Mitschrift Einf. in die Num. Mathematik

Vorlesung SS10 Prof. Dr. H. Yserentant

Technische Universität Berlin

Vorwort

Dies ist eine Mitschrift, *kein* Skript! Für eventuelle Fehler übernimmt niemand die Verantwortung, sollten allerdings welche entdeckt werden freue ich mich über Hinweise.

WICHTIG: Auch dient die Mitschrift nicht als Ersatz für die Vorlesung, wichtige Bemerkungen und Erklärungen des Dozenten tauchen nicht auf. Ich empfehle euch also trotzdem regelmäßig zur Vorlesung zu gehen.

Tipp: Nicht immer gleich ausdrucken wenns online ist, im aktuellen Teil sind immer massenhaft Tippfehler die ich erst dann korrigiere wenn sie mir auffallen(d.h. wenn ich die Woche darauf damit arbeite).

Inhaltsverzeichnis

1. Rechnerarithmetik und Rundungsfehler

Normalisierte Gleitkommazahl

$$\pm \underbrace{\overset{=0}{0}\,,\overset{\downarrow\neq 0}{1}2345}_{\text{Mantisse}} \cdot 10^{\overbrace{-14}^{\text{Exponent}}}$$

$\pm 0,12345E-14$ und die Null!

Unsere Beispiele:

Dezimalsystem, 5 Dezimalstellen

$\pm 0,12345E+122$

bel großer Exponent. (schließt "underflow" und "overflow" aus)

IEEE-Standard

Dualsystem

32 bit (real) : Mantisse 23 bit, Exponent 8 bit, Vorzeichen 1 bit

64 bit (double) : Mantisse 52 bit, Exponent 11 bit, Vorzeichen 1 bit

Rundungsoperator

G =Menge der Gleitkommazahlen

$\mathrm{rd} : \mathbb{R} \to G$ Rundungsoperator

$|x - \mathrm{rd}(x)| \leq |x - g|, g \in G$

Beispiel 1.1:

$\mathrm{rd}(7,12341234...) = \mathrm{rd}(0.712341234...10^1) = 0.71234E+1$

$\mathrm{rd}(7,12345678...) = \mathrm{rd}(0.712345678...10^1) = 0.71235E+1$

$\mathrm{rd}(0,999999999...) = 0.10000E+1$

relativer Fehler:

$(x \neq 0)$

$\left|\frac{\mathrm{rd}(x)-x}{x}\right| \leq \left|\frac{5\cdot 10^{-6}}{\text{Mantisse}}\right| \leq 5 \cdot 10^{-5}$

(bei 5 Dezimalstellen)
$\mathrm{rd}(x) = x(1+\varepsilon)$, $|\varepsilon| \leq 5 \cdot 10^{-5}$ (gilt auch für $\mathrm{rd}(0) = 0$)

ideale Gleitkommaoperationen
x, y Gleitkommazahlen
$x \oplus y = \mathrm{rd}(x+y)$
$x \ominus y = \mathrm{rd}(x-y)$
$x \odot y = \mathrm{rd}(x \cdot y)$
$x \oslash y = \mathrm{rd}(x/y),\ y \neq 0$

reale Maschinenoperationen
x, y Gleitkommazahlen
$x \oplus y = (x+y)(1+\varepsilon_1)$
$x \ominus y = (x-y)(1+\varepsilon_2)$
$x \odot y = (x \cdot y)(1+\varepsilon_3)$
$x \oslash y = (x/y)(1+\varepsilon_4)$
$|\varepsilon_i| = |\varepsilon_i(x,y)| \leq \varepsilon^*$
ε^* Maschinengenauigkeit

Bemerkung:
Diese Beziehungen sind Grundlage jeder Rundungsfehleranalyse.

Vorsicht:
Das Assoziativgesetz der Addition und Multiplikation und das Distributivgesetz gelten nicht exakt.

Beispiel 1.2

Berechnung des Skalarprodukts $s = \sum_{i=1}^{n} x_i y_i$ nach dem Algorithmus.
$s_1 = x_1 y_1$
$s_{k+1} = s_k + x_{k+1} y_{k+1}, k = 1, \ldots, n-1$
$s = s_n$

Behauptung:
Für das berechnete Skalarprodukt $\widetilde{s}$ gilt

$\widetilde{s} = s + f, |f| \leq \underbrace{((1+(2+\varepsilon^*)\varepsilon^*)^n - 1)}_{=2n\varepsilon^* + O(\varepsilon^{*2})} \sum_{i=1}^{n} |x_i y_i|$

Beweis: Sei $\widetilde{s_k} = s_k + f_k$ leite Rekursion für den Fehler f_k her!

Es gilt:

$\widetilde{s_1} = x_1 y_1 (1+\varepsilon_1) = x_1 y_1 + f_1$ mit $f_1 = \varepsilon_1 x_1 y_1, |\varepsilon_1| \leq \varepsilon^*$

Weiter ist $\widetilde{s_k} = s_k + f_k$

$\widetilde{s}_{k+1} := (\widetilde{s_k} + x_{k+1} y_{k+1} \varepsilon_1)(1+\varepsilon_2) = s_{k+1} + f_{k+1}$

$f_{k+1} = \varepsilon_2 s_{k+1} + (1+\varepsilon_2) f_k + x_{k+1} y_{k+1} \varepsilon_2 (1+\varepsilon_2), |\varepsilon_1|, |\varepsilon_2| \leq \varepsilon^*$

Daraus folgt:

$|f_1| \leq \varepsilon^* \sum_{i=1}^{n} |x_i y_i| \leq (2+\varepsilon^*)\varepsilon^* \sum_{i=1}^{n} |x_i y_i|$

$|f_{k+1}| \leq (1+(2+\varepsilon^*)\varepsilon^*)|f_k| + \varepsilon^*(2+\varepsilon^*) \sum_{i=1}^{k+1} |x_i y_i|$

Mit Induktion folgt daraus

$|f_k| \leq \Big((1+(2+\varepsilon^*)\varepsilon^*)^k - 1\Big) \sum_{i=1}^{k} |x_i y_i|$

□

Interpolation:

Der relative Fehler kann sehr groß werden, falls

$|\sum_{i=1}^{n} x_i y_i| \ll \sum_{i=1}^{n} |x_i y_i|$

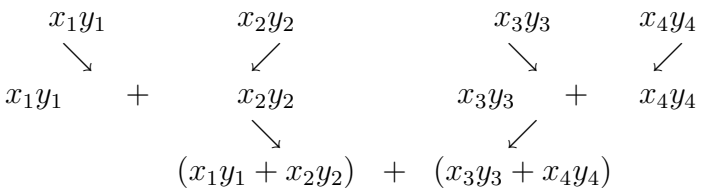
Es gilt aber

$\widetilde{s} = \sum_{i=1}^{n} (1+\varepsilon_i) x_i y_i$ mit $|\varepsilon_i| \leq (1+(2+\varepsilon^*)\varepsilon^*)^n - 1$

Das berechnete Ergebnis ist gleich dem exakten Ergebnis bei leicht gestörten Eingangsdaten.

Der Algorithmus ist rückwärtsstabil!

günstigere Strategie binärer Baum



2. Skalare Gleichungen

gegeben: $f : [a, b] \to \mathbb{R}$
gesucht: $\overline{x} \in [a, b]$ mit $f(\overline{x}) = 0$

Definition 2.1

Der Punkt $a \leq \overline{x} \leq b$ ist eine m-fache Nullstelle der stetigen Funktion $f : [a, b] \to \mathbb{R}$, wenn sich f in der Form $f(x) = (x - \overline{x})^m g(x), g(\overline{x}) \neq 0$ mit einer stetigen Funktion $g : [a, b] \to \mathbb{R}$ darstellen lässt.

Beispiel 2.2

$f(x) = 2x^3 - 4x^2 + 2x = 2x(x - 1)^2$
$\overline{x} = 0$ ist einfache Nullstelle,
$\overline{x} = 1$ ist doppelte Nullstelle
Bemerkung: Vorzeichenwechsel finden nur in Nullstellen ungerader Vielfachheit statt.

Satz 2.3

Die Funktion f sei in einem Intervall um den Punkt $x = \overline{x}$ m-mal stetig differenzierbar. Dann ist $\overline{x}$ eine m-fache Nullstelle von f, wenn $f(\overline{x}) = f'(\overline{x}) = ... = f^{(m-1)}(\overline{x}) = 0$ und $f^{(m)}(\overline{x}) \neq 0$ ist.

Beweis: Nach dem Taylorschen Satz ist

$$f(x) = \sum_{k=0}^{m-1} \frac{1}{k!} f^{(k)}(\overline{x})(x - \overline{x})^k + (x - \overline{x})^m g(x)$$

mit $g(x) = \frac{1}{(m-1)!} \int_0^1 (1 - \theta)^{m-1} f^{(m)}(\overline{x} + \theta(x - \overline{x}))\, d\theta$

und $g(\overline{x}) = \frac{1}{m!} f^{(m)}(\overline{x})$

□

Satz 2.4

Die Funktion f sei auf einer Umgebung ihrer m-fachen Nullstelle mindestens $(m + 1)$-mal stetig differenzierbar.
Dann ist $\overline{x}$ eine $(m - 1)$-fache Nullstelle von $f'(x)$ und eine einfache Nullstelle von $F(x) = \frac{f(x)}{f'(x)}$.

Beweis: Die Funktion f hat die Darstellung $f(x) = (x - \overline{x})^m g(x)$ mit der stetig differenzierbaren Funktion

$g(x) = \frac{1}{(m-1)!} \int_0^1 (1-\theta)^{m-1} f^{(m)}(\overline{x} + \theta(x - \overline{x})\, d\theta$

Daher ist $f'(x) = m(x - \overline{x})^{m-1} g(x) + (x - \overline{x})^m g'(x) = (x - \overline{x})^{m-1} h(x)$

$h(x) = m\, g(x) + (x - \overline{x}) g'(x)$. Wegen $h(\overline{x}) = m\, g(\overline{x}) \neq 0$ ist $\overline{x}$ daher eine $(m-1)$-fache Nullstelle von $f'(x)$.

Weiter gilt $F(x) = \frac{f(x)}{f'(x)} = \frac{(x-\overline{x})^m g(x)}{(x-\overline{x})^{m-1} h(x)} = (x - \overline{x}) \frac{g(x)}{h(x)}$

Wegen $g(\overline{x})/h(\overline{x}) \neq 0$ ist $\overline{x}$ daher einfache Nullstelle von F.

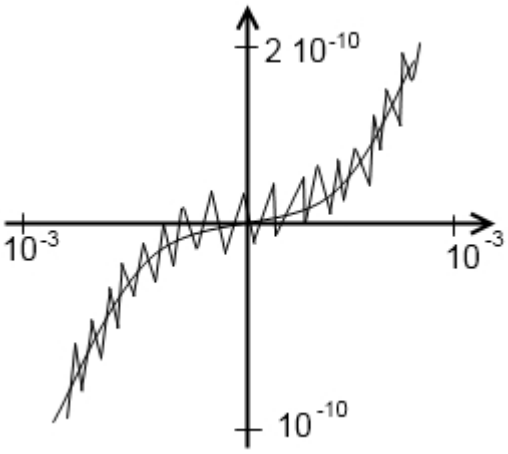
□

Soweit die Theorie! Zur Praxis:

Beispiel 2.5:

$f(x) = e^x - \left(1 + x + \frac{1}{2}x^2\right) \approx x^3 \left(\frac{1}{6} + \frac{x}{24} + \frac{x^2}{120}\right)$

$\overline{x} = 0$ ist eine 3-fache Nullstelle.



2.1 Das Lokalasieren von Nullstellen mit Vorzeichenwechsel

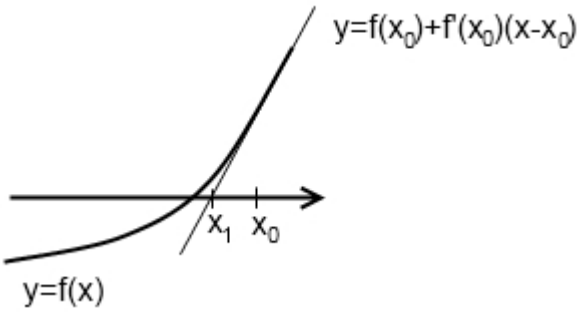
Mache vom Zwischenwertsatz Gebrauch!

Satz 2.6 (Bisektion)
Sei $f : [a, b] \to \mathbb{R}$ stetig und $f(a_0)f(b_0) < 0$. Die Folge $(x_n)_{n=0}^{\infty}$,
deren Glieder durch $x_n = (a_n + b_n)/2$ mit
$a_{n+1} = a_n, b_{n+1} = x_n$, falls $f(a_n)f(x_n) \leq 0$,
$a_{n+1} = x_n, b_{n+1} = b_n$, sonst
gegeben sind, erfüllt:
(i) $\lim_{n\to\infty} x_n = \overline{x}, f(\overline{x}) = 0$
(ii) $|x_n - \overline{x}| \leq \frac{1}{2}(b_n - a_n) = \frac{1}{2^{n+1}}(b_0 - a_0)$.
Jedes Intervall $[a_n, b_n]$ enthält die Nullstelle $\overline{x}$.

Vorteile:
-absolut sicher, falls man Rundungsfehler vernachlässigt
-Konvergenzgeschwindigkeit unabhängig von f
Nachteil:
-sehr langsame Konvergenz verglichen mit anderen Verfahren

2.2 Das Newton-Verfahren

Idee: Gegeben sei eine Näherung x_0 für $\overline{x}$. Linearisiere f in x_0 :
$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$
Berechne eine verbesserte(?) Näherung x_1 als Lösung der
linearisierten Gleichung:
$f(x_0) + f'(x_0)(x - x_0) = 0$
Das ergibt $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$
Berechne ausgehend von x_1 die Näherung $x_2, ...$



Newton-Verfahren

$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

iteratives Verfahren $x_{n+1} = f(x_n)$.
Setze x_n in die Iterationsvorschrift ein, um x_{n+1} zu erhalten.

Beispiel 2.8:

$f(x) = x^5 - 3 = 0, \overline{x} = \sqrt[5]{3}$
Die Anzahl der sicheren Stellen scheint in jedem Schritt um einen festen Faktor zuzunehmen.

Satz 2.9:

Sei $\overline{x}$ eine einfache Nullstelle der zweimal stetig differenzierbaren Funktion f.Es sei $\delta > 0$ und für alle x mit $|x - \overline{x}| \leq \delta$ sei $f(x)$ definiert und $f'(x) \neq 0$.
Sei $c_1 := \min\limits_{|x-\overline{x}|\leq\delta} |f'(x)| > 0, c_2 := \max\limits_{|x-\overline{x}|\leq\delta} |f''(x)| > 0$
Ist dann $|x - \overline{x}| \leq \min\left(\delta, \frac{c_1}{c_2}\right)$, so ist die Folge der Iterierten
$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, n = 0, 1, 2, ...$ wohldefiniert.
Die x_n konvergieren gegen $\overline{x}$. Es gilt:
$|x_{n+1} - \overline{x}| \leq \frac{1}{2}\frac{c_2}{c_1}|x_n - \overline{x}|^2$.

Beweis: Sei $|x_n - \overline{x}| \leq \min\left(\delta, \frac{c_1}{c_2}\right)$.
Dann ist x_{n+1} wohldefiniert und erfüllt die Gleichung
$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n)$.
Nach dem Taylorschen Satz ist
$0 = f(\overline{x}) = f(x_n) + f'(x_n)(\overline{x} - c_n) + \frac{1}{2}f''(\eta)(\overline{x} - x_n)^2$ oder
$x_{n+1} - \overline{x} = \frac{1}{2}\frac{f''(\eta)}{f'(x_n)}(x_n - \overline{x})^2$
Daraus folgt
$|x_{n+1} - \overline{x}| \leq \frac{1}{2}\frac{c_2}{c_1}|x_n - \overline{x}|^2$
Wegen $|x_n - \overline{x}| \leq \frac{c_1}{c_2}$ ist speziell
$|x_{n+1} - \overline{x}| \leq \frac{1}{2}\frac{c_2}{c_1}|x_n - \overline{x}||x_n - \overline{x}| \leq \frac{1}{2}|x_n - \overline{x}| \leq \min\left(\delta, \frac{c_1}{c_2}\right)$
Daher ist die Folge der Iterierten wohldefiniert und

$$\lim_{n\to\infty} x_n = \overline{x}$$

gesucht: Nullstelle $\overline{x}$ eine Funktion $f : [\overline{x} - \delta, \overline{x} + \delta] \to \mathbb{R}$
Annahmen: $f'(x) \neq 0$ auf $[\overline{x} - \delta, \overline{x} + \delta]$ (wesentlich!)
$f''(x)$ existiert und ist beschränkt auf $[\overline{x} - \delta, \overline{x} + \delta]$.
(kann eingeschränkt werden)

Newton Verfahren:
Linearisiere f um Näerung $\overline{x} - \delta < x_n < \overline{x} + \delta$:
$f(x) \approx f(x_n) + f'(x_n)(x - x_n)$
Löse $f(x_n) + f'(x_n)(x - x_n) = 0$ nach x auf.
$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
Folgerungen aus Satz 2.9
-Konvergenz, falls $|x_n - \overline{x}| < \delta, \delta$ klein genug.
-asymptotisch sehr schnelle Konvergenz
$|x_{n+1} - \overline{x}| \leq K|x_n - \overline{x}|^2 = \underbrace{K|x_n - \overline{x}|}_{\to 0}|x_n - \overline{x}|$

$\limsup_{n\to\infty} \frac{|x_{n+1}-\overline{x}|}{|x_n-\overline{x}|^2} \leq K$ "(lokal) quadratische Konvergenz"

Satz 2.10
Ist $f : [a, b] \to \mathbb{R}$ zweimal stetig differenzierbar, $f(a) < 0, f(b) > 0$ und $f'(x) > 0$ und $f''(x) \geq 0$ für alle $a \leq x \leq b$, so konvergiert das Newton Verfahren für Startwerte $\overline{x} \leq x_0 \leq b$ monoton gegen die einzige Nullstelle $\overline{x}$ von f in $[a, b]$.

Beweis: Man beweise, dass für die Schrittfunktion $f(x) := x - \frac{f(x)}{f'(x)}$ für $\overline{x} \leq x \leq b$ die Abschränkung $\overline{x} \leq \varphi(x) \leq x$ gilt.
Die Folge der Iterierten konvergiert dann monoton gegen eine (die!) Nullstelle von f.
□

Frage: Verhalten bei mehrfachen Nullstellen?

Beispiel 2.11
Die Funktion $f(x) = (\sin x)^3$ hat an der Stelle $\overline{x} = \pi = 3,14159265358...$ eine dreifache Nullstelle. Die Iterationsvorschrift des Newton-Verfahrens lautet hier:
$x_{n+1} = x_n - \frac{(\sin x_n)^3}{3(\sin x_n)^2 \cos x_n} = x_n - \frac{1}{3}\frac{\sin(x_n)}{\cos(x_n)}$

Beobachtung: Das Newtonverfahren konvergiert für mehrfache Nullstellen nicht mehr quadratisch.

Modifikation: Wende das Newtonverfahren auf die transformierte Funktion $F(x) = \frac{f(x)}{f'(x)}$ aus Satz 2.4 an:

$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} = x_n - m(x_n)\frac{f(x_n)}{f'(x_n)}$

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

Beispiel 2.12:

Für die Funktion $f(x) = (\sin x)^3$ erhält man $x_{n+1} = x_n - \frac{\tan x_n}{1+\tan x_n}$

Satz 2.13

Der Faktor $m(x_n)$ strebt gegen die Vielfachheit der Nullstelle $\overline{x}$.

Beweis: Sei j die Vielfachheit der Nullstelle. Dann folgt aus der Darstellung $f(x) = (x-\overline{x})^j g(x), g(\overline{x}) \neq 0$, wegen

$f'(x) = j(x-\overline{x})^{j-1}g(x) + (x-\overline{x})^j g'(x) == (x-\overline{x})^{j-1}[j\ g(x) + (x-\overline{x})g'(x)]$

$f''(x) = (x-\overline{x})^{j-2}[j(j-1)g(x) + j(x-\overline{x})g'(x) + (x-\overline{x})^2 g''(x)]$

$\lim\limits_{x\to\overline{x}} m(x) = \lim\limits_{x\to\overline{x}} \left(1 - \frac{f(x)f''(x)}{(f'(x))^2}\right)^{-1} = \left(1 - \frac{g(\overline{x})j(j-1)g(\overline{x})}{(j\ g(\overline{x}))^2}\right)^{-1} = j$

□

2.3 Allgemeine Iterationsverfahren

Verfahren der Form $x_{n+1} = \varphi(x_n), n = 0, 1, 2, \ldots\ x_0$ gegeben zur Brechnung eines Fixpunkts $\overline{x} = \varphi(\overline{x})$.

allgemeine Konvergenzaussage
Banachscher Fixpunktsatz

Lässt sich unter sehr viel allgemeineren Voraussetzungen beweisen!

Satz 2.14 (Banachscher Fixpunktsatz)

Die Funktion $\varphi : [a, b] \to \mathbb{R}$ bilde dass Intervall $[a, b]$ in sich ab.
Sie sei Lipschitz-stetig auf $[a, b]$, für alle $x, y \in [a, b]$ sei also
$|\varphi(x) - \varphi(y)| \leq L|x - y|$ mit einer festen Konstanten L. Es sei $L < 1$.

Dann besitzt φ genau einen Fixpunkt $\overline{x}$ in $[a,b]$. Für alle $x \in [a,b]$
strebt die durch $x_{n+1} = \varphi(x_n), n = 0,1,2,...$ definierte Folge gegen $\overline{x}$.
Es gelten die Abschätzungen $|x_n - x| \leq \frac{L}{1-L}|x_n - x_{n-1}|$,
$|x_n - \overline{x}| \leq \frac{L^n}{1-L}|x_1 - x_0|$.

Beweis:
Schritt 1: Für alle k ist $|x_{k+1} - x_k| \leq L^k|x_1 - x_0|$, denn:
$k = 0$ klar, $k \to k+1$:
$|x_{k+2} - x_{k+1}| = |\varphi(x_{k+1}) - \varphi(x_k)| \leq L|x_{k+1} - x_k| \leq L\, L^k|x_1 - x_0|$
Schritt 2: Für alle n und k ist $|x_{n+k} - x_n| \leq \frac{L^n}{1-L}|x_1 - x_0|$, denn:

$$|x_{n+k} - x_n| = |\sum_{j=0}^{k-1}(x_{n+j+1} - x_{n+j})| \leq \sum_{j=0}^{k-1}|x_{n+j+1} - x_{n+j}|$$

$$\leq \sum_{j=0}^{k-1} L^{n+j}(x_1 - x_0) = L^n(\sum_{j=0}^{k-1} L^j)(x_1 - x_0)$$

$$\leq L^n(\sum_{j=0}^{\infty} L^j)(x_1 - x_0) = L^n \tfrac{1}{1-L}(x_1 - x_0)$$

Schritt 3: Die Folge der x_n strebt gegen einen Fixpunkt $\overline{x}$ von φ, denn:
Nach Schritt 2 ist sie eine Cauchyfolge in $\mathbb{R}$. Sie konvergiert als solche
gegen ein $\overline{x} \in \mathbb{R}$. Da die x_n in $[a,b]$ liegen und $[a,b]$ abgeschlossen ist
liegt der Grenzwert $\overline{x}$ in $[a,b]$.
Die φ als Lipschitz-stetige Funktion stetig ist, ist

$$\overline{x} = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \varphi(x_n) = \varphi(\lim_{n\to\infty} x_n) = \varphi(\overline{x}),$$

$\overline{x}$ ist also Fixpunkt von φ.
Schritt 4: Für alle n ist $|x_n - \overline{x}| \leq \frac{L^n}{1-L}|x_1 - x_0|$,
$|x_n - \overline{x}| \leq \frac{L}{1-L}|x_n - x_{n-1}|$
denn: Die erste Abschätzung erhält man, wenn man in der Abschätzung aus Schritt 2 k gegen Unendlich streben lässt.
Die zweite Abschätzung erhält man aus den ersten, wenn man mit x_{n-1}
statt x_0 startet.
Schritt 5: Es gibt nur einen Fixpunkt, denn:
Ist $\overline{x} = \varphi(\overline{x})$, $\overline{\overline{x}} = \varphi(\overline{\overline{x}})$, so ist $|\overline{x} - \overline{\overline{x}}| = |\varphi(\overline{x}) - \varphi(\overline{\overline{x}})| \leq L|\overline{x} - \overline{\overline{x}}|$
was wegen $L < 1$ nur für $|\overline{x} - \overline{\overline{x}}| = 0$ möglich ist.
□

Bemerkung: Ist φ stetig differenzierbar, und $L = \max_{a\leq x\leq b} |\varphi'(x)|$
so gilt nach dem Mittelwertsatz für alle $x, y \in [a,b]$

$|\varphi(x) - \varphi(y)| = |\varphi'(\eta)(x-y)| \leq L|x-y|$.

Satz 2.15

Die Funktion φ sei auf einer Umgebung ihres Fixpunkts $\overline{x}$ q-mal stetig differenzierbar. Es sei $q \geq 2$ und $\varphi^{(k)}(\overline{x}) = 0, k = 1, \dots, q-1$. Dann ist die durch $x_{n+1} = \varphi(x_n), n = 0, 1, 2, \dots$ gegebene Folge für alle genügend nahe an $\overline{x}$ gelegenen Startwerte wohldefiniert und konvergent gegen $\overline{x}$. Überdies ist $|x_{n+1} - \overline{x}| \leq c|x_n - \overline{x}|^q$ mit einer festen Konstante c.

Beweis: Nach dem Taylorschen Satz ist

$\varphi(x) = \sum_{k=1}^{n-1} \frac{1}{k!}\varphi^{(k)}(\overline{x})(x-\overline{x})^k + \frac{1}{q!}\varphi^{(q)}(\eta)(x-\overline{x})^\eta$

$= \overline{x} + \frac{1}{q!}\varphi^{(q)}(\eta)(x-\overline{x})^q$

und damit für alle x aus einer festen δ-Umgebung um $\overline{x}$

$|\varphi(x) - \overline{x}| \leq c|x-\overline{x}|^q, c = \max_{|\eta-\overline{x}|\leq\delta} \left|\frac{1}{q!}\varphi^{(q)}(\eta)\right|$

Daraus folgt für alle x mit

$|x-\overline{x}| \leq \min\left(\delta, \left(\frac{1}{2c}\right)^{\frac{1}{q-1}}\right) =: \delta^*$

die Abschätzung

$|\varphi(x) - \overline{x}| \leq \underbrace{c|x-\overline{x}|^{\eta-1}}_{\leq c\frac{1}{2c}}|x-\overline{x}| \leq \frac{1}{2}|x-\overline{x}|$

Damit bildet φ die δ^*-Umgebung von $\overline{x}$ in sich ab. Für $|x_0 - \overline{x}| \leq \delta^*$ sind die x_n wohldefiniert und es ist $|x_{n+1} - \overline{x}| \leq \left(\frac{1}{2}\right)^n |x_1 - \overline{x}|$.
Die x_n konvergieren also gegen $\overline{x}$ und es ist
$|x_{n+1} - \overline{x}| = |\varphi(x_n) - \overline{x}| \leq c|x_n - \overline{x}|^q$
□

Satz 2.16(modifiziertes Newtonverfahren)

Die Funktion sei auf einer Umgebung ihrer m-fachen Nullstelle hinreichend oft stetig differenzierbar. Dann konvergiert die durch
$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}, n = 0, 1, 2, \dots$
gegebene Folge für alle hinreichend nahe bei $\overline{x}$ gelegenen x_0 quadratisch gegen $\overline{x}$.

