ist daher das Maximum der Zielfunktion

$$z(x) = \sum_{i=1}^{n} c_i x_i$$

zu finden unter den Nebenbedingungen

$$\sum_{i=1}^{n} a_{ji} x_i \le b_j, \quad j = 1, \dots, m$$

$$x_i \ge 0, \quad i = 1, \dots, n.$$

Zahlenbeispiel: Der Besitzer einer kleinen Schuhfabrik will je ein Modell eines Damen- und eines Herrenschuhs herstellen. Er verfügt über 40 Angestellte und über einen Maschinenpark von 10 Maschinen. Die (pro Monat) zur Verfügung stehende Arbeitszeit bzw. Ledermenge ist in der folgenden Tabelle enthalten:

Angaben zum Produktionsproblem

	Damenschuh	Herrenschuh	verfügbar
Herstellungszeit [h]	20	10	8000
Maschinenbearbeitung [h]	4	5	2000
Lederbedarf [dm²]	6	15	4500
Reingewinn [Fr]	16	32	–

Mit den Optimierungsvariablen

x_1: Zahl der produzierten Damenschuhe

x_2: " " " Herrenschuhe

lautet die Optimierungsaufgabe:

maximiere $z = 16 x_1 + 32 x_2$

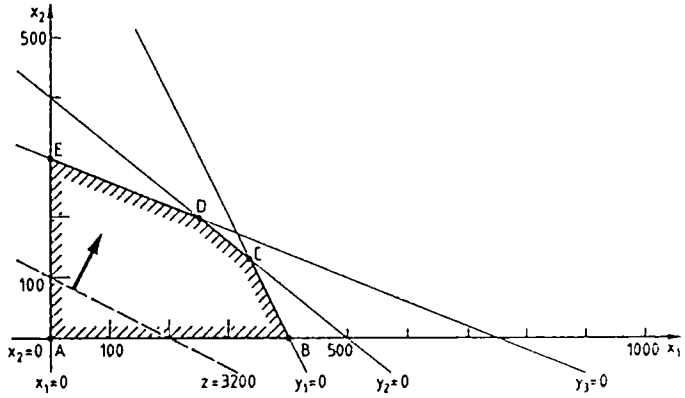
unter

$$
\begin{aligned}
20 x_1 + 10 x_2 &\leq 8000 \\
4 x_1 + 5 x_2 &\leq 2000 \\
6 x_1 + 15 x_2 &\leq 4500 \qquad (7.1) \\
x_1 &\geq 0 \\
x_2 &\geq 0
\end{aligned}
$$

Der durch die obigen Nebenbedingungen (Ungleichungen) beschriebene zulässige Bereich K ist ein konvexes Polyeder. Durch Einführung von Schlupfvariablen y_1, y_2, y_3 für die ersten drei Ungleichungen geht K über in

$$
\begin{aligned}
20 x_1 + 10 x_2 + y_1 &= 8000 \\
4 x_1 + 5 x_2 + y_2 &= 2000 \\
6 x_1 + 15 x_2 + y_3 &= 4500 \qquad (7.2) \\
x_1, x_2 &\geq 0 \\
y_1, y_2, y_3 &\geq 0
\end{aligned}
$$

Die graphische Lösung des Optimierungsproblems ist in der folgenden Figur angegeben:



Die Niveaulinien der linearen Zielfunktion

$$z = c_1 x_1 + c_2 x_2 = \text{const}$$

bestehen aus einer Schar von parallelen Geraden. Der Pfeil gibt die Richtung an, in welcher der Wert von z zunimmt. Offenbar nimmt die Zielfunktion z ihren maximalen Wert in der Ecke D des zulässigen Bereiches K an. Die Ecken von K sind A, B, C, D, E; z.B. berechnet man die Ecken B, C, D mit zugehörigem Wert z:

Ecke	z	x_1	x_2	y_1	y_2	y_3
B	6 400	400	0	0	400	2100
C	9600	1000/3	400/3	0	0	500
D	10 400	250	200	100	0	0

In jeder Ecke haben genau zwei Variable den Wert 0 (Basislösungen). Bei Durchlaufung der Ecken $B \to C \to D$ vergrößert sich der Wert der Zielfunktion. Das Optimum in der Ecke D ist

$$x_1 = 250 \text{ (Damenschuhe)}, \quad x_2 = 200 \text{ (Herrenschuhe)} : z = 10400 .$$

7.2 Das Standardproblem

Das Standardproblem der linearen Optimierung lautet

(7.3)

> Minimiere
> $z(x) = c\,x$
> unter den Nebenbedingungen
> $A\,x = b, \quad x \geq 0$

Dabei sind

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \quad c = (c_1, \ldots, c_n) \in \mathbb{R}^n,$$

$A = (a_{ji})$ $m \times n$ Matrix, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m$

Im allgemeinen gilt hierbei $\operatorname{rang}(A) = m$ und $m < n$.

Viele ursprünglich anders formulierte Probleme lassen sich auf diese Standardform zurückführen.

(1) Ungleichungen: Zu einer $m \times n$ Matrix A (nicht notwendig $m < n$) betrachten wir das Problem

(7.4)

> minimiere $c\,x$
> unter $A\,x \leq b, \quad x \geq 0$

Die Ungleichung $Ax \leq b$ bedeutet komponentenweise

$$\sum_{i=1}^{n} a_{ji}\, x_i \leq b_j, \quad j = 1, \ldots, m .$$

Durch Einführung der Schlupf-Variablen

$$y := b - Ax \in \mathbb{R}^m$$

sind die Ungleichungen äquivalent zu (vgl. Beispiel (7.1), (7.2))

$$Ax + y = b, \quad y \geq 0 .$$

Mittels

$$\tilde{c} = (c, 0) \in \mathbb{R}^{n+m}, \quad \tilde{x} = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{n+m},$$

$$\tilde{A} = \begin{pmatrix} a_{11} \cdots a_{1n} & 1 & & 0 \\ \vdots & & \ddots & \\ a_{m1} \cdots a_{mn} & 0 & & 1 \end{pmatrix} = (A \mid I_m)$$

$m \times (n+m)$ Matrix

wird (7.4) überführt in die Standardform

(7.5)

minimiere	$\tilde{c}\tilde{x}$
unter	$\tilde{A}\tilde{x} = b, \ \tilde{x} \geq 0$

.

Für Ungleichungen der Form

$$Ax \geq b$$

definiert man Schlupfvariablen $y \in \mathbb{R}^m$ gemäß

$$Ax - y = b, \quad y \geq 0,$$

d.h.

$$-Ax + y = -b .$$

(2) Freie Variable: Wenn die Variable x_i keiner Vorzeichenbeschränkung $x_i \geq 0$ unterliegt, so ersetzt man

$$x_i = u_i - v_i$$
$$u_i = x_i^+ = \max\{0, x_i\} \geq 0$$
$$v_i = x_i^- = \max\{0, -x_i\} \geq 0$$

und erhält ein LP in den $n+1$ Variablen $x_1, \ldots, x_{i-1}, u_i, v_i, x_{i+1}, \ldots, x_n$.

(3) Die Maximierung einer Funktion $z(x) = cx$ ist äquivalent zur Minimierung von $-z(x) = (-c)x$.

Die Menge

$$K = \{x \in \mathbb{R}^n \mid Ax = b, \ x \geq 0\}$$

der Vektoren, die den Nebenbedingungen des Standardproblems genügen, heißt zulässige Menge. K ist ein konvexes Polyeder. Die Eigenschaft der Konvexität

$\alpha x_1 + (1-\alpha) x_2 \in K$ für alle $x_1, x_2 \in K$, $0 \leq \alpha \leq 1$,

prüft man leicht nach.

Es ist geometrisch plausibel (vgl. Fig. 7.1), daß die lineare Funktion $z(x) = cx$ ihr Minimum in einer Ecke des konvexen Polyeders K annimmt: diese Aussage ist der Hauptsatz der linearen Optimierung. Bei der Suche nach der optimalen Lösung kann man sich daher auf die Ecken von K beschränken. Wir verzichten hier auf eine genaue Definition einer Ecke von K und arbeiten direkt mit dem zu einer Ecke äquivalenten Begriff einer zulässigen Basislösung.

7.3 Das Simplexverfahren

Die am häufigsten benutzte Methode zur Lösung des Standardproblems

$$\min \{ cx \mid Ax = b,\ x \geq 0 \},$$
$$\operatorname{rang} A = m < n,$$

ist das Simplexverfahren. In seiner Durchführung unterscheidet man die beiden Phasen:

Phase I: Bestimmung einer zulässigen Basislösung (Ecke) von K.

Phase II: Entscheidung, ob die vorliegende Basislösung optimal ist (Abbruchkriterium). Übergang von einer Basislösung zu einer benachbarten, für die die Zielfunktion verkleinert werden kann.

Wir betrachten zunächst das LGS $Ax = b$, dessen Lösungsgesamtheit wegen $\operatorname{rang} A = m$ ein $(n-m)$-dimensionaler affiner Unterraum ist. Die Spalten von A seien a^j, $j = 1, \dots, n$.

(7.6) Definition: Ein Indexvektor

$$B = (i_1, \dots, i_m)$$

von m verschiedenen Indizes $i_R \in \{1, \dots, n\}$ heißt Basis, wenn die Spalten a^j, $j \in B$, linear unabhängig sind. Ein zu B komplementärer Indexvektor

$$N = (j_1, \dots, j_{n-m}),\ j_k \in \{1, \dots, n\},\ j_k \notin B,$$

heißt Nichtbasis; d.h. es gilt die

Zerlegung $B \oplus N = \{1, \dots, n\}$.

Bezeichnungen:

x_j, $j \in B$, heißen Basis-Variable,
x_j, $j \in N$, heißen Nicht-Basis-Variable,
$x_B := (x_j)_{j \in B}$, $x_N := (x_j)_{j \in N}$,
A_B: $m \times m$ Matrix mit Spalten a^j, $j \in B$,
A_N: $m \times (n-m)$ Matrix mit Spalten a^j, $j \in N$.

Beispiel: $B = (4, 2, 3, 1)$, $x_B = (x_4, x_2, x_3, x_1)^T$.
Für $B = (1, \dots, m)$ hat man die direkte Zerlegung

$$A = (A_B \mid A_N), \quad x = \begin{pmatrix} x_B \\ x_N \end{pmatrix}.$$

Bei einer beliebigen Basis B geht das LGS $Ax = b$ über in

$$Ax = A_B x_B + A_N x_N = b.$$

Damit erhält man eine Parametrisierung des $(n-m)$-dim. Lösungsraumes von $Ax = b$ durch

$$(7.7) \quad \boxed{x_B = A_B^{-1} b - A_B^{-1} A_N x_N, \quad x_N \in \mathbb{R}^{n-m}}$$

x_B: abhängige Variable,
x_N: unabhängige Variable.

Wir zerlegen c entsprechend in $c_B \in \mathbb{R}^m$, $c_N \in \mathbb{R}^{n-m}$. Durch Einsetzen des Ausdruckes (7.7) ergibt sich die Zielfunktion zu

$$(7.8) \quad \boxed{\begin{aligned} z(x) = cx &= c_B x_B + c_N x_N \\ &= c_B A_B^{-1} b - (c_B A_B^{-1} A_N - c_N) x_N \\ &=: z_0 - \tau_N x_N \end{aligned}}$$

mit $z_0 := c_B A_B^{-1} b \in \mathbb{R}$,
$\tau_N := c_B A_B^{-1} A_N - c_N = (\tau_j)_{j \in N} \in \mathbb{R}^{n-m}$
Vektor der reduzierten Kosten.

(7.9) Definition: Sei B eine Basis. $x \in \mathbb{R}^n$ heißt Basis-Lösung von $Ax = b$, falls $x_N = 0$, $x_B = A_B^{-1} b$. Eine Basis-Lösung x heißt

a) <u>zulässig</u>, wenn $x_B \geq 0$,
b) <u>nicht-entartet</u>, wenn $x_B > 0$.

Die Darstellung (7.9) der Zielfunktion zeigt sofort das folgende Optimalitäts-kriterium:

(7.10) <u>Satz:</u> (<u>Hinreichendes Optimalitäts-kriterium.</u>)

Sei B eine Basis mit den Eigenschaften:

(1) die zuhörige Basis-Lösung x ist zulässig, d.h. $x_B \geq 0$,

(2) $\tau_N = c_B A_B^{-1} A_N - c_N \leq 0$.

Dann ist x optimal für das LP (7.6), und der optimale Wert ist $z_0 = c_B A_B^{-1} b$.

<u>Beweis:</u> Für jeden zulässigen Punkt $\tilde{x}$ ist $\tilde{x}_N \geq 0$, und daher folgt aus (7.9) wegen $\tau_N \leq 0$

$$z(\tilde{x}) = c\tilde{x} = z_0 - \tau_N \tilde{x}_N \geq z_0 = z(x). \quad \blacksquare$$

(7.11) <u>Spezialfall</u>: <u>LP mit Ungleichungen:</u>

Das Problem

$$\min\{ cx \mid Ax \leq b,\ x \geq 0 \}$$

ist äquivalent zu

$$\text{minimiere } cx + 0y = \tilde{c}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$\text{unter } Ax + y = \tilde{A}\begin{pmatrix} x \\ y \end{pmatrix} = b$$

$$x \geq 0,\ y \geq 0$$

mit $\tilde{A} = (A \mid I_m)$. Für die Basis

$$B = (n+1, \ldots, n+m), \quad N = (1, \ldots, n)$$

ist die Basislösung $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix} \in \mathbb{R}^{n+m}$. Diese ist zulässig, wenn $b \geq 0$.

Die reduzierten Kosten sind wegen $\tilde{A}_B = I_m$, $\tilde{c}_B = 0$, $\tilde{c}_N = c$:

$$\tau_N = \tilde{c}_B \tilde{A}_B^{-1} \tilde{A}_N - \tilde{c}_N = -c \in \mathbb{R}^n,$$

$$z_0 = 0.$$

Das <u>Simplex-Verfahren</u> basiert auf dem hinreichenden Optimalitäts-kriterium (7.10). Man hat also eine Basis zu finden mit

(1) $x_B \geq 0$, $x_N = 0$,

(2) $\tau_N \leq 0$.

Man startet dabei mit einer Basis B, für die (1) erfüllt ist. Wenn das Kriterium (2) verletzt ist, so geht man zu einer benachbarten Basis B' über (Basistausch), in der die Zielfunktion "verbessert" wird. Zur Basis B benötigt man die allgemeine Darstellung (7.7)

$$x_B = A_B^{-1} b - A_B^{-1} A_N x_N \,, \quad x_N \in \mathbb{R}^{n-m}.$$

Die Multiplikation mit A_B^{-1} kann mittels elementarer Zeilenoperationen (GAUSS-JORDAN-Elimination) durchgeführt werden. Das folgende Beispiel hierzu dient gleichzeitig als Illustration eines Austauschschrittes.

(7.12) Beispiel:

minimiere $z = -x_1 - 2x_2 - 3x_3$

unter

$$Ax = \begin{pmatrix} 2 & 1 & 5 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \end{pmatrix} = b \,, \quad x \geq 0.$$

Wir wählen die Basis $B = (1, 2)$ und

führen die Rechnungen in Tableau-Form an der Matrix $[b \mid A]$ durch.

b	x_1	x_2	x_3	
5	2	1	5	$\cdot 1/2$
4	1	2	1	
$\frac{5}{2}$	1	$\frac{1}{2}$	$\frac{5}{2}$	
4	1	2	1	$-$ 1. Zeile
$\frac{5}{2}$	1	$\frac{1}{2}$	$\frac{5}{2}$	
$\frac{3}{2}$	0	$\frac{3}{2}$	$-\frac{3}{2}$	$\cdot \frac{2}{3}$
$\frac{5}{2}$	1	$\frac{1}{2}$	$\frac{5}{2}$	$-\frac{1}{2} \cdot$ 2. Zeile
1	0	1	-1	
2	1	0	3	
1	0	1	-1	

$$x_B = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ -1 \end{pmatrix} x_3 = A_B^{-1} b - A_B^{-1} A_N x_N \,.$$

Die Basis-Lösung

$$x_1 = 2 \,, \quad x_2 = 1 \,, \quad x_3 = 0$$

ist zulässig, aber nicht optimal, denn die reduzierten Kosten sind

$$r_N = r_3 = c_B A_B^{-1} A_N - c_3$$
$$= (-1,-2)\begin{pmatrix}3\\-1\end{pmatrix} - (-3) = 2 > 0.$$

Für $0 \le x_3 \le 2/3$ ist $x_B = \begin{pmatrix}x_1\\x_2\end{pmatrix} \ge 0$ zulässig: hier läuft man geometrisch entlang einer Kante von K.
Für $x_3 = 2/3$ erhält man die neue zulässige Basislösung $x = (0, 5/3, 2/3)^T$. Der Übergang $x_3 = 0 \to x_3 = 2/3$ entspricht einem Basistausch

$$B = (1,2) \to B' = (3,2), \quad N = (3) \to N' = (1);$$

die Nichtbasisvariable x_3 wird dabei mit der Basisvariablen x_1 getauscht. Dies ergibt die Darstellung in der neuen Basis B'

$$x_{B'} = \begin{pmatrix}x_3\\x_2\end{pmatrix} = \begin{pmatrix}2/3\\5/3\end{pmatrix} - \begin{pmatrix}1/3\\1/3\end{pmatrix} x_1$$

mit $r_{N'} = r_1' = (-3,-2)\begin{pmatrix}1/3\\1/3\end{pmatrix} - (-1) = -2/3 < 0$.
Daher ist die gefundene Basislösung optimal. ∎

Vorüberlegung zum Basistausch:
Bei der Beschreibung eines Basistausches

$B \to B'$ treffen wir zur Vereinfachung die Annahme

$$A_B = I_m, \quad B = (1,2,\dots,m), \quad N = (m+1,\dots,n).$$

Aus (7.7), (7.8) wird dann

(7.13a) $x_B = b - A_N x_N$,

(7.13b) $z(x) = z_0 - r_N x_N$, $r_N = c_B A_N - c_N$.

Sei nun $b \ge 0$. Ist $r_N \le 0$, so ist $x_B = b$, $x_N = 0$ optimal.

Sei $r_s < 0$ für ein $s \in N$:
Setzt man $x_j = 0$ für $j \in N \setminus \{s\}$, so wird aus (7.13)

(7.14)
$$\boxed{\begin{aligned} x_B &= b - a^s x_s, \\ z(x) &= z_0 - r_s x_s, \quad z_0 = c_B b \end{aligned}}$$

1. Fall: $a^s \le 0$:

$$\left.\begin{aligned} x_B &= b - a^s x_s \ge 0 \quad \text{zulässig} \\ z(x) &= z_0 - r_s x_s \to -\infty \end{aligned}\right\} \text{für } x_s \to \infty$$

Es gibt keine endliche Lösung, die zulässige Menge K ist unbeschränkt.

2. Fall: $a_{is} > 0$ für mindestens ein $i \in \{1,\dots,m\}$.

Definiere Index $p \in B$ durch

$$\text{(7.15)} \quad \boxed{\frac{b_p}{a_{ps}} = \min\left\{\frac{b_i}{a_{is}} \;\middle|\; a_{is} > 0,\ i = 1,..,m\right\}}$$

Um $x_B \geq 0$ in (7.14) zu garantieren, kann x_s höchstens den Wert

$$x_s = \frac{b_p}{a_{ps}} =: b'_p \quad \Rightarrow \quad x_p = 0$$

annehmen. Dies bedeutet einen Tausch der Nicht-Basis-Variablen x_s gegen die Basis-Variable x_p (beachte $B = (1,..,m)$); $a_{ps} > 0$ heißt <u>Pivotelement</u>. Außerdem gilt

$$\text{(7.16)} \quad z(x) = z_0 - \tau_s \frac{b_p}{a_{ps}} \,.$$

Hieraus liest man ab:

(a) $b_p = 0$: Basis-Lösung ist entartet, $z(x) = z_0$.

(b) $b_p > 0$: Hinreichend ist eine nicht-entartete Basislösung, d.h. $x_B = b > 0$, $z(x) < z_0$: Verbesserung der Zielfunktion.

(7.17) <u>Folgerung</u>: (<u>Notwendiges Optimalitäts-kriterium</u>)

Die Basis-Lösung sei optimal und nicht entartet, d.h. $x_B > 0$. Dann gilt $\tau_N \leq 0$.

Der Index $s \in N$ (Pivot-Spalte) wird meist gemäß

$$\text{(7.18)} \quad \boxed{\tau_s = \max_{j \in N} \tau_j}$$

gewählt.

<u>Basis-Tausch</u>: (<u>Jordan-Elimination</u>)

Nach (7.13) gilt

$$\text{(7.19)} \quad x_i = b_i - \sum_{j \in N} a_{ij} x_j \,, \quad i \in B.$$

Die p-te Gleichung kann wegen $a_{ps} > 0$ nach x_s aufgelöst werden:

$$\text{(7.20)} \quad x_s = \frac{b_p}{a_{ps}} - \sum_{\substack{j \in N \\ j \neq s}} \frac{a_{pj}}{a_{ps}} x_j - \frac{1}{a_{ps}} x_p \,.$$

Man setze x_s in die übrigen Gleichungen ein für <u>$i \in B$, $i \neq p$:</u>

$$(7.21)\quad x_i = b_i - a_{is}\frac{b_p}{a_{ps}} - \sum_{\substack{j\in N\\ j\neq s}}\left[a_{ij} - a_{is}\frac{a_{pj}}{a_{ps}}\right]x_j + \frac{a_{is}}{a_{ps}}x_p .$$

Definiere neue Basis und Nichtbasis

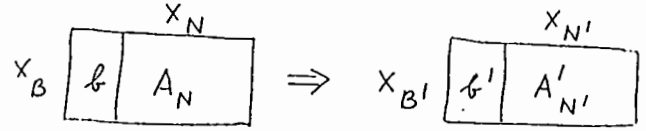
$$B' := B \cup \{s\} \setminus \{p\} = (1,\dots,p-1,s,p+1,\dots,m),$$
$$N' := N \cup \{p\} \setminus \{s\} = (m+1,\dots,s-1,p,s+1,\dots,n).$$

Bem.: Gilt allgemein $B = (i_1,\dots,i_m)$, $N = (j_1,\dots,j_{n-m})$, so ist beim Übergang $B \to B'$ der p-te Index i_p in B durch den s-ten Index j_s von N zu ersetzen.

Die Ausdrücke (7.20), (7.21) sind von der Form

$$x_{B'} = \begin{pmatrix} x_1 \\ \vdots \\ x_{p-1} \\ x_s \\ x_{p+1} \\ x_m \end{pmatrix} =: b' - A'_{N'}\, x_{N'}$$

und haben also die Gestalt (7.19).
Der Übergang der Tableaus



ist dann gegeben durch

Pivotelement: reziproker Wert:

$$a'_{ps} = \frac{1}{a_{ps}}$$

Übrige Zeile p: dividiere durch Pivotelement:

$$a'_{pj} = \frac{a_{pj}}{a_{ps}}, \quad j \neq s$$

$$b'_p = \frac{b_p}{a_{ps}}$$

(7.22)

Übrige Spalte s: dividiere durch negatives Pivotelement

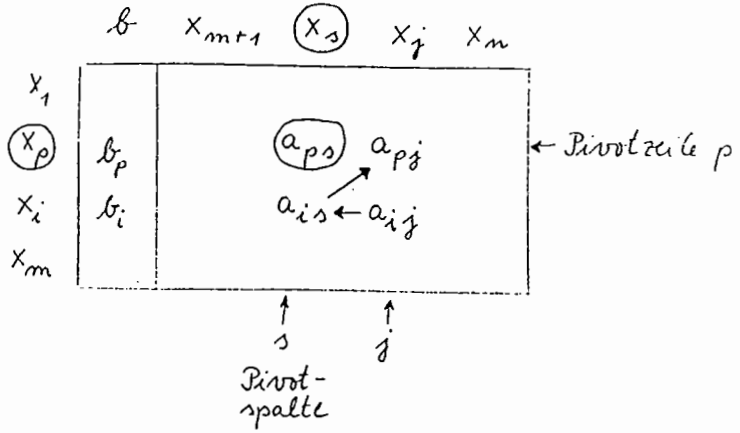
$$a'_{is} = -\frac{a_{is}}{a_{ps}}, \quad i \neq p$$

Sonstige Elemente: subtrahiere das a_{is}-fache der neuen Zeile p von i-ter Zeile

$$a'_{ij} = a_{ij} - a_{is}\frac{a_{pj}}{a_{ps}} = a_{ij} - a_{is}\,a'_{pj} \qquad (i \neq p,\ j \neq s)$$

$$b'_i = b_i - a_{is}\frac{b_p}{a_{ps}} = b_i - a_{is}\,b'_p \qquad (i \neq p)$$

Berechnung im Tableau-Form



Beispiel:

	b	x_4	x_5	x_6
x_1	5	(1)	1	-1
x_2	3	2	-3	1
x_3	-1	-1	2	-1

Tausche Nicht-Basis-Variable x_4 gegen Basis-Variable x_1:

	b'	x_1	x_5	x_6
x_4	5	1	1	-1
x_2	-7	-2	-5	3
x_3	4	1	3	-2

z.B. $b_2' = 3 - 2 \cdot 5 = -7$

$b_3' = -1 - (-1) \cdot 5 = 4$

$x_{B'} = \begin{pmatrix} x_4 \\ x_2 \\ x_3 \end{pmatrix} = b' = \begin{pmatrix} 5 \\ -7 \\ 4 \end{pmatrix}, \quad B' = (4,2,3).$ ∎

Basis-Tausch: Zielfunktion und reduzierte Kosten

Durch Einsetzen des Ausdrucks (7.20) für x_s folgt

$$z(x) = z_0 - \sum_{j \in N} \tau_j x_j$$

$$= z_0 - \sum_{\substack{j \in N \\ j \neq s}} \tau_s x_j - \tau_s \left(\frac{b_p}{a_{ps}} - \sum_{\substack{j \in N \\ j \neq s}} \frac{a_{pj}}{a_{ps}} x_j - \frac{1}{a_{ps}} x_p \right)$$

$$= z_0 - \tau_s \frac{b_p}{a_{ps}} - \sum_{\substack{j \in N \\ j \neq s}} \left(\tau_j - \tau_s \frac{a_{pj}}{a_{ps}} \right) x_j + \frac{\tau_s}{a_{ps}} x_p$$

$$=: z_0' - \sum_{j \in N'} \tau_j' x_j$$

mit

(7.23)

$$z_0' := z_0 - \tau_s \frac{b_p}{a_{ps}} = z_0 - \tau_s b_p'$$

vgl (7.16)

$$\tau_p' := -\frac{\tau_s}{a_{ps}}$$

$$\tau_j' := \tau_j - \tau_s \frac{a_{pj}}{a_{ps}} = \tau_j - \tau_s a_{pj}', \quad j \neq p$$

Die Zeile (z_0, τ_N) wird also wie die Zeilen von $\boxed{b \mid A_N}$ in (7.22) transformiert.

Damit kann die Pivotoperation des Basis-Tausches an der erweiterten Matrix durchgeführt werden:

Simplex-Tableau:

(7.24)

	x_j, $j \in N$
$z(x)$	z_0 | τ_N
x_i, $i \in B$	b | A_N

Beispiel (7.12)

maximiere $z(x) = x_1 + 2x_2 + 3x_3$

unter
$$2x_1 + x_2 + 5x_3 = 5$$
$$x_1 + 2x_2 + x_3 = 4$$
$$x_1, x_2, x_3 \geq 0 .$$

Für $B=(1,2)$, $N=(3)$ wurde das LGS überführt in ein LGS mit $A_B = I_2$:

$$\left[\begin{array}{c|c} b & A_N \end{array}\right] = \left[\begin{array}{c|c} 2 & 3 \\ 1 & -1 \end{array}\right], \quad x_B = b = \begin{pmatrix} 2 \\ 1 \end{pmatrix} > 0,$$

$c = (-1, -2, -3)$,

$z_0 = c_B b = (-1, -2)\begin{pmatrix} 2 \\ 1 \end{pmatrix} = -4$,

$\tau_3 = c_B A_N - c_3 = 2 > 0$.

Simplex-Tableau

	x_3
$z(x)$	-4 | 2
x_1	2 | $\textcircled{3}$
x_2	1 | -1

$p=1$, $s=3$, $a_{ps} = 3 > 0$

Tausche x_3 gegen x_1, $B'=(3,2)$, $N'=(1)$:

	x_1
$z(x)$	$-16/3$ | $-2/3$
x_3	$2/3$ | $1/3$
x_2	$5/3$ | $1/3$

Es ist $\tau_1 = -2/3 < 0$, also ist die Basis-Lösung

$x_1 = 0$, $x_3 = 2/3$, $x_2 = 5/3$ optimal. ∎

Anwendung auf ein LP mit Ungleichungen:

Das in (7.11) behandelte LP

$$\min \{ cx \mid Ax + y = b,\ x \geq 0,\ y \geq 0 \}$$

führt mit

$B = (n+1, .., n+m)$, $N = (1, .., n)$

$x = 0$, $y = b \geq 0$, $\tau_N = -c$, $\widetilde{A}_N = A$

auf das Simplex-Tableau

(7.25)

		x
z	0	$-c$
y	b	A

Als Beispiel lösen wir das Problem aus Abschnitt 7.1:

maximiere $z = 16x_1 + 32x_2$

unter

$$20x_1 + 10x_2 + y_1 = 8000$$
$$4x_1 + 5x_2 + y_2 = 2000$$
$$6x_1 + 15x_2 + y_3 = 4500$$

$$x_1, x_2, y_1, y_2, y_3 \geq 0.$$

Der geometrische Übergang der Ecken $A \to B \to C \to D$ wird rechnerisch in Tableau-Form nachvollzogen.

1. Tableau:

		x_1	x_2
	z	0	16	32
→	y_1	8000	(20)	10
	y_2	2000	4	5
	y_3	4500	6	15

↑

Entgegen der Vorschrift (7.18) wählen wir $s = 1$. Dann ist $p = 1$ und wir tauschen

x_1 mit y_1: Übergang $A \to B$.

2. Tableau

		y_1	x_2
	z	-6400	$-\frac{4}{5}$	24
	x_1	400	$\frac{1}{20}$	$\frac{1}{2}$
→	y_2	400	$-\frac{1}{5}$	(3)
	y_3	2100	$-\frac{3}{10}$	12

↑

Tausche x_2 mit y_2: Übergang $B \to C$.

3. Tableau

		y_1	y_2
	z	-9600	$\frac{4}{5}$	-8
	x_1	$\frac{1000}{3}$	$\frac{1}{12}$	$-\frac{1}{6}$
	x_2	$\frac{400}{3}$	$-\frac{1}{15}$	$\frac{1}{3}$
→	y_3	500	$(\frac{1}{2})$	-4

↑

Tausche y_1 mit y_3: Übergang $C \to D$

4. Tableau (optimal):

		y_3	y_2
z	-10400	$-\frac{8}{5}$	$-\frac{8}{5}$
x_1	250	$-\frac{1}{6}$	$\frac{1}{2}$
x_2	200	$\frac{2}{15}$	$-\frac{1}{5}$
y_1	1000	2	-8

Hier ist $r_N = \left(-\frac{8}{5}, -\frac{8}{5}\right) < 0$, also ist

$x_1 = 250$, $x_2 = 200$, $z_{min} = -10400$, d.h.

$z_{max} = 10400$, die optimale Lösung.

Kapitel III

Iterationsverfahren zur Lösung von Gleichungen

§8 Definitionen und Grundbegriffe

8.1 Nullstellen

Sei $D \subset \mathbb{R}^n$ und $f: D \to \mathbb{R}^n$, $f = (f_1, \dots, f_n)^T$.
Zu lösen sei die Gleichung

$$f(x) = 0, \quad x \in D, \tag{8.1}$$

d.h. komponentenweise

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \dots, x_n) &= 0. \end{aligned}$$

Ein Punkt $\bar{x} \in D$ heißt Nullstelle von f, wenn

$$f(\bar{x}) = 0.$$

Kriterien für die lokale Existenz und Eindeutigkeit von Nullstellen: §10.

Beispiele:

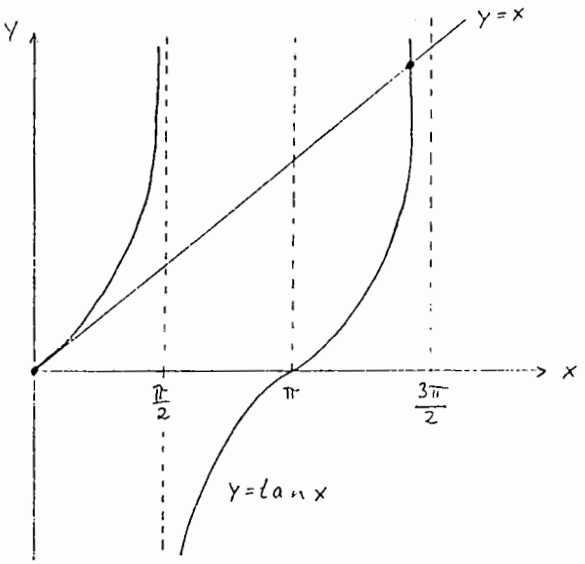
(1) Polynome: $D = \mathbb{R}$.

$$f(x) = a_0 + a_1 x + \dots + a_n x^n = 0, \quad x \in \mathbb{R}.$$

Z.B. sind Eigenwerte von Matrizen Nullstellen des charakteristischen Polynoms.

(2) $$f(x) = x - \tan x = 0.$$

Dieses Problem tritt bei der Berechnung der Schwingungen eines Balkens auf.



$\bar{x}_0 = 0$. In den Intervallen $(k\pi, (k+\frac{1}{2})\pi)$ liegen Nullstellen $\bar{x}_k$, $k = 1, 2, \dots$. Mit $\bar{x}_k$ ist auch $-\bar{x}_k$ Nullstelle.

(3) Optimierungsprobleme

Sei $h: \mathbb{R}^m \to \mathbb{R}$ differenzierbar. Jede Lösung $\bar{x}$ des Minimierungsproblems

$$\min_{x \in \mathbb{R}^m} h(x)$$

ist Nullstelle von

$$f(x) := \nabla h(x) = \left(\frac{\partial h}{\partial x_1}(x), \ldots, \frac{\partial h}{\partial x_m}(x)\right)^T = 0.$$

Minimierungsprobleme mit Nebenbedingungen haben die Gestalt:

minimiere $h(x)$

unter allen $x \in \mathbb{R}^m$ mit

$h_i(x) \leq 0$ für $i = 1, 2, \ldots, r$,

$h_i(x) = 0$ für $i = r+1, \ldots, m$,

mit C^1-Funktionen $h_i: \mathbb{R}^m \to \mathbb{R}$, $i = 1, \ldots, m$.

(4) Eine Gleichgewichtslösung des folgenden chemischen Systems wird gesucht (Gewinnung von Wasserstoff aus Methan):

$$CH_4 + \tfrac{1}{2} O_2 \rightleftharpoons CO + 2H_2$$

$$CH_4 + H_2O \rightleftharpoons CO + 3H_2$$

$$H_2 + CO_2 \rightleftharpoons CO + H_2O$$

Dies führt zu folgendem Gleichungssystem für die Konzentrationen $x \in \mathbb{R}^7$:

$$f_1(x) = \frac{1}{2} x_1 + x_2 + \frac{1}{2} x_3 - \frac{x_6}{x_7} = 0$$

$$f_2(x) = x_3 + x_4 + 2x_5 - \frac{2}{x_7} = 0$$

$$f_3(x) = x_1 + x_2 + x_5 - \frac{1}{x_7} = 0$$

$$f_4(x) = -28837\, x_1 - 139009\, x_2 - 78213\, x_3 + 18927\, x_4 + 8427\, x_5 + \frac{13492}{x_7} - 10690\, \frac{x_6}{x_7} = 0$$

$$f_5(x) = x_1 + x_2 + x_3 + x_4 + x_5 - 1 = 0$$

$$f_6(x) = 400\, x_1 x_4^3 - 1.7837 \times 10^5\, x_3 x_5 = 0$$

$$f_7(x) = x_1 x_3 - 2.6058\, x_2 x_4 = 0,$$

Literatur: Carnahan, Luther, Wilkes: Applied Numerical Methods, New York, Wiley, 1969, S. 321

8.2 Fixpunkte

Sei $D \subset \mathbb{R}^n$ und $g: D \to \mathbb{R}^n$. Gesucht sind Lösungen $x \in D$ der Gleichung

$$x = g(x). \tag{8.2}$$

Ein Punkt $\bar{x} \in D$ heißt Fixpunkt von g, falls

$$\bar{x} = g(\bar{x}).$$

Sei $A(x)$ eine reguläre (n,n)-Matrix, $x \in D$, und $f: D \to \mathbb{R}^n$. Dann ist die Nullstellenbestimmung

$$f(x) = 0$$

äquivalent zur Fixpunktgleichung

$$x = g(x) := x + A(x) f(x). \tag{8.3}$$

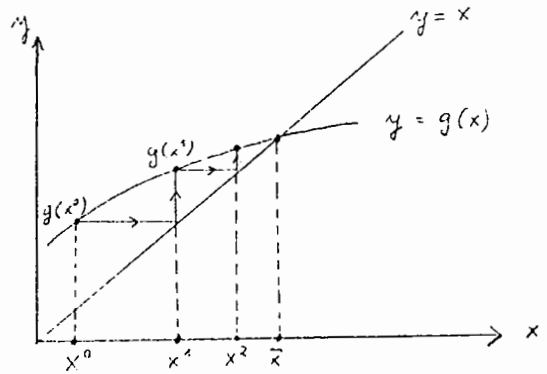
Fixpunkte von (8.2) werden mit Iterationsverfahren der folgenden Form bestimmt:

(8.4)
Startwerte: $x^0, x^1, .., x^s$ gegeben, $s \geq 0$,
Iteration: $x^{k+1} = \phi(x^k, x^{k-1}, .., x^{k-s})$, $k \geq s$

ϕ heißt Iterationsfunktion und hängt von g ab. Oft kann $\phi = g$ gewählt werden, sodaß die Iteration (8.4) lautet:

(8.5)
$x^{k+1} = g(x^k)$, $k = 0, 1, ..$
$x^0 \in D$ gegeben.

Graphische Darstellung für $g: \mathbb{R} \to \mathbb{R}$



$(|g'(x)| < 1)$

Im Zusammenhang mit der Iteration (8.4) stellen sich folgende Fragen:

a) Wie findet man passende Iterationsfunktionen ϕ?

b) Wie findet man passende Anfangspunkte $x^0, .., x^s$?

c) Unter welchen Bedingungen konvergiert die Folge $\{x^k\}$ gegen einen Fixpunkt von g

d) Wie schnell konvergiert die Folge $\{x^k\}$?

Diese Fragen werden in §9 und §10 behandelt.

8.3 Konvergenzgeschwindigkeit

Sei $\|\cdot\|$ irgendeine Norm für $\mathbb{R}^n$ und sei $\{x^k\} \subset \mathbb{R}^n$ eine Folge mit

$$\bar{x} = \lim_{k\to\infty} x^k .$$

Falls $p \geq 1$ existiert mit

$$(8.6)\qquad c := \lim_{k\to\infty} \sup \frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|^p} < \begin{cases} 1 & \text{für } p = 1 \\ \infty & \text{für } p > 1 \end{cases},$$

so heißt $\{x^k\}$ konvergent vom Grade p.
p heißt Konvergenzgrad. Ist $p \geq 1$ die größte Zahl mit (8.6), so heißt p der genaue Konvergenzgrad.
$p = 1$: lineare Konvergenz,
$p = 2$: quadratische Konvergenz

Interpretation von (8.6): Es gibt

$$0 \leq c < \begin{cases} 1 & \text{für } p = 1 \\ \infty & \text{für } p > 1 \end{cases}$$

und $k_0 \in \mathbb{N}$ mit

$$\|x^{k+1} - \bar{x}\| \leq c \|x^k - \bar{x}\|^p \quad \text{für } k \geq k_0$$

Die Konstante c heißt asymptotische Fehlerkonstante. Für den Fehler

$$e_k := \|x^k - \bar{x}\|$$

gilt also

$$e_{k+1} \leq c\, e_k^p ,$$

d.h.

$$\boxed{e_{k+1} = O(e_k^p)} .$$

Im Falle $p=1$ (lineare Konvergenz) erhält man aus $e_{k+1} \le c\, e_k$ die Abschätzung

$$e_{k+m} \le c^m e_k \,, \quad m > 0 \,.$$

Zur Vermeidung von Mißverständnissen werden reelle Folgen $\{x_k\}$ mit einem unteren Index k bezeichnet.

Beispiel: Für $0 < q < 1$ sei x_k der Abschnitt der geometrischen Reihe

$$x_k = \sum_{i=0}^{k} q^i = \frac{1-q^{k+1}}{1-q} \,,$$

$$\bar{x} = \lim_{k\to\infty} x_k = \frac{1}{1-q} \,,$$

$$e_k = \frac{1-q^{k+1}}{1-q} - \frac{1}{1-q} = \frac{-q^{k+1}}{1-q}$$

$$\frac{e_{k+1}}{e_k} = \frac{-q^{k+2}}{-q^{k+1}} = q$$

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k} = q$$

$\Rightarrow$ $p=1$: lineare Konvergenz.

Sei nun $\bar{x}$ ein Fixpunkt von g. Das Iterationsverfahren (8.4) heißt ein

Verfahren von (mindestens) p-ter Ordnung

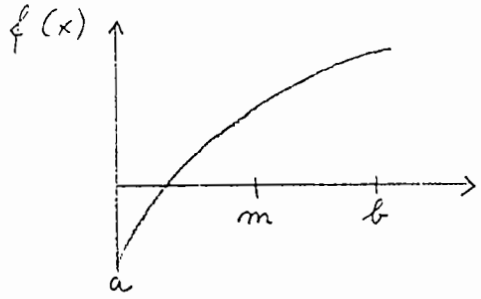
zur Bestimmung von $\bar{x}$, wenn es eine Umgebung $U(\bar{x})$ gibt, so daß für alle Startwerte $x^0, .., x^s \in U(\bar{x})$ die durch (8.4) definierte Folge $\{x^k\}$ konvergent vom Grade p gegen $\bar{x}$ ist.

§9 Nullstellen reeller Funktionen.

Sei $f: D \to \mathbb{R}$ mit $D \subset \mathbb{R}$ gegeben. Iterationsfolgen $\{x_k\}$ zur Bestimmung einer Nullstelle $\bar{x}$ von f haben zur Vermeidung von Mißverständnissen einen unteren Index k.

9.1. Intervallhalbierung

Sei $f: [a, b] \to \mathbb{R}$ stetig mit $f(a) \cdot f(b) < 0$. Dann hat f mindestens eine Nullstelle im Intervall $[a, b]$.



Die Funktion f wird im Mittelpunkt $m = \frac{1}{2}(a+b)$ ausgewertet; dann wählt man das Teilintervall, in dem eine Nullstelle liegen muß.

Das Verfahren ist nur für sehr einfache Probleme geeignet. Für die Folge der iterativ bestimmten Mittelpunkte x_k

gilt offenbar:

$$|x_k - \bar{x}| \leq (b-a)/2^{k+1},$$

d.h. $\{x_k\}$ konvergiert <u>linear</u> gegen $\bar{x}$.

<u>Beispiel</u>: $f(x) = x - \tan x$

$a = 2, \qquad b = 4.6$

$f(a) \doteq 4.18, \qquad f(b) \doteq -4.26$

Damit
$x_0 = 3{,}3$
$x_1 = 3{,}95$
$x_2 = 4{,}2375$
$x_3 = 4{,}437500$
$x_4 = 4{,}5187750$

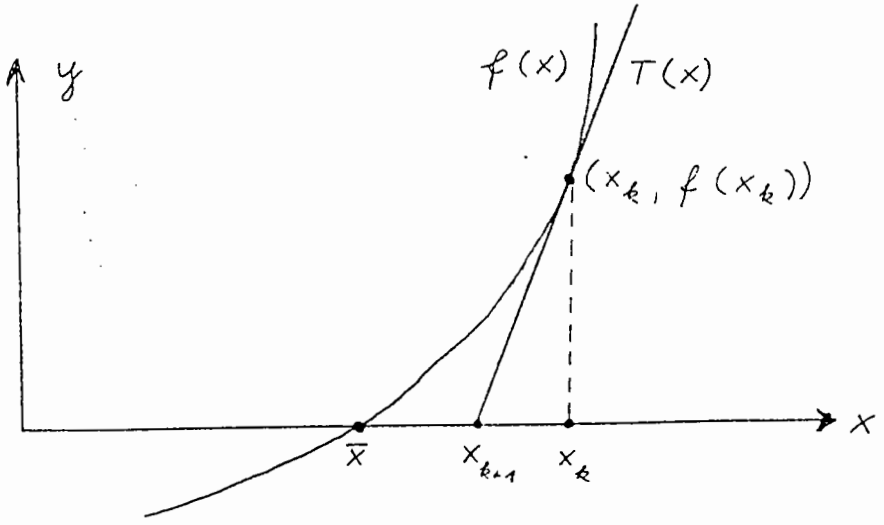
$x_5 = 4.478125, \quad f(x_5) = 2.87 \times 10^{-1}$

$x_{20} = 4.493410, \quad f(x_{20}) = -1.51 \times 10^{-5}$

$\bar{x} = x_{100} = 4.49340946, \quad f(x_{100}) = -1.72294616 \times 10^{-10}.$

9.2. <u>Das Newton-Verfahren</u>

Geometrische Interpretation:



Sei x_k eine Näherung für $\bar{x}$. Im Punkt $(x_k, f(x_k))$ wird die Tangente

$$T(x) = f(x_k) + f'(x_k)(x - x_k)$$

an die Kurve $y = f(x)$ konstruiert. Der Schnittpunkt x_{k+1} von $T(x)$ mit der x-Achse ist dann, falls $f'(x_k) \neq 0$:

(9.1) $$\boxed{x_{k+1} = x_k - f(x_k) / f'(x_k)}$$

Das Newton-Verfahren ist also eine Fixpunktiteration (8.5)

$$x_{k+1} = g(x_k), \quad g(x) := x - f(x)/f'(x),$$

$x_0 \in D$ gegeben.

Beispiele:

(1) $f(x) = x^2 - 2$ (d.h. Berechnung von $\bar{x} = \sqrt{2}$),

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

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{2}{x_k}\right).$$

k	x_k	Anzahl der korrekten Dezimalen (erste falsche Dez. unterstrichen)
0	1	1
1	1.5	1
2	1.417	3
3	1.414216	6
4	1.414213562	10

(quadratische Konvergenz)

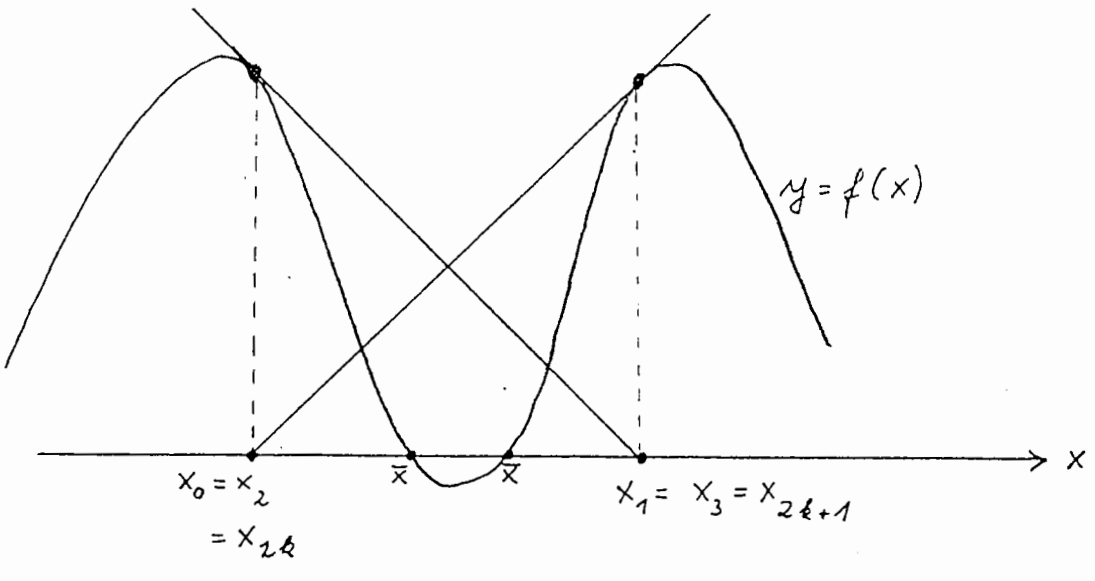
(2) $f(x) = x - \tan x$

Für $x_0 = 2$ bzw. $x_0 = 4$ divergiert $\{x_k\}$.

Für $x_0 = 4{,}5$ erhält man

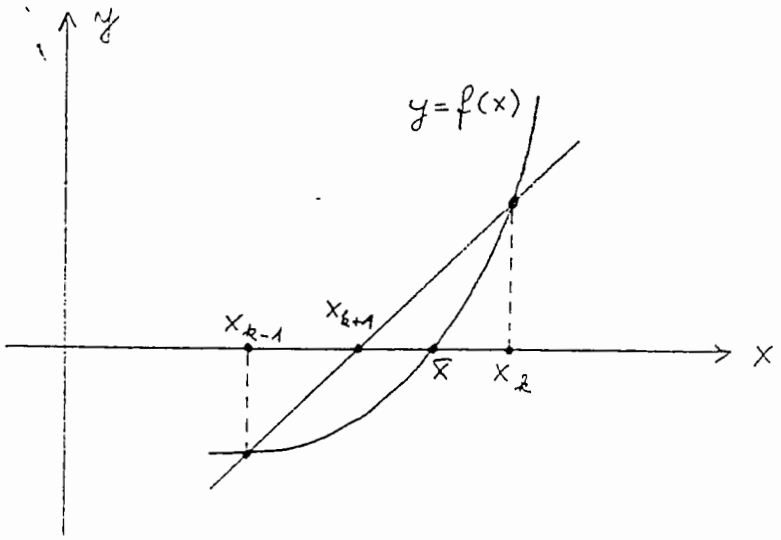
$$\overline{x} = x_3 = 4.493409 45.$$

Die lokale Konvergenz des Newton-Verfahrens wird in §10 diskutiert. Es ist unmittelbar einzusehen, daß das Newton-Verfahren nicht immer konvergiert; z.B.



9.3. Das Sekanten-Verfahren (Regula falsi)

Die Startwerte x_0, x_1 seien gegeben. Die Sekantenmethode ist eine Vereinfachung des Newton-Verfahrens, wobei die Tangente in $(x_k, f(x_k))$ durch die Sekante zwischen $(x_k, f(x_k))$ und $(x_{k-1}, f(x_{k-1}))$ ersetzt wird.



Die Nullstelle x_{k+1} der Sekante ergibt ein Iterationsverfahren der Form (8.4) mit $s=1$:

(9.2)

x_0, x_1 gegeben

$$x_{k+1} = \phi(x_k, x_{k-1}) = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}.$$

Beispiel: $f(x) = x - \tan(x)$

k	x_k	x_{k-1}	$f(x_k)$
1	4.0	4.6	-4.26×10^{0}
5	4.6587	4.3957	1.34×10^{0}
12	4.49340824	4.49340945	1.07×10^{-8}
13	Division durch Null		

(9.3) Satz: Sei $f: \mathbb{R} \to \mathbb{R}$ in einer Umgebung von $\bar{x}$ zweimal stetig differenzierbar und sei $f(\bar{x}) = 0$, $f'(\bar{x}) \neq 0$. Dann konvergiert das Sekantenverfahren gegen $\bar{x}$, falls x_0, x_1 hinreichend nahe bei $\bar{x}$ gewählt wurden. Der Konvergenzgrad ist

$$p = \frac{1}{2}(1 + \sqrt{5}) = 1.618\ldots .$$

Beweis: Schwarz: Satz 5.7., S. 201.

§10 Konvergenzsätze für Iterationsverfahren

Sei $D \subset \mathbb{R}^n$ und $g: D \to \mathbb{R}^n$. Wir untersuchen die Frage, wann die Fixpunktiteration

(10.1) $$x^{k+1} = g(x^k), \quad k \geq 0, \quad x^0 \in D \text{ gegeben},$$

wohldefiniert ist und gegen einen Fixpunkt $\bar{x} \in D$ konvergiert.

(10.2) <u>Definition:</u> $g: D \to \mathbb{R}^n$ heißt <u>kontrahierend in D</u>, falls es eine Zahl $0 \leq q < 1$ gibt mit

$$\|g(x) - g(y)\| \leq q \|x - y\| \quad \text{für alle } x, y \in D.$$

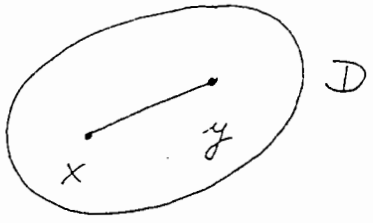
Für differenzierbare Abbildungen g kann ein einfaches Kriterium für die Kontraktion mit der Ableitung

$$g'(x) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix}$$

gegeben werden. D heißt <u>konvex</u>, wenn

für $x, y \in D$

$$\alpha x + (1-\alpha) y \in D \quad \text{für alle} \quad \alpha \in [0,1].$$



(10.3) **Satz:** Sei D konvex, $g: D \to \mathbb{R}^n$ differenzierbar und sei

$$\sup_{x \in D} \| g'(x) \|_\infty \leq q < 1.$$

Dann ist g kontrahierend in D.

Beweis: Für zwei beliebige Punkte $x, y \in D$ betrachten wir $\varphi: [0,1] \to \mathbb{R}^n$:

$$\varphi(\lambda) := g(\lambda x + (1-\lambda) y), \quad \lambda \in [0,1],$$

$$\varphi(1) = g(x), \quad \varphi(0) = g(y),$$

$$\varphi'(\lambda) = g'(\lambda x + (1-\lambda) y)(x-y).$$

Aus dem Mittelwertsatz folgt:

$$|\varphi_i(1) - \varphi_i(0)| \leq \max_{0 \leq \lambda \leq 1} |\varphi_i'(\lambda)|, \quad i = 1, \ldots, n.$$

$$\Rightarrow \|g(x)-g(y)\|_\infty = \|\varphi(1)-\varphi(0)\|_\infty$$

$$\leq \max_{0\leq\lambda\leq 1} \|\varphi'(\lambda)\|_\infty$$

$$= \max_{0\leq\lambda\leq 1} \|g'(\lambda x+(1-\lambda)y)(x-y)\|$$

$$\leq \sup_{z\in D} \|g'(z)\|_\infty \, \|x-y\|_\infty . \quad \blacksquare$$

Für $\underline{n=1}$ ist $D=[a,b]$ konvex und $g \in C^1[a,b]$ kontrahierend, falls

$$\max_{a\leq x\leq b} |g'(x)| = q < 1 .$$

(Vgl. Graphik hinter (8.5)).

(10.4) <u>Satz (Fixpunktsatz von Banach)</u>:

Sei D abgeschlossen und $g: D \to \mathbb{R}^n$ kontrahierend in D mit $g(D) \subseteq D$. Dann konvergiert die Folge

$$x^{k+1} = g(x^k), \quad k = 0,1,2,\ldots, \quad x^0 \in D \text{ beliebig},$$

gegen den eindeutig bestimmten Fixpunkt $\bar{x}$ von g in D und es gilt:

(i) $\|\bar{x}-x^k\| \leq \frac{q}{1-q}\|x^k-x^{k-1}\| \leq \frac{q^k}{1-q}\|x^1-x^0\|$,

(ii) $\|\bar{x}-x^k\| \leq q\,\|\bar{x}-x^{k-1}\|$.

Beweis: Wir zeigen zunächst, daß $\{x^k\}$ eine Cauchy-Folge ist. Für $k \geq 1$ gilt

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

$$\leq q^2 \| x^{k-1} - x^{k-2} \| \leq q^k \| x^1 - x^0 \|$$

und damit für $j > l$

$$\| x^j - x^l \| = \| \sum_{k=l}^{j-1} (x^{k+1} - x^k) \| \leq \sum_{k=l}^{j-1} \| x^{k+1} - x^k \|$$

(10.5)

$$\leq \sum_{k=l}^{j-1} q^k \| x^1 - x^0 \|$$

$$\leq q^l \frac{1}{1-q} \| x^1 - x^0 \| \xrightarrow[l \to \infty]{} 0 .$$

Die x^k bilden damit eine Cauchy-Folge. Sei $\bar{x} \in D$ der Grenzwert der Folge x^k. Dann gilt

$$g(\bar{x}) \xleftarrow[k \to \infty]{} g(x^k) = x^{k+1} \xrightarrow[k \to \infty]{} \bar{x} ,$$

$$\Rightarrow \quad g(\bar{x}) = \bar{x} .$$

In (10.5) ergibt der Grenzwert für $j \to \infty$:

$$\| \bar{x} - x^l \| \leq \frac{q^l}{1-q} \| x^1 - x^0 \| .$$

Mit $\ell = 1$ folgt hieraus nach Ersetzen $x^0 \to x^{k-1}$

$$\| \bar{x} - x^k \| \leq \frac{q}{1-q} \| x^k - x^{k-1} \| .$$

Damit ist (i) gezeigt; (ii) folgt aus der Kontraktionseigenschaft:

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

Zu zeigen bleibt noch die Eindeutigkeit von $\bar{x}$.

Seien $\bar{x}$, x^* Fixpunkte von g:

$$\| \bar{x} - x^* \| = \| g(\bar{x}) - g(x^*) \| \leq q \| \bar{x} - x^* \| , \quad q < 1,$$

$$\Rightarrow \| \bar{x} - x^* \| = 0 . \quad \blacksquare$$

Bem.:

(1) Wegen Teil (ii) konvergiert $\{x^k\}$ linear gegen $\bar{x}$, vgl. § 8.3.

(2) Die Schwierigkeiten bei der Anwendung des Kontraktionssatzes auf ein konkretes Problem bestehen darin:

(a) man finde eine zugehörige kontrahierende Funktion $g : D \to \mathbb{R}^n$,

(b) man prüfe $g(D) \subseteq D$.

Beispiele:

① Gesucht ist die Lösung $\bar{x}$ der Gleichung

$$x = e^{-x} =: g(x), \quad x \in \mathbb{R}.$$

Auf das Intervall $D = [0.5, 0.69]$ trifft die Voraussetzung $g(D) \subset D$ zu. Als Kontraktionszahl q dient nach (10.3) die Zahl

$$\max_{x \in D} |g'(x)| = e^{-0.5} = 0.606531 < 1.$$

Zum Startwert $x^{(0)} = 0.55 \in D$ berechnet man die Iterierten:

k	$x^{(k)}$	k	$x^{(k)}$	k	$x^{(k)}$
0	0.55000000	10	0.56708394	20	0.56714309
1	0.57694981	11	0.56717695	21	0.56714340
2	0.56160877	12	0.56712420	22	0.56714323
3	0.57029086	13	0.56715412	23	0.56714332
4	0.56536097	14	0.56713715	24	0.56714327

Mit der apriori Fehlerabschätzung
$\|\bar{x} - x^{(k)}\| \le \frac{q^k}{1-q} \|x^{(1)} - x^{(0)}\|$ kann die Anzahl k der Iteration geschätzt werden, die nötig

sind, damit z.B. $\|\bar{x} - x^{(k)}\| \le \varepsilon = 10^{-6}$ gilt. Man erhält

$$k \ge \log\left(\frac{\varepsilon(1-q)}{\|x^1 - x^0\|}\right) \Big/ \log q = 22.3,$$

eine gegenüber der Tabelle leichte Überschätzung. Für den Wert $x^{(12)}$ erhält man die aposteriori Fehlerschranke

$$\|\bar{x} - x^{(12)}\| \le \frac{q}{1-q} \|x^{(12)} - x^{(11)}\| = 8.3 \cdot 10^{-5}.$$

(2) $f(x) = x - \tan x$

Die Nullstelle $\bar{x}$ wird in $D = [\pi, \frac{3}{2}\pi]$ gesucht.

Die Funktion $g(x) = \tan x$ ist <u>nicht</u> kontrahierend wegen

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

Umformulierung:

$$x \neq \tan x = \tan(x - \pi) \iff \arctan x = x - \pi .$$

Setze nun

$$g(x) = \pi + \arctan x , \quad D = [\pi, \tfrac{3}{2}\pi] .$$

Offenbar gilt $g(D) \subseteq D$ und

$$q := \max_{x \in D} |g'(x)| = \frac{1}{1+\pi^2} \approx 0.092 < 1 .$$

g ist also kontrahierend in D nach (10.3).

Für $\bar{x} \doteq 4.4934094$ ist $g'(\bar{x}) \doteq 0.04719$.

k	x^k	$\frac{q}{1-q}\lvert x^k - x^{k-1}\rvert$	$\lvert \bar{x} - x^k \rvert$	$\frac{\lvert \bar{x} - x^k \rvert}{\lvert \bar{x} - x^{k-1} \rvert}$
0	3.14159265	–	–	–
1	4.4042 1991	0.1351	0.0892	–
2	4.489 11945	0.008918	0.0043	0.0672
3	4.49320683	0.0004280	0.0002	0.0481
4	4.4933999	—	—	0.0472

(10.6) **Satz: (Lokaler Konvergenzsatz)**

Sei $g: \mathbb{R}^n \longrightarrow \mathbb{R}^n$ mit $g(\bar{x}) = \bar{x}$.
Ist g in einer Umgebung von $\bar{x}$ stetig differenzierbar und $\|g'(\bar{x})\|_\infty < 1$, dann gibt es eine Umgebung D von $\bar{x}$, so daß das Iterationsverfahren

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

gegen $\bar{x}$ konvergiert.

Beweis: Sei D eine Kugel mit Radius r um $\bar{x}$ mit $\|g'(x)\|_\infty \leq q < 1$ für $x \in D$.
Für $x \in D$ gilt

$$\|g(x) - \bar{x}\|_\infty = \|g(x) - g(\bar{x})\|_\infty \leq q\|x - \bar{x}\|_\infty \leq r$$

$$\Rightarrow \quad g(x) \in D.$$

Damit ist g kontrahierend in D und es gilt $g(D) \subseteq D$. Mit dem Fixpunktsatz (10.4) folgt die Behauptung. ∎

Als Anwendung erhält man im Falle $n=1$ einfache Kriterien dafür, daß die Fixpunkt-Iteration ein Verfahren p-ter Ordnung ist.

(10.7) Satz: Sei $g: \mathbb{R} \to \mathbb{R}$ eine C^p-Funktion mit $p \in \mathbb{N}_+$. Sei $\bar{x}$ ein Fixpunkt von g mit

(a) $|g'(\bar{x})| < 1$ für $p=1$,

(b) $g^{(i)}(\bar{x}) = 0$ $(i=1,..,p-1)$ für $p>1$.

Dann gibt es ein Intervall

$I = [\bar{x}-\delta, \bar{x}+\delta]$, $\delta > 0$,

sodaß für alle $x_0 \in I$ die Iteration $x_{k+1} = g(x_k)$, $k=0,..,1$, konvergent vom Grade p ist mit

$$\lim_{k\to\infty} \frac{|x_{k+1}-\bar{x}|}{|x_k-\bar{x}|^p} = \frac{1}{p!} g^{(p)}(\bar{x}).$$

Beweis: Aus den Vor. (a), (b) folgt insbesondere $|g'(\bar{x})| < 1$. Der lokale Konvergenzsatz (10.6) sichert dann die (mindestens) lineare Konvergenz der Folge $x_{k+1} = g(x_k)$ für alle $x_0 \in I = [\bar{x}-\delta, \bar{x}+\delta]$, $\delta > 0$ geeignet.

Die Taylor-Entwicklung ergibt mit Vor. (b) und $\bar{x} = g(\bar{x})$:

$$x_{k+1} = \bar{x} + \frac{1}{p!} g^{(p)}(\bar{x})(x_k - \bar{x})^p + o(|x_k - \bar{x}|^p).$$

Hieraus folgt die Beh.

$$\lim_{k \to \infty, |x_k - \bar{x}|^p} \frac{\| x_{k+1} - \bar{x} \|}{} = \frac{1}{p!} g^{(p)}(\bar{x}). \qquad \blacksquare$$

Als Anwendung betrachten wir das Newton-Verfahren (9.1)

$$x_{k+1} = g(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Ist f eine C^3-Funktion, so ist g eine C^2-Funktion.

1. Fall: $\bar{x}$ ist einfache Nullstelle von f, d.h. $f'(\bar{x}) \neq 0$: man berechnet

$$g'(x) = \frac{f(x) f''(x)}{f'(x)^2}, \quad g'(\bar{x}) = 0,$$

$$g''(\bar{x}) = \frac{f''(\bar{x})}{f'(\bar{x})}.$$

Also ist das Newton-Verfahren (mindestens)

quadratisch konvergent mit der asymptotischen Fehlerkonstanten

$$c = \frac{1}{2} \frac{f''(\bar{x})}{f'(\bar{x})}.$$

2. Fall: $\bar{x}$ sei m-fache Nullstelle von f, d.h. $f^{(i)}(\bar{x}) = 0$ für $i = 0, .., m-1$:

$$f(x) = (x - \bar{x})^m f_0(x), \quad f_0(\bar{x}) \neq 0$$

$$\Rightarrow g'(\bar{x}) = 1 - \frac{1}{m}.$$

Für $m > 1$ ist daher $g'(\bar{x}) \neq 0$ und das Newton-Verfahren ist nur linear konvergent. Für das modifizierte Newton-Verfahren

$$x_{k+1} = g(x_k) := x_k - m \frac{f(x_k)}{f'(x_k)}$$

gilt jedoch $g'(\bar{x}) = 0$, also hat man quadratische Konvergenz.

§ 11 Das Newton-Verfahren im $\mathbb{R}^n$

Gegeben sei eine C^1-Funktion $f: D \to \mathbb{R}^n$. Gesucht ist eine Nullstelle $\bar{x} \in D$ von f. Das Newton-Verfahren zur Berechnung von $\bar{x}$ ist die folgende Fixpunktiteration:

(11.1)
$$\boxed{\begin{aligned} x^{k+1} &= x^k - \left(f'(x^k)\right)^{-1} f(x^k), \quad k \geq 0, \\ x^0 &\in D \text{ gegeben.} \end{aligned}}$$

Das Newton-Verfahren läßt sich auf verschiedene Weise erklären:

(1) Verallgemeinerung von § 9.2 mittels Taylor-Entwicklung. Es gilt

$$0 = f(\bar{x}) = f(x^k) + f'(x^k)(\bar{x} - x^k) + o(\|\bar{x} - x^k\|).$$

Vernachlässigt man $o(\|\bar{x} - x^k\|)$ und ersetzt den unbekannten Punkt $\bar{x}$ durch x^{k+1}, so erhält man

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

und daraus (11.1).

(2) Anwendung des lokalen Konvergenzsatzes (10.6

$f(\bar{x}) = 0$ gilt genau dann, wenn $\bar{x}$ Fix-punkt von

$$g(x) := x + A(x) f(x)$$

ist mit einer geeignet zu wählenden regulären (n,n) C^1-Matrix $A(x)$. Nach Satz (10.6) ist g kontrahierend, falls $\|g'(\bar{x})\|_\infty < 1$ ist. Wegen $f(\bar{x}) = 0$ gilt

$$g'(\bar{x}) = I + A(\bar{x}) f'(\bar{x}).$$

Wählen wir nun

$$A(\bar{x}) = -(f'(\bar{x}))^{-1}$$

so ist $g'(\bar{x}) = 0$. Da $\bar{x}$ unbekannt ist, setzen wir

$$A(x) = -(f'(x))^{-1},$$

d.h.

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

Die Fixpunktiteration $x^{k+1} = g(x^k)$ ergibt gerade (11.1). Satz (10.6) sichert wegen $g'(\bar{x}) = 0$ die lokale Konvergenz.

Bem.: Praktisch benutzt man in höheren Dimensionen das Newton-Verfahren in der Form

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

So muß anstatt der Invertierung von $f'(x^k)$ (n^3 Operationen) nur noch ein LGS gelöst werden ($\frac{1}{3}n^3$ Operationen).

<u>Beispiel:</u> $f: \mathbb{R}^2 \longrightarrow \mathbb{R}^2$,

$$f(x) = \begin{pmatrix} 10^4 x_1 x_2 - 1 \\ e^{-x_1} + e^{-x_2} - 1.0001 \end{pmatrix},$$

$$f'(x) = \begin{pmatrix} 10^4 x_2 & 10^4 x_1 \\ -e^{-x_1} & -e^{-x_2} \end{pmatrix}.$$

$$x^0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad f(x^0) = \begin{pmatrix} -1 \\ 0{,}36 \end{pmatrix},$$

$$\bar{x} = x^{13} = \begin{pmatrix} 1.0981595 \times 10^{-5} \\ 9.10614 \end{pmatrix}.$$

Die Matrix

$$f'(\bar{x}) \doteq \begin{pmatrix} 9.1 \times 10^4 & 0.11 \\ -1 & -1.1 \times 10^{-4} \end{pmatrix}$$

hat die Kondition

$$\| f'(\bar{x}) \|_\infty \cdot \| (f'(\bar{x}))^{-1} \|_\infty = O(10^9).$$

Bei der Berechnung von $f_2(x)$ entsteht Auslöschung; $f_2(x)$ läßt sich in folgender Gestalt besser berechnen:

$$e^{-x_1} + e^{-x_2} - 1.0001 = (e^{-x_1} - 1) + (e^{-x_2} - 10^{-4})$$
$$\doteq (-x_1 + (x_1)^2/2) + (e^{-x_2} - 10^{-4}).$$

Zum Nachweis der lokalen quadratischen Konvergenz des Newton-Verfahrens (11.1) benötigen wir den folgenden Hilfssatz.

(11.2) Hilfssatz: Sei $D_0 \subset D$ konvex. Es gebe $\gamma > 0$ mit

$$\| f'(x) - f'(y) \| \leq \gamma \| x - y \| \quad \text{für } x, y \in D_0.$$

Dann gilt

$$\| f(x) - f(y) - f'(y)(x-y) \| \leq \frac{\gamma}{2} \| x-y \|^2 \quad \forall\, x, y \in D_0.$$

Beweis: Definiere die differenzierbare Funktion $\varphi : [0,1] \to \mathbb{R}^n$ durch

$$\varphi(t) := f(y + t(x-y)), \quad x, y \in D_0,$$
$$\varphi'(t) = f'(y + t(x-y))(x-y) \in \mathbb{R}^n.$$

Mit der Voraussetzung folgt

$$\|\varphi'(t)-\varphi'(0)\| = \|(f'(y+t(x-y))-f'(y))(x-y)\|$$
$$\leq \gamma t \|x-y\| \|x-y\|.$$

Es ist

$$\Delta := f(x)-f(y)-f'(y)(x-y)$$
$$= \varphi(1)-\varphi(0)-\varphi'(0) = \int_0^1 (\varphi'(t)-\varphi'(0))\, dt,$$

$$\Rightarrow \|\Delta\| \leq \int_0^1 \|\varphi'(t)-\varphi'(0)\|\, dt$$
$$\leq \gamma \|x-y\|^2 \int_0^1 t\, dt = \frac{\gamma}{2}\|x-y\|^2. \quad \blacksquare$$

(11.3) <u>Satz:</u> (<u>Newton-Kantorovich</u>)
Es sei eine offene Menge $D \subseteq \mathbb{R}^n$ gegeben, ferner eine konvexe Menge D_0 mit $\overline{D}_0 \subseteq D$ und $f: D \to \mathbb{R}^n$ sei eine für alle $x \in D_0$ differenzierbare und für alle $x \in D$ stetige Funktion.

Für ein $x^0 \in D_0$ gebe es positive Konstanten $\tau, \alpha, \beta, \gamma, h$ mit:

$$S_\tau(x^0) := \{x \mid \|x-x^0\| < \tau\} \subseteq D_0,$$

$$h := \alpha\beta\gamma/2 < 1,$$

$$\tau := \alpha/(1-h).$$

$f(x)$ habe die Eigenschaften

(a) $\|f'(x) - f'(y)\| \leq \gamma \|x-y\|$ für alle $x, y \in D_0$
(Lipschitz-Bedingung für f')

(b) $f'(x)^{-1}$ existiert und es gilt
$\|(f'(x))^{-1}\| \leq \beta$ für alle $x \in D_0$,

(c) $\|(f'(x^0))^{-1} f(x^0)\| \leq \alpha$.

Dann gilt

(i) ausgehend von x^0 ist jedes
$$x^{k+1} = x^k - (f'(x^k))^{-1} f(x^k), \quad k \geq 0,$$
wohldefiniert und es gilt $x^k \in S_r(x^0)$ für alle $k \geq 0$.

(ii) $\bar{x} = \lim_{k \to \infty} x^k$ existiert und es gilt
$$\bar{x} \in \overline{S_r(x^0)} \text{ und } f(\bar{x}) = 0.$$

(iii)
$$\|\bar{x} - x^k\| \leq \alpha \frac{h^{2^k - 1}}{1 - h^{2^k}} \quad \text{für alle } k \geq 0.$$

Wegen $0 < h < 1$ ist also das Newton-Verfahren mindestens quadratisch konvergent.

Beweis: zu (i): Zunächst wird $x^k \in S_r(x^0)$, $k \geq 0$, induktiv gezeigt. Für $k=1$ ist

$$x^1 = x^0 - f'(x^0)^{-1} f(x^0) \Rightarrow \|x^1 - x^0\| \leq \alpha < \frac{\alpha}{1-h} = r.$$

Seien $x^0, \dots, x^k \in S_r(x^0)$:

$$\|x^{k+1} - x^k\| = \|-f'(x^k)^{-1} f(x^k)\| \leq \beta \|f(x^k)\|$$

$$= \beta \| f(x^k) - \underbrace{f(x^{k-1}) - f'(x^{k-1})(x^k - x^{k-1})}_{=0 \text{ nach Def.}} \|$$

$$\leq \frac{1}{2}\beta\gamma \|x^k - x^{k-1}\|^2 \quad \text{nach Hilfssatz (11.}$$

Hiermit zeigen wir nun induktiv

(11.4) $\|x^{k+1} - x^k\| \leq \alpha h^{2^k - 1}$.

Für $k=0$ ist dies bereits gezeigt (s.o.). Ist die Abschätzung für $k \geq 0$ richtig, so auch für $k+1$, denn

$$\|x^{k+1} - x^k\| \leq \frac{\beta\gamma}{2} \|x^k - x^{k-1}\|^2 \leq \frac{\beta\gamma}{2} \alpha^2 h^{2^k - 2} = \alpha h^{2^k - 1}$$

wegen $h = \frac{1}{2}\alpha\beta\gamma$. Nun folgt mit (11.4)

$$\|x^{k+1} - x^0\| \leq \|x^{k+1} - x^k\| + \|x^k - x^{k-1}\| + \dots + \|x^1 - x^0\|$$

$$\leq \alpha(1 + h + h^3 + h^7 + \dots + h^{2^k - 1}) < \alpha/(1-h) = r$$

und daher $x^{k+1} \in S_r(x^0)$.

Zu (iii): $\{x^k\}$ ist eine Cauchy-Folge, denn für $m \geq n$ hat man nach (11.4)

$$\|x^{m+1}-x^n\| \leq \|x^{m+1}-x^m\| + \|x^m-x^{m-1}\| + \ldots + \|x^{n+1}-x^n\|$$

$$\leq \alpha h^{2^n-1}\left(1+h^{2^n}+(h^{2^n})^2+\ldots\right)$$

$$< \frac{\alpha h^{2^n-1}}{1-h^{2^n}} < \varepsilon$$

für genügend großes $n \geq N(\varepsilon)$, da $0<h<1$. Also existiert

$$\lim_{k\to\infty} x^k =: \bar{x} \in \overline{S_+(x^0)} \,,$$

und für $m \to \infty$ ergibt sich die Abschätzung (iii). Zu zeigen ist noch $f(\bar{x}) = 0$:

$$\|f(x^k)\| = \|f'(x^k)(x^{k+1}-x^k)\| \leq \beta \|x^{k+1}-x^k\| \to 0 \text{ für } k\to\infty$$

$\Rightarrow f(\bar{x}) = 0$, da f stetig in $\bar{x} \in D$ ist. ∎

Bem.: (1) Sei $\bar{x}$ eine isolierte Nullstelle von f, sodaß die Vor. (a), (b) in einer Umgebung D_0 von $\bar{x}$ erfüllt sind. Für x^0 in einer genügend kleinen Umgebung E (Einzugsbereich) von $\bar{x}$ ist dann α in Vor. (c) hinreichend klein, sodaß die Vor. von Satz (11.3) erfüllt sind. Dann konvergiert $\{x^k\}$

quadratisch gegen $\bar{x}$ für $x^0 \in E$.

(2) Für eine C^2-Funktion f ist die Lipschitz-Bedingung in Vor. (a) erfüllt.

Erweiterung des Newton-Verfahrens:

1. Approximation von $f'(x)$ durch Differenzen

Die Berechnung von $f'(x^k)$ kann sehr zeitraubend bzw. explizit nicht möglich sein. Man kann $f'(x^k)$ z.B. numerisch approximieren, indem man $f(x)$ in der Nähe von x^k geeignet auswertet:

$$\left.\frac{\partial f_i(x)}{\partial x_j}\right|_{x=x^k} \doteq \frac{f_i(x^k + h e_j) - f_i(x^k)}{h}, \quad h > 0 \text{ klein}$$

Diese Methode erfordert n Berechnungen von f für jeden Schritt des Newton-Verfahrens und dies ist für n groß unpünktig.

2. λ-Strategie, modifiziertes Newton-Verfahren

Der Einzugsbereich des Newton-Verfahrens kann durch Einführung eines konvergenzerzeugenden Faktors $0 < \lambda \leq 1$

vergrößert werden. Sei

$$d^k = f'(x^k)^{-1} f(x^k) \in \mathbb{R}^n .$$

Gewöhnliches Newton-Verfahren:

$$x^{k+1} = x^k - d^k .$$

Modifiziertes Newton-Verfahren:

(11.5) $$\boxed{x^{k+1} = x^k - \lambda_k d^k , \quad 0 < \lambda_k \leq 1}$$

Zur Bestimmung der konvergenzerzeugenden Faktoren λ_k vergleiche man STOER, § 5.4.

§ 12 Nullstellenbestimmung für Polynome

Zu berechnen seien die reellen Nullstellen z_i eines Polynoms

$$p_n(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n ,$$

$$a_i \in \mathbb{R} , \quad a_0 \neq 0 .$$

z.B. sind die Eigenwerte λ_i einer symmetrischen $n \times n$ Matrix A die Nullstellen des charakteristischen Polynoms $p_n(\lambda) := \det(A - \lambda I)$.

Bei der Anwendung des Newton-Verfahrens (11.1)

$$x_{k+1} = x_k - \frac{p_n(x_k)}{p_n'(x_k)} , \quad k = 0, 1, \cdots \tag{12.1}$$

(unterer Iterationsindex k wegen $x \in \mathbb{R}$)

zur Berechnung der Nullstellen z_i treten folgende Probleme auf:

(1) Sparsame Berechnung von $p_n(x_k)$, $p_n'(x_k)$,
(2) Wahl eines guten Startwertes x_0,
(3) Effiziente Berechnung aller Nullstellen z_i.

Zu (1): HORNER-Algorithmus

Sei $z \in \mathbb{R}$ gegeben. Gesucht ist ein Polynom $p_{n-1}(x)$ mit

$$p_n(x) = (x - z) p_{n-1}(x) + b_n , \quad b_n \in \mathbb{R} , \tag{12.2}$$

$$p_{n-1}(x) = b_0 x^{n-1} + \cdots + b_{n-2} x + b_{n-1} .$$

Durch Einsetzen und Koeffizientenvergleich folgt die Rekursion

$$\boxed{\begin{aligned} b_0 &= a_0 \\ b_i &= a_i + b_{i-1} z , \quad i = 1, \ldots, n \end{aligned}} \tag{12.3}$$

Dies entspricht der rekursiven Berechnung

$$p_n(z) = (\cdots((a_0 z + a_1) z + a_2) z + \cdots) z + a_n .$$

Der Ansatz (12.2) ergibt dann

$$(12.4)\quad p_n(z) = b_n, \quad p_n'(z) = p_{n-1}(z).$$

Der Wert von $p_{n-1}(z)$ kann analog bestimmt werden:

$$p_{n-1}(x) = (x-z)\,p_{n-2}(x) + c_{n-1},$$

$$p_{n-2}(x) = c_0 x^{n-2} + c_1 x^{n-3} + \cdots + c_{n-2},$$

$$(12.5)\quad \boxed{\begin{aligned} c_0 &= b_0 \\ c_i &= b_i + c_{i-1} z, \quad i = 1, \ldots, n-1 \end{aligned}}.$$

Wie in (12.4) erhält man

$$p_{n-1}(z) = c_{n-1}, \quad p_{n-1}'(z) = p_{n-2}(z).$$

Zweimalige Differentation von (12.2) liefert

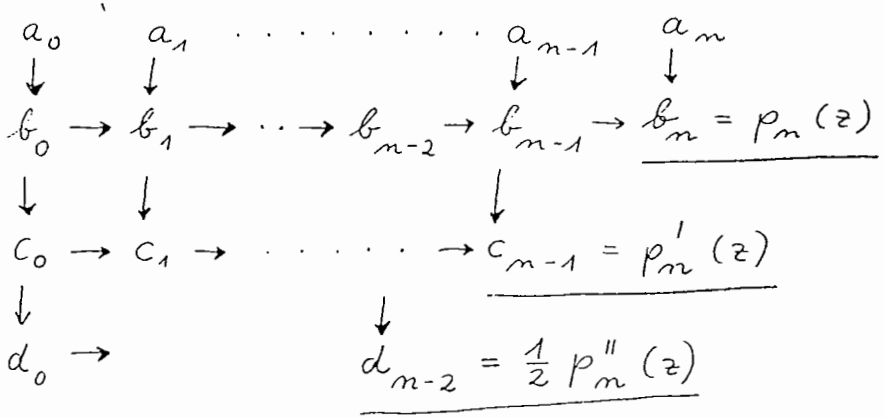
$$p_n''(x) = 2\,p_{n-1}'(x) + (x-z)\,p_{n-1}''(x),$$

$$p_n''(z) = 2\,p_{n-1}'(z) = 2\,p_{n-2}(z).$$

Allgemein erhält man rekursiv:

$$\boxed{\frac{1}{i!}\,p_n^{(i)}(z) = p_{n-i}(z), \quad i = 1, \ldots, n}$$

Dies ergibt das HORNER-Schema:



Allgemein:

$$\begin{aligned} a_k^{(0)} &= a_k \quad (k = 0, \ldots, n) \\ a_0^{(i)} &= a_0 \quad (i = 1, \ldots, n+1) \\ a_k^{(i)} &= a_k^{(i-1)} + z \cdot a_{k-1}^{(i)} \quad (i, k \geq 1) \\ a_{n-k}^{(k+1)} &= \frac{1}{k!}\,p_n^{(k)}(z) \end{aligned}$$

↓ + 1 ·
→ + z ·

Beispiel:

$$p_4(x) = 7x^4 + 5x^3 - 2x^2 + 8, \quad z = 0.5.$$

7	5	-2	0	8
7	8.5	2.25	1.125	$8.5625 = p_4(2)$
7	12.0	8.25	$5.25 = p_4'(2)$	
7	15.5	16.0		
7	19.0			
7				

zu (2): Wahl eines Startwertes x_0

Setze $p(x) = p_n(x)$.

(12.6) Satz: Hat $p(x)$ nur reelle Nullstellen z_i mit $z_n \le z_{n-1} \le \dots \le z_2 \le z_1$, so ist für alle $x_0 > z_1$ die Folge $\{x_k\}$ monoton fallend (streng monoton für $n \ge 2$) mit $\lim_{k\to\infty} x_k = z_1$.

Beweis: Für $n = 1$ trivial wegen $x_1 = z_1$.

Sei $n \ge 2$ und o.E. $p(x_0) > 0$, $a_0 > 0$.

$\Rightarrow p(x) > 0$ für $x > z_1$.

Nach dem Satz von Rolle hat $p'(x)$ Nullstellen α_i mit

$$z_1 \ge \alpha_1 \ge z_2 \ge \alpha_2 \ge \dots \ge \alpha_{n-1} \ge z_n$$

$\Rightarrow p'(x) > 0$ für $x > \alpha_1$.

Die nochmalige Anwendung des Satzes von Rolle ergibt

$$p''(x) > 0 \quad \text{für} \quad x > \alpha_1.$$

Wir zeigen induktiv $z_1 < x_k$:

Wegen $p(x_k) > 0$, $p'(x_k) > 0$ folgt zunächst

$$x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k)} < x_k.$$

Die Beh. $z_1 < x_{k+1}$ folgt dann aus

$$0 = p(z_1) = p(x_k) + (z_1 - x_k)\,p'(x_k) + \frac{1}{2}(z_1 - x_k)^2 p''(\delta) \qquad (z_1 < \delta < x_k)$$

$$> p(x_k) + (z_1 - x_k)\,p'(x_k)$$

$$= \underbrace{p'(x_k)}_{>0}\,(z_1 - x_{k+1}). \qquad \blacksquare$$

Zur Bestimmung eines Startwertes $x_0 > z_1$ kann man die folgenden aus der Algebra bekannten Abschätzungen benutzen.

(12.7) Satz: Für alle Nullstellen z_i von $p(x)$ gilt

$$|z_i| \le \max\left\{ \left|\frac{a_n}{a_0}\right|, 1 + \left|\frac{a_{n-1}}{a_0}\right|, \dots, 1 + \left|\frac{a_1}{a_0}\right| \right\},$$

$$|z_i| \le \max\left\{ 1, \sum_{i=1}^{n} \left|\frac{a_i}{a_0}\right| \right\},$$

$$|z_i| \le \max\left\{ \left|\frac{a_n}{a_{n-1}}\right|, 2\left|\frac{a_{n-1}}{a_{n-2}}\right|, \dots, 2\left|\frac{a_1}{a_0}\right| \right\}.$$

Für große Werte x_k gilt

$$x_{k+1} = x_k - \frac{x_k^n + \dots}{n x_k^{n-1} + \dots} \approx x_k \left(1 - \frac{1}{n}\right).$$

(langsame Konvergenz)

Dies führt zu der Idee, bei großen Werten von $x_k - z_1$ das Newton-Verfahren (12.1)

durch ein Doppelschritt-Verfahren zu ersetzen:

$$x_{k+1} = x_k - 2\frac{p(x_k)}{p'(x_k)}, \quad k = 0, 1, 2, \dots$$

Einzelheiten: Stoer I, Satz (5.5.9), S. 247.

Zu (2): Berechnung der Nullstellen $z_2, \dots, z_n$

1. Methode: Explizites Abdividieren von z_1

Berechne

$$p_{n-1}(x) = \frac{p_n(x)}{x - z_1}$$

und wende dann das Newton-Verfahren zur Bestimmung der größten Nullstelle z_2 an. Diese Vorgehensweise ist numerisch gefährlich. Da man nur eine Näherung z_1^* für z_1 berechnet, bestimmt man die Nullstelle z_2^* eines genäherten Polynoms $p_{n-1}^*(x)$. Wegen der großen Empfind-

lichkeit der Nullstellen gegenüber Störungen in den Koeffizienten (s.u.) können dann die weiteren Nullstellen $z_2, z_3, \ldots$ sehr ungenau sein.

2. Methode: Implizites Abdividieren

Die Nullstellen $z_1, \ldots, z_j$ seien bekannt, $1 \le j \le n-1$. Dann ist

$$p_{n-j}(x) = \frac{p_n(x)}{(x-z_1)\cdots(x-z_j)},$$

$$p'_{n-j}(x) = \frac{p'_n(x)}{(x-z_1)\cdots(x-z_j)} - \frac{p_n(x)}{(x-z_1)\cdots(x-z_j)} \sum_{i=1}^{j} \frac{1}{x-z_i}.$$

Das Newton-Verfahren zur Berechnung von z_{j+1} lautet dann

Machly

$$x_{k+1} = \phi_j(x_k), \quad k = 0, 1, \ldots$$

$$\phi_j(x) := x - \frac{p_n(x)}{p'_n(x) - \sum_{i=1}^{j} \frac{p_n(x)}{x-z_i}} \qquad (12.8)$$

(Implizite Deflation, Methode v. Machly)

Beispiele: ① Das Polynom

$$p_5(x) = x^5 - 5x^3 + 4x + 1$$

$p'_5(x) = 5x^4 - 15x^2 + 4$

hat die gesuchte Nullstelle $z_1 = 1.95408$.

Die Iteration (12.8) liefert mit $j = 1$ und Startwert $x_0 = 2$:

k	x_k	$p_5(x_k)$	$(x_k - z_1)^{-1}$	$x_{k+1} - x_k$
0	2.00000	1.00000	21.7770	$-4.49843 \cdot 10^{-1}$
1	1.55016	−2.47327	−2.47574	$-2.66053 \cdot 10^{-1}$
2	1.28411	−0.959150	−1.49260	$-1.11910 \cdot 10^{-1}$
3	1.17220	−0.151390	−1.27897	$-2.05572 \cdot 10^{-2}$
4	1.15164	−0.00466	−1.24620	$-6.55884 \cdot 10^{-4}$
5	1.15098 = z_2			

② Der Gewinn an numerischer Stabilität durch die Methode von Machly gegenüber dem expliziten Abdividieren wird durch das folgende Polynom demonstriert

$$p_{14}(x) = \prod_{j=0}^{13} (x - 2^{-j}) = \sum_{i=0}^{13} a_i x^{14-i}, \quad z_j = 2^{-j}.$$

Für $\varepsilon_{RD} = 10^{-12}$ erhält man den Fehler

j ($z_j = 2^{-j}$)	Machly ($\text{Fehler} \times 10^{12}$)	Abdivision
0	0	0
1	6.8	$3.7 \cdot 10^2$
2	1.1	$1.0 \cdot 10^6$
3	0.2	$1.4 \cdot 10^9$
4	4.5	
8	10.0	
9	5.3	$> 10^{12}$
10	0	

Die bisherigen Überlegungen bleiben auch für komplexe Koeffizienten a_i und Nullstellen z_j richtig. Das Rechnen mit komplexen Zahlen kann aber vollständig vermieden werden bei reellen Koeffizienten a_i. Ist $z_1 = u + iv$, $v \neq 0$, ist auch

$z_2 = u - iv$ eine Nullstelle von $p_n(x)$. Dann ist

$$(x - z_1)(x - z_2) = x^2 - 2ux + (u^2 + v^2)$$

ein quadratischer Teiler von $p_n(x)$. Die Methode von BAIRSTOW besteht in der Bestimmung eines quadratischen Teilers von $p_n(x)$ in der Form

$$p_n(x) = (x^2 - px - q)\, p_{n-2}(x) + b_{n-1}(x - p) + b_n .$$

Dies führt zu einem doppelzeiligen Horner-Schema: vgl. SCHWARZ, S. 224 - 229.

