Numerische Mathematik

für

Studierende der Mathematik

und Technomathematik

SS 2004 WS 2004/05

W. Hofmann

Universität Hamburg

Vorbemerkung

Dieses Numerik–Skript will, kann und soll kein Lehrbuch ersetzen. Vielmehr möchte es den Studenten ermöglichen, der Vorlesung zu folgen, ohne den Zwang mitschreiben zu müssen. Es soll sie auch anleiten, den behandelten Stoff in unterschiedlichen Lehrbüchern nachzulesen und zu ergänzen und sie damit in Stand setzen, beim Kauf eines Lehrbuchs eine begründete Wahl zu treffen.

Daß dieses Skript geschrieben wurde, liegt auch am speziellen Hamburger Studienplan, der eine Einführung in die Numerische Mathematik parallel zu den Grundvorlesungen über Analysis, Lineare Algebra und Analytische Geometrie vorsieht. Dies hat, neben einer frühen Einführung in die Numerik, den Vorteil, daß theoretische Ergebnisse aus Analysis, Lineare Algebra und Analytische Geometrie sowohl ergänzt als auch motiviert werden können.

Natürlich ergeben sich aus dieser Situation auch inhaltliche Konsequenzen. Die Auswahl des Stoffes muß, so weit als möglich, danach ausgerichtet werden, was an mathematischen Kenntnissen durch die Schule oder die parallel laufenden Grundvorlesungen schon bereitgestellt worden ist. Deshalb können eine Reihe von Themen, die üblicherweise zu einer Einführung in die Numerische Mathematik gehören, nicht, noch nicht oder nur marginal behandelt werden. Auch die Reihenfolge des dargebotenen Stoffes ist diesen Rahmenbedingungen unterworfen. Themenkomplexe, die inhaltlich zusammengehören, müssen zum Teil zeitlich entzerrt werden, bis die notwendigen Vorkenntnisse bereitgestellt worden sind. Aus diesen Gründen sind Ergänzungen durch die Lehrbuchliteratur unverzichtbar.

Die meisten Numerik–Bücher setzen die Kenntnisse aus den Anfänger–Vorlesungen voraus und sind deshalb als einziges Vorlesungsbegleitmaterial nur bedingt geeignet. Auch dies ist ein Grund für die Erstellung dieses Skriptes.

Diese Schrift geht zurück auf ein Skript, das von Christoph Maas angefertigt wurde und von vielen Kollegen, die seither die Numerik gelesen haben (Werner, Hass, Opfer, Geiger, Hofmann, um nur einige zu nennen), ergänzt, umgearbeitet und aktualisiert worden ist.

Inhaltsverzeichnis

Einige Literatur–Hinweise

Analysis

Königsberger, K.: Analysis 1 und 2, Springer–Lehrbuch, Springer Verlag, 1990

Forster, O.: Analysis, Bände 1–3, Vieweg Studium; Friedr. Vieweg u. Sohn, Braunschweig, Wiesbaden, 1983

Lineare Algebra und Analytische Geometrie

Fischer, G.: Lineare Algebra; Vieweg, Braunschweig 1986

Jänich, K.: Lineare Algebra; Springer Verlag 1981

Numerische Mathematik

Stoer, J.: Numerische Mathematik 1, Springer–Lehrbuch, 5. Aufl. 1989

Stoer, J., Bulirsch, R.: Numerische Mathematik 2, Springer–Lehrbuch, 3. Aufl. 1990

Schwarz, H.R.: Numerische Mathematik, Teubner, Stuttgart 1988

Deuflhard / Hohmann: Numerische Mathematik, 2. Aufl., de Gruyter Lehrbuch 1993

Hämmerlin, G., Hoffmann, K.H.: Numerische Mathematik, 2. Aufl. 1991

Werner, J.: Numerische Mathematik, Bände 1–2, Vieweg Studium 1992

Opfer, G.: Numerische Mathematik für Anfänger, 2. Aufl., Vieweg Studium 1994

Collatz, L., Wetterling, W.: Optimierungsaufgaben, Springer–Verlag 1971

Glashoff, K., Gustafson, A.: Einführung in die Lineare Optimierung, Wissenschaftliche Buchgesellschaft, Darmstadt, 1978

Wilkinson, J.H.: Rundungsfehler, Springer–Verlag 1969

§ 1 Zahlendarstellung und Rundungsfehler

Wir wollen hier nur eine vereinfachte, beispielorientierte und keineswegs vollständige Einführung in die Problematik des Rechnens an Computern geben. Eine ausführlichere, gut lesbare Darstellung dieses Themenkreises, die allerdings gewisse Grundkenntnisse der Analysis voraussetzt (Differenzierbarkeit, Taylorreihe), findet man z.B. in Stoer: Numerische Mathematik I, §1.