Beweis: Untersuche $\varphi(x) = x - m\frac{f(x)}{f'(x)}$
Wegen $f(x) = (x-\overline{x})^m g(x), g(\overline{x}) \neq 0$, gilt

$\varphi(x) = x - m\frac{(x-\overline{x})^m g(x)}{m(x-\overline{x})^{m-1}g(x)+(x-\overline{x})^m g'(x)}$
$= x - (x-\overline{x})h(x), h(x) = \frac{mg(x)}{mg(x)+(x-\overline{x})g'(x)}.$
Es ist $f(\overline{x}) = \overline{x}$ und wegen $\varphi'(x) = 1 - h(x) - (x-\overline{x})h'(x)$
$\varphi'(\overline{x}) = 1 - h(\overline{x}) = 0$
□

2.4 Die Regula falsi

Gesucht:
schnell konvergierendes, ableitungsfreies Verfahren
Idee: Ersetze im Newtonverfahren
$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
die Ableitung $f'(x_n)$ durch den Differenzenquotienten
$\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$
Dies ergibt die Regula falsi
$x_{n+1} = x_n - \frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}f(x_n)$

Beispiel 2.17

$f(x) = x^5 - 3, x_0 = 1, x_1 = 2$
$x_{10} = 1,245730940$

Satz 2.18

Die Funktion f sei auf einer δ-Umgebung ihrer Nullstelle $\overline{x}$ zweimal stetig differenzierbar.
Für $|x-\overline{x}| < \delta$ sei $|f'(x)| \geq c_1 > 0, |f''(x)| \leq c_2 \neq 0$.
Die Umgebung U von $\overline{x}$ sie definiert durch $U = \left\{ x \;\middle|\; |x-\overline{x}| \leq \min\left(\delta, \frac{c_1}{c_2}\right) \right\}$
Liegen dann x_0 und x_1 in U, so brich die durch
$x_{n+1} = x_n - \frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}f(x_n)$,$n = 1, 2, 3, ...$
entweder mit der exakten Lösung $\overline{x}$ ab, oder ist für alle n definiert und konvergiert gegen $\overline{x}$. In beiden Fällen liegen alle x_n in U.
Es gibt eine Nullfolge (Δ_n) mit $|x_n - \overline{x}| \leq \frac{2c_1}{c_2}\Delta_n$ für alle berechenbaren x_n und $\lim\limits_{n\to\infty} \frac{\Delta_{n+1}}{\Delta_n^q} = 1, q = \frac{1+\sqrt{5}}{2} = 1,6180...$

Beweis: Nach dem Hauptsatz der Differential- und Integralrechnung ist wegen $f(\overline{x}) = 0$

$f(x) = (x - \overline{x})g(x), g(x) = \int_0^1 f'(\overline{x} + \theta(x - \overline{x}))d\theta$

Ist nun $x_{n-1}, x_n \in U$ und $x_{n-1} \neq x_n$, so ist x_{n+1} wegen
$\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}} = f'(\eta) \neq 0$ wohldefiniert und mit Zwischenstellen
α_n und β_n.

$x_{n+1} - \overline{x} = x_n - \frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})} f(x_n) - \overline{x}$

$= \frac{(x_n-\overline{x})(f(x_n)-f(x_{n-1}))-(x_n-x_{n-1})f(x_n)}{f(x_n)-f(x_{n-1})}$

$= \frac{(x_{n-1}-\overline{x})f(x_n)-(x_n-\overline{x})f(x_{n-1})}{f(x_n)-f(x_{n-1})}$

$= \frac{(x_{n-1}-\overline{x})(x_n-\overline{x})g(x_n)-(x_n-\overline{x})(x_{n-1}-\overline{x})g(x_{n-1})}{f(x_n)-f(x_{n-1})}$

$= \frac{g(x_n)-g(x_{n-1})}{f(x_n)-f(x_{n-1})}(x_{n-1} - \overline{x})(x_n - \overline{x})$

$= \frac{g'(\alpha_n)(x_n-x_{n-1})}{f'(\beta_n)(x_n-x_{n-1})}(x_{n-1} - \overline{x})(x_n - \overline{x})$

$= \frac{g'(\alpha_n)}{f'(\beta_n)}(x_{n-1} - \overline{x})(x_n - \overline{x})$

Wegen $g'(x) = \int_0^1 f(\overline{x} + \theta(x - \overline{x}))\theta \, d\theta$ ist

$|g'(\alpha_n)| \leq \int_0^1 c_2\theta d\theta = \frac{1}{2}c_2$, so dass wir die Abschätzung

$|x_{n+1} - \overline{x}| \leq \frac{c_2}{2c_1}|x_{n-1} - \overline{x}||x_n - \overline{x}|$ $\qquad (*)$

erhalten. Wegen $|x_{n-1} - \overline{x}| \leq \frac{c_1}{c_2}$ und $x_n \in U$ folgt aus $(*)$
$|x_{n+1} - \overline{x}| \leq \frac{1}{2}|x_n - \overline{x}|$ und damit $x_{n+1} \in U$. Ist $x_{n+1} \neq x_n$,
so ist die Ausgangssituation wiederhergestellt und man kann
die Iteration fortsetzen. Ist hingegen $x_{n+1} = x_n$, so ist
$0 = \frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})} f(x_n) = \frac{1}{f'(\eta_n)} f(x_n)$, also x_n Nullstelle,
d.h. $x_n = \overline{x}$. Das Verfahren bricht also mit der exakten Nullstelle ab.

Um die Konvergenz nachzuweisen, setzen wir $\Delta_0 = \frac{c_2}{2c_1}|x_0 - \overline{x}|$,
$\Delta_1 = \frac{c_2}{2c_1}|x_1 - \overline{x}|$ und $\Delta_{n+1} = \Delta_{n-1}\Delta_n, n = 1, 2, 3, \ldots$.
Durch Induktion folgt dann aus $(*)$ $|x_n - \overline{x}| \leq \frac{2c_1}{c_2}\Delta_n$.
Wegen $x_0, x_1 \in U$ ist $\Delta_0 \leq \frac{1}{2}, \Delta_1 \leq \frac{1}{2}$. Durch Induktion folgt
erst $\Delta_n \leq \frac{1}{2}$ und dann $\Delta_n \leq \left(\frac{1}{2}\right)^n$. Damit ist die Konvergenz
der x_n gegen $\overline{x}$ bewiesen.

Wir setzen nun $\delta_n = \ln\Delta_n$. Wegen $\Delta_{n+1} = \Delta_{n-1}\Delta_n$ ist dann
$\delta_{n+1} = \delta_n + \delta_{n-1}, n = 1, 2, 3, \ldots$
Die allgemeine Lösung dieser Differenzengleichung ist

$\delta_n = a\left(\frac{1+\sqrt{5}}{2}\right)^n + b\left(\frac{1-\sqrt{5}}{2}\right)^n$
mit reellen Konstanten a und b. Diese Lösung findet zudem mit
dem Ansatz $\delta_n = \lambda^n$. Wegen
$\left|\frac{1-\sqrt{5}}{2}\right| < 1$ folgt daraus, $\lim_{n\to\infty}\left(\delta_{n+1} - \frac{1+\sqrt{5}}{2}\delta_n\right) = 0$
Setzt man $q = \frac{1+\sqrt{5}}{2}$, ist dies gleichbedeutend mit
$\lim_{n\to\infty} \frac{\Delta_{n+1}}{\Delta_n^q} = 1$
□

Aufwand Newton Verfahren
2 Funktionsauswertungen pro Schritt
Aufwand Regula falsi
1 Funktionsauswertung pro Schritt
Doppelschritt:
$\lim_{n\to\infty} \frac{\Delta_{n+2}}{\Delta_n^{q^2}} = \lim_{n\to\infty} \frac{\Delta_{n+2}}{\Delta_{n+1}^q}\left(\frac{\Delta_{n+1}}{\Delta_n}\right)^2 = 1$
effektive Konvergenzordnung
$\left(\frac{1+\sqrt{5}}{2}\right)^2 = 2,6180...$

Die Regula falsi ist asymptotisch schneller als das Newtonverfahren!

3. Lineare Gleichungssysteme

3.1 Normen

Normen dienen dazu, die Länge von Vektoren und damit indirekt den Abstand zwischen Vektoren zu messen.

Definition 3.1:
Eine reellwertige Funktion $\|.\|$ auf dem $\mathbb{R}^n$ heißt Norm auf dem $\mathbb{R}^n$,
falls für alle $x, y \in \mathbb{R}^n$ und alle $\alpha \in \mathbb{R}$ gilt:
(i) $\|x\| \geq 0, \|x\| = 0 \iff 0$
(ii) $\|\alpha x\| = |\alpha|\|x\|$ (Homogenität)
(iii) $\|x + y\| \leq \|x\| + \|y\|$ (Dreiecksungleichung)
Die Größe $\|x - y\|$ ist dann der Abstand von x und y bezüglich dieser Norm.

Beispiel 3.2:
Sei $x \in (x_1, \dots, x_n) \in \mathbb{R}^n$. Dann ist die euklidische Norm oder auch 2-Norm von x definiert durch:
$\|x\| = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$
Sie induziert den euklidischen Abstand $\|x-y\| = (\sum_{i=1}^n (x_i - y_i)^2)^{\frac{1}{2}}$
Mit dem Skalarprodukt $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$ gilt $\|x\| = \langle x, x\rangle^{\frac{1}{2}}$.
Für alle Vektoren x, y ist $\langle x, y\rangle \leq \|x\|\|y\|$
(Schwarzsche Ungleichung)

Beispiel 3.3:
Die Maximumnorm auf dem $\mathbb{R}^n$ ist definiert durch $\|x\|_\infty = \max_{i=1,..,n} |x_i|$

Beispiel 3.4:
Die 1-Norm auf dem $\mathbb{R}^n$ ist gegeben durch $\|x\|_1 = \sum_{i=1}^n |x_i|$.

Bemerkung: Es gibt viele andere (und wesentlich kompliziertere) Normen auf dem $\mathbb{R}^n$.
So ist z.B. für jede symmetrische positiv definite $(n \times n)$-Matrix Q durch

$$\langle x, y\rangle = x^T Q y$$

ein Skalarprodukt auf dem $\mathbb{R}^n$ gegeben, das die Norm
$\|x\| = \langle x, x\rangle^{\frac{1}{2}}$ induziert.

Satz 3.5:
Sind $\|.\|_1$ und $\|.\|_2$ zwei beliebige (!) Normen auf dem $\mathbb{R}^n$ so gibt es Konstanten $c_1, c_2 > 0$ mit $c_1\|x\|_1 \leq \|x\|_2 \leq c_2\|x\|_1$ für alle $x \in \mathbb{R}^n$.

Beweis: Es genügt die behauptung für die Maximumnorm und eine weitere beliebige Norm $\|.\|$ auf dem $\mathbb{R}^n$ zu zeigen.

Wir zeigen zunächst die Ungleichung $\|x\| \leq c\|x\|_\infty$. seien $e_1, \dots, e_n$ die EInheitsvektoren. Nach der Dreiecksungleichung ist dann
für alle $x \in \mathbb{R}^n$ $\|x\| = \|\sum_{i=1}^n x_i e_i\| \leq \sum_{i=1}^n \|x_i e_i\| = \sum_{i=1}^n |x_i|\|e_i\|$,

also $\|x\| \leq c\|x\|_\infty$ mit $c = \sum_{i=1}^{n} \|e_i\|$.

Zum Beweis der Umkehrung betrachten wir zunächst die Funktion
$f : \mathbb{R}^n \to \mathbb{R} : x \mapsto \|x\|$. Für alle $x, y \in \mathbb{R}^n$ ist
$\|x\| = \|(x-y)+y\| \leq \|x-y\| + \|y\|$, und damit $\|x\| - \|y\| \leq \|x-y\|$.
Entsprechend ist $\|y\| - \|x\| \leq \|x-y\|$, also $|\|x\| - \|y\|| \leq \|x-y\|$
Damit ist für alle $x, y \in \mathbb{R}^n$ $|f(x) - f(y)| = |\|x\| - \|y\|| \leq \|x-y\| \leq c\|x-y\|_\infty$,
so dass f stetig ist.
Die Menge $K = \{x \in \mathbb{R}^n \mid \|x\|_\infty = 1\}$ ist kompakt. Daher gibt es ein
$x^* \in K$ mit $f(x^*) \leq f(x)$ für alle $x \in K$. Dies bedeutet $f(x^*) \leq \left\| \frac{1}{\|x\|_\infty} x \right\|$
$= \frac{1}{\|x\|_\infty} \|x\|$ oder mit $c' = f(x^*) = \|x^*\| > 0, c_1\|x\|_\infty \leq \|x\|$ für alle $x \neq 0$
□

Bemerkung: Praktisch ist diese Aussage oft von nur sehr geringem Wert, da der Satz keine Information über die Größe der Konstanten liefert und die Konstanten in der oberen und unteren Abschätzung leicht um Größenordnungen differieren können.

Beispiel 3.6:
Für alle $x \in \mathbb{R}^n$ ist $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$ mit der Maximumnorm $\|x\|_\infty$ aus Beispiel 3.3, der 1-Norm aus Beispiel 3.4 und der euklidischen Norm $\|x\|_2$.
Umgekehrt gilt:
$\|x\|_2 \leq \sqrt{n}\|x\|_\infty$
$\|x\|_1 \leq \sqrt{n}\|x\|_2$
$\|x\|_1 \leq n\|x\|_\infty$
Alle Konstanten sind optimal.

Definition 3.7:
Einer gegebenen Vektornorm $\|.\|$ ist die Matrixnorm $\|A\| := \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ auf dem Raum der reellen $(n \times x)$-Matrizen zugeordnet:
Für alle $x \in \mathbb{R}^n$ ist damit $\|Ax\| \leq \|A\|\|x\|$, und es gibt $x^* \neq 0$ mit $\|Ax^*\| = \|A\|\|x^*\|$.

Eigenschaften:
(i) $\|A\| \geq 0, \|A\| = 0 \iff A = 0$
(ii) $\|\alpha A\| = |\alpha| \|A\|$
(iii) $\|A + B\| \leq \|A\| + \|B\|$
(deshalb Norm)
(iv) $\|AB\| \leq \|A\| \|B\|$
invertierbare Matrizen:
$\|x\| = \|A^{-1}Ax\| \leq \|A^{-1}\| \|Ax\|$, daher:
$\frac{1}{\|A^{-1}\|} \|x\| \leq \|Ax\| \leq \|A\| \|x\|$
Existiert A^{-1}, so ist
$\frac{1}{\|A^{-1}\|} = \min\limits_{x \neq 0} \frac{\|Ax\|}{\|x\|}$

Satz 3.8

Der Maximumnorm aus Beispiel 3.3 entspricht die Matrixnorm
$\|A\|_\infty = \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}|$ der Matrix $A = (a_{ij})$, die Zeilensummennorm.
Die der 1-Norm aus Beispiel 3.4 zugeordnete Norm ist die
Spaltensummennorm $\|A\|_1 = \max\limits_{j=1,\dots,n} \sum\limits_{i=1}^{n} |a_{ij}|$

Beweis: (für die Maximumnorm)
Für alle $x \in \mathbb{R}^n$ ist $\|Ax\|_\infty = \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij} x_j|$
$\leq \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}| |x_j| \leq \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}| \|x\|_\infty$
Damit ist $\|A\|_\infty \leq \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}|$. Um zu zeigen, dass die Abschätzung
scharf ist, nehmen wir an, dass $\max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}| = \sum\limits_{j=1}^{n} |a_{kj}|$
ist und definieren wir x^* Komponentenweise durch
$x_j^* = \begin{cases} 1, \text{falls } a_{kj} \geq 0 \\ -1, \text{ falls } a_{kj} < 0 \end{cases}$
Dann ist $\|Ax^*\|_\infty \geq |\sum\limits_{j=1}^{n} a_{kj} x_j^*| = \sum\limits_{j=1}^{n} |a_{kj}| = (\max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}|) \|x_j^*\|_\infty$
also $\|A\|_\infty \geq \max\limits_{i=1,\dots,n} \sum\limits_{j=1}^{n} |a_{ij}|$.
□

Beispiel 3.9:

Die Matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ hat die Zeilensummennorm:

$\|A\|_\infty = \max\{|1| + |2|, |3| + |4|\} = 7$

Ihre Inverse $A^{-1} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$

hat die Zeilensummennorm

$\|A^{-1}\|_\infty = \max\{|-2| + |1|, \left|\frac{3}{2}\right| + \left|-\frac{1}{2}\right|\} = 3$

Satz 3.10

Der euklidischen Norm als Vektornorm ist die Spektralnorm als Matrixnorm zugeordnet. Die Spektralnorm der Matrix A ist die Wurzel aus dem größten Eigenwert der Matrix $A^\top A$.

Beweis: Wegen $\left(A^\top A\right)^\top = A^\top (A^\top)^\top = A^\top A$ ist $A^\top A$ eine symmetrische Matrix.

Es gibt daher nach dem Spektralsatz der Linearen Algebra eine Orthonormalbasis $x_1, \dots, x_n$ des $\mathbb{R}^n$ mit $A^\top \mathrm{Ax} := \lambda_i x_i$.

Es ist $\lambda_i \langle x_i, x_i\rangle = \langle x_i, \lambda_i x_i\rangle = \left\langle x_i, A^\top A x_i\right\rangle = \langle Ax_i, Ax_i\rangle = \|Ax_i\|_2^2 \geq 0$

Für den Vektor $x = \sum_{i=1}^{n} \alpha_i x_i$ gilt

$\|Ax\|_2^2 = \langle Ax, Ax\rangle = \left\langle x, A^\top Ax\right\rangle = < \sum_{i=1}^{n} \alpha_i x_i, \sum_{j=1}^{n} \alpha_j \lambda_j x_j >$

$= \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leq (\max_{i=1,\dots,n} \lambda_i) \sum_{i=1}^{n} \alpha_i^2 = (\max_{i=1,\dots,n} \lambda_i) \|x\|_2^2$

Damit ist $\|A\|_2 \leq (\max_{i=1,\dots,n} \lambda_i)^{\frac{1}{2}}$

Da für $x = x_i$

$\|Ax\|_2^2 = \lambda_i = \lambda_i \|x_i\|_2^2$ ist, ist $\|A\|_2 \geq (\max_{i=1,\dots,n} \lambda_i)^{\frac{1}{2}}$, also

$\|A\|_2 = (\max_{i=1,\dots,n} \lambda_i)^{\frac{1}{2}}$

□

Bemerkung: Die Spektralnorm von A^{-1} ist $\|A^{-1}\|_2 = \frac{1}{(\min_{i=1,\dots,n} \lambda_i)^{1/2}}$.

Beispiel 3.11:
Für die Matrix aus Beispiel 3.9 ist
$A^\top A = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 10 & 14 \\ 14 & 20 \end{pmatrix}$
Die Matrix $A^\top A$ hat das charakteristische Polynom
$\begin{vmatrix} 10-\lambda & 14 \\ 14 & 20-\lambda \end{vmatrix} = (10-\lambda)(20-\lambda) - 14^2 = \lambda^2 - 30\lambda + 4$
Die Eigenwerte der Matrix $A^\top A$ sind daher
$\lambda_{1,2} = 15 \pm \sqrt{15^2 - 4} = 15 \pm \sqrt{221}$
Damit:
$\|A\|_2 = \sqrt{15 + \sqrt{221}}$

3.2 Die Kondition von Matrizen

Ziel: Ein Maß für die Empfindlichkeit eines linearen Gleichungssystems gegenüber Störungen der Koeffizientenmatrix und der rechten Seite.

Bezeichnungen: Sei $\widehat{x}$ eine Näherungslösung des linearen Gleichungssystems $Ax = b$ mit der exakten Lösung x.

Der Fehler e und das Residuum r sind dann
$e = x - \widehat{x}, r = b - A\widehat{x}$

Satz 3.12
Es ist $\frac{1}{\|A\|\|A^{-1}\|}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|}$
Diese Abschätzungen sind scharf in dem Sinn, dass sie für bestimmte b und $\widehat{x}$ angenommen werden.

Beweis: Wegen $\|r\| = \|Ae\| \leq \|A\|\|e\|$ ist $\frac{\|r\|}{\|A\|} \leq \|e\|$.
Wegen $\|x\| = \|A^{-1}b\| \leq \|A^{-1}\|\|b\|$ ist $\frac{1}{\|A^{-1}\|\|b\|} \leq \frac{1}{\|x\|}$.
Mit diesen beiden Abschätzungen folgt die weitere Abschätzung
$\frac{1}{\|A\|\|A^{-1}\|}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|}$.
Die Abschätzung wird angenommen, wenn
$\|r\| = \|Ae\| = \|A\|\|e\|, \|x\| = \|A^{-1}b\| = \|A^{-1}\|\|b\|$
Nach Definition von $\|A\|$ und $\|A^{-1}\|$ gibt es derartige Vektoren b^* und e^*. Für dieses b^* und $\widehat{x}^* = A^{-1}b^* - e^*$ wird die Abschätzung

angenommen; die Norm des fehlers e^* kan dabei beliebig gewählt werden.

Wegen $\|b\| = \|Ax\| \leq \|A\|\|x\|$ ist $\frac{1}{\|x\|} \leq \|A\|\frac{1}{\|b\|}$
Wegen $\|e\| = \|A^{-1}r\| \leq \|A^{-1}\|\|r\|$ folgt $\frac{\|e\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|r\|}{\|b\|}$
DIe Abschätzung wird angenommen, falls $\|b\| = \|Ax\| = \|A\|\|x\|$,
$\|e\| = \|A^{-1}r\| = \|A^{-1}\|\|r\|$.
Nach Definition der Matrixnorm gibt es derartige Vektoren x^* und r^*,
wobei $\|r^*\|$ als beliebig klein gewählt werden kann.
Für $b^* = Ax^*$ und $\widehat{x}^* = x^* - A^{-1}r^*$ geht die Ungleichung wieder in
eine Gleichung über.
□

Definition 3.13
Die Größe $\kappa(A) = \|A\|\|A^{-1}\|$ heißt Konditionszahl oder Kondition der
invertierbaren $(n \times n)$-Matrix A bezgl. der gegebenen auf dem $\mathbb{R}^n$
Kurzfassung Satz 3.12:
$\frac{1}{\kappa(A)}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \kappa(A)\frac{\|r\|}{\|b\|}$
Bemerkung:
Für alle Matrizen A gilt wegen $1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|$
stets $\kappa(A) \geq 1$. íst $\kappa(A) \approx 1$, spricht man von einer gut konditionierten
Matrix, und ist $\kappa(A) \gg 1$ ist A schlecht konditioniert.

Satz 3.14:
Löst man statt des exakten Systems $Ax = b$ das gestörte System $\widehat{A}\widehat{x} = b$,
so gilt $\frac{\|x-\widehat{x}\|}{\|\widehat{x}\|} \leq \kappa(A)\frac{\|A-\widehat{A}\|}{\|A\|}$.

Beweis: Es ist $x = A^{-1}b = A^{-1}\widehat{A}\widehat{x} = A^{-1}(A + \widehat{A} - A)\widehat{x} = \widehat{x} + A^{-1}(\widehat{A} - A)\widehat{x}$
und damit $\|x - \widehat{x}\| \leq \|A^{-1}\|\left\|A - \widehat{A}\right\|\|\widehat{x}\|$, das heißt
$\frac{\|x-\widehat{x}\|}{\|\widehat{x}\|} \leq \|A\|\|A^{-1}\|\frac{\|A-\widehat{A}\|}{\|A\|} = \kappa(A)\frac{\|A-\widehat{A}\|}{\|A\|}$
Sei nun $\|A^{-1}y\| = \|A^{-1}\|\|y\|$ und $\widehat{A} = A + \delta I, \widehat{x} = \frac{1}{\delta}y$.
Dann ist $\|x - \widehat{x}\| = \left\|A^{-1}(\widehat{A} - A)\widehat{x}\right\| = \|A^{-1}y\| = \|A^{-1}\|\|y\|$
$= \|A^{-1}\|\left\|A - \widehat{A}\right\|\|\widehat{x}\|$ so daß in diesem Fall Gleichheit gilt.
□

Satz 3.15:
Für alle invertierbaren Matrizen A ist $\min\left\{ \frac{\|A-B\|}{\|A\|} \,\middle|\, B \text{ singulär} \right\} \geq \frac{1}{\kappa(A)}$
Für die Maximumnorm, die 1-Norm und die euklidische Norm gilt Gleichheit.

Beweis: Ist B singulär, geht es eine $x \neq 0$ mit $Bx = 0$. Für dieses x ist
$\|x\| = \|A^{-1}Ax\| \leq \|A^{-1}\|\|Ax\| = \|A^{-1}\|\|(A-B)x\|$
$\leq \|A^{-1}\|\|A-B\|\|x\|$ Daher ist $1 \leq \|A^{-1}\|\|A-B\|$ und damit
$\frac{1}{\kappa(A)} = \frac{1}{\|A\|\|A^{-1}\|} \leq \frac{\|A-B\|}{\|A\|}$

Um zu zeigen, dass für die Maximumnorm auch Gleichheit einreten kann, sei v ein Vektor mit $\|v\| = 1, \|A^{-1}v\| = \|A^{-1}\|\|v\|$.
Weiter sei $y = A^{-1}v, \|y\| = |y_m|$, sowie mit dem m-ten Einheitsvektor e_m
$z = \frac{1}{y_m}e_m, B = A - vz^\top$
Dann ist B wegen $y \neq 0$ und $By = AA^{-1}v - \frac{1}{y_m}y_m v = 0$ singulär.
Für alle Vektoren x gilt
$\|(A-B)x\| = \left\| \frac{1}{y_m}x_m v \right\| = \frac{1}{|y_m|}|x_m|\|r\| = \frac{1}{\|y_m\|}|x_m| = \frac{1}{\|A^{-1}v\|}|x_m|$
$= \frac{1}{\|A^{-1}\|\|v\|}|x_m| = \frac{1}{\|A^{-1}\|}|x_m|$ und damit
$\|A-B\| \leq \frac{1}{\|A^{-1}\|}$, d.h. $\frac{\|A-B\|}{\|A\|} \leq \frac{1}{\kappa(A)}$.
□

3.3 Die LR-Zerlegung einer Matrix

Problem: Wie löst man lineare Gleichungssysteme der Form $Ax = b$
mit einer nichtsingulären $(n \times n)$-Matrix A und bekannter rechter Seite b?

naheliegende Methode: Gaußscher Algorithmus
Durch schrittweise Elimination der Unbekannten $x_1, \ldots, x_n$.