Kondition und Sensitivität der Nullstelle eines Polynoms

Problem: Wie verhalten sich die Nullstellen z von

$$p(x) = a_n x^n + \cdots + a_1 x + a_0$$

bei Störung der Koeffizienten

$$a_i \to a_i + \Delta a_i = a_i(1+\varepsilon).$$

Sei z einfache Nullstelle von $p(x)$ und sei $g(x)$ ein Polynom. Betrachte die Funktion

$$F(x,\varepsilon) := p(x) + \varepsilon g(x).$$

Wegen

$$F(z,0) = 0, \quad \frac{\partial F}{\partial x}(z,0) = p'(z) \neq 0$$

gibt es nach dem Satz über implizite Funktionen ein $\varepsilon_0 > 0$ und eine C^∞-Funktion $z(\varepsilon)$ für $|\varepsilon| \le \varepsilon_0$ mit $z(0) = z$ und

$$F(z(\varepsilon),\varepsilon) = p(z(\varepsilon)) + \varepsilon g(z(\varepsilon)) = 0, \quad |\varepsilon| \le \varepsilon_0.$$

Die Differentation nach ε ergibt

$$p'(z)\, z'(0) + g(z) = 0,$$

also

$$z'(0) = -\frac{g(z)}{p'(z)}.$$

In erster Näherung gilt daher

$$z(\varepsilon) \doteq z - \varepsilon \frac{g(z)}{p'(z)} \tag{12.9}$$

Die Anwendung auf $g(x) = a_i x^i$ liefert

$$z(\varepsilon) \doteq z - \varepsilon \frac{a_i z^i}{p'(z)}.$$

Beispiel: (WILKINSON)

$$p(x) = \prod_{j=1}^{20} (x-j) = \sum_{i=0}^{20} a_i x^i.$$

$z_{20} = 20$: $a_{19} = 1+2+\ldots+20 = 210$, $a_{19} \to a_{19}(1+\varepsilon)$.

$$z_{20}(\varepsilon) - z_{20} \doteq -\varepsilon \frac{210 \cdot 20^{19}}{19!} \doteq -\varepsilon \cdot 0.9 \cdot 10^{10}$$

$z_{16} = 16$: (größte Änderung), $a_{15} \approx 10^{10}$.

$$z_{16}(\varepsilon) - z_{16} \doteq -\varepsilon\, a_{15} \frac{16^{15}}{4!\,15!} \doteq -\varepsilon \cdot 3.7 \cdot 10^{14}.$$

Bei Störungen mit ε können sogar reelle Nullstellen in konjugiert komplexe aufgespalten werden.

§ 13 Iterationsverfahren für lineare Gleichungssysteme

Sei A reguläre (n,n)-Matrix und $b \in \mathbb{R}^n$. Zu lösen sei

$$Ax = b.$$

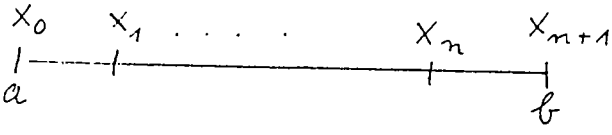
Falls
(1) A schwach besetzt ist,
(2) n sehr groß ist,
so wird der Rechenaufwand beim Gauß'schen Eliminationsverfahren zu groß. Man zieht dann iterative Fixpunktverfahren vor.

Anwendung: Diskretisierungsverfahren zur Lösung von Randwertaufgaben bei gew. und partiellen Differentialgleichungen

Beispiel: Sei $f \in C[a,b]$. Gesucht ist eine Lösung $u \in C^2[a,b]$ der Randwertaufgabe

$$u''(x) = f(x), \quad u(a) = u(b) = 0, \quad a \le x \le b.$$

Diskretisierung:



Schrittweite: $h = \frac{1}{n+1}(b-a)$, $x_i = a + ih$, $i = 0, \ldots, n+1$

Näherungen:

$$u'(x_i) \approx \frac{1}{h}(u(x_i) - u(x_{i-1})), \quad i = 1, \dots, n+1$$

$$f(x_i) = u''(x_i) \approx \frac{1}{h}\{u'(x_{i+1}) - u'(x_i)\}$$

$$\approx \frac{1}{h}\left\{\frac{1}{h}(u(x_{i+1}) - u(x_i)) - \frac{1}{h}(u(x_i) - u(x_{i-1}))\right\}$$

$$= \frac{1}{h^2}\{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\}, \quad i = 1, \dots, n$$

Wegen

$$u(x_0) = u(a) = 0, \quad u(x_{n+1}) = u(b) = 0$$

löst man für $(u_1, \dots, u_n)^T$ das LGS

$$\begin{aligned} u_2 - 2u_1 &= h^2 f(x_1) \\ u_{i+1} - 2u_i + u_{i-1} &= h^2 f(x_i), \quad i = 2, \dots, n-1 \\ -2u_n + u_{n-1} &= h^2 f(x_n). \end{aligned}$$

$$\underbrace{\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \\ 0 & & & 1 & -2 \end{pmatrix}}_{=: A} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = h^2 \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

(negativ definit nach §4)

A erfüllt <u>nicht</u> das starke Zeilensummenkriterium (13.10); jedoch erfüllt A das schwache Zeilensummenkriterium (13.12).

Bei der Entwicklung eines Iterationsverfahrens wird $Ax = b$ äquivalent umgeformt in die Fixpunktgleichung

$$x = Cx + d, \quad C\ (n,n)\text{-Matrix}, \quad d \in \mathbb{R}^n,$$

Dies ergibt die Iteration

(13.1)
$$\boxed{x^{(k+1)} = C x^{(k)} + d, \quad k \geq 0}.$$

<u>Spezialfälle:</u> Zerlege A in

$$A = L + D + R,$$

$$L = \begin{pmatrix} 0 & & 0 \\ a_{21} & \ddots & \\ \vdots & \ddots & \ddots \\ a_{n1} & \cdots & a_{n,n-1} \quad 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & & & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{n,n} \end{pmatrix}$$

(1) GS-Verfahren (Gesamtschritt- oder Jacobiverfahren)

$$A x = (L + D + R) x = b .$$

Iteration für $a_{ii} \neq 0$ $(i = 1, \dots, n)$:

$$D x^{(k+1)} + (L+R) x^{(k)} = b , \qquad k \geq 0 , \tag{13.2}$$

d.h.

$$\boxed{\begin{aligned} x^{(k+1)} &= C x^{(k)} + d , \quad k \geq 0 , \\ C &:= -D^{-1}(L+R) , \quad d := D^{-1} b \end{aligned}} \tag{13.3}$$

Explizit lautet die Iteration (13.2):

$$a_{ii} x_i^{(k+1)} + \sum_{j \neq i} a_{ij} x_j^{(k)} = b_i , \qquad i = 1, \dots, n . \tag{13.4}$$

Für weniger als $\frac{1}{3} n$ Schritte ist das GS-Verfahren schneller als das Eliminationsverfahren.

(2) ES-Verfahren (Einzelschritt- oder Gauß-Seidel-Verfahren)

Idee: man erzielt bessere Konvergenz, falls $x_j^{(k)}$ in (13.4) durch den schon berechneten Wert $x_j^{(k+1)}$ ersetzt wird für $j < i$:

(13.5) $$\sum_{j<i} a_{ij} x_j^{(k+1)} + a_{ii} x_i^{(k+1)} + \sum_{j>i} a_{ij} x_j^{(k)} = b_i, \quad i = 1, \dots, n$$

d. h.

(13.6)
$$\boxed{\begin{aligned} &(L+D)\, x^{(k+1)} + R\, x^{(k)} = b, \\ &x^{(k+1)} = C\, x^{(k)} + (L+D)^{-1} b, \quad C := -(L+D)^{-1} R \end{aligned}}.$$

Bemerkung: $(L+D)^{-1}$ muß bei der Iteration nicht explizit berechnet werden wegen (13.5). Der Unterschied zum GS-Verfahren besteht im Einsparen von Speicherplatz.

Konvergenzsätze:

Die Iteration

(13.7) $$x^{(k+1)} = C x^{(k)} + d, \quad k \geq 0, \ x^{(0)} \text{ gegeben},$$

heißt konvergent, wenn die Folge $\{x^{(k)}\}$ für alle $d, x^{(0)}$ konvergiert. Zum Nachweis der Konvergenz benötigen wir:

(13.8) Satz: Für den Spektralradius gilt

$$\begin{aligned} \rho(C) &= \max \{ |\lambda| : \lambda \text{ Eigenwert von } C \} \\ &= \inf \{ \|C\| : \|\cdot\| \text{ zugeordnet} \} \end{aligned}$$

Sind alle Eigenwerte λ mit $|\lambda| = \rho(C)$ einfach, so gilt $\rho(C) = \|C\|$ mit einer geeigneten Norm $\|\cdot\|$.

Beweis: Wir zeigen

(1) $\rho(C) \leq \|C\|$ für alle zugeordneten Matrixnormen

(2) für alle $\varepsilon > 0$ gibt es eine Matrixnorm mit

$$\|C\| \leq \rho(C) + \varepsilon.$$

Zu (1): Sei x_1 Eigenvektor von C zum Eigenwert λ mit $|\lambda| = \rho(C)$. Dann gilt

$$\rho(C) = \frac{\|Cx_1\|}{\|x_1\|} \leq \sup_{x \neq 0} \frac{\|Cx\|}{\|x\|} = \|C\|.$$

Zu (2): Sei $J = X^{-1} C X$ die Jordan'sche Normalform von C, d.h.

$$J = \begin{pmatrix} \lambda_1 & \theta_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \theta_{n-1} \\ 0 & & & \lambda_n \end{pmatrix},$$

$\theta_1 = \dots \theta_r = 0$, falls
$\rho(C) = |\lambda_1| = \dots = |\lambda_r|$
$> |\lambda_{r+1}| \geq \dots \geq |\lambda_n|$
$\lambda_1, \dots, \lambda_r$ einfach,

wobei die λ_i die Eigenwerte von C und die $\theta_i \in \{0, 1\}$ sind. Mit $\varepsilon > 0$ und

$$E = \begin{pmatrix} \varepsilon & & 0 \\ & \varepsilon^2 & \\ & & \ddots \\ 0 & & \varepsilon^n \end{pmatrix}$$

gilt

$$E^{-1} J E = \begin{pmatrix} \lambda_1 & \varepsilon\theta_1 & & 0 \\ & \lambda_2 & \varepsilon\theta_2 & \\ & & \ddots & \ddots \\ & & & \varepsilon\theta_n \\ 0 & & & \lambda_n \end{pmatrix}.$$

Wir definieren die Vektornorm

$$\|x\| := \|E^{-1} X^{-1} x\|_\infty$$

und berechnen damit

$$\|C\| = \max_{x \neq 0} \frac{\|Cx\|}{\|x\|} = \max_{x \neq 0} \frac{\|E^{-1} X^{-1} X J X^{-1} x\|_\infty}{\|\underbrace{E^{-1} X^{-1} x}_{=: y}\|_\infty}$$

$$= \max_{y \neq 0} \frac{\|E^{-1} J E y\|_\infty}{\|y\|_\infty} = \|E^{-1} J E\|_\infty$$

$$= \max_{i=1,\dots,n} (|\lambda_i| + \varepsilon\theta_i) \leq \rho(C) + \varepsilon .$$

Falls $\rho(G) = |\lambda_1| = \dots = |\lambda_r| > |\lambda_{r+1}| \geq \dots \geq |\lambda_n|$, so wähle man $\theta_1 = \dots = \theta_r = 0$, $\varepsilon = \rho(G) - |\lambda_{r+1}| > 0$.
Dann ist $\|G\| = \max (|\lambda_i| + \varepsilon\theta_i) = \rho(G)$. ∎

(13.9) Satz: Gegeben sei die Iteration

$$x^{(k+1)} = G x^{(k)} + d \quad (k \geq 0), \; x^{(0)} \text{ gegeben.}$$

(i) Hinreichend für die Konvergenz d. Iteration ist $\|C\| < 1$ für irgendeine Norm $\|\cdot\|$.

(ii) Die Iteration konvergiert genau dann, wenn $\rho(C) < 1$.

Beweis: Zu (i): Die Abbildung

$$g(x) := Cx + d$$

ist kontrahierend wegen

$$\|g(x) - g(y)\| = \|C(x-y)\| \leq \|C\| \, \|x-y\|, \quad \|C\| <$$

Der Fixpunktsatz (10.4) liefert dann die Behauptung.

Zu (ii): Sei $\rho(C) < 1$. Nach Satz (13.3) gibt es eine Norm $\|\cdot\|$ mit $\|C\| < 1$. Die Konvergenz folgt dann aus (i).
Die Iteration sei konvergent. Für $|\lambda| = \rho(C)$ mit $Cx = \lambda x$ setzen wir

$$x^{(0)} = d = x$$

Dann gilt

$$x^{(1)} = Cx^{(0)} + d = (\lambda + 1)x,$$

$$x^{(k)} = (\lambda^k + \lambda^{k-1} + \ldots + \lambda + 1)x$$

Da die Folge $\{x^{(k)}\}$ nach Vor. konvergiert, gilt $\rho(C) = |\lambda| < 1$. ∎

Anwendung auf das GS-Verfahren

Nach (13.3) ist

$$A = L + D + R, \quad C = -D^{-1}(L+R),$$

$$\|C\|_\infty = \max_{1 \le i \le n} \frac{1}{|a_{ii}|} \sum_{k \neq i} |a_{ik}|.$$

(Zeilensumme)

(13.10) Satz:

(i) (Starkes Zeilensummenkriterium)
Das GS-Verfahren konvergiert für Matrizen A mit

$$\sum_{k \neq i} |a_{ik}| < |a_{ii}| \quad \text{für } i = 1, \dots, n.$$

(ii) (Starkes Spaltensummenkriterium)
Das GS-Verfahren konvergiert für Matrizen A mit

$$\sum_{i \neq k} |a_{ik}| < |a_{kk}| \quad \text{für } k = 1, \dots, n.$$

Beweis: Zu (i): Die Behauptung folgt aus $\|C\|_\infty < 1$ und Satz (13.9) (i). Nach Satz (10.4) gilt außerdem $\|\bar{x} - x^{(k)}\| \le \frac{q^k}{1-q} \|x^{(1)} - x^{(0)}\|$, $q := \|G\|_\infty$

zu (ii): (Übung) Man wende (i) auf A^T an,

$$A^T = L^T + D + R^T, \quad \tilde{C} := -D^{-1}(L^T + R^T),$$

und zeige $\rho(C) = \rho(\tilde{C}) < 1$. ∎

Die Matrix des Anfangsbeispiels

$$A = \begin{pmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 0 & & & 1 & -2 \end{pmatrix}$$

erfüllt nicht (13.10); aber: A ist unzerlegbar und erfüllt (13.12)!

(13.11) Definition: $A = (a_{ik})$ heißt zerlegbar, wenn es nichtleere Teilmengen $N_1, N_2 \subset \{1, \dots, n\}$ gibt mit

(1) $N_1 \cup N_2 = \{1, \dots, n\}, \quad N_1 \cap N_2 = \emptyset,$

(2) für $i \in N_1$ und $k \in N_2$ gilt $a_{ik} = 0$,

d.h. es gibt eine Permutationsmatrix P mit

$$P^T A P = \begin{pmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{pmatrix} \begin{matrix} \} q \\ \} n-q \end{matrix},$$

$$N_1 = \{1, \dots, q\}, \quad N_2 = \{q+1, \dots, n\}.$$

A heißt unzerlegbar, falls A nicht zerlegbar ist.

(13.12) Satz: (Schwaches Zeilensummenkriterium)

Sei A unzerlegbar und es gelte

$$\sum_{k \neq i} |a_{ik}| \leq |a_{ii}|, \quad i = 1, \ldots, n,$$

$$\sum_{k \neq i_0} |a_{i_0 k}| < |a_{i_0 i_0}| \quad \text{für mindestens ein } i_0.$$

Dann konvergiert das GS-Verfahren.

Beweis: Zu zeigen ist: $\rho(C) < 1$.

Wegen $\|C\|_\infty \leq 1$ gilt $\rho(C) \leq 1$.

Annahme: $\rho(C) = |\lambda| = 1$, $\lambda \in \mathbb{C}$ Eigenwert.

Sei $x \in \mathbb{C}^n$ mit

$$Cx = \lambda x, \quad \|x\|_\infty = 1.$$

Setze

$$N_1 = \{ i \in \{1, \ldots, n\} \mid |x_i| = 1 \}, \quad N_2 = \{1, \ldots, n\} \setminus N_1.$$

Da $c_{ii} = 0$, so ist

$$(Cx)_i = \sum_{k \neq i} c_{ik} x_k = \lambda x_i, \quad i = 1, \ldots, n.$$

Wegen $|\lambda| = 1$ folgt

$$|x_i| = |\lambda x_i| \leq \sum_{k \neq i} \frac{|a_{ik}|}{|a_{ii}|} \underbrace{|x_k|}_{\leq 1} \leq \sum_{k \neq i} \frac{|a_{ik}|}{|a_{ii}|} \leq 1.$$

Für alle $i \in N_1$ ist $|x_i| = 1$ und die letzte Abschätzung ergibt $\sum_{k \neq i} |a_{ik}| = |a_{ii}|$.

Wegen

$$\sum_{k \neq i_0} |a_{i_0 k}| < |a_{i_0 i_0}|$$

ist $i_0 \in N_2$, d.h. $N_2 \neq \emptyset$. Da A unzerlegbar ist, gibt es $i_1 \in N_1$, $k_1 \in N_2$ mit $a_{i_1 k_1} \neq 0$. Damit ergibt sich ein Widerspruch:

$$1 = |x_{i_1}| \leq \sum_{k \neq i_1} \frac{|a_{i_1 k}|}{|a_{i_1 i_1}|} |x_k| < \sum_{k \neq i_1} \frac{|a_{i_1 k}|}{|a_{i_1 i_1}|} \leq 1 .$$

$\uparrow$

wegen $|x_{k_1}| < 1$, $|a_{i_1 k_1}| \neq 0$ ∎

Anwendung auf das ES-Verfahren

$C_G := -D^{-1}(L+R)$ G.S.-Verfahren

$C_E := -(L+D)^{-1} R$ ES-Verfahren

(13.13) Satz: Falls $\|C_G\|_\infty < 1$, so gilt

$$\|C_E\|_\infty \leq \|C_G\|_\infty .$$

Beweis: (Übung)

Idee: Zur Vereinfachung sei $a_{ii} = 1$. Für $x \in \mathbb{R}^n$ und $y := C_E x$ zeigt man induktiv

$$|y_k| \leq \|C_G\|_\infty \|x\|_\infty , \quad k = 1, \ldots, n .$$

Beispiel:

$$A = \begin{pmatrix} 1 & 0.1 \\ 6 & 1 \end{pmatrix} ,$$

$$C_G = -D^{-1}(L+R) = \begin{pmatrix} 0 & -0.1 \\ -6 & 0 \end{pmatrix} ,$$

$$C_E = -(D+L)^{-1} R = \begin{pmatrix} 0 & -0.1 \\ 0 & 0.6 \end{pmatrix} .$$

$$\|C_G\|_\infty = 6 , \quad \|C_E\|_\infty = 0.6 ,$$

$$\rho(C_G) = \sqrt{6} , \quad \rho(C_E) = 0.6 .$$

Das GS- und ES-Verfahren sind also konvergent (obwohl $\|C_G\|_\infty > 1$).

Relaxationsverfahren:

Eine Konvergenzbeschleunigung erreicht man beim ES-Verfahren durch Einführung eines Faktors $\omega > 0$. Hierzu multipliziert man das ES-Verfahren (13.5) mit ω

$$\omega \sum_{j<i} a_{ij} x_j^{(k+1)} + \omega a_{ii} x_i^{(k+1)} + \omega \sum_{j>i} a_{ij} x_j^{(k)} = \omega b_i ,$$
$$(i = 1, \dots, n)$$

und ersetzt den Wert $\omega a_{ii} x_i^{(k+1)}$ durch die "Mittelung"

$$a_{ii}\left(x_i^{(k+1)} - x_i^{(k)}\right) + \omega a_{ii} x_i^{(k)} .$$

Dies ergibt das Verfahren

$$a_{ii} x_i^{(k+1)} = a_{ii} x_i^{(k)} (1-\omega) + \omega \left[- \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} + b_i \right]$$
$$(i = 1, \dots, n; \quad k \geq 0)$$

Programm für N Schritte:

für $k = 1, \dots, N$:

  für $i = 1, \dots, n$:

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

In Matrix-Schreibweise bedeutet dies den Übergang vom ES-Verfahren

$$x^{(k+1)} = Cx^{(k)} + d,$$

$$C = -(L+D)^{-1}R, \quad d = (L+D)^{-1}b,$$

zum <u>Relaxationsverfahren</u>:

(13.14)
$$\boxed{\begin{aligned} x^{(k+1)} &= C(\omega)x^{(k)} + d(\omega) \\ C(\omega) &:= -(\omega L + D)^{-1}((\omega-1)D + \omega R), \\ d(\omega) &:= \omega(\omega L + D)^{-1}b. \end{aligned}}$$

Für $\omega = 1$ erhält man das ES-Verfahren.
Allgemein spricht man bei

$\omega < 1$ von Unterrelaxation,
$\omega > 1$ von Überrelaxation.

Für eine große Klasse von Matrizen A wählt man $\omega > 1$ und (13.14) ergibt das <u>SOR-Verfahren</u> (<u>S</u>uccessive <u>O</u>ver<u>r</u>elaxation).

Ohne Beweis zitieren wir

(13.15) Satz: Für beliebige Matrizen A gilt

$$\rho(C(\omega)) \geq |\omega - 1| \quad \text{für alle } \omega.$$

Das Relaxationsverfahren (13.14) konvergiert also bestenfalls für Parameter ω mit $0 < \omega < 2$.

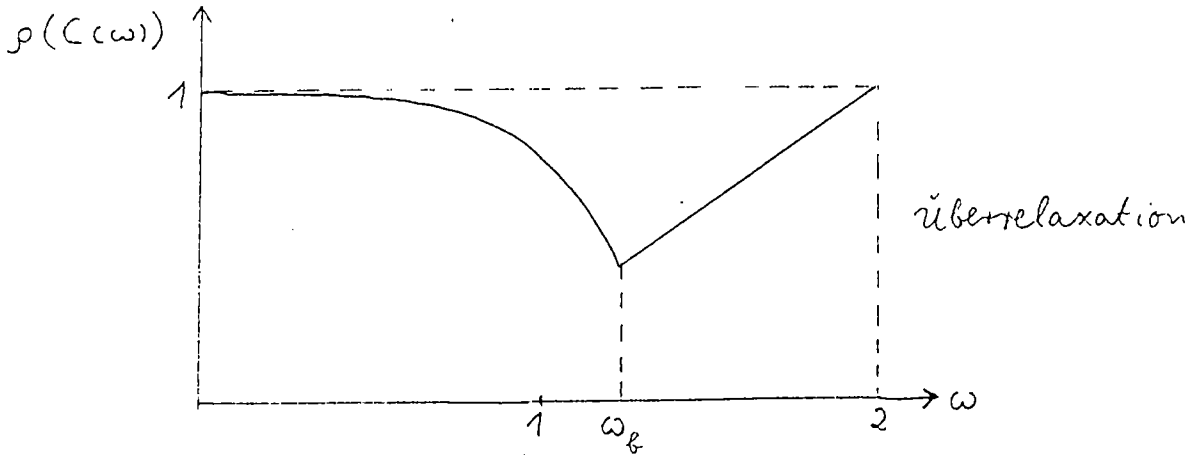
(13.16) Satz: Für positiv definite Matrizen A gilt

$$\rho(C(\omega)) < 1 \quad \text{für alle } 0 < \omega < 2.$$

Der optimale Relaxationsparameter ω_b mit

$$\rho(C(\omega_b)) = \min_{0<\omega<2} \rho(C(\omega))$$

läßt sich für eine wichtige Klasse von Matrizen A ("konsistent geordnet") explizit angeben.
Qualitativ gilt



vgl. Stoer / Bulirsch II, Satz (8.3.17).

Beweis von (13.16): Zu zeigen ist

$\rho(G(\omega)) < 1$.

Sei

$$G(\omega)x = \lambda x, \qquad |\lambda| = \rho(G(\omega)).$$

Dann gilt nach Def.

$$((1-\omega)D - \omega R)x = \lambda(D+\omega L)x,$$

also

$$(1-\omega)(Dx,x) - \omega(Rx,x) = \lambda[(Dx,x) + \omega(Lx,x)].$$

Da A positiv definit ist, so hat man $R = L^T$ und

$$0 < (Dx,x),$$
$$0 < (Ax,x) = (Dx,x) + 2(Lx,x).$$

Daraus folgt

$$q := \frac{(Lx,x)}{(Dx,x)} > -1/2.$$

Aus der obigen Gleichung für λ berechnet man

$$\lambda = \frac{(1-\omega) - \omega q}{1 + \omega q}.$$

Wegen $0 < \omega < 2$, $q > -1/2$, gilt

$$1 + \omega q > 0, \qquad -1 - q < q$$

und daher ist

$$-1 - \omega q < 1 - \omega - \omega q < 1 + \omega q$$

$$\Rightarrow |1 - \omega - \omega q| < 1 + \omega q$$

$$\Rightarrow |\lambda| = \frac{|1-\omega-\omega q|}{1+\omega q} < 1. \quad \blacksquare$$

Kapitel IV

Interpolation

Gegeben sein:

(1) $n+1$ Paare reeller oder komplexer Zahlen

$$(x_j, f_j), \quad j = 0, .., n \quad \text{(Stützpunkte)}.$$

Die x_j heißen Knoten; falls $x_j \in \mathbb{R}$, so hat man die Anordnung $x_0 < x_1 < .. < x_n$.
Z.B. sind $f_j = f(x_j)$ Meßdaten einer unbekannten Funktion f.

(2) Eine durch $n+1$ Parameter $a_0, .., a_n$ bestimmte Familie ϕ von Funktionen

$$\phi(x; a_0, .. a_n).$$

Gesucht: Parameter $a_0, .., a_n$ mit

$$\phi(x_j; a_0, .., a_n) = f_j, \quad j = 0, .., n$$

Dann ist $\phi(x; a_0, .., a_n)$ ein Näherungswert für $f(x)$.

Lineare Interpolationsprobleme

$\phi(x; a_0, .., a_n)$ hänge linear von $a_0, ..., a_n$ ab, d.h.

$$\phi(x; a_0, .., a_n) = a_0 \phi_0(x) + a_1 \phi_1(x) + ... + a_n \phi_n(x),$$

$\phi_i(x)$ heißen Basisfunktionen.

Beispiele:

(1) Interpolation durch Polynome:

$$\phi(x; a_0, .., a_n) = a_0 + a_1 x + .. + a_n x^n,$$

$$\phi_i(x) = x^i.$$

(2) Trigonometrische Interpolation:

$$\phi(x; a_0, .., a_{n-1}) = a_0 + a_1 e^{xi} + a_2 e^{2xi} + ... + a_{n-1} e^{(n-1)xi}$$

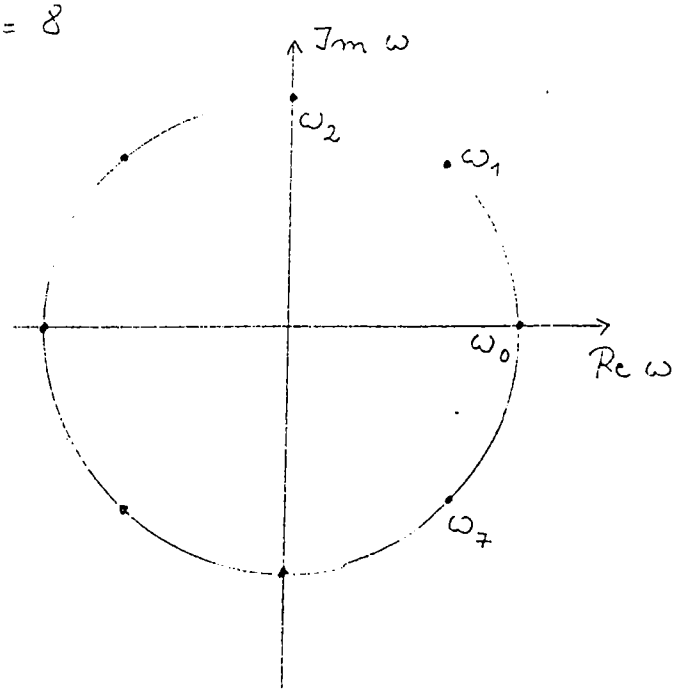
$$= a_0 + a_1 \omega + a_2 \omega^2 + ... + a_{n-1} \omega^{n-1},$$

wobei

$$\omega = e^{xi} = \cos x + i \sin x, \quad i^2 = -1.$$

Knoten: $\omega_j = e^{x_j i}, \quad x_j = \frac{2\pi j}{n}, \quad j = 0, .., n-1.$

z.B. $n = 8$



(3) Spline - Interpolation:

z.B. kubische Spline-Funktionen mit

a) $\phi(\cdot\,; a_0, \ldots, a_n) \in C^2[x_0, x_n]$,

b) $\phi(\cdot\,; a_0, \ldots, a_n) \in C^3(x_i, x_{i+1})$.

Nichtlineare Interpolationsprobleme:

Interpolation durch rationale Funktionen

$$\phi(x; a_0, \ldots, a_n, b_0, \ldots, b_m) = \frac{a_0 + a_1 x + \ldots + a_n x^n}{b_0 + b_1 x + \ldots + b_m x^m}.$$

Interpolation durch Exponentialsummen

$$\phi(x; a_0, \ldots, a_n, \lambda_0, \ldots, \lambda_n) = a_0 e^{\lambda_0 x} + \ldots + a_n e^{\lambda_n x}.$$

§ 15 Interpolation durch Polynome

Π_n : Menge aller reellen oder komplexen Polynome vom Grade $\le n$

$$P(x) = a_0 + a_1 x + \dots + a_n x^n .$$

15.1 Satz: Zu beliebigen $n+1$ Stützstellen

(x_i, f_i) $i = 0 \dots n$, $x_i \neq x_k$ für $i \neq k$,

gibt es genau ein Polynom $P \in \Pi_n$ mit

$$P(x_i) = f_i , \quad i = 0, \dots, n .$$

Beweis: Die Bedingung $P(x_i) = f_i$, $i = 0 \dots n$, ergibt das LGS

$$\sum_{k=0}^{n} a_k x_i^k = f_i , \quad i = 0, \dots, n ,$$

d.h.

$$\underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix}}_{\text{VANDERMONDE-Matrix } X} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix} .$$

Nun ist

$$\det(X) = \det\left(x_j^k\right) = \prod_{\substack{j,k=0 \\ j>k}}^{n} (x_j - x_k) \neq 0 ,$$

da $x_j \neq x_k$ für $j \neq k$.

Also hat das obige LGS genau eine Lösung $a_0, \ldots, a_n$. ∎

Bem.: Die Koeffizienten a_i können explizit mit der Cramer'schen Regel unter Verwendung der Matrix X berechnet werden. Zur praktischen Berechnung des Polynoms P stellen wir jedoch drei weniger aufwendige Methoden vor.

15.1 Die Interpolationsformel von Lagrange

Für das Polynom

$$L_i(x) := \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}$$

gilt

$$L_i \in \Pi_n , \quad L_i(x_j) = \delta_{ij} = \begin{cases} 1 , & i = j \\ 0 , & i \neq j \end{cases} .$$

Das gesuchte Polynom $P(x)$ ist dann wegen der Eindeutigkeit

(15.2)
$$P(x) = \sum_{i=0}^{n} f_i L_i(x) = \sum_{i=0}^{n} f_i \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k} .$$

Beispiel:

j	x_j	f_j
0	0	1
1	1	3
2	3	2

$$L_0(x) = \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{1}{3}(x^2 - 4x + 3),$$

$$L_1(x) = \frac{(x-0)(x-3)}{(1-0)(1-3)} = -\frac{1}{2}(x^2 - 3x),$$

$$L_2(x) = \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{1}{6}(x^2 - x).$$

$$P(x) = \frac{1}{3}(x^2 - 4x + 3) + 3\left(-\frac{1}{2}(x^2 - 3x)\right) + 2\,\frac{1}{6}(x^2 - x)$$

$$= -\frac{5}{6}x^2 + \frac{17}{6}x + 1 .$$

In der Praxis möchte man häufig ein vorhandenes Interpolationspolynom durch Hinzunahme weiterer Stützstellen verbessern. Die Lagrange-

-Methode ist hierfür nicht geeignet, da man bei neu hinzukommenden Stützstellen mit der Rechnung von vorn beginnen muß.

15.2 Der Algorithmus von Aitken und Neville

Gesucht ist ein numerisch sparsamer Algorithmus zur Berechnung von $P(x)$ an einigen wenigen Stellen x.

Für $i_0, \dots, i_k \in \{0, 1, \dots, n\}$ sei $P_{i_0,\dots,i_k} \in \Pi_k$ das Interpolationspolynom zu $x_{i_0}, \dots, x_{i_k}$; insbesonder

$$P_i(x) \equiv f_i, \quad P_{0,\dots,n}(x) = P(x).$$

Es gilt die Rekursionsformel (Aitken)

$$P_{i_0,\dots,i_k}(x) = \frac{(x - x_{i_0}) P_{i_1,\dots,i_k}(x) - (x - x_{i_k}) P_{i_0,\dots,i_{k-1}}(x)}{x_{i_k} - x_{i_0}} \tag{15.3}$$

Beweis: Das Polynom $Q(x) \in \Pi_k$ auf der rechten Seite von (15.3) erfüllt

$$Q(x_{i_0}) = P_{i_0,\dots,i_{k-1}}(x_{i_0}) = f_{i_0},$$

$$Q(x_{i_k}) = P_{i_1,\dots,i_k}(x_{i_k}) = f_{i_k},$$

$$Q(x_{i_j}) = f_{i_j}, \quad j = 1, \dots, k-1.$$

Wegen der Eindeutigkeit der Polynom-Interpolation folgt dann (15.3).

Variante von Neville: Sei x fest.

Erzeuge die Werte $P_{i-k,\dots,i}(x)$ $(k \le i)$ nach dem Schema

.	$k=0$	1	2	3
x_0	$f_0 = P_0(x)$			
x_1	$f_1 = P_1(x)$	$P_{0,1}(x)$		
x_2	$f_2 = P_2(x)$	$P_{1,2}(x)$	$P_{0,1,2}(x)$	
x_3	$f_3 = P_3(x)$	$P_{2,3}(x)$	$P_{1,2,3}(x)$	$P_{0,1,2,3}(x)$

Bei Hinzunahme von weiteren Stützstellen wird das Schema einfach erweitert, ohne daß die bereits berechneten Ergebnisse ungültig werden.

Für festes x gilt mit der Bezeichnung

$$P_{i,k} := P_{i-k,\dots,i}(x) \quad (k \le i):$$

Startwerte: $P_{i,0} := f_i, \quad i = 0,\dots,n,$

Rekursion:

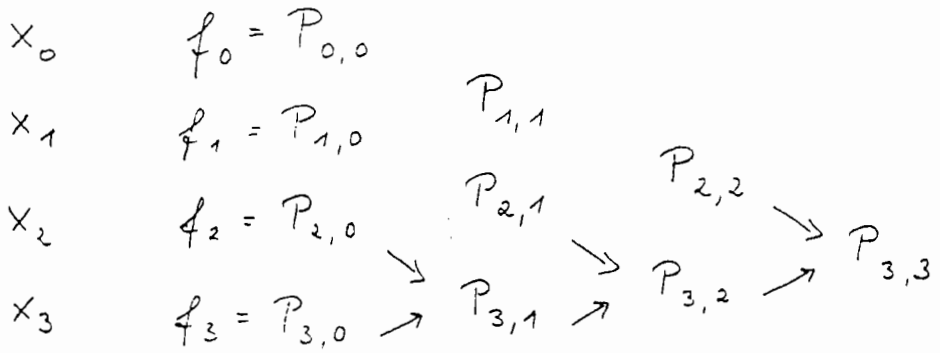
$$P_{i,k} = \frac{(x - x_{i-k})\,P_{i,k-1} - (x - x_i)\,P_{i-1,k-1}}{x_i - x_{i-k}},$$

(15.4)

$$= P_{i,k-1} + \frac{P_{i,k-1} - P_{i-1,k-1}}{\dfrac{x - x_{i-k}}{x - x_i} - 1}, \qquad \begin{matrix} k = 1,\dots,i, \\ i = 1,2,\dots \end{matrix}$$

Resultat: $P_{n,n} = P(x)$.

Das obige Schema lautet dann



Diese Variante wird speziell für $x = 0$ bei Extrapolationsalgorithmen angewandt (vgl. § 27).

15.3 Die Newton'sche Interpolationsformel. Dividierte Differenzen

Es sollen die Koeffizienten des Interpolationspolynoms $P(x)$ berechnet werden. Für $P = P_{0,\ldots,n}$ macht man den Ansatz

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)\cdots(x - x_{n-1}). \tag{15.5}$$

Die Auswertung dieses Ausdruckes für festes x geschieht mit dem Horner-Schema (vgl. § 12):

$$P(x) = [\dots (a_n (x - x_{n-1}) + a_{n-1})(x - x_{n-2}) + \dots + a_1](x - x_0) + a_0$$

d.h. rekursiv

$$b_n = a_n ,$$

$$b_i = b_{i+1}(x - x_i) + a_i , \quad i = n-1, \dots, 0 ,$$

$$b_0 = P(x) .$$

Für die Abschnittspolynome

$$Q_k(x) = \sum_{i=0}^{k} a_i \prod_{j=0}^{i-1} (x - x_j)$$

folgt

(a) $Q_k(x) = P_{0,1,\dots,k}(x) , \quad k = 0, \dots, n ,$

(b) a_k ist Koeffizient von x^k in $P_{0,1,\dots,k}(x)$.

Die Koeffizienten a_k werden nicht mittels der Bedingung $P(x_i) = f_i$ berechnet, sondern mit Hilfe des Differenzenschemas:

x_0	$[f_0]$		
		$[f_0, f_1]$	
x_1	$[f_1]$		$[f_0, f_1, f_2]$
		$[f_1, f_2]$	
x_2	$[f_2]$		
$\vdots$			

Die hier auftretenden "Dividierten Differenzen" $[f_i, \dots, f_k]$ sind rekursiv definiert durch

(a) $[f_i] = f_i$,

(b) $$[f_i, \dots, f_k] = \frac{[f_{i+1}, \dots, f_k] - [f_i, \dots, f_{k-1}]}{x_k - x_i},$$

z.B.

$$[f_0, f_1] = \frac{f_1 - f_0}{x_1 - x_0},$$

$$[f_0, f_1, f_2] = \frac{[f_1, f_2] - [f_0, f_1]}{x_2 - x_0}.$$

(15.6) Satz: Es gilt

$$P_{0,\dots,k}(x) = \sum_{i=0}^{k} [f_0, \dots, f_i] \prod_{j=0}^{i-1} (x - x_j).$$

Beweis: Durch Induktion bzgl. k: für $k = 0$ ist die Beh. offensichtlich richtig. Die Aussage gelte für $k - 1 \geq 0$.

$$P_{0,\dots,k}(x) = P_{0,\dots,k-1}(x) + a(x - x_0) \cdots (x - x_{k-1}).$$

Zu zeigen ist $a = [f_0, \dots, f_k]$.

Koeffizient von x^k in $P_{0,..,k}(x)$: a

" " x^{k-1} " $P_{0,..,k-1}(x)$: $[f_0,..,f_{k-1}]$

" " x^{k-1} " $P_{1,..,k}(x)$: $[f_1,..,f_k]$

(nach Induktionsvoraussetzung)

Nach der Formel von Aitken (15.3) ist

$$P_{0,..,k}(x) = \frac{(x-x_0)\,P_{1,..,k}(x) - (x-x_k)\,P_{0,..,k-1}(x)}{x_k - x_0}.$$

Der Koeffizient von x^k auf der rechten Seite ist

$$a = \frac{[f_1,..,f_k] - [f_0,..,f_{k-1}]}{x_k - x_0} = [f_0,..,f_k]. \quad \blacksquare$$

(Definition)

Beispiel: (vgl. Lagrange-Form)
Das Differenzenschema lautet

$x_0 = 0$	1		
		2	
$x_1 = 1$	3		$-\frac{5}{6}$
		$-1/2$	
$x_2 = 3$	2		

a_0, a_1, a_2

$$P_{0,1,2}(x) = 1 + 2(x-0) - \frac{5}{6}(x-0)(x-1).$$

(15.7) Satz: Für eine beliebige Permutation $i_0,..,i_n$ von $0,...,n$ gilt

$$[f_{i_0},..,f_{i_n}] = [f_0,..,f_n].$$

15.4 Der Interpolationsfehler, Konvergenzfragen

Sei $f \in C^{n+1}[a,b]$ und $x_0, \dots, x_n \in [a,b]$. Wir betrachten den Fehler $f(x) - P(x)$, wobei $P(x)$ das Interpolationspolynom zu den Stützstellen (x_j, f_j), $f_j = f(x_j)$, $j = 0, \dots, n$, ist.

(15.8) Satz: Sei $f \in C^{n+1}[a,b]$ und $\bar{x}, x_j \in [a,b]$, $j = 0, \dots, n$. Dann gibt es $\xi \in [a,b]$ mit

$$f(\bar{x}) - P(\bar{x}) = L(\bar{x}) \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

$$L(x) = (x - x_0) \cdots (x - x_n).$$

Beweis: Betrachte für $\bar{x} \neq x_j$ die Funktion

$$F(x) := f(x) - P(x) - \frac{f(\bar{x}) - P(x)}{L(\bar{x})} L(x) \in C^{n+1}[a,b].$$

F hat die $n+2$ Nullstellen $\bar{x}, x_0, \dots, x_n$ in $[a,b]$. Nach dem Satz von Rolle hat $F^{(n+1)}$ mindestens eine Nullstelle ξ in $[a,b]$:

$$0 = F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(\bar{x}) - P(\bar{x})}{L(\bar{x})} (n+1)!$$

$\Rightarrow$ Beh.. ■

Für den Interpolationsfehler erhalten wir also die Abschätzung (Bez. $\|g\|_\infty := \max_{x\in[a,b]} |g(x)|$)

$$\|f - P\|_\infty \leq \|L\|_\infty \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}$$

Konvergenzfragen: Sei

$$\Delta_m := \{ a = x_0^{(m)} < x_1^{(m)} < \ldots < x_{n_m}^{(m)} = b \}, \quad m = 0, 1, \ldots,$$

eine Folge von Intervallteilungen von $[a, b]$.

$$\|\Delta_m\| = \max_i |x_{i+1}^{(m)} - x_i^{(m)}| .$$

$P_{\Delta_m}(x)$ bezeichne das interpolierende Polynom von f bzgl. Δ_m.

Problem: Gilt $\lim_{m\to\infty} P_{\Delta_m}(x) = f(x)$ für

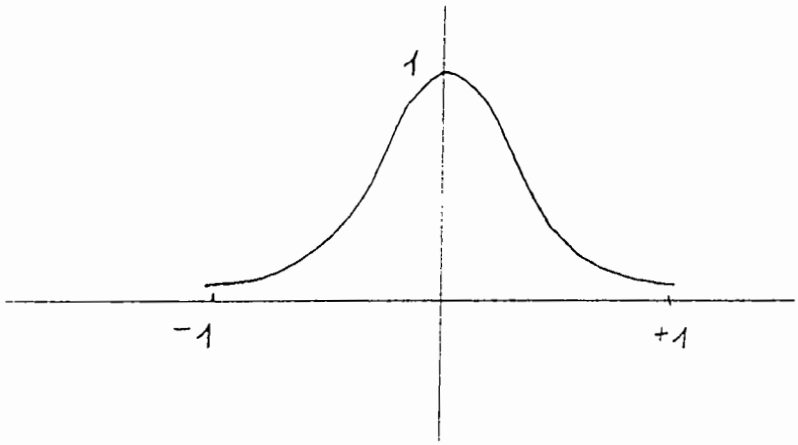
$$\lim_{m\to\infty} \|\Delta_m\| = 0 \quad ?$$

Antwort: i.a. nicht richtig.

Beispiel von Runge:

$$f(x) = (1 + 25x^2)^{-1} \text{ in } [-1, 1]$$

$f(x)$ ist bzgl. $x \in \mathbb{C}$ keine ganze Funktion; vgl. Satz (15.10).



. f wird an den Stellen $x_j = -1 + \frac{2j}{n}$, $j = 0, .., n$, interpoliert durch ein Polynom P_n.

n	$\|f - P_n\|_\infty$
1	0,96
5	0.43
13	1.07
19	8,57

Wir wählen nun die Knoten $x_0, .., x_n$ so, daß $\|L\|_\infty = \max_{-1 \le x \le 1} |L(x)|$ möglichst klein wird. Man erhält

$$L(x) = 2^{-n} T_{n+1}(x), \quad x \in [-1, 1],$$

wobei die Tschebyscheff-Polynome T_n wie folgt definiert sind (vgl. § 19.3):

$$T_n(x) = \cos n\theta, \quad x = \cos\theta, \quad 0 \le \theta \le 2\pi.$$

Es gilt $\| T_n \|_\infty \leq 1$ und die Nullstellen von $L(x) = 2^{-n} T_{n+1}(x)$ sind

$$x_j = \cos \frac{(j + \frac{1}{2})\pi}{n+1}, \quad j = 0, \ldots, n.$$

Bei dieser Wahl der Knoten ergibt sich die Fehlerabschätzung

n	$\| f - P_n \|_\infty$
1	0.93
5	0.56
13	0.12
19	0.04

Die Verbesserung ist erheblich, die Approximation aber immer noch unbefriedigend.

Ohne Beweis geben wir noch an.

(15.9) Satz: Zu jeder Folge $\{\Delta_m\}$ gibt es $f \in C[a,b]$, so daß $\{P_{\Delta_m}\}$ nicht gleichmäßig gegen f konvergiert.

(15.10) Satz: Sei f eine ganze Funktion. Dann gilt $P_{\Delta_m} \to f$ gleichmäßig für alle $\{\Delta_m\}$ mit $\|\Delta_m\| \to 0$.

Bem.: $P(x) = P_{0,\dots,n}(x)$ oszilliert i. a. stark für großes n; die Interpolation ist dann unbrauchbar. Diese Schwierigkeiten werden bei der Spline-Interpolation vermieden; vgl. Konvergenzsatz (17.9).

§16 Trigonometrische Interpolation

Gegeben: $I = [0, 2\pi]$,

Stützpunkte (x_k, f_k), $k = 0, \ldots, n-1$,

$x_k = k \cdot \frac{2\pi}{n} \in I$, $f_k \in \mathbb{C}$.

Gesucht: "Trigonometrisches Polynom"

$$p(x) = \beta_0 + \beta_1 e^{xi} + \beta_2 e^{2xi} + \ldots + \beta_{n-1} e^{(n-1)xi}, \quad \beta_i \in \mathbb{C},$$

$$e^{xi} = \cos x + i \sin x, \quad i = \sqrt{-1},$$

mit

$$p(x_k) = f_k, \quad k = 0, \ldots, n-1.$$

Setze

$$\omega = e^{ix},$$

$$\omega_k = e^{ix_k} = e^{2k\pi i/n}, \quad \omega_j \neq \omega_k \text{ für } j \neq k,$$

$$P(\omega) := \beta_0 + \beta_1 \omega + \ldots + \beta_{n-1} \omega^{n-1}.$$

Nach Satz (15.1) ist die Interpolation bzgl ω

(16.1)
$$P(\omega_k) = \sum_{j=0}^{n-1} \beta_j \omega_k^j = f_k, \quad k = 0, \ldots, n$$

FOURIER-Synthese

eindeutig lösbar. Für die Berechnung der Koeffizienten β_k benötigen wir

(16.2)

$$(a)\quad \omega_k^j = \omega_j^k, \qquad 0 \le j,k \le n-1,$$

$$(b)\quad \sum_{k=0}^{n-1} \omega_k^j \omega_k^{-l} = \begin{cases} n, & j = l \\ 0, & j \neq l, \end{cases} \qquad 0 \le j,l \le n-1$$

<u>Beweis:</u> Die Aussage (a) ist trivial,

<u>Zu (b):</u> ω_k ist n-te Einheitswurzel von

$$\omega^n - 1 = (\omega - 1)(\omega^{n-1} + \omega^{n-2} + \ldots + 1) = 0.$$

Wegen $\omega_0 = 1$ und $\omega_k \neq 1$ für $k \neq 0$ folgt damit

$$\sum_{k=0}^{n-1} \omega_k^j \omega_k^{-l} = \sum_{k=0}^{n-1} \omega_k^{j-l} = \sum_{k=0}^{n-1} \omega_{j-l}^k = \begin{cases} n, & j = l \\ 0, & j \neq l. \end{cases}$$

(16.3) <u>Satz:</u> Für die Koeffizienten β_j in (16.1) gilt

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k \omega_k^{-j}, \qquad \omega_k^{-j} = e^{-2\pi jki/n},$$

$$j = 0, \ldots, n-1,$$

FOURIER-Analyse

Beweis: Mit (19.2) (b) folgt

$$\sum_{k=0}^{n-1} f_k \omega_k^{-j} = \sum_{k=0}^{n-1} \left(\beta_0 + \ldots + \beta_j \omega_k^{j} + \ldots + \beta_{n-1} \omega_k^{n-1} \right) \omega_k^{-j}$$

$$= \beta_j \cdot n . \quad \blacksquare$$

Wir stellen nun die Beziehung zum Titel dieses Paragraphen her und drücken $p(x) = P(\omega)$ durch trigonometrische Funktionen aus: Setze

(16.4) $$A_j := \frac{2}{n} \sum_{k=0}^{n-1} f_k \cos \frac{2\pi k j}{n}, \quad B_j := \frac{2}{n} \sum_{k=0}^{n-1} f_k \sin \frac{2\pi k j}{n}, \quad j = 0, \ldots, n-1$$

Dann gilt

(1) $$\beta_{n-j} = \frac{1}{n} \sum_{k=0}^{n-1} f_k \omega_k^{j-n} = \frac{1}{n} \sum_{k=0}^{n-1} f_k \omega_k^{j},$$

(16.5) (2) $$\beta_j = \frac{1}{2} (A_j - i B_j), \quad \beta_{n-j} = \frac{1}{2} (A_j + i B_j),$$

(3) $$\beta_j \omega_k^{j} + \beta_{n-j} \omega_k^{n-j} = A_j \cos j x_k + B_j \sin j x_k .$$

Beweis:

(1) Wegen $\omega_k^n = 1$,

(2) wegen (16.2),

(3) folgt aus (2).

(16.6) **Satz:** Definiert man A_j, B_j gemäß (16.4) und setzt für ungerades $\underline{n = 2m+1}$

$$\psi(x) := \frac{A_0}{2} + \sum_{j=1}^{m} \left(A_j \cos jx + B_j \sin jx \right)$$

bzw. für gerades $\underline{n = 2m}$

$$\psi(x) := \frac{A_0}{2} + \sum_{j=1}^{m-1} (A_j \cos jx + B_j \sin jx) + \frac{A_m}{2} \cos mx,$$

so gilt mit $x_k = k \cdot 2\pi/n$

$$\psi(x_k) = f_k, \quad k = 0, \ldots, n-1.$$

Beweis: Für gerades $\underline{n = 2m}$ ist

$$B_0 = 0, \quad B_m = 0, \quad \omega_k^m = \cos mx_k$$

und daher gilt mit (16.5)

$$f_k = \sum_{j=0}^{n-1} \beta_j \omega_k^j = \beta_0 + \sum_{j=1}^{m-1} \underbrace{\left(\beta_j \omega_k^j + \beta_{n-j} \omega_k^{n-j} \right)}_{= A_j \cos jx + B_j \sin jx} + \beta_m \omega_k^m$$

$$\beta_0 = \frac{A_0}{2}, \qquad \beta_m = \frac{A_m}{2} \cos mx$$

$$= \psi(x_k).$$

Ebenso folgt die Behauptung für ungerades $n = 2m+1$.

Hinweise zur praktischen Durchführung

Nach (16.1), (16.3) ist zu berechnen

$$f_k = \sum_{j=0}^{n-1} \beta_j \omega_k^j , \qquad \omega_k = e^{2\pi k i/n}$$

(16.6)

$$\beta_j = \frac{1}{n} \sum_{k=0}^{n-1} f_k \omega_k^{-j} .$$

Die trigonometrischen Summen in (16.4), (16.5) können hierauf zurückgeführt werden.

Mit

$$y := (f_0, \dots, f_{n-1})^T , \quad \hat{y} := (\beta_0, \dots, \beta_{n-1})^T$$

und der (n,n)-Matrix

$$T = (t_{jk}) , \quad t_{jk} = \omega_k^j = e^{2\pi jki/n}$$

lautet (19.6):

(16.7)

$$y = T\hat{y} ,$$

$$\hat{y} = \frac{1}{n} P T y ,$$

P Permutationsmatrix, da $\omega_k^{-j} = \omega_k^{n-j}$

Die Berechnung von (16.7) erfordert

$O(n^2)$ Operationen.

Für $\boxed{n = 2^m}$ kann (16.7) sehr effektiv mit dem Algorithmus von Cooley und Tukey (FFT = Fast Fourier Transform) berechnet werden.

Nach Stoer I, Satz (2.3.3.4), läßt sich die Matrix T faktorisieren zu

$$\therefore T = Q\,S\,P\,(D_{m-1}\,S\,P)\,..\,(D_1\,S\,P) = T_m \cdot T_{m-1} \,..\, T_1\,,$$

$$S = \begin{pmatrix} \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} & & & 0 \\ & \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} & & \\ & & \ddots & \\ 0 & & & \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} \end{pmatrix} \quad n \times n\,,$$

Q, P: geeignete Permutationsmatrizen

D_i: " Diagonalmatrizen.

Also

$T_i z$: $O(n)$ Operationen

$T \hat{y}$: $O(n \cdot m)$ " , $m = \log_2 n$.

$$\boxed{O(n \log_2 n) \text{ Operationen}}$$

§ 17 Spline-Funktionen

Spline-Funktionen, die sich stückweise aus Polynomen zusammensetzen, verbinden den Vorteil einer glatten Interpolation mit denjenigen, die der Umgang mit Polynomen niedrigen Grades mit sich bringt.

17.1 Polynom-Splines

Sei $\Delta = \{ a = x_0 < x_1 < \cdots < x_n = b \}$ eine Zerlegung des Intervalls $[a,b] \subset \mathbb{R}$ mit inneren Knoten $x_1, \ldots, x_{n-1}$ und Randknoten x_0, x_n.

(17.1) Definition: Eine Funktion $s : [a,b] \to \mathbb{R}$ heißt Polynom-Spline vom Grad ℓ ($\ell = 0, 1, 2, \ldots$) zur Zerlegung Δ, wenn sie folgende Eigenschaften besitzt:

(a) $s \in C^{\ell-1}[a,b]$

(b) $s \in \Pi_\ell$ für $x_j \le x < x_{j+1}$, $j = 0, 1, \ldots, n-1$.

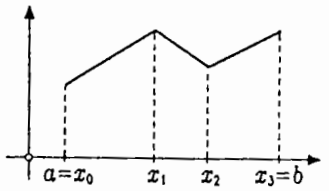
Hierbei ist $C^{-1}[a,b]$ der Raum der auf $[a,b]$ stückweise stetigen Funktionen. Die Menge aller Polynom-Splines vom Grade ℓ zur Zerlegung Δ bezeichnen wir mit $S_\ell(\Delta)$. Fortan wird schlechthin von Splines gesprochen.

(17.2) Beispiele

(1) Lineare Splines: Sind $(n+1)$ Punkte

$$(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$$

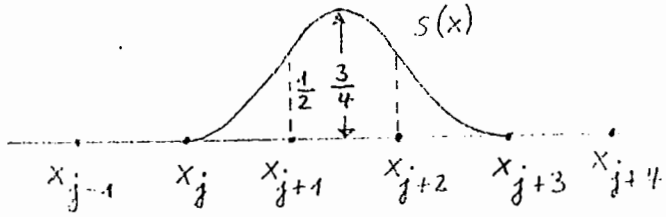
gegeben, so stellt der Polygonzug durch diese Punkte einen Spline $s \in S_1(\Delta)$ dar.



(2) Quadratische Splines: Bei äquidistanten Knoten $x_j = a + jh$, $j = 0, \ldots, n$, ist

$$s(x) = \frac{1}{2h^2} \begin{cases} (x - x_j)^2 & , x_j \le x < x_{j+1} \\ h^2 + 2h(x - x_{j+1}) - 2(x - x_{j+1})^2 & , x_{j+1} \le x < x_{j+2} \\ (x_{j+3} - x)^2 & , x_{j+2} \le x < x_{j+3} \end{cases}$$

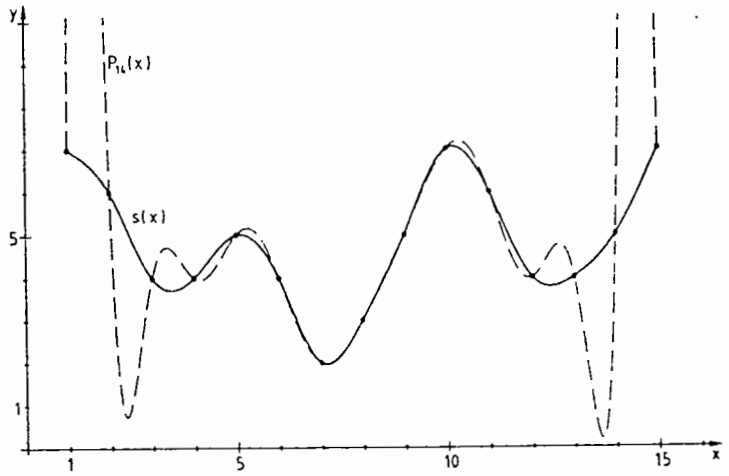
ein quadratischer B-Splinie $s \in S_2(\Delta)$.



(3) Kubische Splines: Zu 15 äquidistanten Knoten $x_j = j$

	x_j	y_j	Konstante	M_j
0	1	7	–	0
1	2	6	6	−2.306437
2	3	4	−12	3.225748
3	4	4	− 6	1.403445
4	5	5	12	−2.839529
5	6	4	6	−2.045331
6	7	2	−18	5.020853
7	8	3	− 6	−0.038080
8	9	5	0	1.131468
9	10	7	18	−4.487791
10	11	6	6	−1.180305
11	12	4	−12	3.209010
12	13	4	− 6	0.344264
13	14	5	− 6	1.413934
14	15	7	–	0

erhält man mit den Methoden von Abschnitt 17.3, Fall (17.5)(a) den folgenden kubischen Spline $s \in S_3(\Delta)$:

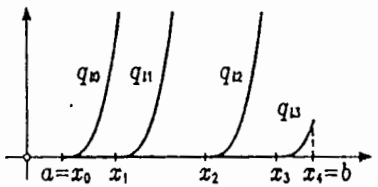


Im Vergleich zum stark oszillierenden Interpolationspolynom $P_{14}(x)$ zeigt sich die glättende Eigenschaft des kubischen Splines.

(4) Die Funktionen $q_{\ell j} : [a,b] \to \mathbb{R}$, $j = 0, 1, \ldots, n-1$,

$$q_{\ell j}(x) = (x - x_j)_+^\ell := \begin{cases} (x-x_j)^\ell, & x \geq x_j \\ 0, & x < x_j \end{cases}$$

sind Splines vom Grade ℓ zu Δ.



Man beachte, daß $q_{\ell j}$ keine Polynome in $[a,b]$ sind.

Der Spline-Raum $S_\ell(\Delta)$ ist ein linearer Teilraum von $C^{\ell-1}(\Delta)$. Der folgende Satz gibt Auskunft über eine

Basis von $S_\ell(\Delta)$.

(17.3) Satz: Die Menge $S_\ell(\Delta)$ ist ein linearer Raum der Dimension $n+\ell+1$. Die Elemente $p_i(x) = x^i$ $(i = 0, \ldots, \ell)$, $q_{\ell j}(x) = (x - x_j)_+^\ell$ $(j = 0, \ldots, n-1)$ bilden eine Basis von $S_\ell(\Delta)$.

Beweis: Wir haben zu zeigen, daß es zu $s \in S_\ell(\Delta)$ eine eindeutige Darstellung

$$s(x) = \sum_{i=0}^{\ell} a_i x^i + \sum_{j=0}^{n-1} b_j (x - x_j)_+^\ell, \quad x \in [a,b],$$

gibt. Dies erkennt man durch Induktion bzgl. des Index j der Zerlegung Δ. Im Intervall $I_1 = [x_0, x_1]$ ist s ein Polynom $s(x) = a_0 + a_1 x + \ldots + a_\ell x^\ell$, also gilt die Darstellung

$$s(x) = \sum_{i=0}^{\ell} a_i x^i + \sum_{j=1}^{k-1} b_j (x - x_j)_+^\ell$$

für $k = 1$ auf $I_k = [x_0, x_k]$. Wir betrachten nun

$$d(x) := s(x) - \sum_{i=0}^{\ell} a_i x^i - \sum_{j=1}^{k-1} b_j (x-x_j)_+^\ell .$$

Dann ist $d \in C^{\ell-1}(I_{k+1})$ und $d(x) = 0$ für $x \in I_k$. Außerdem ist $d \in \Pi_\ell$ in $[x_k, x_{k+1}]$. Also genügt d auf $[x_k, x_{k+1}]$ der Differentialgleichung

$$d^{(\ell+1)}(x) = 0, \quad x_k \le x \le x_{k+1}$$

$$d^{(i)}(x_k) = 0, \quad i = 0, \dots, \ell-1$$

Die Lösung dieser Anfangswertaufgabe ist

$$d(x) = b_k (x-x_k)_+^\ell , \quad x_k \le x , \quad b_k \in \mathbb{R}.$$

Damit ist die Beh. für den Index $k+1$ gezeigt. Für $k = n$ bilden daher die $n+\ell$ linear unabhängigen Elemente
$p_i(x) = x^i$ $(i = 0, \dots, \ell)$, $q_{\ell j}(x) = (x-x_j)_+^\ell$ $(j = 0, \dots, n-1)$ eine Basis von $S_\ell(\Delta)$. ∎

Wir untersuchen nun die folgende Interpolationsaufgabe: bestimme zu Stützwerten y_j $(j = 0, \dots, n)$ einen interpolierenden Spline $s \in S_\ell(\Delta)$ mit

$$s(x_j) = y_j , \quad j = 0, \dots, n .$$

Die Stützwerte y_j sind dabei als Funktionswerte $y_j = f(x_j)$ einer hinreichend glatten Funktion f aufzufassen. Wegen $\dim S_\ell(\Delta) = n+\ell$ können über die $n+1$ Interpolationsbedingungen hinaus noch

$$n + \ell - (n+1) = \ell - 1$$

freie Parameter bestimmt werden. Für ungerades $\ell = 2m-1$ ist dies eine gerade Anzahl $2m-2$ von Parametern, welche sich symmetrisch in den Randknoten x_0, x_n anordnen lassen. Wir beschränken uns auf den in der Praxis wichtigen Fall $m = 2$, $\ell = 3$, der kubischen Splines $s \in S_3(\Delta)$.

17.2 Kubische Splines

Sei $f \in C^2[a,b]$. Neben den Interpolationsbedingungen

$$s(x_j) = f(x_j), \quad j = 0, \dots, n \tag{17.4}$$

betrachten wir die drei folgenden Typen von Endbedingungen zur Festlegung der

2 freien Parameter:

(a) Natürliche Endbedingungen:

$$s''(a) = 0, \quad s''(b) = 0$$

(b) Hermite-Endbedingungen:

$$s'(a) = f'(a), \quad s'(b) = f'(b)$$

(17.5)

(c) Periodische Endbedingungen:

$$s^{(i)}(a) = s^{(i)}(b), \quad i = 0, 1, 2,$$

falls f periodisch mit

$$f^{(i)}(a) = f^{(i)}(b), \quad i = 0, 1, 2.$$

Wir können nun zeigen, daß die Interpolationsaufgabe (17.4), (17.5) eindeutig lösbar ist. Zusätzlich erfüllt der interpolierende Spline eine Minimum-Norm-Eigenschaft bzgl. der Norm in $C^2[a,b]$

(17.6) $$\|f\|_2 := \left(\int_a^b (f''(x))^2 dx \right)^{1/2}, \quad f \in C^2[a,b].$$

Für diese Norm gilt

$$\|f\|_2 = 0 \iff f''(x) = 0 \text{ für } x \in [a,b]$$
$$\iff f \text{ linear in } [a,b].$$

(17.7) Existenz, Eindeutigkeit und Extremaleigenschaft von Splines

Sei $f \in C^2[a,b]$. Dann gibt es genau einen Spline $s \in S_3(\Delta)$, der (17.4) und eine der Interpolationsbedingungen (17.5) erfüllt. Dieser interpolierende Spline genügt der Minimum-Norm-Bedingung

$$0 \le \|f - s\|_2^2 = \|f\|_2^2 - \|s\|_2^2.$$

Beweis: Mit einem Spline $s \in S_3(\Delta)$ berechnet man für die Abweichung $d(x) = f(x) - s(x)$:

$$\|f - s\|_2^2 = \int_a^b (f''(x) - s''(x))^2 dx$$
$$= \|f\|_2^2 - \|s\|_2^2 - 2 \int_a^b d''(x)\, s''(x)\, dx.$$

Da nur $s \in C^2[a,b]$ gilt, müssen wir für die partielle Integration aufspalten:

$$\int_a^b d''(x)\, s''(x)\, dx = \sum_{j=1}^n \int_{x_{j-1}}^{x_j} d''(x)\, s''(x)\, dx$$

$$= \sum_{j=1}^n \Big\{ [d'(x)\, s''(x) - d(x)\, s^{(3)}(x)]_{x_{j-1}}^{x_j} + \int_{x_{j-1}}^{x_j} d(x)\, s^{(4)}(x)\, dx \Big\}.$$

Nun ist $s^{(4)}(x) \equiv 0$ auf $[x_{j-1}, x_j]$. Die Interpolationsforderungen (17.4), (17.5) bewirken gerade, daß

$$\sum_{j=1}^{n} \left[d'(x) s''(x) - d(x) s^{(3)}(x) \right]_{x_{j-1}}^{x_j}$$

$$= \left[d'(x) s''(x) - d(x) s^{(3)}(x) \right]_a^b = 0$$

in den Fällen (17.5) (a)-(c). Damit ist die Minimum-Norm-Bedingung

$$0 \le \|f - s\|_2^2 = \|f\|_2^2 - \|s\|_2^2$$

gezeigt, mit der sich die Eindeutigkeit von s folgendermaßen ergibt:
ist $\tilde{s}$ ein weiterer interpolierender Spline, so kann man $f = \tilde{s}$ in der letzten Ungleichung nehmen:

$$0 \le \|\tilde{s} - s\|_2^2 = \|\tilde{s}\|_2^2 - \|s\|_2^2 .$$

Durch Vertauschung von s und $\tilde{s}$ sieht man $\|s - \tilde{s}\|_2 = 0$. Also ist $s - \tilde{s}$ linear. Wegen $s(x) = \tilde{s}(x) = 0$ für $x = a$ und $x = b$ muß dann $s = \tilde{s}$ gelten.

Zum Nachweis der Existenz greifen wir auf die Basis-Darstellung in Satz (17.3) zurück. Die Interpolationsforderungen stellen ein LGS in $(n+3)$ Unbekannten $a_0, \dots, a_3, b_1, \dots, b_{n-1}$ dar. In den Fällen (11.5) (a)-(c) wird das System homogen, wenn $f \equiv 0$ zu interpolieren ist. Dann ist aber $s = 0$ interpolierender Spline, und nach den obigen Überlegungen auch der einzige. ∎

Die Extremaleigenschaft des kubischen Splines

$$\int_a^b (s''(x))^2 dx \le \int_a^b (f''(x))^2 dx$$

erlaubt die folgende

geometrische und mechanische Interpretation:

Die Krümmung $k(x)$ einer Kurve $y = f(x)$ in der (x,y)-Ebene ist gegeben durch

$$k(x) = \frac{f''(x)}{(1 + (f'(x))^2)^{3/2}} .$$

Unter der Annahme $|f'(x)| \ll 1$ wird die mittlere Gesamtkrümmung

$$\|k\|_2^2 \approx \int_a^b (f''(x))^2 dx .$$

Der kubische Spline minimiert also die Norm $\|k\|_2$ unter allen interpolierenden Funktionen.

Das <u>Biegemoment</u> $M(x)$ eines homogenen, isotropen Stabes, dessen Biegelinie durch $y = f(x)$ beschrieben wird, ist $M(x) = c_1 k(x)$, $c_1 > 0$. Die Biege-Energie ist dann näherungsweise

$$E(f) = c_2 \int_a^b M(x)^2 dx \approx c_3 \int_a^b (f''(x))^2 dx .$$

Wird ein gebogener Stab durch Lager in "Interpolationspunkten" fixiert, so wird die minimale Biege-Energie durch einen kubischen Spline realisiert. Außerhalb von $[a, b]$, wo der Stab nicht fixiert ist, nimmt er die spannungsfreie "natürliche" Lage $s''(x) = 0$ an. In diesem Sinne sind die Endbedingungen $s''(a) = 0$, $s''(b) = 0$ in (17.5)(a) als "natürlich" zu verstehen.

17.3 Die Berechnung von Spline-Funktionen

Zu berechnen sei die Spline-Funktion $s(x)$ mit $s(x_j) = y_j$, $j = 0, \dots, n$, welche zusätzlich eine der Eigenschaften (a), (b), (c) hat. Wir setzen

$$h_j := x_j - x_{j-1}, \quad j = 1, \dots, n,$$

$$M_j := s''(x_j), \quad j = 0, \dots, n, \quad \text{(Momente)}.$$

Da s'' linear in $[x_{j-1}, x_j]$ ist, gilt

$$s''(x) = \frac{1}{h_j}\left(M_j(x - x_{j-1}) + M_{j-1}(x_j - x)\right), \quad x_{j-1} \le x \le x_j.$$

Durch Integration erhält man für $x \in [x_{j-1}, x_j]$

$$s'(x) = \frac{1}{2h_j}\left(M_j(x - x_{j-1})^2 - M_{j-1}(x_j - x)^2\right) + a_j,$$

$$s(x) = \frac{1}{6h_j}\left(M_j(x - x_{j-1})^3 + M_{j-1}(x_j - x)^3\right) + a_j(x - x_{j-1}) + b_j.$$

mit $a_j, b_j \in \mathbb{R}$. Für die Koeffizienten a_j, b_j erhält man aus $s(x_{j-1}) = y_{j-1}$, $s(x_j) = y_j$ die Gleichungen

$$M_{j-1}\frac{h_j^2}{6} + b_j = y_{j-1},$$

$$M_j\frac{h_j^2}{6} + a_j h_j + b_j = y_j,$$

und daraus

$$b_j = y_{j-1} - M_{j-1}\frac{h_j^2}{6},$$

$$a_j = \frac{y_j - y_{j-1}}{h_j} - \frac{h_j}{6}(M_j - M_{j-1}).$$

Damit ergibt sich

$$s'(x_{j-1}+) = \frac{1}{2h_j} M_j h_j^2 + a_j = \frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3} M_j + \frac{h_j}{6} M_{j-1}$$

$$s'(x_{j-1}+) = \frac{-1}{2h_j} M_{j-1} h_j^2 + a_j = \frac{y_j - y_{j-1}}{h_j} - \frac{h_j}{6} M_j + \frac{h_j}{3} M_{j-1}$$

$$s'(x_j+) = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1}.$$

Wegen $s'(x_j-) = s'(x_j+)$ folgt dann

(17.8) $$\boxed{\mu_j M_{j-1} + M_j + \lambda_j M_{j+1} = d_j, \quad j = 1, \ldots, n-1}$$

mit

$$\mu_j := \frac{h_j}{2(h_j + h_{j+1})},$$

$$\lambda_j := \frac{h_{j+1}}{2(h_j + h_{j+1})}, \qquad \mu_j + \lambda_j = \frac{1}{2},$$

$$d_j := \frac{3}{h_j + h_{j+1}} \left\{ \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right\}.$$

Fall (17.5)(a): Natürliche Endbedingungen.

Vorgegeben: Werte M_0, M_n

Die "natürliche" Bedingung $M_0 = M_n = 0$ ist darin enthalten. Dann stellt (17.8) ein LGS für $M_1, \dots, M_{n-1}$ dar:

$$\begin{pmatrix} 1 & \lambda_1 & & & \\ \mu_2 & 1 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-2} & 1 & \lambda_{n-2} \\ & & & \mu_{n-1} & 1 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} d_1 - \mu_1 M_0 \\ d_2 \\ \vdots \\ d_{n-2} \\ d_{n-1} - \lambda_{n-1} M_n \end{pmatrix}$$

Die Matrix ist tridiagonal und wegen $\mu_j + \lambda_j = 1/2$ diagonal-dominant, also LR-zerlegbar nach Satz (4.2). Das LGS kann daher mit Algorithmus (4.11) gelöst werden; vgl. Beispiel (17.2)(3). Bei äquidistanten Knoten x_j gilt $\mu_j = \lambda_j = 1/4$, also ist die Matrix symmetrisch und damit positiv definit.

Fall (17.5)(b): Hermite-Endbedingungen

Vorgegeben: $s'(a) = y_0'$, $s'(b) = y_n'$.

Aus der Darstellung von $s'(x)$ folgt

$$\frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 = \frac{y_1 - y_0}{h_1} - y_0'$$

$$\frac{h_n}{6} M_{n-1} + \frac{h_n}{3} M_n = y_n' - \frac{y_n - y_{n-1}}{h_n}.$$

Mit

$$\lambda_0 = \frac{1}{2}, \quad d_0 = \frac{3}{h_1}\left(\frac{y_1 - y_0}{h_1} - y_0'\right),$$

$$\mu_n = \frac{1}{2}, \quad d_n = \frac{3}{h_n}\left(y_n' - \frac{y_n - y_{n-1}}{h_n}\right)$$

erhalten wir ein LGS in den $(n+1)$ Unbekannten $M_0, \dots, M_n$:

$$\begin{pmatrix} 1 & \lambda_0 & & & \\ \mu_1 & 1 & \lambda_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 1 & \lambda_{n-1} \\ & & & \mu_n & 1 \end{pmatrix} \begin{pmatrix} M_0 \\ \\ \\ \\ M_n \end{pmatrix} = \begin{pmatrix} d_0 \\ \\ \\ \\ d_n \end{pmatrix}.$$

Die Matrix ist ebenfalls tridiagonal und diagonal-dominant, also LR-zerlegbar.

Beispiel: Für die nicht äquidistanten Knoten

j	x_j	y_j	M_j
0	0	0	0.022181
1	8.2	0.5	−0.000665
2	14.7	1.0	−0.010253
3	17.0	1.1	−0.006909
4	21.1	1.2	−0.000613
5	35.0	1.4	−0.000691
6	54.1	1.5	−0.000040
7	104	1.6	−0.000014
8	357	1.7	0.000004

$$y_0' = 0.00012566\,,$$

$$y_8' = 0.0001$$

liefert Algorithmus (4.11) die angegebenen Werte M_j.

Fall (17.5)(c): Periodische Endbedingungen

Hier ist

$$M_0 = M_n\,, \quad \text{da } s''(a) = s''(b)$$

$$y_0 = y_n\,, \quad \text{da } s(a) = s(b).$$

Die weitere Gleichung $s'(a) = s'(b)$ ergibt eine Beziehung

$$\mu_n M_{n-1} + M_n + \lambda_n M_{n+1} = d_n\,,$$

wenn man setzt $h_{n+1} = h_1$, $M_{n+1} = M_1$, $y_{n+1} = y_n$. Für $M_1, \dots, M_n$ ist das LGS zu

Lösen

$$\begin{pmatrix} 1 & \lambda_1 & & & \mu_1 \\ \mu_2 & 1 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_{n-1} & 1 & \lambda_{n-1} \\ \lambda_n & & & \mu_n & 1 \end{pmatrix} \begin{pmatrix} M_1 \\ \vdots \\ M_n \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}.$$

Die Matrix ist nicht mehr tridiagonal, aber immer noch diagonal-dominant, also LR-zerlegbar. Bei symmetrischen Matrizen kann jedoch das CHOLESKY-Verfahren (4.6) durch eine kleine zusätzliche Betrachtung modifiziert werden, auf die wir hier nicht eingehen.

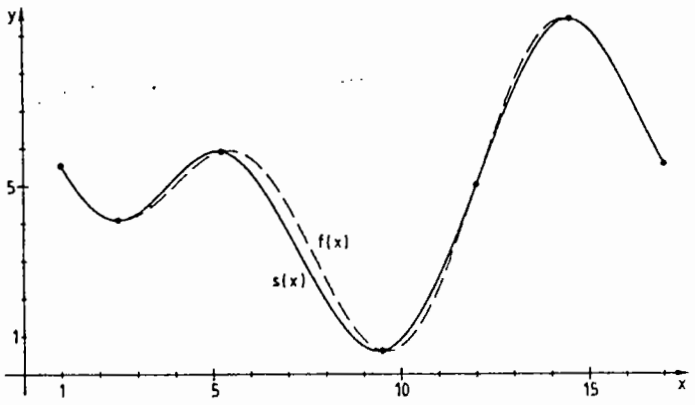
Beispiel: Die Funktion

$$f(x) = 2.5\,[\cos(2\pi x/16) - \sin(4\pi x/16)] + 5$$

mit der Periode $T = 16$ soll an den symmetrischen Knoten

j	x_j	y_j	$y''_j = M_j$
0	1.0	5.541932	0.819490
1	2.5	4.079227	1.555705
2	5.25	5.900182	-1.683242
3	9.5	0.611627	1.846591
4	12.0	5.000000	0.089237
5	14.5	9.388373	-2.203537
6	17.0	5.541932	0.819490

interpoliert werden. Die folgende Figur zeigt die gute Übereinstimmung:



17.4 Konvergenzeigenschaften

Entgegen dem Grenzverhalten von Interpolations-Polynomen konvergieren Spline-Funktionen gegen die Funktion, die sie interpolieren, bei Verfeinerung der Unterteilungen Δ. Sei

$$\Delta_m = \{ a = x_0^{(m)} < x_1^{(m)} < \dots < x_{n_m}^{(m)} = b \}$$

eine Folge von Unterteilungen des Intervalls $[a,b]$. Mit $\|\Delta_m\| := \max_j (x_{j+1}^{(m)} - x_j^{(m)})$ gilt:

(17.9) **Satz:** Sei $f \in C^4[a,b]$ mit $L = \|f^{(4)}\|_\infty$ und sei Δ_m eine Zerlegungsfolge von $[a,b]$ mit

$$\sup_{m,i} \frac{\|\Delta_m\|}{x_{j+1}^{(m)} - x_j^{(m)}} \le K < +\infty.$$

Seien s_m die zu f gehörigen Spline-F. mit

$$s_m(\xi) = f(\xi) \quad \text{für } \xi \in \Delta_m,$$

$$s_m'(x) = f'(x) \quad \text{für } x = a, b.$$

Dann gibt es von Δ_m unabhängige Konstanten C_i (≤ 2), so daß für $x \in [a,b]$ gilt

$$|f^{(i)}(x) - s_m^{(i)}(x)| \le C_i L K \|\Delta_m\|^{4-i}, \quad i = 0,1,2,3.$$

Beweis: Sei

$$\Delta := \Delta_m = \{ a = x_0 < \dots < x_n = b \}$$

eine feste Zerlegung. Für die Momente $M_j = s''(x_j)$ einer Spline-Funktion $s(x)$ mit $s'(x) = f'(x)$ für $x = x_0, x_n$ gilt nach (17.8) die Gleichung

$$A M = d$$

mit

$$\lambda_0 = \mu_n = \frac{1}{2},$$

$$\lambda_j = \frac{1}{2} \frac{h_{j+1}}{h_j + h_{j+1}}, \quad \mu_j = \frac{1}{2} - \lambda_j$$

$$(j = 1, \dots, n-1)$$

$$d_0 = \frac{3}{h_1} \left(\frac{y_1 - y_0}{h_1} - f'(x_0) \right),$$

$$d_n = \frac{3}{h_n} \left(f'(x_n) - \frac{y_n - y_{n-1}}{h_n} \right)$$

$$d_j = \frac{3}{h_j + h_{j+1}} \left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right)$$

$$(j = 1, \dots, n-1).$$

Man zeigt leicht (Übung), daß die Matrix A die Eigenschaft hat

(17.10) $$\|x\|_\infty \le 2 \|Ax\|_\infty \quad \text{für } x \in \mathbb{R}^n.$$

Für die Vektoren

$$F := (f''(x_0), f''(x_1), \dots, f''(x_n))^T,$$

$$r := d - AF = A(M - F)$$

wird zunächst gezeigt

(17.11) $$\|r\|_\infty \le \frac{3}{8} L \|\Delta\|^2.$$

Beweis von (17.11):

Die Taylor-Entwicklung um x_0 ergibt

$$\tau_0 = d_0 - f''(x_0) - \frac{1}{2} f''(x_1)$$

$$= \frac{3}{h_1}\left(\frac{f(x_1)-f(x_0)}{h_1} - f'(x_0)\right)$$

$$- f''(x_0) - \frac{1}{2} f''(x_1)$$

$$= \frac{h_1^2}{8} f^{(4)}(\tau_1) - \frac{h_1^2}{4} f^{(4)}(\tau_2)$$

mit $\tau_1, \tau_2 \in [x_0, x_1]$.

$$\Rightarrow \quad |\tau_0| \le \frac{3}{8} L \, \|\Delta\| .$$

Analog erhält man für

$$\tau_n = d_n - \frac{1}{2} f''(x_{n-1}) - f''(x_n)$$

die Abschätzung

$$|\tau_n| \le \frac{3}{8} L \, \|\Delta\| .$$

Entsprechend ergibt sich durch Taylor-Entwicklung um x_j für $j = 1, \dots, n-1$:

$$\tau_j = d_j - u_j f''(x_{j-1}) - f''(x_j) - \lambda_j f''(x_{j+1})$$

$$= \frac{1}{2(h_j + h_{j+1})}\left[\frac{h_{j+1}^3}{4} f^{(4)}(\tau_1)\right.$$

$$+ \frac{h_j^3}{4} f^{(4)}(\tau_2) - \frac{h_j^3}{4} f^{(4)}(\tau_3)$$

$$\left. - \frac{h_{j+1}^3}{2} f^{(4)}(\tau_4)\right]$$

mit $\tau_1, \dots, \tau_4 \in [x_{j-1}, x_{j+1}]$.

Also

$$|\tau_j| \le \frac{3}{8} L \, \frac{h_j^3 + h_{j+1}^3}{h_j + h_{j+1}} \le \frac{3}{8} L \, \|\Delta\|^2 .$$

Insgesamt gilt

$$\|\tau\|_\infty \le \frac{3}{8} L \|\Delta\|^2$$

und damit wegen $\tau = A(M-F)$ und (17.10)

$$\|M-F\|_\infty \le 2\|\tau\|_\infty \le \frac{3}{4} L \|\Delta\|^2. \tag{17.12}$$

Wir zeigen nun die Beh. des Satzes für $\underline{i=3}$:

Für $x \in [x_{j-1}, x_j]$ ist

$$s^{(3)}(x) - f^{(3)}(x)$$

$$= \frac{M_j - M_{j-1}}{h_j} - f^{(3)}(x)$$

$$= \frac{M_j - f''(x_j)}{h_j} - \frac{M_{j-1} - f''(x_{j-1})}{h_j}$$

$$+ \frac{f''(x_j) - f''(x) - (f''(x_{j-1}) - f''(x))}{h_j} - f^{(3)}(x)$$

Taylor-Entwicklung um x ergibt mit der Abschätzung (17.12)

$$|s^{(3)}(x) - f^{(3)}(x)|$$

$$\le \frac{3}{2} L \frac{\|\Delta\|^2}{h_j} + \frac{L}{2} \frac{\|\Delta\|^2}{h_j} \tag{17.13}$$

$$\le 2 L K \|\Delta\|,$$

da $\|\Delta\|/h_j \le K$ nach Vor..

Die Beh. für $\underline{i=2}$ folgt so:
für $x \in [a,b]$ gibt es $x_j = x_j(x)$ mit

$$|x_j(x) - x| \le \frac{1}{2} \|\Delta\|.$$

Es ist

$$f''(x) - s''(x) = f''(x_j(x)) - s''(x_j(x))$$

$$+ \int_{x_j(x)}^{x} (f^{(3)}(t) - s^{(3)}(t))\, dt$$

erhält man wegen (17.13) und $K \ge 1$:

$$|f''(x) - s''(x)| \le \frac{3}{4} L \|\Delta\|^2 + LK \|\Delta\|^2$$

$$= \frac{7}{4} LK \|\Delta\|^2.$$

Nun zeigen wir die Beh.

für $\underline{i=1}$:

es gilt

$$f(x_j) = s(x_j), \quad j = 0, \dots, n.$$

Außer $\xi_0 := a$, $\xi_{n+1} := b$ gibt es daher nach dem Satz von Rolle n Punkte $\xi_j \in (x_{j-1}, x_j)$ mit

$$f'(\xi_j) = s'(\xi_j), \quad j = 0, \dots, n+1.$$

Zu jedem $x \in [a, b]$ kann man also $\xi_j(x)$ wählen mit

$$|\xi_j(x) - x| \le \|\Delta\|.$$

Damit erhält man

$$|f'(x) - s'(x)| = \left| \int_{\xi_j(x)}^{x} (f''(t) - s''(t))\, dt \right|$$

$$\le \frac{7}{4} LK \|\Delta\|^3.$$

Schließlich ergibt sich die Beh. für $\underline{i=0}$:

$$|f(x) - s(x)| = \left| \int_{x_j(x)}^{x} (f'(t) - s'(t))\, dt \right|$$

$$\le \frac{7}{8} LK \|\Delta\|^4. \quad \blacksquare$$

Kapitel VI

§ 16

Eigenwertprobleme bei Matrizen

§ 21 Theoretische Grundlagen. Die Potenzmethode

Sei $A = (a_{ik})$ eine $n \times n$ Matrix mit $a_{ik} \in \mathbb{C}$. Eine Zahl $\lambda \in \mathbb{C}$ heißt Eigenwert (EW) von A, falls es einen Vektor $x \in \mathbb{C}^n$, $x \neq 0$, gibt mit

$$Ax = \lambda x, \quad \text{d.h.} \quad (A - \lambda I)x = 0.$$

Ein solcher Vektor x heißt (Rechts-) Eigenvektor (EV) von A zum Eigenwert λ.

Die Menge

$$L(\lambda) = \{ x \in \mathbb{C}^n \mid (A - \lambda I)x = 0 \}$$

bildet einen linearen Teilraum des $\mathbb{C}^n$ der Dimension

$$\rho(\lambda) = n - \operatorname{rang}(A - \lambda I)$$

(Vielfachheit des EW λ).

Jeder EW λ ist Nullstelle des charakteristischen Polynoms

$$\varphi(\lambda) = \det(A - \lambda I) = (-1)^n (\lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_0),$$

$$\varphi(\lambda) = (-1)^n (\lambda - \lambda_1)^{\sigma_1} \cdots (\lambda - \lambda_k)^{\sigma_k}.$$

Für die Zahlen $\sigma(\lambda_i) = \sigma_i$ gilt nach Linearer Algebra

$$1 \leq \rho(\lambda_i) \leq \sigma(\lambda_i) \leq n.$$

Beispiel: Für

$$J(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \quad n \times n, \ \lambda \in \mathbb{C},$$

gilt $\varphi(\mu) = (\lambda - \mu)^n$. λ ist einziger EW mit

$$\sigma(\lambda) = n,$$
$$\rho(\lambda) = 1, \quad \text{da } \operatorname{rang}(J(\lambda) - \lambda I) = n - 1.$$

Sei

$$A^H = \bar{A}^T.$$

Wegen

$$\det(A-\lambda I) = \det(A^T-\lambda I),$$

$$\overline{\det(A-\lambda I)} = \det(A^H-\overline{\lambda} I)$$

folgt: ist λ EW von A, so ist

λ EW von A^T,

$\overline{\lambda}$ EW von A^H.

Zwischen den zugehörigen EV x, y, z mit

$$Ax = \lambda x,$$
$$A^T y = \lambda y,$$
$$A^H z = \overline{\lambda} z$$

besteht die Beziehung $\overline{y} = z$, $y^T = z^H$ und

$$z^H A = \lambda z^H,$$

d.h. z^H, y^T sind <u>Links-Eigenvektoren</u>.

Sei T eine reguläre $n \times n$ Matrix. Die Transformation

$$B := T^{-1} A T$$

heißt <u>Ähnlichkeits-Transformation</u>.

Aus

$$Ax = \lambda x$$

folgt

$$By = \lambda y, \quad y := T^{-1} x,$$

$$\begin{aligned}\det(B-\lambda I) &= \det(T^{-1}(A-\lambda I)T) \\ &= \det(T^{-1}) \det(A-\lambda I) \det T \\ &= \det(A-\lambda I).\end{aligned}$$

Also haben B und A die gleichen EW und das gleiche charakteristische Polynom. Ebenso bleiben die Zahlen $S(\lambda), G(\lambda)$ erhalten.