Beispiel 3.16

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 4 & 1 & 0 & 6 \\ -2 & 2 & 1 & 5 \end{array}\right)$$

$(\quad A \quad |b)$

Vielfaches der 1. Zeile von Zeile 2 und 3 abziehen, so dass in der 1. Spalte von Zeile 2 und 3 die Null steht.

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 0 & -1 & -2 & -8 \\ 0 & 3 & 2 & 12 \end{array}\right)$$

geeignetes Vielfaches der 2. Zeile von 3. Zeile abziehen:

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 0 & -1 & -2 & -8 \\ 0 & 0 & -4 & -12 \end{array}\right)$$

Ende Eliminationsphase.

Dreieckssystem auflösen:

$$-4x_3 = -12$$
$$-x_2 = -8 + 2x_3$$
$$2x_1 = 7 - x_2 - x_3$$
$$x_3 = 3$$
$$x_2 = 2$$
$$x_1 = 1$$

Bemerkung:

(1) Der Gaußsche Algorithmus ist ein starres Rechenschema.

(2) Dieses Rechenschema kann zusammenbrechen, ohne dass die Matrix singulär ist.

Beispiel 3.17

Die erste Gleichung des eindeutig lösbaren Systems

$$\left(\begin{array}{cc|c} 0 & 1 & 1 \\ 1 & 1 & 2 \end{array}\right)$$

kann nicht genutzt werden, um die erste Variable zu eliminieren!

Abhilfe: Gleichungen in systematischer Weise vertauschen: Spaltenpivotsuche.

Beispiel 3.18

Gegeben wieder das System

$$\left(\begin{array}{ccc|c} 2 & 1 & 1 & 7 \\ 4 & 1 & 0 & 6 \\ -2 & 2 & 1 & 5 \end{array}\right)$$

Zunächst wird die Gleichung mit dem betragsmäßig größten Quotienten vor der ersten Variablen mit der ersten Gleichung vertauscht.

$$\left(\begin{array}{ccc|c} 4 & 1 & 0 & 6 \\ 2 & 1 & 1 & 7 \\ -2 & 2 & 1 & 5 \end{array}\right)$$

Jetzt wird ein geeignetes Vielfaches der 1. Gleichung von der 2. bzw 3. Gleichung abgezogen und aus diesen Gleichungen so dei 1. Variable eliminiert.

$$\left(\begin{array}{ccc|c} 4 & 1 & 0 & 6 \\ 0 & \frac{1}{2} & 1 & 4 \\ 0 & \frac{5}{2} & 1 & 8 \end{array}\right)$$

Nun wird die 3. mit der 2. Gleichung vertauscht, denn mit den verbliebenen Gleichungen hat die 3. Gleichung den größten Koeffizienten vor der 2. Variablen.

$$\left(\begin{array}{ccc|c} 4 & 1 & 0 & 6 \\ 0 & \frac{5}{2} & 1 & 8 \\ 0 & \frac{1}{2} & 1 & 4 \end{array}\right)$$

Eliminatiuon der 2. Variablen aus der 3. Gleichung ergibt

$$\left(\begin{array}{ccc|c} 4 & 1 & 0 & 6 \\ 0 & \frac{5}{2} & 1 & 8 \\ 0 & 0 & \frac{4}{5} & \frac{12}{2} \end{array}\right)$$

Damit ist die Eliminationsphase beendet. Auflösen des Dreieckssystems ergibt $x_3 = 3$,$x_2 = 2$,$x_1 = 1$.

□

Gaußscher Algorithmus ohne Pivotsuche

Der Gaußsche Algorithmus formt das Ausgangssystem $Ax = b$ oder $A^{(0)}x = b^{(0)}$ Schrittweise um in Systeme $A^{(k)}x = b^{(k)}$,$k = 1, \dots, n-1$ mit unveränderter Lösung x

Für die Einträge $a_{ij}^{(k)}$ von $A^{(k)}, k \geq 1$, gilt $a_{ij}^{(k)} = a_{ij}^{(k-1)}$ für $i = 1, \dots, k$

$a_{ij}^{(k)} = 0$ für $i = k+1, \dots, n, j = 1, \dots, k$.

Die restlichen Einträge ergeben sich gemäß:

Für $k = 1, \dots, n-1$ berechne

Für $i = k+1, \dots, n$ berechne

Für $j = k+1, \dots, n$ berechne

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}$$

$$b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}$$

Endergebnis ist das Dreieckssystem

$$\left(\begin{array}{cccc|c} a_{11}^{(n-1)} & a_{12}^{(n-1)} & \cdots & a_{1n}^{(n-1)} & b_1^{(n-1)} \\ 0 & a_{22}^{(n-1)} & & \vdots & b_2^{(n-1)} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{(n-1)} & b_n^{(n-1)} \end{array}\right)$$

Dieses Dreieckssystem lässt sich einfach auflösen:

$x_n = \frac{1}{a_{nn}^{(n-1)}} b_n^{(n-1)}$

Für $i = n-1, \dots, 1$ berechne

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

Bemerkungen:

1) Im Rechner können die Elemente $a_{ij}^{(k-1)}$ der Matrix $A^{(k-1)}$ mit den Elementen $a_{ij}^{(k)}$ der Matrix $A^{(k)}$ überschrieben werden.

2) Man kann die Eliminationsfaktoren $l_i = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$ auf den Plazt der mit Null überschriebenen Matrixelemente $a_{ik}^{(k-1)}$ abspeichern und die rechte Seite dann nachträglich bearbeiten.
Dies ist insbesondere vorteilhaft, wenn das Gleichungssystem für mehrere rechte Seiten gelöst werden soll.

3) Die Eliminationsfaktoren l_{ik} sollten außerhalb der innersten Schleife berechnet werden.

Matrixdarstellung

Führt man die Eliminationsfaktoren

$$l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \Bigg\} \begin{array}{c} k = 1, 2, \dots, n-1 \\ i = k+1, \dots, n \end{array}$$

und die Elinimationsmatrizen

$$G_k|_{ij} = \begin{cases} l_{ik}\,, j = k, i = k+1, \dots, n \\ 0 \quad \text{, sonst} \end{cases}$$

ein, so ist

$A^{(k)} = (I - G_k)A^{(k-1)}, b^{(k)} = (I - G_k)b^{(k-1)}$

Beweis: Da man die k-te Spalte von G_k von Null verschieden ist, gilt

$$G_k A^{(k-1)}|_{ij} = \sum_{m=1}^{n} G_k|_{im} A^{(k-1)}|_{mj} = G_k|_{ik} A^{(k-1)}|_{kj}$$

Für $i = 1, \dots, k$ ist $G_k|_{ik} = 0$ und daher $G_k A^{(k-1)}|_{ij}, j = 1, \dots, n$

Für $i = k+1, \dots, n$ ist $G_k A^{(k-1)}|_{ij} = l_{ik} a_{kj}^{(k-1)}, j = 1, \dots, n$

□

Satz 3.19

Ist der Gaußsche Algorithmus für die $(n \times n)$-Matrix A durchführbar und ergeben sich dabei die Eliminationsmaztrizen $G_1, \dots, G_{n-1}$ so gilt: (i) $(I - G_{n-1})(I - G_{n-2}) \cdots (I - G_1)A = R$

(ii) $G_k G_l = 0, k \leq l$

(iii) $(I - G_k)^{-1} = I + G_k$

(iv) $((I - G_{n-1}) \cdots (I - G_1))^{-1} = I + \sum_{k=1}^{n-1} G_k = I + L$

Dabei ist I die $(n \times n)$-Einheitsmatrix,

$$R = \begin{pmatrix} a_{11}^{(n-1)} & \cdots & a_{1n}^{(n-1)} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{nn}^{(n-1)} \end{pmatrix}$$

eine obere Dreiecksmatrix und

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

Beweis:

(i) Nach Konstruktion ist $A^{(k)} = (I - G_k)A^{(k-1)}$ (siehe oben)

(ii) Für alle Vektoren x mit $x|_k = 0$ ist $G_k x = 0$

(iii) $(I + G_k)(I - G_k) = I - G_k + G_k - G_k G_k = I$

$(I - G_k)(I + G_k) = I + G_k - G_k - G_k G_k = I$

(iv) $((I - G_{n-1}) \cdots (I - G_1))^{-1}$

$= (I - G_1)^{-1} \cdots (I - G_{n-1})^{-1}$

$= (I + G_1) \cdots (I + G_{n-1})$ wegen (ii)

$= I + \sum_{k=1}^{n-1} G_k$ wegen (ii)

□

Folgerung:

Der Gaußsche Algorithmus liefert bei exakter Rechnung eine LR-Zerlegung

$$A = LR$$

Für Matrix A in eine untere $\triangle$-Matrix mit Einsen in der Diagonale und eine obere Dreiecksmatrix R, die LR-Zerlegung von A.

Das Gleichungssystem $Ax = b$ lasst sich dann wie folgt lösen:

1) Löse $Ly = b$
2) Löse $Rx = y$

Spaltenpivotsuche

Führt man eine Spaltenpivotsuche durch, so liefert der Gaußsche Algorithmus eine Zerlegung $LR = PA$ einer Matrix PA, deren Zeilen gegenüber A vertauscht sind.

Begründung:
Der Gaußsche Algorithmus mit Spaltenpivotsuche liefert Eliminationsmatrizen $G_1, \dots, G_{n-1}$ Permutationsmatrizen, $P_1, \dots, P_{n-1}$ und eine obere Dreiecksmatrix R mit
$(I - G_{n-1})P_{n-1} \cdots (I - G_1)P_1 A = R$
Die Permutationsmatrix P_i vertauscht die i-te Komponente eines Vektors mit einer Komponente $k \geq i$. Daher ist $P_i^{-1} = P_i$ und wegen
$(I - G_i)^{-1} = I + G_i, A = P_i(I + G_1) \cdot \dots \cdot P_{n-1}(I + G_{n-1})R$
Sei nun $Q_k = P_{n-1} \cdot \dots \cdot P_k, P = P_{n-1} \cdot \dots \cdot P_1$
Wegen $P_k P_k = I$ gilt dann
$Q_k P_k(I + G_k) = Q_{k+1} P_k P_k(I + G_k) = Q_{k+1}(I + G_k)$
$= (I + Q_{k+1} G_k Q_{k+1}^{-1}) Q_{k+1}$
Da die Spalten $k+1, \dots, n$ von G_k ausschließlich mit Nullen besetzt sind, gilt $G_k Q_{k+1}^{-1} = G_k P_{k+1} \cdot \dots \cdot P_{n-1} = G_k$ und damit
$Q_k P_k(I + G_k) = (I + Q_{k+1} G_k) Q_{k+1}$
Die Matrix $I + Q_{k+1} G_k$ ist eine Matrix mit gleicher Besetzungsstruktur wie $I + G_k$. Daher ist
$L := P P_1(I + G_1) \dots P_{n-1}(I + G_{n-1})$
$= (I + Q_2 G_1) \dots (I + Q_{n-1} G_{n-2})(I + G_{n-1})$
$= I + Q_2 G_1 + \dots + Q_{n-1} G_{n-2} + G_{n-1}$
eine untere Dreiecksmatrix mit Einsen in der Diagonale und
$PA = LR$

Beispiel 3.20

ohne Spaltensuche

$$\begin{pmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{pmatrix} \to \begin{pmatrix} 2 & 1 & 1 \\ \overline{2\ |} & -1 & -2 \\ -1| & 3 & 2 \end{pmatrix} \to \begin{pmatrix} 2 & 1 & 1 \\ \overline{2\ |} & -1 & -2 \\ -1 & \overline{-3|} & -4 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{pmatrix}$$

mit Spaltensuche

$$\begin{pmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 4 & 1 & 0 \\ 2 & 1 & 1 \\ -2 & 2 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 4 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \\ -\frac{1}{2} & \frac{5}{2} & 1 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 4 & 1 & 0 \\ -\frac{1}{2} & \frac{5}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \leftarrow \text{erste Spalte mitvertauschen!}$$

$$\rightarrow \begin{pmatrix} 4 & 1 & 0 \\ -\frac{1}{2} & \frac{5}{2} & 1 \\ \frac{1}{2} & \frac{1}{5} & \frac{4}{5} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{1}{5} & 0 \end{pmatrix} \begin{pmatrix} 4 & 1 & 0 \\ 0 & \frac{5}{2} & 1 \\ 0 & 0 & \frac{4}{5} \end{pmatrix} = \begin{pmatrix} 4 & 1 & 0 \\ -2 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}$$

Aufwand für die Berechnung der LR-Zerlegung

Anzahl der Multiplikationen und Divisionen zur Elimination der k-ten Variablen:

$(n-k)$ Berechnung der $l_{ik}, i = k+1, \dots, n$

$(n-k)^2$ Berechnung der $a_{ij}^{(k)}, i, j = k = 1, .., n$

Gesamtzahl: $(k = n-k)$

$$\sum_{k=1}^{n-1}((n-k)+(n-k)^2) = \sum_{k=1}^{n-1}(k+k^2) = \frac{1}{3}n^3 - \frac{1}{3}n$$

Aufwand für die Lösung der Dreieckssysteme

$$\sum_{k=2}^{n}(k-1) \qquad |Ly = b$$

$$+\sum_{k=1}^{n-1}(n-k)+n \qquad |Rx = y$$

Gesamt: n^2

Aufwand für die Lösung von $Ax = b$

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

Aufwand für die Berechnung von A^{-1}:

$(\frac{1}{3}n^3 - \frac{1}{3}n) + n \cdot n^3 = \frac{4}{3}n^3 - \frac{1}{3}n$

Aufwand für die Lösung von k Systemen mit gleicher Matrix A

$(\frac{1}{3}n^3 - \frac{1}{3}n) + kn^2$ mit LR Zerlegung
$(\frac{1}{3}n^3 - \frac{1}{3}n) + n \cdot n^2 + kn^2$ mit A^{-1}
Merke: Man berechnet nie die Inverse einer Matrix, es sei denn
man benötigt sie explizit.

3.4 Eine Fehleranalyse für den Gaußschen Algorithmus

Frage: Wie stark können Rundungsfehler das Ergebnis verfälschen?
Idee: Rückwärtsnanlyse
Zeige, dass die berechnete Lösung $\overline{x}$ des Systems $Ax = b$ die
exakte Lösung eines leicht(?) gestörten Systems $\widehat{A}\widehat{x} = \widehat{b}$ ist.
Schätze die Größe der Störung $\widehat{A} - A, \widehat{b} - b$ ab.
Ihr Einfluss hängt von der Kondition von A ab.

Bezeichnung: Ist A eine relle Matrix mit den Matrixelementen a_{ij}, so
bezeichnet $|A|$ die Matrix gleicher Dimension mit den Matrixelementen $|a_{ij}|$.

Bezeichnung: $A \leq B \iff a_{ij} \leq b_{ij}, i, j = 1, \dots, n$

Satz 3.21

Ist der Gaußsche Algorithmus(ohne Pivotsuche) für die Matrix A durchführbar,
so liefert er eine LR Zerlegung $\widehat{L}\widehat{R} = A + H$ einer gestörten Matrix $A + H$.
Die Störungsmatrix H genügt der komponentenweisen Abschätzung
$|H| \leq C(n)\varepsilon^*(|A| + |\widehat{L}||\widehat{R}|)$, wobei die Konstante $C(n)$ gegeben ist durch
$C(n) = \frac{(1+(2+\varepsilon^*)\varepsilon^*)^n - 1}{\varepsilon^*} \approx 2n$ und ε^* die Maschinengenauigkeit bezeichnet.

Satz 3.22

Die berechnete Lösung der beiden Dreieckssysteme $\widehat{L}y = b, \widehat{R}x = y$
sind die exakten Lösungen zweier gestörter Dreieckssysteme
$(\widehat{L} + F)\widehat{y} = b, (\widehat{R} + G)\widehat{x} = \widehat{y}$ mit Matrizen F und G mit
$|F| \leq C(n)\frac{\varepsilon^*}{1-(2+\varepsilon^*)\varepsilon^*}\left|\widehat{L}\right|, |G| \leq C(n)\frac{\varepsilon^*}{1-(2+\varepsilon^*)\varepsilon^*}\left|\widehat{R}\right|$

Folgerung:
Die Auflösung von Dreieckssystemen ist ein gutartiger Prozeß.
Ob die LR-Zerlegung selbst gutartig ist, hängt davon ab, ob man die
Größenordnung der Matrixelemente von $\widehat{L}$ und $\widehat{R}$ kontrollieren kann.

Spaltenpivotisierung

$\widehat{L}\widehat{R} = PA + H$

$|H| \leq C(n)\varepsilon^*(|PA| + \left|\widehat{L}\right|\left|\widehat{R}\right|)$

Begründung:

Der Gaußsche Algorithmus ohne Pivotsuche angewandt auf PA liefert identische Matrizen $\widehat{L}, \widehat{R}$ wie der Gaußsche Algorithmus mit Pivotsuche angewandt auf A.

Satz 3.23

Die Spaltenpivotsuche gilt für die Matrixelemente der Dreiecksmatrizen L und R bei exakter Rechnung. Die Abschätzung $(l_{ik}) \leq 1, |r_{kl}| \leq 2^{k-1}a$, wobei $a = \max\limits_{k,l} |a_{kl}|$ sein soll.

Beweis: (einfache) Induktion über k.

Bemerkung: Diese Schranken sind scharf, wie das Beispiel der Matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -1 & \cdots & -1 & 1 \end{pmatrix}$$

zeigt.

Meistens ist diese Abschätzung für die Elemente von R aber viel zu pessimistisch. Der Gaußalgorithmus verhält sich fast immer gutartig.

3.5 Symmetrisch positiv-definite Matrizen

Definition 3.24

Eine $(m \times n)$-Matrix A heißt symmetrisch, falls $A^T = A$ ist.

Sie heißt positiv definit, falls $\langle x, Ax\rangle > 0$ für alle $x \in \mathbb{R}^n$,$x \neq 0$ gilt, oder wenn es äquivalent dazu eine Konstante $\delta > 0$ gibt mit $\langle x, Ax\rangle \geq \delta\|x\|_2^2$ für alle $x \in \mathbb{R}^n$.

Satz 3.25

Jede symmetrische und positiv definite Matrix A besitzt eine LR-Zerlegung von der Form $A = LR = LDL^T$ mit einer unteren Dreiecksmatrix L mit Einsen in der Diagonale und einer Diagonalmatrix D mit positiven Diagonalelementen.

Beweis: Wegen $\langle e_1, Ae_1\rangle = a_{11}$ ist $a_{11} > 0$. Sei G_1 die Gaußsche Eliminationsmatrix mit

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

Dann ist wegen $a_{ij} = a_{j1}$

$(I - G_1)A(I - G_1)^T$

$$= \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 \\ 0 & B \end{pmatrix}$$

Die $(n-1) \times (n-1)$-Restmatrix B ist wegen $((I - G_1)A(I - G_1)^T)^T$

$= \left((I - G_1)^T\right)^T A^T (I - G_1)^T = (I - G_1)A^T(I - G_1)^T$

$= (I - G_1)A(I - G_1)^T$symmetrisch.

Ist $x = (x_2, \dots, x_n)^T \in \mathbb{R}^{n-1}$ und $\widetilde{x} \in \mathbb{R}^n$ gegeben durch

$\widetilde{x} = (0, x_2, \dots, x_n)^T$, so gilt $(x, Bx) = (\widetilde{x}, (I - G_1)A(I - G_1)^T\widetilde{x})$

$= ((I - G_1)^T\widetilde{x}, A(I - G_1)^T\widetilde{x})$

Da $(I - G_1)^T\widetilde{x}$ genau dann der Nullvektor ist, wenn $x = 0$ ist, ist für $x \neq 0$

$\langle x, Bx\rangle > 0$. Damit Restmatrix B wieder symmetrisch und positiv definit.

Durch Induktrion folgt aus diesen Überlegungen, dass es Eliminationsmatrizen $G_1, .., G_{n-1}$ gibt derart, dass

$D := (I - G_{n-1}) \cdots (I - G_1)A(I - G_1)^T \cdots (I - G_{n-1})^T$

eine Diagonalmatrix mit positiven Diagonalelementen ist. Wegen

$L := ((I - G_{n-1}) \cdots (I - G_1))^{-1} = I + \sum_{k=1}^{n-1} G_k$ und

$((I - G_1)^T \cdots (I - G_{n-1})^T)^{-1} = (((I - G_{n-1}) \cdots (I - G_1))^T)^{-1}$

$= (((I - G_{n-1}) \cdots (I - G_1))^{-1})^T = L^T$

folgt daraus die Existenz der Zerlegung.

□

Folgerung:
Eine symmetrische Matrix ist genau dann positiv definit, wenn sich der Gaußsche Algorithmus ohne Spaltenpivotsuche für sie als durchführbar erweist und auf eine obere Dreiecksmatrix $R(= DL^T)$ mit positiven Diagonalelementen liefert.
Wegen $\langle x, LDL^T x\rangle = \langle L^T x, DL^T x\rangle > 0$

Satz 3.26
Ist die Matrix A symmetrisch und positiv definit und liefert der Gaußsche Algorithmus bei exakter Rechnung die LR-Zerlegung $A = LR$, so gilt für die Elemente der Matrix $|L||R|$ die Abschätzung
$|L||R|\,|_{ij} \leq \sqrt{a_{ii}a_{jj}}$

Beweis: Sei $R = DL^T$, $C = LD^{\frac{1}{2}}$. Dann ist $A = LR = LDL^T = CC^T$ und $|L||R| = |C||C|^T$. Für die Elemente der Matrix $C = (c_{ij})$ gilt
$\sum_{j=1}^{n} c_{ij}^2 = CC^T|_{ii} = a_{ii}$
Daher ist
$|L||R|\,|_{ij} = |C||C|^T|_{ij} = \sum_{k=1}^{n} |c_{ik}||c_{jk}|$
$\leq (\sum_{k=1}^{n} |c_{ik}|^2)^{1/2} (\sum_{k=1}^{n} |c_{jk}|^2)^{1/2}$ (Schwarzsche Ungleichung)
$= \sqrt{a_{ii}a_{jj}}$ $\square$

Folgerung:
Der Gaußsche Algorithmus ist für symmetrische-positiv definite Matrizen nicht nur immer durchführbar, sondern auch numerisch extrem stabil.

Bemerkung:
Die Zerlegung $A = LBL^T$ einer symmetrisch positiv definiten Matrix A lässt sich durch Koeffizientenvergleich auf direktem Weg beechnen.
Gegenüber der allgemeinen Form des Gaußschen Algorithmus halbieren sich dabei Rechenaufwand und Speicherplatbedarf.
Anzahl Punktoperationen:
$\frac{1}{6}n^3 + \mathcal{O}(n^2)$

Cholesky Zerlegung:
$A = CC^T$ $\quad (C = LD^{1/2}$ untere Dreiecksmatrix)
Lässt sich in gleicher Weise durch Koeffizientenvergleich bestimmen.
Rechenaufwand:
$\frac{1}{6}n^3 + \mathcal{O}(n^2)$ Punktoperationen, n Quadratwurzeln.

3.6 Eliminationsverfahren für schwach besetzte Matrizen

Schwach besetzte Matrix:
sehr wenige Nichtnullelemente im Vergleich zur Geamtzahl der Matrixelemente.
typisch $\mathcal{O}(n)$ Nichtnullelemente bei $n \times n$-Matrix.
$n = 1\,000, 10\,000, 100\,000, 1\,000\,000$

Direkte Verfahren

Nur die Nichtnullelemente von A, L und R werden abgespeichert.
Von besonderer Bedeutung sind in diesem Zusammenhang die benutzten Datenstrukturen. Für wesentliche Problemklassen wie z.B. symmetrische positiv definite Matrizen kann man auf eine Pivotsuche verzichten.
Dann läuft ein direktes Verfahren in vier Phasen ab.

Phase 1: Analysephase
In dieser Phase wird eine Permutationsmatrix P bestimmt, so dass die LR-bzw. LDL^T-Zerlegung der Matrix PAP^T möglichst wenig Nichtnullelemente enthält.
Dabei wird allein von der Besetzungsstruktur, d.h. der Position der Nichtnullelemente von A ausgegangen.
Algorithmen zur Bestimmung solcher Permutationsmatrizen beruhen auf Graphentheoretischen Überlegungen.

Phase 2: symbolische Zerlegung
In dieser Phase wird die Position der Nichtnullelemente von L und R bzw. nur von L bestimmt und eine entsprechende Datenstruktur aufgebaut.

Phase 3: numerische Zerlegung
In dieser Phase werden L und R bzw. L und D berechnet.

Bemerkung: Phase 2 und Phase 3 sind am aufwändigsten. In der Regel wächst der Speicherplatz und Rechenaufwand überproportional zur Anzahl der Unbekannten.

Phase 4: Lösungsphase
Auflösung der Dreieckssysteme.

Definition 3.27 (Bandmatrix)
Eine $(n \times n)$-Matrix $A = (a_{ik})$ heißt Bandmatrix der Bandbreite $m, m < n$, falls $a_{ik} = 0$ für $|i - k| > m$ gilt.
Im Fall $m = 1$ spricht man von einer Tridiagonalmatrix.

Satz 3.28
Besitzt eine Bandmatrix A der Bandbreite m eine LR-Zerlegung und ist A invertierbar, so sind auch L und R Bandmatrizen der Bandbreite m.

Beweis: (und Konstruktion von L und R)
durch Koeffizientenvergleich.
□

Folgerung:
Berechnet man eine solche LR-Zerlegung durch Koeffizientenvergleich, braucht man $nm^2 - \frac{2}{3}m^3 + mn$ Punktoperationen.
Für $m \ll n$ ist dies eine sehr große Ersparnis gegenüber den $\frac{1}{3}n^3 - \frac{1}{3}n$ Punktoperationen bei voller Abspeicherung der Matrix.
bei Spaltenpivotsuche: ähnliche Einsparungen

3.7 Iterationsverfahren

Problem: Bei direkten Verfahren(auch wenn sie auf schwach besetzte Matrizen zugeschnitten sind) wächst die Rechenzeit und Speicherplatzaufwand in der Regel überproportiional zur Anzahl der Unbekannten.

Ausweg: Iterative Verfahren
gegeben: A nichsinguläre $(n \times n)$ Matrix
B nichtsinguläre $(n \times n)$ Matrix
Iterationsverfahren:
$x^{(i+1)} = x^{(i)} + B^{-1}(b - Ax^{(i)})$
zur Lösung des Gleichungssystems
$Ax = b$
mit der exakten Lösung $\overline{x}$.

wichtige Bemerkung:
Im Gegensatz etwa zum Gaußschen Eliminationsverfahren sind iterative Verfahren keine universellen Verfahren. Zur Konstruktion einer guten Näherungsinversen B^{-1} für A muss die Problemstruktur ausgenutzt werden.

Beispiel:
Mehrgitterverfahren für diskretisierte partielle Differentialgleichungen.

Beispiel 3.29
Die Matrix A habe die Darstellung
$A = L + D + R$ mit
$L|_{ij} = 0$ für $j \geq i$
$L|_{ij} = 0$ für $j \leq i$
$D|_{ij} = 0$ für $i \neq j$

D sei nichtsingulär. Dann ist $x^{(i+1)} = x^{(i)} + D^{-1}(b - Ax^{(i)})$ das Gesamtschritt oder Jacobi-Verfahren.
In Koordinatenschreibweise:
$x_k^{(i+1)} = x_k^{(i)} + \frac{1}{a_{kk}}(b_k - \sum_{j=1}^{n} a_{kj}x_j^{(i)})$
$= \frac{1}{a_{kk}}(b_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj}x_j^{(i)})$

Beispiel 3.30

(Bezeichnungen wie im Beispiel 3.29)

Das verfahren

$x^{(i+1)} = x^{(i)} + (L + D)^{-1}(b - Ax^{(i)})$ ist das Einzelschritt- oder

Gauß-Seidel-Verfahren.

Wegen $Dx^{(i+1)} = b - Lx^{(i+1)} - Rx^{(i)}$ lautet es in Komponentenschreibweise

$x_k^{(i+1)} = \frac{1}{a_{kk}}(b_k - \sum_{j<k} a_{kj}x_j^{(i+1)} - \sum_{j>k} a_{kj}x_j^{(i)})$ (leere Summe=0)

einfacher zu realisieren als das Jacobi-Verfahren!

Satz 3.31

Für die Fehler $x^{(i)} - \overline{x}$ gilt die Rekursion:

$x^{(i+1)} - \overline{x} = (I - B^{-1}A)(x^{(i)} - \overline{x})$

Die Matrix $I - B^{-1}A$ ist die Fehlerfortpflanzungsmatrix.

Beweis: klar

□

Satz 3.32

Gilt in der durch die Vektornorm $\|.\|$ induzierte Matrixnorm

$\|I - B^{-1}A\| \leq q < 1$ so ist $\left\|x^{(i+1)} - \overline{x}\right\| \leq q\left\|x^{(i)} - \overline{x}\right\| \leq q^{(i)}\left\|x^{(0)} - \overline{x}\right\|$,

speziell also $\left\|x^{(i)} - \overline{x}\right\| \to 0$.

Beweis: klar

□

Beispiel 3.33

Gibt es eine Konstante $\theta < 1$ und

$\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \leq \theta |a_{ii}|, i = 1, \ldots, n$

so gilt bezüglich der Matrixnorm als Vektornorm für die Fehlerfortpflanzungsmatrix des Jacobiverfahrens $\|I - D^{-1}A\| \leq \theta$.

Beweis: Für $i = 1, \dots, n$ ist

$|x_i - \frac{1}{a_{ii}} \sum_{j=1}^{n} a_{ij} x_j|$

$= |\frac{1}{a_{ii}} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j|$

$\leq (\frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|) \|x\|$

$\leq \theta \|x\|$

$\Rightarrow \|(I - D^{-1}A)x\| \leq \theta \|x\|$

$\square$

Beispiel 3.34

Unter den Voraussetzungen aus Beispiel 3.33
gilt für die Fehlerfortpflanzungsmatrix des Gauß-Seidel Verfahrens bezüglich der Maximumnorm $\|I - (L + D)^{-1}A\| \leq \theta$.

Beweis: Wir zeigen durch Induktion über k

$|x_k^{(i+1)} - \overline{x}_k| \leq \theta \left\| x^{(i)} - \overline{x} \right\|$

Es ist

$Dx^{(i+1)} = b - Lx^{(i+1)} - Rx^{(i)}$

$D\overline{x} = b - L\overline{x} - R\overline{x}$, also

$x^{(i+1)} - \overline{x} = D^{-1}L\left(\overline{x} - x^{(i+1)}\right) + D^{-1}R(\overline{x} - x^{(i)})$

oder in Komponentenschreibweise

$x_k^{(i+1)} - \overline{x}_k = \frac{1}{a_{kk}} \sum_{j<k} a_{kj}(\overline{x}_j - x_j^{(i+1)}) + \frac{1}{a_{kk}} \sum_{j>k} a_{kj}(\overline{x}_j - x_j^{(i)})$

Daraus folgt

$|x_k^{(i+1)} - \overline{x}_k| \leq \left[\frac{1}{|a_{kk}|} \sum_{j<k} |a_{kj}| \theta + \frac{1}{|a_{kk}|} \sum_{j>k} |a_{kj}| \right] \left\| \overline{x} - x^{(i)} \right\|$

$\leq \left[\frac{1}{|a_{kk}|} \sum_{j \neq k} |a_{kj}| \right] \left\| \overline{x} - x^{(i)} \right\|$

und endlich

$\left\| x^{(i+1)} - \overline{x} \right\| \leq \theta \left\| x^{(i)} - \overline{x} \right\|$

Die $x^{(i)} - \overline{x}$ ein beliebiger Vektor sein kann folgt daraus

$\|I - (L + D)^{-1}A\| \leq \theta$.

Verfahren: $x^{(i+1)} = x^{(i)} + B^{-1}(b - Ax^{(i)})$

Fehlerfortpflanzung:
$x^{(i+1)} - \overline{x} = (I - B^{-1}A)(x^{(i)} - \overline{x})$

Satz 3.35
Sind $\|.\|$ und $|||.|||$ Normen auf dem $\mathbb{R}^n$ für die mit paarweise verschiedenen Konstanten K_1, K_2, $K_1\|x\| \leq K_2\|x\| \quad (x \in \mathbb{R}^n)$ gilt, und gilt bezühlich der durch die Norm $|||.|||$ induzierte Matrixnorm $|||I - B^{-1}A||| \leq q < 1$,
so ist $\left\|x^{(i)} - \overline{x}\right\| \leq \frac{K_2}{K_1} q^i \left\|x^{(i)} - \overline{x}\right\|$.

Beweis:
$\left\|x^{(i)} - \overline{x}\right\| \leq \frac{1}{K_1}\left|\left\|x^{(i)} - \overline{x}\right\|\right| \leq \frac{1}{K_1} q^i \left|\left\|x^{(i)} - \overline{x}\right\|\right| \leq \frac{1}{K_1} q^i K_2 \left\|x^{(i)} - \overline{x}\right\|$
□

Bemerkung: Nach Satz 3.5 existieren solche Konstanten K_1 und K_2 immer; praktisch ist die Größenordnung von K_2/K_1 entscheidend.

Ziel: Charakterisierung der Konvergenz
Spektralraidus $\rho(A)$ einer Matrix A.:
Betrag des größten Eigenwertes $\lambda \in \mathbb{C}$ von A.

Satz 3.36
Zu jeder $(n \times n)$-Matrix A und jedem $\varepsilon > 0$ gibt es eine Vektornorm auf dem $\mathbb{R}^n$, so dass für die induzierte Matrixnorm $|||A||| \leq \rho(A) + \varepsilon$ gilt.

Beweis: Sei $A = T^{-1}JT$ und

$$J = \begin{pmatrix} J_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_n \end{pmatrix}, J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix}$$

(Jordansche Normalform)
Sei weiter $D = \mathrm{diag}(1, \varepsilon, ..., \varepsilon^{n-1})$
Dann ist

$$D^{-1}TAT^{-1}D = \begin{pmatrix} J_1' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_n' \end{pmatrix}, J_i' = \begin{pmatrix} \lambda_i & \varepsilon & & 0 \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ 0 & & & \lambda_i \end{pmatrix}$$

Wählt man jetzt als Vektornorm $|||x||| := \|D^{-1}Tx\|_\infty$, so ist wegen
$D^{-1}T(Ax) = (D^{-1}TAT^{-1}D)(D^{-1}Tx)$
die zugehörige Matrixnorm die Zeilensummennorm von
$D^{-1}TAT^{-1}D$, also $|||A||| \leq \max_i |\lambda_i| + \varepsilon = \rho(A) + \varepsilon$
□

Satz 3.37
Ist $\rho(I - B^{-1}A) < \delta < 1$, so gilt in jeder Norm auf dem $\mathbb{R}^n$
$\lim_{i\to\infty} \delta^{-i}\left\|x^{(i)} - \overline{x}\right\| = 0$
Ist umgekehrt $\rho(I - B^{-1}A) \geq 1$ so konvergiert $x^{(i)}$ nicht für alle
Startvektoren $x^{(i)}$ gegen $\overline{x}$.

Beweis: Sei $\rho(I - B^{-1}A) < q < \delta < 1$. Dann gibt es nach Satz 3.36 eine
Norm $|||.|||$ auf dem $\mathbb{R}^n$, so daß bezüglich der induzierten Matrixnorm
$|||I - B^{-1}A||| \leq q$ ist. Nach Satz 3.35 folgt $\left|\left\|x^{(i)} - \overline{x}\right\|\right| \leq q^i \left|\left\|x^{(i)} - \overline{x}\right\|\right|$.
Da sämtliche Normen auf dem $\mathbb{R}^n$ äquivalent sind, gibt es eine
Konstante K mit $\|x\| \leq K|||x|||$ für alle x, also $\left\|x^{(i)} - \overline{x}\right\| \leq Kq^{(i)}\left|\left\|x^{(0)} - \overline{x}\right\|\right|$
Daraus folgt wegen $q < \delta$ $\lim_{i\to\infty} \delta^{-i}\left\|x^{(i)} - \overline{x}\right\| = 0$

Ist $\rho(I - B^{-1}A) \geq 1$ so gibt es $z \in \mathbb{C}, z \neq 0$, und ein $\lambda \in \mathbb{C}$ mit
$(I - B^{-1}A)z = \lambda z, |\lambda| \geq 1$
Damit kann $(I - B^{-1}A)^i z = \lambda^i z$ nicht gegen 0 streben.
Ist $z = u + iv, n, v \in \mathbb{R}^n$, so gilt für ein $e \in \{u, v\}$
$(I - B^{-1}A)e \nrightarrow 0$ für $i \to \infty$.

4. Nichtlineare Gleichungssysteme

$\sin x_1 - x_2 = 0$
$x_1 - \cos x_2 = 0$

4.1 Der Banachsche Fixpunktsatz

Satz 4.1 (BFP)

Die Funktion $\varphi : G \subseteq \mathbb{R}^n \to \mathbb{R}^n$ bilde die abgeschlossene Menge G in sich ab. Sei sie Lipschitz-stetig bezüglich der Norm $\|.\|$, d.h. für alle $x, y \in G$ sei $\|\varphi(x) - \varphi(y)\| \leq L\|x - y\|$ mit einer Konstanten $L < 1$.

Dann besitzt φ genau einen Fixpunkt $\overline{x} \in G$. Für alle $x_0 \in G$ strebt da durch $x_{k+1} = \varphi(x_k), k = 0, 1, 2, ...$ gegebene Folge gegen $\overline{x}$.
Es gelten die a posteriori Abschätzung
$\|x_k - \overline{x}\| \leq \frac{L}{1-L}\|x_k - x_{k-1}\|$
und die a priori Abschätzung
$\|x_k - \overline{x}\| \leq \frac{L^k}{1-L}\|x_1 - x_0\|$.
Beweis: genau wie Satz 2.14
□

Bemerkung: Ist G konvex, ist die Schrittfunktion φ auf einer offenen Obermenge von G stetig differenzierbar und ist $L := \sup_{x \in G} \|\varphi'(x)\| < 1$, so gilt nach dem Hauptsatz der Differential- und Integralrechnung

$$\|\varphi(x) - \varphi(y)\| = \|\int_0^1 \varphi'(x + t(y - x))(y - x)\, \mathrm{dt}\|$$
$$\leq \int_0^1 \|\varphi'(x + t(y - x))(y - x)\mathrm{dt}\|$$
$$\leq \int_0^1 \|\varphi'(x + t(y - x))\|\|y - x\|\mathrm{dt}$$
$$\leq \int_0^1 L\|y - x\|\mathrm{dt}$$
$$= L\|y - x\|$$
□

Beispiel 4.2
Iterationsverfahren der Form $x_{k+1} = \varphi(x_k), \varphi(x) = x + B^{-1}(b - Ax)$
zur Lösung des linearen Gleichungssystems $Ax = b$.
In diesem Falls ist $L = \|I - B^{-1}A\|$

weiteres Beispiel:
Fixpunktiteration der Form
$x_{k+1} = x_k - f'(x_0)^{-1} f(x_k)$
zur Lösung des Systems $f(x) = y$

In diesem Fall ist $\varphi(x) = x - f'(x_0)^{-1}(f(x) - y))$ und
$\varphi'(x) = I - f'(x_0)^{-1} f'(x)$

anderes Beispiel:
$\varphi(x) = a + \varepsilon\psi(x)$

4.2 Das Newton Verfahren

gegeben: $f : G \subseteq \mathbb{R}^n \to \mathbb{R}^n$
gesucht: $\overline{x} \in G$ mit $f(\overline{x}) = 0$
Idee: lokale Linearisierung
$0 = f(\overline{x}) = f(x_k) + f'(x_k)(\overline{x} - x_k) + ...$
$0 = f(x_k) + f'(x_k)(\overline{x} - x_k)$
$x_{k+1} = x_k - f'(x_k)^{-1} f(x_k)$

rechnerisch:
$f'(x_k)(x_{k+1} - x_k) = -f(x_k)$
(Löse lineares Gleichungssystem für $\delta_2 = x_{k+1} - x_k$)

affine Invarianz
A invertierbare Matrix, $g(x) = Af(x)$
$x - g'(x)^{-1} g(x)$
$= x - (Af'(x))^{-1}(Af(x))$
$= x - (f'(x)^{-1}A^{-1})(Af(x))$
$= x - f'(x)^{-1}(A^{-1}A)f(x)$
$= x - f'(x)^{-1} f(x)$

Beispiel 4.3

$\sin x_1 - x_2 = 0$

$x_1 - \cos x_2 = 0$

Kurzform: $f(x) = 0$ mit $f(x) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} \sin x_1 - x_2 \\ x_1 - \cos x_2 \end{pmatrix}$

Es ist

$f'(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial f_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \cos x_1 & -1 \\ 1 & \sin x_2 \end{pmatrix}$

$f'(x)^{-1} = \frac{1}{1+\cos x_1 \sin x_2} \begin{pmatrix} \sin x_2 & 1 \\ -1 & \cos x_1 \end{pmatrix}$

Iterationsvorschrift:

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \frac{1}{1+\cos x_1 \sin x_2} \begin{pmatrix} \sin x_2 & 1 \\ -1 & \cos x_1 \end{pmatrix} \begin{pmatrix} \sin x_1 - x_2 \\ x_1 - \cos x_2 \end{pmatrix}$

Beobachtung: sehr schnelle Konvergenz

□

Satz 4.4

Die Funktion f sei auf einer Kugel $\mathcal{K}$ mit ihrer Nullstelle $\overline{x}$ als Mittelpunkt definiert und dort stetig differenzierbar.
$\forall x \in \mathcal{K}$ sei $f'(x)$ nichtsingulär.

Es geben eine Konstante $\omega > 0$ mit $\|f'(x)^{-1}[f'(y) - f'(x)]\| \leq \omega \|y - x\|$ für alle $x, y \in \mathcal{K}$. Für ein gegebenes $x_0 \in \mathcal{K}$ ist $q := \frac{\omega}{2}\|x_0 - \overline{x}\| < 1$.
Dann ist die Folge der Newton Iterierten $x_{k+1} = x_k - f'(x_k)^{-1} f(x_k)$, $k = 0, 1, 2, ...$, wohldefiniert. Sie konvergiert gegen $\overline{x}$.
Alle x_k liegen in $\mathcal{K}$ und es gilt die Abschätzung
$\|x_{k+1} - \overline{x}\| \leq \frac{\omega}{2}\|x_k - \overline{x}\|^2$

Beweis: Für $x \in \mathcal{K}$ sei $\varphi(x) := x - f'(x)^{-1} f(x)$.
Hält man $x \in \mathcal{K}$ fest und wendet den Hauptsatz der Differential- und Integralrechnung auf die Komponenten der Hilfsfunktion
$\psi(t) = f(x + t(\overline{x} - x)), 0 \leq t \leq 1$ an, so erhält man

$$f(\overline{x}) = f(x) + \int_0^1 f'(x + t(\overline{x} - x))(\overline{x} - x)$$

Wegen $f(\overline{x}) = 0$ folgt daraus

$$f(x) = \int_0^1 f'(x + t(\overline{x} - x))(x - \overline{x})\mathrm{dt}$$

Daher ist $\overline{x} - \varphi(x) = \overline{x} - x + f'(x)^{-1} f(x)$
$= f'(x)^{-1}[f(x) - f'(x)(x - \overline{x})]$
$= \int_0^1 f'(x)^{-1}[f'(x + t(\overline{x} - x)) - f'(x)](x - \overline{x})\mathrm{dt}$
Daraus folgt
$\|\varphi(x) - \overline{x}\| \leq \int_0^1 \|f'(x)^{-1}[f'(x + t(\overline{x} - x)) - f'(x)](x - \overline{x})\|\mathrm{dt}$
$\leq \int_0^1 \|f'(x)^{-1}[f'(x + t(\overline{x} - x)) - f'(x)]\|\|x - \overline{x}\|\mathrm{dt}$
$\leq \int_0^1 \omega\|(x + t(\overline{x} - x)) - x\|\|x - \overline{x}\|\mathrm{dt}$
$= \int_0^1 \omega t\|\overline{x} - x\|\|x - \overline{x}\|\mathrm{dt}$
$= \frac{\omega}{2}\|x - \overline{x}\|^2$
Für alle $x \in \mathcal{K}$ ist also
$\|\varphi(x) - \overline{x}\| \leq \frac{\omega}{2}\|x - \overline{x}\|^2$ (1)
Ist daher $x_k \in \mathcal{K}$ und gilt
$\frac{\omega}{2}\|x_k - \overline{x}\| \leq q$ (2)
so erfüllt $x_{k+1} = \varphi(x_k)$ wegen (1)
$\|x_{k+1} - \overline{x}\| \leq \frac{\omega}{2}\|x_k - \overline{x}\|^2 \leq q\|x_k - \overline{x}\|$

Damit ist die Folge der x_k wohldefiniert. Alle x_k liegen in $\mathcal{K}$ und erfüllen (2).
Aus (1) folgt weiter $\|x_{k+1} - \overline{x}\| \leq \frac{\omega}{2}\|x_k - \overline{x}\|^2$
Das bedeutet $\frac{\omega}{2}\|x_k - \overline{x}\| \leq q^{2^k}$ also Konvergenz der x_k gegen $\overline{x}$.
□

Folgerung: lokal quadratische Konvergenz
Bemerkung: Die Konstante ω ist so etwas wie eine Lipschitzkonstante von f'.
Sie ist affin invariant: Sei also $g(x) = Af(x)$, A invertierbar, nichtsingulär.
Dann gilt wegen $g'(x) = Af'(x)$
$g'(x)^{-1}[g(y) - g(x)] = f'(x)^{-1}[f(y) - f(x)]$

globales Konvergenzverhalten:
sehr kompliziert

Beispiel 4.5

Nullstellen des konvergenten Polynoms $z^3 - 1$:
äquivalentes reelles System für x, y $(z = x + iy)$
$x^3 - 3xy^2 - 1 = 0$
$-3xy^2 + y^3 = 0$

Vergrößerung des Einzugsbereichs
Benutze eine Dämpfungsstrategie, um zu garantieren, dass eine vorgegebene, durch ein Skalarprodukt induzierte Norm des Residuums $f(x_k)$ von Schritt zu Schritt hinreichend stark abnimmt. Solche Strategien beruhen auf der Beobachtung, dass die Funktion
$\psi(\theta) = \|f(x - \theta f'(x)^{-1} f(x))\|^2$
bei vorgegebener durch ein Skalarprodukt induzierter Norm die Ableitung
$\psi'(0) = -2\|f(x)\|^2$
hat. Eine ganz einfache (nicht die beste) Version eines solchen Verfahrens sieht wie folgt aus:
Start: (Schritt 0)
Gebe x_0 vor. Berechne $f(x_0), \|f(x_0)\|$

Schritt k+1:
$x_k, f(x_k), \|f(x_k)\|$ sind bekannt
1) Berechne $f'(x_k)$
2) Berechne die Korrekturrichtung d als Lösung des GLS $f'(x_k)d = -f(x_k)$
Setze $\theta = 1$
3) Berechne $x_k + \theta d, f(x_k + \theta d), \|f(x_k + \theta d)\|$
4) $\|f(x_k + \theta d)\| \leq \left(1 - \frac{\theta}{2}\right)\|f(x_k)\|$?
ja: Setze $x_{k+1} = x_k + \theta d$
nein: Halbiere θ und springe auf 3)

Konvergenz:
bei Start in kompakter Niveaumenge $\{x | \|f(x)\| \leq \|f(x_0)\|\} =: N(x_0)$, falls $f'(x)$ auf $N(x_0)$ nichtsingulär ist.

affin invariante Version:

$x_k, f'(x_k)^{-1}f(x_k), \|f(x_k)^{-1}f(x_k)\|_2$ gegeben.

1) setze $\theta = 1$

2) Berechne mit $d := -f'(x_k)^{-1}f(x_k), x_k + \theta d, f(x_k + \theta d)$ sowie

$\overline{d} = f'(x_k)^{-1}f(x_k + \theta d), \|\overline{d}\|_2$

3) $\|\overline{d}\|_2 \leq \|d\|$?

ja: Setze $x_{k+1} = x_k + \theta d, f(x_{k+1}) = f(x_k + \theta d)$

Berechne $f'(x_{k+1})$ und $f'(x_{k+1})^{-1}f(x_{k+1}), \|f'(x_{k+1})^{-1}f(x_{k+1})\|$

nein: halbiere θ und gehen auf 2)

Abwandlungen des Newtonverfahrens:

1) $f'(x)$ wird über mehrere Schritte konstant gehalten

2) $f'(x)$ wird nicht exakt, sondern durch numerische Differentiation berechnet (Differenzenquiotienten)

3) $f'(x)^{-1}$ wird in singulären Punkten durch $f'(x)^+$ (Pseudoinverse) ersetzt

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

4) Rangmodifikationsverfahren

$x_{k+1} = x_k - H_k f(x_k)$

$\text{rang}(H_k - H_{k+1}) = 1 \text{ oder } 2$

5) näherungsweise Lösung des linearen Gleichungssystems

$f'(x_k)(x_{k+1} - x_2) = -f(x_k)$

5. Lineare Ausgleichsprobleme

5.1 Lineare Ausgleichsprobleme in unitären Räumen

Definition 5.1

Sei V ein reeller Vektorraum. Eine Abbildung $\langle\ ,\ \rangle : V \times V \to \mathbb{R}$ heißt Skalarprodukt auf V, wenn für alle $x, y, z \in V$ und alle $\alpha, \beta \in \mathbb{R}$ die folgenden Bedingungen erfüllt sind.

(i) $\langle x, y\rangle = \langle y, x\rangle$

(ii) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$

(iii) $\langle x, x\rangle \geq 0, \langle x, x\rangle = 0 \iff x = 0$

Zusammen mit einem solchen Skalarprodukt wird V zu einem unitären Raum.

Satz 5.2

Ist $\langle\ ,\ \rangle$ ein Skalarprodukt auf dem reellen Vektorraum V, so gilt für alle $x, y \in V$ die Schwarze Ungleichung.

$\langle x, y\rangle \leq \sqrt{\langle x, x\rangle\langle y, y\rangle}$

Durch $\|x\| := \sqrt{\langle x, x\rangle}$ wird eine Norm auf V induziert.

Es ist $\|x - y\|^2 + \|x + y\|^2 = 2\|x\|^2 + 2\|y\|^2$ und

$\langle x, y\rangle = \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2)$

Beispiel 5.3

Der $\mathbb{R}^n$ und dem Standardskalarprodukt

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

und der durch es induzierten euklidischen Norm

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

Beispiel 5.4

Der Raum $C[a, b]$ aller stetigen Funktionen $f : [a, b] \to \mathbb{R}$ mit dem Skalarprodukt $\langle f, g\rangle = \int_a^b f(x)g(x)\mathrm{dx}$ und der L_2-Norm

$$\|f\| = (\int_a^b f(x)^2 \mathrm{dx})^{\frac{1}{2}}$$

Beispiel 5.5

Der Raum $C_0^1[a,b]$ aller stetig differenzierbaren Funktionen $f : [a,b] \to \mathbb{R}$ und $f(a) = 0, f(b) = 0$ und dem Skalarprodukt.

$\langle f, g\rangle = \int_a^b f'(x)g'(x)\mathrm{dx}$

Definition 5.6

Sei V ein reeller Vektorraum mit der Norm $\|.\|$ und M ein linearer Teilraum von V. Dann heißt $y^* \in M$ ein Element bester Approximation an $x \in V$ bezüglich der gegebenen Norm, wenn für $y \in M$ $\|x - y^*\| \leq \|x - y\|$ gilt.

Satz 5.7

Die Norm werde durch ein Skalarprodukt $\langle .\,, .\rangle$ induziert. Dann ist $y^* \in M$ genau dann ein Element bester Approximation an $x \in V$, wenn für alle $v \in M$ $\langle x - y^*, v\rangle = 0$ ist.Es gibt höchstens ein Element bester Approximation.