Bei den wichtigsten Verfahren zur Berechnung von EW und EV einer Matrix A werden zunächst eine Reihe von Ähnlichkeits-Transformationen vorgenommen

$$\text{(21.1)} \quad \begin{aligned} A_1 &:= A \\ A_{k+1} &:= T_k^{-1} A_k T_k, \quad k = 1, 2, \ldots, \end{aligned}$$

um die Matrix A schrittweise in eine Matrix einfacherer Gestalt zu transformieren, deren EW und EV man leichter bestimmen kann.

Wegen der Empfindlichkeit der Nullstellen eines Polynoms gegenüber Störungen der Koeffizienten (vgl §12) muß das charakteristische Polynom als Hilfsmittel zur Berechnung von EW i.a. vermieden werden.

Eine wichtige Klasse von Matrizen sind die <u>normalen</u> Matrizen: eine $n \times n$ Matrix heißt normal, wenn gilt

$$A^H A = A A^H .$$

(21.2) <u>Satz:</u> Eine $n \times n$ Matrix A ist genau dann normal, wenn es eine unitäre Matrix U gibt mit

$$U^{-1} A U = U^H A U = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} .$$

Normale Matrizen sind diagonalisierbar und besitzen n linear unabhängige zueinander orthogonale Eigenvektoren x_i, $i = 1, \dots, n$, $A x_i = \lambda_i x_i$, nämlich die Spalten der Matrix $U = (x_1, \dots, x_n)$.

Insbesondere sind hermitesche und symmetrische Matrizen normal. In der Praxis stellen sich folgende Eigenwert-Aufgaben:

(1) Bestimme den größten EW

$$\rho(A) = \max \{ |\lambda| \mid \lambda \text{ EW von } A \}$$
$$= \text{Spektralradius von } A .$$

z.B. ist die Iteration

$$x^{(k+1)} = A x^{(k)} + b$$

nach Satz (13.9) genau dann konvergent, wenn $\rho(A) < 1$.

(2) Bestimmung aller EW

(3) Bestimmung mehrerer EV

Das Eigenwert-Problem kann schlecht konditioniert sein:

Beispiel:

$$A_\varepsilon := \begin{pmatrix} 0 & \varepsilon \\ 1 & 0 \end{pmatrix}.$$

A_0 hat einen EW $\lambda = 0$ und einen EV

$$x_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Für $\varepsilon > 0$ hat A_ε zwei EW

$$\lambda_1(\varepsilon) = \sqrt{\varepsilon}, \quad \lambda_2(\varepsilon) = -\sqrt{\varepsilon}$$

und zwei reelle EV

$$x_1(\varepsilon) = \begin{pmatrix} \sqrt{\varepsilon} \\ 1 \end{pmatrix}, \quad x_2(\varepsilon) = \begin{pmatrix} -\sqrt{\varepsilon} \\ 1 \end{pmatrix}.$$

Es gilt $\Delta A = A_\varepsilon - A_0 = O(\varepsilon)$, aber

$$\Delta\lambda_1 = \lambda_1(\varepsilon) - \lambda_1(0) = O(\sqrt{\varepsilon}),$$
$$\Delta x_1 = x_1(\varepsilon) - x_1(0) = O(\sqrt{\varepsilon}).$$

Für ε klein ist $|\Delta\lambda| / \|\Delta A\| = O(\varepsilon^{-1/2})$ sehr groß. ∎

Da das Eigenwertproblem für allgemeine Matrizen schlecht konditioniert sein kann, beschränken wir uns auf zwei Klassen von Matrizen, für die gute Algorithmen existieren:

(1) Normale Matrizen mit n verschiedenen EW $\lambda_1, \ldots, \lambda_n$.

(2) Hermitesche Matrizen A, d.h. $A^H = A$.

In der Praxis kommen überwiegend reelle Matrizen vor; daher wird im folgenden A als reell angenommen.

Praktische Anwendungen:

Stabilitätsprobleme bei gewöhnlichen und partiellen DGL.
Schwingungen und Gleichgewichtspunkte (Mechanik, VWL-BWL).

Das einfachste Verfahren zur Berechnung der EW λ_i von A und der EV x_i ist die Potenzmethode. Ausgehend von einem Vektor $x^{(0)}$ wird die Folge von Vektoren

$$x^{(k+1)} = Ax^{(k)} = A^{k+1}x^{(0)}, \quad k = 0, 1, \ldots$$

gebildet. Im folgenden beschränken wir uns auf den Fall

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| .$$

Dann hat A n lin. unabh. EV $x_1, \ldots, x_n$, und es gilt

$$x^{(0)} = \sum_{i=1}^{n} c_i x_i ,$$

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

$$= \lambda_1^k (c_1 x_1 + r_k) ,$$

$$r_k = \sum_{i=2}^{n} c_i \left(\frac{\lambda_i}{\lambda_1}\right)^k x_i = O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right) .$$

Wir wollen nun λ_1 und x_1 berechnen und setzen voraus

$$c_1 \neq 0 ,$$

$$x_{1,j} \neq 0 \quad \text{für ein } j \in \{1, \ldots, n\}$$

(j-te Komponente von x_1).

Damit erhält man

$$x_j^{(k+1)} / x_j^{(k)} = \lambda_1 \frac{(c_1 x_1 + r_{k+1})_j}{(c_1 x_1 + r_k)_j}$$

$$= \lambda_1 + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right) ,$$

$$x^{(k)} / x_j^{(k)} = \frac{x_1}{x_{1,j}} + O\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right) .$$

Beispiel:

$$A = \begin{pmatrix} 90 & 231 & 70 \\ 110 & 336 & 110 \\ 70 & 231 & 90 \end{pmatrix}$$

$$x^{(0)}, x^{(1)}, x^{(2)} = \begin{matrix} 1 & 391 & 190756 \\ 1 & 556 & 272836 \\ 1 & 391 & 190756 \end{matrix}$$

Für $j=2$ ergibt sich

$$x_2^{(1)}/x_2^{(0)} = 556, \quad x_2^{(2)}/x_2^{(1)} = 490.71$$

und die normierten Vektoren $x^{(k)}/x_2^{(k)}$ lauten

1	0.703237	0.699160
1	1	1
1	0.703237	0.699160

Die exakten Werte sind

$$\lambda_1 = 490, \quad \lambda_2 = 20, \quad x_1 = \begin{pmatrix} 0.7 \\ 1 \\ 0.7 \end{pmatrix}.$$

Aus dem kleinen Verhältnis $\lambda_2/\lambda_1 = 0.04$ erklärt sich die schnelle Konvergenz.

Zur Berechnung der weiteren EW bildet man die Matrix

$$T = (A - \mu I)^{-1}.$$

Diese hat die Eigenwerte $(\lambda_i - \mu)^{-1}$ mit

den EV x_i. Zur Berechnung von λ_2 wählt man μ so, daß

$$|\lambda_2 - \mu| < |\lambda_i - \mu|, \quad i \neq 2.$$

Dann ist $(\lambda_2 - \mu)^{-1}$ betragsgrößter EW von T. Diesen kann man nach der Potenzmethode berechnen. Zur Bildung von

$$x^{(k+1)} = T x^{(k)},$$

$$(A - \mu) x^{(k+1)} = x^{(k)}$$

muß man bei jedem Schritt ein LGS mit ein und derselben Matrix lösen. Man braucht also die LR-Zerlegung nur einmal durchzuführen. Dieses Verfahren heißt "inverse Potenzmethode" oder Wielandt-Iteration.

Bem.: (1) Man kann zeigen, daß die Potenzmethode konvergiert im Falle eines betragsgrößten EW

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

(2) Eine Variante der Potenzmethode, mit der alle λ_i gleichzeitig berechnet werden können, ist das LR-Verfahren von Rutishauser; dies ist ein Vorläufer des QR-Verfahrens in § 23.

§ 22 Transformations-Methoden

Man bilde eine Sequenz von Ähnlichkeits-Transformationen gemäß (21.1)

$$A = A_1 \to A_2 \to \dots \to A_{m+1},$$

$$(22.1)\quad A_{k+1} = T_k^{-1} A_k T_k, \quad k = 1, 2, \dots,$$

$$B := A_{m+1} = T^{-1} A T, \quad T := T_1 T_2 \dots T_m$$

mit dem Ziel, daß

(1) B möglichst einfache Gestalt hat,

(2) das Eigenwert-Problem für B nicht schlechter konditioniert ist als das für A.

Zu (1): B ist entweder Hessenberg-Matrix

$$B = \begin{pmatrix} * & \cdots & * \\ * & \ddots & \vdots \\ 0 & * & * \end{pmatrix} = (b_{ij}),\ b_{ij} = 0 \text{ für } i > j+1,$$

oder symmetrische Tridiagonalmatrix (falls A symmetrisch)

$$B = \begin{pmatrix} \delta_1 & \gamma_2 & & 0 \\ \gamma_2 & \delta_2 & \gamma_3 & \\ & \ddots & \ddots & \gamma_n \\ 0 & & \gamma_n & \delta_n \end{pmatrix}.$$

Zu (2): $B = T^{-1} A T$,

$B + \Delta B = T^{-1}(A + \Delta A) T$, $\Delta A = T \Delta B T^{-1}$

$$\Rightarrow \|B\| \le \operatorname{cond}(T) \|A\|,$$
$$\|\Delta A\| \le \operatorname{cond}(T) \|\Delta B\|,$$
$$\frac{\|\Delta A\|}{\|A\|} \le (\operatorname{cond}(T))^2 \frac{\|\Delta B\|}{\|B\|},$$
$$\operatorname{cond}(T) = \operatorname{cond}(T_1 \cdots T_m) \le \operatorname{cond}(T_1) \cdots \operatorname{cond}(T_m).$$

Daher muß man T_k so wählen, daß $\operatorname{cond}(T_k)$ nicht zu groß wird. Dies ist z.B. der Fall bei Eliminationsmatrizen

$$(22.2)\quad L_k = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & 1 & \\ & \ell_{k+1,k} & 1 & \\ & \vdots & & \ddots \\ & \ell_{n,k} & & 1 \end{pmatrix}, \quad |\ell_{ik}| \le 1,$$
$$\operatorname{cond}_\infty(L_k) \le 4,$$

oder für Householder-Matrizen

$$T_k = I - 2 w_k w_k^T, \quad w_k^T w_k = 1,$$
$$\operatorname{cond}_2(T_k) = 1.$$

Auslassen

22.1 Transformation auf Hessenberg-form

Eine beliebige $n \times n$ Matrix A soll mittels der Sequenz

$$A = A_1 \to A_2 \to \cdots \to A_{n-1} = B,$$
$$A_{k+1} = T_k^{-1} A_k T_k$$

auf eine Hessenberg-Matrix B transformiert werden. Sei

$$P_{\tau s} = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 0 & \cdots 1 & \cdots \tau \\ & & \vdots & & \\ & & 1 & \cdots 0 & \cdots s \\ 0 & & & & \\ & & & 1 & \\ & & & & \ddots \\ & & & & 1 \end{pmatrix}$$

eine Permutations-Matrix. Es gilt $P_{\tau s} = P_{\tau s}^{-1}$ und

$P_{\tau s} A$: vertauscht Zeilen τ und s von A

$A P_{\tau s}$: " Spalten " .

Mit der Eliminations-Matrix L_k in (22.2) wird der

Übergang $A_k \to A_{k+1}$:

$$A_k = \begin{pmatrix} * & \cdots & * & * & \cdots & * \\ * & \ddots & \vdots & \vdots & & \vdots \\ & \ddots & * & * & & \vdots \\ & 0 & & * & & \vdots \\ & & & \vdots & & \vdots \\ & & & * & & * \end{pmatrix} = \begin{pmatrix} B_k & & * \\ & & \\ & a & \tilde{A}_k \end{pmatrix}$$

$k-1$ k $\qquad$ k

Ziel:

$$a = \begin{pmatrix} \alpha_{k+1,k} \\ \vdots \\ \alpha_{nk} \end{pmatrix} \longrightarrow \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Bestimme $\tau \geq k+1$ mit

$$|\alpha_{\tau k}| = \max_{k+1 \leq j \leq n} |\alpha_{jk}| .$$

Fall (a): $\alpha_{\tau k} = 0$: setze $T_k = T_{k-1}$.

Fall (b): $\alpha_{\tau k} \neq 0$: vertausche Zeile und Spalte τ und $k+1$ in A_k, d.h. setze

$$A' := P_{\tau, k+1}^{-1} A_k P_{\tau, k+1} = (\alpha'_{ij}).$$

Definiere Eliminations-Matrix L_{k+1} mit

$$\ell_{j,k+1} = \frac{\alpha'_{jk}}{\alpha'_{k+1,k}} , \quad j = k+1, \ldots, n, \quad |\ell_{j,k+1}| \leq 1 ,$$

und setze

$$A_{k+1} = L_{k+1}^{-1} A' L_{k+1} = T_k^{-1} A_k T_k , \quad T_k := P_{\tau, k+1} L_{k+1} ,$$

$$B = A_{n-1} .$$

Die Methode ist numerisch gutartig und benötigt

$$\frac{5}{6}n^3 + O(n^2) \quad \text{Operationen.}$$

Zur Berechnung der Nullstellen des charakteristischen Polynoms

$$p(\lambda) = \det(B - \lambda I)$$

einer Hessenberg-Matrix B kann man das Newton-Verfahren (9.1) anwenden. Sei o.E. $b_{i+1,i} \neq 0$, d.h. B ist unzerlegbar.

Sei λ fest gewählt. Bestimme $\alpha, x_1, \ldots, x_{n-1}$ so, daß

$$x = (x_1, \ldots, x_{n-1}, x_n)^T, \; x_n := 1$$

Lösung von

$$(B - \lambda I)\, x = \alpha e_1$$

ist, d.h.

$$\begin{aligned}
&(b_{11} - \lambda)\, x_1 + b_{12} x_2 + \ldots + b_{1n} x_n = \alpha \\
&b_{21} x_1 + (b_{22} - \lambda) x_2 + \ldots + b_{2n} x_n = 0 \qquad (22.3) \\
&\qquad\qquad \vdots \\
&b_{n,n-1} x_{n-1} + (b_{nn} - \lambda) x_n = 0 .
\end{aligned}$$

Hieraus kann man rekursiv $x_{n-1}, x_{n-2}, \ldots, x_1, \alpha$ berechnen. Nach der Cramer'schen Regel gilt

$$1 = x_n = \frac{\det \begin{pmatrix} b_{11}-\lambda & b_{12} & & \alpha \\ b_{21} & b_{22}-\lambda & & 0 \\ & \ddots & \ddots & \vdots \\ & & b_{n,n-1} & 0 \end{pmatrix}}{p(\lambda)} = \alpha \frac{(-1)^{n-1} b_{21} \cdots b_{n,n-1}}{p(\lambda)}$$

$$\Rightarrow \quad \alpha = \alpha(\lambda) = \text{const.}\; p(\lambda).$$

Durch Differentation der Funktionen $x_i = x_i(\lambda)$ in (22.3) erhält man

$$\begin{aligned}
&(b_{11} - \lambda) x_1'(\lambda) - x_1(\lambda) + b_{12} x_2'(\lambda) + \\
&\qquad + \ldots + b_{1,n-1} x_{n-1}'(\lambda) = \alpha'(\lambda) \\
&\qquad\qquad \vdots \\
&b_{n,n-1} x_{n-1}'(\lambda) - x_n = 0 .
\end{aligned}$$

Damit kann man $\alpha'(\lambda)$ berechnen und das Newton-Verfahren anwenden wegen

$$p(\lambda)/p'(\lambda) = \alpha(\lambda)/\alpha'(\lambda).$$

(Methode von Hyman)

22.2 Transformation einer symmetrischen Matrix auf Tridiagonalgestalt

Sei A symmetrisch, $A_1 = A = A^T$.
In der Sequenz (22.1) wähle man Householder-Matrizen (vgl. §6)

$$T_k = I - 2 w_k w_k^T = I - \beta_k u_k u_k^T, \quad \|w_k\|_2 = 1,$$

$$T_k^T = T_k^{-1} = T_k .$$

Dann ist

$$A_{k+1} = T_k^{-1} A_k T_k \quad \text{symmetrisch.}$$

Übergang $A_k \to A_{k+1}$:

$$A_k = \left(\begin{array}{c|c|c} J_k & 0 & \\ \hline & a_k^T & \\ \hline 0 & a_k & \widetilde{A}_k \end{array}\right), \quad J_k = \begin{pmatrix} \delta_1 & \gamma_2 & 0 \\ \gamma_2 & \ddots & \ddots \\ & \ddots & \gamma_k \\ 0 & \gamma_k & \delta_k \end{pmatrix},$$

$$a_k = \begin{pmatrix} \alpha_{k+1,k} \\ \vdots \\ \alpha_{n,k} \end{pmatrix}.$$

Nach (6.2) gibt es eine $(n-k)\times(n-k)$ Householder-Matrix

$$\widetilde{T}_k = I - \beta u u^T, \quad u \in \mathbb{R}^{n-k},$$

$$\widetilde{T}_k a_k = c e_1 \in \mathbb{R}^{n-k}, \quad c \in \mathbb{R}.$$

Die $n \times n$ Matrix

$$T_k = \left(\begin{array}{c|c} I & 0 \\ \hline 0 & \widetilde{T}_k \end{array}\right)$$

ist orthogonal und es gilt mit $\gamma_{k+1} := c$

$$A_{k+1} = T_k^{-1} A_k T_k = T_k A_k T_k = \left(\begin{array}{c|c|c} J_k & 0 & \\ \hline & \gamma_{k+1} & 0 \dots 0 \\ \hline 0 & 0 \\ \vdots & \vdots & \widetilde{T}_k \widetilde{A}_k \widetilde{T}_k \\ 0 & 0 & \end{array}\right).$$

Mit

$$p := \beta \widetilde{A}_k u, \quad q := p - \tfrac{1}{2}\beta (p^T u) u$$

gilt

$$\widetilde{T}_k \widetilde{A}_k \widetilde{T}_k = \widetilde{A}_k - u q^T - q u^T.$$

Damit ist $B = A_{n-1}$ tridiagonal.

statt Householder-Matrizen kann man auch ebene Rotationen (Givens-Rotation) benutzen; vgl. Stoer, § 6.5.2.

Eine naheliegende Methode zur Berechnung des EW einer Tridiagonal-Matrix

$$J = J_n = \begin{pmatrix} \delta_1 & \gamma_2 & & 0 \\ \gamma_2 & \ddots & \ddots & \\ & \ddots & \ddots & \gamma_n \\ 0 & & \gamma_n & \delta_n \end{pmatrix}$$

ist das Newton-Verfahren zur Berechnung der Nullstellen von

$$p_n(\lambda) = \det(J_n - \lambda I).$$

Für das Polynom

$$p_k(\lambda) := \det(J_k - \lambda I)$$

der k-ten Abschnitts-Matrix J_k von J_n gilt die Rekursion

$$p_0(\lambda) = 1, \quad p_1(\lambda) = \delta_1 - \lambda,$$

(22.4)
$$p_k(\lambda) = (\delta_k - \lambda)\, p_{k-1}(\lambda) - \gamma_k^2\, p_{k-2}(\lambda), \quad k = 2, \dots, n.$$

Sei o.E. $\gamma_k \neq 0$, $k = 2, \dots, n$, d.h. J_n sei unzerlegbar. Setze

$$q(\lambda) = \begin{pmatrix} q_0(\lambda) \\ \vdots \\ q_{n-1}(\lambda) \end{pmatrix} \in \mathbb{R}^n, \quad q_0(\lambda) = 1,$$

$$q_k(\lambda) = \frac{(-1)^k\, p_k(\lambda)}{\gamma_2 \cdots \gamma_{k+1}}, \quad k = 1, \dots, n, \quad \gamma_{n+1} = 1.$$

Dann ist (22.4) äquivalent zu

(22.5)
$$(J_n - \lambda I)\, q(\lambda) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -q_n(\lambda) \end{pmatrix}.$$

Für die EW λ_k von J_n gilt

$$p_n(\lambda_k) = 0 \Rightarrow q_n(\lambda_k) = 0,$$

also

$$(J_n - \lambda_k I)\, q(\lambda_k) = 0, \quad q(\lambda_k) \neq 0.$$

Daher ist $q(\lambda_k)$ EV zum EW λ_k. Die Differentation von (22.5) ergibt

$$-q(\lambda) + (J_n - \lambda I)\, q'(\lambda) = \begin{pmatrix} 0 \\ 0 \\ -q_n'(\lambda) \end{pmatrix}.$$

Die Multiplikation mit $q(\lambda_k)^T$ liefert für $\lambda = \lambda_k$ wegen $q(\lambda_k)^T (J_n - \lambda_k I) = 0$:

$$0 < q(\lambda_k)^T q(\lambda_k) = q_{n-1}(\lambda_k)\, q_n'(\lambda_k)$$

$$= -\frac{p_{n-1}(\lambda_k)\, p_n'(\lambda_k)}{\gamma_2^2 \cdots \gamma_n^2}.$$

Also ist $p_n'(\lambda_k) \neq 0$ und daher λ_k eine einfache Nullstelle.

(22.6) Satz: Sei $\gamma_k \neq 0$, $k = 2, \ldots, n$. Dann besitzt J_n n verschiedene reelle EW $\lambda_n < \ldots < \lambda_2 < \lambda_1$.

Die Ableitung $p_n'(\lambda)$ kann durch Differentation von (22.4) berechnet werden:

$$p_0'(\lambda) = 0, \quad p_1'(\lambda) = -1,$$

$$p_k'(\lambda) = -p_{k-1}(\lambda) + (\delta_k - \lambda)\, p_{k-1}'(\lambda) - \gamma_k^2\, p_{k-2}'(\lambda),$$

Wegen Satz (22.6) kann die Variante von Maehly (12.8) zur Berechnung sämtlicher Nullstellen λ_k benutzt werden. Ein Startwert $\lambda^{(0)} \geq \lambda_1$ ergibt sich aus der Abschätzung

$$|\lambda_k| \leq \rho(J_n) \leq \|J_n\|_\infty$$

$$= \max_k \{|\gamma_k| + |\delta_k| + |\gamma_{k+1}|\},$$

wobei $\gamma_1 = \gamma_{n+1} = 0$.

Beispiel: Die Eigenwerte der symmetrischen Tridiagonalmatrix

$$A = \begin{pmatrix} 12 & 1 & & & \\ 1 & 9 & 1 & & \\ & 1 & 6 & 1 & \\ & & 1 & 3 & 1 \\ & & & 1 & 0 \end{pmatrix}$$

liegen symmetrisch zu $\lambda_3 = 6$, d.h. $\lambda_i + \lambda_{5-i} = 12$, $i = 1, 2$. Die Methode von Maehly liefert mit dem Startwert $\lambda^{(0)} = 13$ die Werte $\lambda_1 = 12.316876$, $\lambda_2 = 9.0161363$, $\lambda_4 = 12 - \lambda_1$, $\lambda_5 = 12 - \lambda_2$.

§ 23 Das QR-Verfahren

Das QR-Verfahren (FRANCIS, 1961) ist heutzutage das meistbenutzte Verfahren für Eigenwertprobleme. Es ist eine Verbesserung des LR-Verfahrens von RUTISHAUSER. Das QR-Verfahren wird hauptsächlich auf Hessenberg-Matrizen bzw. tridiagonale Matrizen angewandt (vgl. § 22).

Im QR-Verfahren erzeugt man eine Folge von Matrizen

(23.1)
$$
\begin{aligned}
&A_1 := A \\
&A_k = Q_k R_k , \\
&\qquad Q_k \text{ orthogonal, d.h. } Q_k^T Q_k = I , \\
&\qquad R_k \text{ obere Dreiecksmatrix,} \\
&A_{k+1} := R_k Q_k , \quad k = 1, 2, \ldots
\end{aligned}
$$

Die QR-Zerlegung $A_k = Q_k R_k$ existiert nach Satz (6.4). Zu beachten ist, daß die

QR-Zerlegung nicht eindeutig bestimmt ist.

(23.2) Lemma: Sei A regulär und $A = QR$ mit einer orthogonalen Matrix Q und einer oberen Dreiecksmatrix R. Dann sind Q und R bis auf die Multiplikation mit einer orthogonalen Diagonalmatrix D eindeutig bestimmt; d.h. sei

$$A = Q_1 R_1 = Q_2 R_2 ,$$

dann gibt es $D = \operatorname{diag}(\pm 1)$ mit

$$Q_1 = Q_2 D , \quad R_1 = D R_2 .$$

Beweis: Übung.

Bei dem QR-Verfahren (23.1) handelt es sich um eine Folge von Ähnlichkeitstransformationen:

(23.3) Lemma: Die Matrizen Q_k und R_k seien gemäß (23.1) definiert. Mit

$$P_k := Q_1 Q_2 \cdots Q_k, \quad U_k := R_k \cdots R_1.$$

gilt:

(a) A_{k+1} ist ähnlich zu A_k; $A_{k+1} = Q_k^{-1} A_k Q_k$,

(b) $A_{k+1} = P_k^{-1} A_1 P_k$,

(c) $A^k = P_k U_k$.

Beweis: zu (a): Aus den Beziehungen

$$A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k$$

folgt sofort

$$Q_k^{-1} A_k Q_k = R_k Q_k = A_{k+1}$$

Rekursiv erhält man daraus die Aussage (b):

$$A_{k+1} = Q_k^{-1} A_k Q_k = \cdots$$
$$= (Q_1 \cdots Q_k)^{-1} A_1 Q_1 \cdots Q_k = P_k^{-1} A P_k$$

zu (c): Nach (b) gilt

$$P_{k-1} A_k = A P_{k-1}$$

und dies ergibt die Zerlegung

$$P_k U_k = Q_1 \cdots Q_{k-1} Q_k R_k R_{k-1} \cdots R_1$$
$$= P_{k-1} A_k U_{k-1}$$
$$= A P_{k-1} U_{k-1} = \cdots$$
$$= A^{k-1} P_1 U_1 = A^{k-1} Q_1 R_1$$
$$= A^k. \qquad \blacksquare$$

(23.4) Konvergenzsatz für das QR-Verfahren

Die reelle Matrix A habe betragsmäßig verschiedene Eigenwerte

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0$$

und der Faktor R_k von A_k habe positive Diagonalelemente. Dann gibt es eine Permutation P der EW

$$\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = P \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}$$

mit

$$\lim_{k\to\infty} A_k = \begin{pmatrix} \mu_1 & & * \\ & \ddots & \\ 0 & & \mu_n \end{pmatrix}.$$

Beweis: Die EV von A seien $x_1, \dots, x_n$. Mit $X = (x_1, \dots, x_n)$ gilt dann

$$A = XDY, \quad Y = X^{-1}, \quad D = \operatorname{diag}(\lambda_i).$$

Die Matrix Y hat eine Dreieckszerlegung

$$PY = LU,$$

P: Permutationsmatrix,

L: untere Dreiecksmatrix mit $l_{ii} = 1$,

U: obere Dreiecksmatrix.

O.E. sei $P = I$. Die Matrix X besitze die QR-Zerlegung

$$X = QR.$$

Aus diesen Zerlegungen folgt

$$A^k = XD^kY = QRD^kLU$$

$$= QR\,(D^kLD^{-k})\,D^kU.$$

Mit der Vor. $0 < |\lambda_n| < \dots < |\lambda_1|$ folgt

$$D^kLD^{-k} = I + E_k,$$

$$(23.5)\quad (E_k)_{ij} = O\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right) \quad \text{für } i > j$$

$$\to 0 \quad \text{für } k \to \infty.$$

Damit erhält man

$$A^k = Q\,(I + RE_kR^{-1})\,RD^kU$$

$$=: Q\,(I + F_k)\,RD^kU,$$

$$F_k \to 0 \quad \text{für } k \to \infty.$$

Aus der QR-Zerlegung

$$I + F_k = \tilde{Q}_k\,\tilde{R}_k$$

folgt $\tilde{Q}_k \to I$, $\tilde{R}_k \to I$ für $k \to \infty$.

Also ist

$$A^k = (Q\tilde{Q}_k)(\tilde{R}_kRD^kU).$$

Sei weiter

$$D = |D|\,D_1, \quad D_1^2 = I,$$

$$U = D_2\,(D_2^{-1}U), \quad D_2^2 = I,$$

wobei $|D|$ und $D_2^{-1}U$ positive Diagonalelemente haben. Dann gilt

$$A^k = Q\tilde{Q}_k D_2 D_1^k \{ (D_2 D_1^k)^{-1} \tilde{R}_k R (D_2 D_1^k) |D|^k D_2^{-1} U \}.$$

Aus Lemma (23.2) und (23.3) ergibt sich mit der Vor., daß R_k positive Diagonalelemente hat:

$$P_k = Q\tilde{Q}_k D_2 D_1^k,$$

$$U_k = (D_2 D_1^k)^{-1} \tilde{R}_k R (D_2 D_1^k) |D|^k D_2^{-1} U.$$

Mit der Def. von P_k folgt

$$\begin{aligned} Q_k &= P_{k-1}^{-1} P_k \\ &= D_1^{-k-1} D_2^{-1} \tilde{Q}_{k-1}^{-1} Q^{-1} Q \tilde{Q}_k D_2 D_1^k \\ &= D_1 + D_1^{-k-1} D_2^{-1} (\tilde{Q}_{k-1}^{-1} \tilde{Q}_k - I) D_2 D_1^k \\ &\longrightarrow D_1 \text{ für } k \to \infty, \text{ da } \tilde{Q}_k \to I, D_1^2 = I. \end{aligned}$$

In der gleichen Weise folgt:

$$\begin{aligned} \operatorname{diag}(R_k) &= \operatorname{diag}(U_k U_{k-1}^{-1}) \\ &= \operatorname{diag}(\tilde{R}_k) \operatorname{diag}(\tilde{R}_{k-1}^{-1}) |D| \\ &\longrightarrow |D| \text{ für } k \to \infty, \\ &\quad \text{da } \tilde{R}_k \to I \text{ für } k \to \infty. \blacksquare \end{aligned}$$

Die Beweisanalyse mit (23.5) zeigt, daß das QR-Verfahren linear konvergiert. Eine Konvergenzverbesserung kann durch Shift-Techniken erreicht werden. Darunter versteht man die Spektralverschiebung

$$A_k - s_k I = Q_k R_k,$$

$$A_{k+1} = R_k Q_k + s_k I.$$

Folgende Shift-Techniken sind üblich:

(a) $s_k = a_{nn}^{(k)}$,

(b) s_k sei derjenige EW λ von

$$\begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{nn}^{(k)} \end{pmatrix},$$

für den $|a_{nn}^{(k)} - \lambda|$ am kleinsten ist. Bei konjugiert komplexen EW läßt sich die Rechnung rein reell in einem Doppelschritt $A_k \to A_{k+2}$ durchführen.

Für unzerlegbare symmetrische Tridiagonalmatrizen (vgl. Satz (22.6)) konvergiert das QR-Verfahren mit Shift-Technik <u>quadratisch</u>.

Numerisches Beispiel: vgl. Stoer II, S. 66-67.

Sind mehrere EW betragsgleich, d.h.

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_t| = |\lambda_{t+1}| = |\lambda_s| > |\lambda_{s+1}| > \ldots > |\lambda_n|,$$

so gilt eine modifizierte Konvergenzaussage:

$$A_k \xrightarrow{k} \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & * \\ & & \lambda_{t-1} & & \\ & & & \boxed{A_k^{s,t}} & \\ 0 & & & & \lambda_{s+1} \\ & & & & & \ddots \\ & & & & & & \lambda_n \end{pmatrix}$$

Die Elemente von $A_k^{s,t}$ konvergieren i.a. nicht, jedoch konvergieren die Eigenwerte von $A_k^{s,t}$ gegen $\lambda_t, \ldots, \lambda_s$. Dieser Fall kann bei reellen, nichtsymmetrischen Matrizen auftreten, die EW sind dann konjugiert zueinander.

§ 24 Eigenwert-Abschätzungen

Sei $\|\cdot\|$ eine Norm in $\mathbb{R}^n$ und

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

die zugeordnete Matrix-Norm. Eine triviale EW-Abschätzung ist

$$\rho(A) = \max\{|\lambda| \mid \lambda \text{ EW von } A\} \leq \|A\|.$$

Eine bessere EW-Abschätzung erhält man durch:

(24.1) Satz: Sei B $n \times n$ Matrix. Dann gilt für alle EW λ von A, die nicht EW von B sind

$$1 \leq \|(\lambda I - B)^{-1}(A-B)\| \leq \|(\lambda I - B)^{-1}\| \, \|A-B\|.$$

Beweis: Sei $Ax = \lambda x$

$\Rightarrow (A-B)x = (\lambda I - B)x$

$\Rightarrow (\lambda I - B)^{-1}(A-B)x = x$

$\Rightarrow \|(\lambda I - B)^{-1}(A-B)\| \geq 1.$ ∎

Folgerung: Wähle

$$B = A_D := \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}.$$

Für $\lambda \neq a_{ii}$, $i = 1, \ldots, n$, folgt

$$1 \leq \|(\lambda I - A_D)^{-1}(A - A_D)\|_\infty = \max_{1 \leq i \leq n} \frac{1}{|\lambda - a_{ii}|} \sum_{k \neq i} |a_{ik}|.$$

(24.2) Satz: (Gerschgorin)

(i) Die Vereinigung aller Kreisscheiben

$$K_i := \{\lambda \mid |\lambda - a_{ii}| \leq \sum_{k \neq i} |a_{ik}| =: r_i\}$$

enthält alle EW von A.

(ii) Ist die Vereinigung M_1 von k Kreisen K_i disjunkt von der Vereinigung M_2 der übrigen Kreise, so enthält M_1 genau k und M_2 genau $n-k$ EW von A.

Beweis: zu (ii): Setze

$$A = A_D + R, \quad A_t := A_D + tR, \quad 0 \le t \le 1,$$

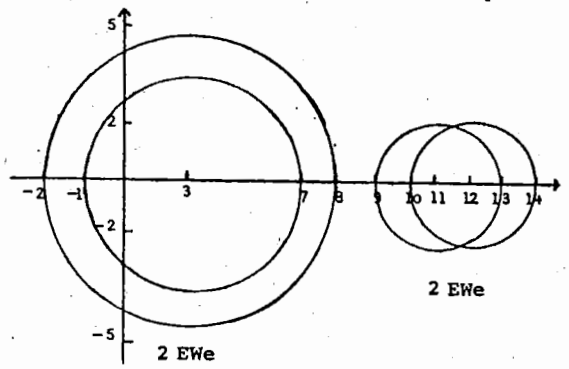
$$A_0 = A_D, \quad A_1 = A.$$

Die EW von A_t sind stetig bzgl. t und liegen in den Kreisen K_i.
Für $t=0$ ist die Beh. offenbar richtig; für $t=1$ folgt dann die Beh. aus Stetigkeitsgründen. ∎

Beispiel:

$$A = \begin{pmatrix} 3 & 2 & 1 & -2 \\ 1 & 11 & 0 & 1 \\ -1 & 0 & 12 & -1 \\ -3 & 1 & 0 & 3 \end{pmatrix}$$

$r_1 = 5$
$r_2 = 2$
$r_3 = 2$
$r_4 = 4$



Eine Verbesserung der Gerschgorin-Abschätzung erreicht man ggf. durch eine Skalierung:

$$A \to A' = D^{-1} A D,$$

$$D = \operatorname{diag}(d_1, \ldots, d_n), \quad d_i > 0,$$

$$K_i' = \{ \lambda \mid |\lambda - a_{ii}| \le \sum_{k \ne i} |a_{ik}| \frac{d_k}{d_i} =: \rho_i \}.$$

Die Verkleinerung von ρ_i ist ggf. durch Wahl von d_j möglich.

Der Satz von Gerschgorin ist ein Beispiel für einen sog. Lokalisierungssatz. Sätze dieser Art waren für die Praxis früher sehr wertvoll, da sie ohne Rechnung einen Überblick über die ungefähre Lage der EW geben. Heute sind die Rechenanlagen so leistungsfähig, daß die Lokalisierungssätze ihre praktische Bedeutung

weitgehend verloren haben (z.B. QR-Verfahren in §23).

Eine Norm $\|\cdot\|$ mit

$$\|\,|x|\,\| = \|x\| \quad \text{für alle } x \in \mathbb{R}^n$$

heißt <u>absolute Norm</u>. Z.B. sind $\|\cdot\|_\infty$, $\|\cdot\|_2$ absolute Normen. Für absolute Normen gilt

$$\|\operatorname{diag}(d_1,\dots,d_n)\| = \max_{1\le i\le n} |d_i|.$$

Aus (24.1) gewinnt man den folgenden <u>Störungssatz</u>:

(24.3) <u>Satz</u>: Seien A, B $n\times n$ Matrizen. Sei B diagonalisierbar mit $B = PDP^{-1}$, $D = \operatorname{diag}(\lambda_1(B),\dots,\lambda_n(B))$. Dann gibt es zu jedem EW $\lambda(A)$ von A einen EW $\lambda_i(B)$ von B mit

$$|\lambda(A) - \lambda_i(B)| \le \operatorname{cond}(P)\,\|A-B\|$$

für absolute Normen $\|\cdot\|$.

<u>Beweis</u>: Für $\lambda \neq \lambda_i(B)$, $i=1,\dots,n$, gilt

$$\begin{aligned}\|(\lambda I - B)^{-1}\| &= \|P(\lambda I - D)^{-1}P^{-1}\| \\ &\le \max_{1\le i\le n} \frac{1}{|\lambda-\lambda_i(B)|}\operatorname{cond}(P) \\ &= \frac{1}{\min_i |\lambda-\lambda_i(B)|}\operatorname{cond}(P).\end{aligned}$$

Die Beh. ergibt sich dann aus (24.1). ∎

Ist B <u>normal</u>, so ist $\operatorname{cond}_2(P)=1$ und (24.3) liefert die Abschätzung

$$|\lambda(A)-\lambda(B)| \le \|A-B\|_2 \tag{24.4}$$

mit einem EW $\lambda(B)$. Insbesondere ist das EW-Problem für <u>symmetrische Matrizen gut konditioniert</u>: für symmetrische $A, \Delta A$ gilt also

$$|\lambda(A+\Delta A)-\lambda(A)| \le \|\Delta A\|_2.$$

Bei nicht-symmetrischen Matrizen A erhält man bei einer Störung

$$A \to A + \varepsilon C \,, \quad C \; n \times n \text{ Matrix}$$

eine Abschätzung der Form

$$|\lambda(A+\varepsilon C) - \lambda(A)| \le K\,|\varepsilon|^{1/\nu}, \tag{24.5}$$

$\nu :=$ max. Dimension eines zu $\lambda(A)$ gehörenden Jordan-Blockes;

vgl. Stoer II, § 6.8; vgl. auch Beispiel mit $\nu = 2$ in § 21.

Für Diagonalmatrizen $D = \operatorname{diag}(d_i)$ gilt

$$\min_{x \neq 0} \frac{\|Dx\|_2}{\|x\|_2} = \min_i |d_i| \,.$$

Nun sei A <u>normal</u>. Also gibt es U orthogonal mit

$$A = U^T D U \,, \quad D = \operatorname{diag}(\lambda_i(A)) \,.$$

Für ein beliebiges Polynom $f(\lambda)$ folgt dann

$$f(A) = U^T f(D)\, U \,.$$

Wegen $\|Ux\|_2 = \|x\|_2$ hat man

$$\frac{\|f(A)x\|_2}{\|x\|_2} = \frac{\|U^T f(D) U x\|_2}{\|Ux\|_2} = \frac{\|f(D)Ux\|_2}{\|Ux\|_2}$$

$$\ge \min_{y \neq 0} \frac{\|f(D)y\|_2}{\|y\|_2} = \min_i |f(\lambda_i(A))|$$

Also gilt

(24.6) <u>Satz</u>: Sei A normal, $x \neq 0$, und $f(\lambda)$ ein Polynom. Dann gibt es einen EW $\lambda(A)$ mit

$$|f(\lambda(A))| \le \frac{\|f(A)x\|_2}{\|x\|_2} \,.$$

<u>Folgerung</u>: Sei $x \neq 0$ und f das lineare Polynom

$$f(\lambda) = \lambda - \frac{x^T A x}{x^T x}$$

↑ Rayleigh-Quotient

$$\Rightarrow \|f(A)x\|_2^2 = x^T \left(A^T - \frac{x^T A x}{x^T x} I\right) \cdot$$

$$\cdot \left(A - \frac{x^T A x}{x^T x} I\right) x$$

$$= x^T A^T A x - \frac{(x^T A^T x)(x^T A x)}{x^T x}.$$

Für symmetrisches A erhält man mit (24.6)

(24.7) Satz: (Bogolyubov – Krylov, Weinstein)
Ist A symmetrisch und $x \neq 0$, so gilt

$$\min_i \left| \lambda_i(A) - \frac{x^T A x}{x^T x} \right|$$

$$\leq \left(\frac{x^T A^2 x}{x^T x} - \left(\frac{x^T A x}{x^T x} \right)^2 \right)^{1/2}.$$

Statistische (Quantenmechnische) Deutung:

x : Zustände, o.E. $\|x\|_2 = 1$

$\lambda_i(A)$: Meßwerte

$x^T A x$: Erwartungswert von A bzgl. x,
$= \lambda$, falls x EV zu λ

$(x^T A^2 x - (x^T A x)^2)^{1/2}$
$= \|(A - x^T A x)x\|_2$: Unschärfe von A bzgl. x;
$= 0$, falls x EV

Die Abschätzung (24.7) besagt

|Meßwert - Erwartungswert| ≤ Unschärfe.

§ 18 Approximation in normierten Räumen

18.1 Funktionalanalytische Grundlagen

Sei V ein Vektorraum über $\mathbb{R}$ (oder $\mathbb{C}$).
Eine Norm für V ist eine Abbildung
$\|\cdot\| : V \to \mathbb{R}$ mit den Eigenschaften:

(a) $\|f\| > 0$ für alle $f \in V$, $f \neq 0$,

(b) $\|\alpha f\| = |\alpha| \, \|f\|$ für alle $f \in V$, $\alpha \in \mathbb{R}$ (oder $\mathbb{C}$),

(c) $\|f+g\| \leq \|f\| + \|g\|$ für alle $f, g \in V$.

Das Paar $(V, \|\cdot\|)$ heißt normierter Raum.
Eine Norm $\|\cdot\|$ heißt streng (strikt), wenn gilt

$$\|f+g\| = \|f\| + \|g\|$$

$\Rightarrow f$ und g linear abhängig.

Beispiel:

$$H = C[a,b] = \{ f : [a,b] \to \mathbb{R} \text{ (oder } \mathbb{C}) \text{ stetig} \}.$$

Für $1 \leq p < \infty$ ist

$$\|f\|_p := \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}$$

eine Norm für $C[a,b]$. Den wichtigsten Fall erhält man für $p \to \infty$

$$\|f\|_\infty = \lim_{p \to \infty} \|f\|_p = \max_{x \in [a,b]} |f(x)| .$$

Die Norm $\|\cdot\|_2$ ist streng (Cauchy-Schwarz'sche Ungleichung), die Norm $\|\cdot\|_\infty$ hingegen nicht. Letzteres sieht man an dem Beispiel $f \equiv 1$, $g = x$ in $C[0,1]$.

Eine Folge $\{f_n\} \subset V$ heißt CAUCHY-Folge, wenn es zu jedem $\varepsilon > 0$ ein $n(\varepsilon) \in \mathbb{N}$ gibt mit $\|f_k - f_\ell\| < \varepsilon$ für alle $k, \ell \geq n(\varepsilon)$. Konvergiert jede CAUCHY-Folge eines normierten Vektorraumes $(V, \|\cdot\|)$ gegen ein Element von V, so heißt V vollständig oder Banach-Raum. Jeder endlich-dimensionale normierte Vektorraum $(V, \|\cdot\|)$ ist ein Banach-Raum. Ein Beispiel eines unendlich-dimensionalen Banach-Raumes ist $(C[a,b], \|\cdot\|_\infty)$.

Diejenigen Räume, deren Norm durch ein inneres Produkt induziert wird, zeichnen sich durch besondere Eigenschaften aus. Eine Abbildung

$$(\cdot,\cdot) : V \times V \to \mathbb{R} \text{ (oder } \mathbb{C})$$

heißt inneres Produkt, wenn für alle $f, g, h \in V$ und $\alpha \in \mathbb{C}$ gilt

$(f+g, h) = (f,h) + (g,h)$	Linearität
$(\alpha f, g) = \alpha (f,g)$	Homogenität
$(f,g) = \overline{(g,f)}$	Symmetrie
$(f,f) > 0$ für $f \neq 0$	Positivität.

Dann wird durch

$$\|f\| = \sqrt{(f,f)}$$

eine Norm auf V erklärt.

Zur Nachprüfung der Dreiecksungleichung $\|f+g\| \leq \|f\| + \|g\|$ in (c) benötigt man die Schwarzsche Ungleichung:

$$|(f,g)| \leq \|f\| \, \|g\| \quad \text{für alle } f, g \in V.$$

Gleichheit gilt hier genau dann, wenn f, g linear abhängig sind. Die Abschätzung ergibt sich, wenn man auswertet

$$(\alpha f + g, \alpha f + g) \geq 0 \quad \text{mit } \alpha := -\frac{(g,f)}{(f,f)}.$$

Der normierte Vektorraum $(V, \|\cdot\|)$ heißt Prä-Hilbertraum. Ist darüberhinaus $(V, \|\cdot\|)$ vollständig, so heißt V Hilbertraum.

Prä-Hilberträume sind stets streng (strikt) normierte Räume. Gleichheit in der Dreiecksungleichung kann vermöge

$$\|f+g\|^2 = (f+g, f+g)$$
$$= \|f\|^2 + \|g\|^2 + (f,g) + (g,f)$$
$$\le \|f\|^2 + \|g\|^2 + 2\,|(f,g)|$$
$$\le (\|f\| + \|g\|)^2$$

nur eintreten, wenn $|(f,g)| = \|f\|\,\|g\|$, also wenn f, g linear abhängig sind.

Beispiele:

(a) Der Raum $(\mathbb{C}^n, \|\cdot\|_2)$ ist ein Hilbertraum, wobei $\|\cdot\|_2$ durch

$$(x,y) = \sum_{i=1}^{n} x_i \bar{y}_i \quad , \quad x, y \in \mathbb{C}^n$$

induziert wird.

(b) Sei $w : [a,b] \to \mathbb{R}$, $w(x) > 0$ für $a < x < b$, eine stetige Gewichtsfunktion. Auf $C[a,b]$ wird durch

$$(f,g) = \int_a^b f(x)\, g(x)\, w(x)\, dx, \quad f, g \in C[a,b],$$

ein inneres Produkt definiert mit der Norm

$$\|f\|_2 = \left(\int_a^b f(x)^2\, w(x)\, dx \right)^{1/2}.$$

18.2 Das allgemeine Approximationsproblem

Sei $(V, \|\cdot\|)$ ein normierter Raum und $T \subset V$ eine Teilmenge. Zu einem gegebenen Element $v \in V$ suchen wir eine beste Näherung (Proximum) $\tilde{u} \in T$ von v bzgl. T mit

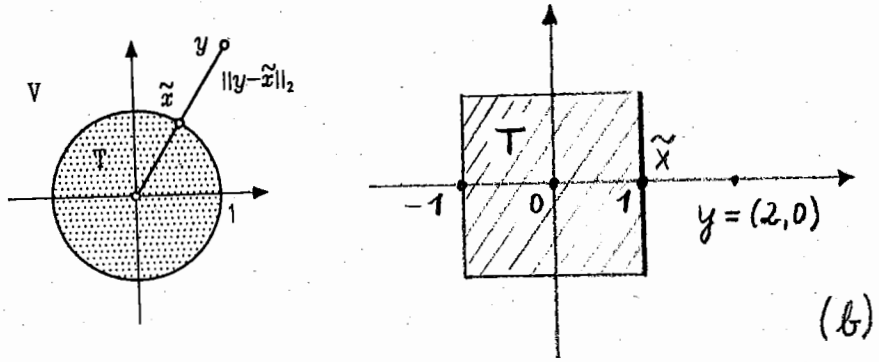
(18.1) $\quad \|v - \tilde{u}\| \le \|v - u\|$ für alle $u \in T$.

Die Existenz und Eindeutigkeit der besten Näherung $\tilde{u} \in T$ hängt wesentlich von Eigenschaften der Norm $\|\cdot\|$ und der Menge T ab.

Beispiele:

(a) Sei $V = \mathbb{R}^2$ mit der euklidischen Norm $\|\cdot\|_2$ und sei $T = \{x \in V \mid \|x\|_2 \le 1\}$. Zu jedem $y \in \mathbb{R}^2$ gibt es genau eine beste Näherung $\tilde{x} \in T$; vgl. Skizze (a).

(b) Sei $V = \mathbb{R}^2$ mit der Maximum-Norm $\|\cdot\|_\infty$ und sei $T = \{x \in V \mid \|x\|_\infty \le 1\}$.



Zu $y=(2,0)\in\mathbb{R}^2$ ist das Proximum $\tilde{x}\in T$ nicht eindeutig bestimmt. Der Abstand

$$\|x-y\|_\infty = \max\{|x_1-2|, |x_2|\}$$

wird nämlich minimal für alle Punkte der Kante $\{(1,x_2)\mid |x_2|\le 1\}$ von T.

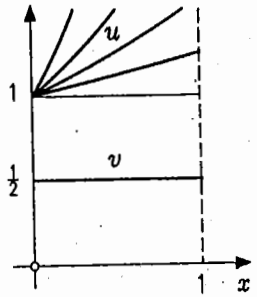
(c) In $(C[0,1], \|\cdot\|_\infty)$ sei

$$T=\{u\in V \mid u(x)=e^{\beta x},\ \beta>0\}.$$

Gefragt ist nach einem Proximum $\tilde{u}\in T$ an die konstante Funktion $v(x)=1/2$. Für $u(x)=e^{\beta x}$, $\beta>0$, ist

$$\|u-v\|_\infty = \max_{0\le x\le 1} |e^{\beta x}-\tfrac{1}{2}| = e^{\beta}-\tfrac{1}{2}$$

und daher wird das Minimum von keinem Element $\tilde{u}\in T$ angenommen; vgl. Skizze.



In Beispiel (a) ist die benutzte Norm $\|\cdot\|_2$ streng, während die Norm $\|\cdot\|_\infty$ in Beispiel (b) nicht streng ist. Die Existenz in Beispiel (c) ist nicht gesichert, da T nicht kompakt ist. Die nachfolgenden Existenz- und Eindeutigkeitsaussagen verdeutlichen diese Situation.

Zu $v\in V$ heißt

$$e_T(v) = \inf_{u\in T} \|v-u\| \tag{18.2}$$

der Minimalabstand von v zu T. Eine beste Näherung $\tilde{u} \in T$ an v erfüllt also $e_T(v) = \|v - \tilde{u}\|$. Nach Definition von $e_T(v)$ gibt es eine Minimalfolge

$$\{u_n\} \subset T, \quad e_T(v) = \lim_{n \to \infty} \|v - u_n\|.$$

(18.3) Hilfssatz:

(a) Jede Minimalfolge ist beschränkt.

(b) Jeder in T liegende Häufungspunkt einer Minimalfolge ist ein Proximum.

Beweis: zu (a): Es gilt

$$e_T(v) \le \|v - u_n\| \le e_T(v) + 1$$

für alle $n \ge n_0$. Also ist

$$\|u_n\| \le \|v - u_n\| + \|v\| \le e_T(v) + 1 + \|v\| =: K_2$$

für $n \ge n_0$. Dies ergibt die Abschätzung $\|u_n\| \le K := \max\{K_1, K_2\}$ für alle $n \in \mathbb{N}$ mit $K_1 = \max\{\|u_n\|, n = 1, \dots, n_0\}$.

zu (b): Die Teilfolge $\{u_{k(n)}\}$ konvergiere gegen $\tilde{u} \in T$. Dann folgt die Abschätzung $\|v - \tilde{u}\| \le \|v - u_k\| + \|u_k - \tilde{u}\|$ für alle k, also $\|v - \tilde{u}\| \le e_T(v)$ wegen $\|v - u_k\| \to e_T(v)$ und $\|u_k - \tilde{u}\| \to 0$. Also ist $\|v - \tilde{u}\| = e_T(v)$. ∎

Die Menge T heißt streng konvex, wenn für alle $u_1, u_2 \in T$:

$$\alpha u_1 + (1-\alpha) \in T \quad \text{für } 0 \le \alpha \le 1,$$
$$\alpha u_1 + (1-\alpha) \in \operatorname{int} T \quad \text{für } 0 < \alpha < 1.$$

Die Menge T in Beispiel (b) ist nicht streng konvex.

(18.4) 1. Existenz- und Eindeutigkeitssatz:
Sei $T \subset V$ eine kompakte Teilmenge. Dann gibt es zu jedem $v \in V$ ein Proximum $\tilde{u} \in T$. Ist außerdem T streng konvex, so ist $\tilde{u} \in T$ eindeutig bestimmt.

Beweis: Existenz: Eine Minimalfolge $\{u_n\} \subset T$ enthält wegen der Kompaktheit

von T eine gegen $\tilde{u} \in T$ konvergente Teilfolge. Nach (18.3)(b) ist $\tilde{u}$ ein Proximum.

Eindeutigkeit: Seien u_1, u_2 mit $u_1 \neq u_2$ Proxima in T an v. Dann gilt

$$\left\| \tfrac{1}{2}(u_1+u_2) - v \right\| \leq \tfrac{1}{2}\|u_1 - v\| + \tfrac{1}{2}\|u_2 - v\| = e_T(v).$$

Somit ist $\frac{1}{2}(u_1+u_2) \in T$ (T konvex) ein Proximum. Da T streng konvex ist, gibt es Werte $\alpha \in (0,1)$ mit

$$\tilde{u} := \tfrac{1}{2}(u_1+u_2) + \alpha\left(v - \tfrac{1}{2}(u_1+u_2)\right) \in T.$$

Für solche Werte α folgt ein Widerspruch:

$$\|\tilde{u} - v\| = \left\| \tfrac{1}{2}(1-\alpha)(u_1+u_2) - (1-\alpha)v \right\| = (1-\alpha)\, e_T(v) < e_T(v).$$

Also ist die Annahme $u_1 \neq u_2$ falsch. ∎

Neben der Kompaktheit von T ist in den Anwendungen vor allem der Fall wichtig, daß $T =: U$ ein endlich-dimensionaler linearer Unterraum von V ist.

(18.5) 2. Existenz- und Eindeutigkeitssatz

Sei U ein endlich-dimensionaler Unterraum von V. Dann gibt es zu jedem $v \in V$ ein Proximum $\tilde{u} \in U$. Ist außerdem V streng normiert, so ist $\tilde{u} \in U$ eindeutig bestimmt.

Beweis: Existenz: Nach Hilfssatz (18.3)(a) ist jede Minimalfolge beschränkt und besitzt daher einen Häufungspunkt $\tilde{u}$. Da U als endlich-dimensionaler Unterraum abgeschlossen ist, liegt $\tilde{u}$ in U und ist somit ein Proximum nach (18.3)(b).

Eindeutigkeit: Sei $v \in U$. Sind u_1, u_2 Proxima, so folgt wie im Beweis von (18.4):

$$\left\| \tfrac{1}{2}(u_1+u_2) - v \right\| = e_U(v).$$

Demnach ist

$$\|(v-u_1) + (v-u_2)\| = \|v-u_1\| + \|v-u_2\|,$$

also aufgrund der strengen Norm

$$v - u_1 = \alpha (v - u_2) \quad \text{für ein } \alpha \in \mathbb{R},$$

d.h.

$$(1-\alpha) v = u_1 - \alpha u_2 \in U.$$

Wegen $v \notin U$ ist diese Beziehung nur für $\alpha = 1$ erfüllt, und damit folgt $u_1 = u_2$. ∎

Mit einer Basisdarstellung

$$U = \operatorname{span}(u_1, u_2, \cdots, u_n)$$

$$= \left\{ \sum_{k=1}^{n} \alpha_k u_k \mid \alpha_k \in \mathbb{R} \text{ (oder } \mathbb{C}) \right\}$$

und der Funktion

(18.6 a) $$F(\alpha) = \| v - \sum_{k=1}^{n} \alpha_k u_k \|, \quad \alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n,$$

geht die Approximationsaufgabe (18.1) über in die Optimierungsaufgabe.

(18.6 b)

> Bestimme $\tilde{\alpha} \in \mathbb{R}^n$ mit
> $$F(\tilde{\alpha}) = \min_{\alpha \in \mathbb{R}^n} F(\alpha)$$

.

In dem praktisch wichtigen Fall $V = C[a,b]$ nennt man das Problem (18.6)

- TSCHEBYCHEV - Approximation, falls
  $$\| \cdot \| = \| \cdot \|_\infty ;$$
- GAUSS - Approximation, Approximation im quadratischen Mittel, falls
  $$\| \cdot \| = \| \cdot \|_2 .$$

§19 Approximation in Prä-Hilberträumen

19.1 Die Normalgleichungen

Sei V ein Prä-Hilbertraum mit dem inneren Produkt $(\cdot,\cdot)$ und der durch

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

induzierten Norm. Sei

$$U = \operatorname{span}(u_1,\dots,u_n) \subset V$$

ein endlich-dimensionaler Unterraum. Wegen der Strenge der Norm gibt es zu $f \in V$ nach Satz (18.5) genau ein $u \in U$ mit

$$\|f-\tilde{u}\| = \min_{u\in U} \|f-u\|.$$

Diese Aufgabe ist äquivalent zu

$$\|f-\tilde{u}\|^2 = \min_{u\in U} \|f-u\|^2$$

$$= \min_{u\in U} (f-u, f-u).$$

In der Basisdarstellung $u = \sum_{k=1}^{n} \alpha_k u_k$ ist also die Optimierungsaufgabe (18.6) zu lösen:

$$\min_{\alpha\in\mathbb{R}^n} \Big\| v - \sum_{k=1}^{n} \alpha_k u_k \Big\|^2, \quad \alpha = (\alpha_1,\dots,\alpha_n).$$

(19.1) Orthogonalität und Normalgleichungen für Proximum

$\tilde{u} \in U$ ist genau Proximum an $f \in V$, wenn gilt,

$$(f-\tilde{u}, u) = 0 \quad \text{für alle } u \in U.$$

In der Basisdarstellung $\tilde{u} = \sum_{k=1}^{n} \tilde{\alpha}_k u_k$ sind die Koeffizienten $\tilde{\alpha}_k$ die eindeutig bestimmte Lösung der Normalgleichungen

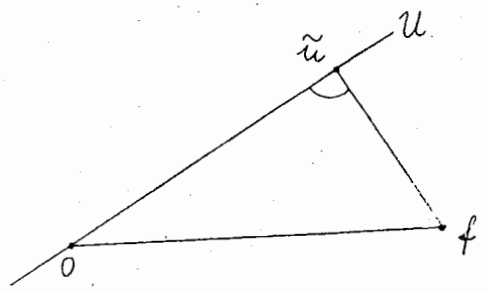
$$\boxed{\sum_{i=1}^{n} \tilde{\alpha}_i (u_i, u_k) = (f, u_k), \quad k = 1,\dots,n}.$$

Die Abweichung erfüllt

$$\|f-\tilde{u}\|^2 = \|f\|^2 - \|\tilde{u}\|^2$$

$$= \|f\|^2 - \sum_{k=1}^{n} \tilde{\alpha}_k (f, u_k).$$

Beweis: Wir wollen den folgenden geometrischen Sachverhalt (Satz von Pythagoras) zeigen:



Es gelte die Orthogonalitätsrelation

$$(f-\tilde{u}, u)=0 \quad \text{für alle } u\in U.$$

Dann ist für jedes $u\in U$ wegen $(f-\tilde{u}, u-\tilde{u})=0$:

$$\begin{aligned}\|f-u\|^2 &= \|f-\tilde{u}-(u-\tilde{u})\|^2\\ &= \|f-\tilde{u}\|^2+\|u-\tilde{u}\|^2\\ &\geq \|f-\tilde{u}\|^2.\end{aligned}$$

Also ist $\tilde{u}$ das eindeutig bestimmte Proximum.

Wir haben jetzt noch zu zeigen, daß die obige Orthogonalitätsrelation für ein $\tilde{u}$ lösbar ist. In der Basisdarstellung $\tilde{u}=\sum_{i=1}^{n}\tilde{\alpha}_i u_i$ lautet diese

$$(f-\sum_{i=1}^{n}\tilde{\alpha}_i u_i, u_k)=0, \quad k=1,\dots,n.$$

Daher müssen die Normalgleichungen gelten:

$$\sum_{i=1}^{n}\tilde{\alpha}_i(u_i,u_k)=(f,u_k), \quad k=1,\dots,n.$$

Wegen der linearen Unabhängigkeit von $u_1,\dots,u_n$ ist die GRAMsche Matrix $((u_i,u_k))_{i,k=1,\dots,n}$ positiv definit, also gibt es eine eindeutig bestimmte Lösung $\tilde{\alpha}_1,\dots,\tilde{\alpha}_n$. Für die Abweichung berechnet man wegen $(f-\tilde{u},\tilde{u})=0$

$$\|f\|^2=\|f-\tilde{u}\|^2+\|\tilde{u}\|^2,$$

$$\|u\|^2=(\tilde{u}-f+f,\tilde{u})=(f,\tilde{u})=\sum_{k=1}^{n}\tilde{\alpha}_k(f,u_k). \quad \blacksquare$$

Beispiel: $V=C[-1,1]$, $f(x)=e^x$.

$$U=\operatorname{span}(1,x,x^2)=\operatorname{span}(u_0,u_1,u_2).$$

$$(u_i,u_k)=\int_{-1}^{1}u_i(x)u_k(x)\,dx$$

$$=\int_{-1}^{1}x^{i+k}\,dx=\frac{1+(-1)^{i+k}}{i+k+1},$$

$$(f,u_0)=\int_{-1}^{1}e^x\,1\,dx=e-1/e,$$

$$(f,u_1)=\int_{-1}^{1}e^x x\,dx=2/e,$$

$$(f,u_2)=\int_{-1}^{1}e^x x^2\,dx=e-5/e.$$

Die Normalgleichungen lauten damit

$$\begin{pmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{pmatrix} = \begin{pmatrix} e - 1/e \\ 2/e \\ e - 5/e \end{pmatrix}$$

Für die Lösung

$\tilde{\alpha}_0 = 0.99629$, $\tilde{\alpha}_1 = 1.10364$, $\tilde{\alpha}_2 = 0.53672$

gilt

$$\max_{-1 \le x \le 1} |e^x - (\tilde{\alpha}_0 + \tilde{\alpha}_1 x + \tilde{\alpha}_2 x^2)| = 0.082.$$

Im Falle eines Orthonormalsystems (ONS) $u_1, \dots, u_n$, d.h.

$$(u_i, u_k) = \delta_{ik} \quad (1 \le i, k \le n),$$

vereinfachen sich die Normalgleichungen zu

(19.2)
$$\boxed{\begin{aligned} \tilde{\alpha}_k &= (f, u_k) \\ \tilde{u} &= \sum_{k=1}^{n} (f, u_k)\, u_k \end{aligned}}$$

Die $\tilde{\alpha}_k$ heißen verallgemeinerte FOURIER-Koeffizienten von f bzgl. $u_1, \dots, u_n$. Aus

der Abweichung

(19.3) $$\|f - \tilde{u}\|^2 = \|f\|^2 - \sum_{k=1}^{n} (\tilde{\alpha}_k)^2$$

in Satz (19.1) erhalten wir die BESSELsche Ungleichung

(19.4) $$\sum_{k=1}^{n} (\tilde{\alpha}_k)^2 \le \|f\|^2 .$$

Gegeben sei nun ein unendliches ONS $\{u_n\}_{n \in \mathbb{N}}$.

(19.5) Definition: Das ONS $\{u_1, u_2, \dots\}$ heißt vollständig, wenn es zu jedem $f \in V$ eine Folge $\{f_n\} \subset V$ gibt mit

$$f_n \in \operatorname{span}(u_1, \dots, u_n), \quad \lim_{n \to \infty} \|f - f_n\| = 0.$$

(19.6) Vollständigkeitsrelation

Notwendig und hinreichend für die Vollständigkeit des ONS $\{u_1, u_2, \dots\}$ ist die Vollständigkeitsrelation (PARSEVALsche Gleichung)

$$\sum_{k=1}^{\infty} (\tilde{\alpha}_k)^2 = \|f\|^2 .$$

Der Beweis dieser Aussage mittels (19.3) ist elementar.

Wir werden nun Beispiele für ONS kennenlernen.

19.2 Trigonometrische Approximation

Gegeben sei eine stückweise stetige Funktion $f: \mathbb{R} \to \mathbb{R}$ mit der Periode 2π:

$$f(x+2\pi) = f(x) \quad \text{für alle } x \in \mathbb{R}.$$

Zur Approximation von f betrachten wir den Raum $V = C^{-1}[-\pi, \pi]$ der stückweise stetigen Funktionen mit der Norm

$$\|f\|_2 = \left(\int_{-\pi}^{\pi} f(x)^2 dx \right)^{1/2}.$$

Man prüft leicht nach, daß die Funktionen $u_0, u_1, u_2, \ldots, u_{2m}$, erklärt durch

$$u_0(x) = \frac{1}{\sqrt{2\pi}}, \quad u_{2k-1}(x) = \frac{1}{\sqrt{\pi}} \sin(kx)$$

$$u_{2k}(x) = \frac{1}{\sqrt{\pi}} \cos(kx), \quad (k = 1, \ldots, m),$$

ein ONS in V bzgl. der Norm $\|\cdot\|_2$

bilden. Das Proximum

$$u \in \mathcal{U} = \operatorname{span}(u_0, u_1, u_2, \ldots, u_{2m})$$

an $f \in V$ ist nach (19.2) die FOURIER-Reihe

(19.7)
$$\tilde{u}(x) = \frac{a_0}{2} + \sum_{k=1}^{m} (a_k \cos(kx) + b_k \sin(kx))$$

mit
$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx) dx \quad (k = 0, 1, \ldots, m)$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx) dx \quad (k = 1, \ldots, m)$$

Die BESSELsche Ungleichung (19.4) wird zu

$$\frac{a_0^2}{2} + \sum_{k=1}^{m} (a_k^2 + b_k^2) \leq \frac{1}{\pi} \|f\|_2^2.$$

Über das Verhalten der Reihe (19.7) gibt der folgende Satz Auskunft.

(19.8) **Satz:** Sei $f \in C^{-1}[-\pi, \pi]$ periodisch mit der Periode 2π.

(i) Die Folge (19.7) der Proxima bzgl. $\|\cdot\|_2$ konvergiert für $m \to \infty$ im Mittel gegen f.

(ii) Existiert zusätzlich $f' \in C^{-1}[-\pi, \pi]$,

so konvergiert die Reihe (19.7) punktweise gegen den Wert

$$\lim_{h\to 0+} \frac{1}{2}\left(f(x+h)+f(x-h)\right), \quad x\in[-\pi,\pi].$$

Beispiele:

(1) Für die stetige gerade Funktion

$$f(x)=|x|, \quad -\pi\le x\le\pi, \quad f(x+2\pi)=f(x),$$

erhält man

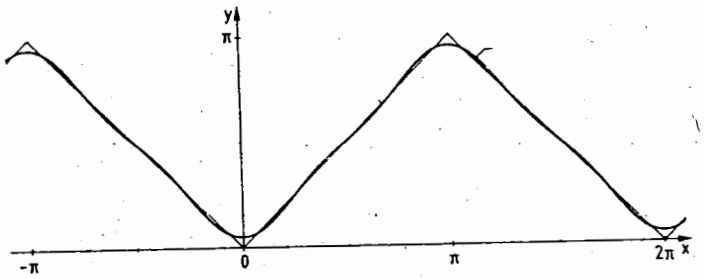
$$a_0=\frac{1}{\pi}\int_{-\pi}^{\pi}|x|\,dx=\pi$$

$$a_k=\frac{2}{\pi}\int_0^{\pi}x\cos(kx)\,dx=\frac{2}{\pi k^2}\left[(-1)^k-1\right], \quad k>0,$$

$$b_k=0.$$

Die FOURIER-Reihe lautet daher

$$\frac{1}{2}\pi-\frac{4}{\pi}\left\{\frac{\cos(x)}{1^2}+\frac{\cos(3x)}{3^2}+\frac{\cos(5x)}{5^2}+\cdots\right\}.$$

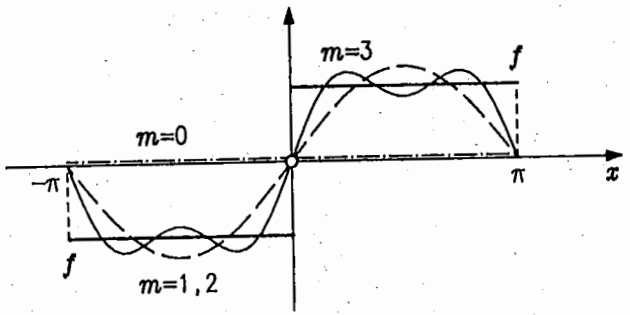


(2) Die Funktion

$$f(x)=\begin{cases}-1, & -\pi\le x<0\\ 0, & x=0\\ 1, & 0<x\le\pi\end{cases}$$

ist unstetig in $x=0$. Da f ungerade ist, gilt $a_k=0$, und man berechnet

$$b_k=\frac{2}{\pi}\int_0^{\pi}\sin(kx)\,dx=\begin{cases}\frac{4}{\pi k}, & k \text{ ungerade}\\ 0, & k \text{ gerade}.\end{cases}$$



19.3 Orthogonalpolynome

Der Raum $V = C[a,b]$ mit dem inneren Produkt

$$(f,g) = \int_a^b f(x)\,g(x)\,w(x)\,dx,$$

$$w \in C(a,b),\quad w(x) > 0 \text{ für } a < x < b,$$

ist ein Prä-Hilbertraum. Wendet man das SCHMIDTsche Orthogonalisierungsverfahren auf die Monome $u_n(x) = x^n$ an, so erhält man Orthogonalpolynome mit Höchstkoeffizient $a_n = 1$:

$$\tilde{p}_n \in \tilde{\Pi}_n := \left\{ x^n + \sum_{k=0}^{n-1} a_k x^k \right\} \quad (n = 0,1,2,\ldots),$$

$$(\tilde{p}_i, \tilde{p}_k) = 0 \text{ für } i \neq k.$$

Die Polynome $\tilde{p}_0, \ldots, \tilde{p}_n$ bilden eine Basis von Π_n und genügen der folgenden Drei-Term-Rekursion (vgl. STOER, Satz (3.5.3)):

(19.9)
$$\tilde{p}_0(x) = 1,\quad \tilde{p}_1(x) = x - \delta_1$$
$$\tilde{p}_{n+1}(x) = (x - \delta_{n+1})\,\tilde{p}_n(x) - \gamma_{n+1}^2\,\tilde{p}_{n-1}(x),\quad n \geq 1,$$
$$\delta_{n+1} = \frac{(x\tilde{p}_n, \tilde{p}_n)}{(\tilde{p}_n, \tilde{p}_n)},\quad \gamma_{n+1}^2 = \frac{(\tilde{p}_n, \tilde{p}_n)}{(\tilde{p}_{n-1}, \tilde{p}_{n-1})}$$

Darüberhinaus gilt der bemerkenswerte

(19.10) Nullstellensatz:

Das Orthogonalpolynom $\tilde{p}_n \in \tilde{\Pi}_n$ hat in (a,b) genau n einfache Nullstellen.

Beweis: Seien $x_1, \ldots, x_m$ $(m \geq 0)$ die Zeichenwechsel von $\tilde{p}_n$ in (a,b). Wir zeigen, daß $m = n$ gilt. Mit dem Polynom

$$q(x) = \prod_{i=1}^{m} (x - x_i) \in \Pi_m$$

hat das Polynom $q \cdot \tilde{p}_n$ konstantes Vorzeichen. Für $m < n$ würde folgen

$$(q, \tilde{p}_n) = \int_a^b q(x)\,\tilde{p}_n(x)\,w(x)\,dx = 0,$$

also $q \cdot \tilde{p}_n \equiv 0$ in (a,b): Widerspruch. ∎

Die Proxima von $f \in C[a,b]$ in

$$\mathcal{U} = \operatorname{span}(\tilde{p}_0, \tilde{p}_1, \ldots, \tilde{p}_n)$$

haben wegen $(\tilde{p}_i, \tilde{p}_k) = 0$ $(i \neq k)$ die FOURIER-Darstellung

(19.11)
$$\tilde{u} = \sum_{k=0}^{n} \tilde{\alpha}_k\,\tilde{p}_k$$
$$\tilde{\alpha}_k = (f, \tilde{p}_k) / (\tilde{p}_k, \tilde{p}_k) \quad (k \geq 0)$$

(19.12) Konvergenzsatz: Für jede Gewichtsfunktion $w \in C[a,b]$, $w(x) > 0$ für $a < x < b$, sind die Funktionen $\{p_0, p_1, \dots, p_n, \dots\}$, $p_n := \tilde{p}_n / \sqrt{(\tilde{p}_n, \tilde{p}_n)}$, ein vollständiges ONS in $C[a,b]$. Die Folge der Proxima (19.11) konvergiert für $n \to \infty$ im Mittel gegen f.

Zum Beweis vgl. man HÄMMERLIN/HOFFMANN, Kap. 4, § 5.6.

$$\int_{-1}^{1} \tilde{L}_n(x)\, \tilde{L}_m(x)\, dx$$

Beispiele:

(1) $[a,b] = [-1,1]$, $w = 1$: LEGENDRE-Polynome

Man prüft leicht nach, daß die LEGENDRE-Polynome

$$\tilde{L}_n(x) = \frac{n!}{(2n)!} \frac{d^n (x^2-1)^n}{dx^n} = x^n + \dots, \quad n \geq 0 \tag{19.13}$$

ein Orthogonalsystem bilden mit

n	0	1	2	3
$\tilde{L}_n$	1	x	$x^2 - \frac{1}{3}$	$x^3 - \frac{3}{5}x$
$(\tilde{L}_n, \tilde{L}_n)$	2	$\frac{2}{3}$	$\frac{8}{45}$	$\frac{8}{175}$

$$(\tilde{L}_n, \tilde{L}_n) = \left(\frac{2n+1}{2} \frac{1}{(2^n n!)^2} \left(\frac{(2n)!}{n!}\right)^2\right)^{-1} = \left(\frac{n!}{(2n)!}\right)^2 \frac{2(2^n n!)^2}{2n+1}$$

Zum Beispiel ist Normierungsfaktor: $L_n(x) = \tilde{L}_n(x) \sqrt{\frac{2n+1}{2}} \frac{(2n)!}{n!} \frac{1}{2^n n!}$ (19.15)

$$\tilde{L}_1(x) = x, \quad \tilde{L}_2(x) = x^2 - \frac{1}{3}, \quad \tilde{L}_3(x) = x^3 - \frac{3}{5}x.$$

Hierfür bestätigt man den Nullstellensatz (19.10).

(2) $[a,b] = [-1,1]$, $w(x) = 1/\sqrt{1-x^2}$:

TSCHEBYCHEV-Polynome

Die Tschebychev-Polynome $T_n \in \Pi_n$ werden rekursiv definiert durch

$$T_{n+1}(x) = 2x\, T_n(x) - T_{n-1}(x) \quad (n = 1, 2, \dots), \tag{19.14}$$
$$T_0(x) = 1, \quad T_1(x) = x.$$

Mit der Substitution

$$x = \cos\theta, \quad \theta = \arccos x$$

gelangt man zur Darstellung

$$T_n(x) = \cos(n\theta), \quad \theta = \arccos x,$$

denn die Rekursionsformel

$$\cos((n+1)\theta) = 2\cos\theta \cos(n\theta) - \cos((n-1)\theta)$$

entspricht gerade der Rekursion (19.14). Damit bestätigt man die Orthogonalität bzgl. $w(x) = 1/\sqrt{1-x^2}$:

$$\int_{-1}^{1} T_i(x)\, T_k(x) \frac{dx}{\sqrt{1-x^2}} = \int_0^{\pi} \cos(i\theta)\cos(k\theta) \frac{\sin\theta}{\sin\theta} d\theta$$

$$= \begin{cases} 0 & \text{für } i \neq k \\ \pi & \text{für } i = k = 0 \\ \frac{\pi}{2} & \text{für } i = k \neq 0 \end{cases}.$$

Die Darstellung $T_n(x) = \cos(n\theta)$ zeigt, daß $T_n(x)$ die Nullstellen (TSCHEBYCHEV-Abszissen)

$$x_k = \cos\left(\frac{2k-1}{n}\,\frac{\pi}{2}\right) \in (-1,1) \quad , \quad k = 1, \dots, n,$$

und die Extremalstellen

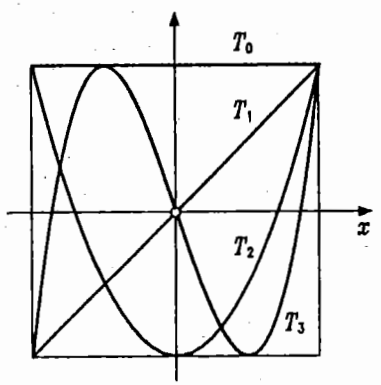
$$x_k^{(\ell)} = \cos\left(\frac{k\pi}{n}\right) \quad (k = 0, 1, \dots, n) \, , \; n \geq 1,$$

$$T_n\left(x_k^{(\ell)}\right) = (-1)^k,$$

besitzt. Z.B. liefert die Rekursion (19.14)

$$T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3, \dots,$$

$$T_n(x) = 2^{n-1} x^n - \dots \, .$$



Zu normierten Polynomen $\tilde{T}_n \in \tilde{\Pi}_n$ gelangt man durch die Normierung

$$\tilde{T}_n(x) = \frac{1}{2^{n-1}} T_n(x).$$

Die FOURIER-Entwicklung einer Funktion $f \in C[a,b]$ lautet wegen (19.11)

$$\tilde{u} = \frac{c_0}{2} + \sum_{k=1}^{n} c_k T_k(x),$$

$$c_k = \frac{2}{\pi} \int_{-1}^{1} f(x) T_k(x) \frac{dx}{\sqrt{1-x^2}} \, , \; k \in \mathbb{N},$$

$$= \frac{2}{\pi} \int_0^{\pi} f(\cos\theta) \cos(k\theta) \, d\theta \, .$$

Also sind c_k gerade die FOURIER-Koeffizienten (19.7) der 2π-periodischen Funktion $F(\theta) = f(\cos\theta)$:

Die Konvergenz für $n \to \infty$ gilt bzgl. der Norm $\|\cdot\|_2$; für $f \in C^2[a,b]$ ist diese Konvergenz sogar gleichmäßig bzgl. der Norm $\|\cdot\|_\infty$.

Weitere Orthogonalpolynome bzgl. anderer Gewichte $w(x)$ finden sich in M. ABRAMOVITZ, F. STEGUN: Handbook of Mathematical Functions.

§25 Die Integrationsformeln von Newton-Cotes

Die Stützstellen x_i seien äquidistant und enthalten die Randpunkte des Intervalls, d.h.

$$x_i = a + i \cdot h, \quad h = \frac{b-a}{n}, \quad i = 0, \dots, n.$$

Sei P_n das f interpolierende Polynom mit

(a) Grad $P_n \le n$,

(b) $P_n(x_i) = f_i := f(x_i), \quad i = 0, \dots, n.$

Als Näherungswert für I nehmen wir den Ausdruck

$$I_n := \int_a^b P_n(x)\,dx = \sum_{i=0}^{n} A_i f(x_i)$$

mit dem Fehler

$$R_n(f) = I - I_n = \int_a^b f(x)\,dx - \int_a^b P_n(x)\,dx.$$

Es gilt $R_n(f) = 0$, falls f Polynom vom Grade $\le n$ ist.

In der Form von Lagrange lautet P_n:

$$P_n(x) = \sum_{i=0}^{n} f_i L_i(x), \quad L_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}.$$

Damit gilt

$$I_n = \int_a^b P_n(x)\,dx = \sum_{i=0}^{n} f_i \int_a^b L_i(x)\,dx,$$

$$A_i = \int_a^b L_i(x)\,dx = \int_a^b \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}\,dx.$$

Mit der Substitution $x = a + sh$, $s \in [0, n]$, $dx = h\,ds$, erhalten wir die Formeln von Newton-Cotes:

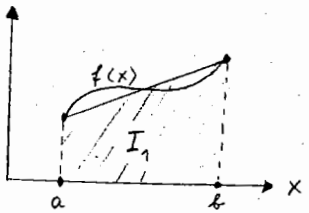
(25.1)
$$I_n = \sum_{i=0}^{n} A_i f_i, \quad f_i = f(a + ih),$$

$$A_i = h \int_0^n \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{s-k}{i-k}\,ds =: h\,a_i, \quad a_i \in \mathbb{Q}$$

Beispiele: $n = 1$: Trapezregel

$h = b - a$, $a_0 = a_1 = \frac{1}{2}$,

$I_1 = \frac{h}{2}(f(a) + f(b))$



$n = 2$: Simpson-Regel

$h = \frac{b-a}{2}$, $a_0 = \frac{1}{3}$, $a_1 = \frac{4}{3}$, $a_2 = \frac{1}{3}$,

$I_2 = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$

Allgemein gilt $a_i \in \mathbb{Q}$ und

$$\sum_{i=0}^{n} a_i = n, \quad \text{da } P_n \equiv 1 \text{ zu } f \equiv 1.$$

Die folgende Tabelle enthält die Koeffizienten a_i für $n \le 4$:

n	a_0	a_1	a_2	a_3	a_4		Bezeichnung
1	1	1				$\cdot \frac{1}{2}$	Trapezregel
2	1	4	1			$\cdot \frac{1}{3}$	Simpson - Regel
3	1	3	3	1		$\cdot \frac{3}{8}$	Newton'sche $\frac{3}{8}$ - Regel
4	7	32	12	32	7	$\cdot \frac{2}{45}$	Milne - Regel

Für $n \ge 8$ können <u>negative</u> Gewichte auftreten und die Formeln werden dann aus Rundungsfehlergründen numerisch unbrauchbar. Formeln höherer Genauigkeit kann man konstruieren, indem man die oben angegebenen Regeln auf Teilintervalle anwendet.

<u>Beispiel</u>:

$$I = \int_0^1 e^x dx = e - 1 = 1.7183$$

$$I_1 = \frac{1}{2}(1+e) = 1.8591$$

$$I_2 = \frac{1}{6}(1 + 4e^{\frac{1}{2}} + e) = 1.7189$$

$$I_3 = \frac{1}{8}(1 + 3e^{\frac{1}{3}} + 3e^{\frac{2}{3}} + e) = 1.7185$$

Man sieht, daß der Übergang von I_1 nach I_2 einen großen Gewinn an Genauigkeit ergibt.

Der folgende Satz gibt eine Abschätzung für den Fehler $R_n(f) = I - I_n$.

(25.2) <u>Satz</u>:

(<u>i</u>) Für $f \in C^{n+1}[a,b]$ gilt

$$|R_n(f)| \le h^{n+2} c_n \max_{[a,b]} |f^{(n+1)}(x)|$$

mit $c_n = \frac{1}{(n+1)!} \int_0^n \prod_{i=0}^{n} |s-i| \, ds$.

(<u>ii</u>) Für n gerade und $f \in C^{n+2}[a,b]$ gilt

$$|R_n(f)| \le h^{n+3} c_n^* \max_{[a,b]} |f^{(n+2)}(x)|, \quad c_n^* = \frac{n}{2} c_n,$$

<u>Bem.</u>: Bei geradem n gewinnt man durch den Übergang zu $n+1$ keine Potenz von h.

<u>Beweis</u>: <u>zu (i)</u>: Es gilt

$$R_n(f) = I - I_n = \int_a^b (f - P_n)(x) dx.$$

Nach der Restgliedformel der Polynominterpolation (vgl. (15.8)) ist

$$(f-P_n)(x) = \frac{L(x)}{(n+1)!} f^{(n+1)}(\xi), \quad \xi \in [a,b], \quad L(x) = \prod_{i=0}^{n} (x-x_i).$$

Hieraus folgt

$$|R_n(f)| \le \frac{1}{(n+1)!} \int_a^b |L(x)|\,dx \max_{[a,b]} |f^{(n+1)}(x)|.$$

Mit der Substitution $x = a+sh$, $s \in [0,n]$ ergibt sich

$$\int_a^b |L(x)|\,dx = \int_a^b \prod_{i=0}^{n} |x-x_i|\,dx = h^{n+2} \int_0^n \prod_{i=0}^{n} |s-i|\,ds.$$

Zusammen mit der vorigen Abschätzung erhält man die Behauptung.

zu (ii): Für gerades n ist L schiefsymmetrisch bezüglich der Intervallmitte $c = \frac{a+b}{2}$, es gilt also

$$\int_a^b L(x)\,dx = 0$$

Durch Taylor-Entwicklung von $f^{(n+1)}(\xi)$ bzgl. c erhält man

$$\int_a^b (f-P_n)(x)\,dx = \frac{1}{(n+1)!} \int_a^b L(x) f^{(n+1)}(\xi)\,dx$$

$$= \frac{1}{(n+1)!} \int_a^b L(x) \left\{ f^{(n+1)}(c) + (\xi-c) f^{(n+2)}(\eta) \right\} dx$$

$$\left(\int_a^b L(x)\,dx = 0 \right) \qquad = \frac{1}{(n+1)!} \int_a^b L(x)(\xi-c) f^{(n+2)}(\eta)\,dx.$$

Wegen $|\xi - c| \le \frac{b-a}{2} = \frac{nh}{2}$ gilt

$$|R_n(f)| \le \frac{1}{(n+1)!} \int_a^b |L(x)|\,dx \cdot \frac{nh}{2} \max_{[a,b]} |f^{(n+2)}(x)|$$

$$= h^{n+2} c_n \frac{nh}{2} \max_{[a,b]} |f^{(n+2)}(x)|$$

$$= h^{n+3} c_n^* \max_{[a,b]} |f^{(n+2)}(x)|. \quad \blacksquare$$

Bem.: Da das Maximum hoher Ableitungen von f sehr schwer zu bestimmen ist, sind diese Formeln zur praktischen Abschätzung des Fehlers i.a. unbrauchbar. Ihr Nutzen liegt in der Information, mit welcher Potenz von h der Fehler abfällt.

Beispiele:

(1) $n=1$: Trapezregel

$$|R_1(f)| \le \frac{h^3}{12} \max_{[a,b]} |f^{(2)}(x)|.$$

(2) $n=2$: Simpson-Regel

$$|R_2(f)| \le \frac{h^5}{90} \max_{[a,b]} |f^{(4)}(x)|.$$

(3) $n=3$: Newton'sche $\frac{3}{8}$-Regel

$$\int_0^1 \frac{\sin x}{x}\,dx, \quad f(x)=\frac{\sin x}{x},$$

$$I_3 = \frac{1}{8}\left(f(0)+3f\left(\tfrac{1}{3}\right)+3f\left(\tfrac{2}{3}\right)+f(1)\right) = 0.946111$$

$$|R_3(f)| \le \frac{3}{80}\left(\frac{1}{3}\right)^5 \max_{0\le x\le 1}|f^{(4)}(x)| \le 3.1\cdot 10^{-5}$$

$$I = \int_0^1 \frac{\sin x}{x}\,dx = 0.946\,083\,07$$

$$|I - I_3| \le 2.8\cdot 10^{-5} \quad \text{(exakt)}.$$

§26 Die zusammengesetzte Trapezregel und Extrapolationsverfahren

Die Stützstellen seien wieder äquidistant:

$$x_i = a + i\cdot h, \quad \therefore h = \frac{b-a}{n}, \quad i = 0,\dots,n.$$

Die Anwendung der Trapezregel auf das Teilintervall $[x_i, x_{i+1}]$ ergibt die Approximation

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \cong \frac{h}{2}\left(f(x_i)+f(x_{i+1})\right), \quad i=0,\dots,n-1.$$

Durch Summation erhalten wir die zusammengesetzte Trapezregel:

(26.1)
$$\int_a^b f(x)\,dx \cong T(h) := \sum_{i=0}^{n-1} \frac{h}{2}\left(f(x_i)+f(x_{i+1})\right)$$
$$= h\cdot\left[\frac{f(a)}{2} + f(a+h) + \dots + f(b-h) + \frac{f(b)}{2}\right]$$

Der Gesamtfehler beträgt nach Satz (25.2)(i) mit $n=1$:

$$|T(h) - \int_a^b f(x)\,dx| \le \frac{1}{12}\sum_{i=0}^{n-1} h^3 \max_{[x_i,x_{i+1}]} |f''(x)|$$

(26.2)
$$\le \frac{n}{12}h^3 \max_{[a,b]}|f''(x)|$$

$$= \frac{b-a}{12}h^2 \max_{[a,b]}|f''(x)| = O(h^2).$$

Durch Anwendung der Extrapolation auf die zusammengesetzte Trapezregel (26.1) wollen wir nun Formeln konstruieren, deren Fehler mit einer hohen Potenz von h abfällt. Grundlage dieser Extrapolationsverfahren ist die folgende asymptotische Entwicklung von $T(h)$ nach Potenzen von h^2.