Beweis: Sei $y^* \in M$ ein Element bester Approximation an $x \in V$. Sei $v \in M$. Dann nimmt die skalare Funktion $F(x) := \|x - (y^* + \alpha v)\|^2 = \|(x - y^*) - \alpha v\|$ $= \|x - y^*\|^2 - 2\langle x - y^*, v\rangle\alpha + \|v\|^2\alpha^2$ wegen $y^* + \alpha v$ in $\alpha = 0$ ein globales Minimum an. Wegen $F'(\alpha) = -2\langle x - y^*, v\rangle + 2\|v\|\alpha$ und $F'(0) = -2\langle x - y^*, v\rangle$ folgt daraus $\langle x - y^*, v\rangle = 0$.
Sei nun umgekehrt $\langle x - y^*, v\rangle = 0$ für alle $v \in M$. Dann ist für alle $v \in M$: $\|x - (y^* + v)\|^2 = \|x - y^*\|^2 - 2\langle x - y^*, v\rangle + \|v\|^2$ $= \|x - y^*\|^2 + \|v\|^2$ also $\|x - (x^* + v)\| > \|x - y^*\|$ für $v \neq 0$.
Damit ist y^* ein Element bester Approximation und $\|x - y^*\| < \|x - y\|$ für alle $y \neq y^*$ aus M.□

Satz 5.8

Ist M endlichdimensional und bilden die Vektoren $v_1, \ldots, v_n$ eine Basis von M, so ist $y^* = \sum_{i=1}^{n} a_i v_i$ genau dann ein Element bester Approximation an $x \in V$, wenn die Koeffizienten α_i das Normalgleichungssystem $\sum_{i=1}^{n} \langle v_i, v_j\rangle\, \alpha_i = \langle x, v_j\rangle, j = 1, \ldots, n$ erfüllen.
Der Koeffizientenmatrix dieses Systems ist symmetrisch und positiv definit.
Das Normalgleichungssystem ist damit eindeutig lösbar.

Beweis: Die charakterisierende Bedingung aus Satz 5.7 ist genau dann erfüllt, wenn für $j = 1, \dots, n$
$< x - \sum_{i=1}^{n} \alpha_i v_i, v_j > = 0$
gilt, das heißt wenn α_i das Gleichungssystem
$\sum_{i=1}^{n} \langle v_i, v_j \rangle \alpha_i = < x, v_j >$, $j = 1, \dots, n$ erfüllen.
Die Koeffizientenmatrix A mit $A|_{ij} = \langle v_j, v_i \rangle$ dieses Systems ist wegen der Symmetrie des Skalarprodukts symmetrisch. Für alle Vektoren
$\alpha = (\alpha_1, \dots, \alpha_n)^T \in \mathbb{R}^n$ gilt $\alpha^T A \alpha = \sum_{i,j=1}^{n} \langle v_i, v_j \rangle \alpha_i \alpha_j$
$= < \sum_{i=1}^{n} \alpha_i v_i, \sum_{j=1}^{n} \alpha_j v_j > = \| \sum_{i=1}^{n} \alpha_i v_i \|^2 \geq 0$ und genau dann
$\alpha^T A \alpha = 0$, wenn $\sum_{i=1}^{n} \alpha_i v_i = 0$ ist. Da die
v_i als Basis linear unabhängig sind, ist dies genau dann der Fall,
wenn $\alpha_1, \dots, \alpha_n = 0$ oder $\alpha = 0$ ist.
□

Beispiel 5.9

Sei $V = C[0,1]$, versehen mit dem Skalarprodukt $\langle f, g \rangle = \int_0^1 f(x) g(x) \mathrm{dx}$.
Sei M der Raum aller Polynome n-ten Grades. Zu einer gegebenen Funktion $f : [0,1] \to \mathbb{R}$ ist das Polynom $P = P^*$ vom Grad n gesucht für das der Abstand $\|f - P\| = (\int_0^1 |f(x) - P(x)|^2 \mathrm{dx})^{1/2}$ minimal wird.
Eine Basis des Raums aller Polynome n-ten Grades bilden die Monome $\varphi_i(x) = x^i, i = 0, \dots, n$. Stellt man das optimale Polynom P^* in der Form $P^*(x) = \sum_{i=0}^{n} \alpha_i x^i = \sum_{i=0}^{n} \alpha_i \varphi_i(x)$ dar, müssen die α_i das System
$\sum_{i=0}^{n} < \varphi_i, \varphi_j > \alpha_i = < f, f_j >$, $j = 0, \dots, n$ erfüllen.
Wegen $< \varphi_i, \varphi_j > = \int_0^1 x^i x^j \mathrm{dx} = \frac{1}{i+j+1}$ ist die Koeffizientenmatrix dieses Systems die Hilbertmatrix H der Dimension $n+1$ mit den Koeffizienten $H|_{ij} = \frac{1}{i+j+1}$
Die Hilbertmatrizen sind berüchtigt für ihre schlechte Kondition.
Das beschriebene Gleichungssystem ist daher ungeeignet, um das gegebene Ausgleichsproblem zu lösen.

Ausweg: man muss zu einer besser konditionierten Basis übergehen.
Ideal ist eine Orthonormalbasis mit Basiselementen $\varphi_i, i = 0, \dots, n$
mit $\langle \varphi_i, \varphi_j \rangle = \delta_{ij}$. Dann ist $P^*(x) = \sum_{i=0}^{n} < f, \varphi_i > \varphi_i$.

Fourieranalyse

Gesucht ist die beste Approximation einer 2π-periodischen Funktion
$f : \mathbb{R} \to \mathbb{C}$ durch ein trigonometrisches Polynom
$\sum_{k=-n}^{n} a_k e^{ikx}$
n-ten Grades im Sinne der durch das Skalarprodukt
$< f, g > = \int_{-\pi}^{\pi} f(x)\overline{g(x)}\mathrm{dx}$
induzierten Norm. Die Monome $f_k(x) = e^{ikx}$ bilden eine Basis dieses Raumes.
Wegen $\langle \varphi_k, \varphi_l \rangle = 2\pi\delta_{kl}$ ist $\sum_{k=-n}^{n} \widehat{f}(k)e^{ikx}, \widehat{f}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-ikx}\mathrm{dx}$
die optimale Lösung. Die Koeffizienten $\widehat{f}(k)$,$k \in \mathbb{Z}$, sind die
Fourierkoeffizienten der Funktion f.

Beispiel 5.10

Sei V der Vektorraum $C[a, b]$ mit dem Skalarprodukt
$\langle f, g \rangle = \int_a^b f(x)g(x)\mathrm{dx}$. Mit $h = \frac{b-a}{n}, n \in \mathbb{N}$ fest, sei
$x_k = a + kh = a + \frac{b-a}{n}k, k = 0, \dots, n$
Der Raum M besteh aus allen Funktionen aus $C[a, b]$, die auf den
Teilintervallen $[x_k, x_{k+1}], k = 0, \dots, n-1$ linear sind.
Eine Funktion M ist durch ihre Werte in den Punkten $x_0, \dots, x_n$
eindeutig festgelegt. Eine Basis von M bilden die Funktionen $\varphi_0, \dots, \varphi_n$
mit $\varphi_i(x_k) = \delta_{ik}$. Diese Funktionen haben lokale Träger.

Damit ist $\langle \varphi_i, \varphi_j \rangle = 0$, falls $|i - j| > 1$ ist. Die Koeffizientenmatrix
ist damit eine Tridiagonalmatrix. Sie ist

$$A = h \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & 0 & \cdots & 0 \\ \frac{1}{6} & \frac{2}{3} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \frac{2}{3} & \frac{1}{6} \\ 0 & \cdots & 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix}$$

Sie ist sehr gut Konditioniert. Ihre Zeilensummennorm ist
$\|A\| = \frac{h}{6} + \frac{2}{3}h + \frac{h}{6} = h$
Es ist
$\frac{h}{3}x_1 = Ax|_1 - \frac{h}{6}x_2$
$\frac{2}{3}hx_i = Ax|_i - \frac{h}{6}x_{i-1} - \frac{h}{6}x_{i+1}, i = 1, \dots, n-1$
$\frac{h}{3}x_n = Ax|_n - \frac{h}{6}x_{n-1}$
Daraus folgt
$|x_1| \leq \frac{3}{h}\|Ax\|_\infty + \frac{1}{2}\|x\|_\infty$
$|x_i| \leq \frac{3}{2h}\|Ax\|_\infty + \frac{1}{2}\|x\|_\infty$
$|x_n| \leq \frac{3}{h}\|Ax\|_\infty + \frac{1}{2}\|x\|_\infty$
und damit
$\|x\|_\infty \leq \frac{3}{h}\|Ax\| + \frac{1}{2}\|x\|$ oder
$\|x\| \leq \frac{6}{h}\|Ax\|$, d.h. $\|A^{-1}\| \leq \frac{6}{h}$
$\kappa_\infty(A) = \|A\|\|A^{-1}\| \leq 6$

5.2 Lineare Ausgleichsprobleme im $\mathbb{R}^n$

Die QR Zerlegung einer Matrix

Beispiel: Gesucht ist die beste Approximation einer Funktion f im Sinne des Abstandsmaßes
$(\sum_{i=1}^{n}(f(x_i) - \varphi(x_i))^2)^{\frac{1}{2}}$
durch eine Funktion φ in einem endlich dimensionalen Teilraum.
Abstrakter: gesucht ist die beste Approximation des Vektors
$(f(x_1), \dots, f(x_n))^T \in \mathbb{R}^n$ durch einen Vektor $(\varphi(x_1), \dots, \varphi(x_n))^T$
aus einem gegebenen Teilraum des $\mathbb{R}^n$.

allgemein: Gegeben sei eine $(m \times n)$-Matrix A (mit m Zeilen und $n \leq m$ Spalten) eines Rangs $\leq n$ und ein Vektor $b \in \mathbb{R}^n$. Gesucht ist die beste Approximation von b durch eine Linearkombination der Spalten von A.
Summe des euklidischen Skalarprodukts, also ein Vektor $x^* \in \mathbb{R}^n$ mit
$\|Ax^* - b\|_2 \leq \|Ax - b\|_2, x \in \mathbb{R}^n$.

Satz 5.11

Hat die Matrix A den Rang n, sind also die Spalten von A linear unabhängig, so ist x^* eindeutig bestimmt und die Lösung des Gleichungssystems

$A^\top Ax^* = A^\top b.$

Beweis: Seien $a_1, \dots, a_n \in \mathbb{R}^m$ die Spalten von A und M der von diesen Spalten ausgespannte Teilraum des $\mathbb{R}^m$, d.h. das Bild des $\mathbb{R}^n$ unter A.

Gesucht sind dann die eindeutig bestimmten Koeffizienten x_i in der Linearkombination $\sum_{i=1}^{n} x_i a_i \in M$, für die für alle $v \in M$
$\|\sum_{i=1}^{n} x_i a_i - b\|_2 \leq \|v - b\|_2$ ist.
Nach Satz 5.8 erfüllen die x_i das Normalgleichungssystem
$\sum_{i=1}^{n} < a_i, a_j > x_i = < b, a_j >, j = 1, \dots, n$
das in Matrixform mit $x = \langle x_1, \dots, x_n \rangle^\top$
$A^* Ax = A^* b$
lautet
□

Bemerkung: Numerisch ist es nicht immer günstig diese Charakterisierung zu nutzen.

Beispiel 5.12
Für die (6×5) Matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ \varepsilon & 0 & \cdots & \cdots & 0 \\ 0 & \varepsilon & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \varepsilon \end{pmatrix}$$

vom Rang 5 ist

$$A^\top A = \begin{pmatrix} 1+\varepsilon^2 & 1 & \cdots & \cdots & 1 \\ 1 & 1+\varepsilon^2 & \ddots & & \vdots \\ \vdots & \ddots & 1+\varepsilon^2 & \ddots & \vdots \\ \vdots & & \ddots & 1+\varepsilon^2 & 1 \\ 1 & \cdots & \cdots & 1 & 1+\varepsilon^2 \end{pmatrix}$$

Für $\varepsilon = \frac{1}{2} \cdot 10^{-5}$ ist $\varepsilon^2 = \frac{1}{4} \cdot 10^{-10}$ und deshalb in 10 stelliger Arithmetik:

$$\mathrm{gl}(A^\top A) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

eine Matrix vom Rang 1.
□

Weiteres Problem:
Der Rang von A kann kleiner als n sein. Die optimale Approximation von b durch die Spalten von A ist dann nach wie vor eindeutig, nicht aber die Koeffizientenmatrix.

Satz 5.13
Unter allen Vektoren $x \in \mathbb{R}^n$ mit $\|Ax - b\|_2 = \min_{y \in \mathbb{R}^n} \|Ay - b\|_2$ gibt es genau einen Vektor x^* kürzester euklidischer Länge.

Beweis: Sei $M \subseteq \mathbb{R}^m$ der Raum, der von den Spalten von A aufgespannt wird. Dann gibt es nach Satz 5.8 genau ein y^* mit $\|y^* - b\|_2 \leq \|y - b\|_2$ für alle $y \in M$. Es gilt genau dann $\|Ax^* - b\|_2 \leq \|Ax - b\|_2$ für alle $x \in \mathbb{R}^n$, wenn $Ax^* = y^*$ ist. Ist x_0^* ein derartiges x^*, so ist die Menge aller dieser x^* von der Form $x^* = x_0^* - v$,$v \in \mathrm{Kern}(A)$. Der Kern von A ist ein endlichdimensionaler Teilraum des $\mathbb{R}^n$.
Die Länge von x^* wird also genau dann minimal, wenn v die eindeutig bestimmte beste Approximation von x_0^* durch ein Element von $\mathrm{Kern}(A)$ ist.

Aufgabe: Man löse das Ausgleichsproblem $\|Ax - b\|_2 \to \min$ in numerisch stabiler Weise!

Definition 5.14
Eine QR-Zerlegung der $(m \times n)$-Matrix A ist eine Zerlegung von A in das Produkt $A = QR$ einer orthogonalen $(m \times m)$-Matrix Q und eine obere Dreiecksmatrix der Dimension $m \times n$.

Beobachtung 1: Die ersten k Spalten, $k \leq n$ von A lassen sich als Linearkombination der ersten k Spalten von Q darstellen.
Beobachtung 2: $\mathrm{Rang}(A) = \mathrm{Rang}(R)$

Beobachtung 3: Hat A vollen Rang, so sind die ersten n Spalten von Q bis auf das Vorzeichen eindeutig bestimmt.

Berechnung der ersten n Spalten von Q:
durch Orhtogonalisierung nach Gram-Schmidt?
leider instabil!

Lösung des Ausgleichsproblems mittels einer QR-Zerlegung von A:
$\|Ax-b\|_2 = \|QRx-b\|_2$
$= \|Q(Rx-Q^\top b)\|_2$ (Wegen $QQ^\top = I$)
$= \|Rx-Q^\top b\|_2$ (wegen $\|Qv\|_2 = \|v\|_2$)
Sei nun
$R = \begin{pmatrix} R_1 \\ 0 \\ n \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix}, Q^\top b = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix}$
Dann ist genau $\|Ax^*-b\|_2 \leq \|Ax-b\|_2$ für alle $x \in \mathbb{R}^n$, wenn#
$\|R_1x^*-c_1\|_2 \leq \|R_1x-c_1\|_2$ für alle $x \in \mathbb{R}^n$ und genau dann
$\|Ax^*-b\|_2 \geq \|c_2\|_2$, wenn $R_1x^* = c_1$ ist.

Bemerkung: Ist A eine $(n \times n)$-Matrix $A = QR$ eine QR Zerlegung von A, so kann man das lineare Gleichungssystem $Ax = b$ wie folgt lösen:
1) Löse $Qy = b$, d.h. setze $y = Q^\top b$
2) Löse $Rx = y$

Definition 5.15
Sei $v \in \mathbb{R}^m$ und $\|v\|_2 = 1$. Dann definiert $Hx = x - 2< v, x >v, x \in \mathbb{R}^m$, die zu v gehörige Householdertransformation.
Sie hat die Matrixdarstellung $H = I - 2vv^\top$

geometrische Interpretation:
H ist die Spiegelung an der Hyperebene
$\mathcal{H} = \{x \in \mathbb{R}^m | \langle v, x\rangle = 0\}$
Für alle $x \in \mathbb{R}^m$ ist nämlich
$x = \underbrace{x - \langle v, x\rangle v}_{\in \mathcal{H}} + \underbrace{\langle v, x\rangle v}_{\in \mathcal{H}^\perp}$
$Hx = \underbrace{x - \langle v, x\rangle v}_{\in \mathcal{H}} - \underbrace{\langle v, x\rangle v}_{\in \mathcal{H}^\perp}$

Satz 5.16

Householdertransformationen sind orthogonal und selbstadjungiert:
Ist $\|v\|_2 = 1$, so gelten für die Matrix $H = I - 2vv^\top$ die Beziehungen
$H^\top H = I, H^\top = H$

Beweis:
$H|_{ij} = \delta_{ij} - 2v_i v_j = H|_{ij}$
$H^\top Hx = HHx = (x - 2< v, x >v) - 2\langle x - 2\langle v, x\rangle v, v\rangle v$
$= x - 2\langle v, x\rangle v - 2\langle v, x\rangle v + 4\langle v, x\rangle\langle v, v\rangle v = x$
Wegen $\langle v, v\rangle = 1$.
□

Satz 5.17

Sei $a \in \mathbb{R}^m, a \neq 0$, gegeben. Sei $w = a + \sigma\|a\|_2 e_1$ mit $\sigma = +1$,
falls $a|_1 = a_1 \geq 0$ ist, und $\sigma = -1$, falls $a|_1 < 0$ ist. Dann ist
$\|w\|_2^2 = 2\|a\|_2(\|a\|_2 + |a_1|)$. Setzt man nun $H = I - 2vv^\top, v = \frac{1}{\|w\|_2}w$,
so gilt: $Ha = -\sigma\|a\|_2 e_1$

Beweis: $\|w\|^2 = \langle a + \sigma\|a\|e_1, a + \sigma\|a\|e_1\rangle$
$= \langle a, a\rangle + 2\sigma\|a\|\langle a, e_1\rangle + \sigma^2\|a\|^2\langle e_1, e_1\rangle$
$= \|a\|^2 + 2|a_1|\|a\| + \|a\|^2$
$= 2\|a\|(\|a\| + |a_1|)$
$\geq 2\|a\|^2$
> 0

$Ha = a - 2\langle a, v\rangle v$
$= a - 2\frac{\langle a,v\rangle}{\|w\|^2}w$
$= a - 2\frac{\langle a,a\rangle + \sigma\|a\|\langle a,e_1\rangle}{2\|a\|(\|a\|+|a_1|)}w$
$= a - 2\frac{\|a\|^2 + |a_1|\|a\|}{2\|a\|(\|a\|+|a_1|)}w$
$= a - w$
$= a - (a + \sigma\|a\|e_1)$
$= -\sigma\|a\|e_1$
□

Bemerkung: Bei der Berechnung dieser Transformation kann
keine Auslöschung durch Addition von Zahlen unterschiedlichen Vorzeichens
auftreten!

Berechnung der QR-Zerlegung einer Matrix

Falls die erste Spalte von A ungleich 0 ist, wähle eine Householder-
transformation H, die sie in ein Vielfaches des 1. Einheitsvektors
transformiert.
(Ist die erste Spalte =0, setze man $H = I - 2vv^\top, v = 0$)
Schritt $k+1, k < n$
Man wähle eine Householdermatrix $H_{k+1} = I - 2vv^\top$ mit einem Vektor
$v = (0, \dots, 0, v_{k+1}, \dots, v_m)^\top$, die die erste Spalte der Restmatrix in einen
Vektor $x = (x_1, \dots, x_k, x_{k+1}, 0, \dots, 0)^\top$ transformiert.
(Gehe nach Satz 5.17 vor, falls die erste Spalte der Restmatrix $\neq 0$ ist,
sonst setze $v = 0$)

Ergebnis: Orthogonale Matrizen $H_1, \dots, H_n$ der Form $H_k = I - 2vv^\top$
mit $\|\sigma\|_2 = 1$ oder $\|v\|_2 = 0$ und eine obere Dreiecksmatrix R und
$H_n \cdots H_1 A = R$ oder $A = (H_n \cdots H_1)^1 R = H_1 \cdots H_n R = QR$

Bemerkung: Q wird nicht explizit berechnet. Man speichert Stattdessen
die Vektoren v ab, die die Householdertransformationen bestimmen.

Speicheraufwand:
$mn + n$ Speicherplätze für die Faktoren von Q und für R.

Aufwand k-ter Schritt:
Berechnung von $H = I - 2vv^\top$ und $Ha = -\text{sgn}(a_1)\|a\|_2 e_1$:
$\|a\|_2$: $m - k + 1$ Multiplikationen, 1 Quadratwurzel
$w = a + \text{sgn}(a_1)\|a\|_2 e_1$: $-$
$\|w\|_2 = (2\|a\|_2(\|a\| + |a_1|))^{1/2}$:1 Quadratwurzel
$r = \frac{1}{\|w\|_2} w$: $m - k + 1$ Divisionen
$\sum = 2(m - k + 1)$ Punktoperationen,2 Quadratwurzeln
Multiplikation der restlichen $n - k$ Spalten mit H
$Ha = a - 2\langle a, v\rangle v$ jeweils $2(m - k + 1)$ Punktoperationen

Gesamtaufwand:
$\sum_{k=1}^{n}[2(m-k+1)+2(m-k+1)(n-k)]$
$=(m-\frac{n-1}{3})(n^2+n)$

numerische Stabilität:
ganz hervorrangend! Bei quadratischen Matrizen viel besser als LR-Zerlegung, ja sogar besser als bei der Zerlegung symmetrischer positiv definiter Matrizen.

Lösung linearer Gleichungssysteme
$Ax=b \iff x=R^{-1}Q^\top b$
Aufwand:
QR-Zerlegung: $\frac{2}{3}n^3+n^2+\mathcal{O}(N)$
$Q^\top b =: y \quad : n^2$
$R^{-1}y \quad : \frac{1}{2}n^2$
asymptotisch doppelt so hoch wie bei der LR-Zerlegung

aber:
ungemein stabil!

Householder-Orthogonalisierung mit Pivotsuche:
$H_n \cdots H_1 A P_1 \cdots P_{n-1} = R$
$A = H_1 \cdots H_n R P_{n-1} \cdots P_1$
Wähle die Transpositionsmatrix P_{k+1} (Vertauschung zweier Komponenten) so, dass die erste Spalte der Restmatrix die längste Spalte der Restmatrix ist.

5.3 Singulärwertzerlegung und Pseudoinverse

Problem: Wie findet man den Vektor kürzester euklidischer Länge, der $\|Ax-b\|_2$ minimiert?

Satz 5.18

Zu jeder $(m \times n)$-Matrix A,$n \leq m$ gibt es eine orthogonale $(m \times m)$-Matrix U und eine orthogonale $(n \times n)$-Matrix V sowie eine Diagonalmatrix D mit $A = UDV^\top$. Diese Zerlegung heißt eine Singulärwertzerlegung von A.
$D = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$

Beweis: Seien $x \in \mathbb{R}^n$ und $y \in \mathbb{R}^m$ Vektoren mit $\|x\|_2 = 1, \|y\|_2 = 1$ und $A\,x = ry$, wobei $\sigma = \max\limits_{n \neq 0} \frac{\|Av\|_2}{\|v\|_2}$ die Norm von A sein soll.
Sei $x, v_2, \dots, v_n$ eine orthonormalbasis des $\mathbb{R}^n$und V_1 die $(n \times n)$-Orthogonalmatrix, deren Spalten diese Vektoren sind.

Sei $y, u_2, \dots, u_m$ eine Orthonormalbasis des $\mathbb{R}^m$ und U_1 die $(n \times n)$-Orthogonalmatrix, deren Spalten diese Vektoren sind. Dann ist wegen $Ax = \sigma y$:

$$U_1^\top A V_1 = \begin{pmatrix} \sigma & | & w^\top \\ \hline 0 & | & B \end{pmatrix} \begin{matrix} 1 \\ m-1 \end{matrix}$$
$$\begin{matrix} 1 & n-1 \end{matrix}$$

Bezeichne $\binom{\sigma}{w}$ die Aufspaltung des Vektors w in 2 Teile
Wegen $\left\|\binom{\sigma}{w}\right\|_2^2 = \sigma^2 + \|w\|_2^2$
$\left\|V_1^\top A V_1 \binom{\sigma}{w}\right\|_2^2 = \left\|\binom{\sigma^2 + w^\top w}{*}\right\|_2^2 \geq \sigma^2 + \|w\|_2^2$ ist
$\|A\|_2^2 = \left\|U_1^\top \mathrm{A V_1}\right\|_2^2 \geq \sigma^2 + \|w\|^2$
Wegen $\|A\|_2 = \sigma$ folgt daraus $\|w\| = 0$, also

$$U_1^\top \mathrm{A V_1} = \begin{pmatrix} \sigma & | & 0 \\ \hline 0 & | & B \end{pmatrix} \begin{matrix} 1 \\ m-1 \end{matrix}$$
$$\begin{matrix} 1 & n-1 \end{matrix}$$

Durch induktion folgt daraus die Existenz einer orthogonalen $(m \times m)$-Matrix U, einer orthogonalen $(n \times n)$-Matrix V und einer Diagonalmatrix D mit $U^\top A V = D$.
Bemerkung: Der Spektralsatz besagt, dass sich jede symmetrische Matrix $A \in \mathbb{R}^{n \times n}$ in der Form $A = UDU^\top$ einer diagonalmatrix D und einer orthogonalen Matrix U schreiben lässt.
Daher lässt sich satz 5.18 als Verwandter des Spektralsatzes deuten.

Definiton 5.19
Ist $A = UDV^\top$ eine Singulärwertzerlegung der $(m \times n)$ Matrix A, so ist die $(n \times m)$ Matrix $A^+ = VD^+U^\top$ eine Pseudoinverse von A.
Also ist $D = \begin{pmatrix} D_1 \\ \hline 0 \end{pmatrix}$ mit $D_1 = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$ so ist D^+ gegeben durch
$D^+ = \left(D_1^+, 0\right)$ mit $D_1^+ = \mathrm{diagl}(\mu_1, \dots, \mu_n)$
und $\mu_i = \begin{cases} \frac{1}{\sigma_i}, \sigma_i \neq 0 \\ 0 \ , \text{sonst} \end{cases}$

Jetzt können wir die definitive Lösung unsere Ausgleichsproblems angeben.

Satz 5.20

Ist A^+ eine Pseudoinverse der $(m \times n)$ Matrix A und $b \in \mathbb{R}^m$, so ist
$x^* = A^+ b$ der Vektor x kürzester euklidischer Länge mit
$\|Ax - b\|_2 = \inf\limits_{y \in \mathbb{R}^n} \|Ay - b\|_2$

Beweis: Wegen $\|Ax - b\|_2 = \left\|UDV^\top x - b\right\|_2 = \left\|U\left(DV^\top x - U^\top b\right)\right\|_2$
$= \left\|DV^\top x - U^\top b\right\|_2$
ist x^* genau dann die gesuchte Lösung kürzerster Länge, wenn $y^* = V^\top x^*$
den Vektor kürzester Länge mit $\left\|Dy - U^\top b\right\|_2 = \min\limits_{z \in \mathbb{R}^n} \left\|Dz - U^\top b\right\|_2$ ist.
Wegen $\left\|Dy - U^\top b\right\|_2 = \text{min!}$ Gilt also $y^* = D^+ U^\top b$ und somit
$x^* = Vy^* = VD^+U^\top b = A^+ b$
□

Satz 5.21

Jede $(m \times n)$ Matrix $A, n \leq m$ besitzt genau eine Pseudoinverse.

Beweis: Nach Satz 5.20 ist für jedes $b \in \mathbb{R}^n$ $x^* = A^+ b$ ein Vektor x
kürzester Länge mit $\|Ax - b\|_2 = \inf\limits_{y \in \mathbb{R}^n} \|Ay - b\|_2$ und nach Satz 5.13
ist dieser Vektor eindeutig bestimmt.
□

Bemerkung: Für $n > m$ ist die Pseudoinverse von $A^\top$ eindeutig und damit
auch von A.

Alternative Darstellung der Singulärwertzerlegung

$A = UDV^\top = \sum\limits_{i=1}^{n} \sigma_i u_i v_i^\top$
falls $U - [u_1, \dots, u_n, *, \dots, *]$
$V = [v_1, \dots, v_n]$
$D = \text{diag}(\sigma_1, \dots, \sigma_n)$

Annahme:
$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1}, \dots, \sigma_n = 0$
Dann gilt

$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$

$A^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^\top$

Frobeniusnorm: $\|A\|_F = (\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2)^{1/2}$

Satz 5.22

Die bestmögliche Approximation der $(m \times n)$ Matrix A mit der SWZ $A = \sum_{i=1}^{\min(n,m)} \sigma_i u_i v_i^\top$ mit $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\min(n,m)} \geq 0$ durch eine Matrix B vom Rang k,$k \leq \min(n, m)$, im Sinne der Frobeniusnorm ist.

$B^* = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$.

Beweis: Es ist $\|A - B\|_F = \|UDV^\top - B\|_F = \|U(D - U^\top BV)V^\top\|_F$.
Da für orthogonale Matrizen U und V
$\|UA\|_F = \|A\|_F, \|AV\|_F = \|A\|_F$
gilt, folgt $\|A - B\|_F = \|D - U^\top BV\|_F$

Die Beste Approximation von D durch eine Matrix vom Rang k ist offensichtlich $U^\top BV = \sum_{i=1}^{k} \sigma_i e_i e_i^\top$ woraus
$B = \sum_{i=1}^{k} \sigma_i U e_i e_i^\top V^\top = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$ folgt.
□

Berechnung einer SWZ
Schritt 1
Reduktion der Matrix A auf Bidiagonalform mit Hilfe von Householder Matrizen. Finiter Prozess
Schritt 2
Berechnung der SWZ der Bidiagonalmatrix mit mit dem QR-Algorithmus
iterativer Prozess

Householderschritt 1

$$\begin{pmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{pmatrix} \xrightarrow{L} \begin{pmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{pmatrix} \xrightarrow{R} \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{pmatrix} \xrightarrow{L} \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \end{pmatrix} \xrightarrow{R} \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & x & 0 \\ 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \end{pmatrix}$$

$$\xrightarrow{L} \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & x & 0 \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \\ 0 & 0 & 0 & x \end{pmatrix} \xrightarrow{L} \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & x & 0 \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

L, R :Multiplikation mit Householdermatrizen von links bzw. rechts.

Anwendung: "information retrieval"
Finde aus einer Sammlung von n Dokumenten die Dokumente heraus, die Informationen zu einem gegebenen Thema enthalten. Dazu muss die Information durch Schlagworte oder Suchtherme kodifiziert werden.
Vektorraummodell
Stelle eine Kollektion von m möglichen Suchtermen zusammen.
Dokumente wie Anfragen werden als Vektoren im $\mathbb{R}^m$ aufgefasst.
Die Koponente a_i eine solchen Vektors gibt an, wie oft der Suchterm, im gegebenen Dokument vorkommt, bzw. ob dieser Term gesucht wird.
Datenbasis
Eine $(m \times n)$ Matrix A; die Spalten von A entsprechen den einzelnen Dokumenten in der Datenbasis.

Bsp.: Internet Suchmaschine
$m = 300.000$ (Worte in engl. Sprache)
$n \approx 3 \cdot 10^9$ (Seiten im Netz)

ideale Suchmaschine
Berechne zur Anfrage $q \in \mathbb{R}^m$ den Zeilenvektor $q^\top A \in \mathbb{R}^n$.
Liste die von Null verschiedenen Komponenten des Vektors aus.
Faktisch unmöglich!

Stattdessen: Ersetze A durch Rang k Approximation $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$ und

berechne $q^\top A_k = \sum_{i=1}^{k} \sigma_i (q^\top u_i) v_i^\top$

5.4 Das Gauß-Newton-Verfahren

nichtlineares Ausgleichsproblem

Sei $f : \mathbb{R}^n \to \mathbb{R}^m, n \leq m$ gegeben. Finde $x^* \in \mathbb{R}^n$ mit
$\|f(x^*)\|_2 \leq \|f(x)\|_2, x \in \mathbb{R}^n$

Spezialfall
$f(x) = Ax - b$, lin. Ausgleichsproblem

Linearisierung um Näherungslösung x_k:
$f(x) \approx f(x_k) + f'(x_k)(x - x_k)$

Gauß-Newton-Verfahren
Finde neue Näherung x_{k+1} als Lösung von
$\|f(x_k) + f'(x_k)(x - x_k)\|_2 = \min!$

formal:
$x_{k+1} = x_k - f'(x_k)^+ f(x_k)$

6. Interpolation

"Darstellung von Funktionen durch ihre Werte in vorgegebenen Punkten."

gegeben:
eine von reellen Parametern $\alpha_0, \ldots, \alpha_n$ abhängige Funktionenklasse

$$\varphi(x; \alpha_0, \ldots, \alpha_n)$$

und $n + 1$ Stützstellen $x_0, \ldots, x_n$ im Definitionsbereich von φ.

gesucht:
(die?) Werte $\alpha_0, \ldots, \alpha_n$ mit $\varphi(x_i; \alpha_0, \ldots, \alpha_n) = f_i$ für $i = 0, \ldots, n$ für vorgegebene Werte f_i $(= f(x_i))$

hier:
linearer Ansatzfunktionenraum

$$\varphi(x;\alpha_0,\ldots,\alpha_n)=\sum_{i=1}^{n}\alpha_i\varphi_i(x)$$

$\varphi_0,\ldots,\varphi_n$ gegeben. $\rightarrow$ lineares Interpolationsproblem

Interpolation durch Polynome und stückweise Polynome (Splines)

Existenz?
Eindeutigkeit?
Fehler?

Satz 6.1
Folgende Aussagen sind äquivalent
i) Zu beliebig vorgegebenen Werten f_i gibt es eindeutig bestimmte Werte α_i mit

$$\sum_{i=0}^{n}\alpha_i\varphi_i(x_j)=f_j,\ \ j=0,\ldots,n$$

ii) Zu beliebig vorgegebenen Werten f_i gibt es Werte mit

$$\sum_{i=0}^{n}\alpha_i\varphi_i(x_j)=f_j,\ \ j=0,\ldots,n$$

iii) Aus

$$\sum_{i=0}^{n}\alpha_i\varphi_i(x_j)=0,\ \ j=0,\ldots,n$$

folgt $\alpha_0,\ldots,\alpha_n=0$.

Beweis: Für die Werte $\alpha_0,\ldots,\alpha_n$ gilt genau dann

$$\sum_{i=0}^{n}\alpha_i\varphi(x_j)=f_j$$

für $j=0,\ldots,n$, wenn die α_i das lineare Gleichungssysten

$$\begin{pmatrix}\varphi_0(x_0) & \cdots & \varphi_n(x_0)\\ \vdots & \ddots & \vdots\\ \varphi_0(x_n) & \cdots & \varphi_n(x_n)\end{pmatrix}\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_n\end{pmatrix}=\begin{pmatrix}f_0\\ \vdots\\ f_n\end{pmatrix}$$

erfüllen. Damit lässt sich die Behauptung auf eine fundamentale Strukturaussage über lineare Gleichungssysteme zurückführen. □

6.1 Interpolation durch Polynome

Satz 6.2

Die Stützstellen $x_0, \ldots, x_n \in \mathbb{R}$ seien paarweise verschieden. Dann gibt es ein eindeutig bestimmtes Polynom

$$P(x) = \sum_{i=0}^{n} \alpha_i x^i$$

vom Grad $\leq n$, das in den Stützstellen x_j vorgegebene Werte f_j annimmt. Dieses Polynom besitzt die Darstellung

$$P(x) = \sum_{j=0}^{n} f_j l_j(x)$$

mit den Polynomen

$$l_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} \quad , \quad j = 0, \ldots, n$$

den *Lagrangeschen Grundpolynomen* zu den Stützstellen $x_0, \ldots, x_n$.

Beweis: Es ist $l_j(x_j) = 1$ und $l_j(x_k) = 0$ für $k \neq j$. Damit ist

$$\sum_{j=0}^{n} f_j l_j(x_k) = \sum_{j=0}^{n} f_j \delta_{jk} = f_k, \quad k = 0, \ldots, n$$

also eine Lösung des Interpolationsproblems gefunden. Die Eindeutigkeit folgt aus Satz 6.1. □

Beispiel 6.3

Das Interpolationspolynom durch die Wertepaare $(-5, f(-5)), (-1, f(-1)), (5, f(5)), (1, f(1))$ und der Funktion
$f(x) = \frac{1}{1+x^2}$ lautet

$$P(x) = \frac{1}{26} l_0(x) + \frac{1}{2} l_1(x) + \frac{1}{2} l_2(x) + \frac{1}{26} l_3(x)$$

mit

$$l_0(x) = \frac{(x+1)(x-1)(x-5)}{(-5+1)(-5-1)(-5-5)} = \frac{1}{240}(x+1)(x-1)(x-5)\,.$$

anderer Beweis für die Eindeutigkeit:
Gilt

$$\sum_{i=0}^{n} \alpha_i x_j^i = \sum_{i=0}^{n} \widetilde{\alpha}_i x_j^i = f_j,\ \ j = 0, \ldots, n$$

so hat das Polynom

$$\sum_{i=0}^{n} (\alpha_i - \widetilde{\alpha}_i) x^i$$

$n+1$ Nullstellen $x_1, \ldots, x_n$. Nach dem Fundamentalsatz der Algebra ist daher $\alpha_i - \widetilde{\alpha}_i = 0$, $i = 1, \ldots, n$.

Newtonsche Darstellung

$$P(x) = a_0 + \sum_{i=1}^{n} a_i \prod_{j=0}^{i-1} (x - x_j)$$

vom Rechenaufwand her wesentlich günstiger!

Voraussetzung: Für

$$P_m(x) = a_0 + \sum_{i=1}^{m} a_i \prod_{j=0}^{i-1} (x - x_j)$$

gelte

$$P_m(x_k) = f_k,\ \ k = 0, 1, \ldots, m\,.$$

Beobachtung:
Dann gilt für jedes Polynom $P(x) = P_m(x) + a \prod_{j=0}^{m} (x - x_j)$ für $k = 0, \ldots, m$: $P(x_k) = f_k$.
Bestimme daher a so, dass $P(x_{m+1}) = f_{m+1}$ ist: also

$$a = a_{m+1} = \frac{f_{m+1} - P_m(x_{m+1})}{\prod_{j=0}^{m} (x_{m+1} - x_j)}\,.$$

Mit $a_0 = f_0$ folgt die Existenz und die Eindeutigkeit der Darstellung.

Schreibweise: $a_i = f[x_0, \ldots, x_i]$
Die dividierte Differenz $f[x_0, \ldots, x_m]$ ist der Leitkoeffizient des Polynoms

$$P_m(x) = f[x_0] + \sum_{i=1}^{m} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j) \, .$$

Satz 6.4
Das Interpolationsproblem zu den paarweise verschiedenen Stützstellen $x_0, \ldots, x_n$ und den Werten $f_0, \ldots, f_n$ besitzt die Newtonsche Darstellung

$$f[x_0] + \sum_{i=1}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j) \, .$$

Die dividierten Differenzen $f[x_0, \ldots, x_i]$ lassrn sich ausgehend von

$$f[x_i] = f_i, \;\; i = 0, \ldots, n$$

rekursiv über die Formel

$$f[x_k, \ldots, x_i] = \frac{f[x_{k+1}, \ldots, x_i] - f[x_k, \ldots x_{i-1}]}{x_i - x_k},$$

$0 \leq k < i \leq n$, berechnen.

Beweis: Sei Q_1 das Interpolationspolynom zu den Punkten $x_0, \ldots, x_{m-1}$ und Q_2 das Interpolationspolynom zu den Punkten $x_1, \ldots x_m$. Dann ist

$$P(x) = \frac{1}{x_m - x_0} \{(x - x_0) Q_2(x) + (x_m - x) Q_1(x)\}$$

das Interpoaltionspolynom zu den Punkten $x_0, \ldots, x_m$. Sein Leitkoeffizient ist einerseits $f[x_0, \ldots, x_m]$ und andererseits

$$\frac{f[x_1, \ldots, x_m] - f[x_0, \ldots, x_{m-1}]}{x_m - x_0}$$

woraus die angegebene Rekursionsformel folgt. □

Bemerkung:
Die dividierte Differenz $f[x_0, \ldots, x_m]$ ist von der Anordnung der Stützstellen unabhängig.
Beweis: Sie ist der Leitkoeffizient des Polynoms

$$\sum_{i=0}^{m} f_i \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{x - x_j}{x_i - x_j}$$

und damit gegeben durch

$$\sum_{i=0}^{m} f_i \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{1}{x_i - x_j}.$$

□

Dreiecksschema zur Berechnung der dividierten Differenzen

x_0	$f[x_0]$			
x_1	$f[x_1]$	$f[x_0,x_1] = \frac{f[x_1]-f[x_0]}{x_1-x_0}$		
x_2	$f[x_2]$	$f[x_1,x_2] = \frac{f[x_2]-f[x_1]}{x_2-x_1}$	$f[x_0,x_1,x_2] = \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}$	
x_3	$f[x_3]$	$f[x_2,x_3] = \frac{f[x_3]-f[x_2]}{x_3-x_2}$	$f[x_1,x_2,x_3] = \frac{f[x_2,x_3]-f[x_1,x_2]}{x_3-x_1}$	$f[x_0,x_1,x_2,x_3]$

neues Wertepaar ⇒ neue Zeile hinzufügen.

Beispiel 6.5 ....

Auswertung mit dem Horner-Schema:

$$P(x) = f[x_0] + \sum_{i=1}^{n} f[x_0,\ldots,x_i] \prod_{j=0}^{i-1} (x - x_j)$$

$$S_n = f[x_0,\ldots,x_n]$$

$$S_i = f[x_0,\ldots,x_i] + (x - x_i)S_{i+1}, \;\; i = n-1,\ldots,0$$

$$P(x) = S_0$$

Satz 6.6

Die Funktion $f : [a, b] \to \mathbb{R}$ sei $(n+1)$-mal stetig differenzierbar. Das Polynom P n-ten Grades interpoliere f in dem $n+1$ Punkten $a = x_0 < x_1 < ... < x_n = b$

Dann gilt in $x \in [a, b]$ für den fehler $f(x) - P(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!} \prod_{j=1}^{n} (x - x_j)$

mit einer Zwischenstelle $\eta \in (a, b)$, die von der Wahl der Stützstellen der Funktion f und dem betrachteten Punkt x abhängt.

Beweis: Sei x fest und ohne Einschränkung $x \neq x_j$ für $j = 0, ..., n$.

Betrachte die Hilfsfunktion $g(t) = (f(t) - P(t)) - (\prod_{j=0}^{n} \frac{t-x_j}{x-x_j})(f(x) - P(x))$.

Die Funktion g ist wie f selbst $(n+1)$-mal differenzierbar und verschwindet in den $n+2$ Punkten x und $x_0, ..., x_n$.

Nach dem Satz von Rolle verschwindet $g'(t)$ dann in mindestens $n+1$ von einander verschiedenen Punkten, die zwischen je zwei Nullstellen von $g(t)$ liegen.
Entsprechend verschindet g'' in mindestens n von einander verschiedenen Punkten aus (a, b) und $g^{(n+1)}$ schließlich noch in einem Punkt $\eta \in (a, b)$.
Wegen $g^{(n+1)}(t) = f^{(n+1)}(t) - \frac{(n+1)!}{\prod_{j=0}^{n} (x-x_j)}(f(x) - P(x))$ folgt daraus die Behauptung.
□

Folgerung:

$|f(x) - P(x)| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} |\prod_{j=0}^{n} (x - x_j)|$, falls $\|f^{(n+1)}\|_\infty = \sup_{a<i<b} |f^{(n+1)}(t)|$ existiert. Falls $|x - x_j| \leq Kh$ ist, folgt $|f(x) - P(x)| = \mathcal{O}(h^{n+1})$ für $h \to 0+$.

Beispiel 6.7

Die Funktion $f(x) = \ln(x)$ sei in $x_j = 1 + jh$,$j = 0, 1, 2, ...$ tabelliert.

stückweise lineare Interpolation

$\frac{x_{j+1}-x}{x_{j+1}-x_j} f(x_j) + \frac{x-x_j}{x_{j+1}-x_j} f(x_{j+1})$ für $x_j \leq x \leq x_{j+1}$

$|\frac{f''(\eta)}{2!}(x-x_j)(x-x_{j+1})| \leq \frac{1}{2}|(\frac{x_j+x_{j+1}}{2}-x_j)(\frac{x_j+x_{j+1}}{2}-x_{j+1})||f''(\eta)|$
$= \frac{1}{8}(x_{j+1}-x_j)^2|f''(\eta)| = \frac{1}{8}(x_{j+1}-x_j)^2|-\frac{1}{\eta^2}| \leq \frac{h^2}{8}$

stückweise quadratische Inteprolation
Interpolationspolynom zweiten Grades mit den Stützstellen x_{j-1}, x_j, x_{j+1}
für $x_j - \frac{h}{2} \leq x \leq x_j + \frac{h}{2}$.
$|\frac{f^{(3)}(\eta)}{3!}(x-x_{j-1})(x-x_j)(x-x_{j+1})| \leq \frac{h^3}{16}|f^{(3)}(\eta)|$
$= \frac{h^3}{16}\left|\frac{2}{\eta^3}\right| \leq \frac{h^3}{8} = h \cdot \frac{h^2}{8}$

Satz 6.8
Die Funktion f sei auf $[a,b]$ m-mal differenzierbar. Die Stützstellen $x_0, \ldots, x_n \in [a,b]$ seien von einander verschieden. Es sei $f[x_2] = f(x_k)$ für $k = 0, 1, \ldots, m$. Dann ist $f[x_0, \ldots, x_m] = \frac{f^{(m)}(\eta)}{m!}, \eta \in (a,b)$.

Beweis: Einerseits gilt nach Satz 6.3 für das Interpolationspolynom P zu den Stützstellen $x_0, \ldots, x_m$ und jeden von diesen Stützstellen verschiedenen Punkt x :
$f(x) = f[x_0] + \sum_{i=1}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1}(x-x_j) + f[x_0, \ldots, x_n, x] \prod_{j=0}^{n}(x-x_j)$
$= P(x) + f[x_0, \ldots, x_n, x] \prod_{j=0}^{n}(x-x_j)$
Andererseits ist nach Satz 6.6
$f(x) = P(x) + \frac{f^{(n+1)}(\eta)}{(n+1)!} \prod_{j=0}^{n}(x-x_j)$
Daraus folgt
$f[x_0, \ldots, x_n, x] = \frac{f^{(n+1)}(\eta)}{(n+1)!}$
also die Behauptung für $m = n+1$ und $x_{k+1} = x$.

Satz 6.9
Die Funktion $f : [a,b] \to \mathbb{R}$ sei $(n+1)$-mal differenzierbar und P sei das Interpolationspolynom von f in $a = x_0 < x_1 < \ldots < x_n = b$.
Dann gilt für $m \leq n$ und beliebiges $x \in [a,b]$
$f^{(m)}(x) - P^{(m)}(x) = \frac{f^{(n+1)}(\eta)}{(n-m+1)!} \prod_{j=0}^{n-m}(x-\theta_j)$
mit Punkten $\eta, \theta_j \in (a,b)$.

Beweis: Da $f(x) - P(x)$ in den $n+1$ Punkten $x_0, \dots, x_n$ verschwindet, gibt es nach dem Satz von Rolle voneinander verschiedene Punkte $\theta_0, \dots, \theta_{n-m}$ mit $f^{(m)}(\theta_i) - P^{(m)}(\theta_i) = 0$.
Mit der Hilfsfunktion
$g(t) = \left[f^{(m)}(t) - P^{(m)}(t)\right] - \left(\prod_{j=0}^{n-m} \frac{t-\theta_i}{x-\theta_j}\right)\left[f^{(m)}(x) - P^{(m)}(x)\right]$
kann man argumentieren wie im Beweis von Satz 6.6 und findet durch mehrmalige Anwendung des Satzes von Rolle die Fehlerdarstellung.
□

Folgerung:

$|f^{(m)}(x) - P^{(m)}(x)| \leq \frac{\|f^{(n+1)}\|}{(n-m+1)!} |\prod_{j=0}^{n-m} (x - \theta_j)|$

$|f^{(m)}(x) - P^{(m)}(x)| = 0(h^{n-m+1})$ falls $|x - x_j| = \mathcal{O}(h)$
Verlust einer Fehlerordnung pro Differentiationsordnung!

Bemerkung:
Wertet man die Ableitungen von Interpolationspolynomen an relativ zur Lage der Interpolationspunkte festen Stellen aus, erhält man Differenzenformeln zur Approximation von Ableitungen.

Beispiel 6.10
Differenziert man das Interpolationspolynom
$P(t) = \frac{(t-x)(t-x-h)}{(x-h-x)(x-h-x-h)} y(x-h) + \frac{(t-x)(t-x-h)}{(x-x+h)(x-x-h)} y(x) + \frac{(t-x+h)(t-x)}{(x+f-x+h)(x+h-x)} y(x+h)$
von $y(t)$ zu den Stützstellen $x-h$,x und $x+h$, so erhält man die Näherung
$P'(t) = ...$
für $y'(t)$, speziell also die Näherung $P'(x) = \frac{y(x+h)-y(x-h)}{2h}$ für $y'(x)$.
Der Fehler ist von der Größenordnung $\mathcal{O}(h^2)$.
□

Bemerkung: Neben den Funktionswerten kann man auch die Ableitungen der Funktion mit interpolieren: dann spricht man von HERMITE-Interpolation.

Beispiel 6.11
Gesucht ist ein Polynom P vom Grad $\leq 2n+1$ mit
$P(x_k) = f(x_k), k = 0, \dots, n$
$P'(x_k) = f'(x_k), k = 0, \dots, n$

Beispiel 6.12

Gesucht ist ein Polynom P n-ten Grades mit
$P^{(k)}(x_k) = f^{(k)}(x_0), k = 0, \dots, n$

6.2 Beste Approximation und optimale Stützstellen

Globales Verhalten

Bei Vergrößerung der Stützstellenzahl neigen Interpolationspolynome zu Überschwungeffekten.

Beispiel 6.13

$f(x) = |x|,\ -2 \leq x \leq 2$

$x_j = -2 + j\frac{4}{14},\ j = 0, \dots, 14$

Beispiel 6.14

$f(x) = \frac{1}{1+x^2}, \quad -5 \leq x \leq 5$

$x_j = -5 + j\frac{10}{14},\ j = 0, \dots, 14$

Definition 6.15

Für eine stetige Funktion $f : [a, b] \to \mathbb{R}$ ist

$$E_n(f) = \inf \left\{ \max_{a \leq x \leq b} |f(x) - P(x)| \;\middle|\; P \text{ Polynom } n - \text{ten Grades} \right\}$$

Ein Polynom P^* vom Grade n heißt Polynom bester Approximation an f, wenn $\max_{a \leq x \leq b} |f(x) - P^*(x)| = E_n(f)$ gilt. E_n heißt Fehler der besten Approximation.

Satz 6.16

Zu jeder stetigen Funktion f existiert genau ein eindeutig bestimmtes Polynom n-ten Grades bester Approximation an f.

Bemerkung: Dieses Polynom hängt nichtlinear von f ab.

Satz 6.17

Interpoliert das Polynom P_n die Funktion f in den Stützstellen $a \leq x_0 < x_1 < \dots < x_n \leq b$ so gilt $\max_{a \leq x \leq b} |f(x) - P(x)| \leq (1 + \lambda_n) E_n(f)$

Dabei ist $\lambda_n = \max_{a \leq x \leq b} \sum_{j=0}^{n} |l_j(x)|, l_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}$ die Lebesgue-Konstante

des Interpolationsoperators.

Beweis: Für alle Polynome P n-ten Grades ist

$f(x) - P_n(x) = f(x) - P(x) + P(x) - P_n(x)$

$= f(x) - P(x) + \sum_{i=0}^{n}(P(x_i) - P_n(x_i))l_i(x)$

$= f(x) - P(x) + \sum_{i=0}^{n}(P(x_i) - f(x_i))l_i(x)$

Damit

$|f(x) - P_n(x)| \leq |f(x) - P(x)| + \sum_{i=0}^{n}|P(x_i) - f(x_i)||l_i(x)|$

$\leq |1 + \sum_{i=0}^{n}|l_i(x)|| - \max_{a\leq x\leq b}|f(x) - P(x)|$

Bilde nun das Infimum über alle P. Dann folgt die Behauptung.

□

$\|f - P_n\| \leq (1 - \lambda_n)E_n(f)$

Frage1: Wie verhält sich $E_n(f)$?

Frage2: Wie verhält sich λ_n?

Für welche Stützstellenwahl wird λ_n minimal?

Satz 6.18 (Weierstraß)

Für alle stetigen Funktionen f gilt:

$\lim_{n\to\infty} E_n(f) = 0$

Satz 6.19 (Jackson)

Ist f r-mal stetig differenzierbar, so gilt für $n \geq r + 1$:

$E_n(f) \leq \frac{\pi}{2}\left(\frac{b-1}{2}\right)^r \left\|f^{(r)}\right\|_\infty \left(\frac{1}{n+1}\right)^r$.

Satz 6.20

Es gibt eine von n unabhängige Konstante $c > 0$ mit $\lambda_n \geq \frac{2}{\pi}\ln(n) - c$ für jede beliebige Stützstellenwahl.

Satz 6.21

Bei konstantem Stützstellenabstand wächst λ_n exponentiell.

Satz 6.22

Wählt man als Stützstellen die auf das Intervall Transformierten Tschebyscheff-Knoten.

$x_j = \frac{b-a}{2}\cos\left(\frac{2j+1}{2}\frac{\pi}{n+1}\right) + \frac{a+b}{2}$

$j = 0, \dots, n$, so gilt $\lambda_n \leq \frac{2}{\pi}\ln(n) + 1$

Bemerkung: Interpoliert man eine einmal stetig differenzierbare Funktion $f : [a, b] \to \mathbb{R}$ an dne Tschebyscheff-Knoten von Satz 6.22, so Konvergieren die Interpolytionspolynome P_n für $n \to \infty$ in der Supremumsnorm gegen f.