(26.3) Satz: (Euler-Maclaurin'sche Summenformel)

Für $f \in C^{2m+2}[a,b]$ gilt die Entwicklung

$$\boxed{T(h) = \tau_0 + \tau_1 h^2 + \tau_2 h^4 + \ldots + \tau_m h^{2m} + \alpha_{m+1}(h) h^{2m+2}}$$

mit

(1) $\tau_0 = \int_a^b f(x)\,dx$,

(2) $\tau_k = \frac{(-1)^{k+1} B_k}{(2k)!}\left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right)$, $k = 1, \ldots, m$

(3) $\alpha_{m+1}(h) = \frac{1}{(2m+2)!} \int_a^b f^{(2m+2)}(x)\, K_{2m+2}\left(\frac{x-a}{h}\right) dx$,

mit $K_{2m+2} \in C[0,n]$ und

$$\int_a^b K_{2m+2}\left(\frac{x-a}{h}\right) dx = (-1)^m B_{m+1}(b-a).$$

Hierbei sind B_k die Bernoulli-Zahlen

$$B_1 = \frac{1}{6},\ B_2 = \frac{1}{30},\ B_3 = \frac{1}{42}, \ldots.$$

Beweis:

Nach dem vorigen Satz gilt

$$T(h) = \underbrace{\tau_0 + \tau_1 h^2 + \ldots + \tau_m h^{2m}}_{\text{Polynom vom Grade } m \text{ in } h^2} + \alpha_{m+1}(h) h^{2m+2}.$$

Es interessiert die Größe

$$\tau_0 = \int_a^b f(x)\,dx = \lim_{h \to 0} T(h).$$

Idee der Extrapolation: Zu $m+1$ Schrittweiten

$$h_0 = b-a,\ h_1 = \frac{h_0}{n_1}, \ldots, h_m = \frac{h_0}{n_m};\quad n_i < n_{i+1}\ (n_i \in \mathbb{N})$$

bestimme man Trapezsummen

$$T_{i0} := T(h_i),\quad i = 0, \ldots, m,$$

und dann durch Interpolation dasjenige Polynom in h^2

(26.4) $$\tilde{T}_{mm}(h) := a_0 + a_1 h^2 + \ldots + a_m h^{2m}$$

mit

$$\tilde{T}_{mm}(h_i) = T(h_i),\quad i = 0, \ldots, m.$$

Dann

$$\tilde{T}_{mm}(0) = a_0 \approx \tau_0 \quad \text{(Extrapolation)}.$$

Beispiel: $h_0 = b-a$, $h_1 = \frac{b-a}{2}$.

Dann ist

$$T_{11} := \tilde{T}_{11}(0) = L_0(0)\,T(h_0) + L_1(0)\,T(h_1)$$

mit $\tilde{T}_{mn}(h) = L_0(h)\,T(h_0) + L_1(h)\,T(h_1)$

$$L_i(x) = \prod_{\substack{k=0 \\ k\neq i}} \frac{x - x_k}{x_i - x_k}, \quad x_i = h_i^2, \quad i = 0,1.$$

Für $x = 0$ erhält man

$$L_0(0) = \frac{-h_1^2}{h_0^2 - h_1^2} = -\frac{1}{3}, \qquad L_1(0) = \frac{-h_0^2}{h_1^2 - h_0^2} = \frac{4}{3}$$

und daher

$$T_{11} = -\frac{1}{3}\underbrace{\frac{b-a}{2}\left(f(a)+f(b)\right)}_{=T(h_0)} + \frac{4}{3}\underbrace{\frac{b-a}{2}\left(\frac{f(a)}{2} + f\left(\frac{a+b}{2}\right) + \frac{f(b)}{2}\right)}_{=T(h_1)}$$

$$= \frac{h_1}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) \quad \text{Simpson-Regel!}$$

Die Berechnung des Wertes $\tilde{T}_{mm}(0) = a_0$ in (26.4) erfolgt mit dem Algorithmus von Neville; vgl. Teil I, (15.4). Dazu sei $1 \le k \le i \le m$ und $\tilde{T}_{ik}(h)$ dasjenige Polynom in h^2 mit

$$\tilde{T}_{ik}(h_j) = T_{j0} := T(h_j) \quad \text{für} \quad j = i-k, i-k+1, \ldots, i.$$

Die Rekursion für $T_{ik} := \tilde{T}_{ik}(0)$ ergibt sich aus dem Algorithmus von Neville mit $x_i = h_i^2$, $x = 0$, zu

$$= T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{2^{2k} - 1}$$

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^2 - 1}, \quad 1 \le k \le i \le m.$$

$\frac{h_{i-k}}{h_i} = \frac{2^{k-i}}{2^{-i}} = 2^k$

Die Berechnung des Tableaus mit den Größen T_{ik} erfolgt spaltenweise, z.B.

h_0	T_{00}			
h_1	T_{10}	T_{11}		
h_2	T_{20}	T_{21}	T_{22}	
h_3	T_{30}	T_{31}	T_{32}	T_{33}

Beispiel: $I = \int_0^1 e^x\,dx = 1.718281828$, $m = 3$

i	h_i	T_{i0}	T_{i1}	T_{i2}	T_{33}
0	1	1.859140914			
1	$\frac{1}{2}$	1.753931092	1.718861151		
2	$\frac{1}{4}$	1.727221904	1.718318841	1.718282687	
3	$\frac{1}{8}$	1.720518592	1.718284155	1.718281842	1.718281828

In der Praxis haben sich zwei Schrittweitenfolgen bewährt:

Romberg-Folge: $h_0 = b-a$, $h_i = \frac{h_0}{2^i}$, $i = 0, 1, \ldots$

Bulirsch-Folge: $h_0 = b-a$, $h_1 = \frac{h_0}{2}$, $h_2 = \frac{h_0}{3}$, $h_3 = \frac{h_0}{4}$, $h_4 = \frac{h_0}{6}$, $h_5 = \frac{h_0}{8}$, ...

$$h_i := \begin{cases} \frac{h_0}{3 \cdot 2^{h-1}} & i = 2h \\ \frac{h_0}{2^h} & i = 2h-1 \end{cases}$$

Die Bulirsch-Folge hat den Vorteil, daß bei ihr die Rechenarbeit für die Berechnung neuer $T(h_i)$ nicht so rasch ansteigt wie bei der Romberg-Folge. Bei der praktischen Durchführung beachte man, daß bei der Berechnung von $T(h_{i+1})$ auf die schon bei $T(h_i)$ berechneten Funktionswerte zurückgegriffen wird.

Im folgenden soll ein Ausdruck für den Fehler

$$T_{mm} - \int_a^b f(x)\,dx$$

angegeben werden.

(26.5) Hilfssatz: Seien x_i, $i = 0, \ldots, m$, paarweise verschiedene Zahlen und sei

$$L_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{m} \frac{x - x_k}{x_i - x_k}, \quad i = 0, \ldots, m. \qquad L_i(x_k) = \begin{cases} 0 & k \neq i \\ 1 & k = i \end{cases}$$

Dann gilt

$$\sum_{i=0}^{m} x_i^j L_i(0) = \begin{cases} 1 & , \ j = 0 \\ 0 & , \ j = 1, \ldots, m \\ (-1)^m x_0 x_1 \ldots x_m & , \ j = m+1 \end{cases}$$

Beweis: Man setze $x = 0$ in den beiden folgenden Identitäten ein:

$$x^j \equiv \sum_{i=0}^{m} x_i^j L_i(x), \quad j = 0, \ldots, m,$$

$$x^{m+1} \equiv \sum_{i=0}^{m} x_i^{m+1} L_i(x) + (x-x_0)(x-x_1)\ldots(x-x_m).$$

Die zweite Identität folgt daraus, daß die rechte Seite gleich der linken Seite in den Punkten x_i, $i = 0, \ldots, m$, ist und die Koeffizienten von x^{m+1} auf beiden Seiten übereinstimmen; vgl. auch die Restgliedformel (15.8). ∎

Die Substitution $x = h^2$, $x_i = h_i^2$, in (26.5) ergibt die Beziehung

$$(26.6) \quad \sum_{i=0}^{m} h_i^{2j} L_i(0) = \begin{cases} 1 & , \ j = 0 \\ 0 & , \ j = 1, \ldots, m \\ (-1)^m h_0^2 h_1^2 \ldots h_m^2 & , \ j = m+1 \end{cases}.$$

Das Polynom $\tilde{T}_{mm}(h)$ in (26.4) interpolierte die Werte $T(h_i)$, $i = 0, \ldots, m$. Also gilt nach der Formel von Lagrange

$$(26.7) \quad T_{mm} = \tilde{T}_{mm}(0) = \sum_{i=0}^{m} L_i(0)\, T(h_i).$$

Die asymptotische Entwicklung (26.3) von $T(h)$ ergab

$$T(h) = \int_a^b f(x)\,dx + \tau_1 h^2 + \dots + \tau_m h^{2m} + \alpha_{m+1}(h) h^{2m+2},$$

$$\alpha_{m+1}(h) = \frac{1}{(2m+2)!} \int_a^b f^{(2m+2)}(x)\, K_{2m+2}\left(\frac{x-a}{h}\right) dx,$$

$$\int_a^b K_{2m+2}\left(\frac{x-a}{h}\right) dx = (-1)^m B_{m+1}(b-a). \tag{26.8}$$

Mit (26.6), (26.7) folgt dann

$$T_{mm} = \int_a^b f(x)\,dx + \frac{1}{(2m+2)!} \int_a^b f^{(2m+2)}(x)\, K(x)\,dx, \tag{26.9}$$

$$K(x) := \sum_{i=0}^{m} L_i(0)\, h_i^{2m+2}\, K_{2m+2}\left(\frac{x-a}{h_i}\right). \tag{26.10}$$

Man kann zeigen: Die Funktion $K(x)$ hat gleiches Vorzeichen in $[a,b]$ für die Romberg-Folge und die Bulirsch-Folge h_i. Daher gilt

$$\int_a^b f^{(2m+2)}(x)\, K(x)\,dx = f^{(2m+2)}(\xi) \int_a^b K(x)\,dx, \quad \xi \in [a,b],$$

und mit (26.6), (26.8), (26.10) haben wir

$$\int_a^b K(x)\,dx = \sum_{i=0}^{m} L_i(0)\, h_i^{2m+2} \int_a^b K_{2m+2}\left(\frac{x-a}{h_i}\right) dx$$

$$= (-1)^m h_0^2 h_1^2 \dots h_m^2 (-1)^m B_{m+1}(b-a)$$

Insgesamt ergibt sich dann aus (26.9) und den vorigen Bezeichnungen der Fehler

$$\boxed{T_{mm} - \int_a^b f(x)\,dx = (b-a)\, h_0^2 \dots h_m^2 \frac{B_{m+1}}{(2m+2)!} f^{(2m+2)}(\xi).} \tag{26.11}$$

Bei Interpolation mit den Schrittweiten $h_{i-k}, \dots, h_i$ erhält man auf ähnliche Weise

$$T_{ik} - \int_a^b f(x)\,dx = (b-a)\, h_{i-k}^2 \dots h_i^2 \frac{B_{k+1}}{(2k+2)!} f^{(2k+2)}(\xi). \tag{26.12}$$

Für $k=0$ gewinnt man hieraus die Abschätzung (26.2) zurück wegen $B_1 = \frac{1}{6}$.

Wegen (26.12) verhält sich der Fehler von T_{im} in der $(i+1)$-ten Spalte des Tableaus wie h_{i-m}^{2m+2}, also wie der Fehler eines Verfahrens $(2m+2)$-ter Ordnung. Aus Gründen der Auslöschung geht man in der Praxis nicht über $m=6$ hinaus. Man beendet die Rechnung, falls das erste Mal

$$|T_{i,6} - T_{i+1,6}| \leq \varepsilon \cdot s$$

erfüllt ist, wobei

ε: gewünschte relative Genauigkeit,

s: grober Näherungswert für $\int_a^b f(x)\,dx$.

§27 Allgemeines über Extrapolationsverfahren

Bei der näherungsweisen Berechnung der Lösung eines Problems verwendet man häufig Diskretisierungsverfahren mit folgenden Gegebenheiten:

Wahl einer Schrittweite: $h \neq 0$

Resultat der Rechnung: $T(h)$

Asymptotische Entwicklung des Resultates:

$$T(h) = \tau_0 + \tau_1 h^\gamma + \tau_2 h^{2\gamma} + \ldots + \tau_m h^{m\gamma} + \alpha_{m+1}(h) h^{(m+1)\gamma}, \quad \gamma > 0 \tag{27.1}$$

Hierbei ist

τ_i unabhängig von h,

$\alpha_{m+1}(h)$ beschränkt in h, $\alpha_{m+1}(h) = \tau_{m+1} + O(h)$,

$\tau_0 = \lim_{h \to 0} T(h)$ exakter Wert.

Z.B. besitzt die zusammengesetzte Trapezregel nach Satz (26.3) eine solche Entwicklung mit $\gamma = 2$.

Beispiel: Numerische Differentiation

(1) Für $h \neq 0$ ist

$$T(h) := \frac{f(x+h) - f(x)}{h}$$

eine Approximation für $f'(x)$. Durch Taylorentwicklung um den Punkt x findet man für Funktionen $f \in C^{m+1}[x-a, x+a]$ und $|h| \leq |a|$ eine Entwicklung

$$T(h) = \tau_0 + \tau_1 h + \tau_2 h^2 + \ldots + \tau_m h^m + h^{m+1}(\tau_{m+1} + O(h)),$$

$$\tau_k = \frac{f^{(k+1)}(x)}{(k+1)!}, \quad k = 0, 1, 2, \ldots, m+1.$$

Also gilt (27.1) mit $\gamma = 1$.

(2) Eine bessere Approximation von $f'(x)$ ist die zweiseitige Differenz

$$T(h) := \frac{f(x+h) - f(x-h)}{2h}.$$

Für Funktionen $f \in C^{2m+3}[x-a, x+a]$ und $|h| \leq |a|$ erhält man durch Taylorentwicklung

$$T(h) = \tau_0 + \tau_1 h^2 + \ldots + \tau_m h^{2m} + h^{2m+2}(\tau_{m+1} + O(h)),$$

$$\tau_k = \frac{f^{(2k+1)}(x)}{(2k+1)!}, \quad k = 0, 1, \ldots, m+1.$$

Also ist (27.1) mit $\gamma = 2$ erfüllt.

Extrapolationsverfahren: Man wähle eine Schrittweitenfolge

$$h_0 > h_1 > h_2 > \ldots > 0$$

und berechne die zugehörigen $T(h_i)$, $i = 0, 1, 2, \ldots$ Für $i \geq k$ bezeichnet man dann mit $\tilde{T}_{ik}(h)$ dasjenige Polynom in $x = h^\gamma$

$$\tilde{T}_{ik}(h) = b_0 + b_1 h^\gamma + \ldots + b_m h^{m\gamma}$$

mit

$$\widetilde{T}_{ik}(h_j) = T(h_j), \quad j = i-k, i-k+1, \ldots, i.$$

Die extrapolierten Werte

$$T_{ik} := \widetilde{T}_{ik}(0)$$

nimmt man als Näherungswerte für $T_0 = \lim_{h \to 0} T(h)$.

Die Berechnung der Werte T_{ik} erfolgt wie in §26 mit dem Algorithmus (15.4) von Neville; hier hat man $x_i = h_i^\gamma$ zu setzen:

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^\gamma - 1} \qquad (1 \le k \le i \le m) \tag{27.2}$$

Für die Rombergfolge $h_i = h_0/2^i$ lautet der Nenner explizit

$$(h_{i-k}/h_i)^\gamma - 1 = 2^{\gamma k} - 1.$$

Beispiele:

① Für die Funktion $f(x) = \tan \frac{\pi}{2} x$ soll die Ableitung $f'(0) = \pi/2 = 1.5707963$

durch Extrapolation aus den Werten

(a) $T(h) = \frac{1}{h}(f(h) - f(0))$ $\quad (\gamma = 1)$

(b) $T(h) = \frac{1}{2h}(f(h) - f(-h))$ $\quad (\gamma = 2)$

berechnet werden. Zu den Schrittweiten $h_0 = 0.5$, $h_i = h_0/2^i$ $(i = 1,2,3)$ ergibt die Rekursion (27.2):

(a) $\gamma = 1$

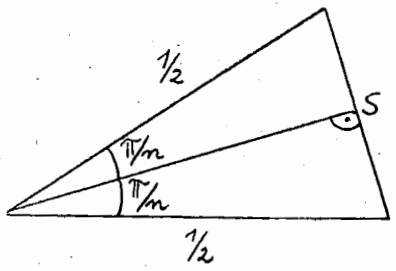
i	h_i	$T(h_i)$	T_{i1}	T_{i2}	T_{i3}
0	0.5	2.000			
1	0.25	1.65685	1.31370		
2	0.125	1.59130	1.52575	1.59643	
3	0.0625	1.57586	1.56042	1.57197	1.56382

(b) $\gamma = 2$

i	h_i	$T(h_i)$	T_{i1}	T_{i2}	T_{i3}
0	0.5	2.000			
1	0.25	1.65685	1.54247		
2	0.125	1.59130	1.56945	1.57125	
3	0.0625	1.57586	1.57072	1.57080	1.57079

② Genäherte Berechnung von π

mit Hilfe der Umfänge von einbeschriebenen, regulären n-Ecken im Kreis vom Durchmesser Eins:



Die Länge der Sehne ist $s = \sin(\pi/n)$, und deshalb ist der Umfang

$$U_n = n \sin(\pi/n)$$
$$= \pi - \frac{\pi^3}{3!}\left(\frac{1}{n}\right)^2 + \frac{\pi^5}{5!}\left(\frac{1}{n}\right)^4 - \frac{\pi^7}{7!}\left(\frac{1}{n}\right)^6 + \dots .$$

Mit $h = 1/n$ hat der Umfang gerade die Form (27.1) mit $\gamma = 2$. Elementar berechenbare Umfänge (ohne Verwendung trigonometrischer Funktionen) sind

$$U_2 = 2, \quad U_3 = \frac{3}{2}\sqrt{3}, \quad U_4 = 2\sqrt{2},$$
$$U_6 = 3, \quad U_8 = 4\sqrt{2-\sqrt{2}}.$$

Mit $x_i = h_i^2$ liefert der Neville-Algorithmus bei 10-stelliger Rechnung

x_i	T_{i0}	...	T_{i3}	T_{44}
1/4	2.000000000			
1/9	2.598076211			
1/16	2.828427125			
1/36	3.000000000		3.141588849	
1/64	3.061467459	...	3.141592411	3.141592648

54 = π

Zur Herleitung einer Fehlerabschätzung für $T_{ik} - \tau_0$ gehen wir wie in §26 vor. Aus der Lagrange'schen Interpolationsformel folgt

$$T_{ik} = \sum_{j=i-k}^{i} L_j(0)\, T(h_j)$$

mit

$$L_j(x) = \prod_{\substack{l=i-k \\ l\neq j}}^{i} \frac{x - x_l}{x_j - x_l}, \qquad x := h^\gamma, \quad x_j := h_j^\gamma .$$

Der Hilfssatz (26.5) ergibt die Beziehung

$$\sum_{j=i-k}^{i} x_j^\tau L_j(0) = \left\{ \begin{array}{ll} 1 & ; \ \tau = 0 \\ 0 & ; \ \tau = 1, \dots, k \\ (-1)^k x_{i-k} \cdots x_i & ; \ \tau = k+1 \end{array} \right\}.$$

Mit der asymptotischen Entwicklung (27.1) folgt dann:

$$T_{ik} = \sum_{j=i-k}^{i} L_j(0) \left\{ \tau_0 + \tau_1 x_j \dots + \tau_k x_j^k + x_j^{k+1} (\tau_{k+1} + O(h_j)) \right\}$$

$$= \tau_0 + (-1)^k x_{i-k} \cdots x_i \, (\tau_{k+1} + O(h_{i-k})).$$

Also gilt in Analogie zu (26.12)

$$\boxed{T_{ik} - \tau_0 = (-1)^k h_{i-k}^\gamma \cdots h_i^\gamma \, (\tau_{k+1} + O(h_{i-k}))} \tag{27.3}$$

Für festes k und für $i \to \infty$ ist also

$$T_{ik} - \tau_0 = O(h_{i-k}^{(k+1)\gamma}),$$

d.h. die Elemente T_{ik} der $(k+1)$-ten Spalte des Tableaus konvergieren gegen τ_0 wie die Resultate eines Verfahrens der Ordnung $(k+1)\gamma$. Je größer γ ist, desto bessere Resultate liefert der Extrapolationsalgorithmus.

Mit einer zusätzlichen Modifikation lassen sich Ausdrücke gewinnen, welche das Resultat τ_0 einschließen: vgl. Stoer I, S. 117.

§24 (§28.) Die Gaußsche Integrationsmethode

Sei $f \in C[a,b]$ und sei $w \in C[a,b]$ eine positive Gewichtsfunktion mit $w(x) > 0$ für $x \in (a,b)$. Wir suchen eine Integrationsformel für das Integral

$$I(f) = \int_a^b w(x) f(x) \, dx. \tag{28.1}$$

Im folgenden benutzen wir die Bezeichnungen:

Π_j: Polynome vom Grade $\le j$ $(j = 0, 1, 2, \dots)$

$\tilde{\Pi}_j := \{ p \in \Pi_j \mid p(x) = x^j + a_{j-1} x^{j-1} + \dots + a_0 \}$.

Eine Integrationsformel für $I(f)$ der Form

$$G_n(f) = \sum_{i=1}^{n} A_i f(x_i) \tag{28.2}$$

hat die $2n$ freien Parameter A_i und x_i. Die Formeln von Newton-Cotes (25.1) mit äquidistanten Stützstellen sind exakt in Π_{n-1}. Wir wollen nun fordern, daß

$$G_n(f) = I(f) \quad \text{für alle} \quad f \in \Pi_{2n-1},$$

d.h. $G_n(f)$ ist exakt in Π_{2n-1}. Dies ergibt gerade $2n$ Bedingungen für die $2n$ Parameter. Der folgende Satz zeigt, daß diese Forderung maximal ist.

(28.3) Satz: Es gibt keine Formel $G_n(f)$ des Typs (28.2), die in Π_{2n} exakt ist.

Beweis: Annahme: $G_n(f) = \int_a^b w(x) f(x)\,dx$ für $f \in \Pi_{2n}$.

Mit

$$f := \prod_{i=1}^{n} (x - x_i)^2 \in \Pi_{2n}$$

erhält man einen Widerspruch wegen

$$G_n(f) = 0 \neq \int_a^b w(x) f(x)\,dx > 0. \quad \blacksquare$$

Zur Konstruktion einer in Π_{2n-1} exakten Formel $G_n(f)$ benutzt man die zur Gewichtsfunktion w gehörenden Orthogonalpolynome $\tilde{p}_n$ bzgl. des Skalarproduktes (vgl. § 19.3)

$$(f,g) = \int_a^b f(x) g(x) w(x)\,dx, \quad f, g \in C[a,b].$$

Das Polynom $\tilde{p}_n \in \tilde{\Pi}_n$ hat nach (19.10) n Nullstellen $x_1, \dots, x_n \in (a,b)$. Damit gelangen wir zum Hauptresultat dieses Abschnittes:

(28.4) Satz: Seien $x_1, \dots, x_n$ die Nullstellen des Orthogonalpolynoms $\tilde{p}_n$ und sei

$$L_i(x) = \prod_{\substack{k=1 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}, \quad i = 1, \dots, n.$$

Dann ist die Integrationsformel

$$G_n(f) = \sum_{i=1}^{n} A_i f(x_i), \quad A_i := \int_a^b w(x) L_i(x)\,dx$$

exakt in Π_{2n-1} und es gilt

$$A_i = \int_a^b w(x) L_i(x)^2\,dx > 0.$$

Beweis: Nach der Interpolationsformel von Lagrange gilt

$$f(x) = \sum_{i=1}^{n} L_i(x) f(x_i) \quad \text{für alle } f \in \Pi_{n-1}.$$

Daher ist

$$G_n(f) = \sum_{i=1}^{n} \int_a^b w(x) L_i(x)\,dx\, f(x_i)$$

exakt in Π_{n-1}. Ein Polynom $f \in \Pi_{2n-1}$ faktorisieren wir in

$$f = q\cdot\tilde{p}_n + r \quad \text{mit } q, r \in \Pi_{n-1}$$

$$\Rightarrow \quad I(f) = \int_a^b w(x) f(x)\,dx = \underbrace{\int_a^b w(x) q(x) \tilde{p}_n(x)\,dx}_{=0,\ \text{da } q\in\Pi_{n-1}} + \int_a^b w(x) r(x)\,dx$$

$$= G_n(r) \qquad , \text{ da } r \in \Pi_{n-1}$$

$$= G_n(r) + \underbrace{G_n(q\cdot\tilde{p}_n)}_{=0,\ \text{da } \tilde{p}_n(x_i)=0,\ i=1,\dots,n}$$

$$= G_n(r + q\cdot\tilde{p}_n)$$

$$= G_n(f).$$

Also ist $G_n(f)$ exakt in Π_{2n-1} und die Formeln für die Gewichte A_i folgen so: es ist $L_i^2 \in \Pi_{2n-2}$, also

$$\int_a^b w(x) L_i(x)^2\,dx = G_n(L_i^2) = \sum_{k=1}^n A_k L_i(x_k)^2 = \sum_{k=1}^n A_k \delta_{ik} = A_i. \quad \blacksquare$$

Beispiel: $[a,b] = [-1,1]$, $w(x) \equiv 1$,

$$\tilde{p}_n(x) = \frac{n!}{(2n)!}\frac{d^n}{dx^n}(x^2-1)^n, \qquad n = 0, 1, 2, \dots,$$

Legendre-Polynome bis auf Normierungsfaktoren:

$$\tilde{p}_1(x) = x, \quad \tilde{p}_2(x) = x^2 - \tfrac{1}{3}, \quad \tilde{p}_3(x) = x^3 - \tfrac{3}{5}x$$

n	x_1	x_2	x_3	A_1	A_2	A_3
1	0			2		
2	$-\sqrt{\frac{1}{3}}$	$+\sqrt{\frac{1}{3}}$		1	1	
3	$-\sqrt{\frac{3}{5}}$	0	$\sqrt{\frac{3}{5}}$	$\frac{5}{9}$	$\frac{8}{9}$	$\frac{5}{9}$

$$I = \int_{-1}^1 e^x\,dx = 2.350402.$$

Die Simpson-Regel liefert

$$I_2 = 2.362054.$$

Dagegen ist mit gleich vielen Funktionsauswertungen

$$G_3 = 2.350337.$$

Für den Fehler der Gauß'schen Integrationsmethode gilt die folgende Abschätzung: vgl. Stoer I, S. 126.

(28.5) Satz: Sei $f \in C^{2n}[a,b]$ und sei $G_n(f)$ die in (28.4) definierte Integrationsformel, dann gilt

$$I(f) - G_n(f) = \frac{f^{(2n)}(\xi)}{(2n)!}(p_n, p_n)$$

mit einem $\xi \in (a,b)$.

Die Gauß-Formeln liefern im Vergleich zu den Newton-Cotes-Formeln bzw. den Extrapolationsverfahren die genauesten Resultate (gemessen an der Zahl der Funktionsauswertungen). Im Gegensatz zu Extrapolationsverfahren können jedoch beim Übergang von einem Index n zu $n+1$ die bis dahin berechneten Funktionswerte $f(x_i)$ nicht weiter verwendet werden. Daher sind in der Praxis Extrapolationsverfahren vorzuziehen.

§ 29 Theoretische Grundlagen gewöhnliche Differentialgleichungen

1 Typen von DGL

(1) Explizite DGL n-ter Ordnung

Sei $D \subset \mathbb{R}^{n+1}$ $(n \geq 1)$ und $f: D \to \mathbb{R}$. Sei $I \subset \mathbb{R}$ ein Intervall. Eine n-mal stetig differenzierbare Funktion $y: I \to \mathbb{R}$ (d.h. $y \in C^n(I, \mathbb{R})$) heißt Lösung der expliziten DGL n-ter Ordnung

(29.1) $$\boxed{y^{(n)} = f(x, y, y', \ldots, y^{(n-1)})}$$

im Intervall I, wenn gilt

$$y^{(n)}(x) = f(x, y(x), \ldots, y^{(n-1)}(x)),$$
$$(x, y(x), \ldots, y^{(n-1)}(x)) \in D \text{ für alle } x \in I.$$

Die DGL (29.1) heißt linear, falls

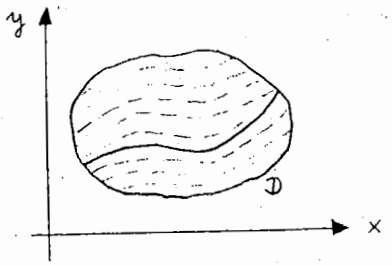
(29.2) $$y^{(n)} = a_0(x)y + a_1(x)y' + \ldots + a_{n-1}(x)y^{(n-1)}(x) + b(x)$$

mit skalaren Funktionen $a_i(x)$, $i = 0, \ldots, n-1$, $b(x)$.

Eine DGL 1. Ordnung, d.h. $n=1$ in (29.1),

(29.3) $$y' = f(x,y),$$

gestaltet eine anschauliche geometrische Deutung mit Hilfe des Richtungsfeldes:



Die Richtung φ des Rasters ist durch $\operatorname{tg}\varphi = f(x,y)$ gegeben.

(2) Systeme von DGL:

Sei $D \subset \mathbb{R}^{n+1}$ $(n \geq 1)$ und $f: D \to \mathbb{R}^n$, $f = (f_1, \dots, f_n)^T$. Sei $I \subset \mathbb{R}$ ein Intervall. Eine stetig differenzierbare Funktion $y: I \to \mathbb{R}^n$ (d.h. $y \in C^1(I, \mathbb{R}^n)$) heißt Lösung der DGL

(29.4) $\quad y' = f(x,y)$

im Intervall I, wenn gilt

$$y'(x) = f(x, y(x)), \quad (x, y(x)) \in D \text{ für alle } x \in I.$$

Komponentenweise bedeutet die DGL (29.4) mit $y(x) = (y_1(x), \dots, y_n(x))^T$:

$$\begin{aligned} y_1'(x) &= f_1(x, y_1(x), \dots, y_n(x)) \\ y_2'(x) &= f_2(x, y_1(x), \dots, y_n(x)) \\ &\vdots \\ y_n'(x) &= f_n(x, y_1(x), \dots, y_n(x)) \end{aligned}$$

Es liegt ein System linearer DGL vor, falls die Funktionen f_i affin linear in y_j sind:

$$f_i(x, y_1, \dots, y_n) = a_{i1}(x) y_1 + \dots + a_{in}(x) y_n + b_i(x), \quad (i = 1, \dots, n)$$

mit

$$A(x) = \begin{pmatrix} a_{11}(x) \dots a_{1n}(x) \\ \vdots \\ a_{n1}(x) \dots a_{nn}(x) \end{pmatrix}, \quad b(x) = \begin{pmatrix} b_1(x) \\ \vdots \\ b_n(x) \end{pmatrix}$$

lautet eine lineare DGL in vektorieller Form

(29.5) $\quad y' = A(x)\,y + b(x).$

Die explizite skalare DGL n-ter Ordnung (29.1)

$$y^{(n)} = f(x, y, y', \dots, y^{(n-1)})$$

ist äquivalent zu einem System von n DGL (29.4)

sehe dazu

$$y_1 := y, \ldots, y_k := y^{(k-1)}, \ldots, y_n := y^{(n-1)}.$$

Dann ist (29.1) äquivalent zu

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{n-1}' &= y_n \\ y_n' &= f(x, y_1, \ldots, y_n) \end{aligned}$$

(3) <u>Anfangswertaufgaben (AWA) für Systeme von DGL</u>

Gegeben sei das System von DGL $y' = f(x,y)$. Sei $I \subset \mathbb{R}$ ein Intervall und seien $x_0 \in I$, $y_0 \in \mathbb{R}^n$ mit $(x_0, y_0) \in D$. Eine Funktion $y \in C^1(I, \mathbb{R}^n)$ heißt <u>Lösung</u> der AWA

$$(29.6) \qquad \boxed{y' = f(x,y), \quad y(x_0) = y_0}\ ,$$

wenn $y(x)$ eine Lösung der DGL $y' = f(x,y)$ im Intervall I ist und die Anfangswertbedingung $y(x_0) = y_0$ erfüllt.

Für eine explizite skalare DGL n-ter Ordnung

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)})$$

lauten dann die Anfangsbedingungen

$$y^{(i)}(x_0) = y_{i0} \in \mathbb{R}, \quad i = 0, \ldots, n-1.$$

Bei der Beschreibung <u>dynamischer Systeme</u> mittels DGL führt man meistens andere Bezeichnungen ein:

$x \longrightarrow t$: Zeit
$y(x) \longrightarrow x(t)$: Zustand eines Systems zur Zeit t.

Dann geht die AWA (29.6) über in

$$(29.6\,a) \qquad \boxed{\dot{x} = \frac{dx}{dt} = f(t,x), \quad x(t_0) = x_0}\ .$$

Für die AWA (29.6) ergeben sich die folgenden Probleme:

(a) Man gebe Bedingungen an für die Existenz und Eindeutigkeit der Lösung in einem "maximalen" Existenzintervall I.

(b) Nur wenige spezielle DGL lassen sich

29.2 Existenz und Eindeutigkeit, Abhängigkeit von Anfangswerten

Sei $D \subset \mathbb{R}^{n+1}$ und sei $f: D \to \mathbb{R}^n$ stetig. Mit $\|.\|$ werde irgendeine Norm des $\mathbb{R}^n$ bezeichnet, z.B.

$$\|x\| := \max_{i=1,\dots,n} |x_i| .$$

(29.7) Definition:

(i) f genügt auf D einer Lipschitz-Bedingung bzgl. y, wenn $L \geq 0$ existiert mit

$$\|f(x,y_1) - f(x,y_2)\| \leq L \|y_1 - y_2\|$$

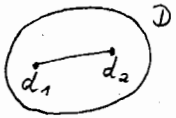
für alle $(x,y_1), (x,y_2) \in D$.

(ii) f heißt lipschitz-stetig bzgl. y auf D, wenn es zu jedem Punkt von D eine Umgebung U gibt, so daß die Einschränkung $f | D \cap U$ einer Lipschitz-Bedingung bzgl. y auf $D \cap U$ genügt.

Bem.: Die Funktion $f: \mathbb{R}^2 \to \mathbb{R}$, $f(x,y) = |y|^{\frac{1}{2}}$, $y \in \mathbb{R}$, ist lipschitz-stetig auf $D := \mathbb{R} \times (0,\infty)$, genügt aber keiner Lipschitz-Bedingung auf D.

Eine Menge $D \subset \mathbb{R}^{n+1}$ heißt konvex, falls

$$\alpha d_1 + (1-\alpha) d_2 \in D \quad \text{für alle } d_1, d_2 \in D,\ \alpha \in [0,1].$$



(29.8) Satz: (Kriterium für Lipschitz-Bedingung)

(i) Ist D konvex und sind die partiellen Ableitungen $\frac{\partial f_i}{\partial y_j}$ stetig und beschränkt in D, so genügt f in D einer Lipschitz-Bedingung.

(ii) Ist D ein Gebiet und f in D stetig differenzierbar, so ist f lipschitz-stetig in D.

Beweis: Zu (i): Aus dem Mittelwertsatz folgt

$$f_i(x,\tilde{y}) - f_i(x,y) = \sum_{j=1}^{m} \frac{\partial f_i}{\partial y_j}(x,y^*)(\tilde{y}_j - y_j),$$

mit einem Zwischenpunkt

$$y^* = \alpha \tilde{y} + (1-\alpha) y, \quad \alpha \in [0,1].$$

Mit

$$L_{ij} := \sup_{(x,y) \in D} \left| \frac{\partial f_i}{\partial y_j}(x,y) \right| < \infty, \quad L := \max_i \sum_j L_{ij} < \infty,$$

erhält man

$$\| f(x,\tilde{y}) - f(x,y) \| \le L \, \| \tilde{y} - y \| .$$

(ii) folgt sofort aus (i).

(29.9) Lokaler Existenz- und Eindeutigkeitssatz von PICARD-LINDELÖF

Die Funktion f sei auf der Menge
$Q = \{ (x,y) \mid \| x - x_0 \| \le a, \ \| y - y_0 \| \le b \}$ stetig und genüge dort einer Lipschitz-Bedingung bzgl. y. Es gelte

$$a M \le b \quad \text{mit} \quad M := \max \{ \| f(x,y) \| \mid (x,y) \in Q \}.$$

Dann existiert genau eine Lösung de AWA

$$y' = f(x,y), \quad y(x_0) = y_0$$

im Intervall $I = [x_0 - a, \ x_0 + a]$.

Beweis: Knobloch, Kappel, Abschnitt I.9 bzw. Walter (elementar bzw. mit BANACH'schem Fixpunktsatz)

Interpretation der Voraussetzungen:
In der Praxis gibt man Zahlen $a, b > 0$ vor und bestimmt

$M = \max \{ \| f(x,y) \| \mid (x,y) \in Q \}$, $Q = \{ (x,y) \mid \| x - x_0 \| \le a, \ \| y - y_0 \| \le b \}$.

Falls $aM > b$, so existiert die Lösung in $|x - x_0| \le \alpha$ mit

$$\alpha := \min \left(a, \frac{b}{M} \right).$$

Beispiel: $y' = e^{-x^2} + y^3$, $y(0) = 1$.

$a = 1$, $b = 1$, $Q = [0,1] \times [0,2]$,

$$M = \max_{(x,y) \in Q} \left(e^{-x^2} + y^3 \right) = 1 + 2^3 = 9,$$

$$\alpha = \min \left(1, \frac{1}{9} \right) = \frac{1}{9}.$$

Die Lösung existiert für $|x| \le \frac{1}{9}$ und verläuft in Q.

(29.10) Satz:

Sei D offen und sei $f \in C(D, \mathbb{R}^n)$ lipschitz-stetig auf D bzgl. y. Zu $(x_0, y_0) \in D$ gibt es ein eindeutig bestimmtes maximales Existenzintervall $I^* = (x^-, x^+)$ mit $-\infty \le x^- < x_0 < x^+ \le \infty$, so daß die AWA

$$y' = f(x,y), \quad y(x_0) = y_0$$

genau eine Lösung $y(x)$ in I^* besitzt. Die Lösung $y(x)$ kommt nach links

und rechts "dem Rand von D beliebig nahe", d.h. etwa für x^+ liegt einer der drei Fälle vor:

(1) $x^+ = \infty$

(2) $x^+ < \infty$ und $\lim_{x \to x^+} \| y(x) \| = \infty$

(3) $x^+ < \infty$ und $\lim_{x \to x^+} \inf d((x, y(x)), \partial D) = 0$.

Beweis: Knobloch, Kappel, Abschnitt III.2

(29.11) Satz: (Stetige Abhängigkeit von den Anfangswerten)
Sei D offen, $(x_0, s_0) \in D$, und sei $y(x; s_0)$ eine Lösung der AWA

$$y' = f(x,y), \quad y(x_0) = s_0$$

im kompakten Intervall I. Es existiere $\alpha > 0$ mit

$$S_\alpha := \{(x,y) \mid x \in I, \ \| y - y(x; s_0) \| \le \alpha\} \subset D,$$

so daß f auf S_α einer Lipschitz-Bedingung bzgl. y mit der Konstanten L genügt. Dann gibt es $\varepsilon > 0$, so daß die AWA

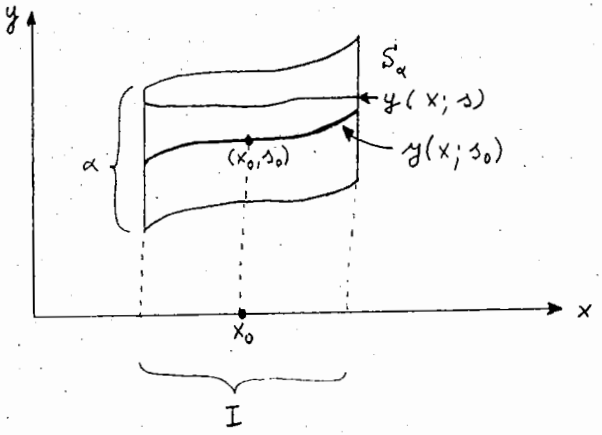
$$y' = f(x,y), \quad y(x_0) = s$$

für s mit $\| s - s_0 \| \le \varepsilon$ genau eine

Lösung $y(x; s)$ in I hat. Es gilt die Abschätzung

$$\boxed{\| y(x; s_1) - y(x; s_2) \| \le e^{L|x - x_0|} \| s_1 - s_2 \|}, \quad x \in I$$

für alle s_i mit $\| s_i - s_0 \| \le \varepsilon$, $i = 1,2$.



Beweis: Knobloch, Kappel, Abschnitt III.3;
Stoer, Bulirsch, Satz (7.1.4) (zu starke Voraussetzungen).

Die obige Abschätzung ist bestmöglich, denn die AWA $y' = Ly$, $y(x_0) = s \in \mathbb{R}$, hat die Lösung $y(x; s) = s e^{L(x - x_0)}$. Für große Werte L bzw. $x - x_0$ hängt daher die Lösung $y(x; s)$ sehr empfindlich von den Anfangswerten $s \in \mathbb{R}^n$ ab. Dies hat Konsequenzen für die Konvergenzgeschwindigkeit und die

Wahl der Schrittweite bei numerischen Verfahren: vgl. spätere Abschnitte.

Die Lösung $y(x;s)$ erfüllt nach dem letzten Satz die Identität

(29.12) $$y'(x;s) = f(x, y(x;s)), \quad y(x_0;s) = s$$

für $\|s - s_0\| \le \varepsilon$. Unter zusätzlichen Voraussetzungen kann gezeigt werden, daß die Ableitungen

$$Z(x;s) := \frac{\partial y}{\partial s}(x;s), \quad (n,n)\text{-Matrix.}$$

$$z'(x;s) := \frac{\partial^2 y}{\partial x \partial s}(x;s), \quad (n,n)\text{-Matrix}$$

existieren. Durch formale Differentation von (29.12) erhält man:

(29.13) Satz: (Differenzierbare Abhängigkeit von den Anfangswerten)

Falls zusätzlich zu den Voraussetzungen von Satz (29.11) die Funktion f stetige partielle Ableitungen auf der Menge S_α besitzt, dann existieren die Ableitungen

$$Z(x;s) := \frac{\partial y}{\partial s}(x;s), \quad z'(x;s) := \frac{\partial^2 y}{\partial x \partial s}(x;s)$$

für alle $x \in I$, $\|s - s_0\| \le \varepsilon$. Die (n,n)-Matrix

$Z(x;s)$ ist Lösung der linearen DGL (Variationsgleichung)

$$Z' = \frac{\partial f}{\partial y}(x, y(x;s))\, Z, \quad Z(x_0;s) = E_n.$$
(Einheitsmatrix)

Beweis: vgl. Knobloch, Kappel, Abschnitt III.4

Die Aussage des vorigen Satzes wird bei der numerischen Lösung von Randwertproblemen benötigt. Das Ergebnis des Satzes läßt sich leicht erweitern auf parameterabhängige AWA der Form

$$y' = f(x, y, s), \quad y(x_0) = y_0(s), \quad (s \in \mathbb{R}^k)$$

mit einer differenzierbaren Funktion $y_0 : \mathbb{R}^k \to \mathbb{R}^n$.

Literatur:

Knobloch/Kappel: Gewöhnliche Differentialgleichungen, Teubner-Verlag 1974.

Walter: Gewöhnliche DGL, HTB, Band Springer-Verlag

26

§ 30 Einschrittverfahren. Grundbegriffe

Sei $D \subset \mathbb{R}^{n+1}$ und sei $f: D \to \mathbb{R}^n$ eine C^p-Funktion, $p \in \mathbb{N}_+$. Für $(x_0, y_0) \in D$ sei $y(x)$ die Lösung der AWA

$$(30.1) \qquad y' = f(x,y), \quad y(x_0) = y_0$$

in einem geeigneten Intervall $I = [a,b]$. Zur Vereinfachung wählen wir im folgenden $\underline{n=1}$ und $D = \mathbb{R}^2$. Alle numerischen Methoden lassen sich aber unmittelbar auf den Fall $n>1$ übertragen. Auf dem Gitter

$$I_h := \{ x_i = x_0 + ih \mid i = 0,1,2,\dots \}$$

zur Schrittweite $h \neq 0$ sollen Näherungswerte η_i für $y_i = y(x_i)$ an den äquidistanten Punkten $x_i = x_0 + ih$ berechnet werden.

Motivation für Einschrittverfahren: Das Polygonzug-Verfahren von Euler

Es gilt näherungsweise

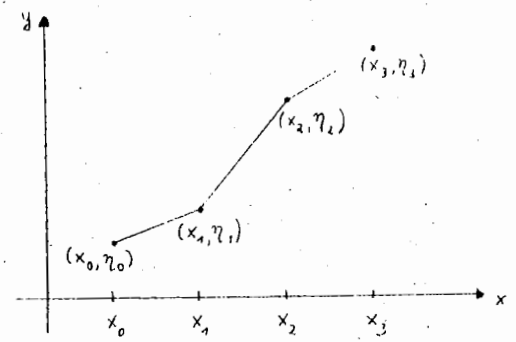
$$\frac{y(x+h) - y(x)}{h} \approx y'(x) = f(x, y(x)),$$

d.h.

$$y(x+h) \approx y(x) + h f(x, y(x)).$$

Man erhält so an den Stellen $x_i = x_0 + ih$ Näherungswerte η_i für $y_i = y(x_i)$:

$$(30.2) \qquad \boxed{\begin{aligned} &\eta_0 := y_0, \\ &\eta_{i+1} := \eta_i + h f(x_i, \eta_i), \quad i = 0,1,2,\dots \end{aligned}}$$



Allgemeine Einschrittverfahren:

Mit einer gegebenen Funktion

$$\phi(x, y, h) = \phi(x, y, h, f)$$

erhält man Näherungswerte η_i durch

$$\text{(30.3)} \qquad \boxed{\begin{aligned} &\eta_0 := y_0 \,, \\ &\eta_{i+1} := \eta_i + h\,\phi(x_i, \eta_i, h), \quad i = 0,1,2,\ldots \end{aligned}}\,.$$

Beim Euler'schen Polygonzug-Verfahren ist beispielsweise $\phi(x, y, h) = f(x, y)$; hier ist ϕ von h unabhängig.

Zur Verdeutlichung schreibt man $\eta(x_i, h) := \eta_i$. Die "Näherungslösung" $\eta(x, h)$ ist also nur für

$$x \in I_h := \{ x_0 + ih \mid i = 0,1,2,\ldots \}$$

bzw. für

$$h \in H_x := \left\{ \frac{x - x_0}{n} \;\middle|\; n = 1,2,\ldots \right\}$$

definiert, und zwar rekursiv durch

$$\text{(30.3a)} \qquad \begin{aligned} &\eta(x_0, h) := y_0 \\ &\eta(x+h, h) := \eta(x,h) + h\cdot\phi(x, \eta(x,h), h)\,. \end{aligned}$$

Seien nun x und y fest gewählt und sei $z(t)$ die exakte Lösung der AWA

$$z'(t) = f(t, z(t)), \qquad z(x) = y\,.$$

Der lokale Diskretisierungsfehler des Verfahrens (30.3a) an der Stelle (x, y) wird definiert durch

$$\text{(30.4)} \qquad \boxed{\begin{aligned} \tau(x, y, h) &= \frac{1}{h}\,(z(x+h) - \eta(x+h;\, h)) \\ &= \frac{1}{h}\,(z(x+h) - y) - \phi(x, y, h)\,. \end{aligned}}$$

Die Größe $\tau(x, y, h)$ gibt an, wie gut die exakte Lösung der DGL die Gleichung des Einschrittverfahrens erfüllt. Für eine vernünftige Einschrittmethode wird man

$$0 = \lim_{h \to 0} \tau(x, y, h) = f(x, y) - \phi(x, y, 0)$$

verlangen; ϕ sei stetig.

(30.5) **Definition:**

(i) Das Einschrittverfahren (30.3a) heißt **konsistent**, wenn gilt

$$\Phi(x,y;0) = f(x,y) \quad \text{für alle } x\in[a,b], y\in\mathbb{R}.$$

(ii) Das Einschrittverfahren hat die **Ordnung** p, falls

$$\tau(x,y,h) = O(h^p) \quad \text{für alle } x\in[a,b], y\in\mathbb{R}.$$

Zur Bestimmung der Ordnung p entwickelt man $z(t)$ in eine Taylor-Reihe um $t=x$:

$$z(x+h) = z(x) + h\,z'(x) + \frac{h^2}{2} z''(x) + \dots + \frac{h^p}{p!} z^p(x+\theta h), \qquad 0<\theta<1.$$

Hierbei ist

$$z(x) = y, \quad z'(x) = f(x,y),$$

$$z''(x) = \frac{d}{dt} f(t,z(t))\Big|_{t=x} = f_x(x,y) + f_y(x,y) f(x,y).$$

Daher gilt

$$\frac{z(x+h)-y}{h} = z'(x) + \frac{h}{2!} z''(x) + \dots + \frac{h^{p-1}}{p!} z^{(p)}(x+\theta h,$$

(30.6)

$$= f(x,y) + \frac{h}{2}\left[f_x(x,y) + f_y(x,y) f(x,y)\right] +$$

Für das Euler'sche Verfahren, $\phi(x,y,h) = f(x,y)$ folgt

$$\tau(x,y,h) = \frac{h}{2}\left[f_x(x,y) + f_y(x,y) f(x,y)\right] + \dots$$

$$= O(h).$$

Also ist die Ordnung $p=1$.

Die Entwicklung (30.6) zeigt, wie man einfach Verfahren höherer Ordnung gewinnen kann. Man nehme als $\phi(x,y,h)$ Abschnitte der Taylor-Reihe in (30.6); z.B. ergibt

$$\phi(x,y,h) := f(x,y) + \frac{h}{2}\left[f_x(x,y) + f_y(x,y) f(x,y)\right]$$

ein Verfahren 2. Ordnung.

Einfachere Verfahren höherer Ordnung erhält man z.B. mit Hilfe des Ansatzes

$$\phi(x,y,h) := a_1 \cdot f(x,y) + a_2 \cdot f(x+p_1 h, y+p_2 h f(x,y)).$$

Die Taylor-Entwicklung bzgl. h liefert

$$\phi(x,y,h) = (a_1+a_2) f + a_2 h [p_1 f_x + p_2 f_y f] + O(h^2).$$

Durch Vergleich mit (30.6) ergibt sich die Ordnung $p=2$, falls gilt

$$a_1 + a_2 = 1, \quad a_2 p_1 = 1/2, \quad a_2 p_2 = 1/2.$$

Zwei Lösungen dieser Gleichungen sind

(1) <u>das Verfahren von Heun</u>: $a_1 = a_2 = \frac{1}{2}$, $p_1 = p_2 = 1$,

$$\Phi(x,y,h) = \tfrac{1}{2}\left[f(x,y) + f(x+h, y+h f(x,y))\right], \tag{30.7}$$

(2) <u>das modifizierte Euler-Verfahren</u> (Collatz 1960)

$$a_1 = 0, \quad a_2 = 1, \quad p_1 = p_2 = \tfrac{1}{2},$$

$$\Phi(x,y,h) := f\left(x + \frac{h}{2}, y + \frac{h}{2} f(x,y)\right). \tag{30.8}$$

Das Verfahren von Runge-Kutta (1895) hat die Form

$$\Phi(x,y;h) := \tfrac{1}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right], \tag{30.9}$$

wobei
$$k_1 := f(x,y)$$
$$k_2 := f\left(x + \tfrac{h}{2}, y + \tfrac{1}{2} h k_1\right)$$
$$k_3 := f\left(x + \tfrac{h}{2}, y + \tfrac{1}{2} h k_2\right)$$
$$k_4 := f(x+h, y+h k_3).$$

Das Runge-Kutta-Verfahren hat die Ordnung <u>$p=4$</u>.

Hängt $f(x,y)$ nicht von y ab, so ist die Lösung der AWA

$$y' = f(x), \quad y(x_0) = y_0$$

gerade das Integral $y(x) = y_0 + \int_{x_0}^{x} f(t)\,dt$. Das Verfahren von Heun (30.7) entspricht dann der Trapez-Regel, das Verfahren von Runge-Kutta entspricht der Simpson-Regel in §25.

§ 31 Konvergenz und Rundungsfehlereinfluß bei Einschrittverfahren

Sei $f: [a,b] \times \mathbb{R} \to \mathbb{R}$ eine C^p-Funktion, $p > 0$, und sei $y(x)$, $x \in [a,b]$, die Lösung der AWA

$$y' = f(x,y), \quad y(x_0) = y_0.$$

Gegeben sei das Einschrittverfahren (30.3) bzw. (30.3a) mit der Näherungslösung $\eta(x,h)$. Bei festem $x \in [a,b]$ interessieren wir uns für das Verhalten des globalen Diskretisierungsfehlers

(31.1) $$e(x,h) := \eta(x,h) - y(x)$$

für $h \to 0$ mit

$$h = h_n \in H_x := \left\{ \frac{x - x_0}{n} \;\middle|\; n = 1,2,\dots \right\}.$$

Das Einschrittverfahren heißt konvergent, falls

$$\lim_{n\to\infty} e(x,h_n) = 0, \quad h_n := \frac{x-x_0}{n},$$

für alle $x \in [a,b]$.

Wir wollen zeigen, daß für ein Einschrittverfahren der Ordnung $p > 0$ (vgl. Def. (30.5)) gilt

$$e(x,h_n) = O(h_n^p).$$

(31.2) Hilfssatz: Genügen die Zahlen c_i einer Abschätzung der Form

$$|c_{i+1}| \le (1+\delta)|c_i| + B, \quad \delta > 0, \quad B \ge 0, \quad i = 0,1,$$

so gilt

$$|c_n| \le e^{n\delta}|c_0| + \frac{e^{n\delta}-1}{\delta} B.$$

Beweis: Aus den Vor. folgt rekursiv

$$|c_1| \le (1+\delta)|c_0| + B$$

$$|c_2| \le (1+\delta)^2|c_0| + B(1+\delta) + B$$

$$\dots$$

$$|c_n| \le (1+\delta)^n|c_0| + B\left[1 + (1+\delta) + (1+\delta)^2 + \dots + (1+\delta)^{n-1}\right]$$

$$= (1+\delta)^n|c_0| + B\,\frac{(1+\delta)^n - 1}{\delta}$$

$$\le e^{n\delta}|c_0| + B\,\frac{e^{n\delta}-1}{\delta}$$

wegen $1 + \delta \le e^{\delta}$ für $\delta > 0$. ∎

Damit können wir folgenden Hauptsatz beweisen, wobei $\tau(x,y,h)$ der lokale Diskretisierungsfehler in (30.4) ist.

(31.3.) Satz: Die Funktion $\Phi(x,y,h)$ sei stetig auf

$$G := \{(x,y,h) \mid a \le x \le b,\ |y-y(x)| \le \alpha,\ |h| \le h_0\},$$

$h_0 > 0$, $\alpha > 0$, und es gebe positive Konstanten M und N, so daß

$$|\Phi(x,y_1,h) - \Phi(x,y_2,h)| \le M|y_1-y_2|$$

für alle $(x,y_i,h) \in G$, $i=1,2$,

und

$$|\tau(x,y(x),h)| \le N|h|^p, \quad p>0,$$

für alle $x \in [a,b]$, $|h| \le h_0$. Dann gilt für den globalen Diskretisierungsfehler

$$|e(x,h_n)| \le |h_n|^p N \frac{e^{M|x-x_0|}-1}{M}$$

für alle $x \in [a,b]$ und $h_n = \frac{x-x_0}{n}$, $n=1,2,\ldots$, mit $|h_n| \le \bar{h} \in h_0$ für ein $\bar{h} > 0$.

Beweis: Sei $x \in [a,b]$ fest gewählt, $x \neq x_0$, und $h = h_n = (x-x_0)/n$. Nach Def. des lokalen Diskretisierungsfehlers (30.4) gilt

$$y_{i+1} = y_i + h\,\phi(x_i,y_i,h) + h\,\tau(x_i,y_i,h)$$

Das Verfahren lautet

$$\eta_{i+1} = \eta_i + h\,\phi(x_i,\eta_i,h).$$

Für den Fehler $e_i := \eta_i - y_i$ erhält man durch Subtraktion die Rekursion

$$e_0 = 0, \quad e_n = e(x,h_n),$$

$$e_{i+1} = e_i + h\,[\phi(x_i,\eta_i,h) - \phi(x_i,y_i,h)] - h\,\tau(x_i,y_i,h).$$

Falls $|e_i| = |\eta_i - y_i| \le \alpha$, so gilt nach Vor.

$$|\phi(x_i,\eta_i,h) - \phi(x_i,y_i,h)| \le M|\eta_i - y_i| = M|e_i|,$$

$$|\tau(x_i,y_i,h)| \le N|h|^p.$$

Damit folgt die Abschätzung

$$|e_{i+1}| \le (1+|h|\cdot M)\,|e_i| + N\cdot|h|^{p+1}.$$

Hilfssatz (31.2) ergibt dann mit $e_0 = 0$, $\delta := |h|\,M$, $n\delta = n|h|M = M|x-x_0|$

$$|e(x,h_n)| = |e_n| \le N\,|h_n|^p\,\frac{e^{M|x-x_0|}-1}{M}.$$

Die rechte Seite hierin ist nun kleiner als α für alle $|h_n| \le \bar{h}$ mit $\bar{h} > 0$ geeignet. Damit sind die Vor. der obigen Abschätzung erfüllt und die Beh. ist bewiesen. ∎

Die Konstanten N und M können mittels höherer Ableitungen von f abgeschätzt werden und sind daher in der Praxis kaum zugänglich; z.B. ist beim Euler-Verfahren mit $\phi(x,y,h) = f(x,y)$:

$$N \approx \tfrac{1}{2}\,|f_x(x,y(x)) + f_y(x,y(x))\,f(x,y(x))|,$$

$$M \approx \left|\frac{\partial \phi}{\partial x}\right| = |f_y(x,y(x))|.$$

Bei der praktischen Durchführung eines Einschritt-Verfahrens erhält man statt der Werte η_i gerundete Werte $\tilde{\eta}_i$. Die Rekursionen für η_i und $\tilde{\eta}_i$ lauten dann

$$\eta_{i+1} = \eta_i + h\,\phi(x_i,\eta_i,h), \qquad \eta_0 = y_0$$

(31.4)

$$\tilde{\eta}_{i+1} = \tilde{\eta}_i + h\,\phi(x_i,\tilde{\eta}_i,h) + \varepsilon_{i+1}, \quad \tilde{\eta}_0 = y_0$$

Der Einfachheit halber gelte

$$|\varepsilon_{i+1}| \le \varepsilon \quad \text{für alle } i \ge 0$$

und ϕ erfülle die Lipschitz-Bedingung von Satz (31.3). Für den Fehler

$$\tau_i := \tilde{\eta}_i - \eta_i$$

folgt dann aus (31.4) durch Subtraktion

$$\tau_{i+1} = \tau_i + h\left(\phi(x_i, \tilde{\eta}_i, h) - \phi(x_i, \eta_i, h)\right) + \varepsilon_{i+1}$$

und daher

$$|\tau_{i+1}| \le (1 + |h| \cdot M)\,|\tau_i| + \varepsilon.$$

Hilfssatz (31.2) ergibt mit $\tau_0 = 0$ und $x = x_n = x_0 + nh$, $\tau(x_i, h) = \tau_i$:

$$|\tau(x,h)| \le \frac{\varepsilon}{|h|}\,\frac{e^{M|x-x_0|} - 1}{M}.$$

Für den Gesamtfehler

$$v(x_i, h) := \tilde{\eta}_i - y_i = \tau(x_i, h) + e(x_i, h)$$

folgt dann unter den Vor. des Satzes (31.3)

$$(31.5)\quad |v(x,h)| \le \left\{ N|h|^p + \frac{\varepsilon}{|h|} \right\} \frac{e^{M|x-x_0|} - 1}{M}.$$

Bei Verkleinerung von $|h|$ über eine gewisse Grenze hinaus wächst also der Gesamtfehler $v(x,h)$ wieder; vgl. Beispiel in Stoer / Bulirsch, S. 114).

Der Konvergenzsatz (31.3) legt die Vermutung nahe, daß die Näherungslösung $\eta(x,h)$ eine <u>asymptotische Entwicklung</u> nach Potenzen von h der Form

$$(31.6)\quad \eta(x,h) = y(x) + h^p e_p(x) + h^{p+1} e_{p+1}(x) + \dots$$

besitzt. Es gilt nach GRAGG:

(31.7) <u>Satz:</u> Sei f eine C^{N+2}-Funktion und sei $\eta(x,h)$ die von einem Einschritt-Verfahren der Ordnung $p \le N$ gelieferte Näherungslösung

für die Lösung $y(x)$ der AWA

$$y' = f(x,y), \quad y(x_0) = y_0, \quad x_0 \in [a,b].$$

Dann gilt

$$\eta(x,h) = y(x) + h^p e_p(x) + h^{p+1} e_{p+1}(x) + \dots + h^N e_N(x) + h^{N+1} E_{N+1}(x,h)$$

für alle $x \in [a,b]$ und alle $h = h_n = (x - x_0)/n$, $n = 1, 2, \dots$

Dabei ist das Restglied $E_{N+1}(x,h)$ bei festem x für alle $h = h_n$, $n = 1, 2, \dots$, beschränkt.

Asymptotische Entwicklungen sind aus zwei Gründen praktisch bedeutsam:

(1) zur Anwendung von Extrapolations-Verfahren; die Entwicklung (31.6) ist von der Form (27.1) mit $\gamma = 1$,

(2) zur automatischen Schrittweiten-steuerung, vgl. STOER / BULIRSCH, §7.2.5.

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 1 , Abgabe: Di, 28.10.1997, 11.00 Uhr

Übungstermine: Klausur, Mi, 11.2 13-17 Sprache: Turbo

Gruppe 1:	Fr.	9.00 - 11.00 Uhr	SR4	BK	41
Gruppe 2:	Fr.	9.00 - 11.00 Uhr	SR5	BK	42
Gruppe 3:	Fr.	13.00 - 15.00 Uhr	SRC	BK	43
Gruppe 4:	Fr.	13.00 - 15.00 Uhr	SR4	BK	(62)

Die in den ersten beiden Aufgaben geforderten Informationen sollen mit Hilfe von Tabellenwerken, Büchern, Zeitschriften oder Computern (Netscape) in der *Mathematischen Bibliothek*, der *Rechenzentrumsbibliothek* oder der Bibliothek des *Institutes für Numerische und instrumentelle Mathematik* gefunden werden.

In jedem Fall muß die Quelle zitiert werden.

Aufgabe 1: (1+1+2+1 Punkte)

(a) Wie lautet die kleinste Primzahl, die größer als 8.000.000 ist?
Wie lautet die größte Primzahl, die kleiner als 8.000.000 ist?

(b) Was ist die Strouhal Nummer?

(c) Nennen Sie eine NAg-Routine zur Optimierung einer quadratischen Funktion mit linearen Nebenbedingungen. Beschreiben Sie die Routine und die Parameter.

(d) Nennen Sie 2 Bücher zum Thema " Operations Research", die nach 1985 erschienen sind.

Aufgabe 2: (2 Punkte)
Bestimmen Sie

$$\int_0^1 \frac{\sin(ax)}{\sqrt{1-x^2}}\,dx \qquad \text{für } a > 0 .$$

maple: int(sin(a*x)/SQRT(1-x^2), x=0..1
Mathematica: Integrate[Sin[a*x]/Sqrt[1-x^2], {x,0

Aufgabe 3: (3 Punkte)
Für

$$A := \begin{pmatrix} 2 & -1 & 2 \\ 1 & -2 & 3 \\ 3 & 3 & -1 \end{pmatrix}, \qquad b := \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}$$

bestimme man die Lösung x der Gleichung $Ax = b$ mit Hilfe des Gaußschen Eliminationsverfahrens exakt "von Hand".

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 2 , Abgabe: Di, 04.11.1997, 11.00 Uhr

Übungstermine:

Gruppe 1:	Fr.	9.00 - 11.00 Uhr	SR4	BK	41
Gruppe 2:	Fr.	13.00 - 15.00 Uhr	SRC	BK	43
Gruppe 3:	Fr.	13.00 - 15.00 Uhr	SR4	BK	62

Aufgabe 5: (3+1 Punkte)
Für $\alpha, \beta \in \mathbb{R}$, $-1 < \alpha < 1$ seien

$$A = \begin{pmatrix} \alpha & 1 & 0 \\ 1 & 0 & 2 \\ 1 & \beta & 4 \end{pmatrix} \quad , \qquad b = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} .$$

(a) Berechnen Sie mit Spaltenpivotsuche die Zerlegung $PA = LR$ und $PA = LDR$, wobei P: Permutationsmatrix, L, R: linke normierte bzw. rechte Dreiecksmatrix und D: Diagonalmatrix bedeutet.

(b) Lösen Sie $Ax = b$.

Aufgabe 6: (1+3 Punkte)
Gegeben sei das LGS

$$\begin{pmatrix} 10^{-4} & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

(a) Man berechne die exakte Lösung.

(b) Man berechne mit 3-stelliger Gleitpunktarithmetik die Lösung durch Gaußelimination

(i) ohne Pivotsuche,

(ii) mit Spaltenpivotsuche.

Geben Sie jeweils die numerisch berechnete Dreieckszerlegung an.

Aufgabe 7: (1+3 Punkte)
Sei

$$A = \begin{pmatrix} 2 & 17 & 9 \\ 4 & 2 & 2 \\ 2 & 9 & 9 \end{pmatrix} .$$

(a) Zeigen Sie: Es gibt eine Permutationsmatrix P, so daß PA positiv definit ist.

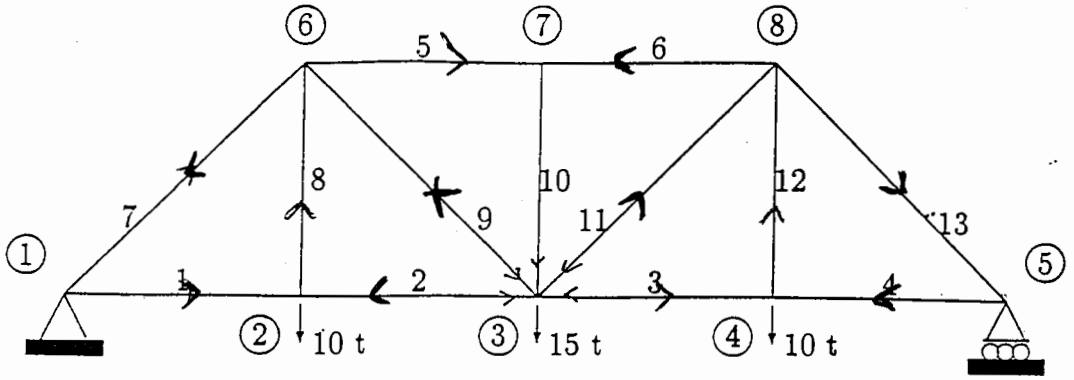
(b) Berechnen Sie die Zerlegung

$$PA = \tilde{L}\tilde{L}^T .$$

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 3 , Abgabe: Di, 11.11.1997, 11.00 Uhr

Aufgabe 9: (3 Punkte)



Die Figur stellt eine Brücke dar, die linksseitig fest montiert und rechtsseitig auf einem beweglichen Gerüst gelagert ist.

Die Knotenpunkte (1) bis (8) sind statische Gleichgewichtspunkte, d.h. die Summe der in x-Richtung wirkenden Kräfte F_x ist dort gleich Null, ebenso wie die Summe der in y-Richtung wirkenden Kräfte F_y.

Stellen Sie das zugehörige Gleichungssystem für die Kräfte f_1 bis f_{13} auf, indem Sie die Gleichgewichtszustände der Knotenpunkte ausnutzen. Beachten Sie hierbei, daß durch die feste Verankerung von Knotenpunkt (1) zwei Gleichungen entfallen und für Knotenpunkt (5) nur eine Kraftgleichung in x-Richtung benötigt wird.

Beispielsweise gilt für den Knotenpunkt (3):

$$\begin{array}{rcl} 0 = \sum F_x & = & -\alpha f_9 - f_2 + \alpha f_{11} + f_3 \\ 0 = \sum F_y & = & \alpha f_9 - f_{10} + \alpha f_{11} + 15 \end{array}$$

mit $\alpha = \sin(45°)$.

Aufgabe 10: (2 Punkte)

Betrachten Sie die Matrix

$$A = \begin{pmatrix} c & 1 & 1 \\ 1 & c & 1 \\ 1 & 1 & c \end{pmatrix}.$$

Für welche $c \in \mathbb{R}$ ist A positiv definit?

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 4 , Abgabe: Di, 18.11.1997, 11.00 Uhr

Aufgabe 13: (2 + 2 Punkte)
Formen Sie die Ausdrücke

(a) $y = \frac{1+x^2}{2+x} - \frac{2-x^2}{2(2+x)}$

(b) $y = \frac{x-\sin x}{x^3}$

so um, daß für kleine (positive) x keine Auslöschung auftritt.

Aufgabe 14: (2 + 1 Punkte)
Sei $f : \mathbb{R}^2 \to \mathbb{R}$ gegeben durch $f(x_1, x_2) = \cos(x_1)\cos(x_2)$.

(a) Bestimmen Sie die Verstärkungsfaktoren $k_{1i}(x)$, $i = 1, 2$.

(b) Sei $\tilde{x} = x + \Delta x$ eine Näherung von $x = (1.570796, 1.570797)$ mit $|\Delta x_i| \leq 0.005$. Schätzen Sie mit Hilfe der Verstärkungsfaktoren den relativen Fehler von f ab.

Aufgabe 15: (4 Punkte)
Der Ausdruck $y = \ell n(x - \sqrt{x^2 - 1})$ soll für $x = 40$ berechnet werden. Die Wurzel $w = \sqrt{x^2 - 1}$ werde hierbei einer 5-stelligen Tafel entnommen:

$$w = \sqrt{1599} \approx 39.987 .$$

(a) Wie groß wird der absolute Fehler?

(b) Man stabilisiere den Ausdruck durch Umformung und bestimme den absoluten Fehler bei der Auswertung des umgeformten Ausdruckes.

Hinweis: Betrachten Sie geeignete Funktionen $f(w)$ mit $w = \sqrt{x^2 - 1}$.

Aufgabe 16: (4 Punkte)
Man berechne die den Normen

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| , \qquad \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

zugeordneten Matrixnormen.

Aufgabe 17: (Programmieraufgabe)

Schreiben Sie ein Programm zur Berechnung von

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} = 1.644934066848\,.$$

(a) Warum erhält man bei einfacher Genauigkeit ein Ergebnis, das nur auf 4 Stellen mit dem exakten Ergebnis übereinstimmt, wenn man die Summe

$$\sum_{k=1}^{N} \frac{1}{k^2}$$

für hinreichend großes N betrachtet? Bestimmen Sie das kleinste N, ab dem keine weitere Veränderung in der Summe auftritt.

(b) Verbessern Sie den Algorithmus, indem Sie in umgekehrter Reihenfolge summieren, d.h. die Summe

$$\sum_{k=N}^{1} \frac{1}{k^2}$$

mit geeignetem N bilden. Bestimmen Sie N experimentell derart, daß der verbesserte Algorithmus einen auf 6 Stellen genauen Wert liefert.

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 5 , Abgabe: Di, 25.11.1997, 11.00 Uhr

Aufgabe 18: (3 Punkte)
Sei

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} .$$

Berechnen Sie $\text{cond}(A)$ für die Normen $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_F$, $\|\cdot\|_\infty$.

Aufgabe 19: (2+2+1+1 Punkte)
Mit $\|\cdot\|$ werde eine Vektornorm des $\mathbb{R}^n$ und die zugeordnete Matrix-Norm bezeichnet.
Beweisen Sie:

(a) $\|AB\| \leq \|A\| \, \|B\|$

(b) $\text{cond}\,(AB) \leq \text{cond}(A)\text{cond}(B)$

(c) $\text{cond}_2(A) = 1$, wenn A orthogonal

(d) $\text{cond}_2(UA) = \text{cond}_2(A)$, wenn U orthogonal

Aufgabe 20: (4 Punkte)
Berechnen Sie mit dem Householder-Verfahren (Handrechnung) die Lösung des LGS $Ax = b$ mit

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix} , \quad b = \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix} .$$

Aufgabe 21: (Programmieraufgabe)
Schreiben Sie ein Programm $\text{cond}(A, n)$ zur Berechnung der Konditionszahl einer regulären (n, n)-Matrix A zur Norm $\|\cdot\|_\infty$. Testen Sie Ihr Programm an der Matrix

$$A = \begin{pmatrix} 0.39412 & 0.51176 & 0.62941 & 0.78824 \\ 0.52857 & 0.67143 & 0.81429 & 0.74286 \\ 0.46842 & 0.62634 & 0.52105 & 0.46842 \\ 0.48462 & 0.63846 & 0.56154 & 0.48462 \end{pmatrix} .$$

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 6 , Abgabe: Di, 09.12.1997, 11.00 Uhr

Aufgabe 22: (3 Punkte)
Bestimmen Sie die lineare Ausgleichsgerade $y = \alpha + \beta t$ zu den Meßwerten

t_i	-1	1	2
y_i	3.9	1	-1.1

Aufgabe 23: (3 Punkte)
Berechnen Sie unter Verwendung der Normalgleichung die optimale kubische Funktion $y = \alpha t + \beta t^3$ zu den Meßpunkten

t_i	-2	-1	1
y_i	-2	0.7	1.7

Aufgabe 24: (1+2+3 Punkte)

a) Seien A, $B \in \mathrm{Mat}(n \times m; \mathbf{R})$, zeigen Sie:
Ist B "genügend klein" gegenüber A, so ist

$$[(A+B)^T(A+B)]^{-1} \approx (I - (A^TA)^{-1}[A^TB + B^TA])(A^TA)^{-1} .$$

Hinweis: $(I+F)^{-1} \approx I - F$, wenn F "klein" gegenüber I ist.

b) Es sei das lineare Ausgleichsproblem

$$\min_{x \in \mathbf{R}^n} \|y - Ax\|_2$$

gegeben. Zeigen Sie, daß für kleine Störungen ΔA und Δy in A und y für den Fehler Δx die folgende Abschätzung gefunden werden kann:

$$\Delta x \approx -(A^TA)^{-1}[A^T\Delta A x - \Delta A^T(y - Ax) - A^T\Delta y]$$

c) Zeigen Sie: Für eine geeignete obere Dreiecksmatrix R kann für den relativen Fehler die folgende Abschätzung gefunden werden.

$$\frac{\|\Delta x\|}{\|x\|} \le \mathrm{cond}(R)\left[\left(1 + \mathrm{cond}(R)\frac{\|y - Ax\|}{\|A\|\|x\|}\right)\frac{\|\Delta A\|}{\|A\|} + \frac{\|y\|}{\|A\|\|x\|}\frac{\|\Delta y\|}{\|y\|}\right]$$

Aufgabe 25: (Programmieraufgabe)

Schreiben Sie ein Programm zur Lösung des linearen Ausgleichsproblems

$$\min\{\|y - Ax\|_2 \mid x \in \mathbf{R}^n\} \quad \text{für} \quad y \in \mathbf{R}^m$$

und A (m,n)-Matrix mit $m > n$. Wenden Sie das Programm in §6 zur QR-Zerlegung von A auf die erweiterte Matrix $(A|y)$ an. Testen Sie das Programm an den Daten:

(a)

$$A = \begin{pmatrix} 1 & 2 & 3 \\ -1 & 0 & -1 \\ -1 & -2 & -1 \\ 1 & 0 & -1 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}; \quad x = \begin{pmatrix} 1 \\ -0.5 \\ -0.5 \end{pmatrix} \quad \text{ist optimal},$$

(b) Bestimmen Sie die Ausgleichsparabel

$$u(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$$

zu den Meßpunkten:

t_i	0.04	0.32	0.51	0.73	1.03	1.42	1.61
y_i	2.63	1.18	1.16	1.54	2.65	5.41	7.67

Abgabe von Aufgabe 25: Di 16.12.97, 11.00 Uhr

$$A = \begin{pmatrix} 1 & 0{,}04 \\ 1 & 0{,}32 \\ 1 & 0{,}51 \\ 1 & 0{,}73 \\ 1 & 1{,}03 \\ 1 & 1{,}42 \\ 1 & 1{,}61 \end{pmatrix}$$

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 7 , Abgabe: Di, 09.12.1997, 11.00 Uhr

Aufgabe 26: (4 Punkte)
Ein Versuch mit m Messungen führt auf das lineare Ausgleichsproblem mit $A \in M(m,n)$. A liege in QR-Zerlegung $A = QR$ vor. Es werde

a) ein (erster) Meßwert hinzugeführt oder

b) der (erste) Meßwert weggelassen.

Geben Sie Formeln zur Berechnung von $\tilde{Q}\tilde{R} = \tilde{A}$

$$\text{für} \quad \tilde{A} = \begin{pmatrix} \omega^T \\ A \end{pmatrix} \quad \text{bzw.} \quad A = \begin{pmatrix} z^T \\ \tilde{A} \end{pmatrix}$$

unter Benutzung der QR-Zerlegung von A an.

Aufgabe 27: (4 Punkte)
Für $\alpha > 0$ sei

$$f(x) = \begin{cases} x^\alpha & \text{für} \quad x \geq 0 \\ -|x|^\alpha & \text{für} \quad x < 0 \,. \end{cases}$$

Sei $x_0 > 0$ Startwert des Newton-Verfahrens zur Bestimmung der Nullstelle $\overline{x} = 0$ von f.

(a) Für welche $\alpha > 0$ konvergiert das Newton-Verfahren?

(b) Zeigen Sie, daß das Newton-Verfahren für $\alpha > \frac{1}{2}$, $\alpha \neq 1$, nur linear konvergiert.

Aufgabe 28: (4 Punkte)
Bei der Bahnbestimmung von Planeten ist die "KEPLERsche Gleichung" zu lösen:
Gesucht wird die "exzentrische Anomalie" $\overline{x}$ als Lösung der Gleichung

$$x = g(x) = e\sin(x) + \frac{2\pi}{u}t \,.$$

Dabei ist u die Umlaufzeit, t die seit dem Perihelddurchgang vergangene Zeit in Tagen und e die numerische Exzentrizität der Bahnellipse. Für die realistischen Werte $e = 0.1$ und $\frac{2\pi t}{u} = 0.85$ berechne man die ersten 3 Iterationen zur Lösung obiger Gleichung mit Hilfe a) der Fixpunktiterationen, b) des Newton-Verfahrens, c) der Sekantenmethode "von Hand". Geben Sie in jedem Iterationsschritt die Näherung für $\overline{x}$ an.
(Zum Vergleich: $\overline{x} = 0.9301722932$)

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 8 , Abgabe: Di, 06.01.1998, 11.00 Uhr

Aufgabe 30: (4 Punkte)
Sei $g : \mathbf{R}^2 \to \mathbf{R}^2$ die Abbildung mit

$$g(x_1, x_2) = \frac{1}{5} \begin{pmatrix} 2x_2 & - & x_1 x_2 & + & 1 \\ x_1^2 & - & x_2 & + & 3 \end{pmatrix} .$$

(a) Zeigen Sie, daß die Abbildung g auf der Menge $D = [-1, 1] \times [-1, 1] \subset \mathbf{R}^2$ genau einen Fixpunkt $\overline{x} \in D$ besitzt.

(b) Berechnen Sie $x^1 = g(x^0)$ mit $x^0 = (0.5, 0.5)^T$. Wieviele Iterationen k benötigt man, um die Genauigkeit $\|\overline{x} - x^k\|_\infty \leq 0.01$ für den Fixpunkt $\overline{x} \in D$ zu erzielen?

Aufgabe 31: (3+5 Punkte)

(a) Konvergenzverbesserung nach AITKEN:
Die Folge $\{x_k\} \subset \mathbf{R}$ sei linear konvergent gegen $\overline{x} \in \mathbf{R}$, d.h. es gelte

$$x_{k+1} - \overline{x} = (q + \varepsilon_k)(x_k - \overline{x}) , \quad |q| < 1 , \quad \varepsilon_k \to 0 .$$

Zeigen Sie: Gilt $x_k \neq \overline{x}$, so ist für genügend großes k die Folge

$$z_k := x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k}$$

erklärt und es gilt

$$\lim_{k \to \infty} \frac{z_k - \overline{x}}{x_k - \overline{x}} = 0 .$$

Hinweis: Setzen Sie $e_k = x_k - \overline{x}$ und überlegen Sie

$$x_{k+1} - x_k = e_{k+1} - e_k , \quad x_{k+2} - 2x_{k+1} + x_k = e_{k+2} - 2e_{k+1} + e_k .$$

(b) Verfahren von STEFFENSEN: Sei $f \in C^2(\mathbf{R})$ und sei $\overline{x}$ ein Fixpunkt von f von $|f'(\overline{x})| < 1$. Durch Kombination der Fixpunktiteration $x_{k+1} = f(x_k)$ mit dem Verfahren von AITKEN in (a) erhält man die Iteration

$$x_{k+1} = g(x_k) := x_k - \frac{(f(x_k) - x_k)^2}{f(f(x_k)) - 2f(x_k) + x_k} .$$

Zeigen Sie, daß dieses Verfahren (mindestens) die Ordnung $p = 2$ hat:

Hinweis: Sei o.E. $\overline{x} = 0$. Benutzen Sie $f(x) = f'(0)x + O(x^2)$ und zeigen Sie $g(x) = O(x^2)$.

Berechnen Sie z_k $(k = 0, 1, 2, 3)$ für die Iterationsfolge $x_{k+1} = g(x_k) = \exp(-x_k)$, $x_0 = 0.55$.

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 9 , Abgabe: Di, 20.01.1998, 11.00 Uhr

Aufgabe 34: (4 Punkte)
Berechnen Sie mit dem Verfahren von Maehly die beiden größten Nullstellen des Polynoms

$$p(x) = x^5 - 7x^3 + 6x - 2\,.$$

Aufgabe 35: (3 Punkte)
Für das Polynom $p(x) = x^4 - 7x^3 - 9x^2 + 5x + 2$ berechne man mit Hilfe des HORNER-Schemas die Ableitungen $p^{(i)}(x_0)$, $i = 0, 1, 2, 3, 4$, an der Stelle $x_0 = -\frac{1}{2}$.

Aufgabe 36: (4 Punkte)
Zur Lösung des LGS

$$\begin{pmatrix} 9 & 0.6 \\ 0.6 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 9 \\ 10 \end{pmatrix}$$

führe man ausgehend von $x^0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ zwei Schritte des Gesamt- und Einzelschrittverfahrens durch.

Aufgabe 37: (4 Punkte)
Sei A eine (m, n)-Matrix mit $n \leq m$ und $\operatorname{rang}(A) = n$. Sei $0 < r < \frac{2}{\sigma_1^2}$, wobei σ_1^2 der größte Eigenwert von $A^T A$ ist.
Zeigen Sie: Die Iteration

$$x^{(k+1)} = x^{(k)} - rA^T(Ax^{(k)} - y)$$

konvergiert für jede Wahl von $x^{(0)}$, y gegen eine Lösung des linearen Ausgleichsproblems

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2\,.$$

Bitte denken Sie an die Abgabe der Programmieraufgabe 33.

Aufgabe 34:

$p(x) = x^5 - 7x^3 + 6x - 2$

$p'(x) = 5x^4 - 21x^2 + 6$

Laut Satz (12.7) gilt für alle Nullstellen $z \in \mathbb{R}$.

$$|z| \leq \max\{2, 7, 8\} = 8$$

$$\equiv \max\left\{\left|\frac{a_n}{a_0}\right| + 1 + \left|\frac{a_{n-1}}{a_0}\right| + \dots + \left|\frac{a_1}{a_0}\right|\right\}$$

$x_0 := 8$ soll Startwert der Newton-Iteration sein.

$$x_1 = x_0 - \frac{p(x_0)}{p'(x_0)} = 8 - \frac{8^5 - 7\cdot 8^3 + 6\cdot 8 - 2}{5\cdot 8^4 - 21\cdot 8^2 + 6}$$

$$= 6,47233\dots$$

$$x_2 = x_1 - \frac{p(x_1)}{p'(x_1)} = 5,27079\dots$$

$x_3 = 4,33447\dots$

$x_4 = 3,61958\dots$

$x_5 = 3,09496\dots$

$\vdots$

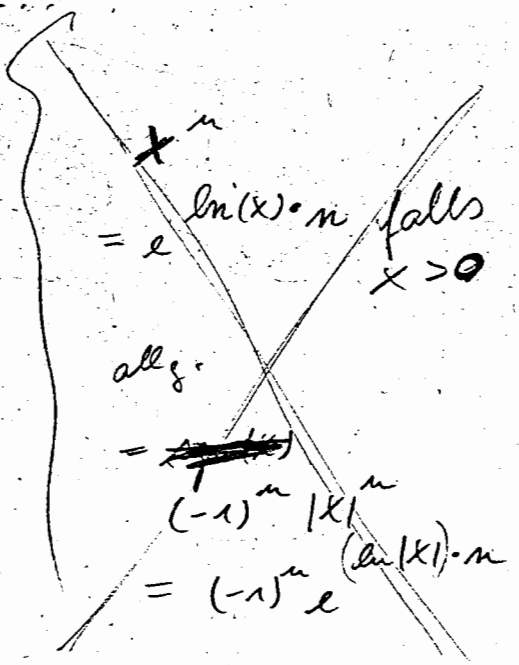
$x_n = 2,48119430411 \approx z_1$

$z_2 = 0,6888921825$

$z_3 = 0,41421356241$

$z_4 = -1,17008648661$

$z_5 = -2,41421356241$



Einzelschrittverfahren:

Iteration: $\begin{pmatrix} 9 & 0 \\ 0{,}6 & 5 \end{pmatrix} x^{k+1} + \begin{pmatrix} 0 & 0{,}6 \\ 0 & 0 \end{pmatrix} x^k = \begin{pmatrix} 9 \\ 10 \end{pmatrix}$

Berechnung der Inversen davon:

$$\left(\begin{array}{cc|cc} 1 & 0 & 9 & 0 \\ 0 & 1 & 0{,}6 & 5 \end{array}\right) \quad | : 9$$

$$\rightarrow \left(\begin{array}{cc|cc} \frac{1}{9} & 0 & 1 & 0 \\ 0 & 1 & 0{,}6 & 5 \end{array}\right) \quad \text{II} - \frac{3}{5}\,\text{I}$$

$$\rightarrow \left(\begin{array}{cc|cc} \frac{1}{9} & 0 & 1 & 0 \\ -\frac{3}{45} & 1 & 0 & 5 \end{array}\right) \quad | : 5$$

$$\rightarrow \left(\begin{array}{cc|cc} \frac{1}{9} & 0 & 1 & 0 \\ -\frac{1}{75} & \frac{1}{5} & 0 & 1 \end{array}\right)$$

$$x^{k+1} = -\begin{pmatrix} \frac{1}{9} & 0 \\ -\frac{1}{75} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} 0 & \frac{3}{5} \\ 0 & 0 \end{pmatrix} x^k + \begin{pmatrix} 1 \\ 2{,}12 \end{pmatrix} 1{,}88$$

$$= -\begin{pmatrix} 0 & \frac{3}{45} \\ 0 & -\frac{1}{125} \end{pmatrix} x^k + \begin{pmatrix} 1 \\ 1{,}88 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1{,}88 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{15} \\ 0 & -\frac{1}{125} \end{pmatrix} x^k$$

$$x^0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$x^1 = \begin{pmatrix} 1 \\ 1{,}88 \end{pmatrix} - \begin{pmatrix} \frac{2}{15} \\ -\frac{2}{125} \end{pmatrix} = \begin{pmatrix} 0{,}8\overline{6} \\ 1{,}896 \end{pmatrix}$$

$$x^2 = \begin{pmatrix} 1 \\ 1{,}88 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{15} \\ 0 & -\frac{1}{125} \end{pmatrix} \begin{pmatrix} 0{,}8\overline{6} \\ 1{,}896 \end{pmatrix} = \begin{pmatrix} 0{,}8736 \\ 1{,}895168 \end{pmatrix}$$

Aufgabe 36:

Gesamtschrittverfahren:

Iteration: $\begin{pmatrix} 9 & 0 \\ 0 & 5 \end{pmatrix} x^{k+1} + \begin{pmatrix} 0 & 0{,}6 \\ 0{,}6 & 0 \end{pmatrix} x^{k} = \begin{pmatrix} 9 \\ 10 \end{pmatrix}$

$$\Rightarrow \quad x^{k+1} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 & \frac{0{,}6}{9} \\ \frac{0{,}6}{5} & 0 \end{pmatrix} x^{k}$$

$$= \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 & \frac{1}{15} \\ \frac{3}{25} & 0 \end{pmatrix} x^{k}$$

$$= \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 & 0{,}0\overline{6} \\ 0{,}12 & 0 \end{pmatrix} x^{k}$$

$$x^0 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$x^1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 & 0{,}0\overline{6} \\ 0{,}12 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0{,}1\overline{3} \\ 0{,}12 \end{pmatrix}$$

$$= \underline{\underline{\begin{pmatrix} 0{,}8\overline{6} \\ 1{,}88 \end{pmatrix}}}$$

$$x^2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0 & 0{,}0\overline{6} \\ 0{,}12 & 0 \end{pmatrix} \begin{pmatrix} 0{,}8\overline{6} \\ 1{,}88 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 0{,}125\overline{3} \\ 0{,}104 \end{pmatrix}$$

$$= \underline{\underline{\begin{pmatrix} 0{,}87\overline{46} \\ 1{,}896 \end{pmatrix}}}$$

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 10 , Abgabe: Di, 27.01.1998, 11.00 Uhr

Aufgabe 37: (4 Punkte)
Sei A eine (m,n)-Matrix mit $n \leq m$ und $\text{rang}(A) = n$. Sei $0 < r < \frac{2}{\sigma_1^2}$, wobei σ_1^2 der größte Eigenwert von A^TA ist.

Zeigen Sie: Die Iteration

$$x^{(k+1)} = x^{(k)} - rA^T(Ax^{(k)} - y)$$

konvergiert für jede Wahl von $x^{(0)}$, y gegen eine Lösung des linearen Ausgleichsproblems

$$\min_{x \in \mathbb{R}^n} \|y - Ax\|_2 .$$

Aufgabe 38: (3+1 Punkte)
Betrachten Sie zu

$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 5 \end{pmatrix} \quad , \quad b = \begin{pmatrix} 1 \\ 1 \\ 6 \end{pmatrix}$$

das Gesamtschrittverfahren $x^{(k+1)} = C_G x^{(k)} + d$ mit $C_G = -D^{-1}(L+R)$, $d = D^{-1}b$.