Es gilt nach Satz 6.17 $\max\limits_{a\leq x\leq b} |P_n(x) - f(x)| \leq (1 - \lambda_n)E_n(f)$ und nach Satz 6.19 $E_n(f) \leq C \cdot \frac{1}{n+1}$. Zusammen mit Satz 6.20 also
$\|P_n - f\|_\infty \leq \big(1 + \frac{2}{\pi}\ln(n) + 1\big) - C\frac{1}{n+1} \to 0$
für $n \to \infty$.

Berechnung des Interpolationspolynoms

$g(t) = f\big(\frac{b-a}{2}\cos(t) + \frac{a+b}{2}\big),\ t \in \mathbb{R}$
ist eine $2a$-periodische Funktion. Berechne die Koeffizienten a_r des trigonometrischen Interpolationspolynoms
$(S_n g)(t) - \sum\limits_{k=0}^{n} a_k \cos(kt)$
von g in den Punkten
$t_j = \frac{2j+1}{2}\frac{\pi}{n}, j \in \mathbb{Z}$
mit Hilfe der schnellen Fouriertransformation(FFT). Dazu benötigt man asymptotisch $n\log(n)$ Operationen. Dann ist
$(*)\ P(x).\sum\limits_{k00}^{n} a_k T_k\big(\frac{2}{b+a}x - \frac{b+a}{b-n}\big)$
das gesuchte Interpolationspolynom, wobei
$T_k = \cos(k \arccos(x)),\ -1 \leq x \leq 1$
das Tschebyscheffpolynom k-ten Grades ist. Die Tschebyscheff Polynome genügen der Rekursion $T_0(x) = 1, T_1(x) = x, T_k(x) = 2x\, T_{k-1}(x) - T_{k-2}(x)$
die sich mit Hilfe der trig. Identitär $\cos\alpha + \cos\beta = 2\cos\big(\frac{\alpha+\beta}{2}\big)\cos\big(\frac{\alpha-\beta}{2}\big)$
leicht beweisen lässt.

Auswertung von $(*)$
Setze $x' = \frac{2}{b-a}x - \frac{b+a}{b-a}, T_0 = 1, T_1 = x', s_1 = a_1 x' + a_0$
Berechne für $k = 2, 3, \dots, n$
$T_k = 2x' T_{k-1} - T_{k-2}$
$S_k = S_{k-1} + a_k T_k.$
Dann ist $P(x) = S_n$

Wegen $|x'| < 1$ ist diese Rekursion im Unterschied zum Horner-Schema stabil und benötigt ebenfalls nut $\mathcal{O}(n)$ Operationen.

6.3 Interpolation durch Splines

Definition 6.24

Im Intervall $[a, b]$ seien $N + 1$ Punkte $a = x_0 < x_1 < ... < x_N = k$ gegeben. Eine auf $[a, b]$ definierte Funktion S heißt genau dann m-mal stetig differenzierbare Splinefunktion k-ten Grades zu der gegebenen Unterteilung von $[a, b]$, wenn die Einschränkungen von S auf jedes Intervall $[x_i, x_{i+1}], i = 0, ... , N - 1$. Ein Polynom k-ten Grades ist und wenn S m-mal stetig differenzierbar ist.

Bemerkung: Sinnvoll ist nur der Fall $m < k$. Für $m \geq k$ wäre S wieder ein Polynom und durch die stückweise Definition gewonnene Flexibilität ginge verloren.

hier: Nur 2 mal stetig differenzierbar kubische Splines. Für viele Anwendungszwecke, bspw. in der Grafischen Datenverarbeitung, optimal.

Satz 6.25

Erfüllt die zweimal stetig differenzierbare Splinefunktion S die Interpolationsbedingungen $S(x_j) = f_j, j = 0, 1, ... , N$, so hat sie auf dem Intervall $x_j \leq x \leq x_{j+1}$ die Darstellung.

$S(x) = f_j \frac{x_{j+1}-x}{x_{j+1}-x_j} + f_{j+1} \frac{x-x_j}{x_{j+1}-x_j} - \frac{1}{6} S''(x_j) \frac{(x_{j+1}-x)(x-x_j)}{x_{j+1}-x_j} ((x_{j+1} - x) + (x_{j+1} - x_j)) - \frac{1}{6} S''(x_{j+1}) \frac{(x_{j+1}-x)(x-x_j)}{x_{j+1}-x_j}$

Weiter ist

$S'(x_j) = \frac{f_{j+1}-f_j}{x_{j+1}-x_j} - \left(\frac{1}{3} S''(x_j) + \frac{1}{6} S''(x_{j+1})\right)(x_{j+1} - x_j)$

$S'(x_{j+1}) = \frac{f_{j+1}-f_j}{x_{j+1}-x_j} + \left(\frac{1}{6} S''(x_j) + \frac{1}{3} S''(x_{j+1})\right)(x_{j+1} - x_j)$

Beweis: Da $S''(x)$ auf $[x_j, x_{j+1}]$ linear ist, gilt dort

$S'(x) = S''(x_j) \frac{x_{j+1}-x}{x_{j+1}-x_j} + S''(x_{j+1}) \frac{x-x_j}{x_{j+1}-x_j}$

Damit ist auf diesem Intervall

$S(x) = \frac{1}{6} \frac{(x_{j+1}-x)^3}{x_{j+1}-x_j} S''(x_j) + \frac{1}{6} \frac{(x-x_j)^3}{x_{j+1}-x_j} S''(x_{j+1}) + c_j(x - x_j) + d_j(x_{j+1} - x)$

mit gewissen Integrationskonstanten c_j, d_j.

Die Interpolationsforderung
$S(x_j) = \frac{1}{6}S''(x_j)(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j) = f_j$
$S(x_{j+1}) = \frac{1}{6}S''(x_{j+1})(x_{j+1} - x_j)^2 + c_j(x_{j+1} - x_j) = f_{j+1}$
führen auf
$c_j = \frac{f_{j+1}}{x_{j+1}-x_j} - \frac{1}{6}S''(x_{j+1})(x_{j+1} - x_j)$
$d_j = \frac{f_j}{x_{j+1}-x_j} - \frac{1}{6}S''(x_j)(x_{j+1} - x_j)$

Damit ist für $x \in [x_j, x_{j+1}]$
$S'(x) = \frac{1}{2}S''(x_{j+1})\frac{(x-x_j)^2}{x_{j+1}-x_j} - \frac{1}{2}S''(x_j)\frac{(x_{j+1}-x)^2}{x_{j+1}-x_j} + \frac{f_{j+1}-f_j}{x_{j+1}-x_j}$
$+\big(\frac{1}{6}S''(x_j) - \frac{1}{6}S''(x_{j+1})\big)(x_{j+1} - x_j)$
Wertet man S' an den Stellen x_j und x_{j+1} aus, so folgt die Behauptung.

Für die Funktionen selbst erhält man die Darstellung
$S(x) = \frac{1}{6}S''(x_j)\Big(\frac{(x_{j+1}-x)^2}{x_{j+1}-x_j} - (x_{j+1} - x_j)(x_{j+1} - x)\Big)$
$+\frac{1}{6}S''(x_{j+1})\Big(\frac{(x-x_j)^2}{x_{j+1}-x_j} - (x_{j+1} - x_j)(x - x_j)\Big)$
$+f_{j+1}\frac{x-x_j}{x_{j+1}-x_j} + f_j\frac{x_{j+1}-x}{x_{j+1}-x_j}$
für $x \in [x_j, x_{j+1}]$, die man um auslöschungseffekte zu vermeiden besser in der oben angebenenen Form auswertet.
□

Satz 6.26
Erfüllt die zweimal stetig differenzierbare kubische Splinefunktion f die Interpolationsbedingungen $S(x_j) = f_j, j = 0, \dots, N$, so gilt für $j = 1, \dots, N-1$ $\frac{x_{j+1}-x_j}{x_{j+1}-x_{j-1}}S''(x_{j+1}) + 2S''(x_j) + \frac{x_j-x_{j-1}}{x_{j+1}-x_{j-1}}S''(x_{j-1})$
$= \frac{6}{x_{j+1}-x_{j-1}}\Big(\frac{f_{j+1}-f_j}{x_{j+1}-x_j} - \frac{f_j-f_{j-1}}{x_j-x_{j-1}}\Big) = 6f[x_{j-1}, x_j, x_{j+1}]$.(2)

Beweis: Aus Satz 6.25 angewandt auf das Intervall $[x_{j-1}, x_j]$, folgt
$S'(x_j) = \frac{f_j-f_{j-1}}{x_j-x_{j-1}} + \big(\frac{1}{6}S''(x_{j-1}) + \frac{1}{3}S''(x_j)\big)(x_j - x_{j-1})$
und andererseits ($[x_j, x_{j+1}]$)
$S'(x_j) = \frac{f_{j+1}-f_j}{x_{j+1}-x_j} - \big(\frac{1}{3}S''(x_j) + \frac{1}{6}S''(x_{j+1})\big)(x_{j+1} - x_j)$
Daraus folgt
$\big(\frac{1}{6}S''(x_{j-1}) + \frac{1}{3}S''(x_j)\big)(x_j - x_{j-1}) + \big(\frac{1}{3}S''(x_j) + \frac{1}{6}S''(x_{j+1})\big)(x_{j-1} - x_j)$
$= \frac{f_{j+1}-f_j}{x_{j+1}-x_j} - \frac{f_j-f_{j-1}}{x_j-x_{j-1}}$

oder nach Multiplikation mit $\frac{6}{x_{j+1}-x_j}$ die behauptung.
□

Um die Momente $S''(x_j)$,$j = 0, \dots, N$ und damit den Spline S bestimmen zu können, fehlen nocht zwei Gleichungen. Diese Gleichungen lassen sich nur aus Zusatzbedingungen herleiten.

Definition 6.27
Eine zweimal stetig differenzierbare kubische Splinefunktion S, deren zweite Ableitungen in dne Randpunkten $a = x_0$ und $b = x_N$
(3) $S''(x_0) = 0$,$S''(x_N) = 0$ erfüllt, heißt natürliche Splinefunktion.

Satz 6.28
Die Gleichungen (2) aus Satz 6.26 und (3) bestimmen die Momente $S''(x_0), \dots, S''(x_n)$ eindeutig. Dieses Momentengleichungssystem ist sehr gut konditioniert. Bzgl. der Maximumnorm als Vektornorm ist die Norm seiner Koeffizientenmatrix unabhängig von der Intervalleinteilung durch (3) und die Norm seiner Inversen durch 1 beschränkt.

Beweis: Sei $M_0 = b_0, M_N = b_N$ und für $j = 1, \dots, N-1$
$\frac{x_{j+1}-x_j}{x_{j+1}-x_{j-1}} M_{j+1} + 2N_j + \frac{x_j-x_{j-1}}{x_{j+1}-x_{j-1}} M_{j-1} = b_j$
Sei weiter $M = \max\limits_{j=0,\dots,N} |M_j|, B = \max\limits_{j=0,\dots,N} |b_j|$

Dann ist für $j = 1, \dots, N-1$
$2|M_j| = \left| b_j - \frac{x_{j+1}-x_j}{x_{j+1}-x_{j-1}} M_{j+1} - \frac{x_j-x_{j-1}}{x_{j+1}-x_{j-1}} M_{j-1} \right|$
$\leq B + \frac{x_{j+1}-x_j}{x_{j-1}-x_{j-1}} M + \frac{x_j-x_{j-1}}{x_{j+1}-x_{j-1}} M = B + M$
also
$|M_j| \leq \frac{1}{2}B + \frac{1}{2}M$
Da außerdem
$|M_0| = \frac{1}{2}|b_0| + \frac{1}{2}|M_0| \leq \frac{1}{2}B + \frac{1}{2}M$
$|M_N| = \frac{1}{2}|b_N| + \frac{1}{2}|M_N| \leq \frac{1}{2}B + \frac{1}{2}M$
gilt $M \leq \frac{1}{2}B + \frac{1}{2}M$ oder $M \leq B$.

Damit ist das Gleichungssystem eindeutig lösbar und die Norm der inversen der Koeffizientenmatrix durch 1 beschränkt. Wegen

$\frac{x_{j+1}-x_j}{x_{j+1}-x_{j-1}} + 2 + \frac{x_j-x_{j-1}}{x_{j+1}-x_{j-1}} = 2$
ist die Norm der Matrix gleich 3.
□

Satz 6.29

Ist $f : [a, b] \to \mathbb{R}$ eine zweimal stetig differenzierbare Funktion und S ein zweimal stetig differenzierbarer kubischer Spline mit $S(x_j) = f(x_j)$ für $j = 0, \dots, N$, so gilt

$$\int_a^b |f''(x)|^2 \mathrm{dx} = \int_a^b |S''(x)|^2 \mathrm{dx} + \int_a^b |S''(x) - f''(x)|^2 \mathrm{dx} - 2(S'(x) - f'(x))S''(x)|_{x_0}^{x_N}$$

Beweis: Zunächst gilt:

$$\int_a^b |S''(x)|^2 \mathrm{dx} + \int_a^b |S''(x) - f''(x)|^2 \mathrm{dx} = 2\int_a^b (S''(x) - f''(x))S''(x)\mathrm{dx} + \int_a^b |f''(x)|^2 \mathrm{dx}$$

Weiter gilt:

$$\int_a^b (S''(x) - f''(x))\, S''(x)\mathrm{dx} = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} (S''(x) - f''(x))S''(x)\mathrm{dx}$$

$$= \sum_{j=0}^{N-1} [(S'(x) - f'(x))S''(x)|_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} \big(S'(x) - f'(x)\big)S'''(x)\mathrm{dx}]$$

$$= (S'(x) - f'(x))S''(x)|_{x_0}^{x_N} - \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} \big(S'(x) - f'(x)\big)S'''(x)\mathrm{dx}$$

Da $S'''(x)$ auf den Teilintervallen $x_j \le x \le x_{j+1}$ konstant ist, gilt

$$\sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} \big(S'(x) - f'(x)\big)S'''(x)\mathrm{dx} = \sum_{j=0}^{N-1} S'''\big(\tfrac{x_j+x_{j+1}}{2}\big) \int_{x_j}^{x_{j+1}} (S'(x) - f'(x))\mathrm{dx}$$

$$= \sum_{j=0}^{N-1} S'''\big(\tfrac{x_j+x_{j+1}}{2}\big)[S(x) - f(x)]_{x_j}^{x_{j+1}}$$

$$= 0$$

Einsetzen ergibt die Behauptung.
□

Folgerung:

Für die Interpolierende natürliche kubische Splinefunktion S gilt:

$$\int_a^b |f''(x)|^2 \mathrm{dx} = \int_a^b |S''(x)|^2 \mathrm{dx} + \int_a^b |f''(x) - S''(x)|^2 \mathrm{dx} \quad (*)$$

und damit für $f \neq S$

$$\int_a^b |f''(x)|^2 \mathrm{dx} > \int_a^b |S''(x)|^2 \mathrm{dx}$$

Damit ist S die "glatteste" Funktion mit den Werten $S(x_j) = f_j$.

Bemerkung:
Der Satz 6.29 legt es nahe, statt der Zusatzbedingung $S''(x_0) = 0$, $S''(x_N) = 0$. Die Bedingungen $S'(x_0) = f'(x_0)$,$S'(x_N) = f'(x_N)$ zu betrachten. Dabei bleiben die beschriebenen Optimalitätseigenschaften erhalten.

Satz 6.30
Die Zusatzbedingungen $S'(x_0) = f'(x_0), S'(x_N) = f'(x_N)$ sind äquivalent zu $2S''(x_0) + S''(x_1) = \frac{6}{x_1-x_0}\left(\frac{f_1-f_0}{x_1-x_0} - f_0'\right)$,
$S''(x_{N-1}) + 2S''(x_N) = \frac{6}{x_N-x_{N-1}}\left(f_N' - \frac{f_N-f_{N-1}}{x_N-x_{N-1}}\right)$,
Das Momentengleichungssystem ist auch mit diesen Zusatzbedingungen eindeutig lösbar und die Norm seiner Koeffizientenmatrix durch 3 mal die Norm von deren Inversen durch 1 beschränkt.

Satz 6.31
Für jeden zweimal stetig differenzierbaren kubischen Spline S, der die zweimal stetig differenzierbare Funktion f in dne Punkten $x_0 < ... < x_N$ interpoliert, gilt auf dem Intervall $[x_j, x_{j+1}]$
$|S(x) - f(x)| \leq (x_{j+1} - x_j) \int_{x_j}^{x_{j+1}} |S''(t) - f''(t)|\mathrm{dt}$
$|S'(x) - f'(x)| \leq \int_{x_j}^{x_{j+1}} |S''(t) - f''(t)|\mathrm{dt}$
Falls $|S''(t) - f''(t)| \leq k$ ist, folgt:
$|S(x) - f(x)| \leq k(x_{j+1} - x_j)^2$,
$|S'(x) - f'(x)| \leq k(x_{j+1} - x_j)$

Beweis: Wegen $S(x_j) - f(x_j) = S(x_{j+1}) - f(x_{j+1}) = 0$ gibt es nach dem Satz von Rolle ein $\eta \in (x_j, x_{j+1})$ mit $S'(\eta) - f'(\eta) = 0$. Für $x_j \leq x \leq x_{j+1}$ gilt daher $|S'(x) - f'(x)| = |\int_{\eta}^{x}(S''(t) - f''(t))\mathrm{dt}| \leq \int_{x_j}^{x_{j+1}} |S''(t) - f''(t)|\mathrm{dt}$
Damit ist wegen $S(x_j) - f(x_j) = 0$
$|S(x) - f(x)| = |\int_{x_j}^{x}(S'(t) - f'(t))\mathrm{dt}| \leq (x_{j+1} - x_j) \int_{x_j}^{x_{j+1}} |S''(t) - f''(t)|\mathrm{dx}$
□

Satz 6.32

Ist h die maximale Teilintervalllänge, so gilt für den zweimal stetig differenzierbaren kubischen Spline S, der die viermal stetig differenzierbare Funktion f in den Stützstellen x_j interpoliert und den Zusatzbedingungen $S'(x_0) = f'(x_0), S'(x_N) = f'(x_N)$ genügt, die Abschätzung
$\|S'' - f''\|_\infty \leq \frac{3}{8}h^2 \|f^{(4)}\|_\infty$

Beweisidee: Man betrachtet das lineare Gleichungssystem
$2D_0 + D_1 = B_0$
$\frac{x_{j+1}-x_j}{x_{j+1}-x_{j-1}} D_{j+1} + 2D_j + \frac{x_j - x_{j-1}}{x_{j+1}-x_{j-1}} D_{j-1} = B_j, j = 1, \dots, N-1$
$D_{N-1} + 2D_N = B_N$
für die Differenzen $D_j := S''(x_j) - f''(x_j)$ mit entsprechenden rechten Seiten B_j.
Eine sorgfältige Analyse mit Hilfe des Taylorschen Satzes mit Integralrestglied zeigt:
$\max_{j=0,\dots,N} |B_j| \leq \frac{1}{4}h^2 \|f^{(n)}\|_\infty$
Nach Satz 6.28 ist die Norm der Inversen der Koeffizientenmatrix des Gleichungssystems durch 1 beschränkt. Daher ist
$\max_{j=0,\dots,N} |D_j| \leq \max_{j=0,\dots,N} |B_j|$
Zusammengefasst ergibt sich
$\max_{j=0,\dots,N} |S''(x_j) - f''(x_j)| \leq \frac{1}{4}h^2 \|f^{(4)}\|_\infty \qquad (*)$
Wir betrachten nun die stückweise lineare Interpolation
$(If'')(x) = \frac{x-x_j}{x_{j+1}-x_j} f''(x_{j+1}) + \frac{x_{j+1}-x}{x_{j+1}-x_j} f''(x_j),\ x_j \leq x \leq x_{j+1}$
der zweiten Ableitung von f. Aus $(*)$ folgt dann $\|S'' - If''\|_\infty = \max_{j=0,\dots,N} |S''(x_j) - f''(x_j)| \leq \frac{1}{4}h^2 \|f^{(4)}\|_\infty$
Andererseits ist nach Satz 6.6
$\|f'' - If''\|_\infty \leq \frac{1}{8}h^2 \|f^{(4)}\|$
und damit schließlich
$\|S'' - f''\| \leq \left(\frac{1}{4}h^2 + \frac{1}{8}h^2\right) \|f^{(4)}\|_\infty = \frac{3}{8}h^2 \|f^{(4)}\|_\infty$

Folgerung:
Für viermal stetig differenzierbares f gilt
$\|S - f\|_\infty = \mathcal{O}(h^4)$
$\|S' - f'\|_\infty = \mathcal{O}(h^3)$
$\|S'' - f''\|_\infty = \mathcal{O}(h^2)$

Satz 6.33

Für den natürlichen interpolierenden Spline (mit den Zusatzbedingungen
$S''(x_0) = S''(x_N) = 0$) gilt für $j = 0, \dots, N-1$
$|S''(x_j) - f''(x_j)| \leq \frac{1}{4}h^2 \left\| f^{(4)} \right\|_\infty + \left(\frac{1}{2}\right)^j |f''(x_0)| + \left(\frac{1}{2}\right)^{N-j} |f''(x_N)|$

Beweisidee: dann definiere die Splinefunktion S_L und S_R durch die Bedingungen $S_L^{(2)}(x_0) = 1, S_L^{(2)}(x_N) = 0, S_L(x_j) = 0, j = 0, \dots, N$
$S_R^{(2)}(x_0) = 0, S_R^{(2)}(x_N) = 1, S_R(x_j) = 0, j = 0, \dots, N$ und setze weiter
$\widetilde{S}(x) = S(x) + f''(x_0)S_L(x) + f''(x_N)S_R(x)$
Für den interpolierenden Spline $\widetilde{S}$ gilt wegen $S''(x_0) = f''(x_0)$,
$S''(x_N) = f''(x_N)$ analog zum Beweis von Satz 6.32
$\max\limits_{j=0,\dots,N} |\widetilde{S}''(x_j) - f''(x_j)| \leq \frac{1}{4}h^2 \left\| f^{(4)} \right\|_\infty$
Daraus folgt für $j = 0, \dots, N$
$|S''(x_j) - f''(x_j)| \leq \frac{1}{4}h^2 \left\| f^{(4)} \right\|_\infty + + |f''(x_0)| \left| S_L^{(2)}(x_j) \right| + |f''(x_N)| \left| S_R^{(2)}(x_j) \right|$
Wie man durch Analyse des Momentengleichungssystems zeigen kann, ist
$\left| S_L^{(2)}(x_j) \right| \leq \left(\frac{1}{2}\right)^j, \left| S_R^{(2)}(x_j) \right| \leq \left(\frac{1}{2}\right)^{N-j}$
□

7. Numerische Integration

Problem: näherungsweise Berechnung bestimmter Integrale.

Beispiel 7.1

$\frac{8}{\pi \ln 2} \int_0^1 \frac{\ln(1+x)}{1+x^2} \mathrm{dx} = 1$

7.1 Zusammengesetzte Quadraturformeln

Zu berechnen sei $\int_a^b f(x)\mathrm{dx}$. Gegeben sei eine Zerlegung

$a = x_0 < x_1 < ... < x_n = b$. Dann gilt mit $g_i(t) = f\left(\frac{x_i + x_{i+1}}{2} + t\frac{x_{i+1} - x_i}{2}\right)$

$$\int_a^b f(x)\mathrm{dx} = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\mathrm{dx} = \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{2} \int_{-1}^{1} f\left(\frac{x_i + x_{i+1}}{2} + t\frac{x_{i+1} - x_i}{2}\right)\mathrm{dt}$$

$$= \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{2} \int_{-1}^{1} g_i(t)\mathrm{dt}$$

Idee: Ersetze die Integrale $\int_{-1}^{1} g_i(t)\mathrm{dt}$ durch finite Ausdrücke

$$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{n} a_k g(t_k)$$

wobei die a_k von g unabhängige Gewichte und die $t_k \in [-1, 1]$ von g unabhängige Knoten sind.

Endergebnis:

$$\int_a^b f(x)\mathrm{dx} \to \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{2} \sum_{k=0}^{n} a_k f\left(\frac{x_i + x_{i+1}}{2} + t\frac{x_{i+1} - x_i}{2}\right)$$

Beispiel Rechteckregel

$$\int_{-1}^{1} g(t)\mathrm{dt} \to 2g(-1)$$

$$\int_a^b f(x)\mathrm{dx} \to \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$$

Mittelpunktsregel

$\int_{-1}^{1} g(t)\mathrm{dt} \to 2g(0)$

$\int_{a}^{b} f(x)\mathrm{dx} \to \sum_{i=0}^{n-1}(x_{i+1}-x_i)f(\frac{x_i+x_{i+1}}{2})$

Trapezregel

$\int_{-1}^{1} g(t)\mathrm{dt} \to g(-1)+g(1)$

$\int_{a}^{b} f(x)\mathrm{dx} \to \sum_{i=0}^{n-1}(x_{i+1}-x_i)\frac{f(x_i)+f(x_{i+1})}{2}$

Simpsonregel

$\int_{-1}^{1} g(t)\mathrm{dt} \to \frac{1}{3}g(-1)+\frac{4}{3}g(0)+\frac{1}{3}g(1)$

$\int_{a}^{b} f(x)\mathrm{dx} \to \sum_{i=0}^{n-1}\frac{(x_{i+1}-x_i)}{6}\big(f(x_i)+4f\big(\frac{x_i+x_{i+1}}{2}\big)+f(x_{i+1})\big)$

Beispiel 7.2

Integral aus Beispiel 7.1

$x_i = \frac{i}{n}, i = 0, \dots, n$

Konvergenz für $n \to \infty$?

Beobachtung: Fehler Simpsonregel $\mathcal{O}\big(\frac{1}{n^4}\big)$?

$\ll$Fehler Mittelpunktsregel,Trapezregel $\mathcal{O}\big(\frac{1}{n^2}\big)$?

$\ll$Fehler Rechtecksregel $\mathcal{O}\big(\frac{1}{n}\big)$?

Satz 7.3

Die auf dem Referenzintervall vorgegebene Formel $\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{n} a_k g(t_k)$

sei exakt für alle Polynome vom Grad $q-1 \geq 0$. Die Funktion $f : [a,b] \to \mathbb{R}$

sei q-mal stetig differenzierbar und $a = x_0 < \dots < x_n = b$ sei eine beliebige

Zerlegung des Intervalls $[a,b]$. Dann ist $|\int_{a}^{b} f(x)\mathrm{dx} - \sum_{i=0}^{n-1}\frac{x_{i+1}-x_i}{2}\sum_{k=0}^{m} a_k f\big(\frac{x_i+x_{i+1}}{2}+t_k\frac{x_{i+1}-x_i}{2}\big)|$

$\leq C\sum_{k=0}^{n-1}(x_{i+1}-x_i)^q \int_{x_i}^{x_{i+1}} \big|f^{(q)}(x)\big|\mathrm{dx}$

mit einer von der Wahl der Formel für das Referenzintervall abhängigen

Konstante C.