(a) Überlegen Sie, daß das starke Zeilensummenkriterium nicht erfüllt ist. Zeigen Sie, daß das GS-Verfahren konvergent ist. Berechnen Sie dazu $\rho(C_G)$.

(b) Berechnen Sie $x^{(1)}$ mit $x^{(0)} = (0,1,1)^T$.

Aufgabe 39: (4 Punkte)
Seien $C_G = -D^{-1}(L+R)$, $C_E = -(L+D)^{-1}R$ die Matrizen des Gesamt- und Einzelschrittverfahrens.

Man zeige: Falls

$$\sum_{k \neq i} |a_{ik}| < |a_{ii}| , \quad i = 1, \ldots, n ,$$

so gilt

$$\|C_E\|_\infty \leq \|C_G\|_\infty < 1 .$$

Hinweis: Für $x \in \mathbb{R}^n$, $y := C_E x$ zeige man induktiv

$$|y_k| \leq \|C_G\|_\infty \|x\|_\infty , \quad k = 1, \ldots, n .$$

Move my way

Übungen zur Vorlesung Numerische Mathematik I

Übungsblatt 11 , Abgabe: Donnerstag, 05.02.1998, 11.00 Uhr

Aufgabe 41: (2+2+3 Punkte)
Sei $p \in \Pi_3$ das Polynom, das die Funktion $f(x) = \sqrt{x}$ an den Stützpunkten $x_0 = 0$, $x_1 = 1$, $x_2 = 4$, $x_3 = 9$ interpretiert.

(a) Bestimmen Sie p mit der Formel von Lagrange.

(b) Bestimmen Sie p mit der Newtonschen Interpolationsformel.

(c) Das Polynom $p_2 \in \Pi_2$ interpoliere $f(x) = \sqrt{x}$ in den Knoten 1, 2, 3. Geben Sie eine Abschätzung für den Interpolationsfehler im Intervall $[1, 3]$.

Aufgabe 42: (2 Punkte)
Sei $p \in \prod_2$ das Interpolationspolynom zu gegebenen Stützwerten

x_j	-1	0	2
f_j	-2	-3	1

mit $p(x_j) = f_j$, $j = 0, 1, 2$.
Berechnen Sie mit dem Algorithmus von Neville zu $x = 1$ den Wert $p(x)$.

Aufgabe 43: (5 Punkte)
Der Streckenzug $s(x)$ interpoliere $f(x) = \ell n(x)$ in den äquidistanten Knoten $x_i = 2 + ih$, $h = 1/n$, $i = 0, \ldots, n$, d.h.

$$s(x) = (1 - t) \cdot \ell n(x_i) + t \cdot \ell n(x_{i+1}) , \quad x_i \le x \le x_{i+1} , \quad t = (x - x_i)/h .$$

Bestimmen Sie ein möglichst kleines n so, daß

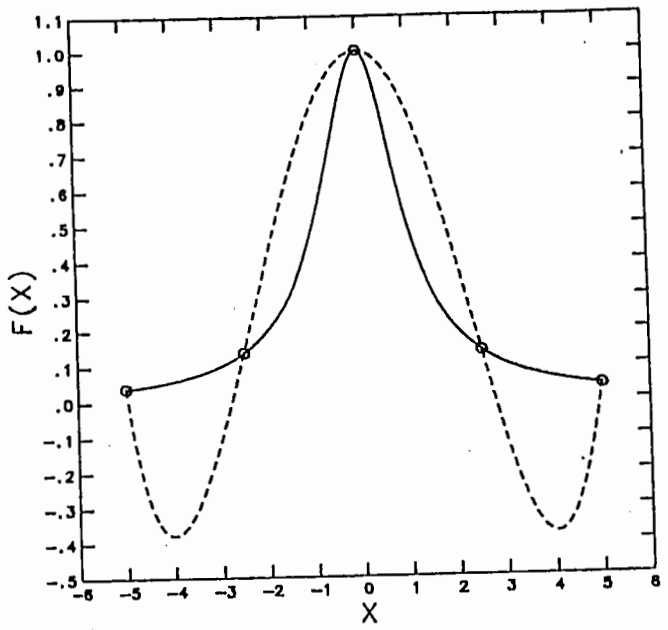
$$\max_{2 \le x \le 3} |s(x) - \ell n(x)| \le 10^{-3} .$$

Hinweis: Restgliedformel für $x_i \le x \le x_{i+1}$.

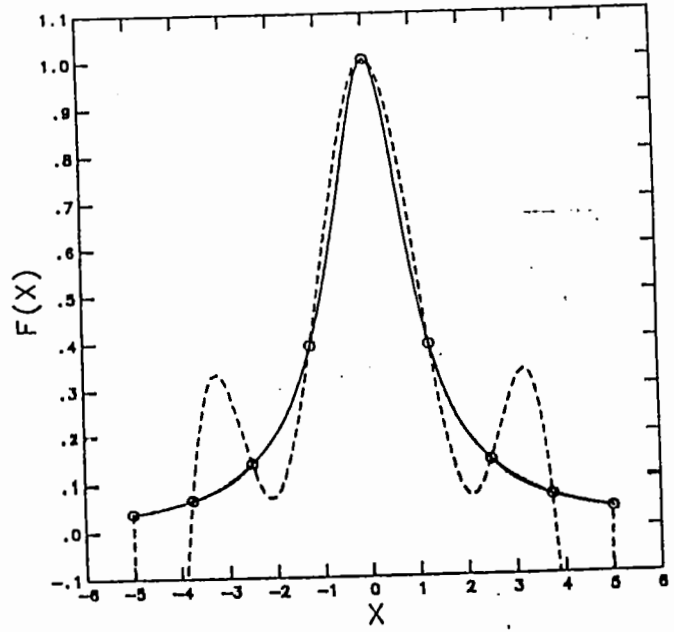
Bitte denken Sie an die Abgabe von Aufgabe 40.

KLAUSUR: Mittwoch, 11. Februar 1998, 13:15 - 16:30 Uhr,
Hörsaal: M 2,
Bitte bringen Sie Ihren Studentenausweis zur Klausur mit.
Erlaubte Hilsmittel: Vorlesungsskriptum (zusammengeheftet),
keine Übungsaufgaben!

POLYNOM-INTERPOLATION VON 1/(1+X*X) :

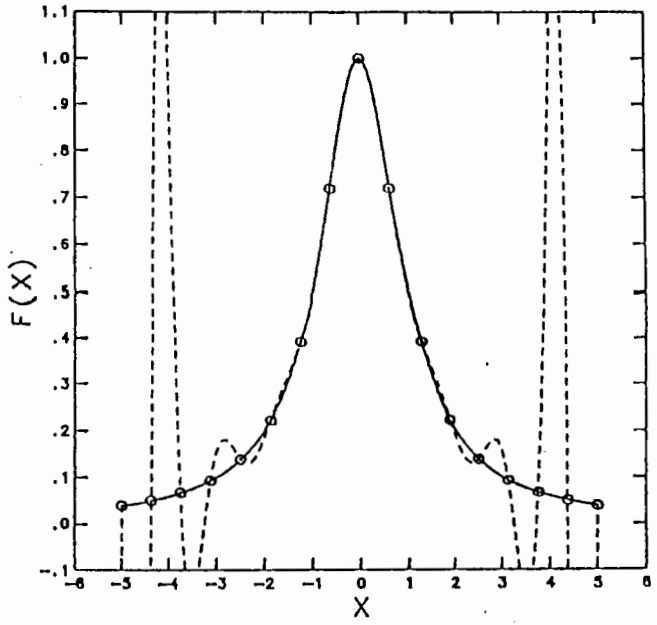


n = 4

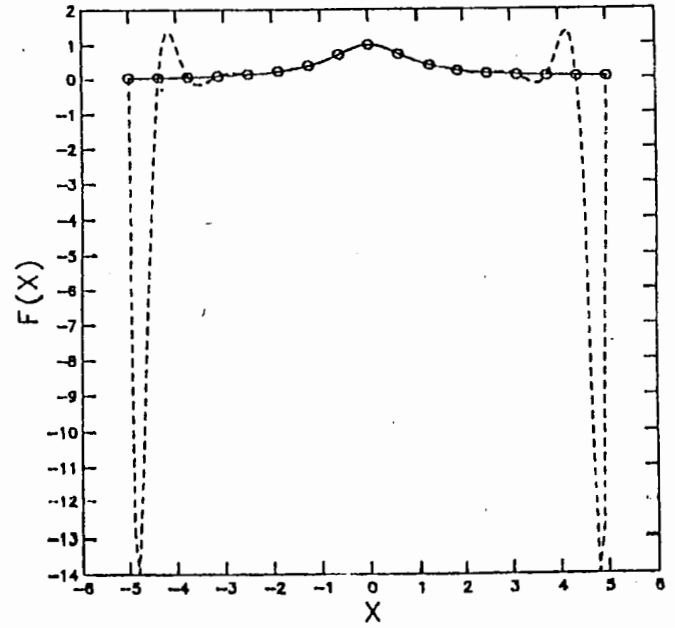


n = 8

POLYNOM-INTERPOLATION VON 1/(1+X*X)



n = 16



n = 16

Aufabe 1: (GAUSS-Elemination)
Gegeben seien

$$A = \begin{pmatrix} 2 & 17 & 9 \\ 4 & 2 & 2 \\ 2 & 9 & 9 \end{pmatrix}, \quad b = \begin{pmatrix} 10 \\ 3 \\ 10 \end{pmatrix}$$

(a) Berechnen Sie mit Spaltenpivotsuche die Zerlegung $PA = LR$ und $PA = LDR$.

(b) Lösen Sie: $Ax = b$

(c) Überlegen Sie, daß PA positiv definit ist und berechnen Sie die Zerlegung $PA = \tilde{L}\tilde{L}^T$.

Aufgabe 2: (Cholesky-Zerlegung)
Berechnen Sie explizit die CHOLESKY-Zerlegung $A = LL^T$ der 2×2-Matrix

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

Sei $a = 2$ und $c = 8$. Für welche b ist A positiv definit?

Aufgabe 3: (Kondition und Norm)
Sei

$$A = \begin{pmatrix} 4 & 5 \\ 5 & 6 \end{pmatrix}.$$

Berechnen Sie $cond(A)$ für die Normen $\|\cdot\|_1, \quad \|\cdot\|_2, \quad \|\cdot\|_F, \quad \|\cdot\|_\infty$.

Aufgabe 4: (QR-Zerlegung, Householder-Verfahren)
Berechnen Sie mit dem Householder-Verfahren (Handrechnung) die Lösung des LGS $A\,x = b$ mit

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Aufgabe 5: (Lineares Ausgleichsproblem, QR-Zerlegung)
Berechnen Sie die Lösung $x \in \mathbf{R}$ des linearen Ausgleichsproblems ($m = 2,\ n = 1$)

$$\min_{x \in \mathbf{R}} \|y - A\,x\|_2, \qquad y = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

mittels

(a) der Normalengleichungen

(b) der QR-Zerlegung von A.

Bestimmen Sie jeweils das Residuum.

Aufgabe 6: (Lineares Ausgleichsproblem)
Durch die Meßpunkte

t_i	0	1	4
y_i	2	3	4

soll eine Ausgleichsfunktion $u(t) = \alpha + \beta\sqrt{t}, \quad \alpha, \beta \in \mathbf{R}$, gelegt werden. Formuliere das zugehörige lineare Ausgleichsproblem und berechne die optimalen Parameter α, β.

Aufgabe 7: (Banachscher Fixpunktsatz)
Sei $g : \mathbf{R}^2 \longrightarrow \mathbf{R}^2$ die Abbildung mit

$$g(x_1, x_2) = \frac{1}{4}\begin{pmatrix} x_2 & - & x_1 x_2 & + & 1 \\ x_1^2 & - & x_2 & + & 2 \end{pmatrix}$$

(a) Zeige, daß die Abbildung g auf der Menge $D = [-1, 1] \times [-1, 1] \subset \mathbf{R}^2$ genau einen Fixpunkt $\bar{x} \in D$ besitzt.

(b) Berechne $x^1 = g(x^0)$ mit $x^0 = (0.5, 0.5)^T$. Wieviele Iterationen k benötigt man, um die Genauigkeit $\|\bar{x} - x^k\|_\infty \leq 0.01$ für den Fixpunkt $\bar{x} \in D$ zu erzielen?

Aufgabe 8: (Iterative Lösung linearer Gleichungssysteme)
Sei A eine symmetrische positiv definite (n, n)–Matrix und sei $b \in \mathbf{R}^n$. Zur iterativen Lösung des LGS $Ax = b$ betrachte man das RICHARDSON-Verfahren

$$x^{(k+1)} = x^{(k)} - Ax^{(k)} + b.$$

Zeige: Dieses Verfahren ist genau dann konvergent, wenn $\lambda_{max}(A) < 2$.

Aufabe 1: (GAUSS-Elemination)
Teil (a):

$$PA = LR: \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 17 & 9 \\ 4 & 2 & 2 \\ 2 & 9 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 4 & 2 & 2 \\ 0 & 16 & 8 \\ 0 & 0 & 4 \end{pmatrix}$$

$$PA = LDR: \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 17 & 9 \\ 4 & 2 & 2 \\ 2 & 9 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1/2 & 1/ \\ 0 & 1 & 1/ \\ 0 & 0 & 1 \end{pmatrix}$$

Teil (b):
Durch Rückwärtselemination:

$$x = \begin{pmatrix} 7/32 \\ 17/16 \end{pmatrix}$$

Teil (c):

$$x^T PAx = x^T LDL^T x = (L^T x)^T D(L^T x) = y^T Dy > 0$$

$$\tilde{L} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 1 & 2 & 2 \end{pmatrix}$$

Aufgabe 2: Nach Cholesky-Algorithmus:

$$L = \begin{pmatrix} \sqrt{a} & 0 \\ \frac{b}{\sqrt{a}} & \sqrt{c - \frac{b^2}{a}} \end{pmatrix}$$

Nach Determinantenkriterium müssen a und $c - b^2/a$ größer als 0 sein. Dieses deckt sich mit der erhaltenen Matrixstruktur. Für $a = 2$ und $c = 8$ ergibt sich: $|b| < 4$.

Aufgabe 3: (Kondition und Norm)
Die Inverse A^{-1} zu A lautet

$$A^{-1} = \begin{pmatrix} -6 & 5 \\ 5 & -4 \end{pmatrix}.$$

Damit lassen sich die Konditionen berechnen zu:

$cond_1(A) = \|A\|_1 \cdot \|A^{-1}\|_1 = 11 \cdot 11 = 121,$
$cond_2(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \sqrt{51 + 10\sqrt{26}} \cdot \sqrt{51 + 10\sqrt{26}} = 51 + 10\sqrt{26}$
$\approx 101.99,$
$cond_F(A) = \|A\|_F \cdot \|A^{-1}\|_F = \sqrt{102} \cdot \sqrt{102} = 102,$
$cond_\infty(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty = 11 \cdot 11 = 121.$

Aufgabe 4: (QR-Zerlegung, Householder-Verfahren)
Folgende Lösungen sind alle auf die Darstellung $a + b\sqrt{c} \quad (a, b \in \mathbb{Q}, \quad c \in \mathbb{Z})$ genormt. Der Ausdruck $\frac{1}{3+\sqrt{3}}$ läßt sich z.B. umformen zu $\frac{1}{2} - \frac{1}{6}\sqrt{3}$. Wie? So:

$$\frac{1}{3+\sqrt{3}} = x + y\sqrt{3} \Rightarrow 1 = (3x + 3y) + (x + 3y)\sqrt{3} \Rightarrow \left\{ \begin{array}{r} 3x + 3y = 1 \\ x + 3y = 0 \end{array} \right\}$$

$$\Rightarrow \left\{ \begin{array}{l} 2x = 1 \\ y = -\frac{1}{3}x \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} x = \frac{1}{2} \\ y = -\frac{1}{6} \end{array} \right\} !$$

$$Q_1 = \begin{pmatrix} -\frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} \\ -\frac{1}{3}\sqrt{3} & \frac{1}{2} + \frac{1}{6}\sqrt{3} & -\frac{1}{2} + \frac{1}{6}\sqrt{3} \\ -\frac{1}{3}\sqrt{3} & -\frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{2} + \frac{1}{6}\sqrt{3} \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2}\sqrt{2} & -\frac{1}{2}\sqrt{2} \\ 0 & -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix}$$

Q ergibt sich damit zu $Q = Q_1 \cdot Q_2 = \begin{pmatrix} -\frac{1}{3}\sqrt{3} & \frac{1}{3}\sqrt{6} & 0 \\ -\frac{1}{3}\sqrt{3} & -\frac{1}{6}\sqrt{6} & -\frac{1}{2}\sqrt{2} \\ -\frac{1}{3}\sqrt{3} & -\frac{1}{6}\sqrt{6} & \frac{1}{2}\sqrt{2} \end{pmatrix}.$

$$R = \begin{pmatrix} -\sqrt{3} & -\frac{2}{3}\sqrt{3} & -\frac{1}{3}\sqrt{3} \\ 0 & -\frac{1}{3}\sqrt{6} & -\frac{1}{6}\sqrt{6} \\ 0 & 0 & \frac{1}{2}\sqrt{2} \end{pmatrix}, \quad c = Q^T b = \begin{pmatrix} -2\sqrt{3} \\ -\frac{1}{2}\sqrt{6} \\ \frac{1}{2}\sqrt{2} \end{pmatrix}.$$

Die Lösung des Gleichungssystems $Ax = b$ bzw. $Rx = c \quad (x = (x_1\ x_2\ x_3)^T)$ ist dann

$$x_3 = \frac{1}{2}\sqrt{2}/\left(\frac{1}{2}\sqrt{2}\right) = 1, \quad x_2 = \left(-\frac{1}{2}\sqrt{6} + \frac{1}{6}\sqrt{6}\right)/\left(-\frac{1}{3}\sqrt{6}\right) = 1,$$

$$x_1 = \left(-2\sqrt{3} + \frac{1}{3}\sqrt{3} + \frac{2}{3}\sqrt{3}\right)/\left(-\sqrt{3}\right) = 1.$$

Aufgabe 5: (Lineares Ausgleichsproblem, QR-Zerlegung)
zu (a): Lösung der Normalengleichung $A^T A x_0 = A^T y$:

$$(1 \quad 2)\begin{pmatrix} 1 \\ 2 \end{pmatrix} x_0 = (1 \quad 2)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3$$

$$\Rightarrow 5\ x_0 = 3 \Rightarrow x_0 = \frac{3}{5}$$

$\Rightarrow$ Residuum $r = \left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \end{pmatrix} \cdot \frac{3}{5} \right\|_2 = \frac{1}{\sqrt{5}}$

zu (b): QR-Zerlegung von A:

$$Q = \begin{pmatrix} -\frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{pmatrix}, \quad R = \begin{pmatrix} -\sqrt{5} \end{pmatrix}.$$

$\Rightarrow Qy = \begin{pmatrix} -\frac{3}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{pmatrix} \Rightarrow r = \left\| -\frac{1}{\sqrt{5}} \right\|_2 = \frac{1}{\sqrt{5}}$

und $x_0 = R^{-1} h_1 = -\frac{1}{\sqrt{5}} \cdot \left(-\frac{3}{\sqrt{5}}\right) = \frac{3}{5}$.

Aufgabe 6 :
Zur Formulierung des linearen Ausgleichproblems werden benötigt:

$$A = \begin{pmatrix} 0^0 & \sqrt{0} \\ 1^0 & \sqrt{1} \\ 4^0 & \sqrt{4} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, y = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}.$$

Die Lösung des linearen Ausgleichsproblems

$$\min_{x \in \mathbb{R}^2} \|Ax - y\|_2^2 = \min_{\alpha,\beta \in \mathbb{R}} \| \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} \|_2^2$$

wird mittels der Normalengleichung $A^TAx = A^Ty$ gelöst zu

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Damit ergibt sich die Ausgleichsfunktion zu $u(t) = 2 + \sqrt{t}$.

Aufgabe 7 :
a) Die Voraussetzungen des BNFS sind erfüllt, da

(1) D abgeschlossen,

(2) wegen

$$|\frac{1}{4}(x_2 - x_1x_2 + 1)| \leq \frac{3}{4} \quad und \quad |\frac{1}{4}(x_1^2 - x_2 + 2)| \leq 1 \quad \forall x \in D$$

gilt: $g(D) \subset D$.

(3) g kontrahierend, da D konvex, g diffbar mit

$$g'(x) = \frac{1}{4}\begin{pmatrix} -x_2 & 1 - x_1 \\ 2x_1 & -1 \end{pmatrix} \quad und \quad \max_{x \in D} \|g'(x)\|_\infty = \frac{3}{4} = q < 1.$$

Damit hat g genau einen Fixpunkt in D.
b) $x^{(1)} = (\frac{5}{16}, \frac{7}{16})^T$ und die a priori Abschätzung des BNFS ergibt, daß weitere 16 Iterationen benötigt werden um die gewünschte Genauigkeit zu erreichen.

Aufgabe 8 :
Betrachte

$$x^{(k+1)} = x^{(k)} - Ax^{(k)} + b = (E - A)x^{(k)} + b =: Cx^{(k)} + b$$

Das Verfahren konvergiert genau dann, wenn $\rho(C) < 1$. Da A positiv definit und damit diagonalisierbar ist, gibt es eine Darstellung mit

$$A = S \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} S^T,$$

wobei $\lambda_i > 0$ die Eigenwerte der Matrix A sind und S eine orthogonale $(n \times n)$-Matrix ist. Damit gilt:

$$C = E - A = SES^T - S \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} S^T$$

$$= S(E - \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix})S^T = S \begin{pmatrix} 1-\lambda_1 & & 0 \\ & \ddots & \\ 0 & & 1-\lambda_n \end{pmatrix} S^T$$

Auf der Diagonalen stehen die Eigenwerte von C. Der Spektralradius $\rho(C)$ ist der Wert des betragsmäßig größten Eigenwertes von C. Damit gilt:
Verfahren konvergiert $\Leftrightarrow \max\limits_{i=1,...,n} |1 - \lambda_i| < 1$
$\Leftrightarrow 0 < \lambda_i < 2 \quad \forall i = 1, ..., n \Leftrightarrow \lambda_{max}(A) < 2.$

Aufgabe 1:

Seien

$$A = \begin{pmatrix} 2 & 5 & 11 \\ 4 & 2 & 2 \\ 2 & 17 & 5 \end{pmatrix} \quad , \quad b = \begin{pmatrix} -7 \\ 6 \\ -1 \end{pmatrix} .$$

(a) Bestimmen Sie mit Spaltenpivotsuche die Zerlegung $PA = LDR$, wobei P: Permutationsmatrix; L, R: linke bzw. rechte normierte Dreiecksmatrix und D: Diagonalmatrix bedeuten.

(b) Berechnen Sie die Lösung des LGS $Ax = b$.

(c) *Überlegen Sie*, daß PA positiv definit ist und bestimmen Sie die Zerlegung $PA = \tilde{L}\tilde{L}^T$.

(8 Punkte)

Aufgabe 2:

Seien

$$A = \begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix} \quad \text{mit} \quad A^{-1} = \begin{pmatrix} -1 & 2 \\ 2 & -3 \end{pmatrix} , \quad b = \begin{pmatrix} 5 \\ 3 \end{pmatrix} , \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

(a) Berechnen Sie die Kondition $\mathrm{cond}(A)$ für die Normen $\|\cdot\|_\infty$ und $\|\cdot\|_2$.

(b) Für Näherungen $\tilde{A}$, $\tilde{b}$, $\tilde{x}$ von A, b, x gelte

$$\tilde{A}\tilde{x} = \tilde{b} \quad \text{mit} \quad \|\tilde{A} - A\|_\infty \leq 0.02 .$$

Wie groß darf der Fehler $\|\tilde{b} - b\|_\infty$ sein, damit $\|\tilde{x} - x\|_\infty \leq 0.5$ gilt?

(7 Punkte)

Aufgabe 3:

Durch die Meßpunkte

t_i	0	1	4
y_i	2	3	4

soll eine Ausgleichsfunktion $u(t) = \alpha + \beta\sqrt{t}$, $\alpha, \beta \in \mathbf{R}$, gelegt werden. Formulieren Sie das zugehörige lineare Ausgleichsproblem und berechnen Sie die optimalen Parameter α, β.

(5 Punkte)

Aufgabe 4:

Sei $g : \mathbf{R}^2 \to \mathbf{R}^2$ die Abbildung mit

$$g(x_1, x_2)^T = \frac{1}{4} \begin{pmatrix} x_2 & - & x_1 x_2 & + & 1 \\ x_1^2 & - & x_2 & + & 2 \end{pmatrix} .$$

(a) Zeigen Sie, daß die Abbildung g auf der Menge $D = [-1, 1] \times [-1, 1] \subset \mathbf{R}^2$ genau einen Fixpunkt $\overline{x} \in D$ besitzt.

(b) Berechnen Sie $x^1 = g(x^0)$ mit $x^0 = (0.5, 0.5)^T$. Wieviele Iterationen k benötigt man, um die Genauigkeit $\|\overline{x} - x^k\|_\infty \leq 0.01$ für den Fixpunkt $\overline{x} \in D$ zu erzielen?

(6 Punkte)

Aufgabe 5:

Sei A eine symmetrische positiv definite (n, n)-Matrix und sei $b \in \mathbf{R}^n$. Zur iterativen Lösung des *LGS* $Ax = b$ betrachte man das RICHARDSON-Verfahren

$$x^{(k+1)} = x^{(k)} - Ax^{(k)} + b .$$

Zeigen Sie: Dieses Verfahren ist genau dann konvergent, wenn $\lambda_{max}(A) < 2$.

(4 Punkte)

Aufgabe 6:

(a) Das Polynom $p \in \Pi_3$ interpoliere die Funktion $f(x) = \frac{x^3}{x^2-1}$ in den Knoten $x_0 = -1$, $x_1 = 0$, $x_2 = 2$ und $x_3 = 3$.

Berechnen Sie die Koeffizienten des Interpolationspolynoms p in der Newton'schen Darstellung.

(b) Das Polynom $p_n \in \Pi_n$ interpoliere die Funktion $f(x) = e^x$ im Intervall $[0, 2]$ in den Knoten $x_i = i \cdot \frac{2}{n}$, $i = 0, 1, \ldots, n$.

Zeigen Sie: $\lim_{n \to \infty} \|p_n - f\|_\infty = 0$.

(6 Punkte)

Aufgabe 7:

Nachfolgend ist ein aus zwei Teilen bestehendes Fortran-Programm aufgelistet. Beantworten Sie für jeweils beide Programmteile die folgenden Fragen:

(a) Welches numerische Verfahren wurde programmiert?

(b) Welches ursprüngliche mathematische Problem wird gelöst?

Hinweis: In der Programmiersprache Fortran entspricht ein Ausdruck *A.gt.B* der logischen Bedeutung $A > B$.

(5 Punkte)

```
      PROGRAM Aufgabe_7
      dimension x(3),y(3)

C-----Teil A-------------------------------------------

      xn = 3.0E0
 100  xa = xn

      xn = xa-(xa*xa-2.0E0*xa)/(3.0E0*xa-4.0E0)

      if (abs(xa-xn).gt.1.0E-8) goto 100
      write(*,*) 'Lcesung :', xn

C-----Teil B-------------------------------------------

      x(1) = 0.0E0
      x(2) = 0.0E0
      x(3) = 0.0E0
 200  y(1) = x(1)
      y(2) = x(2)
      y(3) = x(3)

      x(1) = (1.0E0-y(2))/2.0E0
      x(2) = (-2.0E0+y(1)-y(3))/3.0E0
      x(3) = (5.0E0-y(1))/2.0E0

      vn = sqrt(x(1)*x(1)+x(2)*x(2)+x(3)*x(3))
      va = sqrt(y(1)*y(1)+y(2)*y(2)+y(3)*y(3))

      if (abs(va-vn).gt.1.0E-8) goto 200
      write(*,*) 'Loesung :',x(1), x(2), x(3)

      stop
      end
```

AUFGABE 1: Der Ausdruck $y = \ln\left(x - \sqrt{x^2-1}\right)$ $(= \operatorname{arcosh}(x))$ soll für $x = 30$ berechnet werden. Die Wurzel werde hierbei einer 5-stelligen Tafel entnommen:

$$\sqrt{899} \approx 29.983$$

(a) Wie groß wird der absolute Fehler? (3 Punkte)

(b) Man stabilisiere den Ausdruck und bestimme den absoluten Fehler bei Auswertung des umgeformten Ausdrucks.

(2 Punkte)

Hinweis: Man führe für die Funktion $y = \varphi(z)$ eine Fehleranalyse durch, wobei z die gerundete Variable ist.

AUFGABE 2: Sei

$$A = \begin{pmatrix} 6 & 18 & 30 \\ 18 & 57 & 96 \\ 30 & 96 & 163 \end{pmatrix}$$

(a) Bestimmen Sie die Zerlegungen $A = LR$ und $A = LDR$. (2 Punkte)

(b) Zeigen Sie, daß A positiv definit ist. Bestimmen Sie die Cholesky-Zerlegung $A = LL^T$ (nachdenken!) und schreiben Sie x^TAx, $x \in \mathbb{R}^3$, als Summe von Quadraten. (3 Punkte)

AUFGABE 3: Die Matrix $A = (a_{ik})$ sei diagonaldominant:

$$\sum_{k \neq i} |a_{ik}| < |a_{ii}| \quad , \quad i = 1,\ldots,n \quad .$$

(a) Man zeige: A ist regulär und A besitzt eine LR-Zerlegung $A = L \cdot R$. (3 Punkte)

(b) Man gebe eine möglichst gute Abschätzung für $\|A^{-1}\|_\infty$ an. (2 Punkte)

Hinweis: $A = D(I + F)$, $D = \operatorname{diag}(a_{ii})$.

AUFGABE 4: Die Funktion $f(x) = x^3-x^2-x-1$ hat die einzige positive Nullstelle $\bar{x} = 1.839 \ldots\ldots$.

Man konstruiere eine Iterationsfunktion g (ohne Benutzung von $\bar{x}$), so daß die Iteration $x_{k+1} = g(x_k)$ für jeden Startwert $x_0 \geq 0$ gegen $\bar{x}$ konvergiert.

Hinweis: Benutze $(x + \frac{1}{2})^2 < x^2 + x + 1$. (5 Punkte)

AUFGABE 5: Zur Lösung des LGS

$$\begin{pmatrix} 1 & 0.2 \\ 0.1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

führe man ausgehend von $x^{(0)} = (2,1)^T$ zwei Schritte des Einzelschrittverfahrens durch und schätze den Fehler ab, ohne die explizite Lösung zu benutzen. (4 Punkte)

AUFGABE 6:

(a) Berechnen Sie das Newton'sche Interpolationspolynom $P(x)$ dritten Grades, das die Funktion $f(x) = \sqrt{x}$ in den Knoten 0, 1, 4, 9 interpoliert. Berechnen Sie $P(2)$ mit einem Horner-ähnlichen Schema. (3 Punkte)

(b) Das Polynom $P(x)$ zweiten Grades interpoliere $f(x) = \sqrt{x}$ in den Knoten 1, 2, 3. Geben Sie eine Abschätzung für den Interpolationsfehler im Intervall $[1,3]$ an. (2 Punkte)

Aufgabe 1:

Seien

$$A = \begin{pmatrix} 2 & 26 & 11 \\ 4 & 2 & 2 \\ 2 & 11 & 14 \end{pmatrix} \quad , \quad b = \begin{pmatrix} -13 \\ 4 \\ 5 \end{pmatrix} .$$

(a) Bestimmen Sie mit Spaltenpivotsuche die Zerlegung $PA = LDR$, wobei P: Permutationsmatrix; L, R: linke bzw. rechte normierte Dreiecksmatrix und D: Diagonalmatrix bedeuten.

(b) Berechnen Sie die Lösung des $LGS \quad Ax = b$.

(c) *Überlegen Sie*, daß PA positiv definit ist und bestimmen Sie die Zerlegung $PA = \tilde{L}\tilde{L}^T$.

(7 Punkte)

Aufgabe 2:

Seien

$$A = \begin{pmatrix} -2 & 3 \\ 4 & -5 \end{pmatrix} \quad , \quad b = \begin{pmatrix} -1 \\ 3 \end{pmatrix} .$$

Berechnen Sie die Lösung $x \in \mathbb{R}^2$ des linearen Gleichungssystems

$$Ax = b \, ,$$

mit dem QR-Verfahren

(5 Punkte)

Aufgabe 3:

Seien

$$A = \begin{pmatrix} 8 & 5 \\ 5 & 3 \end{pmatrix} \quad \text{mit} \quad A^{-1} = \begin{pmatrix} -3 & 5 \\ 5 & -8 \end{pmatrix} \quad , \quad b = \begin{pmatrix} 13 \\ 8 \end{pmatrix} \quad , \quad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} .$$

(a) Berechnen Sie die Kondition $\text{cond}(A)$ für die Normen $\| \cdot \|_\infty$ und $\| \cdot \|_2$.

(b) Für Näherungen $\tilde{A}$, $\tilde{b}$, $\tilde{x}$ von A, b, x gelte

$$\tilde{A}\tilde{x} = \tilde{b} \quad \text{mit} \quad \|\tilde{A} - A\|_\infty \leq 0.013 \, .$$

Wie groß darf der Fehler $\|\tilde{b} - b\|_\infty$ sein, damit $\|\tilde{x} - x\|_\infty \leq 0.5$ gilt?

(6 Punkte)

Aufgabe 4:
Durch die Meßpunkte

t_i	$\frac{1}{e}$	1	e
y_i	-1	e	$2+e^2$

soll eine Ausgleichsfunktion $u(t) = \alpha t + \beta ln(t)$, $\alpha, \beta \in \mathbb{R}$, gelegt werden. Formulieren Sie das zugehörige lineare Ausgleichsproblem und berechnen Sie die optimalen Parameter α, β.

(5 Punkte)

Aufgabe 5:
Sei $g : \mathbb{R}^2 \to \mathbb{R}^2$ die Abbildung mit

$$g(x_1, x_2) = \frac{1}{4} \begin{pmatrix} x_1 x_2 & + & \sin(x_1) & + & 2 \\ x_2 \cos(x_1) & + & \frac{1}{2} x_2^2 & + & 2 \end{pmatrix} .$$

(a) Zeigen Sie, daß für ein geeignet zu wählendes $c \in \mathbb{R}^+$ die Abbildung g auf der Menge $D = [-c, c] \times [-c, c] \subset \mathbb{R}^2$ genau einen Fixpunkt besitzt.

(b) Berechnen Sie $x^1 = g(x^0)$ mit $x^0 = (0, 1)^T$. Wieviele Iterationen k benötigt man, um die Genauigkeit $\|\overline{x} - x^k\|_\infty \leq 10^{-3}$ für den Fixpunkt $\overline{x} \in D$ erzielen?

(6 Punkte)

Aufgabe 6:

(a) Das Polynom $p \in \prod_3$ interpoliere die Funktion

$$f(x) = \frac{6}{\sin(\frac{\pi}{2}x) + 2}$$

in den Knoten $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ und $x_3 = 3$.

Berechnen Sie die Koeffizienten des Interpolationspolynoms p in der Newton'schen Darstellung.

(b) Das Polynom $p \in \prod_2$ interpoliere die Funktion

$$f(x) = \sin(\frac{\pi}{4}x) + \cos(\frac{\pi}{4}x) \ ,$$

in den Knoten $x_0 = -2$, $x_1 = 0$, $x_2 = 2$.

Geben Sie eine möglichst gute Abschätzung für den Interpolationsfehler $|f(x) - p(x)|$ im Intervall $[1, 2]$.

(6 Punkte)

Aufgabe 7:

Das Polynom $p_n(x)$ sei rekursiv durch

$$\begin{aligned} p_0(x) &= 1 , \\ p_1(x) &= \alpha_1 - x , \\ p_i(x) &= (\alpha_i - x)p_{i-1}(x) - \beta_i^2 p_{i-2}(x) , \quad i = 2, 3, \ldots \end{aligned}$$

mit α_i, $\beta_i \in \mathbb{R}$ definiert.

Schreiben Sie ein Programm zur Berechnung einer Nullstelle von $p_n(x)$ mit dem Newton-Verfahren.
Wie sieht die Rekursionsformel für $p_i'(x)$ aus?

(5 Punkte)

b) $A = \begin{pmatrix} 2 & 26 & 11 \\ 4 & 2 & 2 \\ 2 & 11 & 14 \end{pmatrix}$; $S = \begin{pmatrix} -13 \\ 4 \\ 5 \end{pmatrix}$

c)

$$\begin{pmatrix} 2 & 26 & 11 \\ (4) & 2 & 2 \\ 2 & 11 & 14 \end{pmatrix}$$

Vertauschen der ersten beiden Zeilen

$$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 26 & 11 \\ 2 & 11 & 14 \end{pmatrix} \begin{matrix} \checkmark \\ -\frac{1}{2} \cdot I \\ -\frac{1}{2} \cdot I \end{matrix} \quad (1)$$

$$\begin{pmatrix} 4 & 2 & 2 \\ \frac{1}{2} & (25) & 10 \\ \frac{1}{2} & 10 & 13 \end{pmatrix} \begin{matrix} \checkmark \\ \\ -\frac{2}{5} \cdot II \end{matrix}$$

$$\begin{pmatrix} 4 & 2 & 2 \\ \frac{1}{2} & 25 & 10 \\ \frac{1}{2} & \frac{2}{5} & 9 \end{pmatrix} \begin{matrix} \checkmark \\ \\ (2) \end{matrix}$$

$$\Rightarrow L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{2}{5} & 1 \end{pmatrix} \quad D = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

Lösung $\begin{pmatrix} 4 & 2 & 2 \\ 0 & 2\sqrt{5} & 10 \\ 2 & 0 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ -15 \\ 9 \end{pmatrix}$

$\Rightarrow x_3 = 1$

$2\sqrt{5}x_2 = -15 - 10x_3 = -25 \Rightarrow x_2 = -1$

$4x_1 = 4 - 2x_2 - 2x_3 = 4 + 2 - 2 = 4 \Rightarrow x_1 = 1$

$\Rightarrow x = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ ✓ (1)

c) Nach a) gilt: $PA = L\bar{D}\bar{D}R$ mit $\bar{D} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & 3 \end{pmatrix}$

und $R^T = L$

✓ (1)

$\Rightarrow \underline{PA} = (L\bar{D})(L^T\bar{D}^T)^T = \underline{(L\bar{D})(L\bar{D})^T}$

$\overset{L.A.}{\Rightarrow}$ PA ist positiv definit

sei $\tilde{L} := L\bar{D} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{2}{5} & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{5} & 0 \\ 1 & 2 & 3 \end{pmatrix}$

$\Rightarrow PA = \tilde{L}\tilde{L}^T$ ≠ PKt

✓ (1)

$$A = \begin{pmatrix} -2 & 3 \\ 4 & -5 \end{pmatrix} \qquad S = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$$

QR-Zerlegung (nach Householder):

1. Schritt:

$$x = \begin{pmatrix} -2 \\ 4 \end{pmatrix} \Rightarrow \|x\| = \sqrt{20} \checkmark$$

$$u = \begin{pmatrix} -2-\sqrt{20} \\ 4 \end{pmatrix} \Rightarrow u u^T = \begin{pmatrix} 24+4\sqrt{20} & -8-4\sqrt{20} \\ -8-4\sqrt{20} & 16 \end{pmatrix}$$

$$\beta = \frac{1}{2\sqrt{20}+20} \checkmark$$

~~$\Rightarrow$ OB~~

$\sqrt{20} = 2\sqrt{5}$

$$\Rightarrow Q_1 = E - \beta u u^T = \begin{pmatrix} 1 - \frac{24+4\sqrt{20}}{2\sqrt{20}+20} & \frac{+8+4\sqrt{20}}{2\sqrt{20}+20} \\ \frac{+8+4\sqrt{20}}{2\sqrt{20}+20} & 1 - \frac{11}{2\sqrt{20}+20} \end{pmatrix}$$

$$= + \frac{4+2\sqrt{20}}{2\sqrt{20}+20} \begin{pmatrix} -1 & 2 \\ 2 & +1 \end{pmatrix} = + \frac{\sqrt{20}}{10} \begin{pmatrix} -1 & 2 \\ 2 & +1 \end{pmatrix} \checkmark$$

$\frac{\sqrt{20}}{10} = \frac{1}{\sqrt{5}}$

$\Rightarrow$ ~~$A_2 = Q_1 A = + \frac{\sqrt{20}}{10} \begin{pmatrix} 10 & -13 \\ 0 & 1 \end{pmatrix}$~~

$$Q = \frac{\sqrt{20}}{10} \begin{pmatrix} -1 & 2 \\ 2 & 1 \end{pmatrix}$$

$$R = QA = \frac{\sqrt{20}}{10} \begin{pmatrix} 10 & -13 \\ 0 & 1 \end{pmatrix} \checkmark$$

3) Sei $A = \begin{pmatrix} 8 & 5 \\ 5 & 3 \end{pmatrix}$; $A^{-1} = \begin{pmatrix} -3 & 5 \\ 5 & -8 \end{pmatrix}$; $b = \begin{pmatrix} 13 \\ 8 \end{pmatrix}$; $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

a) Es gilt: $\|A\|_\infty = 13$

$\|A^{-1}\|_\infty = 13$

$\Rightarrow \text{cond}_\infty(A) = 169$

✓ (1)

Berechnung der Eigenwerte von A: ≥ 0 $\Rightarrow \rho(A) = \|A\|_2$!!!

$$|A - \lambda E| = \begin{vmatrix} 8-\lambda & 5 \\ 5 & 3-\lambda \end{vmatrix} = (8-\lambda)(3-\lambda) - 25 = \lambda^2 - 11\lambda - 1$$

$$\lambda^2 - 11\lambda - 1 = 0$$

$$\Leftrightarrow \left(\lambda - \frac{11}{2}\right)^2 = 1 + \left(\frac{11}{2}\right)^2 = \frac{125}{4}$$

$$\Rightarrow \lambda_{1,2} = \frac{11}{2} \pm \frac{5}{2}\sqrt{5} \qquad \Rightarrow \rho(A) = \frac{11}{2} + \frac{5}{2}\sqrt{5}$$

$$\underset{A \text{ Sym.}}{\Rightarrow} \|A\|_2 = \frac{11}{2} + \frac{5}{2}\sqrt{5}$$

Eigenwerte von A^{-1}: ≥ 0 $\Rightarrow \rho(A^{-1}) = \|A^{-1}\|_2$!!!

(O.B.)

$$|A^{-1} - \lambda E| = \begin{vmatrix} -3-\lambda & 5 \\ 5 & -8-\lambda \end{vmatrix} = \lambda^2 + 11\lambda - 1$$

$$\lambda^2 + 11\lambda - 1 = 0$$

$$\Leftrightarrow \left(\lambda + \frac{11}{2}\right)^2 = 1 + \left(\frac{11}{2}\right)^2 = \frac{125}{4}$$

$$\Rightarrow \lambda_{1,2} = -\frac{11}{2} \pm \frac{5}{2}\sqrt{5} \qquad \Rightarrow \rho(A^{-1}) = \frac{11}{2} + \frac{5}{2}\sqrt{5}$$

$$\underset{A^{-1} \text{ Sym.}}{\Rightarrow} \|A^{-1}\| = \frac{11}{2} + \frac{5}{2}\sqrt{5}$$

$$\Rightarrow \text{cond}_2(A) = \left(\frac{11}{2} + \frac{5}{2}\sqrt{5}\right)^2 = 60,5 + 27,5\sqrt{5}$$

$$\approx 122,99187$$

f. -2

4)

Messpunkte

t_i	$\frac{1}{e}$	1	e
y_i	-1	e	$2+e^2$

Bestimmung der Ausgleichsfunktion $a(t) = \alpha t + \beta \ln(t)$:

$$y = \begin{pmatrix} -1 \\ e \\ 2+e^2 \end{pmatrix} \quad A = \begin{pmatrix} t_1 & \ln t_1 \\ t_2 & \ln t_2 \\ t_3 & \ln t_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{e} & -1 \\ 1 & 0 \\ e & 1 \end{pmatrix}$$

lineares Ausgleichsproblem
Minimiere $\|y - Ax\|_2$
$x \in \mathbb{R}^2$

$$\Rightarrow A^T = \begin{pmatrix} \frac{1}{e} & 1 & e \\ -1 & 0 & 1 \end{pmatrix} \Rightarrow A^T A = \begin{pmatrix} \frac{1}{e^2}+1+e^2 & e-\frac{1}{e} \\ e-\frac{1}{e} & 2 \end{pmatrix}$$

α, β ergeben sich durch Lösung von $A^T A \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = A^T y$:

$$\left(\begin{array}{cc|c} \frac{1}{e^2}+1+e^2 & e-\frac{1}{e} & -\frac{1}{e}+e+2e+e^3 \\ e-\frac{1}{e} & 2 & 3+e^2 \end{array} \right) - \frac{e-\frac{1}{e}}{\frac{1}{e^2}+1+e^2} \cdot \text{I}$$

$$\left(\begin{array}{cc|c} \frac{1}{e^2}+1+e^2 & e-\frac{1}{e} & -\frac{1}{e}+3e+e^3 \\ 0 & 2-\frac{\left(e-\frac{1}{e}\right)^2}{\frac{1}{e^2}+1+e^2} & 3+e^2-\frac{\left(e-\frac{1}{e}\right)\left(-\frac{1}{e}+3e+e^3\right)}{\frac{1}{e^2}+1+e^2} \end{array} \right)$$

$$\Rightarrow \left(2-\frac{\left(e-\frac{1}{e}\right)^2}{\frac{1}{e^2}+1+e^2}\right)\beta = 3+e^2-\frac{\left(e-\frac{1}{e}\right)\left(-\frac{1}{e}+3e+e^3\right)}{\frac{1}{e^2}+1+e^2}$$

5) $f: \mathbb{R}^2 \to \mathbb{R}^2$

$$f(x_1,x_2) = \frac{1}{4}\begin{pmatrix} x_1 x_2 + \sin(x_1) + 2 \\ x_2 \cos(x_1) + \frac{1}{2}x_2^2 + 2 \end{pmatrix}$$

Sei $c = 1$ ✓, $D = [-1,1] \times [-1,1]$ ✓

a) i) D ist abgeschlossen ✓

ii) Für $x \in D$ ges.: (also $|x_1| \le 1, |x_2| \le 1$)

~~$\frac{1}{4}(x_1 x_2 + \sin x_1 + 2) \in$~~

$$\le \frac{1}{4}(|x_1 x_2| + |\sin x_1| + 2) \le \frac{1}{4}(1 + 1 + 2) = 1$$

und $\frac{1}{4}\left(x_2 \cos x_1 + \frac{1}{2}x_2^2 + 2\right) \le \frac{1}{4}\left(|x_2 \cos x_1| + \frac{1}{2}1^2 + 2\right)$

$$\le \frac{1}{4}\left(1 + \frac{1}{2} + 2\right) < 1$$

$\Rightarrow f(D) \subseteq D$ ✓

iii) es gilt:

1) D ist offen und konvex ✓

2) f ist diffbar mit $f'(x) = \frac{1}{4}\begin{pmatrix} x_2 + \cos x_1 & x_1 \\ -x_2 \sin x_1 & \cos x_1 + x_2 \end{pmatrix}$ ✓

$$\Rightarrow \|f'(x)\|_\infty = \frac{1}{4}\max\{|x_2 + \cos x_1| + |x_1|,\ |x_2 \sin x_1| + |\cos x_1| + |x_2|\}$$

für $x \in D$

$$\le \frac{3}{4} =: q$$

$\Rightarrow \sup_{x \in D} \|f'(x)\|_\infty \le q < 1$

6/a) $f(x) := \frac{6}{\sin\left(\frac{\pi}{2}x\right)+2}$

x_i	-1	0	1	3
$f(x_i)$	6	3	2	6

Sei $p \in \Pi_3$, $p(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + a_3(x-x_0)(x-x_1)$

Differenzschema:

$x_0 = -1$	6 ✓ $= a_0$			
		$-3 = a_1$ ✓		
$x_1 = 0$	3 ✓		$1 = a_2$ ✓	
		-1 ✓		$0 = a_3$ ✓
$x_2 = 1$	2 ✓		1 ✓	
		2 ✓		
$x_3 = 3$	6 ✓			

$\Rightarrow a_0 = 6,\ a_1 = -3,\ a_2 = 1,\ a_3 = 0$

Explizites Polynom?: $x^2 - 2x + 3$

2/2,5

b) $p \in \Pi_2$ interpoliere die Funktion

$$f(x) = \sin\left(\frac{\pi}{4}x\right) + \cos\left(\frac{\pi}{4}x\right)$$

Knoten

x_i	-2	0	2
$f(x_i)$	-1	1	1

Sei $I := [1,2]$

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 1 , Abgabe: Donnerstag, 29.04.2004, 8.00 Uhr

Übungstermine:

Gruppe 1:	Di.	9-11 Uhr	SR1	BK27
Gruppe 2:	Di.	11-13 Uhr	SR1	BK28
Gruppe 3:	Di.	15-17 Uhr	SR1	BK29

Aufgabe 1: (4 Punkte)
Bestimmen Sie für die Matrizen

(a) $A = uv^T, \quad u, v \in \mathbb{R}^n,$

(b) $Q = I - 2ww^T, \quad w^T w = 1, \quad w \in \mathbb{R}^n,$

(c)

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

die Eigenwerte λ_i und die Vielfachheiten $\sigma(\lambda_i)$ und $\rho(\lambda_i)$ des charakteristischen Polynoms $\varphi(\lambda)$.

Aufgabe 2: (4 Punkte)
Sei A eine reelle symmetrische $(n \times n)$–Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und den Eigenvektoren $x_1, \ldots, x_n \in \mathbb{R}^n$, $x_i^T x_k = \delta_{ik}$.
Man zeige: $\lambda_j = \max\limits_{y \in \mathbb{R}^n} \{y^T A y \mid y^T y = 1 \ , \ x_i^T y = 0 \quad \text{für} \quad i = 1, \ldots, j-1\}$.

Aufgabe 3: (4 Punkte)
Bei der Berechnung der Grundfrequenzen und Schwingungsformen eines linearen Schwingungssystems der Form (Bild 1), stellt sich die Aufgabe der Berechnung der Eigenwerte und Eigenvektoren einer Matrix

$$A = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix} .$$

(a) Berechnen Sie für $x^{(0)} = (1, 1, 1)^T$ mit Hilfe der Potenzmethode 4 Iterationen zur Bestimmung des größten Eigenwertes von A.

(b) Bestimmen Sie für $c_1 = 8$, $c_2 = 3$, $c_3 = 11$ den größten Eigenwert und zugehörigen Eigenvektor der Matrix A. Benutzen Sie (a) zur Berechnung einer Näherung und

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 1 , Abgabe: Donnerstag, 29.04.2004, 8.00 Uhr

Übungstermine:

Gruppe 1:	Di.	9-11 Uhr	SR1	BK27
Gruppe 2:	Di.	11-13 Uhr	SR1	BK28
Gruppe 3:	Di.	15-17 Uhr	SR1	BK29

Aufgabe 1: (4 Punkte)
Bestimmen Sie für die Matrizen

(a) $A = uv^T, \quad u, v \in \mathbb{R}^n,$

(b) $Q = I - 2ww^T, \quad w^T w = 1, \quad w \in \mathbb{R}^n,$

(c)

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

die Eigenwerte λ_i und die Vielfachheiten $\sigma(\lambda_i)$ und $\rho(\lambda_i)$ des charakteristischen Polynoms $\varphi(\lambda)$.

Aufgabe 2: (4 Punkte)
Sei A eine reelle symmetrische $(n \times n)$–Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und den Eigenvektoren $x_1, \ldots, x_n \in \mathbb{R}^n$, $x_i^T x_k = \delta_{ik}$.
Man zeige: $\lambda_j = \max\limits_{y \in \mathbb{R}^n} \{y^T A y \mid y^T y = 1 \,,\; x_i^T y = 0 \quad \text{für} \quad i = 1, \ldots, j-1\}$.

Aufgabe 3: (4 Punkte)
Bei der Berechnung der Grundfrequenzen und Schwingungsformen eines linearen Schwingungssystems der Form (Bild 1), stellt sich die Aufgabe der Berechnung der Eigenwerte und Eigenvektoren einer Matrix

$$A = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix} .$$

(a) Berechnen Sie für $x^{(0)} = (1, 1, 1)^T$ mit Hilfe der Potenzmethode 4 Iterationen zur Bestimmung des größten Eigenwertes von A.

(b) Bestimmen Sie für $c_1 = 8$, $c_2 = 3$, $c_3 = 11$ den größten Eigenwert und zugehörigen Eigenvektor der Matrix A. Benutzen Sie (a) zur Berechnung einer Näherung und

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 1 , Abgabe: Donnerstag, 29.04.2004, 8.00 Uhr

Übungstermine:

Gruppe 1:	Di.	9-11 Uhr	SR1	BK27
Gruppe 2:	Di.	11-13 Uhr	SR1	BK28
Gruppe 3:	Di.	15-17 Uhr	SR1	BK29

Aufgabe 1: (4 Punkte)
Bestimmen Sie für die Matrizen

(a) $A = uv^T, \quad u, v \in \mathbb{R}^n,$

(b) $Q = I - 2ww^T, \quad w^T w = 1, \quad w \in \mathbb{R}^n,$

(c)

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

die Eigenwerte λ_i und die Vielfachheiten $\sigma(\lambda_i)$ und $\rho(\lambda_i)$ des charakteristischen Polynoms $\varphi(\lambda)$.

Aufgabe 2: (4 Punkte)
Sei A eine reelle symmetrische $(n \times n)$–Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und den Eigenvektoren $x_1, \ldots, x_n \in \mathbb{R}^n$, $x_i^T x_k = \delta_{ik}$.
Man zeige: $\lambda_j = \max\limits_{y \in \mathbb{R}^n} \{ y^T A y \mid y^T y = 1 \, , \; x_i^T y = 0 \quad \text{für} \quad i = 1, \ldots, j-1 \}$.

Aufgabe 3: (4 Punkte)
Bei der Berechnung der Grundfrequenzen und Schwingungsformen eines linearen Schwingungssystems der Form (Bild 1), stellt sich die Aufgabe der Berechnung der Eigenwerte und Eigenvektoren einer Matrix

$$A = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix} .$$

(a) Berechnen Sie für $x^{(0)} = (1, 1, 1)^T$ mit Hilfe der Potenzmethode 4 Iterationen zur Bestimmung des größten Eigenwertes von A.

(b) Bestimmen Sie für $c_1 = 8$, $c_2 = 3$, $c_3 = 11$ den größten Eigenwert und zugehörigen Eigenvektor der Matrix A. Benutzen Sie (a) zur Berechnung einer Näherung und

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 1 , Abgabe: Donnerstag, 29.04.2004, 8.00 Uhr

Übungstermine:

Gruppe 1:	Di.	9-11 Uhr	SR1	BK27
Gruppe 2:	Di.	11-13 Uhr	SR1	BK28
Gruppe 3:	Di.	15-17 Uhr	SR1	BK29

Aufgabe 1: (4 Punkte)
Bestimmen Sie für die Matrizen

(a) $A = uv^T, \quad u, v \in \mathbb{R}^n,$

(b) $Q = I - 2ww^T, \quad w^T w = 1, \quad w \in \mathbb{R}^n,$

(c)

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

die Eigenwerte λ_i und die Vielfachheiten $\sigma(\lambda_i)$ und $\rho(\lambda_i)$ des charakteristischen Polynoms $\varphi(\lambda)$.

Aufgabe 2: (4 Punkte)
Sei A eine reelle symmetrische $(n \times n)$–Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und den Eigenvektoren $x_1, \ldots, x_n \in \mathbb{R}^n$, $x_i^T x_k = \delta_{ik}$.
Man zeige: $\lambda_j = \max\limits_{y \in \mathbb{R}^n} \{ y^T A y \mid y^T y = 1 \, , \; x_i^T y = 0 \quad \text{für} \quad i = 1, \ldots, j-1 \}$.

Aufgabe 3: (4 Punkte)
Bei der Berechnung der Grundfrequenzen und Schwingungsformen eines linearen Schwingungssystems der Form (Bild 1), stellt sich die Aufgabe der Berechnung der Eigenwerte und Eigenvektoren einer Matrix

$$A = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix} .$$

(a) Berechnen Sie für $x^{(0)} = (1, 1, 1)^T$ mit Hilfe der Potenzmethode 4 Iterationen zur Bestimmung des größten Eigenwertes von A.

(b) Bestimmen Sie für $c_1 = 8$, $c_2 = 3$, $c_3 = 11$ den größten Eigenwert und zugehörigen Eigenvektor der Matrix A. Benutzen Sie (a) zur Berechnung einer Näherung und

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 1 , Abgabe: Donnerstag, 29.04.2004, 8.00 Uhr

Übungstermine:

Gruppe 1:	Di.	9-11 Uhr	SR1	BK27
Gruppe 2:	Di.	11-13 Uhr	SR1	BK28
Gruppe 3:	Di.	15-17 Uhr	SR1	BK29

Aufgabe 1: (4 Punkte)
Bestimmen Sie für die Matrizen

(a) $A = uv^T, \quad u, v \in \mathbb{R}^n,$

(b) $Q = I - 2ww^T, \quad w^T w = 1, \quad w \in \mathbb{R}^n,$

(c)

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

die Eigenwerte λ_i und die Vielfachheiten $\sigma(\lambda_i)$ und $\rho(\lambda_i)$ des charakteristischen Polynoms $\varphi(\lambda)$.

Aufgabe 2: (4 Punkte)
Sei A eine reelle symmetrische $(n \times n)$–Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und den Eigenvektoren $x_1, \ldots, x_n \in \mathbb{R}^n$, $x_i^T x_k = \delta_{ik}$.
Man zeige: $\lambda_j = \max\limits_{y \in \mathbb{R}^n} \{y^T A y \mid y^T y = 1 \, , \; x_i^T y = 0 \quad \text{für} \quad i = 1, \ldots, j-1\}$.

Aufgabe 3: (4 Punkte)
Bei der Berechnung der Grundfrequenzen und Schwingungsformen eines linearen Schwingungssystems der Form (Bild 1), stellt sich die Aufgabe der Berechnung der Eigenwerte und Eigenvektoren einer Matrix

$$A = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix} .$$

(a) Berechnen Sie für $x^{(0)} = (1, 1, 1)^T$ mit Hilfe der Potenzmethode 4 Iterationen zur Bestimmung des größten Eigenwertes von A.

(b) Bestimmen Sie für $c_1 = 8$, $c_2 = 3$, $c_3 = 11$ den größten Eigenwert und zugehörigen Eigenvektor der Matrix A. Benutzen Sie (a) zur Berechnung einer Näherung und

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 1 , Abgabe: Donnerstag, 29.04.2004, 8.00 Uhr

Übungstermine:

Gruppe 1:	Di.	9-11 Uhr	SR1	BK27
Gruppe 2:	Di.	11-13 Uhr	SR1	BK28
Gruppe 3:	Di.	15-17 Uhr	SR1	BK29

Aufgabe 1: (4 Punkte)
Bestimmen Sie für die Matrizen

(a) $A = uv^T, \quad u, v \in \mathbb{R}^n,$

(b) $Q = I - 2ww^T, \quad w^T w = 1, \quad w \in \mathbb{R}^n,$

(c)

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

die Eigenwerte λ_i und die Vielfachheiten $\sigma(\lambda_i)$ und $\rho(\lambda_i)$ des charakteristischen Polynoms $\varphi(\lambda)$.

Aufgabe 2: (4 Punkte)
Sei A eine reelle symmetrische $(n \times n)$–Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ und den Eigenvektoren $x_1, \ldots, x_n \in \mathbb{R}^n$, $x_i^T x_k = \delta_{ik}$.
Man zeige: $\lambda_j = \max\limits_{y \in \mathbb{R}^n} \{y^T A y \mid y^T y = 1 \, , \; x_i^T y = 0 \quad \text{für} \quad i = 1, \ldots, j-1\}$.

Aufgabe 3: (4 Punkte)
Bei der Berechnung der Grundfrequenzen und Schwingungsformen eines linearen Schwingungssystems der Form (Bild 1), stellt sich die Aufgabe der Berechnung der Eigenwerte und Eigenvektoren einer Matrix

$$A = \begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix} .$$

(a) Berechnen Sie für $x^{(0)} = (1, 1, 1)^T$ mit Hilfe der Potenzmethode 4 Iterationen zur Bestimmung des größten Eigenwertes von A.

(b) Bestimmen Sie für $c_1 = 8$, $c_2 = 3$, $c_3 = 11$ den größten Eigenwert und zugehörigen Eigenvektor der Matrix A. Benutzen Sie (a) zur Berechnung einer Näherung und

$\Rightarrow Ax = uv^T x = 0$, d.h. $\lambda_1 = 0$ ist EW von A

Aus $\dim(v^\perp) = n-1$ folgt $g(\lambda_1) \geq n-1$ und $\sigma(\lambda_1) \geq g(\lambda_1) \geq n-1$

Ausserdem gilt: $Au = uv^T u = v^T u \, u$ ($v^T u \in \mathbb{R}$)

$\Rightarrow \lambda_2 = v^T u$ ist EW von A mit $g(\lambda_2) \geq 1$ und $\sigma(\lambda_2) \geq 1$.

Aus $\sum_i \sigma(\lambda_i) = n$ folgt $\sigma(\lambda_1) = n-1$, $\sigma(\lambda_2) = 1$

Wegen $\sigma(\lambda_i) \geq g(\lambda_i)$ ist dann $g(\lambda_1) = n-1$, $g(\lambda_2) = 1$

$\Rightarrow \varphi(\lambda) = \lambda^{n-1}(\lambda - v^T u)$

b) Sei $x \in w^\perp \Rightarrow w^T x = 0$

$\Rightarrow Qx = x - 2w\underbrace{w^T x}_{=0} = x$ $\Rightarrow \lambda_1 = 1$ ist EW von Q

mit $g(\lambda_1) \geq n-1$, $\sigma(\lambda_1) \geq n-1$ (s.o.)

Weiterhin ist $Qw = w - 2w\underbrace{w^T w}_{=1} = w - 2w = -w$

$\Rightarrow \lambda_2 = -1$ ist EW von Q mit $g(\lambda_2) \geq 1$, $\sigma(\lambda_2) \geq 1$

$\overset{s.o.}{\Rightarrow} g(\lambda_1) = \sigma(\lambda_1) = n-1$, $g(\lambda_2) = \sigma(\lambda_2) = 1 \Rightarrow \varphi(\lambda) = (\lambda-1)^{n-1}(\lambda+1)$

c) $\det(P - \lambda I) = \begin{vmatrix} -\lambda & 0 & 0 & 1 \\ 1 & -\lambda & 0 & 0 \\ 0 & 1 & -\lambda & 0 \\ 0 & 0 & 1 & -\lambda \end{vmatrix}$ Entwicklung nach der 1. Zeile

$= -\lambda \begin{vmatrix} -\lambda & 0 & 0 \\ 1 & -\lambda & 0 \\ 0 & 1 & -\lambda \end{vmatrix} - \begin{vmatrix} 1 & -\lambda & 0 \\ 0 & 1 & -\lambda \\ 0 & 0 & 1 \end{vmatrix} = \lambda^4 - 1$

$= (\lambda^2 - 1)(\lambda^2 + 1) = (\lambda - 1)(\lambda + 1)(\lambda - i)(\lambda + i)$

$\Rightarrow P$ hat die 4 EW $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = i$, $\lambda_4 = -i$

Aus $n = 4$ folgt: $g(\lambda_i) = \sigma(\lambda_i) = 1$ $\quad i = 1, \ldots, 4$

Sei $u \in v^\perp$. Sei $x \in \mathbb{R}^n$ mit $Ax = \lambda x$ und $x \neq 0$.

Dann gibt es eine eindeutige Zerlegung $x = y + \alpha v$, $y \in v^\perp$, $\alpha \in \mathbb{R}$

$$\Rightarrow \lambda x = Ax = uv^T(y + \alpha v) = u\underbrace{v^T y}_{=0} + \alpha u v^T v$$
$$= \alpha \|v\|_2^2 u \in v^\perp$$

$\Rightarrow \lambda = 0 \quad \vee \quad x \in v^\perp$

Aus $x = y + \alpha v \in v^\perp$ folgt $\alpha = 0$ und somit $\lambda x = 0$

$\Rightarrow \lambda = 0$ ist einziger EW von A.

Da $v^\perp$ im Eigenraum zu $\lambda = 0$ liegt, gilt: $g(0) \geq n-1$

Ist $A = 0$, so folgt $g(0) = \sigma(0) = n$, andernfalls

ist $g(0) = \sigma(0) = n-1$.

$\Rightarrow \varphi(\lambda) = \lambda^n$

(Dialog) In[43]:=

```
Table[λ[3, j] // Simplify, {j, 1, 3}]
```

(Dialog) Out[43]=

$$\left\{\frac{c_1^3+3c_1^2c_2+5c_1c_2^2+c_2^2(4c_2+c_3)}{c_1^2+2c_1c_2+2c_2^2},\ \frac{c_1^2+4c_2^2+3c_2c_3+2c_3^2+c_1(3c_2+c_3)}{c_1+2c_2+c_3},\ c_1+2(c_2+c_3)\right\}$$

(Dialog) In[46]:=

```
Table[EV[3, j] // Simplify // MatrixForm, {j, 1, 3}]
```

(Dialog) Out[46]=

$$\left\{\begin{pmatrix}1\\ -\frac{c_2(c_1+2c_2+c_3)}{c_1^2+2c_1c_2+2c_2^2}\\ \frac{c_2c_3}{c_1^2+2c_1c_2+2c_2^2}\end{pmatrix},\ \begin{pmatrix}-\frac{c_1^2+2c_1c_2+2c_2^2}{c_2(c_1+2c_2+c_3)}\\ 1\\ -\frac{c_3}{c_1+2c_2+c_3}\end{pmatrix},\ \begin{pmatrix}\frac{c_1^2+2c_1c_2+2c_2^2}{c_2c_3}\\ -\frac{c_1+2c_2+c_3}{c_3}\\ 1\end{pmatrix}\right\}$$

Teilaufgabe (b)

Die Eigenwerte und Eigenvektoren der Matrix A für die angegebenen Werte c_i lauten:

(Dialog) In[40]:=

```
B = A /. {c1 → 8, c2 → 3, c3 → 11}
```

(Dialog) Out[40]=

```
{{11, -3, 0}, {-3, 14, -11}, {0, -11, 11}}
```

(Dialog) In[41]:=

```
Eigenvalues[B]
```

(Dialog) Out[41]=

```
{1, 11, 24}
```

(Dialog) In[55]:=

```
EVB = Eigenvectors[B]
```

(Dialog) Out[55]=

```
{{3, 10, 11}, {-11, 0, 3}, {3, -13, 11}}
```

Der Eigenvektor zum größten Eigenwert $\lambda=24$ ist der Vektor (3,–13,11).Die Näherung aus (a) liefert für j=1,2,3

(Dialog) In[58]:=

```
Table[λ[3, j] /. {c1 → 8, c2 → 3, c3 → 11} // Simplify // N, {j, 1, 3}]
```

(Dialog) Out[58]=

```
{12.7308, 24.04, 36.}
```

(15,32 / -16,69 / 11)

(Dialog) In[59]:=

```
Table[
 EV[3, j] /. {c1 → 8, c2 → 3, c3 → 11} // Simplify // N // MatrixForm, {j, 1, 3}]
```

(Dialog) Out[59]=

$$\left\{\begin{pmatrix}1.\\ -0.576923\\ 0.253846\end{pmatrix},\ \begin{pmatrix}-1.73333\\ 1.\\ -0.44\end{pmatrix},\ \begin{pmatrix}3.93939\\ -2.27273\\ 1.\end{pmatrix}\right\}$$

(3.0408 / -13.012 / 11)

Es kommt nicht annähernd der berechnete Eigenvektor raus, obwohl mit 24.04 eine gute Näherung für den berechneten größten Eigenwert rauskommt.
Das liegt vermutlich daran, dass der Quotient vom zweitgrößten Eigenwert 11 durch den größten Eigenwert 24 kaum kleiner als 0.5 ist, deshalb konvergiert
$(\frac{11}{24})^k$ eher schlecht für $k \to \infty$ gegen 0.

(–0,5)

Aufgabe #3

Teilaufgabe (a)

```
In[1]:= A = {{c1 + c2, -c2, 0}, {-c2, c2 + c3, -c3}, {0, -c3, c3}}

Out[1]= {{c1 + c2, -c2, 0}, {-c2, c2 + c3, -c3}, {0, -c3, c3}}

In[2]:= MatrixForm[A]

Out[2]//MatrixForm=
```

$$\begin{pmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{pmatrix}$$

```
In[6]:= x[k_] := A.x[k - 1]

In[7]:= x[0] = {1, 1, 1}

Out[7]= {1, 1, 1}

In[18]:= x[1] // MatrixForm

Out[18]//MatrixForm=
```

$$\begin{pmatrix} c_1 \\ 0 \\ 0 \end{pmatrix}$$

```
In[17]:= x[2] // MatrixForm

Out[17]//MatrixForm=
```

$$\begin{pmatrix} c_1 \ (c_1 + c_2) \\ -c_1 \ c_2 \\ 0 \end{pmatrix}$$

```
In[16]:= x[3] // Simplify // MatrixForm

Out[16]//MatrixForm=
```

$$\begin{pmatrix} c_1 \ (c_2^2 + (c_1 + c_2)^2) \\ -c_1 \ c_2 \ (c_1 + 2 \ c_2 + c_3) \\ c_1 \ c_2 \ c_3 \end{pmatrix}$$

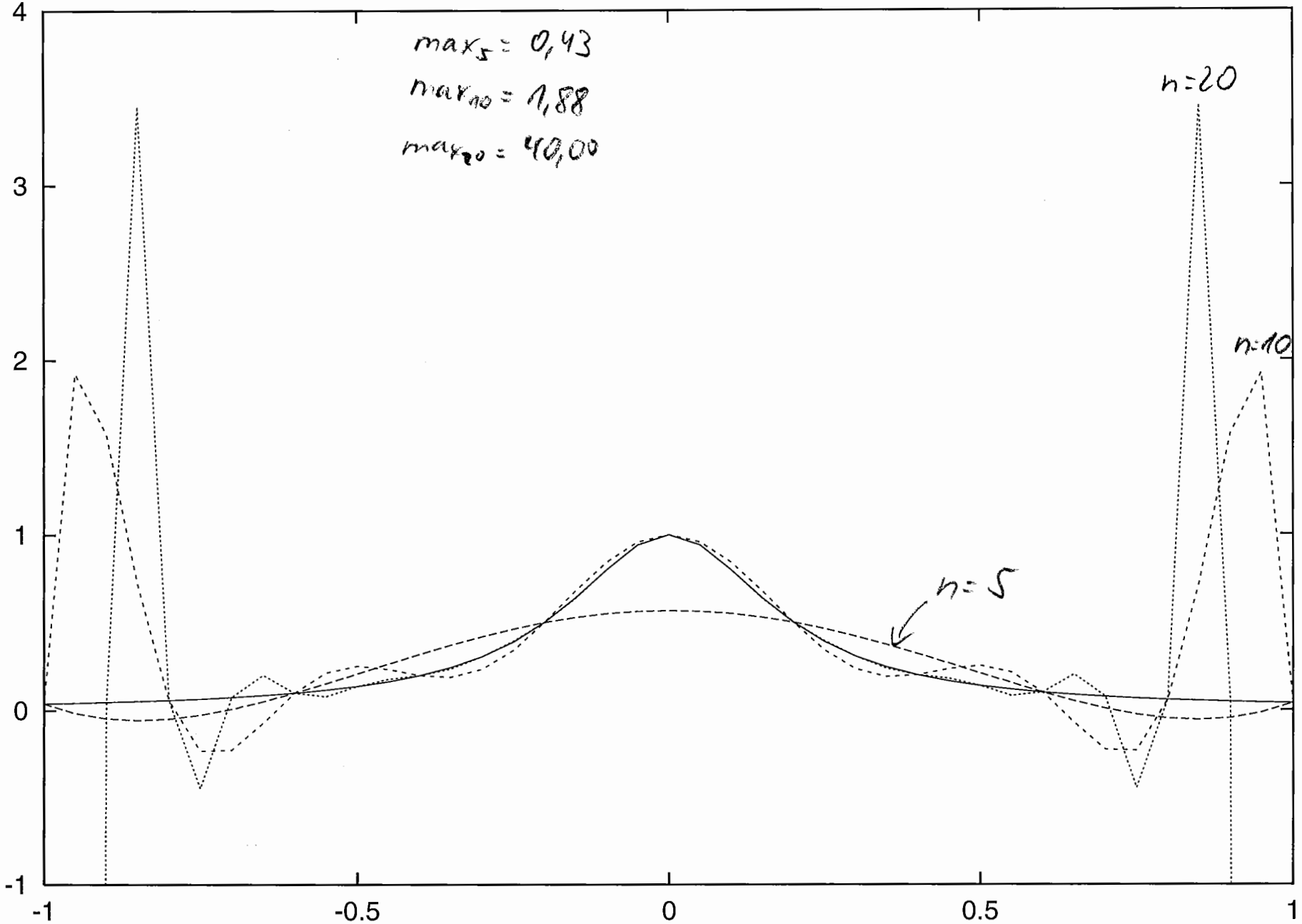
```
In[15]:= x[4] // Simplify // MatrixForm

Out[15]//MatrixForm=
```

$$\begin{pmatrix} c_1 \ (c_1^3 + 3 \ c_1^2 \ c_2 + 5 \ c_1 \ c_2^2 + c_2^2 \ (4 \ c_2 + c_3)) \\ -c_1 \ c_2 \ (c_1^2 + 4 \ c_2^2 + 3 \ c_2 \ c_3 + 2 \ c_3^2 + c_1 \ (3 \ c_2 + c_3)) \\ c_1 \ c_2 \ c_3 \ (c_1 + 2 \ (c_2 + c_3)) \end{pmatrix}$$

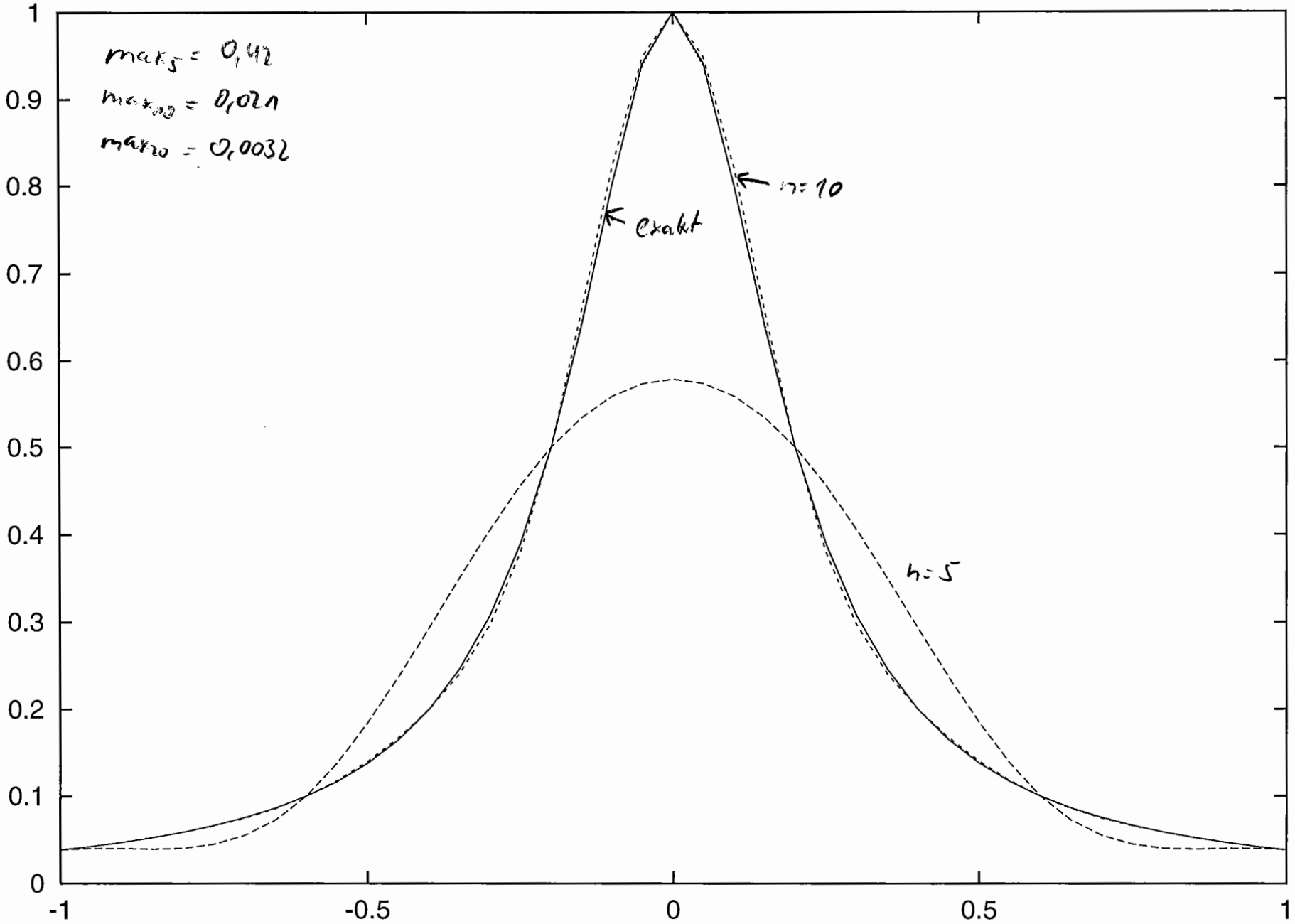
```
In[19]:= λ[k_, j_] := x[k + 1][[j]] / x[k][[j]]

(Dialog) In[44]:=
        EV[k_, j_] := x[k] / x[k][[j]]
```

Für j=1,2,3 nach dem 4. Iterationsschritt lautet die Schätzung für den größten Eigenwert und den dazugehörigen Eigenvektor:

Newton-Interpolation



Kubische Spline-Interpolation



Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 2 , Abgabe: 06.05.2004, 8.00 Uhr

Aufgabe 5: (4 Punkte)
Führen Sie die Matrix

$$A = \begin{pmatrix} 1 & 2 & -1 & 3 \\ -1 & 0 & -4 & 1 \\ 0 & 2 & -1 & 1 \\ 2 & 2 & 0 & 3 \end{pmatrix}$$

durch Ähnlichkeitstransformationen unter Verwendung von Elementarmatrizen in eine Hessenberg-Matrix H über. Geben Sie H explizit an.

Aufgabe 6: (6 Punkte)
Sei

$$A = \begin{pmatrix} \delta_1 & \gamma_2 & & & \mathbf{O} \\ \gamma_2 & \delta_2 & \gamma_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_n \\ \mathbf{O} & & & \gamma_n & \delta_n \end{pmatrix}$$

(a) Zeigen Sie: λ ist Eigenwert von A genau dann, wenn $-\lambda$ Eigenwert von B ist:

$$B := \begin{pmatrix} -\delta_1 & \gamma_2 & & & \mathbf{O} \\ \gamma_2 & -\delta_2 & \gamma_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \gamma_n \\ \mathbf{O} & & & \gamma_n & -\delta_n \end{pmatrix}$$

Hinweis: Betrachten Sie die Rekursion (17.4).

(b) Für die Matrix A gelte

$$\begin{aligned} \delta_i &= -\delta_{n+1-i} && (i = 1, \ldots, n)\,, \\ \gamma_i &= \gamma_{n+2-i} && (i = 2, \ldots, n)\,. \end{aligned}$$

Zeigen Sie: Mit λ ist auch $-\lambda$ Eigenwert von A.

(c) Für die Matrix A gelte

$$\begin{aligned} \delta_i + \delta_{n+1-i} &= 2c \quad , \quad c \in \mathbb{R} && (i = 1, \ldots, n)\,, \\ \gamma_i &= \gamma_{n+2-i} \quad , && (i = 2, \ldots, n)\,. \end{aligned}$$

Was kann man über die Lage der Eigenwerte von A aussagen?

$$A_1 = \left(\begin{array}{cc|cc} -1 & 0 & -4 & 1 \\ 0 & 2 & -1 & 1 \\ \hline 2 & 2 & 0 & 3 \end{array}\right)$$

1. Schritt:

$$|\alpha_{41}| = \max_{2 \le j \le 4} |\alpha_{j1}| = 2$$

$\Rightarrow$ vertausche Zeile und Spalte 2 und 4:

$$A' = P_{4,2}^{-1} A_1 P_{4,2} = P_{4,2}^{-1} \begin{pmatrix} 1 & 2 & -1 & 3 \\ -1 & 0 & -4 & 1 \\ 0 & 2 & -1 & 1 \\ 2 & 2 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 3 & -1 & 2 \\ -1 & 1 & -4 & 0 \\ 0 & 1 & -1 & 2 \\ 2 & 3 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 3 & -1 & 2 \\ 2 & 3 & 0 & 2 \\ 0 & 1 & -1 & 2 \\ -1 & 1 & -4 & 0 \end{pmatrix}$$

Eliminationsmatrix $L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 \end{pmatrix} \Rightarrow L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 1 \end{pmatrix}$

$$\Rightarrow A_2 = L_2^{-1} A' L_2 = L_2^{-1} \begin{pmatrix} 1 & 3 & -1 & 2 \\ 2 & 3 & 0 & 2 \\ 0 & 1 & -1 & 2 \\ -1 & 1 & -4 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 & 2 \\ 2 & 2 & 0 & 2 \\ 0 & 0 & -1 & 2 \\ -1 & 1 & -4 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -1 & 2 \\ 2 & 2 & 0 & 2 \\ 0 & 0 & -1 & 2 \\ 0 & 2 & -4 & 1 \end{pmatrix}$$

$$f_j'(\lambda) = \frac{p_n'(\lambda)}{\prod_{i=1}^{j}(\lambda-\lambda_i)} + p_n(\lambda) \sum_{i=1}^{j} \frac{1}{\prod_{k=1}^{j}(\lambda-\lambda_k)} \cdot \frac{-1}{\lambda-\lambda_i}$$

$$= \frac{1}{\prod_{i=1}^{j}(\lambda-\lambda_i)} \left(p_n'(\lambda) - p_n(\lambda) \sum_{i=1}^{j} \frac{1}{\lambda-\lambda_i} \right)$$

$$\Rightarrow \lambda - \frac{f_j(\lambda)}{f_j'(\lambda)} = \lambda - \frac{p_n(\lambda)}{p_n'(\lambda) - \sum_{i=1}^{j} \frac{p_n(\lambda)}{\lambda-\lambda_i}}$$

6)a) Beh.: $p_{n,B}(-\lambda) = (-1)^n p_{n,A}(\lambda)$

Ber. mit voller Induktion nach n:

I.A.: $n=0$: $p_{0,B}(-\lambda) = 1 = p_{0,A}(\lambda)$

$n=1$: $p_{1,B}(-\lambda) = -\delta_1 + \lambda = -(\delta_1 - \lambda) = (-1)\, p_{1,A}(\lambda)$

I.S.: $n \to n+1$ $(n \geq 1)$

$$p_{n+1,B}(\lambda) = (-\delta_{n+1} + \lambda)\, p_{n,B}(-\lambda) - \gamma_{n+1}^2\, p_{n-1,B}(\lambda)$$

$$= -(\delta_{n+1} - \lambda)(-1)^n p_{n,A}(\lambda) - \gamma_{n+1}^2 (-1)^{n-1} p_{n-1,A}(\lambda)$$

$$= (-1)^{n+1} \left((\delta_{n+1} - \lambda)\, p_{n,A}(\lambda) - \gamma_{n+1}^2 (-1)^{n+1} p_{n-1,A}(\lambda) \right)$$

$$= (-1)^{n+1} p_{n+1,A}(\lambda)$$

$\Rightarrow p_{n,A}(\lambda) = 0 \quad (\Leftrightarrow p_{n,B}(-\lambda) = 0$

```
int n=5;
double *delta, *Gamma, *q, p, dp;
double *dvector(int, int), **dmatrix(int, int, int, int);


double polynom_auswertung(double x) {
  /* liefert p_n(x), p'_n(x) und q(x) */
  int j;
  double p2, p1, dp2, dp1, temp, fak;

  p2 = 1.0;
  p1 = delta[1] - x;
  fak = -1.0/Gamma[2];
  q[0] = 1.0;
  q[1] = p1*fak;
  dp2 = 0.0;
  dp1 = -1.0;
  for (j = 2; j <= n; j++) {
    temp = -p1 + (delta[j] - x)*dp1 - Gamma[j]*Gamma[j]*dp2;
    dp2 = dp1;
    dp1 = temp;
    temp = (delta[j] - x)*p1 - Gamma[j]*Gamma[j]*p2;
    p2 = p1;
    p1 = temp;
    fak /= -Gamma[j+1];
    q[j] = p1*fak;
  }
  p = p1;
  dp = dp1;
}


main() {
  int j, k;
  double *lambda, **x, lambda0, c = 16.0, eps = 1.0e-10, temp, d;
  FILE  *erg, *file_box(char*, char*);

  delta = dvector(1, n);          /* Diagonalwerte */
  Gamma = dvector(2, n+1);        /* Nebendiagonalenwerte */
  lambda = dvector(1, n);         /* Eigenwerte */
  q = dvector(0, n);              /* Eigenvektor */
  x = dmatrix(1, n, 1, n);        /* Matrix mit allen Eigenvektoren */

  /* Einlesen der Matrix */
  for (j = 1; j <= n; j++) {
    delta[j] = c - (j-3)*8.0;
    Gamma[j+1] = 2.0;
  }
  Gamma[n+1] = 1.0;

  /* Startwert für lambda */
  lambda0 = abs(delta[1]) + abs(Gamma[2]);
  for (j = 2; j <= n; j++) {
    temp = abs(Gamma[j]) + abs(delta[j]) + abs(Gamma[j+1]);
    if (temp > lambda0)
      lambda0 = temp;
  }

  /* Bestimmung der ersten Eigenwerte */
  for (j = 1; j <= n/2; j++) {
    d = 1.0;
    while (fabs(d) > eps) {
      polynom_auswertung(lambda0);
```

```
        temp += 1.0/(lambda0 - lambda[k]);
        k++;
      }
      d = p/(dp-p*temp);
      lambda0 -= d;
    }
    lambda[j] = lambda0;
    for (k = 1; k <= n; k++) {
      x[k][j] = q[k-1];
    }
    lambda[n+1-j] = 2*c - lambda0;
    polynom_auswertung(lambda[n+1-j]);
    for (k = 1; k <= n; k++) {
      x[k][n+1-j] = q[k-1];
    }
    /* neuer Startwert */
    lambda0 = lambda[j]-0.5;
  }

  /* Bestimmung des mittleren Eigenwertes */
  if (n % 2 == 1) {
    lambda[n/2+1] = c;
    polynom_auswertung(c);
    for (k = 1; k <= n; k++) {
      x[k][n/2+1] = q[k-1];
    }
  }

  /* Ausgabe */
  erg = file_box("Dateiname:", "w");
  fprintf(erg, "Eigenwerte:\n");
  for (j = 1; j <= n; j++) {
    fprintf(erg, "%13.10g", lambda[j]);
  }
  fprintf(erg, "\n\n");
  fprintf(erg, "Eigenvektoren:\n");
  for (k = 1; k <= n; k++) {
    for (j = 1; j <= n; j++) {
      fprintf(erg, "%13.8g", x[k][j]);
    }
    fprintf(erg, "\n");
  }

  return ;
}
```

Eigenvektoren:

1	1	1	1	1
0.24270851	-3.9927869	-8	-12.007213	-16.242709
0.02974145	-1.0288002	31	95.144313	197.85475
0.0024415978	-0.12983459	8	-369.25632	-1614.6163
0.00015031962	-0.01081305	1	-92.480847	6652.4915

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 3 , Abgabe: 13.05.2004, 8.00 Uhr

Aufgabe 9: (4 Punkte)
Sei

$$A = \begin{pmatrix} 4 & 0 & 2 \\ -2 & 8 & 2 \\ 0 & 2 & -4 \end{pmatrix} .$$

(a) Zeigen Sie mit dem Satz von Gerschgorin, dass A genau einen Eigenwert mit negativem Realteil hat.

(b) Bestimmen Sie drei paarweise disjunkte Gerschgorin-Kreisscheiben, in denen jeweils ein Eigenwert von A liegt.

(c) Geben Sie eine möglichst gute Abschätzung für den größten Eigenwert.

Hinweis zu b) u. c): Betrachten Sie $A' = D^{-1}AD$ mit $D = \mathrm{diag}(1, c, 1)$, $c > 0$.

Aufgabe 10: (6 Punkte)
Gegeben sei die $(n \times n)$-Matrix

$$C(\lambda) = \begin{pmatrix} \lambda & 1 & & & \mathbf{O} \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ \mathbf{O} & & & & \lambda \end{pmatrix}$$

und die $(n \times n)$–Matrix F.

Zeigen Sie: Für die Eigenwerte $\lambda_i(\epsilon)$ der gestörten Matrix $C(\lambda) + \epsilon F$ gilt für genügend kleines ϵ die Abschätzung

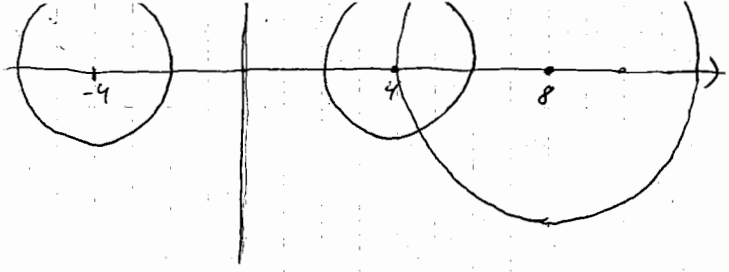
$$|\lambda_i(\epsilon) - \lambda| \leq |\epsilon^{1/n}|(1 + \|F\|_\infty) .$$

Zeigen Sie durch spezielle Wahl der Matrix F, dass der Fall $\lambda_i(\epsilon) - \lambda = O(\epsilon^{1/n})$ tatsächlich auftritt.

Hinweis: Ähnlichkeitstransformation mit $D = \mathrm{diag}\,(1, \delta, \delta^2, \ldots, \delta^{n-1})$, $\delta = \epsilon^{1/n}$. Benutzen Sie den Satz von Gerschgorin.

2. Kreis: $|\lambda - 8| \leq 4$

3. Kreis: $|\lambda + 4| \leq 2$



$K_3 \cap (K_1 \cup K_2) = \emptyset$

$\Rightarrow$ Es liegen zwei EW in $K_1 \cup K_2$ und ein EW in K_3

$\Rightarrow$ genau 1 EW hat negativen Realteil

b) Sei $D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $c > 0$

$$\Rightarrow A' = D^{-1} A D = D^{-1} \begin{pmatrix} 4 & 0 & 2 \\ -2 & 8c & 2 \\ 0 & 2c & -4 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 2 \\ -\frac{2}{c} & 8 & \frac{2}{c} \\ 0 & 2c & -4 \end{pmatrix}$$

$\Rightarrow r_1 = 2$, $r_2 = \frac{4}{c}$, $r_3 = 2c$

Damit die 3 Kreise disjunkt sind, muss gelten:

i) $r_2 < 2 \Leftrightarrow \frac{4}{c} < 2 \Leftrightarrow c > 2$

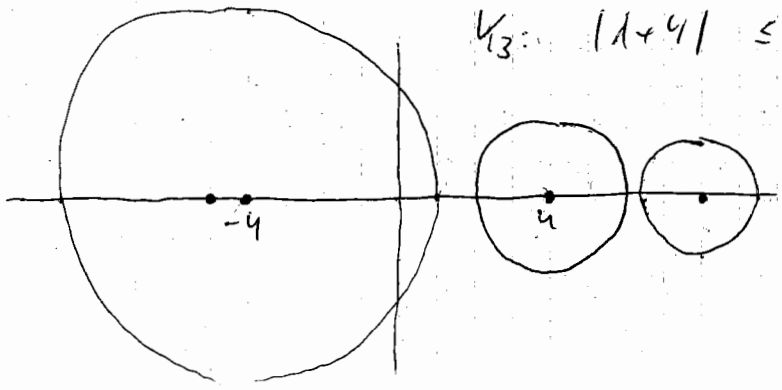
ii) $r_3 < 6 \Leftrightarrow 2c < 6 \Leftrightarrow c < 3$

also $2 < c < 3$

z.B. $c = 2{,}5$: K_1: $|\lambda - 4| \leq 2$

K_2: $|\lambda - 8| \leq 1{,}6$

K_3: $|\lambda + 4| \leq 5$



$$\Rightarrow C(\lambda) = \begin{pmatrix} \lambda & \delta & & 0 \\ & \lambda\delta & \delta^2 & \\ & & \lambda\delta^2 & \ddots \\ & & & \delta^{n-1} \\ 0 & & & \lambda\delta^{n-1} \end{pmatrix}$$

$$D^{-1}C(\lambda) = \begin{pmatrix} \lambda & \delta & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \delta \\ 0 & & & \lambda \end{pmatrix}$$

$$D^{-1}FD = \begin{pmatrix} f_{11} & \delta f_{12} & & & \delta^{n-1} f_{1,n} \\ \delta^{-1} f_{21} & f_{22} & \delta f_{23} & & \\ & \delta^{-1} f_{32} & f_{33} & \ddots & \\ & & \ddots & \ddots & \delta f_{n-1,n} \\ \delta^{1-n} f_{n,1} & & & \delta^{-1} f_{n,n-1} & f_{n,n} \end{pmatrix}$$

Wende Gerschgorin auf $D^{-1}(C(\lambda) + \varepsilon F)D = D^{-1}C(\lambda)D + \varepsilon D^{-1}FD$ an: Die Kreise haben die Mittelpunkte $m_i = \lambda + \varepsilon f_{ii}$

Für die Radien r_i gilt:

$$r_i \leq \delta + \varepsilon \sum_{\substack{j=1 \\ j \neq i}}^{n} \delta^{j-i} |f_{ij}| \underset{\delta < 1}{\leq} \delta + \varepsilon \underbrace{\delta^{1-n}}_{= \delta^n \delta^{1-n} = \delta} \sum_{\substack{j=1 \\ j \neq i}}^{n} |f_{ij}|$$

$$= \delta \left(1 + \sum_{\substack{j=1 \\ j \neq i}}^{n} |f_{ij}|\right)$$

$\Rightarrow$ Die Eigenwerte $\lambda(\varepsilon)$ liegen in der Vereinigung der Kreise

$$|x - (\lambda + \varepsilon f_{ii})| \leq \delta \left(1 + \sum_{\substack{j=1 \\ j \neq i}}^{n} |f_{ij}|\right)$$

mit $A = U^T D U$, D Diagonalmatrix

Für $x \neq 0$ folgt:

$$\frac{\|Ax - \lambda x\|_2}{\|x\|_2} = \frac{\|U^T D U x - \lambda x\|_2}{\|x\|_2} = \frac{\|U^T (D - \lambda I) U x\|_2}{\|x\|_2}$$

$$= \frac{\|(D - \lambda I) U x\|_2}{\|U x\|_2} \quad (\text{da } \|U^T x\|_2 = \|U x\|_2 = \|x\|_2)$$

$$\geq \min_{y \neq 0} \frac{\|(D - \lambda I) y\|_2}{\|y\|_2}$$

$$\underset{(*)}{=} \min_i |d_{ii} - \lambda| = \min_i |\lambda_i - \lambda|$$

zu $(*)$: Für jede Diagonalmatrix D gilt:

$$\|Dy\|_2 = \left(\sum_{i=1}^{n} d_{ii}^2 y_i^2 \right)^{\frac{1}{2}} \geq \min_i |d_{ii}| \, \|y\|_2$$

$$\Rightarrow \frac{\|Dy\|_2}{\|y\|_2} \geq \min_i |d_{ii}|$$

Ausserdem: $\frac{\|De_i\|_2}{\|e_i\|_2} = |d_{ii}| \Rightarrow \min_{y \neq 0} \frac{\|Dy\|_2}{\|y\|_2} = \min_i |d_{ii}|$

b) $$Ax - \lambda x = \begin{pmatrix} 12{,}7 \\ 11{,}9 \\ 11{,}2 \end{pmatrix} - 12 \begin{pmatrix} 0{,}9 \\ 1{,}0 \\ 1{,}1 \end{pmatrix} = \begin{pmatrix} 1{,}9 \\ -0{,}1 \\ -2{,}0 \end{pmatrix}$$

$$\Rightarrow \frac{\|Ax - \lambda x\|_2}{\|x\|_2} = \frac{2{,}7604}{1{,}7378} \approx 1{,}59$$

$\Rightarrow$ A hat ein EW im Intervall $[10{,}41 ; 13{,}59]$

EW von A: $\lambda_1 = 11{,}42$, $\lambda_2 = 12{,}51$, $\lambda_3 = 12{,}07$

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 4 , Abgabe: Freitag, den 21.05.2004, 8.00 Uhr

Aufgabe 12: (3 Punkte)
Seien $\|\cdot\|_a$ und $\|\cdot\|_b$ Normen auf dem Vektorraum V; $\|\cdot\|_a$ sei streng. Zeigen Sie, dass dann auch die durch $\|v\| := \|v\|_a + \|v\|_b$, $v \in V$, definierte Norm streng ist.

Aufgabe 13: (4 Punkte)
Zeigen Sie, dass der Raum $C[-1,1]$ bezüglich der Normen $\|\cdot\|_2$ und $\|\cdot\|_1$ *nicht* vollständig ist. Untersuchen Sie dazu die Konvergenz der Folge $(f_n)_{n\in\mathbb{N}_+}$ mit

$$f_n(x) = \begin{cases} -1 & \text{für} \quad -1 \le x \le -\frac{1}{n}\,, \\ nx & \text{für} \quad -\frac{1}{n} \le x \le \frac{1}{n}\,, \\ 1 & \text{für} \quad \frac{1}{n} \le x \le 1\,. \end{cases}$$

Aufgabe 14: (4 Punkte)
Betrachten Sie in $V = C[0,1]$ mit der Norm $\|\cdot\|_\infty$ die Teilmenge

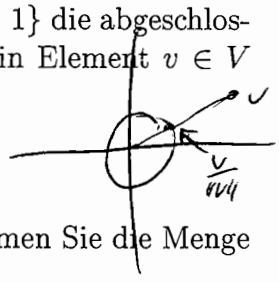
$$T = \{u \in C[0,1] \mid u(0) = 0\}\,.$$

Zeigen Sie, dass die Folge $(u_n)_{n\in\mathbb{N}_+}$, $u_n(x) = x^n$, eine Minimalfolge für das Element $v \in V$ mit $v(x) \equiv 1$ ist, welche <u>nicht</u> gegen ein Element aus T konvergiert.

Aufgabe 15: (4 Punkte)

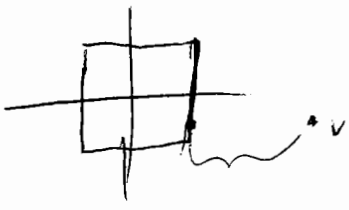
(a) In dem normierten Vektorraum $(V, \|\cdot\|)$ sei $B = \{u \in V \mid \|u\| \le 1\}$ die abgeschlossene Einheitskugel. Man zeige, dass <u>ein</u> Proximum $\tilde{u} \in B$ an ein Element $v \in V$ gegeben ist durch

$$\tilde{u} = \left\{ \begin{array}{ll} v & , \text{ falls } \quad v \in B \\ v/\|v\| & , \text{ falls } \quad v \notin B \end{array} \right\}.$$



(b) Sei $V = \Pi_1$ versehen mit der Norm $\|p\| = |p(0)| + |p(1)|$. Bestimmen Sie die Menge aller Proxima $\tilde{u} \in T = \Pi_0$ an das Polynom $v(x) = x$.

Hinweis: Bitte denken Sie an die Abgabe der Programmieraufgabe 8 !



Es gelte $\|v+w\| = \|v\| + \|w\|$

Zz: v, w sind linear abhängig.

Nach Definition gilt: $\|v+w\|_a + \|v+w\|_b$

$$= \|v\|_a + \|v\|_b + \|w\|_a + \|w\|_b \quad (*)$$

$$\Rightarrow \|v+w\|_a + \|v+w\|_b \leq \|v\|_a + \|w\|_a + \|v+w\|_b \quad \left.\begin{matrix}\\\\\end{matrix}\right\} \text{Dreiecks-ungleichung}$$

$$\leq \|v\|_a + \|w\|_a + \|v\|_b + \|w\|_b$$

$$\overset{(*)}{=} \|v+w\|_a + \|v+w\|_b$$

Somit gilt überall "=", insbesondere ist

$$\|v+w\|_a + \|v+w\|_b = \|v\|_a + \|w\|_a + \|v+w\|_b$$

$$\Rightarrow \|v+w\|_a = \|v\|_a + \|w\|_a$$

$\Rightarrow v, w$ linear abhängig, da $\|\cdot\|_a$ Norm

13) Es ist

$$|f_n(x) - \operatorname{sgn}(x)| = \begin{cases} 0 & -1 \leq x \leq -\frac{1}{n} \\ nx+1 & -\frac{1}{n} \leq x < 0 \\ 0 & x = 0 \\ -nx+1 & 0 < x \leq \frac{1}{n} \\ 0 & \frac{1}{n} \leq x \leq 1 \end{cases}$$

Sei $\alpha \in \{1, 2\}$

$$\Rightarrow \|f_n - \operatorname{sgn}\|_\alpha = \left(\int_{-\frac{1}{n}}^{0} \underbrace{(nx+1)^\alpha}_{\leq 1} dx + \int_{0}^{\frac{1}{n}} \underbrace{(-nx+1)^\alpha}_{\leq 1} dx \right)^{\frac{1}{\alpha}}$$

$$\leq \left(\frac{1}{n} + \frac{1}{n}\right)^{\frac{1}{\alpha}} = \left(\frac{2}{n}\right)^{\frac{1}{\alpha}} \xrightarrow[n \to \infty]{} 0$$

Sei $u \equiv 0 \in T \Rightarrow \|v-u\| = 1$

$\Rightarrow e_T(v) = \inf_{u \in T} \|v-u\| = 1$

Für $u_n(x) = x^n \in T$ gilt:

$$\|v-u_n\| = \max_{x \in [0,1]} |1-x^n| = 1$$

$\Rightarrow e_T(v) = \lim_{n\to\infty} \|v-u_n\|$, d.h. (u_n) ist Minimalfolge

A: Es gibt ein $u \in T$ mit $\lim_{n\to\infty} u_n = u$, d.h.

(u_n) konvergiert glm gegen u

$\Rightarrow (u_n)$ konvergiert punktweise gegen u

Für $x \in [0,1)$ gilt: $\lim_{n\to\infty} u_n(x) = x^n = 0$

Außerdem ist $\lim_{n\to\infty} u_n(1) = 1^n = 1$

$$\Rightarrow u(x) = \lim_{n\to\infty} u_n(x) = \begin{cases} 0 & 0 \le x < 1 \\ 1 & x = 1 \end{cases}$$

$\Rightarrow u \notin T$, da u nicht stetig $\square$

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 5 , Abgabe: 27.05.2004, 8.00 Uhr

Aufgabe 16: (4 Punkte)
Seien $V = C[-1,1]$ und $f(x) = \sin(\pi x) \in V$. Man bestimme das Proximum an f aus Π_k, $0 \leq k \leq 2$, bezüglich der Norm $\| \cdot \|_2$

(a) über die Normalgleichungen;

(b) durch Entwickeln von f nach LEGENDRE-Polynomen.

Aufgabe 17: (4 Punkte)
Gegeben sei der Prä-Hilbertraum $(C[-1,1], \| \cdot \|)$, dessen Norm durch das innere Produkt $(f,g) = \int_{-1}^{1} f(x)g(x)\sqrt{1-x^2}dx$ induziert wird. Zeigen Sie, dass die Funktionen

$$U_n(x) = \sqrt{\frac{2}{\pi}} \frac{\sin((n+1)\arccos(x))}{\sqrt{1-x^2}}$$

ein Orthonormalsystem bilden.

$U_n(x) = U_{n-1}(x)\,x + T_n(x)\sqrt{\frac{1}{\alpha}}$
Un Tschebyscheff-Polynome 2. Art

Aufgabe 18: (4 Punkte)
Die periodische Funktion $f \in C(\mathbb{R})$ sei definiert durch periodische Fortsetzung von $f(x) = x^2$ für $x \in [-\pi, \pi]$.

(a) Berechnen Sie die FOURIER-Entwicklung von f und skizzieren Sie den Verlauf der Proxima (19.7) aus span (u_0, u_1, u_2) und span $(u_0, u_1, u_2, u_3, u_4)$.

(b) Berechnen Sie das Proximum an f aus span $(1, \cos(x), \sin(x), \cos(2x), \sin(2x))$ bezüglich der durch das innere Produkt

$$(f,g) := \sum_{k=1}^{6} f(x_k)g(x_k)\,, \quad x_k = (k-1)\frac{2\pi}{6}$$

induzierten Norm auf $\mathbb{R}^6$. Vergleichen Sie das Ergebnis mit dem Ergebnis in (a).

Aufgabe 19: (5 Punkte)
Sei $V = C[-1,1]$ versehen mit dem inneren Produkt

$$(f,g) = \int_{-1}^{1} f(x)g(x)x^2dx$$

und der Norm $\|f\|_2 = \sqrt{(f,f)}$.
Berechnen Sie eine Orthonormalbasis von $\Pi_2 = \text{span}\{1, x, x^2\}$ und bestimmen Sie das Proximum von x^4 in Π_2. Skizzieren Sie das Proximum und die Funktion x^4.

$$\Rightarrow (u_i, u_k) = \int_{-1}^{1} u_i(x)\, u_k(x)\, dx = \int_{-1}^{1} x^{i+k}\, dx = \frac{1+(-1)^{i+k}}{i+k+1}$$

$$(f, u_0) = \int_{-1}^{1} \sin(\pi x)\, dx = \left[-\frac{1}{\pi} \cos \pi x \right]_{-1}^{1} = 0$$

$$(f, u_1) = \int_{-1}^{1} \sin(\pi x)\, x\, dx = \left[x \, \frac{-\cos \pi x}{\pi} \right]_{-1}^{1} + \int_{-1}^{1} \frac{\cos \pi x}{\pi}\, dx$$

$$= \frac{2}{\pi} + \left[\frac{\sin \pi x}{\pi^2} \right]_{-1}^{1} = \frac{2}{\pi}$$

$$(f, u_2) = \int_{-1}^{1} \sin(\pi x)\, x^2\, dx = 0 \quad \text{(da Integrand ungerade)}$$

$k=0$: Normalgleichung:

$$2\tilde{\alpha}_0 = 0 \quad \Rightarrow \tilde{\alpha}_0 = 0 \quad \Rightarrow \tilde{u}(x) = 0$$

$k=1$: Normalgleichung:

$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{2}{\pi} \end{pmatrix} \quad \Rightarrow \tilde{\alpha}_0 = 0, \ \tilde{\alpha}_1 = \frac{3}{\pi}$$

$$\Rightarrow \tilde{u}(x) = \frac{3}{\pi} x \qquad \left(\|f - \tilde{u}\|_2 = \sqrt{1 - \frac{6}{\pi^2}} \approx 0{,}62 \right)$$

$k=2$: Normalgleichung

$$\begin{pmatrix} 2 & 0 & \frac{2}{3} \\ 0 & \frac{2}{3} & 0 \\ \frac{2}{3} & 0 & \frac{2}{5} \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{2}{\pi} \\ 0 \end{pmatrix}$$

$$\Rightarrow \tilde{\alpha}_0 = \tilde{\alpha}_2 = 0, \ \tilde{\alpha}_1 = \frac{3}{\pi}$$

$$\Rightarrow \tilde{u}(x) = \frac{3}{\pi} x$$

$$U_n(x) = \sqrt{\frac{2}{\pi}} \frac{\sin((n+1)\arccos x)}{\sqrt{1-x^2}}$$

Seien $n, m \geq 0$. Dann gilt:

$$(U_n, U_m) = \frac{2}{\pi} \int_{-1}^{1} \sin((n+1)\arccos x) \sin((m+1)\arccos x) \frac{dx}{\sqrt{1-x^2}}$$

$(y = \arccos x)$
$x = \cos y$
$dx = -\sin y\, dy = -\sqrt{1-\cos^2 y}\, dy$
$\Rightarrow dy = \frac{-1}{\sqrt{1-x^2}} dx$

$$= \frac{2}{\pi} \int_0^{\pi} \sin((n+1)y) \sin((m+1)y)\, dy$$

$$= \frac{2}{\pi} \int_0^{\pi} (n+1) \cos((n+1)y) \frac{\cos((m+1)y)}{m+1}\, dy$$

part. Int.

$$+ \frac{2}{\pi} \underbrace{\left[\sin((n+1)y) \frac{-\cos((m+1)y)}{m+1} \right]_0^{\pi}}_{=0}$$

$$= \frac{2}{\pi} \frac{n+1}{m+1} \int_0^{\pi} \cos((n+1)y) \cos((m+1)y)\, dy$$

part. Int.

$$= \frac{2}{\pi} \frac{n+1}{m+1} \int_0^{\pi} (m+1) \sin((n+1)y) \frac{\sin((m+1)y)}{m+1}\, dy$$

$$+ \underbrace{[\ldots]_0^{\pi}}_{=0}$$

$$= \frac{2}{\pi} \left(\frac{n+1}{m+1}\right)^2 \int_0^{\pi} \sin((n+1)y) \sin((m+1)y)\, dy$$

$$\Rightarrow \frac{2}{\pi} \int_0^{\pi} \sin((n+1)y) \sin((m+1)y)\, dy = \frac{2}{\pi} \left(\frac{n+1}{m+1}\right)^2 \int_0^{\pi} \sin((n+1)y) \underbrace{\sin((m+1)y)}_{\sin((n+1)y)}\, dy$$

$$\Rightarrow \frac{2}{\pi} \underbrace{\left(1 - \left(\frac{n+1}{m+1}\right)^2\right)}_{\neq 0 \text{ falls } n \neq m} \int_0^{\pi} \sin((n+1)y) \sin((m+1)y)\, dy = 0$$

$$\Rightarrow \int_0^{\pi} \sin((n+1)y) \sin((m+1)y)\, dy = 0 \quad \text{falls } n \neq m$$

$$\Rightarrow (U_n, U_m) = 0 \quad \text{falls } n \neq m$$

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} x\,dx = \frac{1}{2\pi}\left[x^2\right]_{-\pi}^{\pi} = \frac{2}{3}\pi$$

$$a_1 = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\cos x\,dx = \frac{1}{\pi}\underbrace{\left[x^2 \sin x\right]_{-\pi}^{\pi}}_{=0} - \frac{2}{\pi}\int_{-\pi}^{\pi} x\sin x\,dx$$

$$= \frac{-2}{\pi}\left(\left[-x\cos x\right]_{-\pi}^{\pi} + \underbrace{\int_{-\pi}^{\pi}\cos x\,dx}_{=0}\right)$$

$$= -\frac{2}{\pi}\,(\pi+\pi) = -4$$

$$a_2 = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\cos 2x\,dx = \frac{1}{\pi}\left(\underbrace{\left[x^2\,\frac{\sin 2x}{2}\right]_{-\pi}^{\pi}}_{=0} - \int_{-\pi}^{\pi} x\sin 2x\,dx\right)$$

$$= -\frac{1}{\pi}\left(\left[x\,\frac{-\cos 2x}{2}\right]_{-\pi}^{\pi} + \frac{1}{2}\int_{-\pi}^{\pi}\cos 2x\,dx\right)$$

$$= -\frac{1}{\pi}\left(-\frac{\pi}{2} - \frac{\pi}{2}\right) - \frac{1}{2\pi}\left[\frac{\sin 2x}{2}\right]_{-\pi}^{\pi} = 1$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x^2\sin(kx)\,dx = 0 \qquad \text{da Integrand ungerade}$$

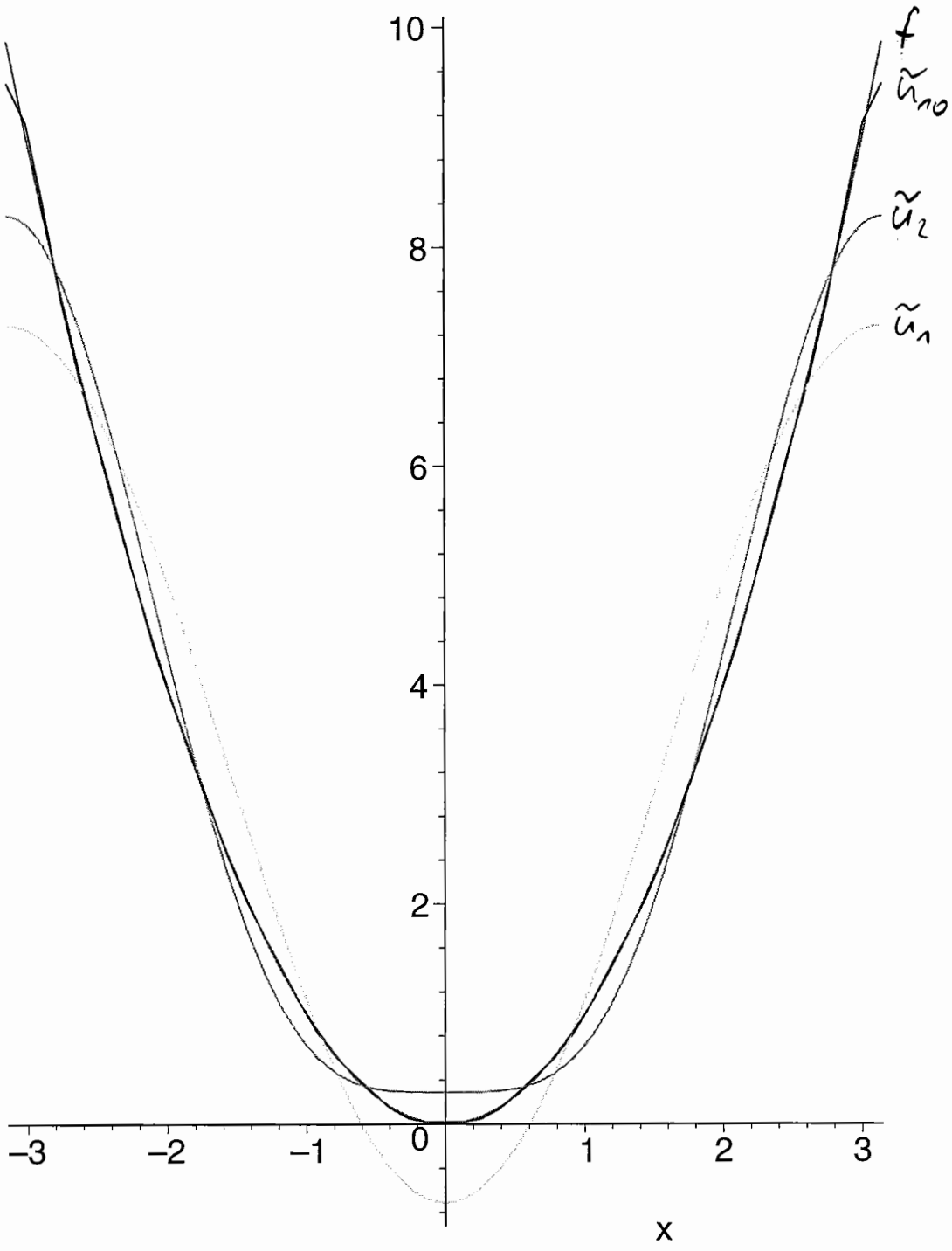
$m=1$: $U_1 = \operatorname{span}(u_0, u_1, u_2)$

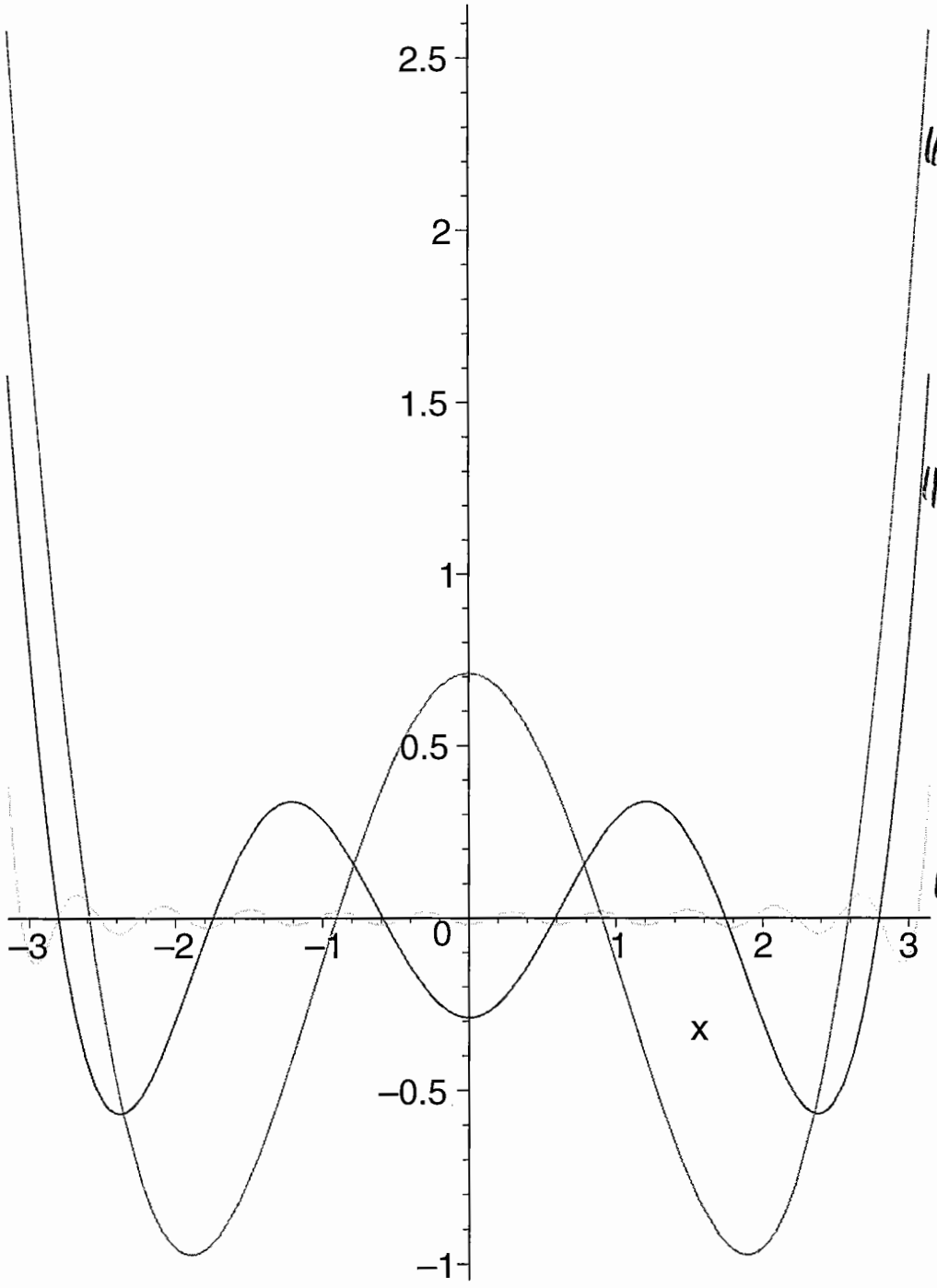
$$\tilde{u}(x) = \frac{\pi^2}{3} - 4\cos x$$

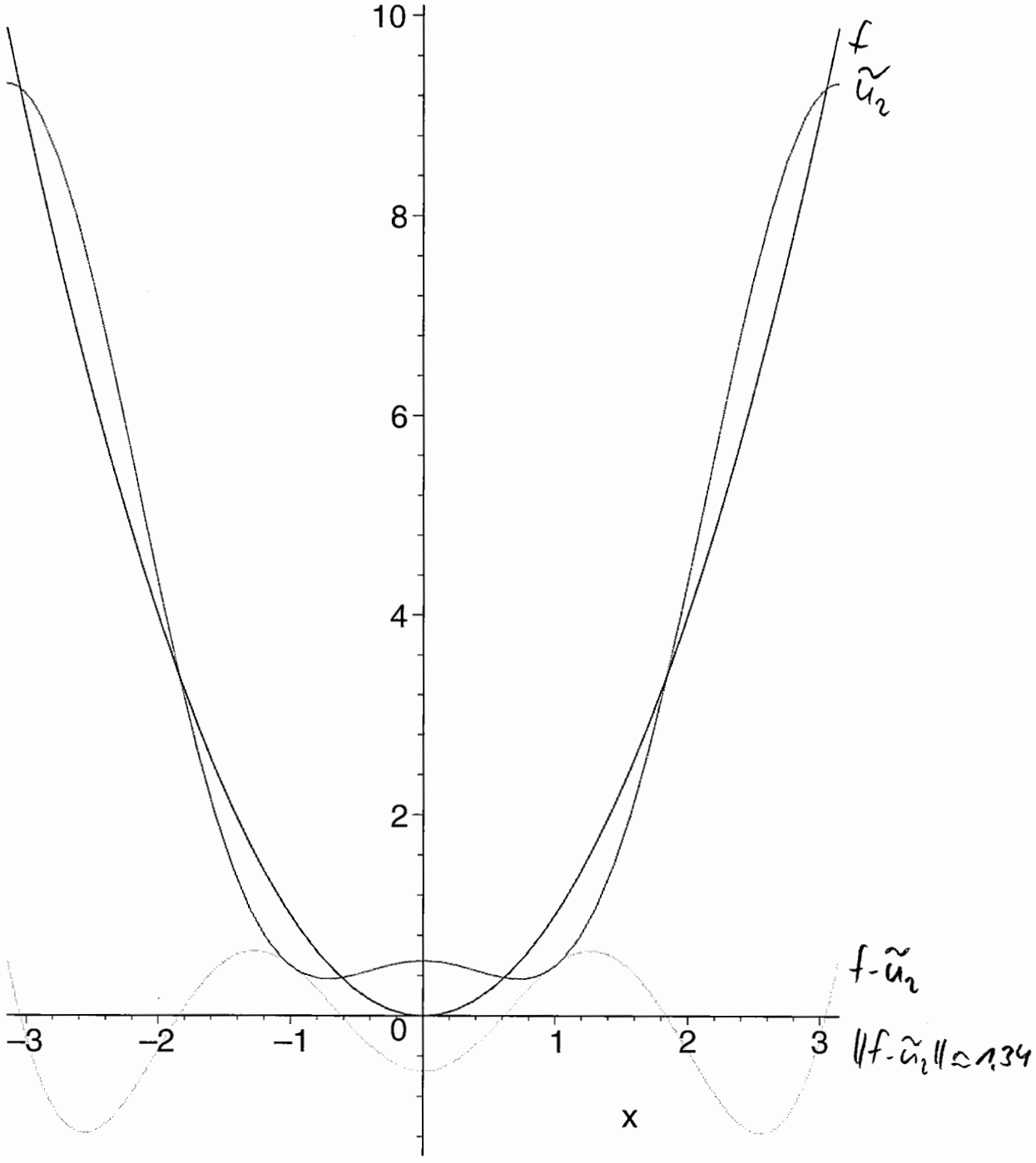
$m=2$: $U_2 = \operatorname{span}(u_0, u_1, u_2, u_3, u_4)$

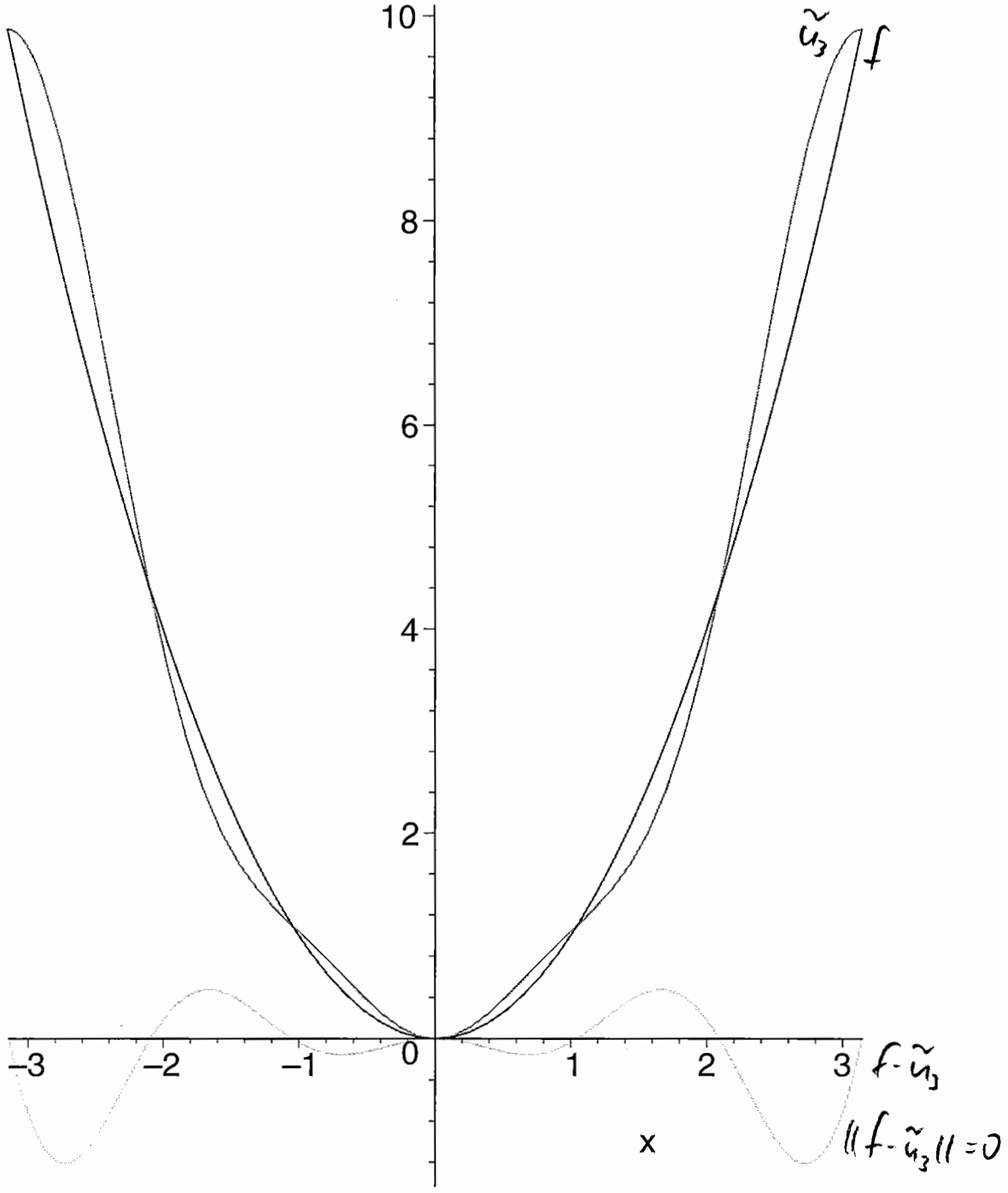
$$\tilde{u}(x) = \frac{\pi^2}{3} - 4\cos x + \cos 2x$$

$$a_k = (-1)^k\,\frac{4}{k^2}$$


10
8
6
4
2
0
−3
−2
−1
1
2
3
x
f
$\tilde{u}_{n0}$
$\tilde{u}_2$
$\tilde{u}_1$


2.5
2
1.5
1
0.5
−0.5
−1
−3
−2
−1
0
1
2
3
x
$f-\tilde{u}_n$
$\|f-\tilde{u}_n\|_2 \approx 2,03$
$f-\tilde{u}_2$
$\|f-\tilde{u}_2\|_2 \approx 1,00$
$f-\tilde{u}_{n0}$
$\|f-\tilde{u}_{n0}\|_2 \approx 0,17$


10
8
6
4
2
0
−3
−2
−1
1
2
3
x
f
$\tilde{u}_2$
$f-\tilde{u}_2$
$\|f-\tilde{u}_2\| \simeq 1{,}34$


10
8
6
4
2
0
−3
−2
−1
1
2
3
x
$\tilde{u}_3$
f
$f-\tilde{u}_3$
$\|f-\tilde{u}_3\| = 0$

$$= \begin{cases} \frac{2}{n+1} & n \text{ gerade} \\ 0 & n \text{ ungerade} \end{cases}$$

Bestimme Orthonormalbasis:

$\tilde{p}_0(x) = 1$

$$(\tilde{p}_0, \tilde{p}_0) = \int_{-1}^{1} x^2 dx = \frac{2}{3}$$

$$\Rightarrow p_0(x) = \frac{\tilde{p}_0(x)}{\sqrt{(\tilde{p}_0, \tilde{p}_0)}} = \sqrt{\frac{3}{2}}$$

$$\tilde{p}_1(x) = x - \delta_1 = x - \frac{(x\tilde{p}_0, \tilde{p}_0)}{(\tilde{p}_0, \tilde{p}_0)} = x - \frac{3}{2} \underbrace{\int_{-1}^{1} x^3 dx}_{=0} = x$$

$$(\tilde{p}_1, \tilde{p}_1) = \int_{-1}^{1} x^4 dx = \frac{2}{5}$$

$$\Rightarrow p_1(x) = \sqrt{\frac{5}{2}}\, x$$

$$\tilde{p}_2(x) = (x - \delta_2)\, \tilde{p}_1(x) - \gamma_2^2\, \tilde{p}_0(x)$$

$$= \left(x - \frac{(x\tilde{p}_1, \tilde{p}_1)}{(\tilde{p}_1, \tilde{p}_1)}\right) x - \frac{(\tilde{p}_1, \tilde{p}_1)}{(\tilde{p}_0, \tilde{p}_0)}$$

$$= x^2 - \frac{5}{2} \underbrace{\int_{-1}^{1} x^5 dx}_{=0} - \frac{3}{5} = x^2 - \frac{3}{5}$$

$$(\tilde{p}_2, \tilde{p}_2) = \int_{-1}^{1} \left(x^6 - \frac{6}{5}x^4 + \frac{9}{25}x^2\right) dx = \frac{2}{7} - \frac{6}{5}\frac{2}{5} + \frac{9}{25}\frac{2}{3} = \frac{8}{175}$$

$$\Rightarrow p_2(x) = \frac{5}{2}\sqrt{\frac{7}{2}}\left(x^2 - \frac{3}{5}\right) = \frac{1}{2}\sqrt{\frac{7}{2}}\left(5x^2 - 3\right)$$

p_0, p_1, p_2 bilden eine ONB von Π_2

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 6+7 , Abgabe: 17.06.2004, 8.00 Uhr

Aufgabe 20: (4 Punkte)
Berechnen Sie das Integral

$$I = \int_0^{2\pi} \sqrt{1 - k^2 \cos^2(\varphi)} d\varphi$$

mit $k^2 = 0{:}84$. ($I = 4.60262252$)

(a) nach der zusammengesetzten Trapezregel mit $h = \pi/8$,

(b) nach der zusammengesetzten Simpsonregel mit $h = \pi/8$.

Hinweis: Betrachten Sie eine geeignete Transformation von $[0, 2\pi]$ auf $[0, \frac{\pi}{2}]$.

Aufgabe 21: (4 Punkte)
Es sei

$$\int_0^1 \frac{4}{1+x^2} dx = \pi = 3.14159\ldots .$$

Bestimmen Sie π näherungsweise, indem Sie für das angegebene Integral 3 Romberg-Schritte ausführen (Schrittweiten $h_i = \frac{1}{2^i}, i = 0, 1, 2, 3$).

Aufgabe 22: (4 Punkte)
Die Ableitung $f'(x)$ einer Funktion $f(x)$ soll durch die Differenzenquotienten

(a) $T(h) = \frac{1}{h}(f(x+h) - f(x))$

(b) $T(h) = \frac{1}{2h}(f(x+h) - f(x-h))$

berechnet werden. Bestimmen Sie $f'(1)$ nach (a) und (b) für $f(x) = e^x$ mittels Extrapolation zu den Schrittweiten $h_0 = 0.2, h_1 = 0.1, h_2 = 0.05$.

Aufgabe 23: (4 Punkte)
Sei $f \in C^4[a, b]$ und $T(h)$ die zusammengesetzte Trapezregel mit der Schrittweite $h = \frac{b-a}{6m}$. Konstruieren Sie aus $T(2h)$ und $T(3h)$ eine Integrationsformel $\hat{T}(h)$, so dass

$$\hat{T}(h) - \int_a^b f(x) dx = O(h^4)$$

gilt.

da $\cos(\varphi+\pi) = -\cos\varphi$
$\Rightarrow \cos^2(\varphi+\pi) = \cos^2\varphi$

$$= 2\int_0^{\pi} \sqrt{1-k^2\cos^2\varphi}\, d\varphi$$

$$= 2\int_0^{\frac{\pi}{2}} \sqrt{1-k^2\cos^2\varphi}\, d\varphi + 2\int_{\frac{\pi}{2}}^{\pi} \sqrt{1-k^2\cos^2\varphi}\, d\varphi \qquad \varphi' = \pi - \varphi$$

$$= 2\int_0^{\frac{\pi}{2}} \sqrt{1-k^2\cos^2\varphi}\, d\varphi + 2\int_0^{\frac{\pi}{2}} \sqrt{1-k^2\cos^2(\pi-\varphi)}\, d\varphi'$$

$\cos(\pi-\varphi) = -\cos\varphi$
$\Rightarrow \cos^2(\pi-\varphi) = \cos^2\varphi$

$$= 4\int_0^{\frac{\pi}{2}} \sqrt{1-k^2\cos^2\varphi}\, d\varphi$$

a) $h = \frac{\pi}{8} = \frac{1}{4}\,\frac{\pi}{2} \quad (\Rightarrow n = 4)$

$$T(h) = 4\cdot\frac{\pi}{8}\left(\frac{1}{2}f(0) + f\left(\frac{\pi}{8}\right) + f\left(\frac{\pi}{4}\right) + f\left(\frac{3}{8}\pi\right) + \frac{1}{2}f\left(\frac{\pi}{2}\right)\right)$$

$$\approx 4{,}602501862$$

b) $$S(h) = 4\cdot\frac{\pi}{8}\cdot\frac{1}{3}\left(f(0) + 4f\left(\frac{\pi}{8}\right) + f\left(\frac{\pi}{4}\right) + f\left(\frac{\pi}{4}\right) + 4f\left(\frac{3}{8}\pi\right) + f\left(\frac{\pi}{2}\right)\right)$$

$$= \frac{\pi}{6}\left(f(0) + 4f\left(\frac{\pi}{8}\right) + 2f\left(\frac{\pi}{4}\right) + 4f\left(\frac{3}{8}\pi\right) + f\left(\frac{\pi}{2}\right)\right)$$

$$\approx 4{,}606\,108\,968$$

$T(h_0) = \frac{1}{h_0}\left(e^{1+h_0} - e\right) = 5\left(e^{\frac{6}{5}} - e\right) = 3,009\,175\,48$

$T(h_1) = \frac{1}{h_1}\left(e^{1+h_1} - e^1\right) = 10\left(e^{\frac{11}{10}} - e\right) = 2,858\,841\,96$

$T(h_2) = \frac{1}{h_2}\left(e^{1+h_2} - e^1\right) = 20\left(e^{\frac{21}{20}} - e\right) = 2,787\,385\,79$

i	h_i	$T_{i0} = T(h_i)$	T_{i1}	T_{i2}
0	$\frac{1}{5}$	3,009175 48		
1	$\frac{1}{10}$	2,858 841 96	2,708 508 44	
2	$\frac{1}{20}$	2,787 385 79	2,715 92962	2,718 403 33

mit $T_{i,k} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left(\frac{h_{i-k}}{h_i}\right)^{\gamma} - 1} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{2^k - 1}$

da $\gamma = 1$ und $h_i = \frac{1}{5 \cdot 2^i}$

i) $T(h_0) = \frac{1}{h_0}\left(e^{1+h_0} - e^{1-h_0}\right) = \frac{5}{2}\left(e^{\frac{6}{5}} - e^{\frac{4}{5}}\right) \approx 2,736\,439\,99$

$T(h_1) = \frac{1}{h_1}\left(e^{1+h_1} - e^{1-h_1}\right) = 5\left(e^{\frac{11}{10}} - e^{\frac{9}{10}}\right) \approx 2,722\,814\,56$

$T(h_2) = \frac{1}{h_2}\left(e^{1+h_2} - e^{1-h_2}\right) = 10\left(e^{\frac{21}{20}} - e^{\frac{19}{20}}\right) \approx 2,719\,414\,59$

Rekursionsformel: $\gamma = 2$, $h_i = \frac{1}{5 \cdot 2^i}$

$\Rightarrow T_{i,k} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{2^{2k} - 1}$

i	h_i	$T_{i0} = T(h_0)$	T_{i1}	T_{i2}
0	$\frac{1}{5}$	2,736 439 99		
1	$\frac{1}{10}$	2,722 814 56	2,718 272 76	
2	$\frac{1}{20}$	2,719 414 59	2,718 281 27	2,718 281 83

a) $T(h) = (1+h)^h = e^{\frac{1}{h}\log(1+h)}$ für $|h|<1$, $h \neq 0$

Sei $\varphi(h) = \frac{1}{h}\log(1+h)$

$$= \frac{1}{h}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} h^k = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} h^{k-1}$$

$$= \sum_{k=0}^{\infty} \underbrace{\frac{(-1)^k}{k+1}}_{=:a_k} h^k$$

Die Reihe $\varphi(h) = \sum_{k=0}^{\infty} a_k h^k$ konvergiert für $|h|<1$ absolut, da sie den Konvergenzradius $R = \lim_{k\to\infty} \left|\frac{a_k}{a_{k+1}}\right| = \lim_{k\to\infty} \frac{k+2}{k+1} = 1$ hat.

$\varphi \in \mathcal{C}^\infty(]-1,1[)$
$T \in \mathcal{C}^\infty(]-1,1[)$

Es gilt: $T(h) = e^{\varphi(h)} = \sum_{n=0}^{\infty} \frac{\varphi(h)^n}{n!}$

$T(h) = \sum_{i=0}^{\infty} \frac{T^{(i)}(0)}{i!}(h-0)^i$

Berechne $\varphi(h)^n$:

Da $\sum_{k=0}^{\infty} a_k h^k$ abs. konv. gilt (für $|h|<1$): Cauchy-Produkt: $\left(\sum_{k=0}^{\infty} a_k x^k\right)\left(\sum_{k=0}^{\infty} b_k x^k\right) = \sum_{k=0}^{\infty} c_k x^k$, $c_k = \sum_{j=0}^{k} a_j b_{k-j}$

$T(0) = e^{\varphi(0)} = e$

$\Rightarrow T(h) = e + \sum_{i=1}^{\infty} \frac{T^{(i)}(0)}{i!}\varphi(h)^i = \left(\sum_{k=0}^{\infty} a_k h^k\right)^n = \sum_{k=0}^{\infty} b_{kn} h^k$

$T'(h) = \varphi'(h)\,T(h)$

$T''(h) = \varphi''(h)\,T(h) + \varphi'(h)^2 T(h)$

mit $b_{kn} = \sum_{c \in \mathbb{N}_0^n,\ |c| = \sum |c_i| = k} \prod_{j=1}^{n} a_{c_j}$, $b_{0n} = 1$

$$\Rightarrow T(h) = \sum_{n=0}^{\infty} \frac{\varphi(h)^n}{n!} = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \frac{b_{kn}}{n!} h^k$$

$$\overset{(*)}{=} \sum_{k=0}^{\infty} \underbrace{\sum_{n=0}^{\infty} \frac{b_{kn}}{n!}}_{=:d_k} h^k \quad \text{für } |h|<1$$

(*) ↑ Begründung s.u.

$\sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \frac{b_{kn}}{n!} h^k$ absolut konvergiert.

Es gilt: $\varphi(h) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} h^k = \sum_{k=0}^{\infty} a_k h^k$

$\varphi(-h) = \sum_{k=0}^{\infty} \frac{1}{k+1} h^k = \sum_{k=0}^{\infty} |a_k| h^k$

$\Rightarrow \max(\varphi(h), \varphi(-h)) = \sum_{k=0}^{\infty} |a_k h^k|$

$\Rightarrow \left(\max(\varphi(h), \varphi(-h))\right)^n = \left(\sum_{k=0}^{\infty} |a_k h^k|\right)^n = \left(\sum_{k=0}^{\infty} |a_k| |h|^k\right)^n$

$= \sum_{k=0}^{\infty} \underbrace{\sum_{c \in \mathbb{N}_0^n, |c|=k} \prod_{j=1}^{n} |a_{c_j}| |h|^k}_{\geq |b_{kn}|} \geq \sum_{k=0}^{\infty} |b_{kn}| |h|^k$

$\Rightarrow \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} \left|\frac{b_{kn}}{n!} h^k\right| \leq \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{\infty} |b_{kn}| |h|^k$

$\leq \sum_{n=0}^{\infty} \frac{1}{n!} \left(\max(\varphi(h), \varphi(-h))\right)^n$

$= e^{\max(\varphi(h), \varphi(-h))}$

d.h. die Doppelreihe konv. absolut!

b) $T(-h) = e + \frac{1}{2} eh + \frac{11}{24} eh^2 + \sum_{k=3}^{\infty} a_k (-h)^k$

$\Rightarrow \tilde{T}(h) = e + \frac{11}{24} h^2 + \sum_{k=2}^{\infty} d_k h^{2k}$

d.h. die ungeraden Potenzen von h fallen weg und $\tilde{T}(h)$ ist eine Potenzreihe in h^2

$\Rightarrow$ schnellere Konvergenz

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 8 , Abgabe: 24.06.2004, 8.00 Uhr

Aufgabe 26: (4 Punkte)
Das Integral

$$I(f) = \int_{-1}^{1} f(x)\,dx\,, \qquad f \in C[-1,1]\,,$$

soll approximiert werden durch den Ausdruck

$$A(f) = A_1 f(-1) + A_2 f(x_1) + A_3 f(1)\,.$$

Bestimmen Sie A_1, A_2, A_3 und $x_1 \in (-1, 1)$, so dass $I(P) = A(P)$ für alle Polynome P vom Grad ≤ 3 gilt. Zeigen Sie, dass $A(f)$ nicht für alle Polynome vom Grad ≤ 4 exakt ist.
Hinweis: Betrachten Sie die Basispolynome.

Aufgabe 27: (4 Punkte)

(a) Die Orthogonalpolynome in $C[a, b]$ mit $[a, b] = [0, \infty)$ und dem Gewicht $w(x) = e^{-x}$ heissen LAGUERRE-Polynome $L_n(x)$. Bestimmen Sie $L_0(x)$, $L_1(x)$, $L_2(x)$ mit dem SCHMIDT'schen Orthogonalisierungsverfahren.

(b) Bestimmen Sie die Stützpunkte x_i und die Gewichte A_i für das GAUSS'sche Integrationsverfahren

$$\int_0^{\infty} e^{-x} f(x)dx \approx A_1 f(x_1) + A_2 f(x_2)\,.$$

Aufgabe 28: (4 Punkte)
Berechnen Sie für $n = 2, 3$ das Integral

$$If = \int_{-1}^{1} \frac{1}{1+x^2}dx$$

mit der Formel von Gauß-Legendre mit der Gewichtsfunktion $\omega(x) = 1$. Vergleichen Sie das Ergebnis mit der exakten Lösung.

Es gilt: $I(p) = A(p) \quad \forall p \in \Pi_3$

$\Leftrightarrow I(p_i) = A(p_i) \quad i = 0,\dots,3$

Es ist: $I(p_0) = \int_{-1}^{1} dx = 2 \qquad A(p_0) = A_0 + A_1 + A_2$

$I(p_1) = \int_{-1}^{1} x\,dx = 0 \qquad A(p_1) = -A_0 + A_1 x_1 + A_2$

$I(p_2) = \int_{-1}^{1} x^2\,dx = \frac{2}{3} \qquad A(p_2) = A_0 + A_1 x_1^2 + A_2$

$I(p_3) = \int_{-1}^{1} x^3\,dx = 0 \qquad A(p_3) = -A_0 + A_1 x_1^3 + A_2$

Wir erhalten das Gleichungssystem:

$$\begin{aligned}
&\text{I} && A_0 + A_1 + A_2 = 2\\
&\text{II} && -A_0 + A_1 x_1 + A_2 = 0\\
&\text{III} && A_0 + A_1 x_1^2 + A_2 = \tfrac{2}{3}\\
&\text{IV} && -A_0 + A_1 x_1^3 + A_2 = 0
\end{aligned}$$

III − I: $A_1 x_1^2 - A_1 = -\frac{4}{3} \Leftrightarrow A_1 \underbrace{(x_1^2 - 1)}_{<0} = -\frac{4}{3}$ ~~$\Rightarrow A_1 = \frac{4}{3}\frac{1}{1-x_1^2} \neq 0$~~

IV − II: $A_1 x_1^3 - A_1 x_1 = 0 \Leftrightarrow x_1 A_1 \underbrace{(x_1^2 - 1)}_{\neq 0} = 0$ ~~$\Rightarrow x_1 = 0$, da $A_1 \neq 0$~~

$\underbrace{\quad = -\frac{4}{3} \quad}$

$\Rightarrow x_1 = 0$

$\Rightarrow A_1 = \frac{4}{3}$

Somit folgt aus II Sew. IV: $A_0 = A_2$

" I Sew III: $A_0 + A_2 = \frac{2}{3}$

$\Rightarrow A_0 = A_2 = \frac{1}{3}$

$\Rightarrow A(f) = \frac{1}{3}\left(f(-1) + 4f(0) + f(1)\right)$

Sei $p(x) = \underbrace{(x-1)}_{\leq 0}\underbrace{(x-x_1)^2}_{\geq 0}\underbrace{(x+1)}_{\geq 0} \leq 0$ für $x \in [-1, 1]$

$\Rightarrow I(p) < 0$ aber $A(p) = 0$

$\Rightarrow A$ ist nicht für alle Polynome aus Π_4 exakt.

$L_0(x) = 1$, $L_1(x) = x$, $L_2(x) = x^2 - \frac{1}{3}$, $L_3(x) = x^3 - \frac{3}{5}x = x\left(x^2 - \frac{3}{5}\right)$

$n=2$: $x_1 = -\frac{1}{\sqrt{3}}$, $x_2 = \frac{1}{\sqrt{3}}$

$$L_1(x) = \frac{x-x_2}{x_1-x_2} = \frac{x-\frac{1}{\sqrt{3}}}{-\frac{2}{\sqrt{3}}} = -\frac{\sqrt{3}}{2}x + \frac{1}{2}$$

$$L_2(x) = \frac{x-x_1}{x_2-x_1} = \frac{x+\frac{1}{\sqrt{3}}}{\frac{2}{\sqrt{3}}} = \frac{\sqrt{3}}{2}x + \frac{1}{2}$$

$$A_1 = \int_{-1}^{1} \left(-\frac{\sqrt{3}}{2}x + \frac{1}{2}\right) dx = 1$$

$$A_2 = \int_{-1}^{1} \left(\frac{\sqrt{3}}{2}x + \frac{1}{2}\right) dx = 1$$

$$\Rightarrow G_2(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = \frac{1}{1+\frac{1}{3}} + \frac{1}{1+\frac{1}{3}} = \frac{3}{2}$$

$f_{rel} = 0{,}045$

$n=3$: $x_1 = -\sqrt{\frac{3}{5}}$, $x_2 = 0$, $x_3 = \sqrt{\frac{3}{5}}$

$$L_1(x) = \frac{x-x_2}{x_1-x_2}\,\frac{x-x_3}{x_1-x_3} = \frac{x}{-\sqrt{\frac{3}{5}}}\,\frac{x-\sqrt{\frac{3}{5}}}{-2\sqrt{\frac{3}{5}}} = \frac{5}{6}x\left(x-\sqrt{\frac{3}{5}}\right)$$

$$L_2(x) = \frac{x-x_1}{x_2-x_1}\,\frac{x-x_3}{x_2-x_3} = \frac{x+\sqrt{\frac{3}{5}}}{\sqrt{\frac{3}{5}}}\,\frac{x-\sqrt{\frac{3}{5}}}{-\sqrt{\frac{3}{5}}} = -\frac{5}{3}\left(x^2-\frac{3}{5}\right) = -\frac{5}{3}x^2+1$$

$$L_3(x) = \frac{x-x_1}{x_3-x_1}\,\frac{x-x_2}{x_3-x_2} = \frac{x+\sqrt{\frac{3}{5}}}{2\sqrt{\frac{3}{5}}}\,\frac{x}{\sqrt{\frac{3}{5}}} = \frac{5}{6}x\left(x+\sqrt{\frac{3}{5}}\right)$$

$$A_1 = \int_{-1}^{1} L_1(x)\,dx = \frac{5}{6}\int_{-1}^{1}\left(x^2-\sqrt{\frac{3}{5}}x\right)dx = \frac{5}{9}$$

$$A_2 = \int_{-1}^{1} L_2(x)\,dx = \int_{-1}^{1}\left(-\frac{5}{3}x^2+1\right)dx = -\frac{10}{9}+2 = \frac{8}{9}$$

$$A_3 = \int_{-1}^{1} L_3(x)\,dx = \frac{5}{6}\int_{-1}^{1}\left(x^2+\sqrt{\frac{3}{5}}x\right)dx = \frac{5}{9}$$

$$\Rightarrow G_3(f) = \frac{1}{9}\left(5f\left(-\sqrt{\tfrac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\tfrac{3}{5}}\right)\right)$$

$$= \frac{1}{9}\left(5\cdot\frac{1}{1+\frac{3}{5}} + 8 + 5\cdot\frac{1}{1+\frac{3}{5}}\right) = \frac{1}{9}\left(\frac{25}{4}+8\right) = \frac{1}{9}\cdot\frac{57}{4} = \frac{19}{12}$$

$f_{rel} = 0{,}008$ $\approx 1{,}58333$ (exakt: $\int_{-1}^{1}\frac{1}{1+x^2}\,dx = \arctan x\Big|_{-1}^{1} = \frac{\pi}{2} \approx 1{,}57079$)

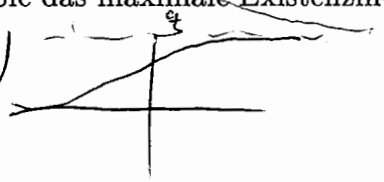
Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 9 , Abgabe: 01.07.2004, 8.00 Uhr

Aufgabe 30: (3 Punkte)
Lösen Sie die *logistische Differentialgleichung*

$$y' = ay - by^2, \quad y(0) = y_0, \qquad (a, b, y_0 > 0)$$

mittels Trennung der Variablen (s. Aufgabe 29) und geben Sie das maximale Existenzintervall an.
Hinweis: Partialbruchzerlegung



Aufgabe 31: (1+2+2 Punkte)
Seien $a, b\colon \mathbb{R} \to \mathbb{R}$ stetig.

(a) Bestimmen Sie die Lösungen der linearen, skalaren und homogenen Differentialgleichung 1. Ordnung

$$y' = a(x)y$$

mittels Trennung der Variablen (s. Aufgabe 29).

(b) Berechnen Sie die Lösungen der inhomogenen Differentialgleichung

$$y' = a(x)y + b(x).$$

Machen Sie den Ansatz $y(x) = c(x)y_h(x)$, wobei $cy_h(x)$ die Lösung der homogenen Differentialgleichung ist (Variation der Konstanten).

(c) Lösen Sie die Anfangswertaufgaben

$$(i)\ y' = 2xy, \quad y(0) = 2 \qquad (ii)\ y' = 2xy + x^3, \quad y(0) = 2$$

Aufgabe 32: (1+1+1+3 Punkte)
Gegeben sei die Differentialgleichung

$$f(x, y) + g(x, y)\frac{dy}{dx} = 0 \tag{1}$$

bzw. in symmetrischer Darstellung

$$f(x, y)dx + g(x, y)dy = 0 \tag{2}$$

mit $f, g \in C^1(D)$, $D \subset \mathbb{R}^2$ Gebiet.

(a) Sei $x = x(t), y = y(x(t)) = y(t)$ eine Parameter-Darstellung mit $\dot{x}(t_0) \neq 0$. Zeigen Sie, dass in einer Umgebung von t_0 gilt:

$$x, y \text{ Lösung von (1)} \iff f(x, y)\dot{x} + g(x, y)\dot{y} = 0$$

Informatik: anmelden für Klausur (Frist 10 Tage vor Klausur)

$$\int_{y_0}^{y} \frac{d\tilde{y}}{a\tilde{y}-b\tilde{y}^2} = \int_0^x d\tilde{x} = x \qquad \left(ay_0 - by_0^2 \neq 0 \text{ da } y_0 \neq \frac{a}{b}\right)$$

$$\int_{y_0}^{y} \frac{d\tilde{y}}{a\tilde{y}-b\tilde{y}^2} = \int_{y_0}^{y} \frac{1}{\tilde{y}(a-b\tilde{y})}\, d\tilde{y} = \frac{1}{a}\int_{y_0}^{y} \left(\frac{1}{\tilde{y}} + \frac{b}{a-b\tilde{y}}\right) d\tilde{y}$$

$$= \frac{1}{a}\left[\log|\tilde{y}| - \log|a-b\tilde{y}|\right]_{y_0}^{y}$$

$$= \frac{1}{a}\left[\log\left|\frac{\tilde{y}}{a-b\tilde{y}}\right|\right]_{y_0}^{y} = \frac{1}{a}\left(\log\left|\frac{y}{a-by}\right| - \log\left|\frac{y_0}{a-by_0}\right|\right)$$

$$= \frac{1}{a}\log\left|\frac{y(a-by_0)}{y_0(a-by)}\right|$$

$$\Rightarrow x = \frac{1}{a}\log\left|\frac{y(a-by_0)}{y_0(a-by)}\right|$$

$$\Rightarrow \left|\frac{y(a-by_0)}{y_0(a-by)}\right| = e^{ax}$$

Wir können annehmen, dass y und y_0 dasselbe Vorzeichen haben, denn es gilt $y(0)=y_0$ und y hat keine Nullstellen, da aus $y(x_0)=0$ $y\equiv 0$ folgt. ($y\equiv 0$ erfüllt die AWA $y' = ay - by^2$, $y(x_0)=0$). Das wäre ein Widerspruch zu $y_0 > 0$. Mit demselben Argument folgt, dass $a-by_0$ und $a-by$ dasselbe Vorzeichen haben, da wir $y_0 \neq \frac{a}{b}$ angenommen haben.

$$\Rightarrow y(a-by_0) = y_0(a-by)e^{ax}$$

$$\Rightarrow y(a-by_0+by_0e^{ax}) = y_0 a e^{ax}$$

$$\Rightarrow y(x) = \frac{ay_0}{(a-by_0)e^{-ax}+by_0} = \frac{a}{b}\,\frac{y_0}{\left(\frac{a}{b}-y_0\right)e^{-ax}+y_0} = \frac{a}{b}\,\frac{1}{\left(\frac{a}{by_0}-1\right)e^{-ax}+1}$$

Diese Lösung stimmt für $y_0 = \frac{a}{b}$ nicht, denn dann gilt offensichtlich $y(x) = \frac{a}{b} = y_0$.

32) a) „⇒" Sei $x(t), y(t)$ Lösung. Es gelte $\dot{y}(t) = \frac{d}{dt} y(x(t)) = \frac{dy}{dx} \dot{x}(t)$

$$\Rightarrow f(x,y)\dot{x} + g(x,y)\dot{y} = f(x,y)\dot{x} + g(x,y)\frac{dy}{dx}\dot{x}$$

$$= \dot{x}\left(f(x,y) + g(x,y)\frac{dy}{dx}\right) = 0$$

„⇐" Es gelte $f(x,y)\dot{x} + g(x,y)\dot{y} = 0$

$$\overset{s.l.}{\Rightarrow} \dot{x}\left(f(x,y) + g(x,y)\frac{dy}{dx}\right) = 0$$

$\dot{x}(t_0) \neq 0 \Rightarrow \dot{x}(t) \neq 0$ in einer Umgebung $\tilde{U}$ von t_0

$$\Rightarrow f(x,y) + g(x,y)\frac{dy}{dx} = 0 \text{ in } U$$

b) ~~Sei $y(x)$ Lösung von (1)~~

$$\frac{d}{dx} H(x, y(x)) = H_x(x, y(x)) + H_y(x, y(x))\frac{dy}{dx}$$

$$\overset{\text{Voraussetzung}}{=} f(x,y) + g(x,y)\frac{dy}{dx} = 0 \quad \text{~~y Lösung~~}$$

$\Rightarrow H(x, y(x)) = \text{const} \Leftrightarrow y(x)$ Lösung von (1)

c) $\dot{x}(t) \neq 0$ für $t \in \mathcal{I} \Rightarrow x(t)$ läßt sich in I nach $t(x)$ auflösen

und es gilt $\frac{d}{dx} t(x)\Big|_{x=x(t)} = \left(\frac{d}{dt} x(t)\right)^{-1} = \frac{1}{\dot{x}(t)}$ (Abl. der Umkehrfkt.)

~~$f(x,y)$~~ Außerdem gilt: $\frac{d}{dx} y(x) = \frac{d}{dx} y(t(x)) = \dot{y}(t(x))\frac{d}{dx} t(x)$

$$\Rightarrow f(x,y)\frac{dy}{dx} - g(x,y) = f(x,y)\dot{y}(t(x))\frac{d}{dx} t(x) - g(x,y)$$

$x = x(t), y = y(t)$

$$= \dot{x}(t)\dot{y}(t)\frac{d}{dx} t(x)\Big|_{x=x(t)} - \dot{y}(t)$$

$$= \dot{x}(t)\dot{y}(t)\frac{1}{\dot{x}(t)} - \dot{y}(t) = 0$$

32) d) $f(x,y) = ax - 5xy$

$g(x,y) = -cy + dxy$

$\Rightarrow$ $y(x) = y(t(x))$ erfüllt die DGL

$$f(x,y)\frac{dy}{dx} - g(x,y) = 0$$

$\Rightarrow$ int. Faktor

$$M(x,y)f(x,y)\frac{dy}{dx} - g(x,y)M(x,y) = 0$$

$\Rightarrow$

$$\left(\frac{a}{y} - 5\right)\frac{dy}{dy} + \frac{c}{x} - d = 0 \qquad (*)$$

$$\frac{d}{dx}\left(\frac{a}{y} - 5\right) = 0 \quad , \quad \frac{d}{dy}\left(\frac{c}{x} - d\right) = 0$$

$\Rightarrow$ $(*)$ ist exakt

$\Rightarrow$ $H_x(x,y) = \frac{c}{x} - d$, $H_y(x,y) = \frac{a}{y} - 5$

$$\Rightarrow H(x,y) = \int_x \left(\frac{c}{\tilde{x}} - d\right) d\tilde{x} + h_1(y) \quad \text{und} \quad H(x,y) = \int_y \left(\frac{a}{\tilde{y}} - 5\right) d\tilde{y} + h_2(x)$$

$$\Rightarrow H(x,y) = c\log|x| - dx + h_1(y)$$
$$H(x,y) = a\log|y| - 5y + h_2(x)$$
$$\Rightarrow H(x,y) = c\log|x| + a\log|y| - dx - 5y + e$$

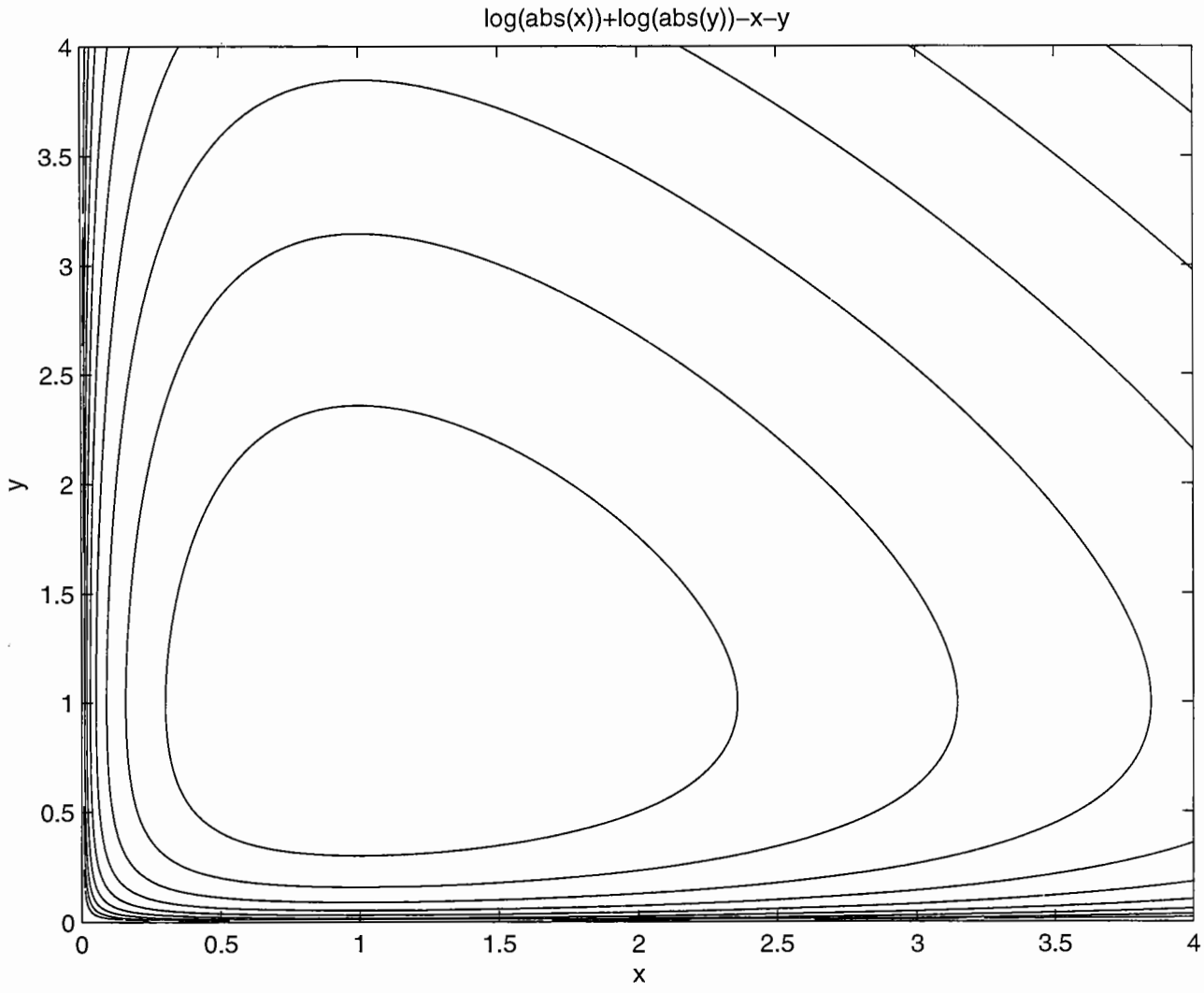
H ist als Stammfunktion nur bis auf eine Konstante bestimmt

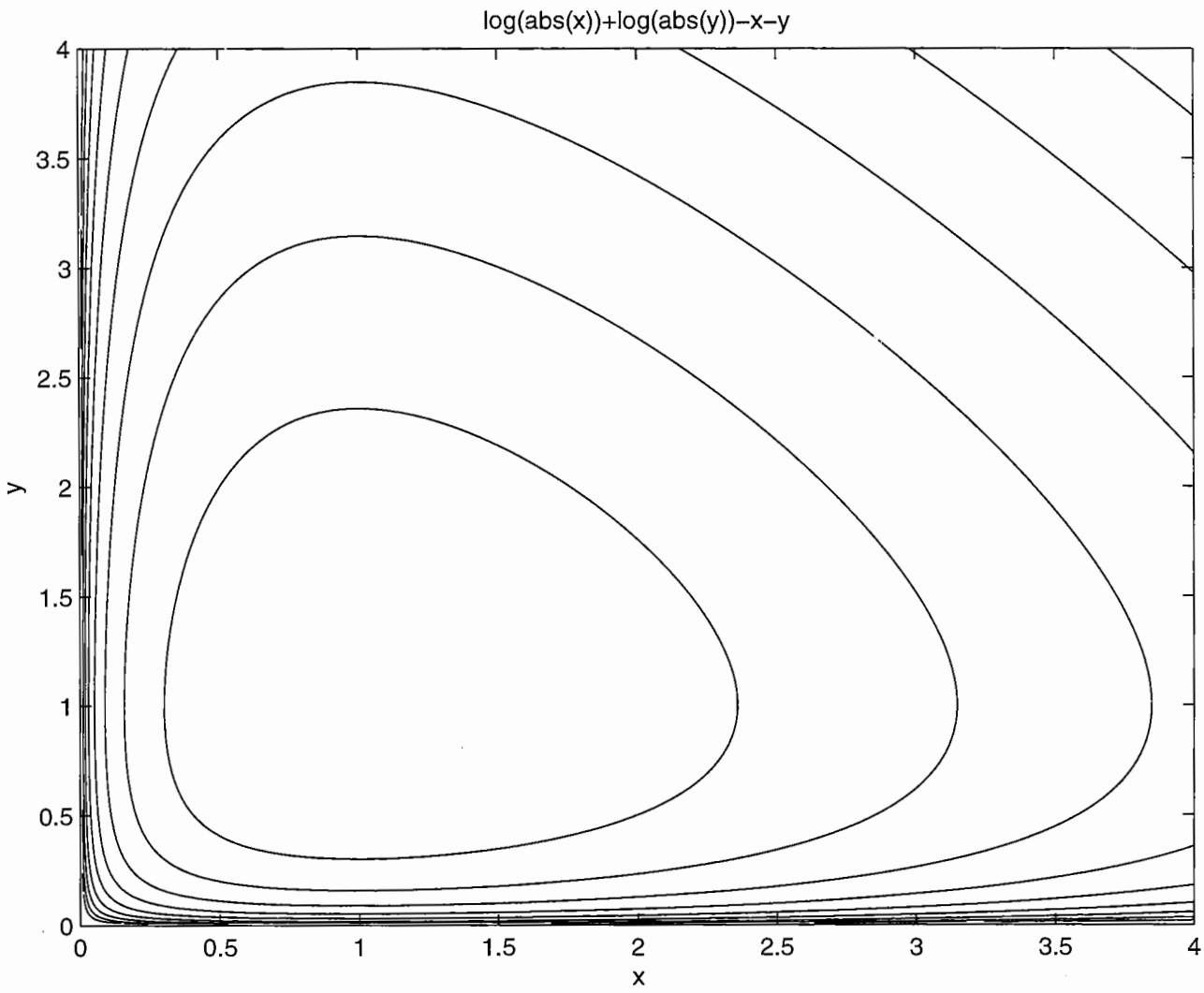
plotten mit Matlab: `ezcontour('H(x,y)',[0,4],1000)`

mit $a = 5 = c = d = 1$

Ausschnitt des 1. Quadranten reicht

Auflösung


log(abs(x))+log(abs(y))−x−y
y
x
0
0.5
1
1.5
2
2.5
3
3.5
4


log(abs(x))+log(abs(y))−x−y
4
3.5
3
2.5
2
1.5
1
0.5
0
y
0
0.5
1
1.5
2
2.5
3
3.5
4
x

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 10 , Abgabe: 08.07.2004, 8.00 Uhr

Aufgabe 34: (3+3+3 Punkte)

(a) Die stetige Funktion $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ genüge auf jedem Streifen $D_a := [-a, a] \times \mathbb{R}^n$, $a > 0$, einer Lipschitzbedingung bzgl. y mit einer von a abhängigen Lipschitzkonstanten $L_a \geq 0$. Zeigen Sie, dass die AWA

$$y' = f(x, y)\,, \qquad y(0) = y_0$$

für alle Anfangswerte $y_0 \in \mathbb{R}^n$ genau eine Lösung auf ganz $\mathbb{R}$ besitzt.
(Hinweis: Satz (26.9))

(b) Zeigen Sie, dass die AWA

$$y' = \frac{y^3 e^x}{1 + y^2} + x^2 \sin y\,, \qquad y(0) = y_0$$

genau eine Lösung auf $\mathbb{R}$ besitzt.

(c) Die Funktion $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ sei stetig differenzierbar und beschränkt:

$$\|f(x, y)\|_\infty \leq M \qquad \forall \quad (x, y) \in \mathbb{R} \times \mathbb{R}^n\,.$$

Zeigen Sie: Die Lösung der AWA

$$y' = f(x, y)\,, \qquad y(x_0) = y_0$$

existiert in ganz $\mathbb{R}$.

Aufgabe 35: (3+2 Punkte)

(a) Man zeige, dass die Lösung der folgenden AWA in dem angegebenen Intervall existiert:

$$y' = y + e^{-y} + e^{-x}\,, \qquad y(0) = 0\,, \quad 0 \leq x \leq 1\,.$$

(b) Bestimmen Sie das größte Existenzintervall $I = [0, a]$, $a > 0$, welches Satz (26.9) für die Lösung der AWA

$$y' = x^2 + y^2\,, \qquad y(0) = 0, \qquad \text{(RICCATI–DGL)}$$

voraussagt.

34) a) Sei $y_0 \in \mathbb{R}$ fest. Ferner sei $a \in \mathbb{R}$.

Für $(x,y) \in D_a$ gilt:

$$\|f(x,y)\| \le \|f(x,y) - f(x,y_0)\| + \|f(x,y_0)\|$$

$$\le L_a \|y - y_0\| + \|f(x,y_0)\|$$

$$\le L_a \|y\| + L_a \|y_0\| + \underbrace{\|f(x,y_0)\|}_{\text{beschränkt, da } f \text{ stetig} + x \in [-a,a]}$$

$$\le c_a$$

$$\le L_a \|y\| + c_a \qquad (*)$$

Sei nun $0 < \alpha < a$ so, dass $\alpha L_a < 1$ gilt.

Auf jedem Intervall $[x_n - \alpha, x_n + \alpha] \subset [-a,a]$ hat dann die AWA

$$y' = f(x,y), \quad y(x_n) = y_n \qquad (y_n \in \mathbb{R})$$

eine eindeutige Lösung:

Sei $Q_\beta = \{(x,y) : \|x - x_n\| \le \alpha, \|y - y_n\| \le \beta\}$

und $M_\beta = \max\{\|f(x,y)\| : (x,y) \in Q_\beta\}$

$$\le L_a (\|y_n\| + \beta) + c_a$$

$(*)$

Wir wählen β so groß, dass $\alpha M_\beta \le \beta$ gilt. Das ist möglich, denn es gilt: $\alpha M_\beta \le \beta \iff \alpha L_a \|y_n\| + \alpha L_a \beta + \alpha c_a \le \beta$

$$\iff \alpha(L_a \|y_n\| + c_a) \le \beta \underbrace{(1 - \alpha L_a)}_{>0}$$

Wir müssen also nur $\beta \ge \dfrac{\alpha(L_a\|y_n\| + c_a)}{1 - \alpha L_a}$ wählen.

Satz 26.9 liefert dann die Behauptung.

Aus $(y_1 \pm y_2)^2 \geq 0$ folgt $|y_1 y_2| \leq \frac{1}{2}(y_1^2 + y_2^2)$ und somit

$$|y_1 y_2 - 1| \leq 1 + |y_1 y_2| \leq 1 + \frac{1}{2}y_1^2 + \frac{1}{2}y_2^2 \leq 1 + y_1^2 + y_2^2 + y_1^2 y_2^2$$

$$\Rightarrow |f(x, y_1) - f(x, y_2)| \leq \underbrace{(2e^a + a^2)}_{=: L_a} |y_1 - y_2|$$

c) ~~Sei Da wie in a)~~

Sei $a > 0$ und $b = Ma$.

Weiter sei $Q = \{(x, y) : \|x - x_0\|_\infty \leq a, \|y - y_0\|_\infty \leq b\}$

Da Q kompakt ist, ist die part. Abl. von f beschränkt.