Hilfssatz: Es gilt

$\int_{-1}^{1} g(t)\mathrm{dt} - \sum_{k=0}^{m} a_k g(t_k = \int_{-1}^{1} K(s) g^{(q)}(s)\mathrm{ds}$

$K(s) = \frac{1}{q!}(1-s)^q - \frac{1}{(q-1)!}\sum_{k=0}^{m} a_k (t_k - s)_t^{q-1}$

mit der abgeschnittenen Potenz

$$(t-s)_t^{q-1} = \begin{cases} (t-s)^{q-1}, \text{ falls } s \le t \\ \\ 0\,, \text{ sonst} \end{cases}$$

Beweis: Nach dem Taylorschen Satz ist $g(t) = \sum_{j=0}^{q-1} \frac{1}{j!} g^{(j)}(-1)(t+1)^j + \frac{1}{(q-1)!}\int_{-1}^{t} (t-s)^{q-1} g^{(q)}(s)\mathrm{ds}$

(Beweis durch Induktion über q mit partieller Integration)

also $g(t) = \sum_{j=0}^{q-1} \frac{1}{j!} g^{(j)}(-1)(t+1)^j + \frac{1}{(q-1)!}\int_{-1}^{+1} (t-s)_t^{q-1} g^{(q)}(s)\mathrm{ds}$

Da der polynomiale Anteil nach Voraussetzung exakt integriert wird folgt:

$\int_{-1}^{1} g(t)\mathrm{dt} - \sum_{k=0}^{m} a_k g(t_k)$

$= \int_{-1}^{1} \left(\frac{1}{(q-1)!}\int_{-1}^{1} (t-s)_t^{q-1} g^{(q)}(s)\mathrm{ds} \right) \mathrm{dt} - \sum_{k=0}^{m} a_k \left(\frac{1}{(q-1)!}\int_{-1}^{1} (t_k - s)_t^{q-1} g^{(q)}(s)\mathrm{ds} \right)$

$= \int_{-1}^{1} \left(\frac{1}{(q-1)!}\int_{-1}^{1} (t-s)_t^{q-1}\mathrm{dt} \right) g^{(q)}(s)\mathrm{ds} - \int_{-1}^{1} \left(\sum_{k=0}^{m} a_k \frac{1}{(q-1)!} (t_k - s)_t^{q-1} \right) g^{(q)}(s)\mathrm{ds}$

$= \int_{-1}^{1} K(s) g^{(q)}(s)\mathrm{ds}$

mit dem Kern

$K(s) = \frac{1}{(q-1)!}\left(\int_{-1}^{1} (t-s)_t^{q-1}\mathrm{dt} - \sum_{k=0}^{m} a_k (t_k - s)_t^{q-1} \right)$

$= \frac{1}{(q-1)!}\left(\frac{1}{q}(1-s)^q - \sum_{k=0}^{m} a_k (t_k - s)_t^{q-1} \right)$

□

Beweis von Satz 7.3:

Setzt man $g_i(t) = f\big(\frac{x_i + x_{i+1}}{2} + t\frac{x_{i+1} - x_i}{2}\big)$, gilt nach dem Hilfssatz

$\int_a^b f(x)\mathrm{dx} - \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{2} \sum_{k=0}^{m} a_k f\big(\frac{x_i + x_{i+1}}{2} + t_k \frac{x_{i+1} - x_i}{2}\big)$

$= \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{2} \left(\int_{-1}^{1} g_i(t)\mathrm{dt} - \sum_{k=0}^{m} a_k g_i(t_k) \right)$

$= \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{2} \int_{-1}^{1} K(s) g_i^{(q)}(s)\mathrm{ds}$

mit $C := \frac{1}{2^{q-1}} \max_{-1\le s\le 1}|K(s)|$ ist $|\int_{-1}^{1} K(s) g_i^{(q)}(s)\mathrm{ds}| \le 2^{q-1} C \int_{-1}^{1} \left|g_i^{(q)}(s)\right|\mathrm{ds}$

$= 2^{q-1} C \int_{-1}^{1} \left|f^{(q)}\left(\frac{x_i+x_{i+1}}{2} + s\frac{x_{i+1}-x_i}{2}\right)\right| \left(\frac{x_{i+1}-x_i}{2}\right)^q \mathrm{ds}$

$= 2^{q-1} C \left(\frac{x_{i+1}-x_i}{2}\right)^{q-1} \int_{x_i}^{x_{i+1}} \left|f^{(q)}(x)\right|\mathrm{dx}$

Summation über n ergibt die Behauptung.

□

äquidistante Zerlegungen: $x_{i+1} - x_i = h$

$|\int_a^b f(x)\mathrm{dx} - \frac{h}{2}\sum_{i=0}^{n-1}\sum_{k=0}^{m} a_k f\left(\frac{x_i+x_{i+1}}{2} + \frac{h}{2}t_k\right)|$

$\le C \sum_{i=0}^{n-1} h^q \int_{x_i}^{x_{i+1}} \left|f^{(q)}(x)\right|\mathrm{dx}$

$= C h^q \int_a^b \left|f^{(q)}(x)\right|\mathrm{dx}$

exakt für Polynome vom Grad $q-1$

⇒Konvergenzordnung $\mathcal{O}(h^q)$

Bemerkung:

Äquidistante Zerlegungen sind nicht immer sinnvoll!

Ziel: $(x_{i+1} - x_i)^q \int_{x_i}^{x_{i+1}} \left|f^{(q)}(x)\right|\mathrm{dx} \approx \frac{x_{i+1}-x_i}{b-a}\varepsilon$

⇒ | Fehler | $\le \varepsilon$

Konvergenz bei Riemann integrierbaren Funktionen

$\int_a^b f(x)\mathrm{dx} - \sum_{i=0}^{n-1} \frac{x_{i+1}-x_i}{2} \sum_{k=0}^{m} a_k f\left(\frac{x_i+x_{i+1}}{2} + t_k\frac{x_{i+1}-x_i}{2}\right)$

$= \frac{1}{2}\sum_{k=0}^{m} a_k \underbrace{\left(\int_a^b f(x)\mathrm{dx} - \sum_{i=0}^{n-1}(x_{i+1}-x_i) f\left(\frac{x_i+x_{i+1}}{2} + t_k\frac{x_{i+1}-x_i}{2}\right)\right)}_{\text{Riemannsche Zwischensumme}}$

⇒Konvergenz

7.2 Interpolatorische Quadraturformeln

Ziel: Konstruiere Formeln

$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(t_k) \quad (*)$
möglichst hoher polynomialer Exaktheit.

Satz 7.4

Gegeben seien die $m+1$ von einander verschiedenen Knoten
$t_0, \dots, t_n \in [-1,1]$. Dann ist die Formel$(*)$ genau dann für alle Polynome
n-ten Grades exakt, falls $a_k = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq k}}^{m} \frac{t-t_j}{t_k-t_j}\mathrm{dt}$

Beweis: Polynome m-ten Grades haben die Darstellung
$P(t) = \sum_{k=0}^{m} l_k(t)P(t_k), l_k(t) = \prod_{\substack{j=0 \\ j\neq k}}^{m} \frac{t-t_j}{t_k-t_j}$

□

Beispiel 7.5

$\int_{-1}^{1} g(t)\mathrm{dt} \to a_0 g(-1) + a_1 g(1)$

$a_0 = \int_{-1}^{1} \frac{t-1}{-1-1}\mathrm{dt} = 1, a_1 = \int_{-1}^{1} \frac{t-(-1)}{1-(-1)}\mathrm{dt} = 1$

Bemerkung: Die Gewichte a_k lassen sich über das lineare Gleichungssystem
$\sum_{k=0}^{m} t_k^j a_k = \int_{-1}^{1} t^j \mathrm{dt}, j = 0, \dots, m$ bestimmen.

Beispiel 7.6

Ist $t_0 = -1, t_1 = 0, t_2 = 1$ ergibt sich

$$\begin{array}{rrrcl|l} a_0 & +a_1 & +a_2 & = & 2 & g(t) = 1 \\ -a_0 & & +a_2 & = & 0 & g(t) = t \\ a_0 & & +a_2 & = & \frac{2}{3} & g(t) = t^2 \end{array}$$

und damit $a_0 = \frac{1}{3}, a_1 = \frac{4}{3}, a_2 = \frac{1}{3}$, also die Simpsonregel
$\int_{-1}^{1} g(t)\mathrm{dt} \to \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1)$,
die auch für $g(t) = t^3$ und damit für alle Polynome 3. Grades exakt ist.

Beispiel 7.7

$\frac{8}{\pi \ln 2} \int_0^1 \frac{\ln(1+x)}{1+x^2} \mathrm{dx} (= 1)$

$x_i = \frac{i}{n}, i = 0, \dots, n$

Beobachtung: Die Simpsonregel ist ungleich genauer als Mittelpunkt- oder Trapezregel.

Newton-Cotes-Formeln

$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(-1 + k\frac{2}{m}) \quad (*)$

exakt für Polynome vom Grad m.

Satz 7.8

Ist m gerade, so ist die Newton-Cotes-Formel (*) sogar für alle Polynome vom Grad $m+1$ exakt.

Beweis: Aus Symmetriegründen ist mit $m = 2l$

$\sum_{k=0}^{m} a_k g(-1 + k\frac{2}{m}) = a_k g(0) + \sum_{k=1}^{l} a_{l+k} \big(g\big(-\frac{k}{l}\big) + g\big(\frac{k}{l}\big) \big)$

Die Formel ist daher für alle ungeraden Funktionen g exakt.

□

Bemerkung: Für $m \geq 7$ treten negative Gewichte a_k auf, was die Brauchbarkeit dieser Formeln dann stark einschränkt.

Satz 7.9

Die Quadraturformel der Form

$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(t_k)$

kann maximal für Polynome $(2m+1)$-ten Grades exakt sein.

Beweis: Das Polynom $P(t) = \prod_{j=0}^{m} (t - t_j)^2$ vom Grad $2m+2$ kann durch diese Formel nicht mehr exakt integriert werden, denn es gilt:

$\int_{-1}^{1} P(t)\mathrm{dt} > 0 = \sum_{k=0}^{m} a_k P(t_k)$

□

Satz 7.10

Ist die Quadraturformel

$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(t_k)$
für alle Polynome vom Grad $2m+1$ exakt, so sind ihre Gewichte a_k positiv.
Beweis: $a_j = \sum_{k=0}^{m} a_k \prod_{\substack{i=0 \\ i\neq j}}^{m} \left(\frac{t_k - t_i}{t_j - t_i}\right)^2$

7.3 Orthogonalpolynome

Gegebene: Skalarprodukt $\langle f, g\rangle = \int_a^b f(x)g(x)w(x)\mathrm{dx}$ auf dem Raum aller stetigen
Funktionen von $[a, b]$ nach $\mathbb{R}$ und einer festen Gewichtsfunktion $\omega > 0$

Definition 7.11

Die Polynome $P_0, P_1, P_2, ...$ bilden ein System von Orthogonalpolynomen,
zu dem gegebenen Skalarprodukt, wenn P_k ein Polynom k-ten Grades und führendem Koeffizienten $\neq 0$ ist und für $l \neq k$ $\langle P_k, P_l\rangle = 0$ ist.

Beispiel 7.12

Ist $[a, b] = [-1, 1]$ und $w(x) = 1$, so erhält man bis auf Normierung die
Legendre-Polynome:
$P_k(x) = \frac{1}{2^k k!}\left(\frac{d}{\mathrm{dx}}\right)^k \left((1-x^2)^k\right)$

Satz 7.13

Zu dem gegebenen Skalarprodukt gibt es eindeutig bestimmte Orthogonalpolynome P_k mit führendem Koeffizienten 1. Sie genügen
einer Dreitermrekursion $P_k(x) = (x - a_k)P_{k-1}(x) - b_k^2 P_{k-2}(x)$
für $k = 1, 2, 3, ...$ mit $P_{-1}(x) = 0$,$P_0(x) = 1$.

Beweis: Für $k \geq 1$ gilt $xP_{k-1} = \sum_{j=0}^{k} c_j P_j$ mit den Koeffizienten
$c_j = \frac{\langle xP_{k-1}, P_j\rangle}{\langle P_j, P_j\rangle} = \frac{\langle P_{k-1}, P_j\rangle}{\langle P_j, P_j\rangle}$
Da xP_j für $j \leq k-3$ ein Polynom vom Grad $\leq k-2$ und damit $c_j = 0$ ist,
folgt $xP_{k-1} = c_k P_k + c_{k-1}P_{k-1} + c_{k-2}P_{k-2}$
unter den gegebenen Normierungsbedingungen also
$P_k = (x - c_{k-1})P_{k-1} - c_{k-2}P_{k-2}$
Da sich xP_{k-2} und P_{k-1} wegen der Normierungsbedingung nur durch ein
Polynom vom Grad $k-2$ unterscheiden, ist

$c_{k-2} = \frac{\langle P_{k-1}, xP_{k-2}\rangle}{\langle P_{k-2}, P_{k-2}\rangle} = \frac{\langle P_{k-1}, P_{k-1}\rangle}{\langle P_{k-2}, P_{k-2}\rangle} > 0$
□

Beispiel 7.14
Die Legendre-Polynome genügen der Rekursion $P_0(x) = 1, P_1(x) = x$ und $(k+1)P_{k+1}(x) = (2k+1)xP_k(x) - kP_{k-1}(x)$.

Satz 7.15
Das Orthogonalpolynom P_k aus Definition 7.11 besitzt k einfache Nullstellen im Innern von $[a, b]$.

Beweis: Wegen $w(x) > 0$ für $a < x < b$ sind $\int_a^b P_k(x)w(x)\mathrm{dx} = 0$ muss P_k für $k \geq 1$ mindestens eine Nullstelle mit Vorzeichenwechsel in (a, b) besitzen. Seien nun $x_1, \ldots, x_m$ die Nullstellen von P_k im Innern von $[a, b]$ in aufsteigender Reihenfolge. Sei
$Q(x) = \prod_{i=1}^{m}(x - x_i)$
Dann ist $P_k(x)Q(x)$ Vorzeichenkonstant und daher
$\int_a^b P_k(x)Q(x)w(x)\mathrm{dx} \neq 0$
Dies ist aber nach Definition von P_k nur dann möglich, wenn $m \geq k$ ist, d.h. $m = k$ ist.
□

interpolatorische Quadraturformel
$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(t_k), a_k = \int_{-1}^{1} \prod_{\substack{j=0 \\ j\neq k}}^{m} \frac{t-t_j}{t_k-t_j}\mathrm{dt}$
mindestens exakt für Polynome vom Grad m und maximal für Polynome vom Grad $2m + 1$.

Frage: Lässt sich der Grad 2m+1 erreichen?

Legendre-polynome $P_k, k = 0, 1, 2, \ldots$ von Grad k
$\int_{-1}^{1} P_k(x)P_l(x)\mathrm{dx} = \delta_{kl}, k, l = 0, 1, 2, \ldots$ haben k Nullstellen im Innern des Intervalls $[-1, 1]$.

7.4 Gaußsche Quadraturformeln

Satz 7.16

Es gibt genau eine Quadraturformel $\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(t_k)$ die für alle Polynome vom Grad $2m+1$ exakt ist. Die Knoten t_k dieser Quadraturformel sind die Nullstellen des Legendre-Polynoms $(m+1)$-ten Grades und liegen im Innern des Intervalls $[-1,1]$. Die Gewichte a_k sind positiv.

Beweis: Jedes Polynom P vom Grad $2m+1$ lässt sich eindeutig in der Form

$$P(t) = Q(t)\prod_{j=0}^{m}(t-t_j) + R(t)$$

mit Polynomen Q und R vom Grad m darstellen. Es gilt dann:

$$\int_{-1}^{1} P(t)\mathrm{dt} = \int_{-1}^{1} Q(t)\prod_{j=0}^{m}(t-t_0)\mathrm{dt} + \int_{-1}^{1} R(t)\mathrm{dt}$$

$$\sum_{k=0}^{m} a_k P(t_k) = \sum_{k=0}^{m} a_k R(t_k)$$

Damit ist die gegenen Quadraturformel genau dann für alle Polynome vom Grad $2m+1$ exakt, wenn sie
1) für alle Polynome m-ten Grades exakt ist, und wenn
2) für alle Polynome Q vom Grad m
$\int_{-1}^{1} Q(t)\prod_{j=0}^{m}(t-t_j)\mathrm{dt} = 0$ ist.

Die Bedingung 2) ist aber genau dann erfüllt, wenn die t_k die Nullstellen des Legendre Polynoms vom Grad $m+1$ sind. Die Positivität der a_k wurde bereits im Satz 7.10 bewiesen.
□

merke: $m+1$ Stützstellen
=exakt für Polynome vom Grad $2m+1$
=Fehlerordnung $O(h^{2m+2})$

Beispiel 7.17

Die Mittelpunktregel

$\int_{-1}^{1} g(t)\mathrm{dt} \to 2g(0)$
ist die Gaußsche Quadraturformel für den Fall $m = 1$.

Beispiel 7.18

$\frac{8}{\pi \ln 2}\int_{0}^{1} \frac{\ln(1+x)}{1+x^2}\mathrm{dx} = 1$
Mit 2 Teilintervallen liefert die 4-Punkt Gaußformel der Fehlerordnung 10 bereits den auf 10 Stellen genauen Wert!

Satz 7.19

Für die Gaußsche Quadraturformel
$\int_{-1}^{1} g(t)\mathrm{dt} \to \sum_{k=0}^{m} a_k g(t_k)$
vom Grad $2m + 1$ gilt:
$|\int_{a}^{b} f(x)\mathrm{dx} - \frac{b-a}{2}\sum_{k=0}^{m} a_k f\big(\frac{a+b}{2} + t_k\frac{b-a}{2}\big)| \leq 2(b-a)E_{2m+1}(f)$
Dabei ist $E_{2m+1}(f)$ der Fehler der besten Approximation der Funktion $f : [a, b] \to \mathbb{R}$ durhc ein Polynom $(2m + 1)$-ten Grades.

Beweis: Für alle Polynome vom Grad $2m + 1$ gilt:
$|\int_{a}^{b} f(x)\mathrm{dx} - \frac{b-a}{2}\sum_{k=0}^{m} a_k f\big(\frac{a+b}{2} + t_k\frac{b-a}{2}\big)|$
$= |\int_{a}^{b} (f-P)(x)\mathrm{dx} - \frac{b-a}{2}\sum_{k=0}^{m} a_k (f-P)\big(\frac{a+b}{2} + t_k\frac{b-a}{2}\big)|$
$\leq \int_{a}^{b} |f(x) - P(x)|\mathrm{dx} + \frac{b-a}{2}\sum_{k=0}^{m} |a_k|\big|(f-P)\big(\frac{a+b}{2} + t_k\frac{b-a}{2}\big)\big|$
$\leq \left((b-a) + \frac{b-a}{2}\sum_{k=0}^{m} |a_k|\right) \max_{a\leq x\leq b} |f(x) - P(x)|$
und daher
$|\int_{a}^{b} f(x)\mathrm{dx} - \frac{b-a}{2}\sum_{k=0}^{m} a_k f\big(\frac{a+b}{2} + t_k\frac{b-a}{2}\big)|$
$\leq (1 + \frac{1}{2}\sum_{k=0}^{m} |a_k|)(b-a)E_{2m+1}(f)$
Wegen der Positivität der a_k ist
$1 + \frac{1}{2}\sum_{k=0}^{m} |a_k| = 1 + \frac{1}{2}\sum_{k=0}^{m} a_k = 1 + \frac{1}{2}\int_{-1}^{1} 1\mathrm{dt} = 2$
□

8. Gewöhnliche Diefferentialgleichungen

Anfangwertproblem 1. Ordnung
Gegeben sei eine Funktion $f : [a, b] \times \Omega \to \mathbb{R}^n$, $\Omega \subseteq \mathbb{R}^n$ und ein
Anfangswert $y_0 \in \Omega$. Gesucht ist eine Funktion $y : [a, b] \to \Omega$ mit
$y'(x) = f(x, y(x))$ für $a \leq x \leq b$
$y(a) = y_0$
Existenz und Eindeutigkeit(oder lokale Ex. und Eind.)
->Analysis II
hier: Konstruktion von Näherungsverfahren

Einschrittverfahren

Gesucht: Näherungswerte Y_k für die Lösungswerte $y(x_k)$ auf dem Gitter
$a = x_0 < x_1 < ... < x_N = b$, $(x_{k+1} = x_k + h_k)$

Euler-Verfahren

$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} y'(x)\mathrm{dx} = y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x))\mathrm{dx} \approx y(x_k) + h_k f(x_k, y(x_k))$
$\frac{y(x_{k+1})-y(x_k)}{h_k} \approx y'(x_k) = f(x_k, y(x_k))$
$Y_{k+1} = Y_k + h_k f(x_k, Y_k)$
sehr grobe Approximation!

verbessertes Euler-Verfahren

$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} f(x, y(x))\mathrm{dx} \approx y(x_k) + h_k f(x_k + \frac{1}{2}h_k, y(x_k + \frac{1}{2}h_k))$
$\frac{y(x_{k+1})-y(x_k)}{h_k} \approx y'(x_k + \frac{1}{2}h_k) = f(x_k + \frac{1}{2}h_k, y(x_k + \frac{1}{2}h_k))$
$f(x_k + \frac{1}{2}h_k, y(x_k + \frac{1}{2}h_k)) \approx f(x_k + \frac{1}{2}h_k, y(x_k) + \frac{1}{2}h_k f(x_k, y(x_k)))$

$Y_{k+1} = Y_k + h_k f(x_k + \frac{1}{2}h_k, Y_k + \frac{1}{2}h_k f(x_k, Y_k))$
Einschrittverfahren:
$Y_{k+1} = Y_k + h_k \phi(x_k, Y_k, h_k)$

Probleme:
-In jedem Schritt wird ein neuer Fehler eingeschleppt

-Einmal vorhandene Fehler pflanzen sich fort

Definition

Die Funktion $y(x)$ sei also die Lösung des gegebenen Anfangswertproblems und $p(x, y, h)$ die Verfahrensfunktion des Einschrittverfahrens. Dann heißt
$\tau(x, h) := \frac{1}{h}(y(x + h) - y(x)) - \varphi(x, y(x), h)$
der Abbruchfehler des Verfahrens in der Lösung y. Es ist
$y(x + h) = y(x) + h\varphi(x, y(x), h) + h\tau(x, h)$

Beispiel Euler

$\tau(x, h)|_i = \frac{1}{h}(y(x + h) - y(x)) - f(x, y(x))|_i$
$= \frac{1}{h}(y(x + h) - y(x)) - y'(x)|_i$
$= \frac{1}{2}y_i''(y_i)h$
$\Rightarrow \tau(h, x) = \mathcal{O}(h)$
verbesserter Euler:
$\tau(h, x) = \mathcal{O}(h^2)$

Ziel:
globalen Verfahrensfehler $\max\|Y_k - y(x_k)\|$ durch de Abbruchfehler $\tau(x_k, h_k)$ abschätzen.

Satz

Die Verfahrensfunktion des Einschrittverfahrens genüge der Lipschitzbedingung $\|\varphi(x, y, h) - \varphi(x, z, h)\| \leq L\|y - z\|$ für alle $y, z \in \mathbb{R}^n, h \in [0, H], x \in [a, b - h]$.

Dann gilt die Fehlerabschätzung $\|y(x_k) - Y_k\| \leq e^{L(x_k - a)} \sum_{j=0}^{k-1} h_j \|\tau(x_j, h_j)\|$,

$k = 1, 2, \dots, N$.

Beweis: Für $k = 0, 1, \dots, N - 1$ gilt:
$\|y(x_{k+1}) - Y_{k-1}\| = \|(y(x_k) + h_k\varphi(x_k, y(x_k), h_k) + h_k\tau(x_k, h_k)) - (Y_k - h_k\varphi(x_k, Y_k, h_k))\| \leq$
$\|y(x_k) - Y_k\| + h_k\|\varphi(x_k, y(x_k), h_k) - \varphi(x_k, Y_k, h_k)\| + h_k\|\tau(x_k, h_k)\|$
$\leq \|y(x_k) - Y_k\| + h_k L\|y(x_k) - Y_k\| + h_k\|\tau(x_k, Y_k)\|$
$\Rightarrow \|y(x_{k+1} - Y_{k+1}\| \leq e^{h_k L}\|y(x_k) - Y_k\| + h_k\|\tau(x_k, h_k)\|$ $(*)$ wegen
$1 + t \leq e^t$. Aus $(*)$ folgt durch Induktion über k die Behauptung $k = 0$ ✓

$$k \to k+1: \ \|y(x_{k+1}) - Y_{k+1}\| \leq e^{h_k L} \underbrace{e^{L(x_k - a)} \sum_{j=0}^{k-1} h_j \|\tau(x_j, h_j)\|}_{\substack{= e^{L(x_{k+1}-a)} \\ \leq e^{L(x_{k+1}-a)} \sum_{j=0}^{k} h_j \|\tau(x_k, h_j)\|}} + h_k \|\tau(x_k, h_k)\| \ \square$$

Folgerung: Wegen $\sum_{j=0}^{k-1} h_j \|\tau(x_j, h_j)\| \leq (x_k - a) \max_{j=0,\dots,k-1} \|\tau(x_j, h_j)\|$ gilt unter der Annahme $\|\tau(x,h)\| \leq K h^q$ mit $h = \max h_j$ die Abschätzung

$\max_{k=0,\dots,N} \|y(x_k) - Y_k\| \leq (b-a) e^{L(b-a)} K h^q = C h^q$

Der Fehler strebt also mit der Ordnung $\mathcal{O}(h^q)$ gegen Null.

Konsitenz (=Verhalten des Abbruchfehlers)
+ Stabilität (Verfahrensfehler lässt sich durch Abbruchfehler abschätzen)
Konvergenz (mit der Ordnung des Abbruchfehlers)

Euler $\mathcal{O}(h)$, verb. Euler $\mathcal{O}(h^2)$

klassisches Runge-Kutta-Verfahren

$\varphi(x,y,h) = \frac{1}{6}K_1 + \frac{1}{3}K_2 + \frac{1}{3}K_3 + \frac{1}{6}K_4$
$K_1 = f(x,y)$
$K_2 = f(x + \frac{h}{2}, y + \frac{h}{2}K_1)$
$K_3 = f(x + \frac{h}{2}, y + \frac{h}{2}K_2)$
$K_4 = f(x + h, y + hK_3)$
Fehlerordnung: $\tau(x,h) = \mathcal{O}(h^4)$
geht für $y'(x) = f(x)$ in die Simpsonregel über.