Q konvex $\Rightarrow$ f genügt in Q einer Lipschitz-Bedingung bzgl. y ~~(sogar)~~ (Satz 26.8 (i))

$\overset{(26.9)}{\Rightarrow}$ Die AWA besitzt auf $[x_0 - a, x_0 + a]$ genau eine Lösung

$\overset{\text{wie in a)}}{\Rightarrow}$ Die AWA besitzt genau eine Lösung auf $\mathbb{R}$.

b) Sei $(x,y) \in Q = \{(x,y) : \|x\| \le a, \|y\| \le b\}$

$\Rightarrow M = \max\{\|f(x,y)\| : (x,y) \in Q\} = a^2 + b^2$

$\Rightarrow$ Lösung existiert auf $[0, \alpha]$, $\alpha = \min\left(a, \frac{b}{M}\right) = \min\left(a, \frac{b}{a^2+b^2}\right)$

Bestimme zu festem $a \in \mathbb{R}$, so dass ~~$\min_{b>0}$~~ $f_a(b) := \frac{b}{a^2+b^2}$ max. wird:

$f_a(0) = 0$, $\lim_{b\to\infty} f_a(b) = 0$, $f_a'(b) = \frac{a^2-b^2}{(a^2+b^2)^2}$

$\Rightarrow f_a(b)$ hat Maximum in $b = a$ mit $f_a(a) = \frac{1}{2a}$

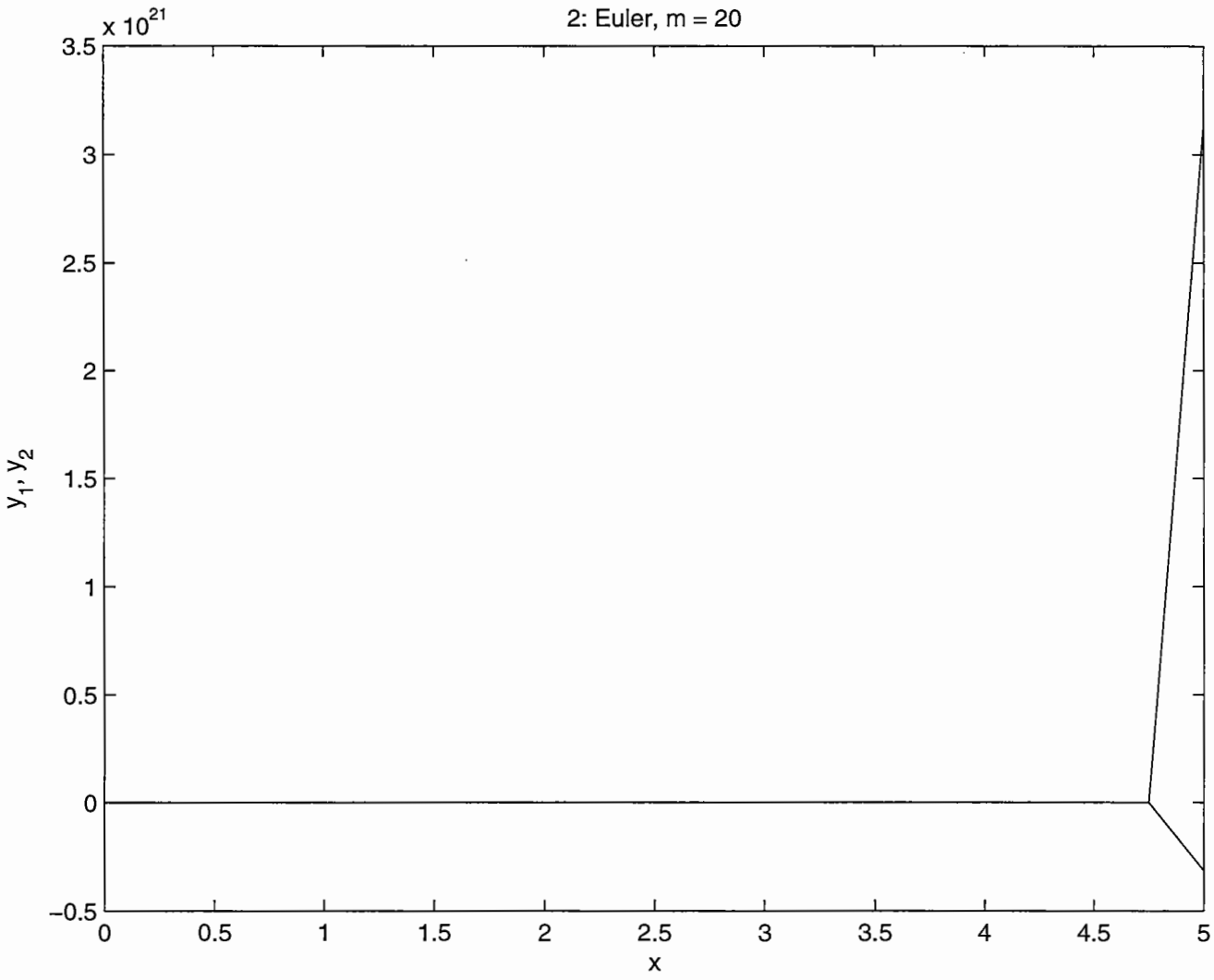
$\Rightarrow h(a) = \max_{b>0} \min(a, f_a(b)) = \min(a, \max_{b>0} f_a(b))$

$= \min\left(a, \frac{1}{2a}\right) = \begin{cases} \frac{1}{2a} & \text{falls } a \ge \frac{1}{\sqrt{2}} \\ a & \text{falls } a < \frac{1}{\sqrt{2}} \end{cases}$

(da $a = \frac{1}{2a} \Leftrightarrow a = \frac{1}{\sqrt{2}}$ für $a > 0$)

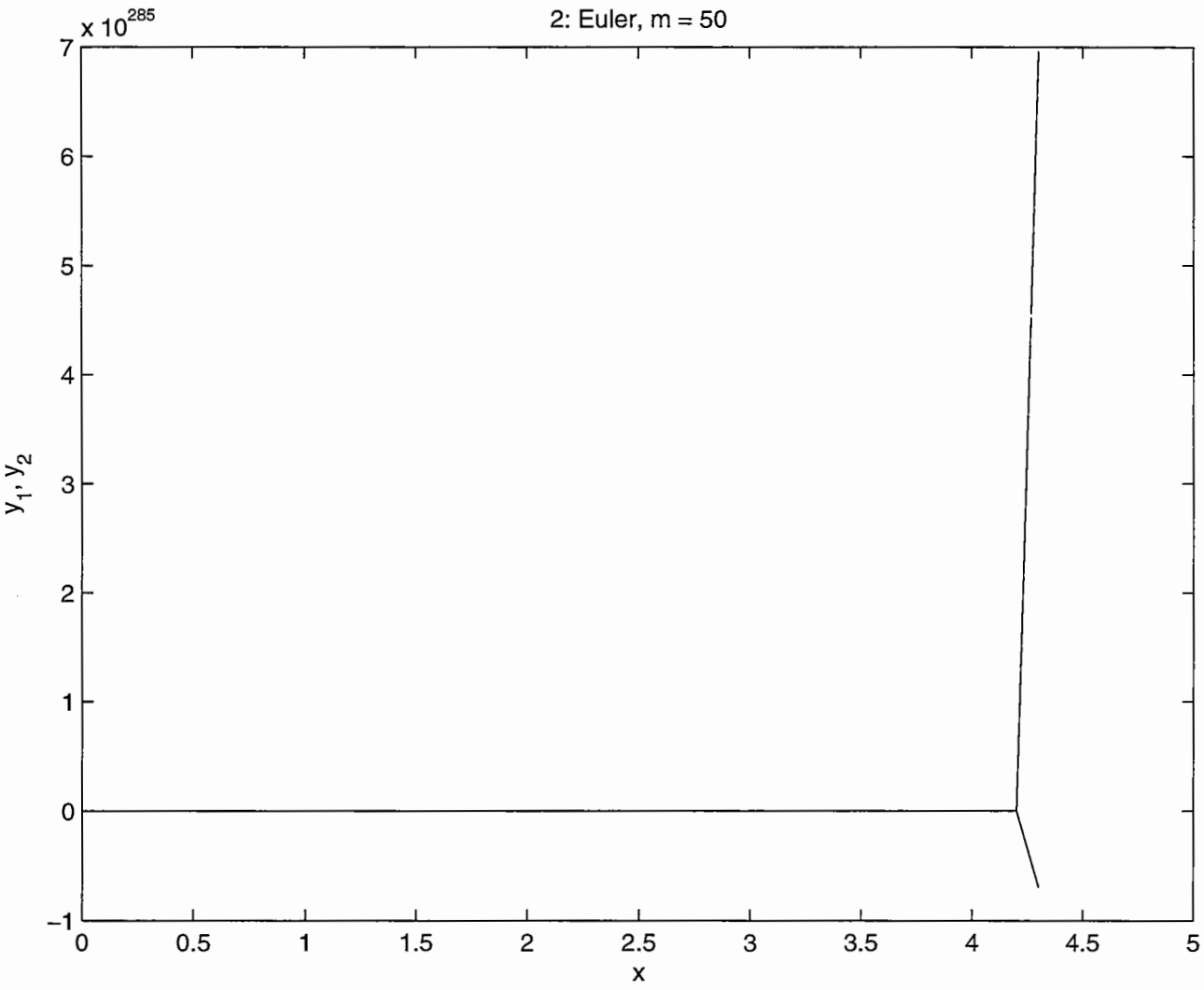
$\Rightarrow \max_{a>0} h(a) = \frac{1}{\sqrt{2}}$

$\Rightarrow$ größte Existenzintervall $I = [0, \frac{1}{\sqrt{2}}]$



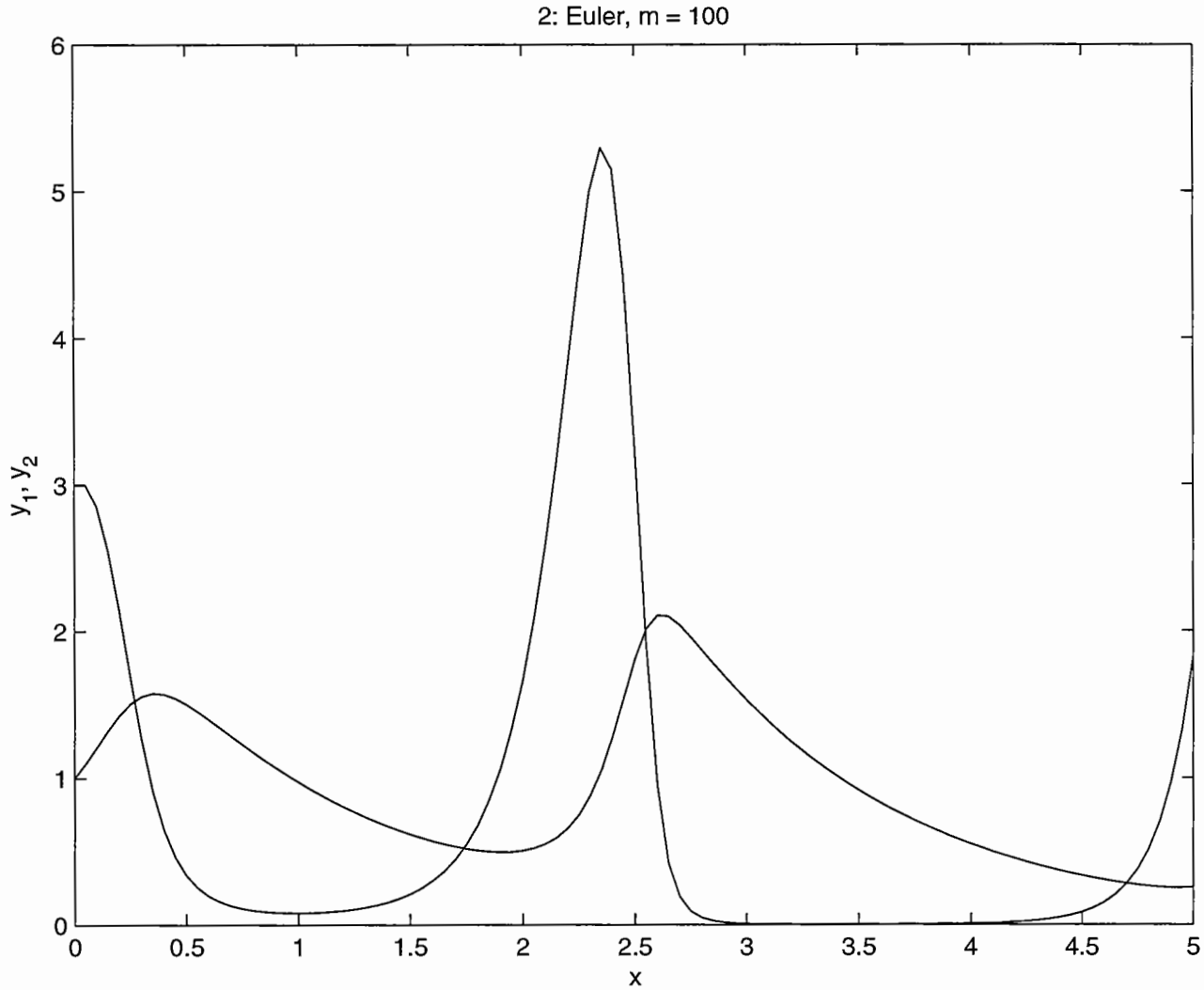
$y_1(x_f) = 3{,}14224149 \cdot 10^{21}$

$y_2(x_f) = -3{,}14224149 \cdot 10^{20}$



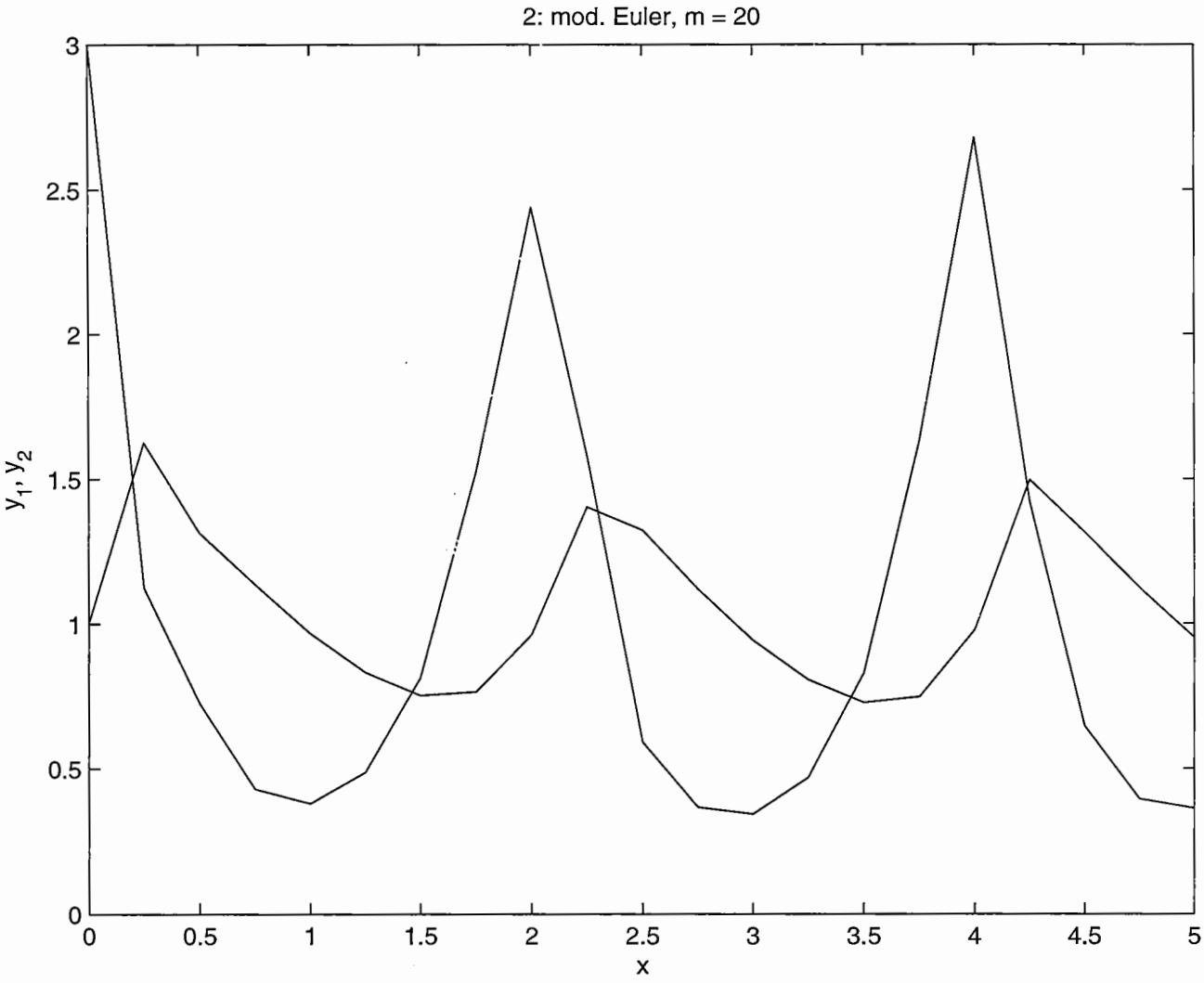
$Y_1(x_0) = \inf$

$Y_2(x_f) = -\inf$



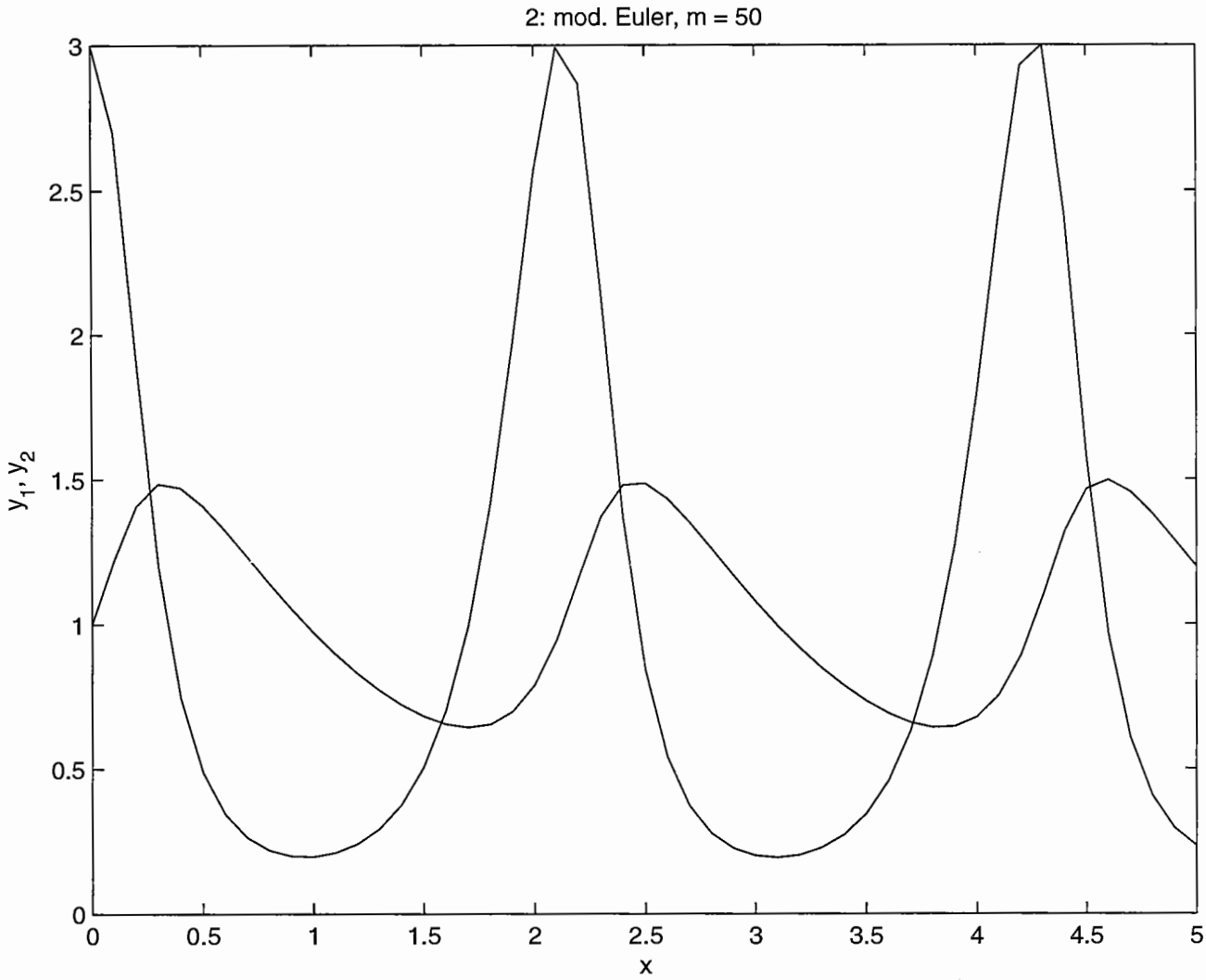
$y_1(x_f) = 1,82699729$

$y_2(x_f) = 0,25499046$



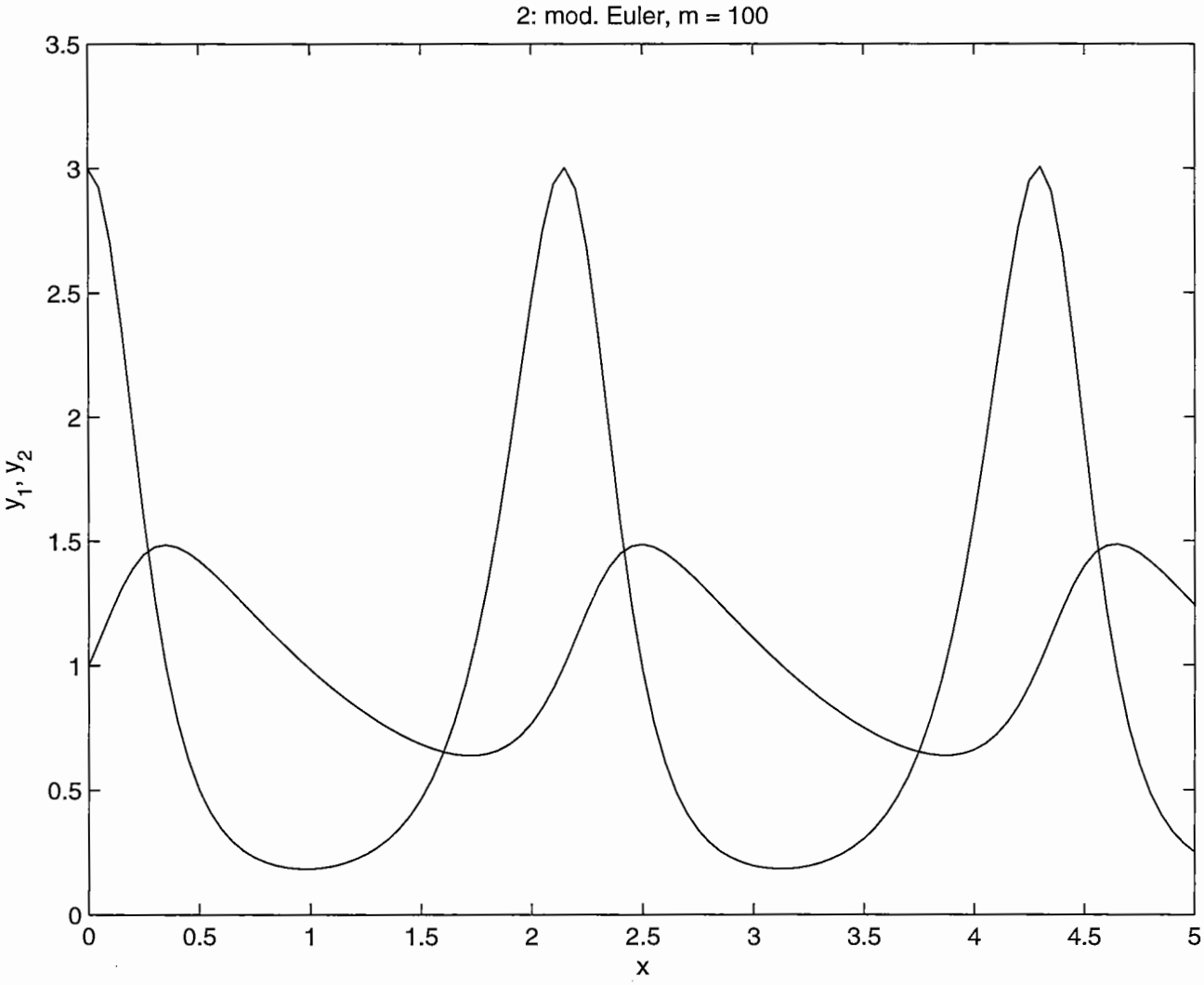
$y_1(x_f) = 0,362\,202\,564$

$y_2(x_f) = 0,950\,563\,328$



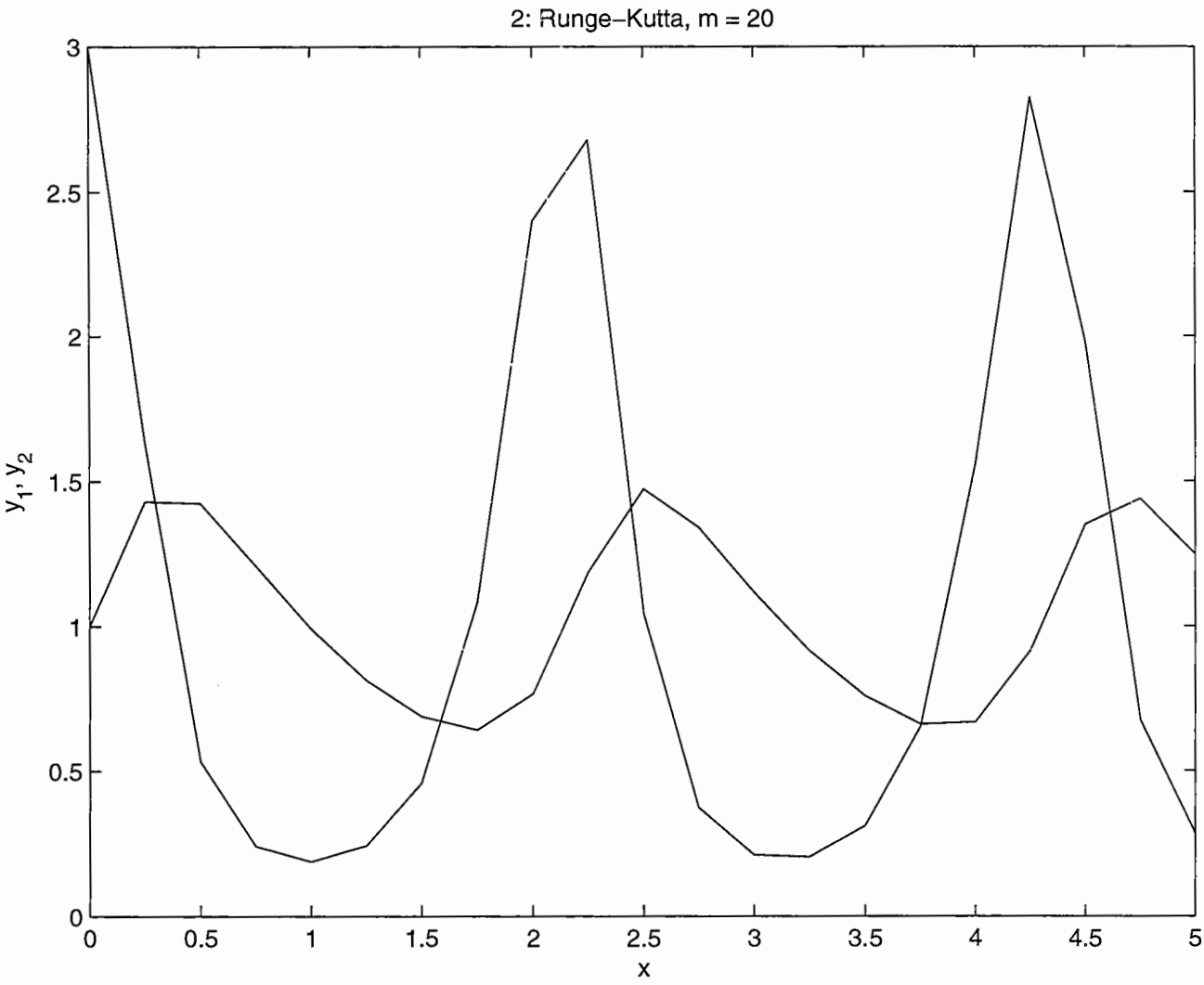
$y_1(x_f) = 0,233\,546\,918$

$y_2(x_f) = 1,196\,589\,81$



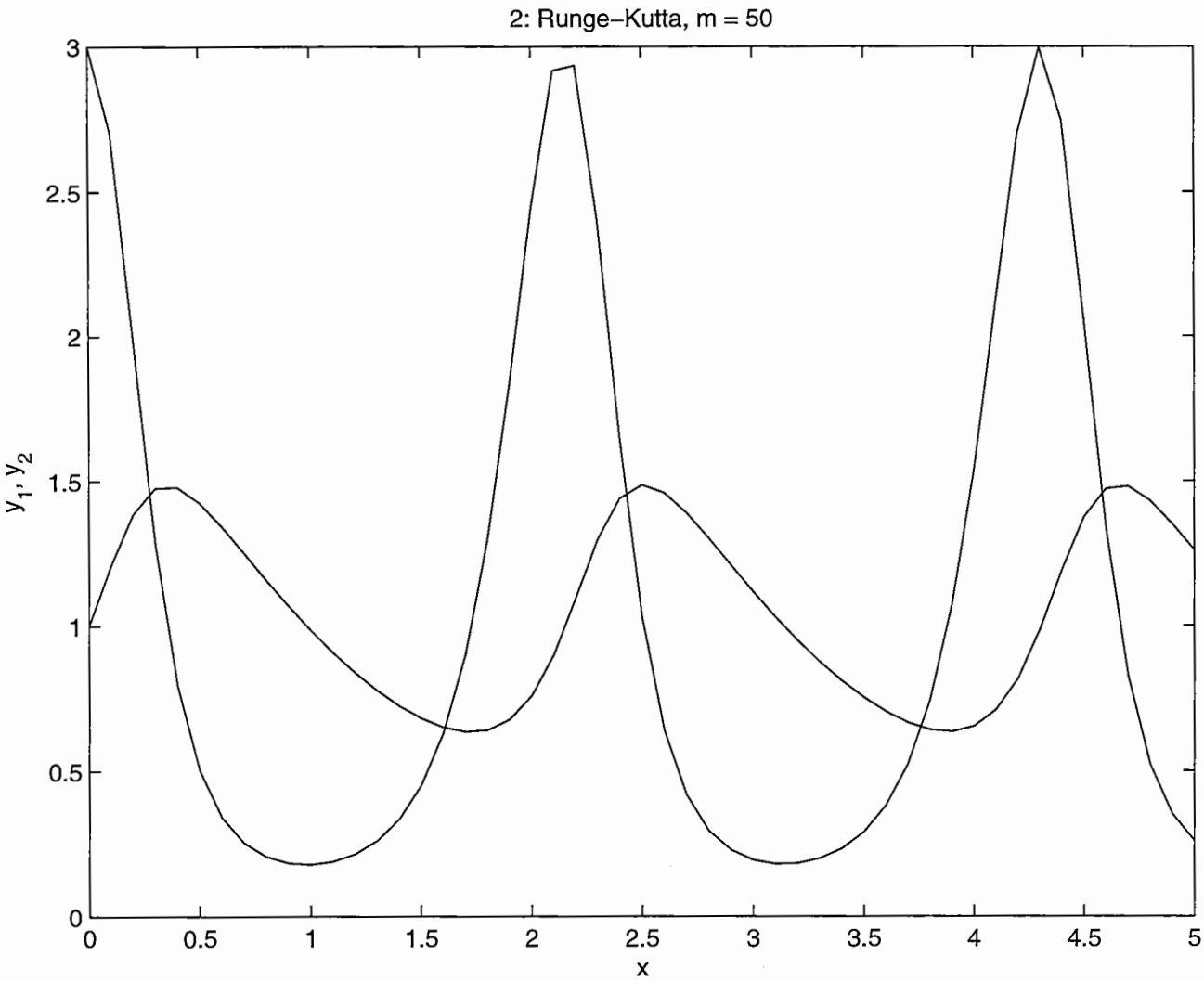
$y_1(x_f) = 0{,}250\,867\,186$

$y_2(x_f) = 1{,}24\,186\,593$



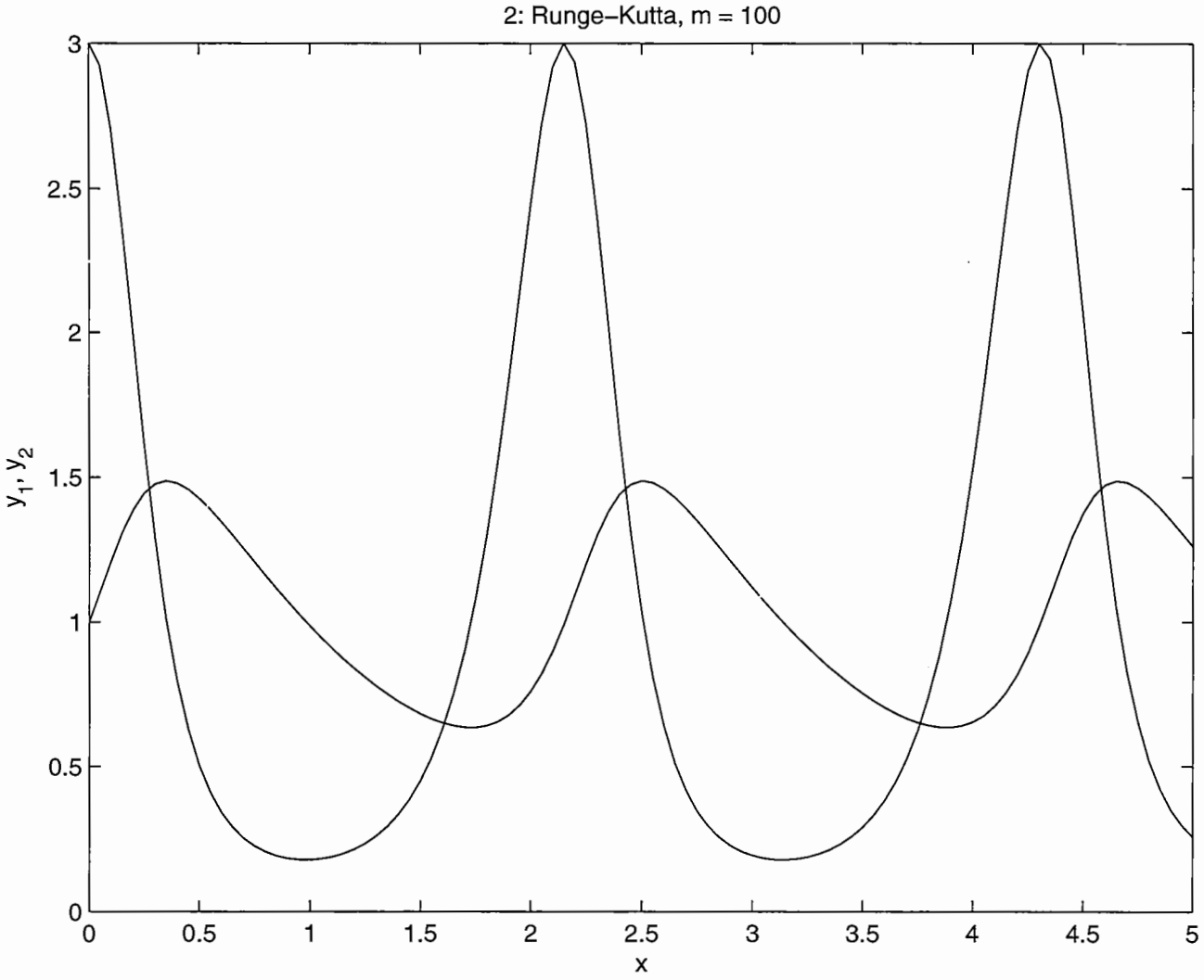
$y_1(x_f) = 0{,}284\,155\,924$

$y_2(x_f) = 1{,}24\,708\,123$



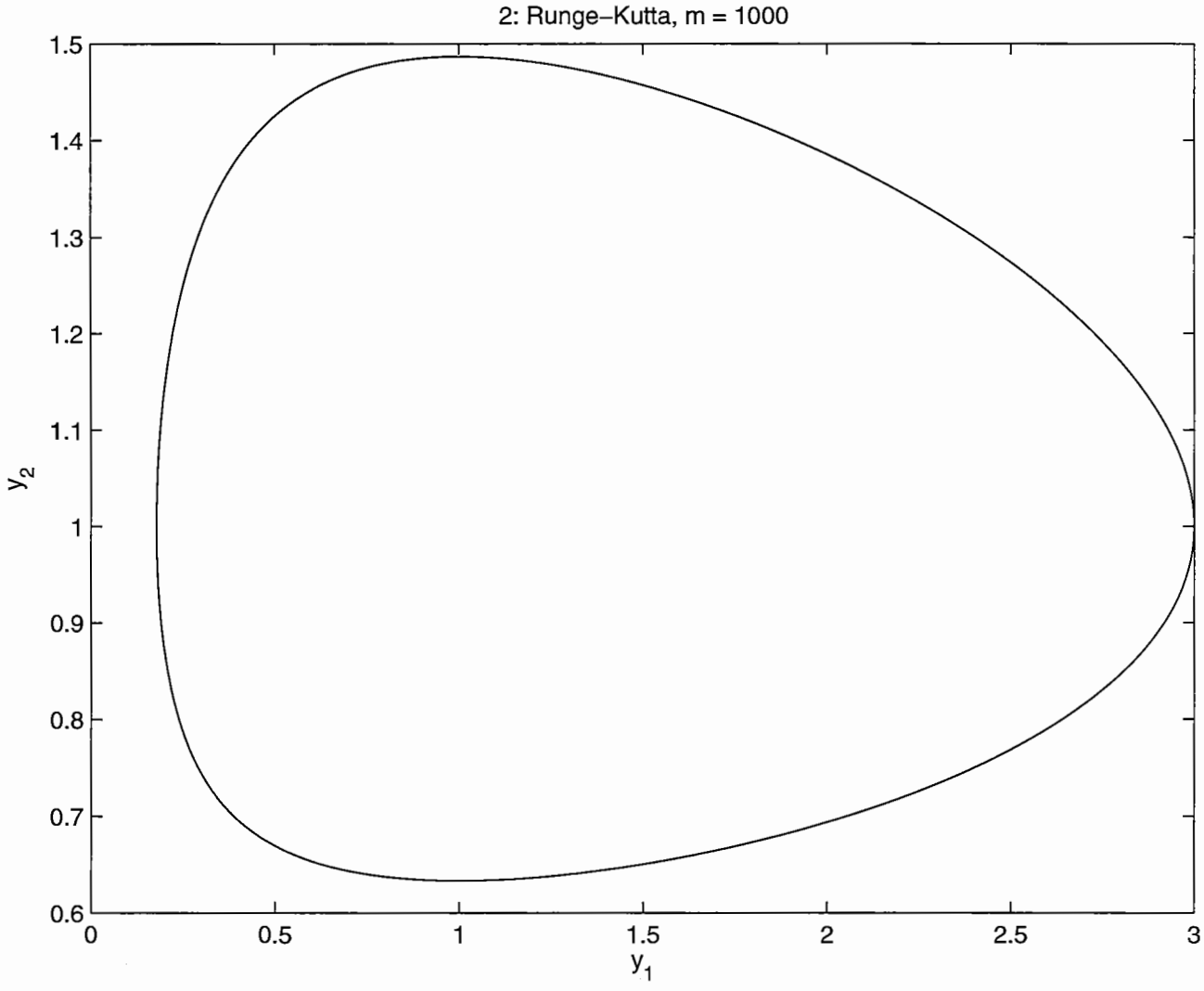
$y_1(x_f) = 0{,}258\ 792\ 891$

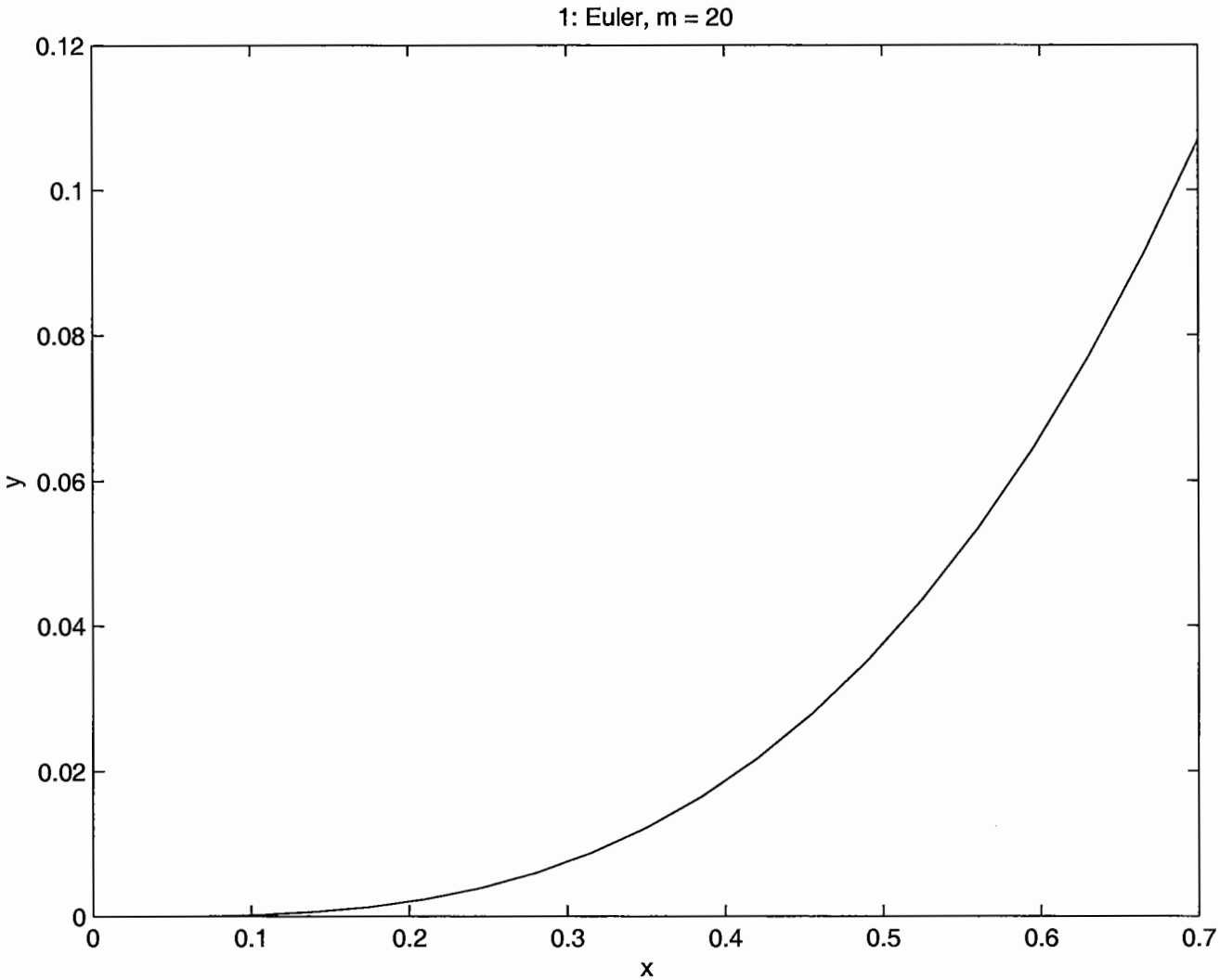
$y_2(x_f) = 1{,}126\ 006\ 315$



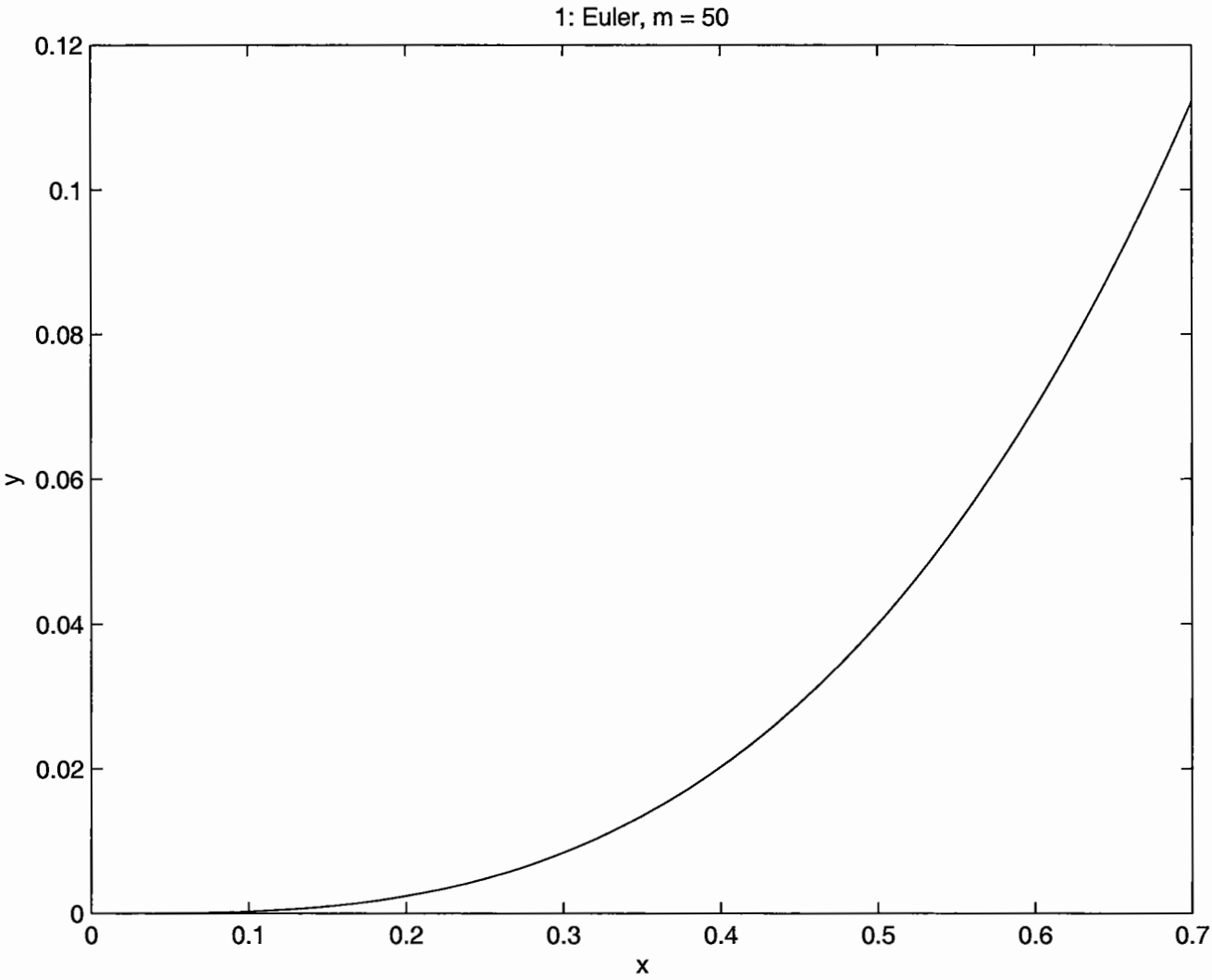
$y_1(x_f) = 0{,}258\ 229\ 939$

$y_2(x_f) = 1{,}26\ 013\ 434$

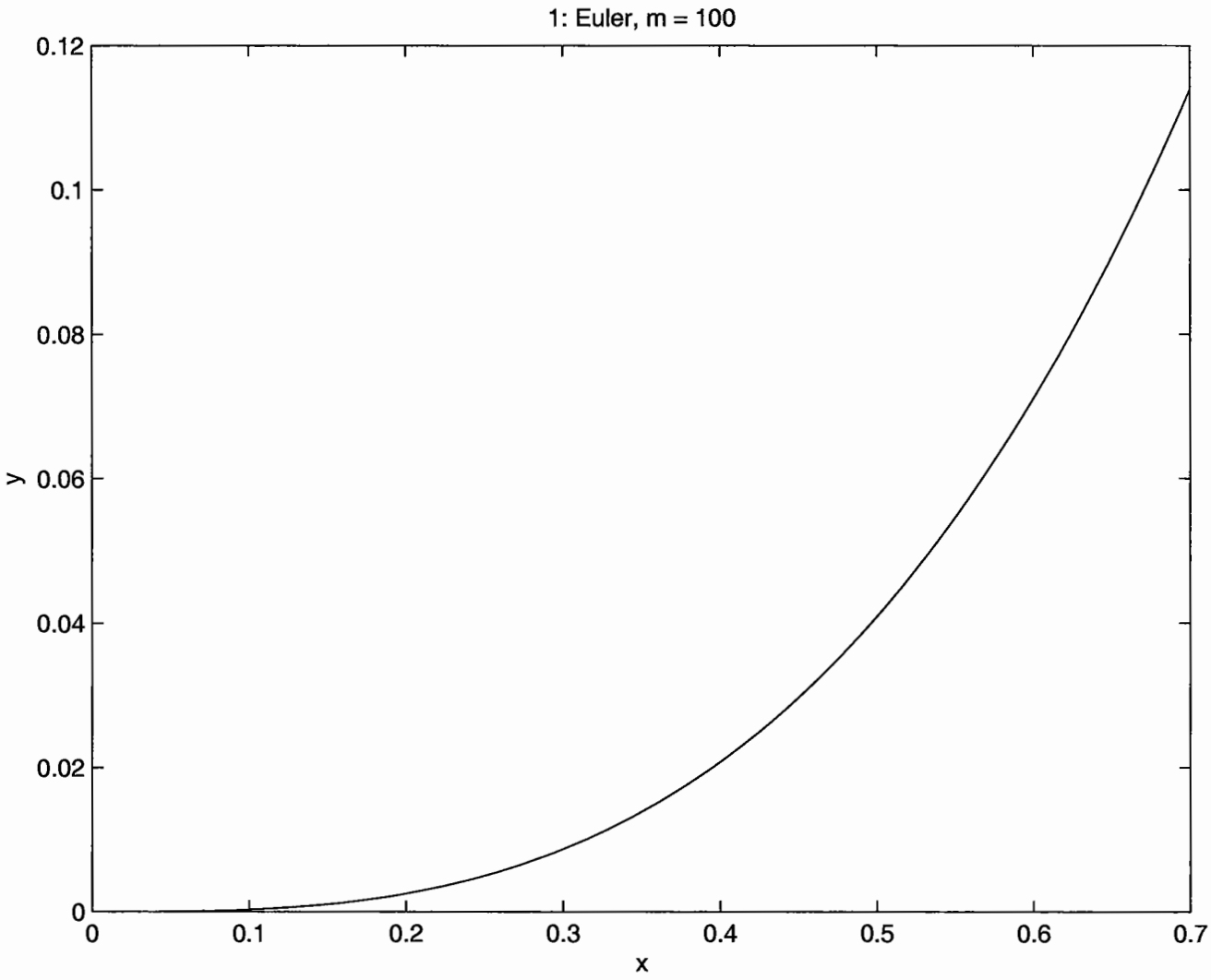

2: Runge–Kutta, m = 1000
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
y_2
0
0.5
1
1.5
2
2.5
3
y_1



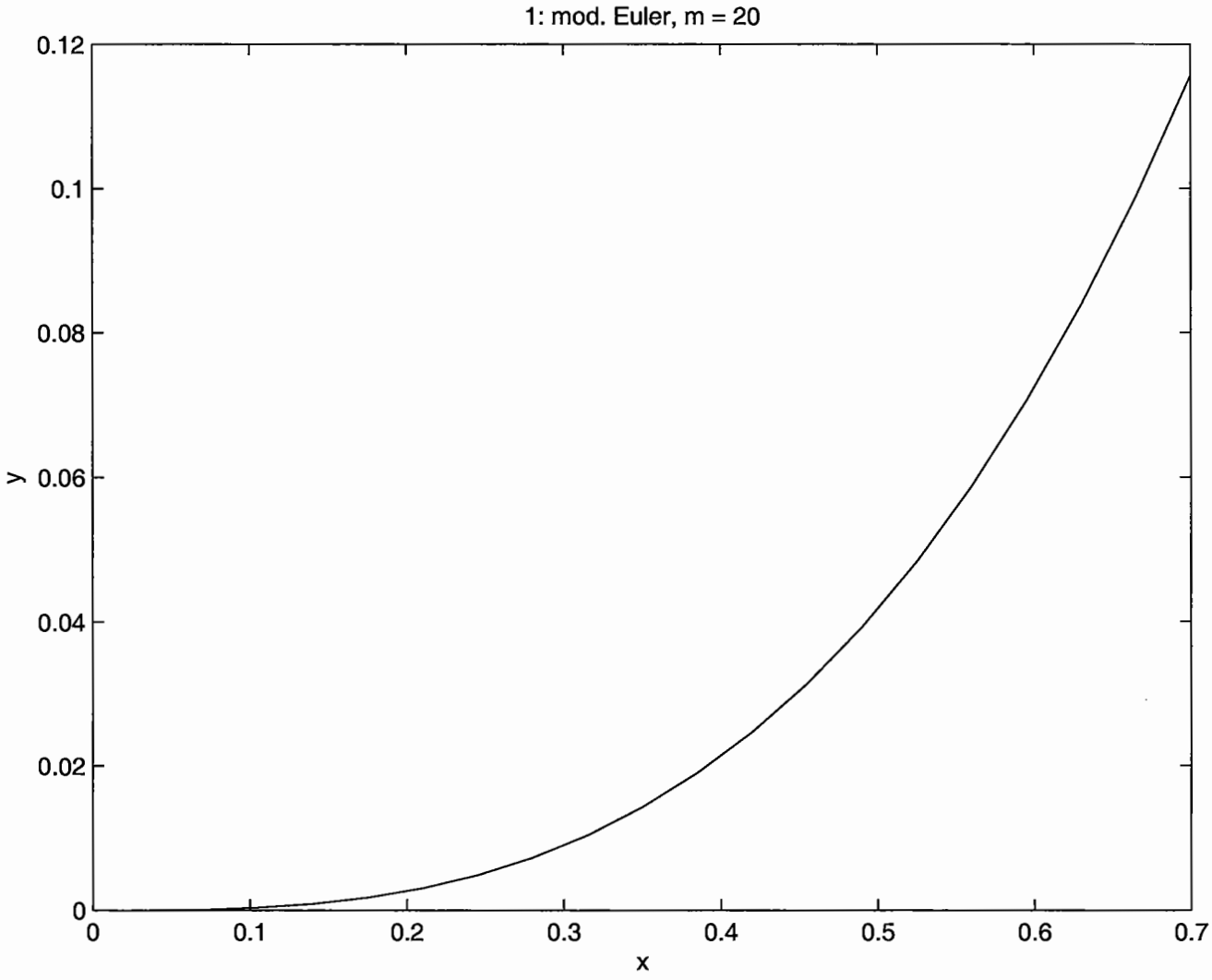
$Y(x_f) = 0{,}106\ 816\ 629$



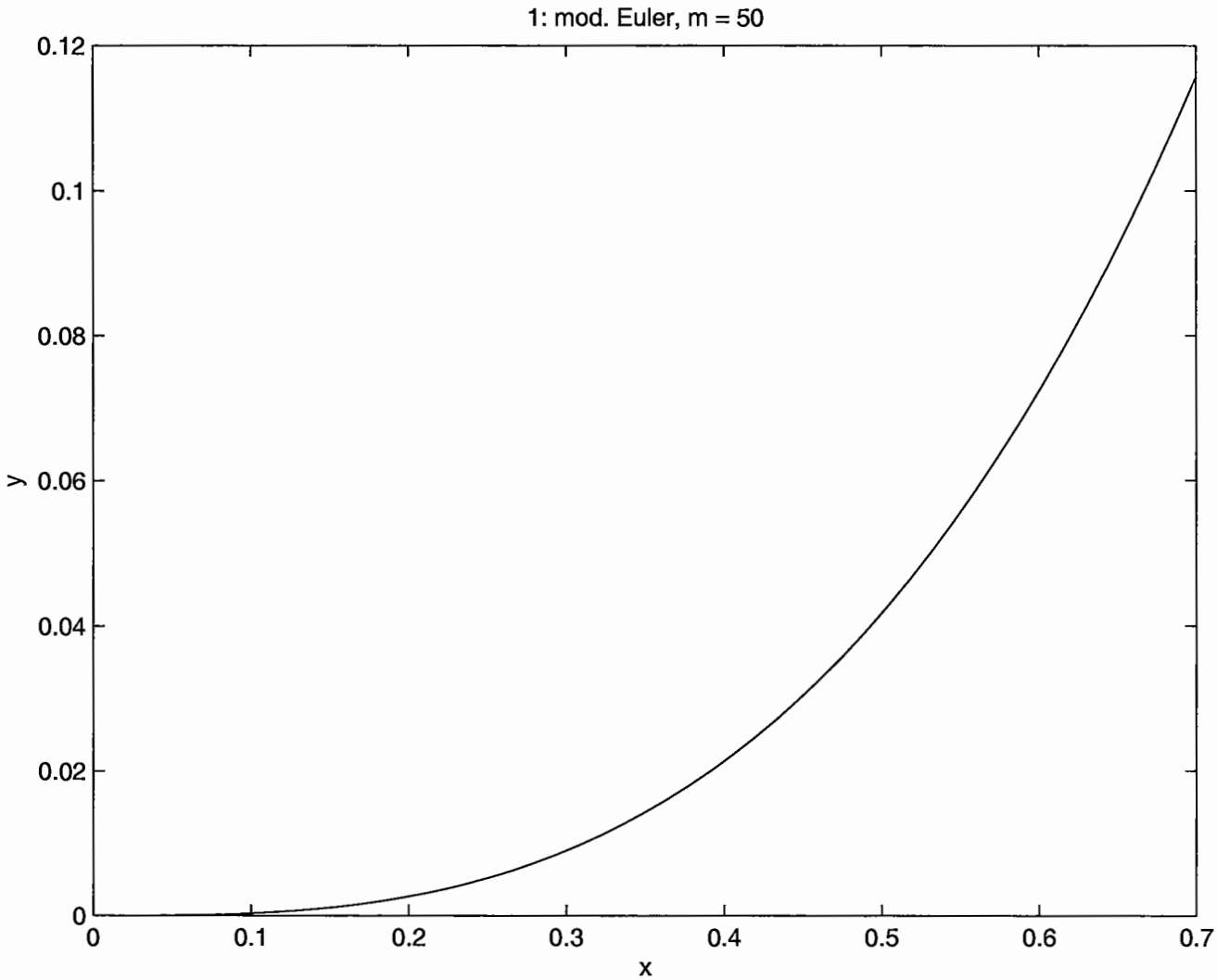
$Y(x_f) = 0,112\,074\,067$



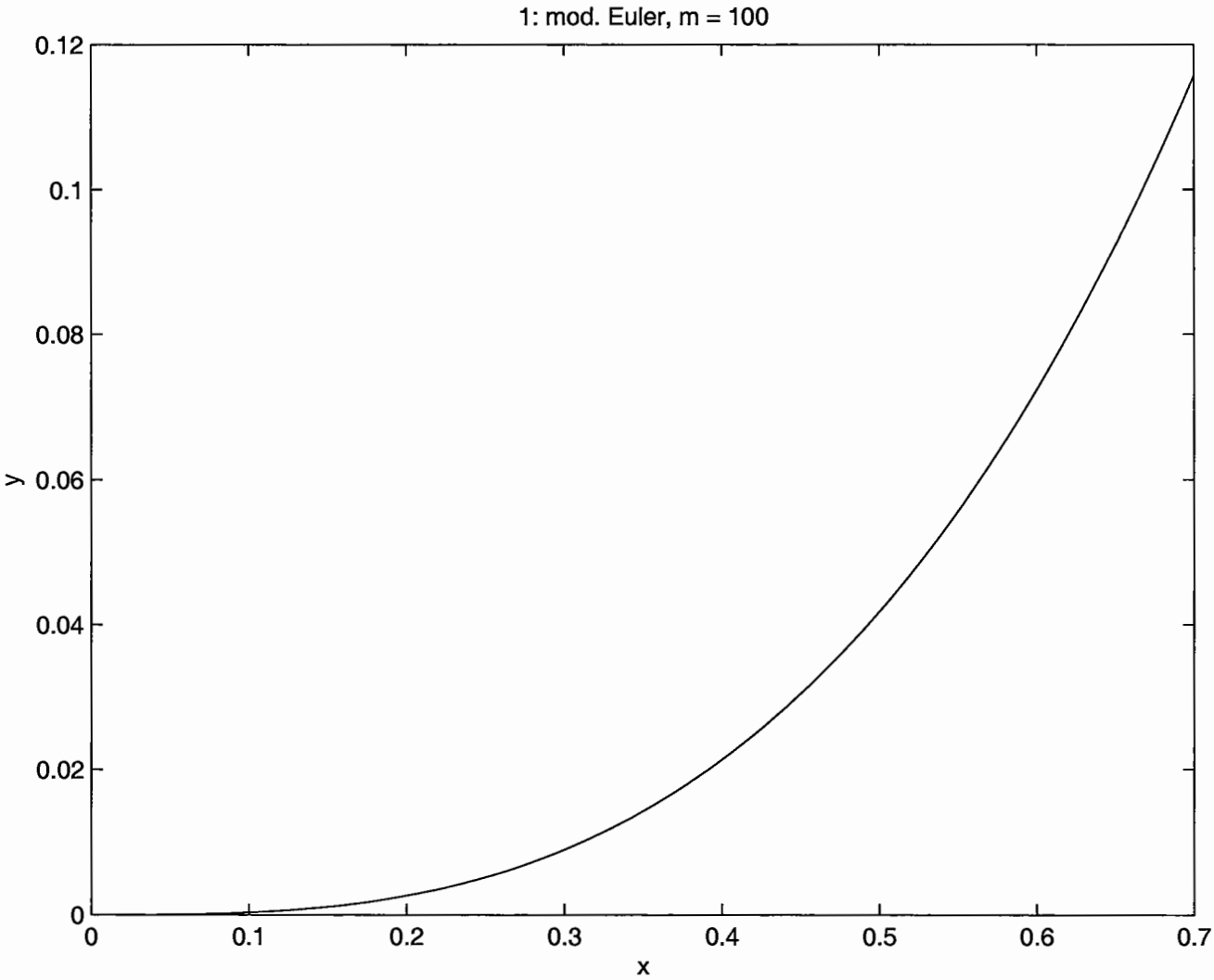
$Y(x_f) = 0{,}113\ 858\ 742$



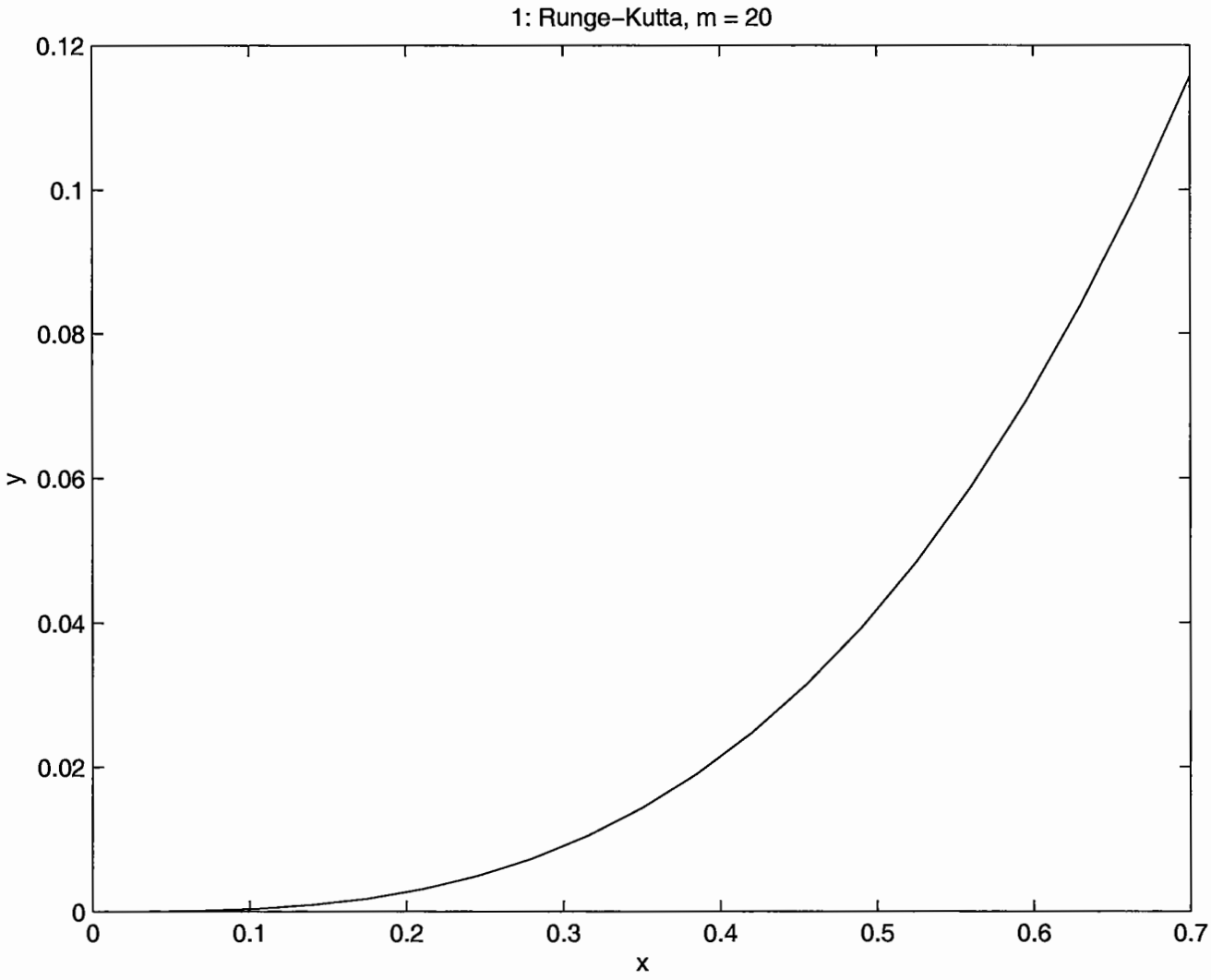
$y(x_f) = 0,115\ 572\ 984$



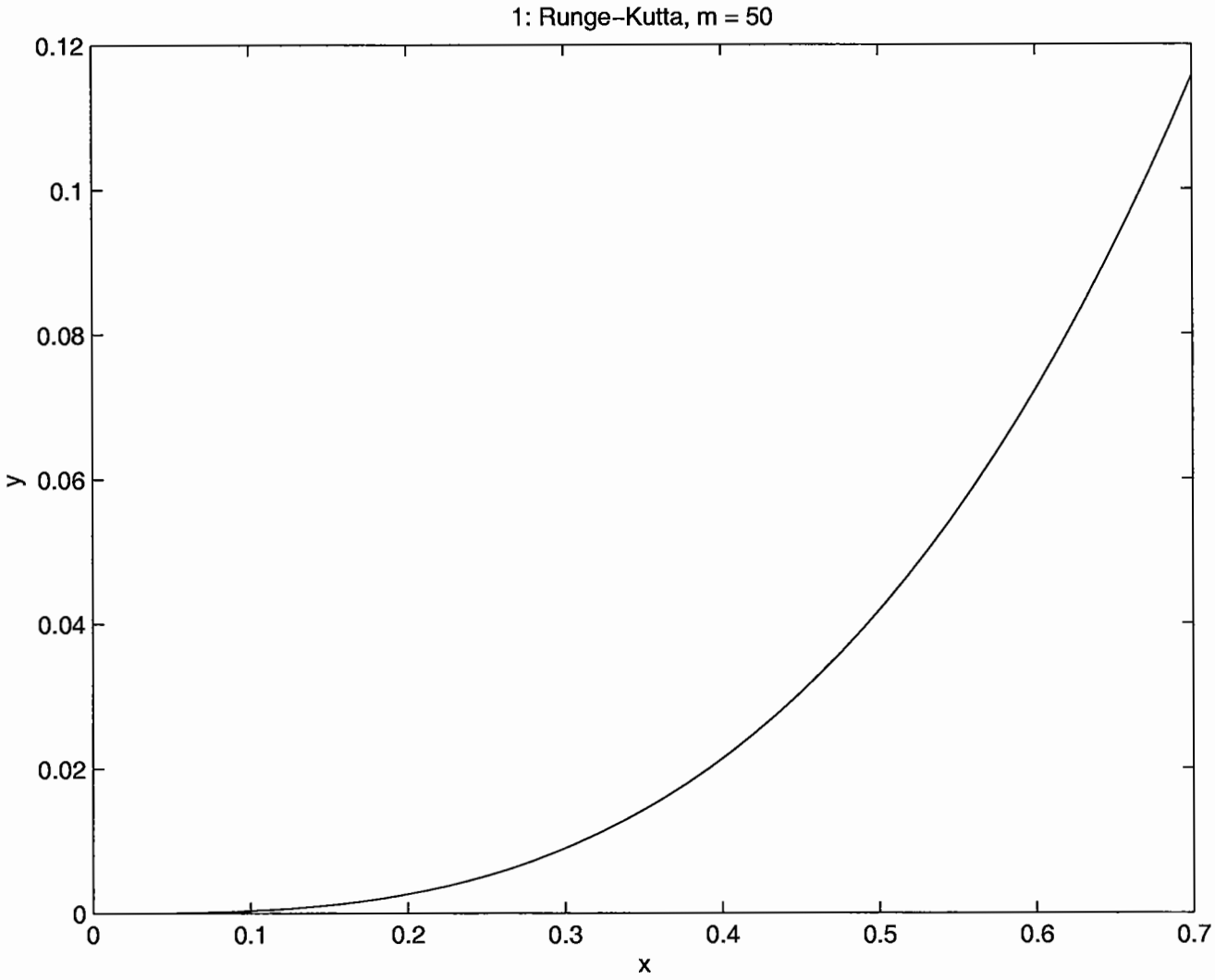
$y(x_j) = 0{,}115695920$



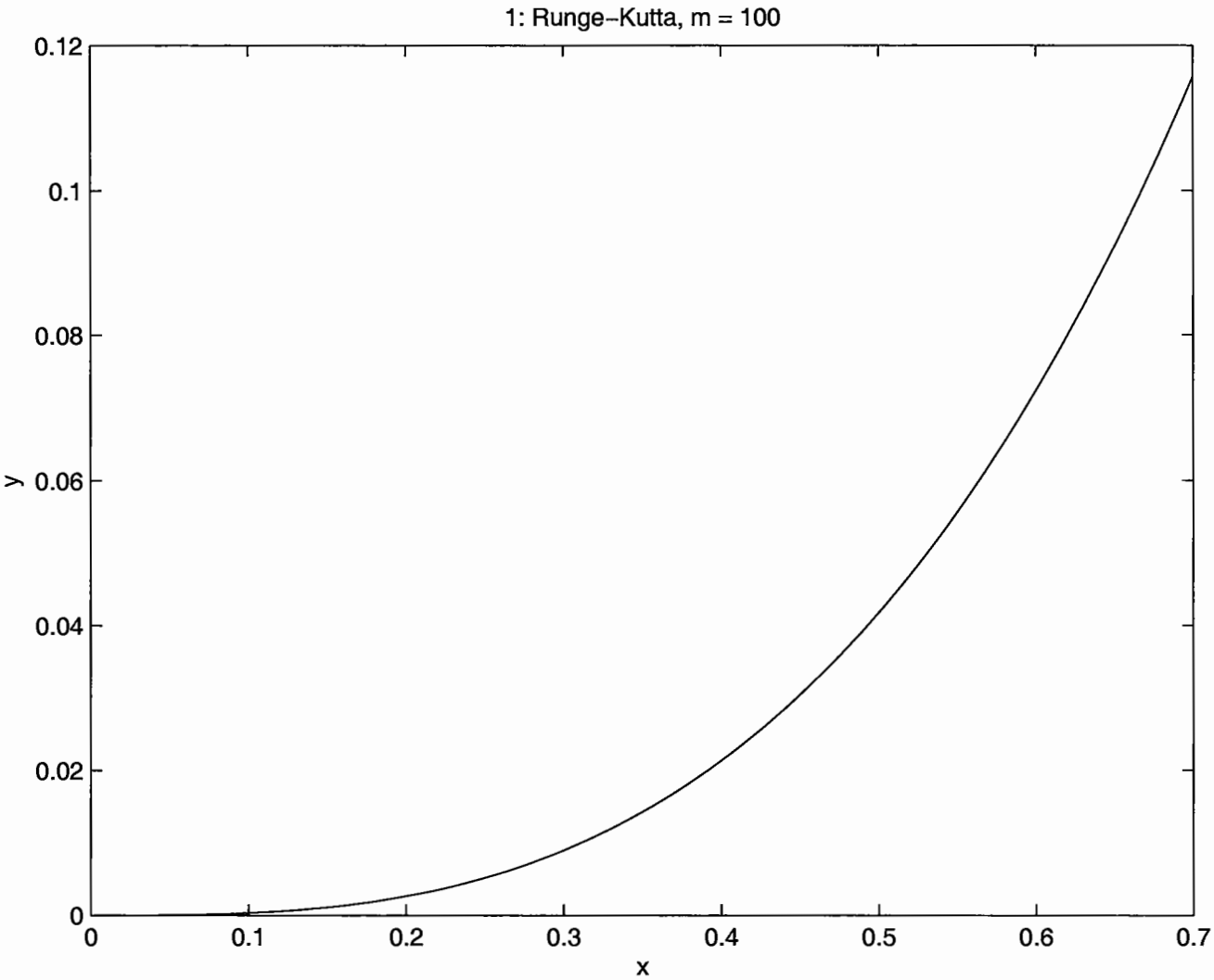
$y(x_g) = 0,115\ 656\ 368$



$y(x_f) = 0,11565\,9860$



$Y(x_f) = 0,115\ 659\ 854$



$y(x_f) = 0{,}11565985$

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 11 , Abgabe: 15.07.2004, 8.00 Uhr

Die Klausur findet am Freitag, dem 23. Juli, von 13:15 bis 16:15 im M3 statt. Als Hilfsmittel sind das Vorlesungsskript und ein Taschenrechner zugelassen. Die Unterlagen zu den Übungen sind NICHT zugelassen!

Aufgabe 37: (2+4 Punkte)

(a) Die Lösung der AWA

$$\dot{x} = t^2 + x^2, \qquad x(0) = x_0$$

sei $x(t, x_0)$. Welcher AWA genügt die Funktion $y(t, x_0) = \frac{\partial x}{\partial x_0}(t, x_0)$?

(b) Die Lösung der parameterabhängigen AWA

$$\dot{x} = t^2 + px^2, \qquad x(0) = 0, \quad (p \in \mathbb{R})$$

sei $x(t, p)$. Bestimmen Sie die AWA'n für die Ableitungen

$$y(t,p) = \frac{\partial x}{\partial p}(t,p), \quad z(t,p) = \frac{\partial^2 x}{\partial p^2}(t,p) = \frac{\partial y}{\partial p}(t,p).$$

Berechnen Sie die Lösungen dieser AWA'n explizit für $p = 0$. Berechnen Sie die TAYLOR-Approximation um $p_0 = 0$

$$x(t,p) \cong x(t,0) + p\,y(t,0) + \frac{1}{2}p^2 z(t,0)$$

in $t = 1/\sqrt{2}$ und $p = 1$ und vergleichen Sie diese mit dem exakten Wert $x(\frac{1}{\sqrt{2}}, 1) = 0.119275711$.

Aufgabe 38: (3 Punkte)
Konstruieren Sie ein Verfahren 2. Ordnung zur Lösung der AWA

$$y' = y\,, \quad y(0) = 1$$

mittels des Taylorreihenansatzes (27.6). Vergleichen Sie Ihr Verfahren mit dem EULER-Verfahren und der exakten Lösung.

Aufgabe 39: (3 Punkte)
Zeigen Sie, dass das RUNGE-KUTTA-Verfahren der 3.Stufe

$$\begin{aligned} y_{k+1} &= y_k + \tfrac{h}{6}(f_1(x_k, y_k) + 4f_2(x_k, y_k) + f_3(x_k, y_k)), \\ f_1(x_k, y_k) &= f(x_k, y_k), \\ f_2(x_k, y_k) &= f(x_k + \tfrac{h}{2}, y_k + \tfrac{h}{2} f_1(x_k, y_k)), \\ f_3(x_k, y_k) &= f(x_k + h, y_k - h f_1(x_k, y_k) + 2h f_2(x_k, y_k)) \end{aligned}$$

ein Einschrittverfahren der Ordnung $p = 3$ ist.

3 Hd

Sei $f(t,x) = t^2 + x^2$.

Nach 26.13 erfüllt $y(t,x_0) = \frac{\partial x}{\partial x_0}(t,x_0)$:

$$\dot{y} = \frac{\partial f}{\partial x}(t, x(t,x_0))\, y = 2x(t,x_0)\, y \quad , \quad y(0,x_0) = 1.$$

(Denn: $\dot{y} = \frac{\partial y}{\partial t} = \frac{\partial}{\partial t}\frac{\partial}{\partial x_0} x(t,x_0) = \frac{\partial}{\partial x_0}\frac{\partial}{\partial t} x(t,x_0) = \frac{\partial}{\partial x_0} f(t, x(t,x_0))$

$= \frac{\partial f}{\partial x}(t, x(t,x_0))\, \frac{\partial x}{\partial x_0} = \frac{\partial f}{\partial x}(t, x(t,x_0))\, y$

und $y(0,x_0) = \frac{\partial x}{\partial x_0}(0,x_0) = \frac{\partial}{\partial x_0} \underbrace{x(0,x_0)}_{=x_0} = 1$)

b)

Es gilt somit $f(t,x,p) = t^2 + px^2$

$$\dot{y} = \frac{\partial}{\partial t}\frac{\partial}{\partial p} x(t,p) = \frac{\partial}{\partial p}\dot{x}(t,p) = \frac{\partial}{\partial p} f(t, x(t,p), p)$$

$$= f_x \frac{\partial}{\partial p} x(t,p) + f_p = f_x\, y + f_p$$

und $y(0,p) = \frac{\partial}{\partial p} \underbrace{x(0,p)}_{=0} = 0$

Somit: $\dot{y} = 2pxy + x^2 \quad , y(0,p) = 0$

Weiterhin gilt:

$$\dot{z} = \frac{\partial}{\partial p}\dot{y} = \frac{\partial}{\partial p}\left(f_x(t, x(t,p), p)\, y(t,p) + f_p(t, x(t,p), p)\right)$$

$$= f_{xx} \underbrace{\frac{\partial x}{\partial p}}_{=y} y + f_{xp}\, y + f_x\, z + f_{px} \underbrace{\frac{\partial x}{\partial p}}_{=y} + f_{pp}$$

$$= f_x z + f_{xx} y^2 + 2 f_{xp} y + f_{pp}$$

und $z(0,p) = \frac{\partial}{\partial p} y(0,p) = \frac{\partial}{\partial p} 0 = 0$

Somit: $\dot{z} = 2px\, z + 2py^2 + 2xy \quad , z(0,p) = 0$

Exakt: $y(x_i) = \left(1+h+\frac{h^2}{2}+O(h^3)\right)^i$

Euler: $|y(x_i) - y_i| = \left(1+h+O(h^2)\right)^i - (1+h)^i = O(h^2)$

2. Ordnung: $|y(x_i) - y_i| = \left(1+h+\frac{h^2}{2}+O(h^3)\right)^i - \left(1+h+\frac{h^2}{2}\right)^i = O(h^3)$

3g) Lokaler Diskretisierungsfehler:

$$\tau_h(x,y) = \frac{y(x+h)-y(x)}{h} - f_h(x, y(x))$$

Taylorreihe von $y(x+h)$:

$$y(x+h) = y(x) + h\,y'(x) + \frac{h^2}{2}y''(x) + \frac{h^3}{6}y'''(x) + O(h^4)$$

$$\Rightarrow \frac{y(x+h)-y(x)}{h} = y'(x) + \frac{h}{2}y''(x) + \frac{h^2}{6}y'''(x) + O(h^3)$$

mit $y'(x) = f(x,y)$

$$y''(x) = \frac{d}{dx} f(x, y(x)) = f_x(x,y) + f_y(x,y) f(x,y)$$

$$y'''(x) = \frac{d}{dx} y''(x) = f_{xx}(x,y) + f_{xy}(x,y) f(x,y)$$
$$+ f_{yx}(x,y) f(x,y) + f_{yy}(x,y) f^2(x,y)$$
$$+ f_y(x,y) f_x(x,y) + f_y^2(x,y) f(x,y)$$

$$\Rightarrow \frac{y(x+h)-y(x)}{h} = f + \frac{h}{2}(f_x + f_y f) + \frac{h^2}{6}(f_{xx} + 2f_{xy}f + f_{yy}f^2$$
$$+ f_y f_x + f_y^2 f) + O(h^3) \quad (*)$$

Berechne nun Taylorentwicklung von $f_h(x,y) = \frac{1}{6}(f_1(x,y) + 4f_2(x,y) + f_3(x,y))$

Dazu berechnen wir die Ableitung $\frac{d}{dh} f_i(x,y)$:

$$\frac{d}{dh} f_1(x,y) = \frac{d}{dh} f(x,y) = 0$$

$$\frac{d}{dh} f_2(x,y) = \frac{d}{dh} f\left(x+\frac{h}{2}, y+\frac{h}{2}f(x,y)\right) = \frac{1}{2} f_x\left(x+\frac{h}{2}, y+\frac{h}{2}f(x,y)\right)$$
$$+ \frac{1}{2} f(x,y) f_y\left(x+\frac{h}{2}, y+\frac{h}{2}f(x,y)\right)$$

$$\Rightarrow \frac{d^2}{dh^2} f_3(x,y)\Big|_{h=0} = f_{xx} + f f_{xy} + f_{yx} f + f_{yy} f^2 + 4 f_y \left(\frac{1}{2} f_x + \frac{1}{2} f f_y\right)$$

$$= f_{xx} + 2 f f_{xy} + f^2 f_{yy} + 2 f_y f_x + 2 f f_y^2$$

$$\Rightarrow f_3(x,y) = f(x,y) + h\left(f_x + f f_y\right) + \frac{h^2}{2}\left(f_{xx} + 2 f f_{xy} + f^2 f_{yy} + 2 f_y f_x + 2 f f_y^2\right) + O(h^3)$$

$$\Rightarrow f_h(x,y) = \frac{1}{6}\left(f_1(x,y) + 4 f_2(x,y) + f_3(x,y)\right)$$

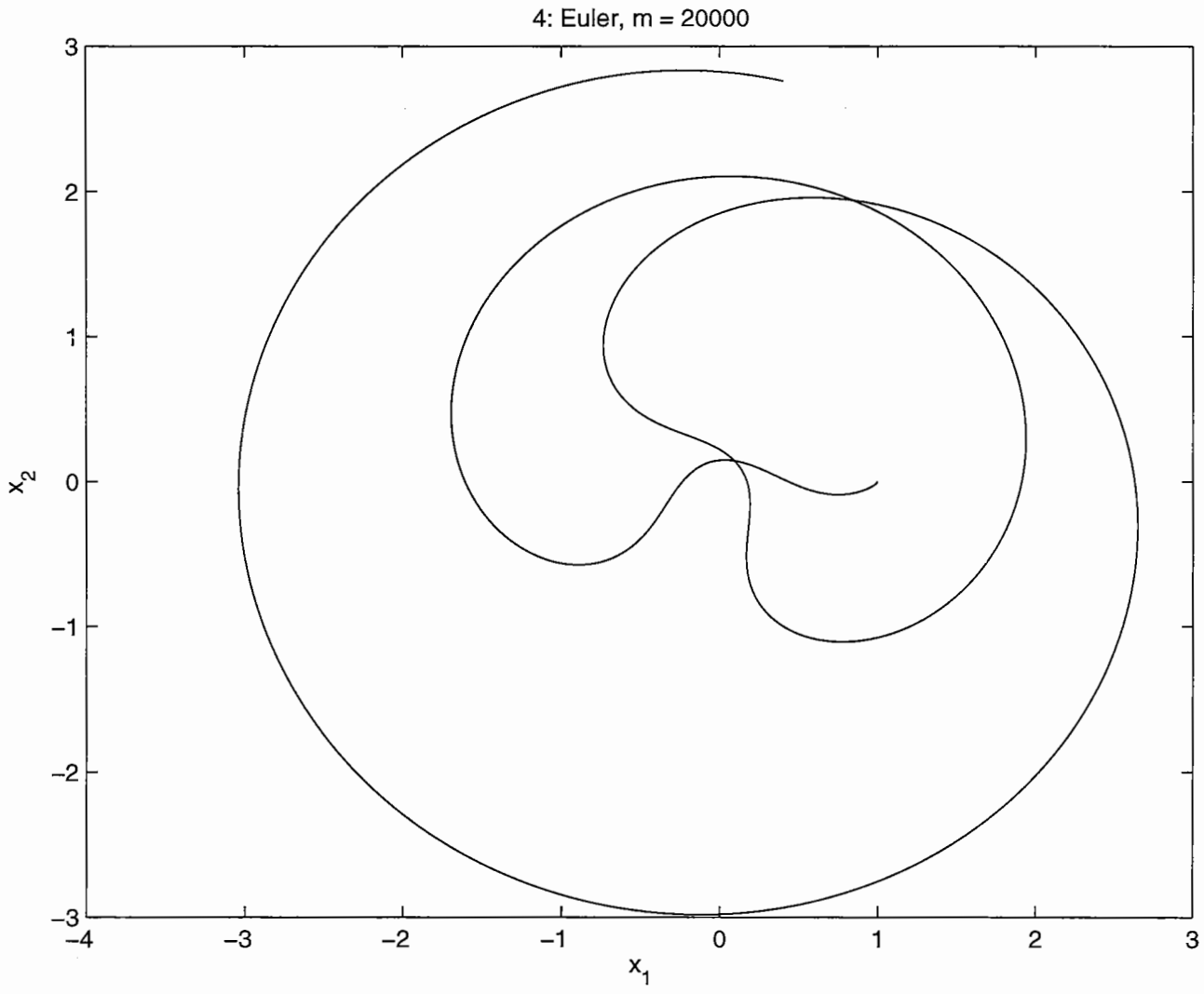
$$= f + \frac{1}{6} h \left(2 f_x + 2 f f_y + f_x + f f_y\right)$$

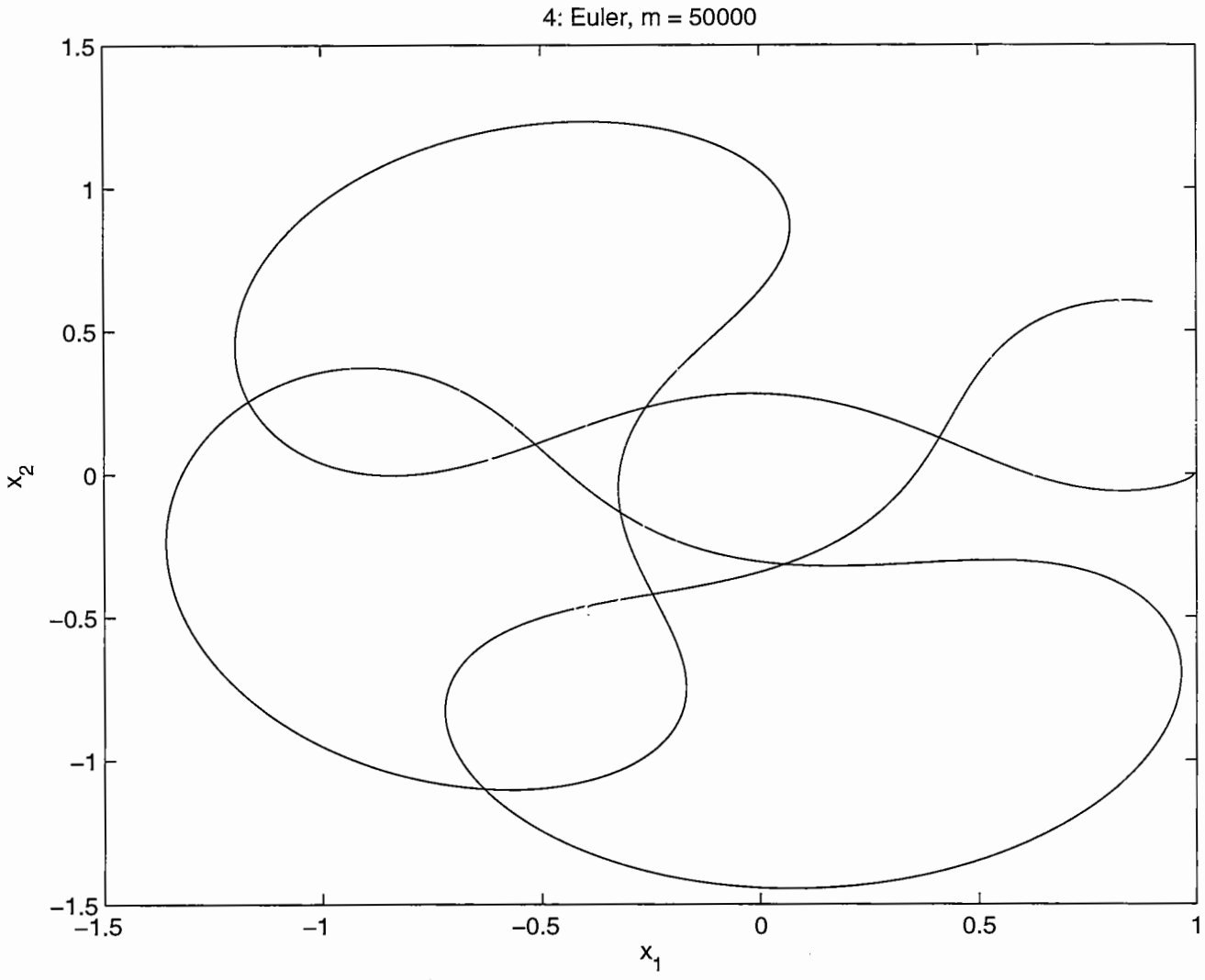
$$+ \frac{h^2}{12}\left(f_{xx} + 2 f f_{xy} + f^2 f_{yy} + f_{xx} + 2 f f_{xy} + f^2 f_{yy} + 2 f_y f_x + 2 f f_y^2\right) + O(h^3)$$

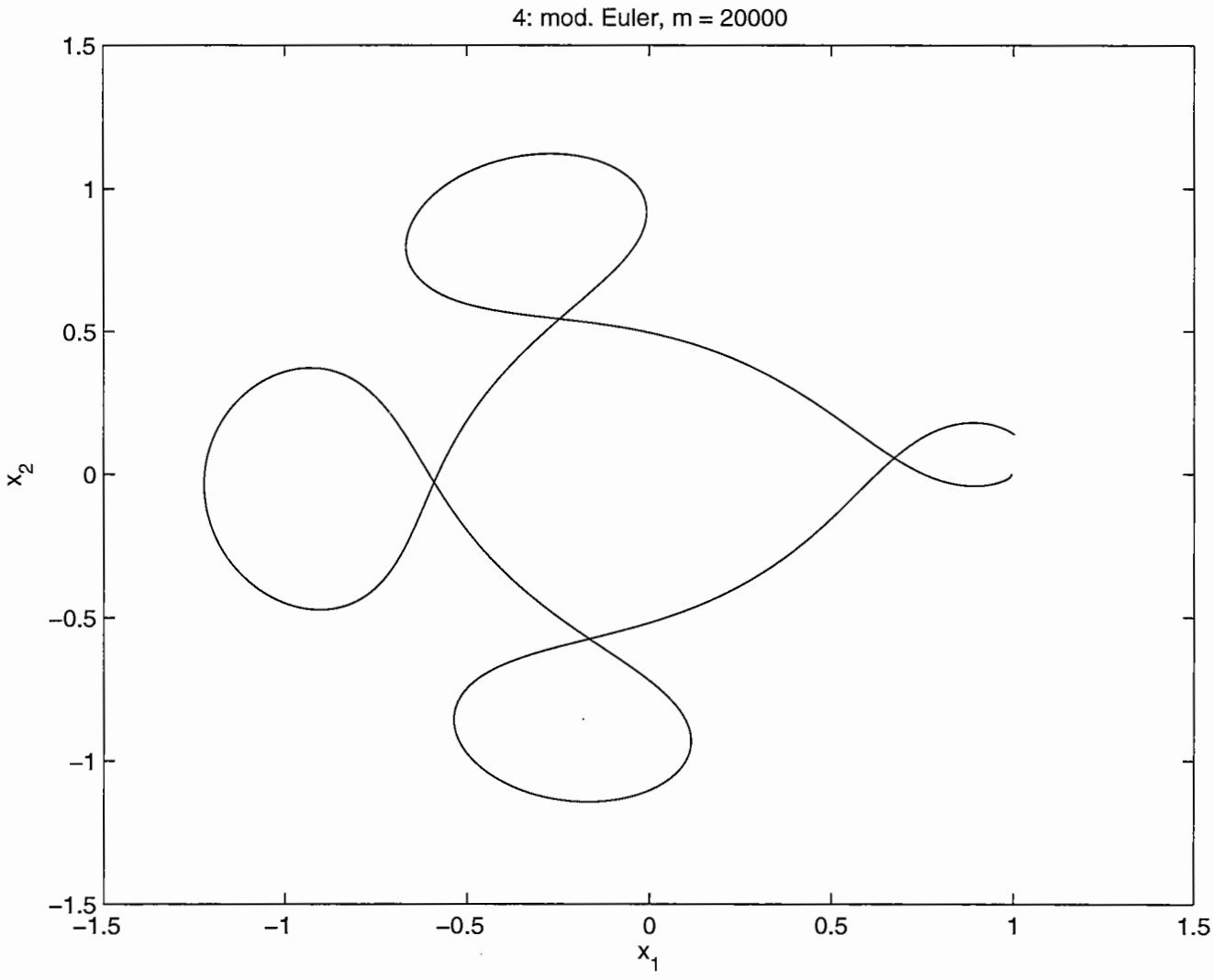
$$= f + \frac{h}{2}\left(f_x + f f_y\right) + \frac{h^2}{6}\left(f_{xx} + 2 f f_{xy} + f^2 f_{yy} + f_y f_x + f f_y^2\right) + O(h^3)$$

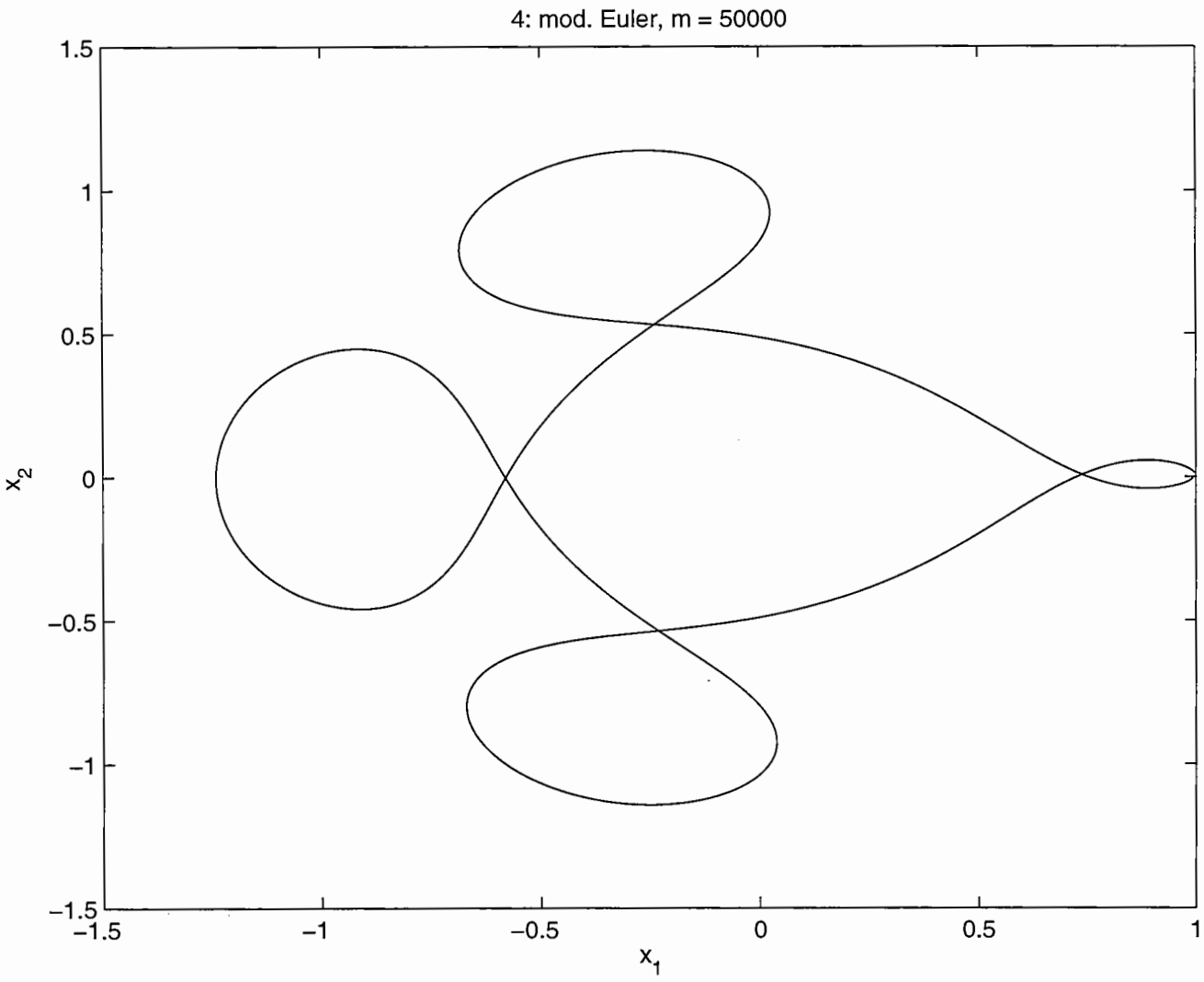
$$\overset{(*)}{\Rightarrow} \tau_h(x,y) = \frac{y(x+h) - y(x)}{h} - f_h(x,y) = O(h^3)$$

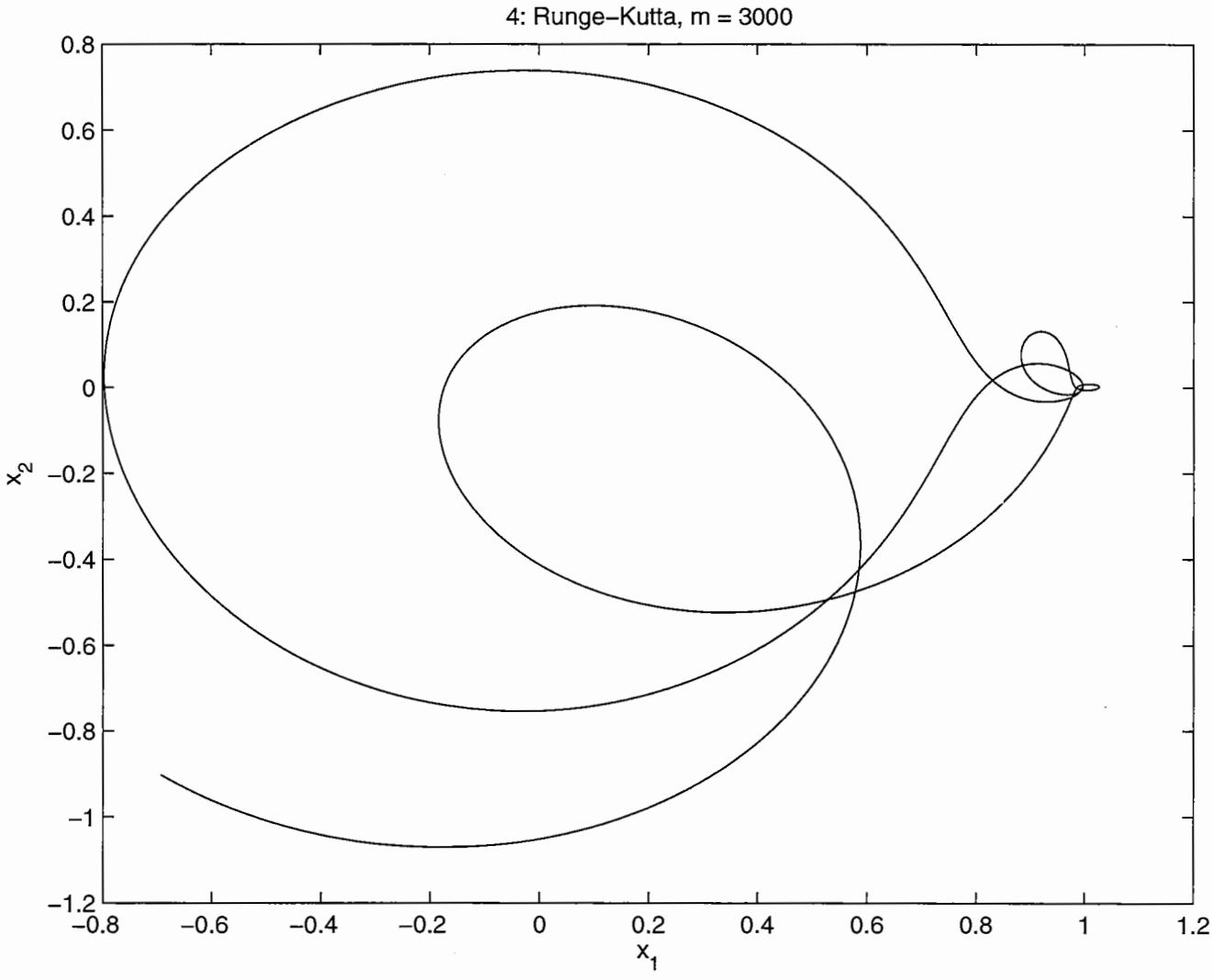
$$\Rightarrow \text{Ordnung } p = 3 .$$

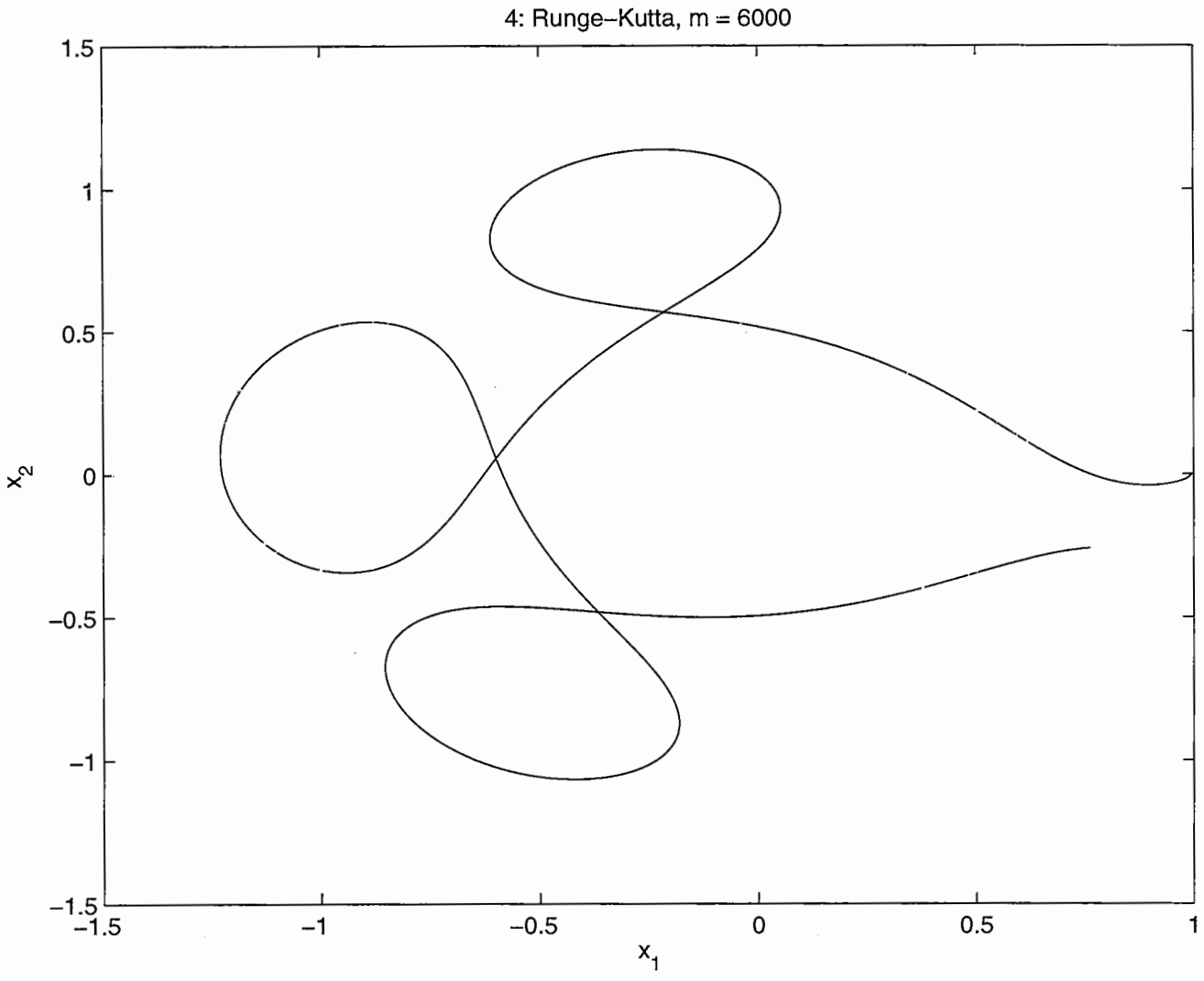

4: Euler, m = 20000
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
x_2
x_1

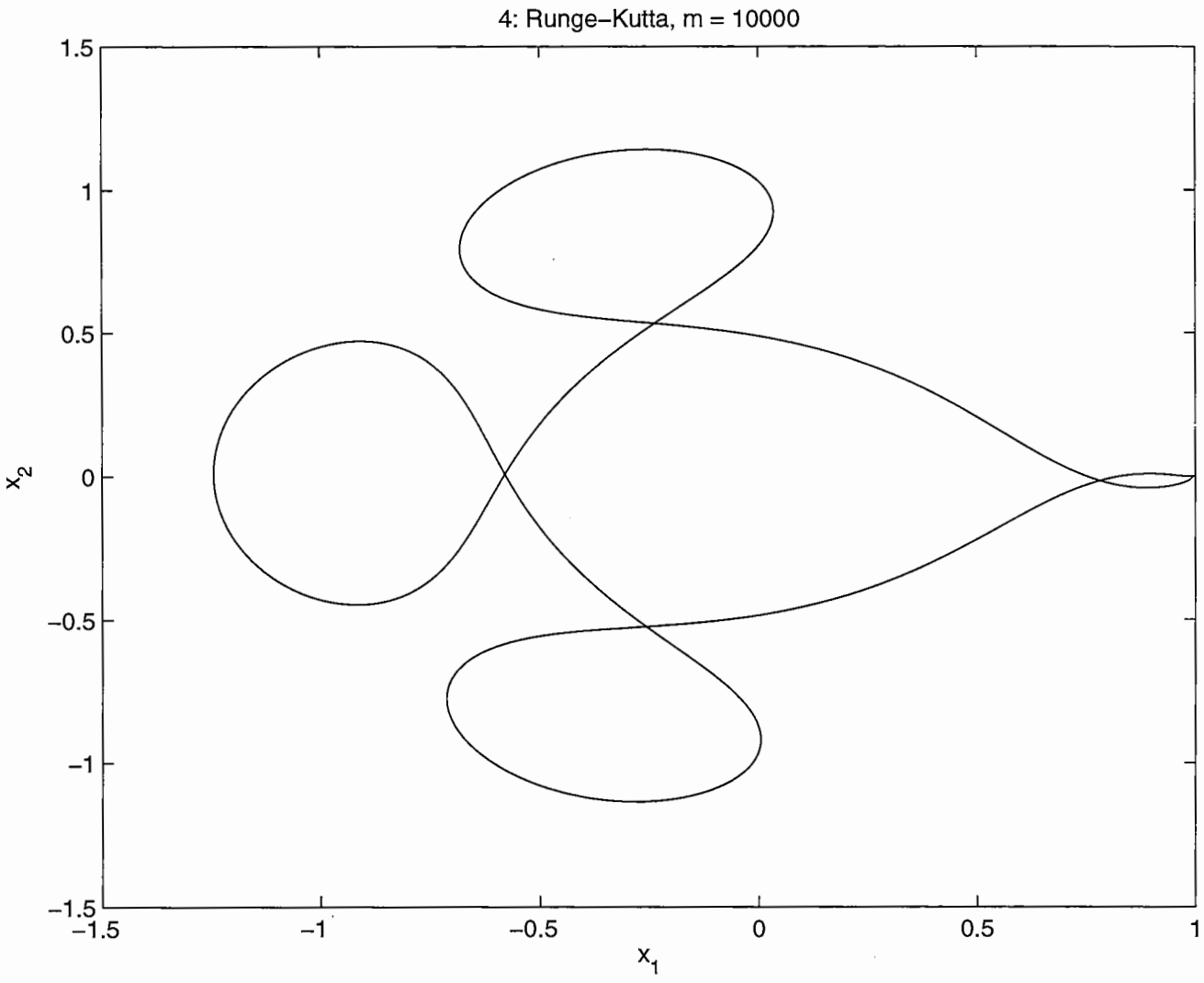

4: Euler, m = 50000
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
x2
x1

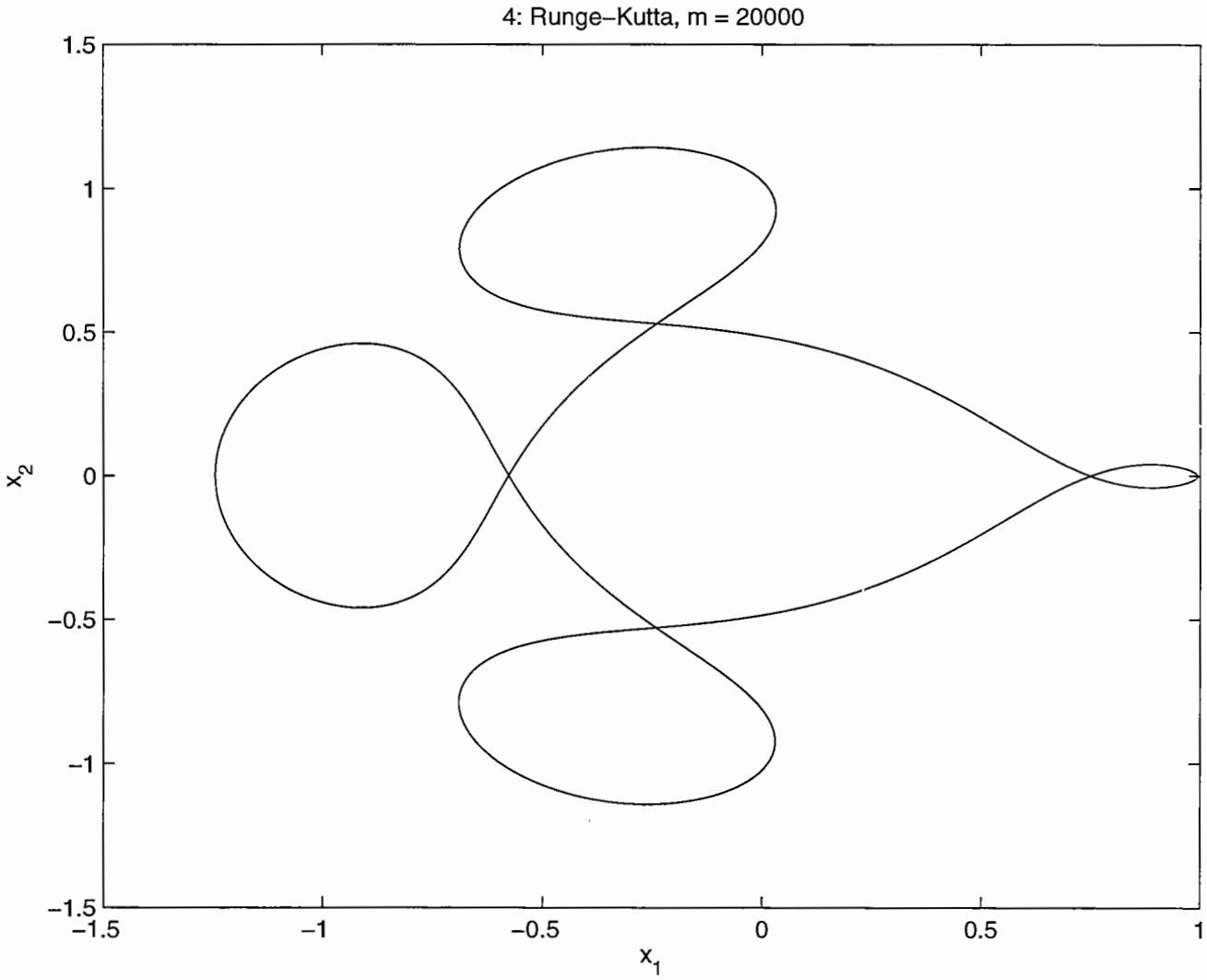

4: mod. Euler, m = 20000
1.5
1
0.5
0
−0.5
−1
−1.5
x2
−1.5
−1
−0.5
0
0.5
1
1.5
x1

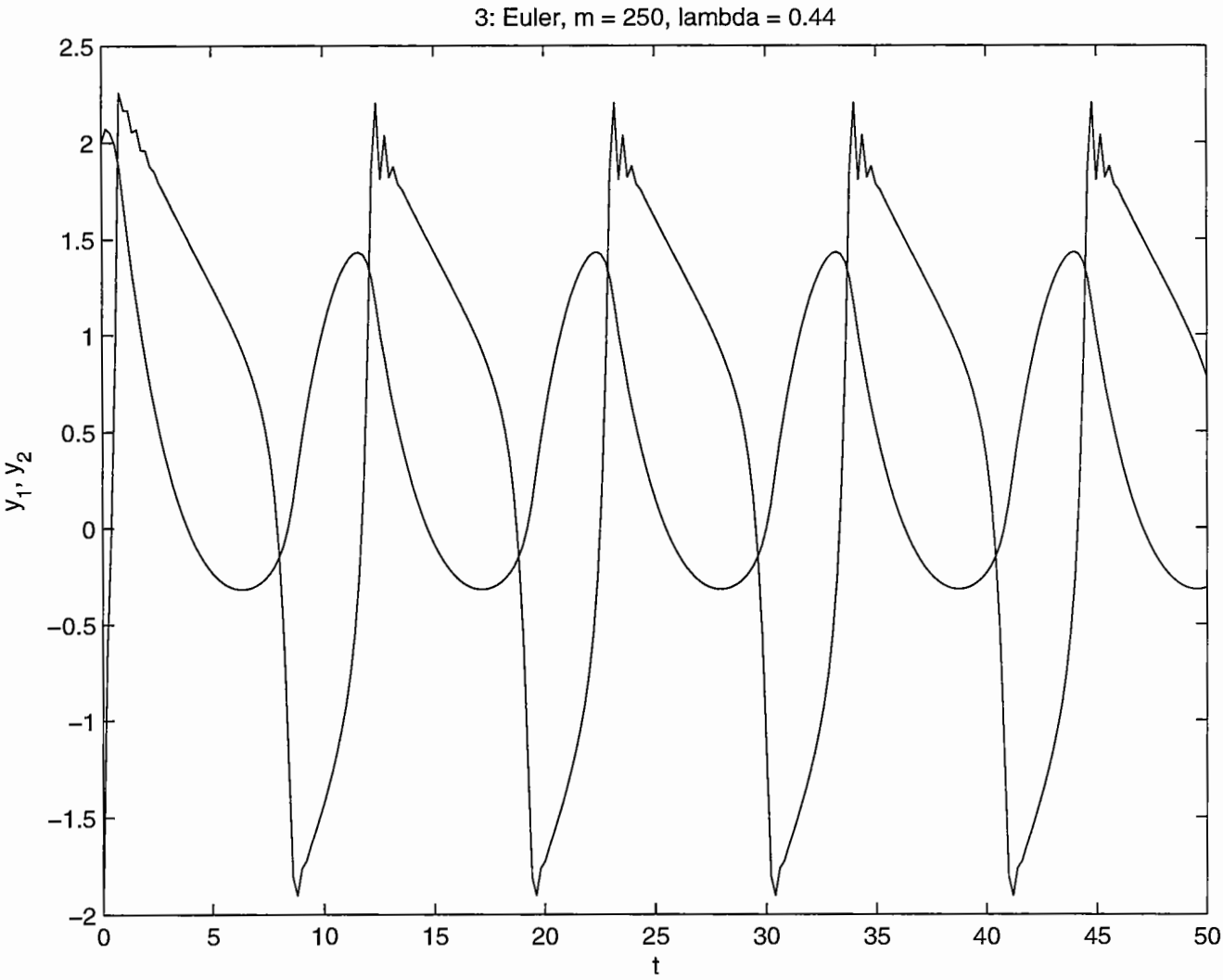

4: mod. Euler, m = 50000
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
x2
x1

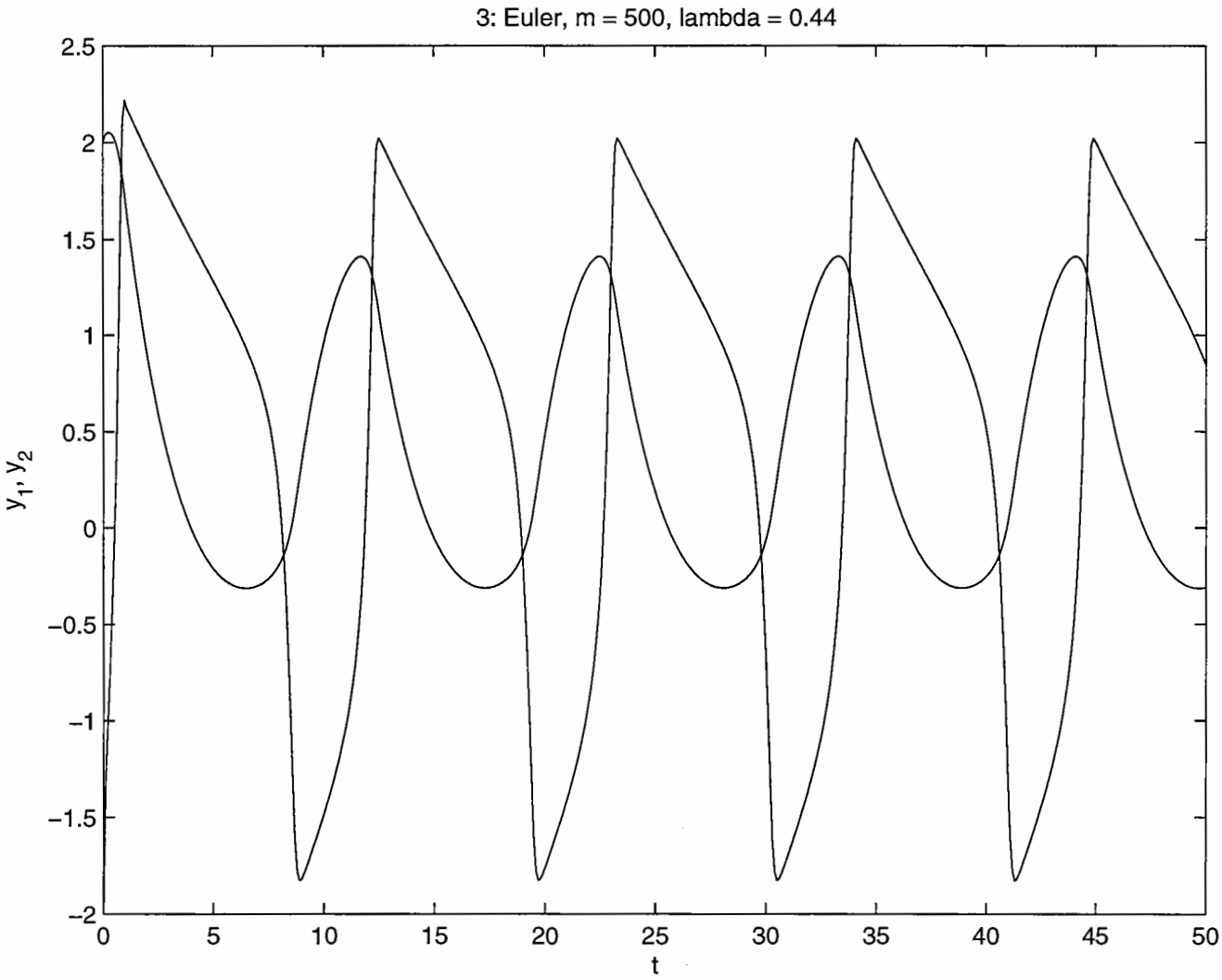

4: Runge–Kutta, m = 3000
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
x_2
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
x_1

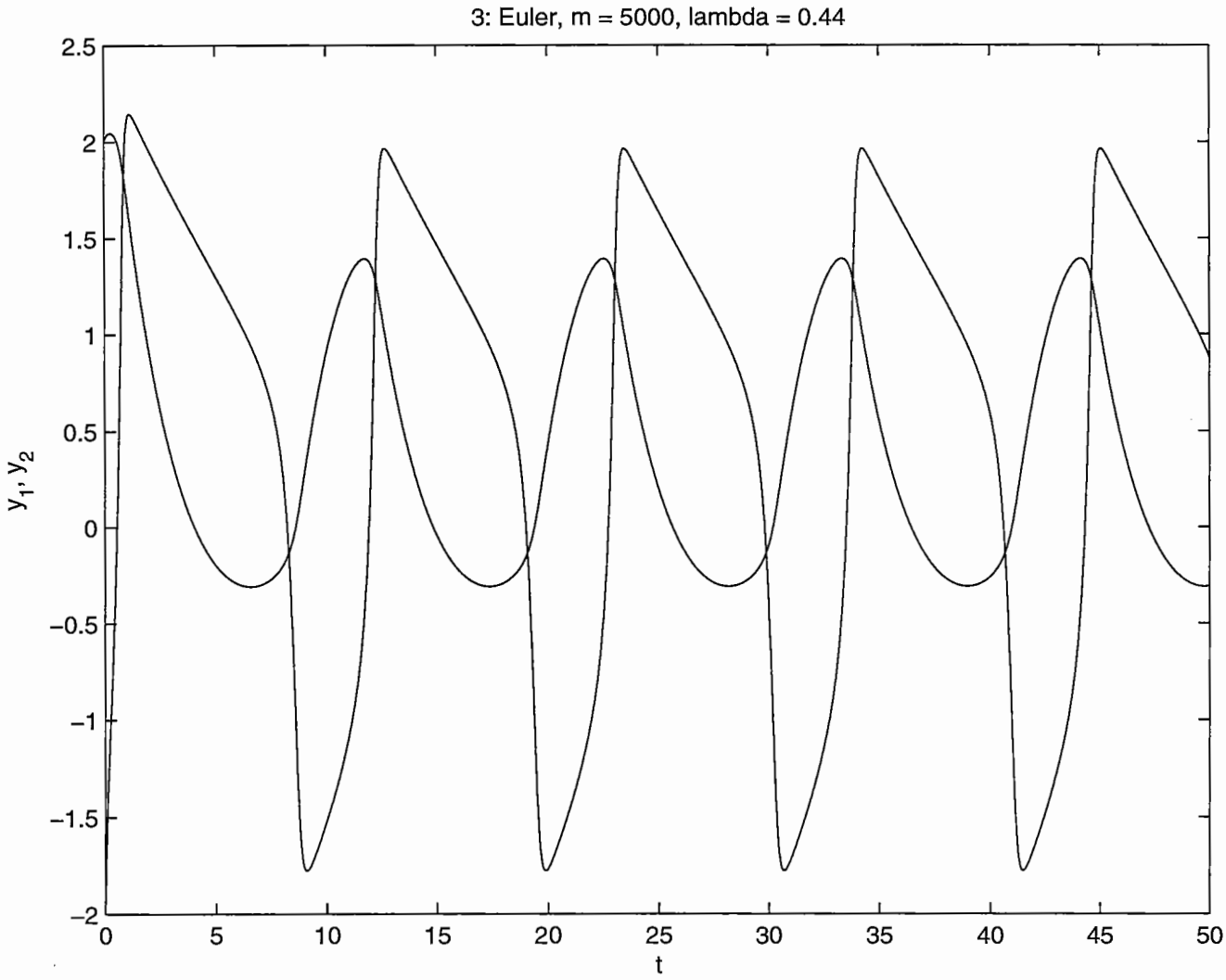

4: Runge–Kutta, m = 6000
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
x2
x1

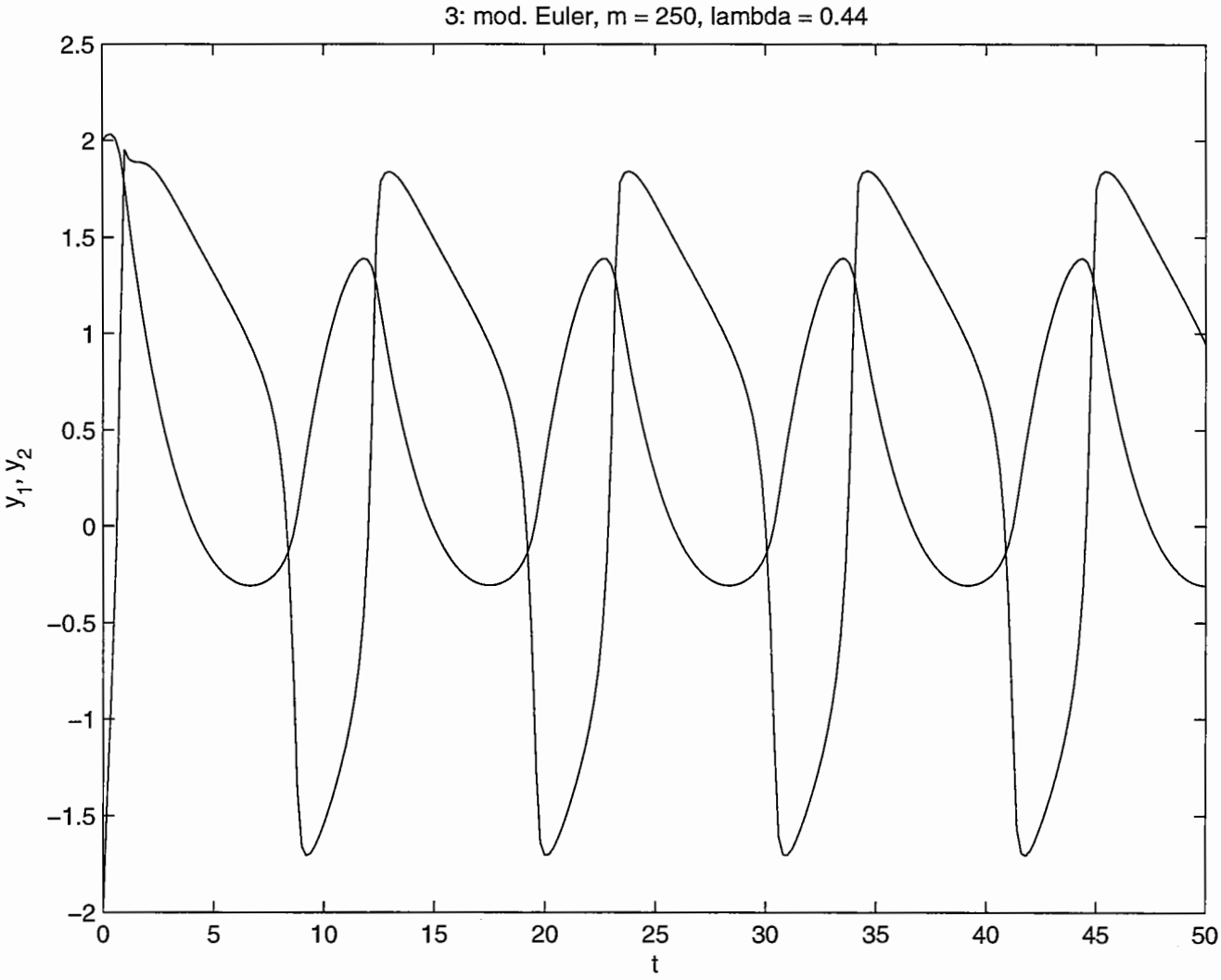

4: Runge–Kutta, m = 10000
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
x2
x1

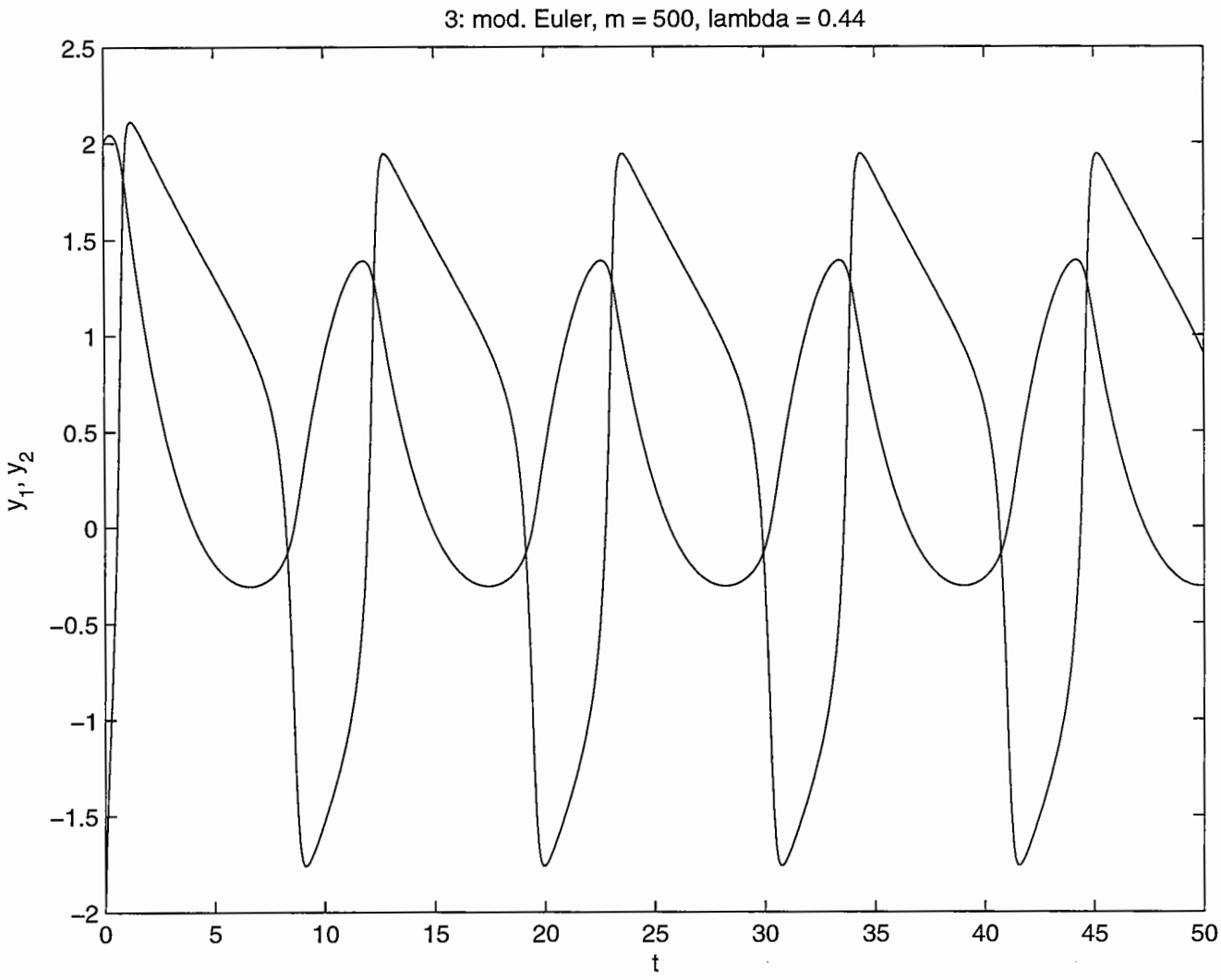

4: Runge–Kutta, m = 20000
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
x2
x1

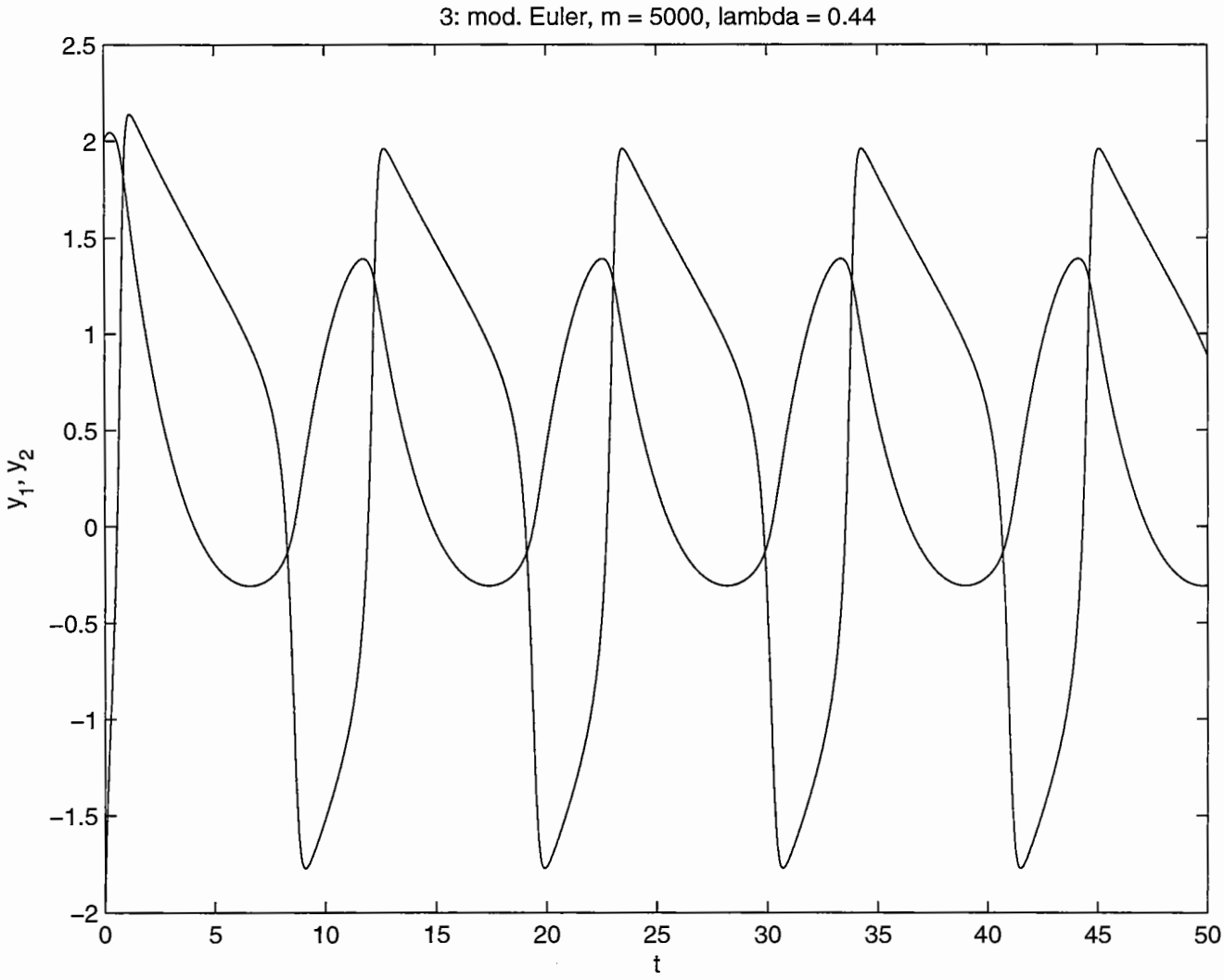

3: Euler, m = 250, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

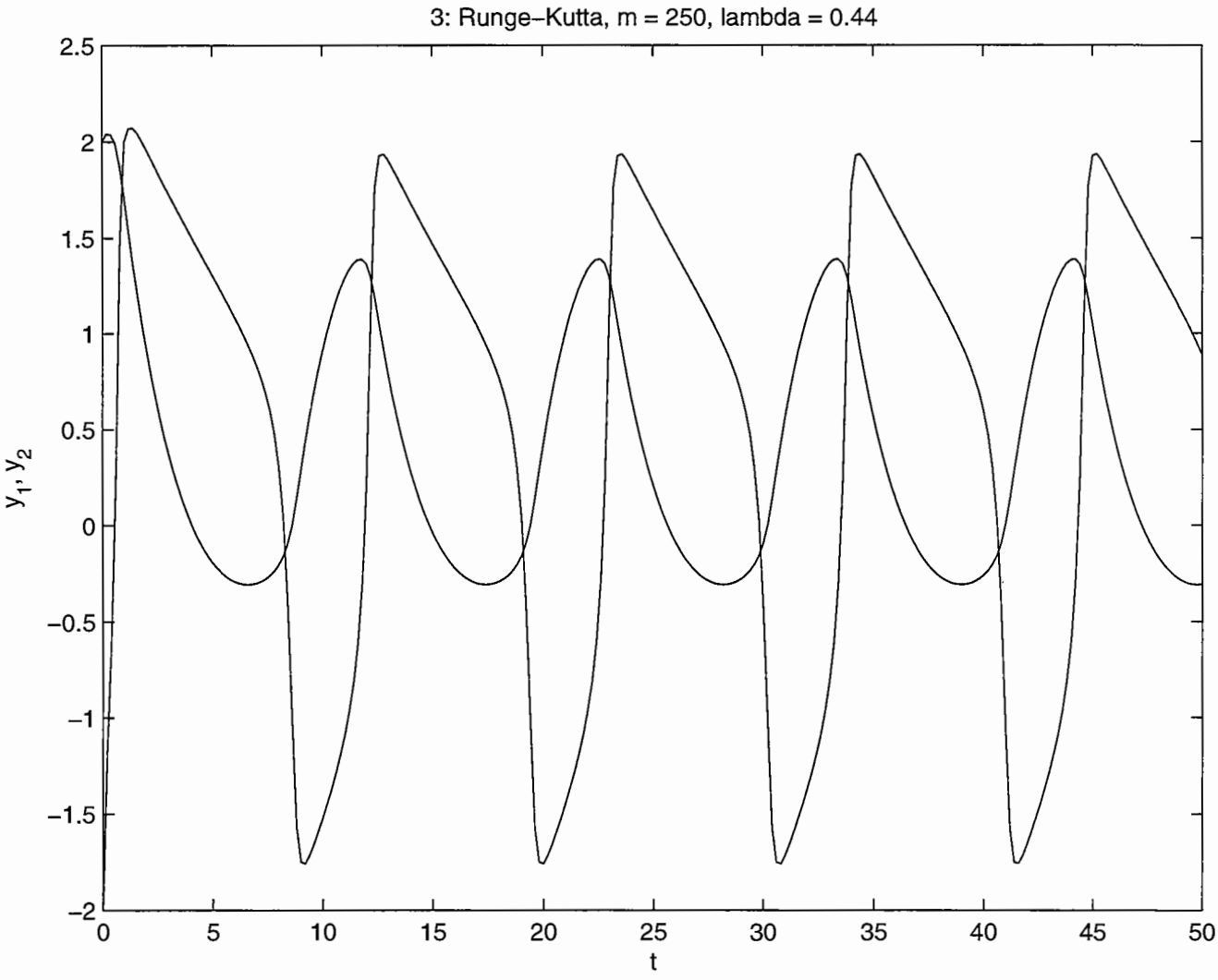

3: Euler, m = 500, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

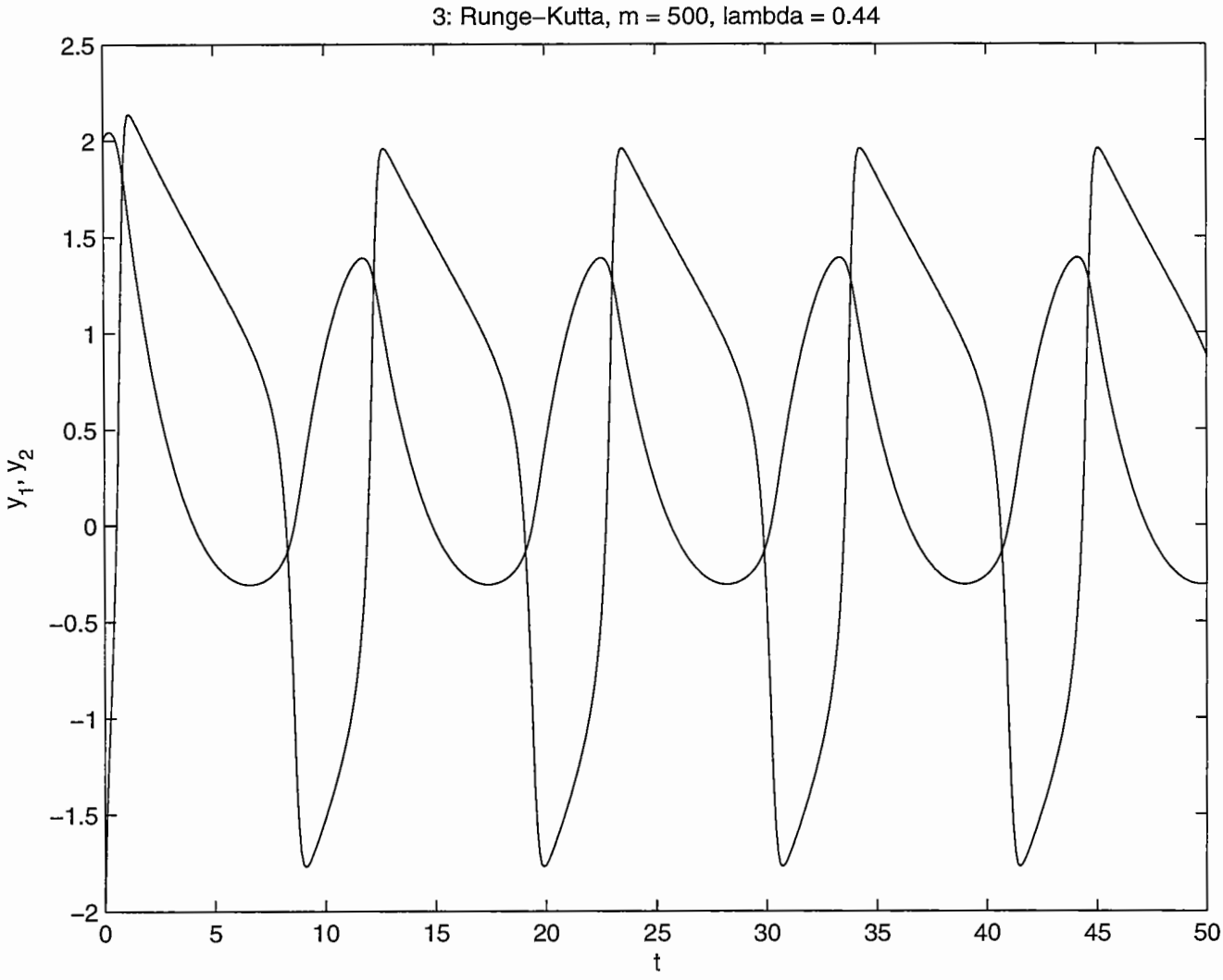

3: Euler, m = 5000, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

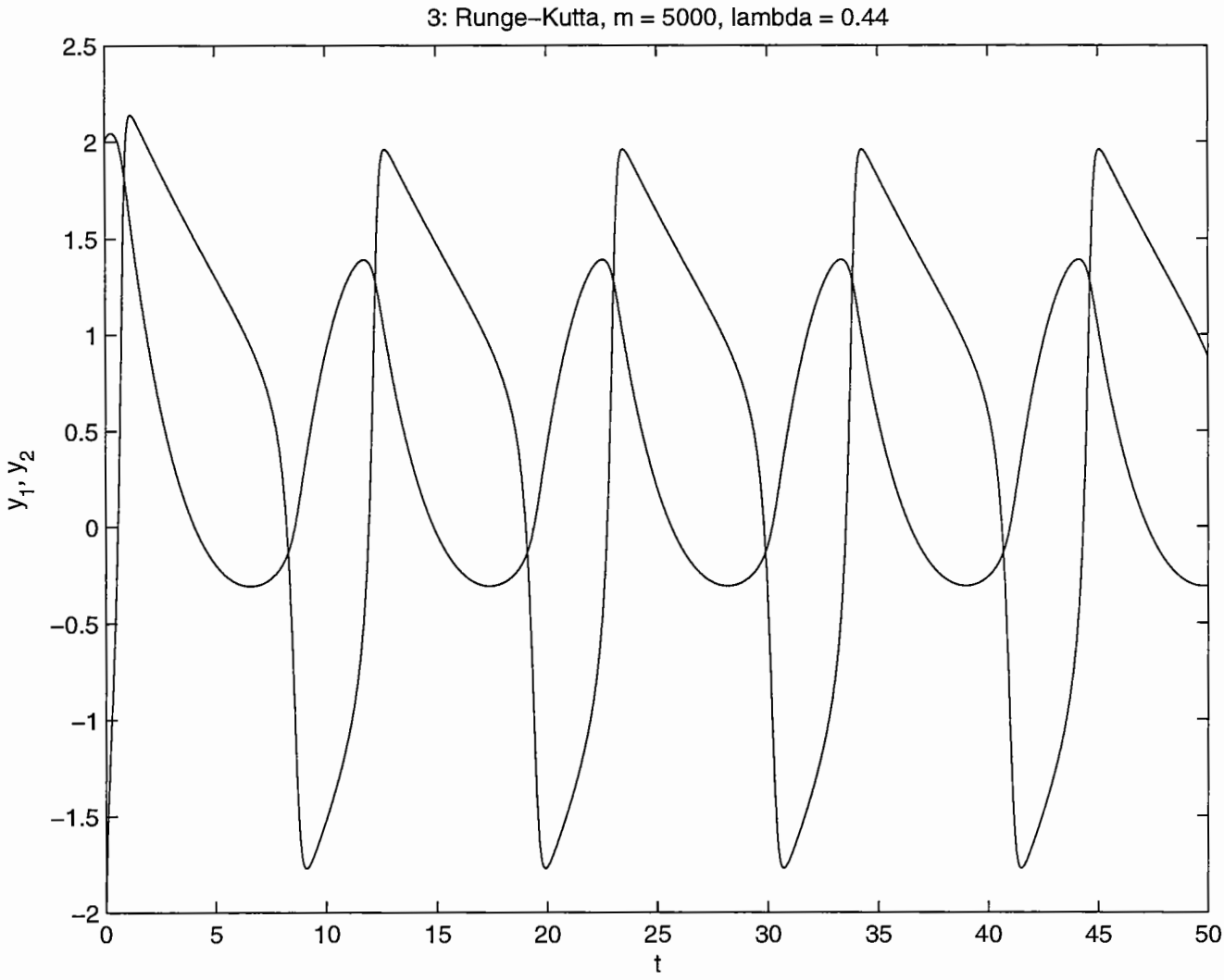

3: mod. Euler, m = 250, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y_1, y_2
0
5
10
15
20
25
30
35
40
45
50
t

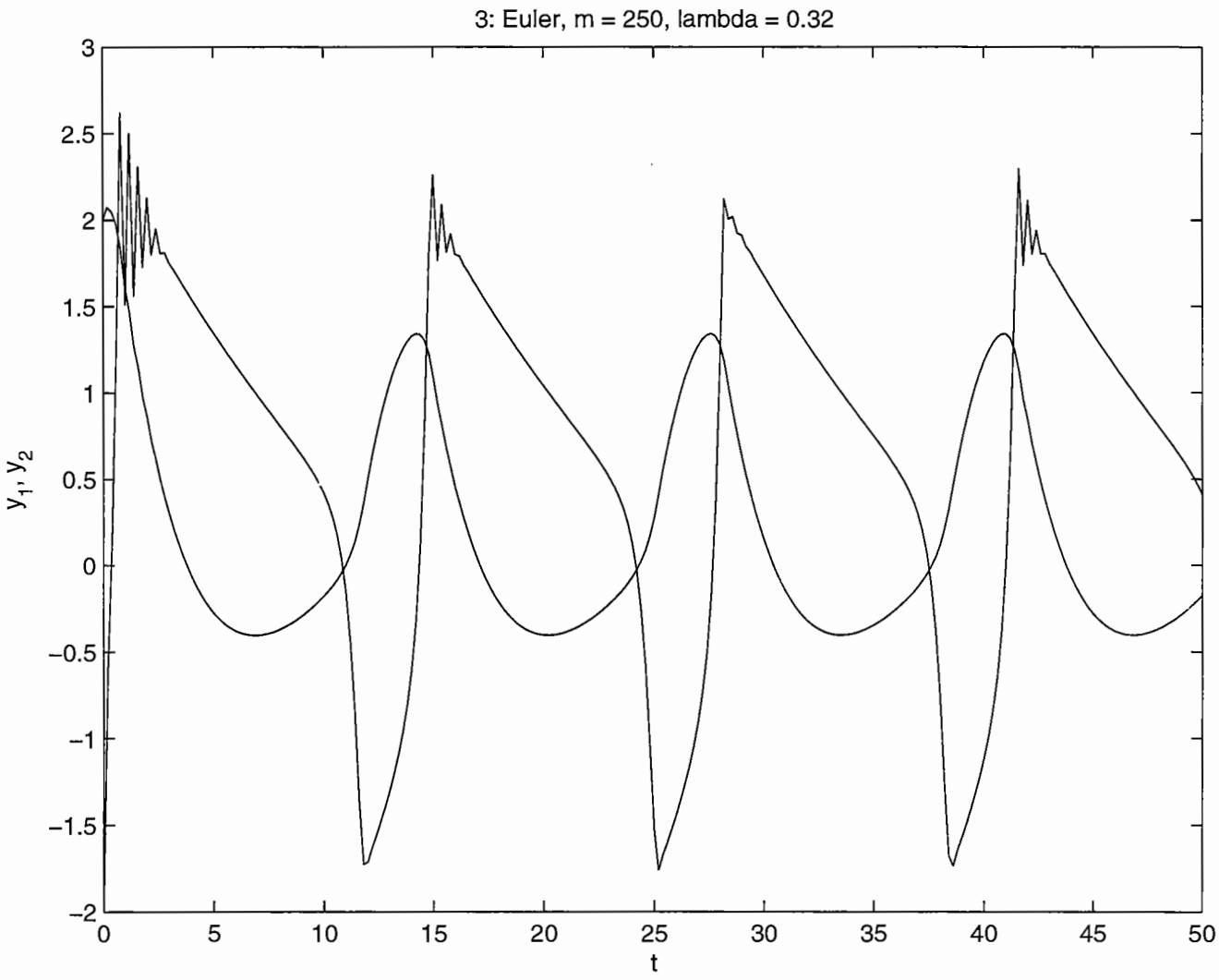

3: mod. Euler, m = 500, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

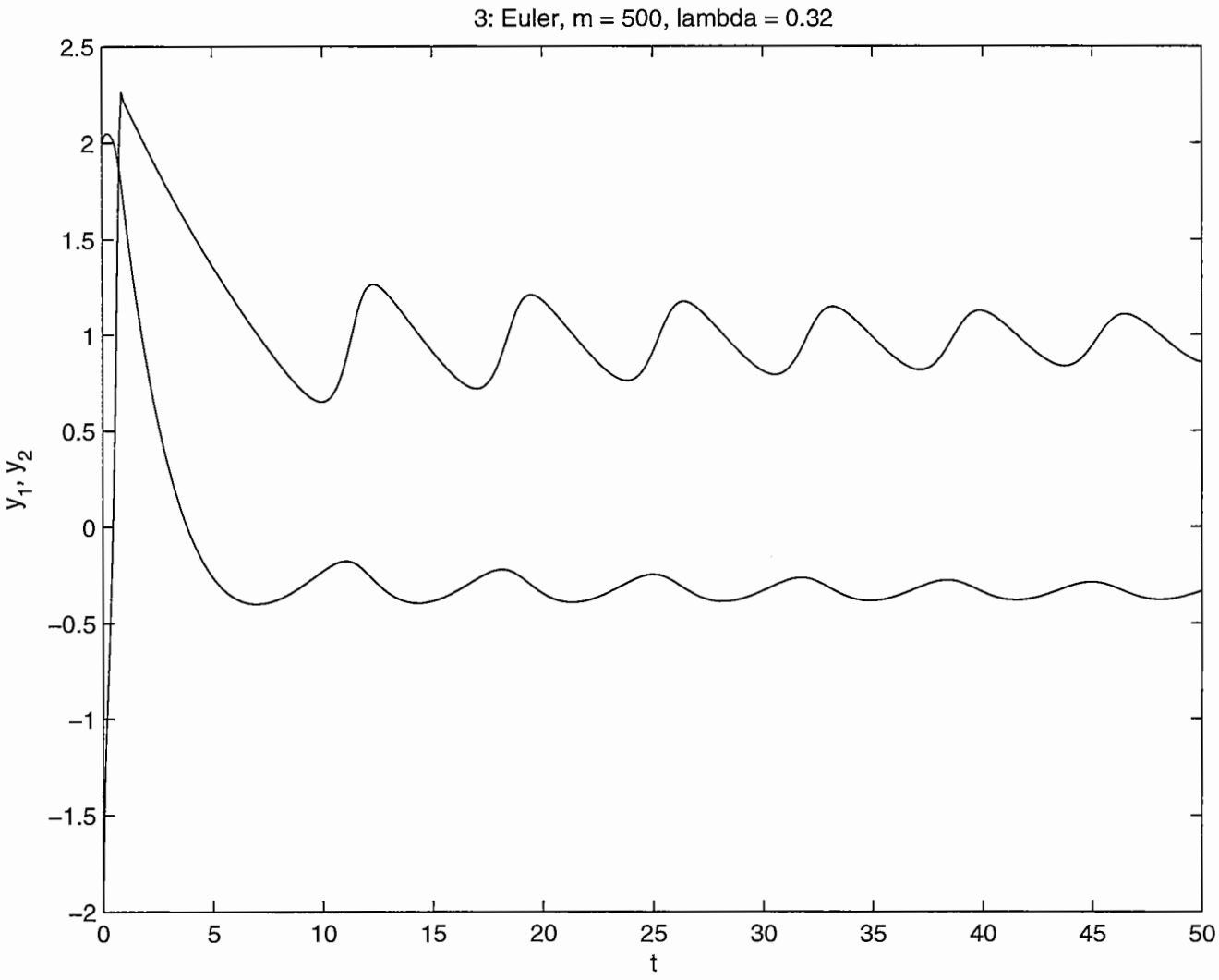

3: mod. Euler, m = 5000, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y_1, y_2
0
5
10
15
20
25
30
35
40
45
50
t

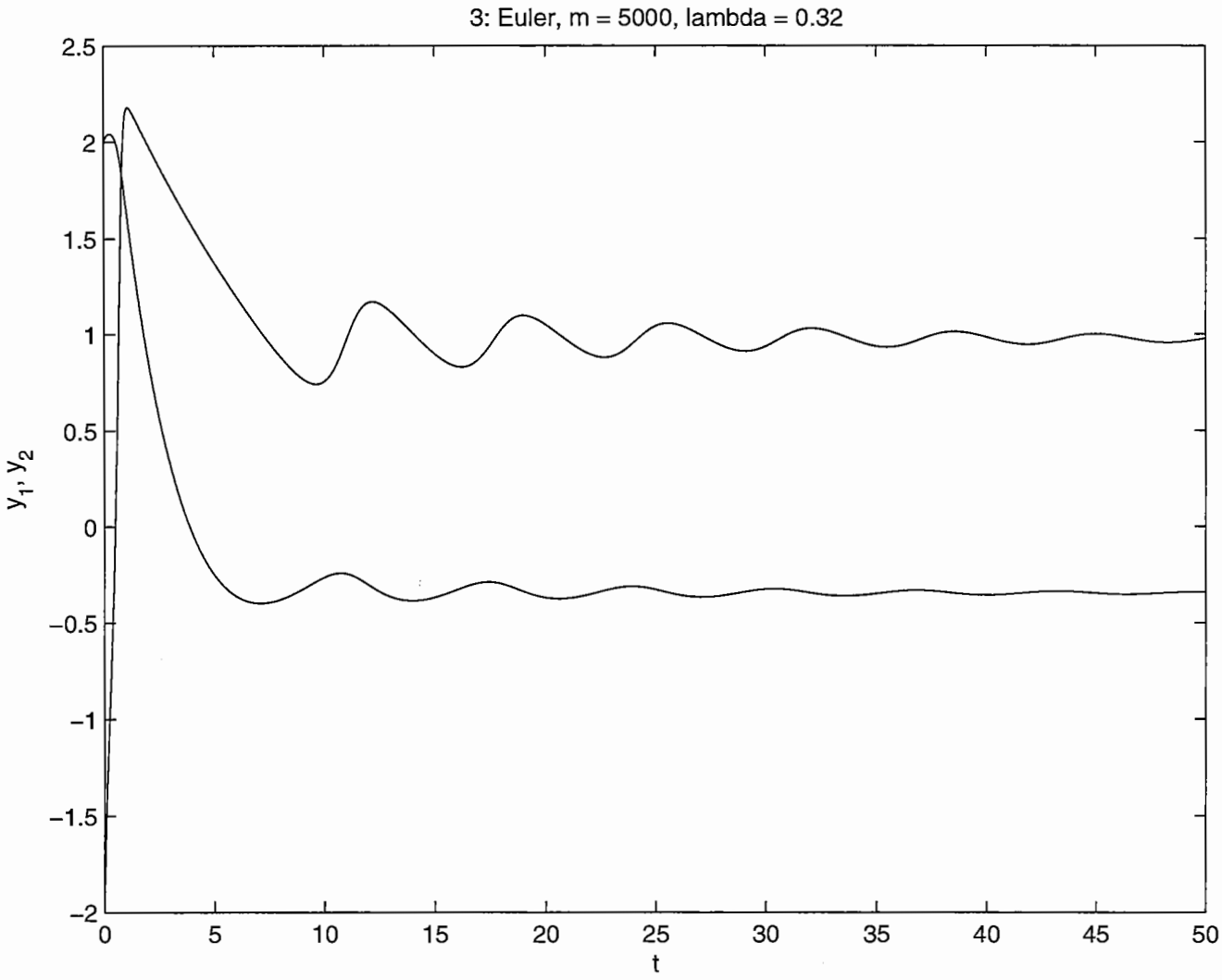

3: Runge–Kutta, m = 250, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y_1, y_2
0
5
10
15
20
25
30
35
40
45
50
t

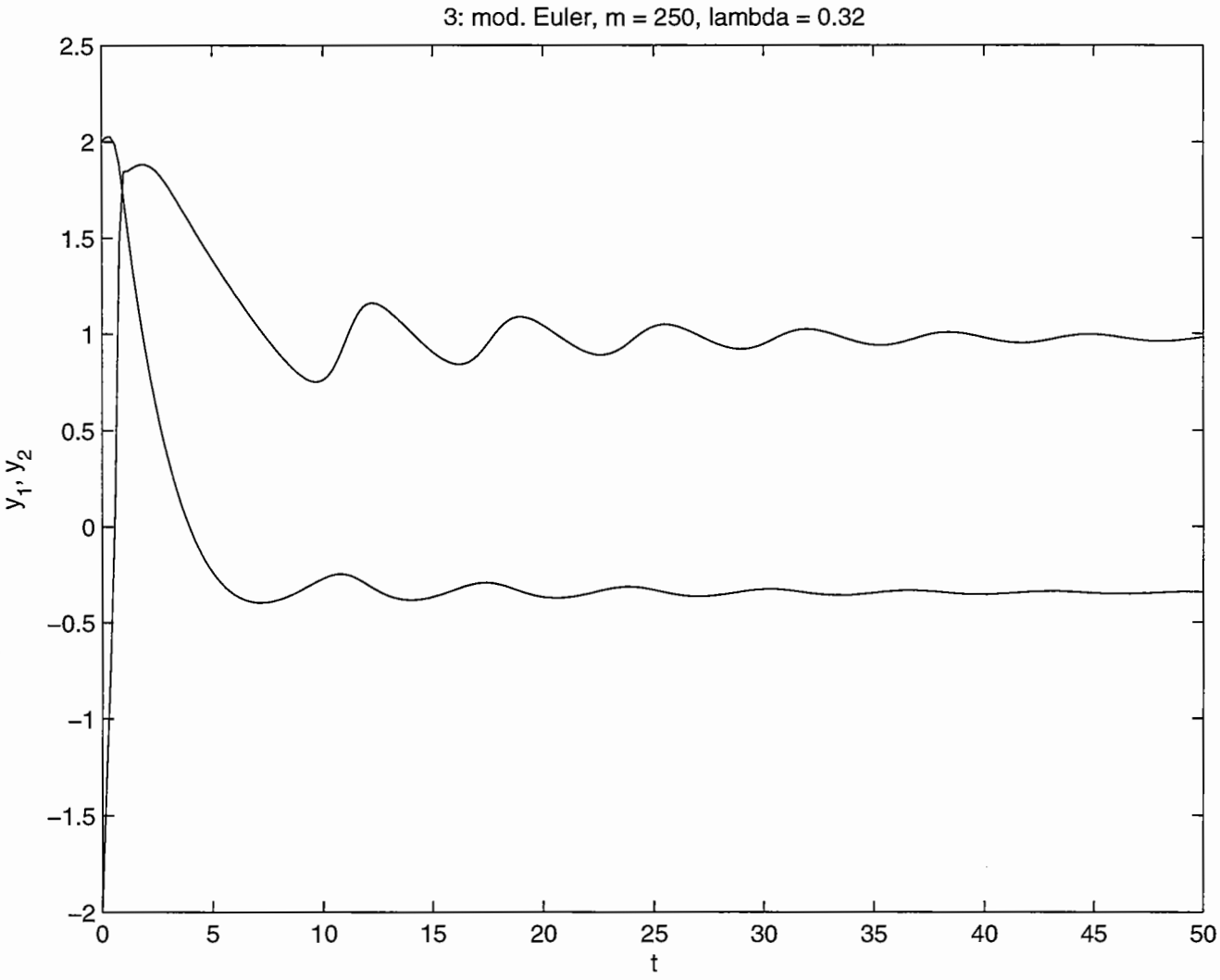

3: Runge–Kutta, m = 500, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

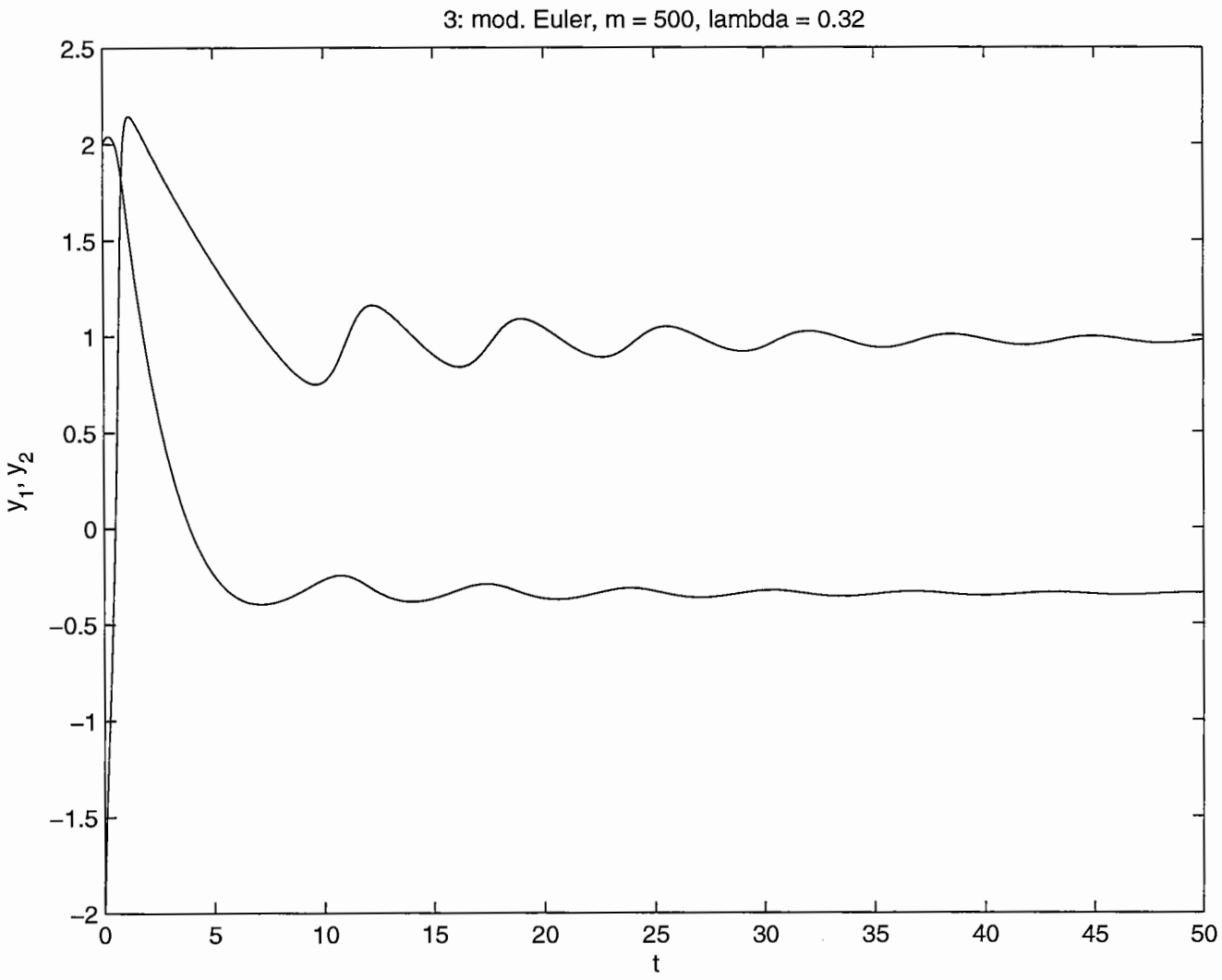

3: Runge–Kutta, m = 5000, lambda = 0.44
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

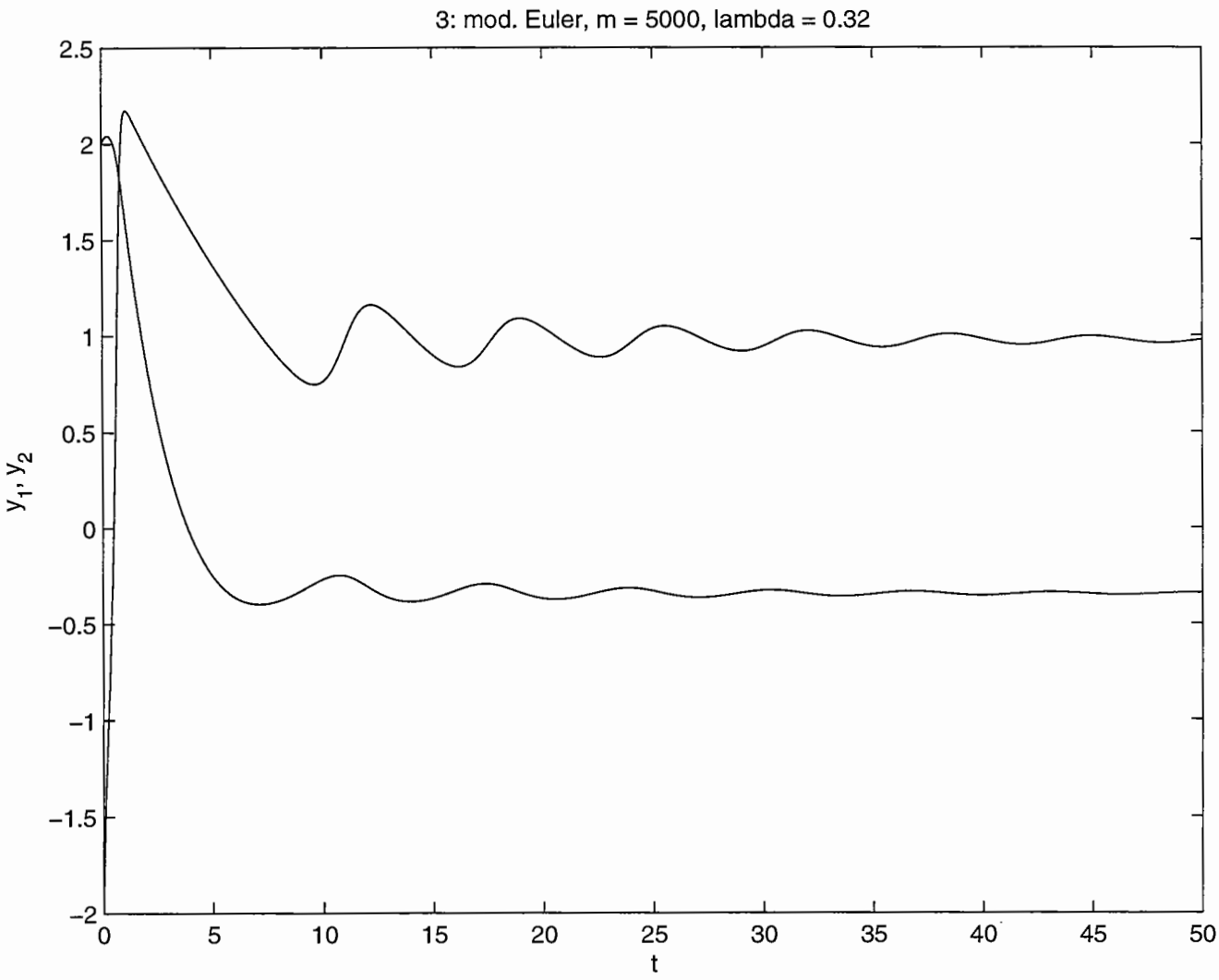

3: Euler, m = 250, lambda = 0.32
3
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

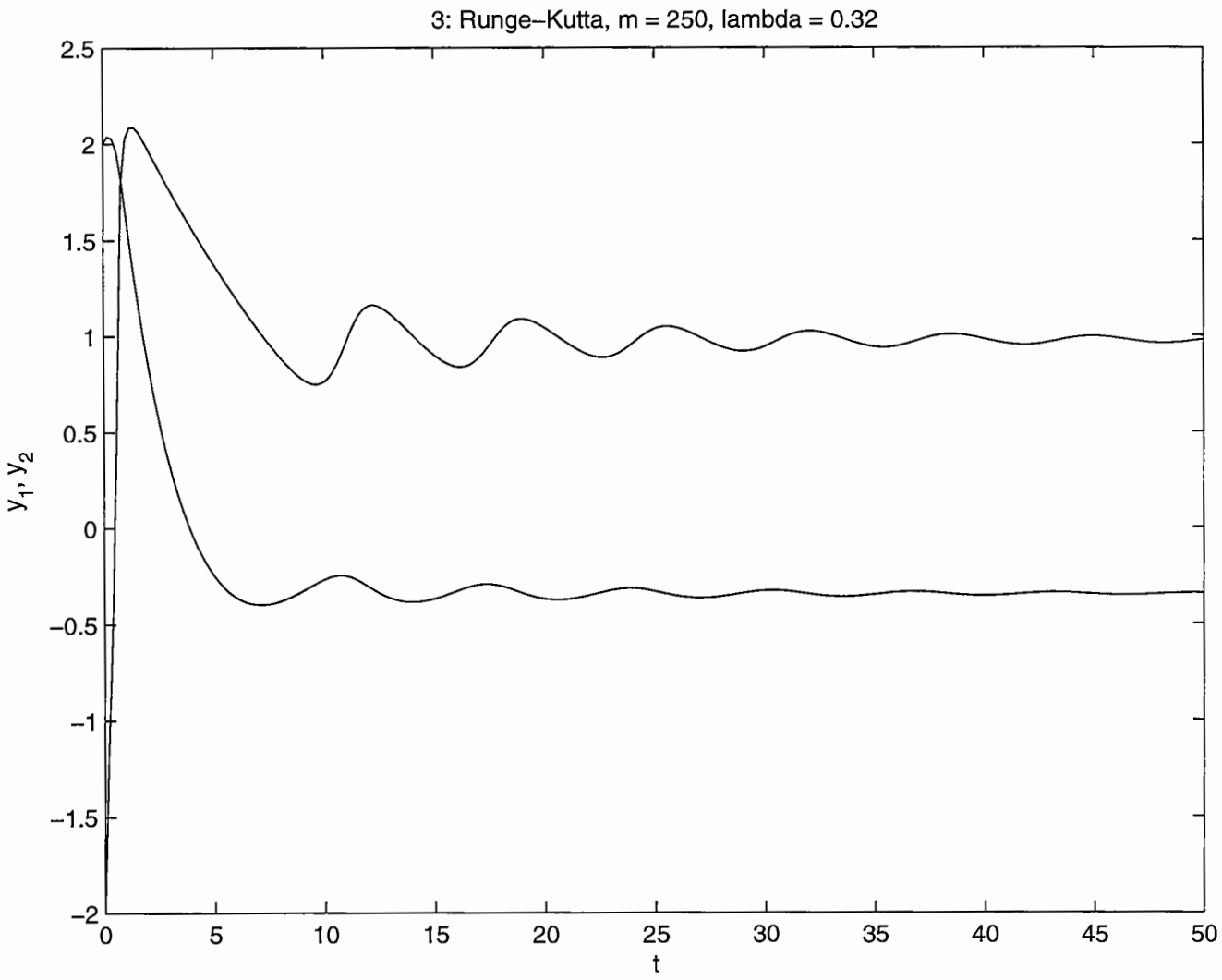

3: Euler, m = 500, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

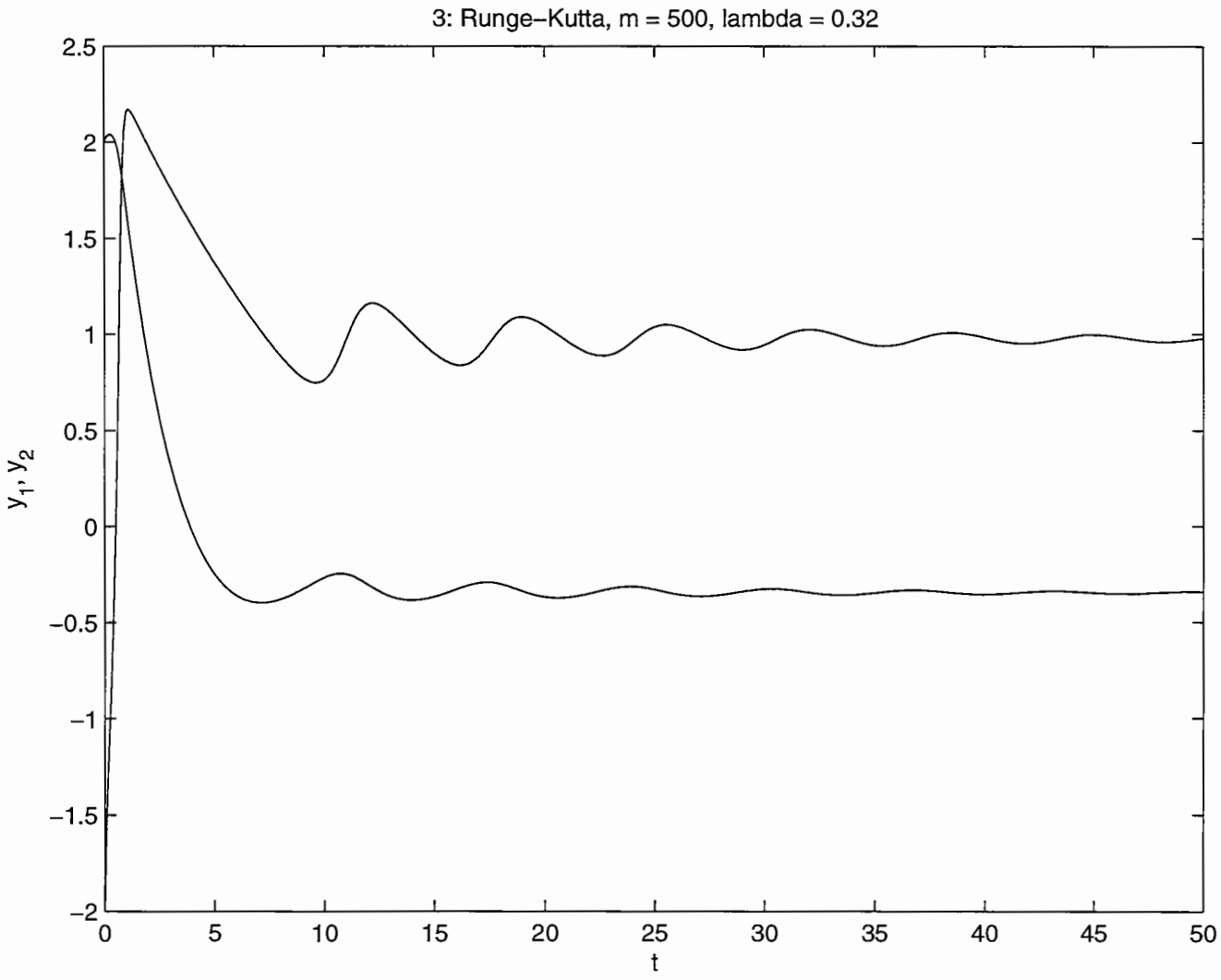

3: Euler, m = 5000, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

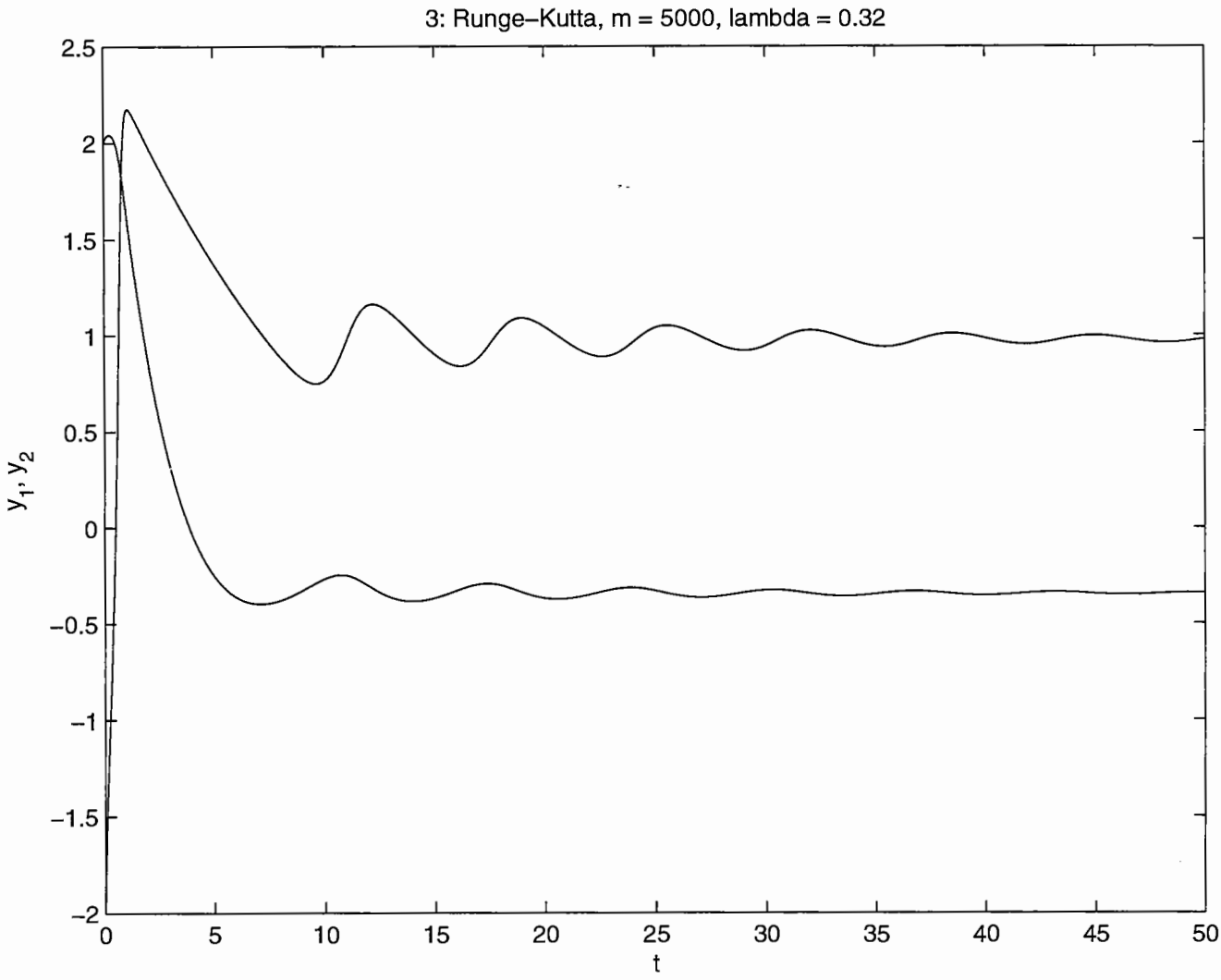

3: mod. Euler, m = 250, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y1, y2
0
5
10
15
20
25
30
35
40
45
50
t


3: mod. Euler, m = 500, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2


3: mod. Euler, m = 5000, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2


3: Runge–Kutta, m = 250, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y_1, y_2
0
5
10
15
20
25
30
35
40
45
50
t


3: Runge–Kutta, m = 500, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
y_1, y_2
0
5
10
15
20
25
30
35
40
45
50
t


3: Runge–Kutta, m = 5000, lambda = 0.32
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
5
10
15
20
25
30
35
40
45
50
t
y1, y2

Übungen zur Vorlesung Höhere Numerische Mathematik

Übungsblatt 12 , Abgabe: 22.07.2004, 8.00 Uhr

Aufgabe 41: (4 Punkte)
Die Lösung der AWA

$$y' = 1 + y\ , \quad y(0) = 1$$

ist $y(x) = 2e^x - 1$. Berechnen Sie Näherungswerte für $y(1)$ mit dem EULER-Verfahren zu den Schrittweiten $h_i = 2^{-i}$ $(i = 0, 1, 2, 3)$. Extrapolieren Sie diese Näherungswerte mit dem Extrapolationsverfahren aus § 24 mit $\gamma = 1$.

Aufgabe 42: (4 Punkte)
Es soll ein Einschrittverfahren folgender Bauart konstruiert werden:

$$y_{j+1} = y_j + h(\mu f(x_j, y_j) + (1 - \mu) f(x_{j+1}, y_{j+1}))$$

Wie muss man $\mu = \mu(\lambda h)$ wählen, damit $y' = \lambda y$, $\lambda \leq 0$, exakt integriert wird.
Zeigen Sie:

$$y_{j+1} = g(\lambda h) y_j\ , \qquad g(z) := \frac{1 + \mu z}{1 - (1 - \mu) z}$$

Aufgabe 43: (4 Punkte)
Gegeben sei das lineare MSV

$$y_{j+3} - y_{j+1} + \alpha(y_{j+2} - y_j) = h(\beta(f_{j+2} - f_j) + \gamma f_{j+1})$$

zur numerischen Lösung von $y' = f(x, y)$. Bestimmen Sie $\alpha, \beta, \gamma \in \mathbb{R}$ so, dass das MSV von 3. Ordnung ist. Berechnen Sie dann die Nullstellen des Polynoms $\rho(\lambda) = \lambda^3 + \alpha\lambda^2 - \lambda - \alpha$.

41)

$y' = 1+y =: f(x,y)$, $y(0)=1$ d.h. $x_0=0$, $y_0=1$

Euler: $y_{j+1} = y_j + h_i f(x_j, y_j)$

$= y_j + h_i(1+y_j) = h_i + (1+h_i)y_j$

$\Rightarrow y_j = (1+h_i)^j (y_0+1) - 1$

$= 2(1+h_i)^j - 1$

Diese Formel braucht man nicht notwendig. Die y_j lassen sich auch ~~[illegible]~~ iterativ berechnen.

$i=0$: $h_0=1$, $y(1) \approx y_1 = 2\,2^1 - 1 = 3$

$i=1$: $h_1=\frac{1}{2}$ $y(1) \approx y_2 = 2\left(\frac{3}{2}\right)^2 - 1 = \frac{7}{2} = 3{,}5$

$i=2$: $h_2=\frac{1}{4}$ $y(1) \approx y_4 = 2\left(\frac{5}{4}\right)^4 - 1 = \frac{497}{128} = 3{,}8828125$

$i=3$: $h_3=\frac{1}{8}$ $y(1) \approx y_8 = 2\left(\frac{9}{8}\right)^8 - 1 \approx 4{,}1315690 28$

Das sind die Startwerte T_{i0} der Extrapolation. Wegen $\gamma=1$ und $h_i = 2^{-i}$ gilt für die T_{ik} $(1 \le k \le i \le 3)$

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{2^k - 1}$$

Das ergibt folgendes Schema:

i	T_{i0}	T_{i1}	T_{i2}
0	3		
1	3,5	4	
2	3,8828125	4,265625	4,354166667
3	4,1315690 28	4,380325556	4,418559074	4,427757989

Exakter Wert: $y(1) = 2e - 1 = 4{,}436563656$

43) MSV:

$$y_{j+3} + \alpha y_{j+2} - y_{j+1} + \alpha y_j = h(\beta f_{j+2} + \gamma f_{j+1} - \beta f_j)$$

Diskretisierungsfehler:

$$\tau_h(x_{j+3}) = \frac{1}{h}(y_{j+3} + \alpha y_{j+2} - y_{j+1} - \alpha y_j) - (\beta f_{j+2} + \gamma f_{j+1} + \beta f_j)$$

$$= \frac{1}{h}(y_{j+3} + \alpha y_{j+2} - y_{j+1} - \alpha y_j) - \beta y'_{j+2} - \gamma y'_{j+1} + \beta y'_j$$

Taylorreihen:

$$y_{j+h} = y_j + hk\, y'_j + \frac{(hk)^2}{2} y''_j + \frac{(hk)^3}{6} y'''_j + O(h^4)$$

$$y'_{j+h} = y'_j + hk\, y''_j + \frac{(hk)^2}{2} y'''_j + O(h^3)$$

Einsetzen und nach Potenzen von h sortieren:

$$\tau_h(x_{j+3}) = (y_j + \alpha y_j - y_j - \alpha y_j)\frac{1}{h}$$
$$+ (3 + 2\alpha - 1 - 2\beta - \gamma + \beta)\, y'_j$$
$$+ \left(\frac{9}{2} + \frac{4}{2}\alpha - \frac{1}{2} - 2\beta - \gamma\right) h\, y''_j$$
$$+ \left(\frac{27}{6} + \frac{8}{6}\alpha - \frac{1}{6} - \frac{4}{2}\beta - \frac{1}{2}\gamma\right) h^2 y'''_j + O(h^3)$$

Der $\frac{1}{h}$ Term verschwindet. Für ein Verfahren 3. Ordnung müssen auch die Terme bis einschließlich h^2 verschwinden. Wir erhalten:

$$\text{I}: \quad 2\alpha - \gamma = -2$$
$$\text{II}: \quad 2\alpha - 2\beta - \gamma = -4$$
$$\text{III}: \quad \frac{4}{3}\alpha - 2\beta - \frac{1}{2}\gamma = -\frac{13}{3}$$

$$\text{II} - \text{III}: \quad \frac{2}{3}\alpha - \frac{1}{2}\gamma = \frac{1}{3} \quad | \cdot(-2) + \text{I}$$

$$\frac{2}{3}\alpha = -\frac{8}{3}$$

$$\Rightarrow \underline{\alpha = -4}$$