In den (von uns benutzten) Digitalrechnern hat man nicht die reellen Zahlen zur Verfügung, sondern nur eine **endliche** Menge A von Zahlen, die sog. Maschinenzahlen. Also gibt es in jedem Intervall zwischen 2 benachbarten Maschinenzahlen unendlich viele Zahlen, die der Computer nicht „kennt". Man muß also überlegen, wie man diese Maschinenzahlen am zweckmäßigsten auswählt und welche Konsequenzen diese Auswahl für die Ergebnisse unserer Rechnungen besitzt.

Rundungsfehler

Unabhängig von der Auswahl der Maschinenzahlen muß eine Zahl $x \notin A$ durch eine gerundete Zahl $rd(x) \in A$ angenähert werden. Vernünftig ist hierbei folgende Optimalitätsforderung

$$|rd(x) - x| \;\leq\; |y - x| \qquad \forall\, y \in A\,.$$

Liegt x genau in der Mitte zwischen 2 Maschinenzahlen, benötigt man eine Zusatzregel: man wählt z.B. die betragsgrößere der beiden möglichen Zahlen.

Wie muß nun eine Fehlergröße festgelegt werden, die Auskunft über die Genauigkeit einer Approximation gibt? Grundsätzlich gibt es zwei Fehlertypen:

1) absoluter Fehler $\quad e_{\text{abs}} \;:=\; rd(x) - x$

2) relativer Fehler $\quad e_{\text{rel}} \;:=\; \dfrac{rd(x) - x}{x}$

Ihre Bedeutung machen wir uns an Beispielen klar.

Die Entfernung der Mittelpunkte von Erde und Mond bis auf einen absoluten Fehler von 5 m zu bestimmen, ist außerordentlich genau. Für die Angabe der Größe einer Parklücke ist eine Fehlermarke von 5 m äußerst ungenügend. Ein absoluter Fehler von ca. 50 cm ist hier angebrachter. Diese Fehlergröße ist wiederum bei der Bestimmung der Wellenlänge des sichtbaren Lichts (etwa $0.4 \cdot 10^{-6}\ m$ bis $0.8 \cdot 10^{-6}\ m$) mehr als wertlos.

Die Tolerierbarkeit eines Fehlers wird also weniger durch seine absolute Größe als durch sein Verhältnis zur Größenordnung des exakten Werts festgelegt (relativer Fehler).

Es hat sich deshalb als sinnvoll erwiesen, die Maschinenzahlen so zu verteilen, daß für jede Zahl x aus den Intervallen $[-M, -m]$ bzw. $[m, M]$ der Zahlen, die man im Rechner darstellen will, (m bzw. M sind die kleinste bzw. größte positive ganze Zahl, die man im Rechner darstellen will) der relative Fehler $\dfrac{rd(x) - x}{x}$ betragsmäßig eine möglichst kleine Schranke nicht übersteigt. Dies führt zur Benutzung der Gleitpunktdarstellung für Zahlen im Rechner, die dieser Bedingung genügt, wie wir noch zeigen werden.

Gleitpunktdarstellung

Eine Zahl wird dargestellt in der Form

$$a \quad = \quad \pm\, a_0 \,.\, a_1\, a_2 \;\ldots\; a_{t-1} \,\cdot\, g^p$$

Hierbei ist: g die Basis ($g = 10$, Dezimalsystem, wird bei der Eingabe von Zahlen in den Rechner und bei der Ausgabe von Ergebnissen benutzt, intern benutzen die Rechner das Dualsystem, $g = 2$ oder das Hexadezimalsystem $g = 16$),

p, der Exponent, ist eine ganze Zahl, die betragsmäßig rechnerabhängig beschränkt wird (z.B. $|p| \le 99$),

$a_0.a_1 \ldots a_{t-1}$, $a_j \in \{0, 1, 2, \ldots, g-1\}$, $j = 0, \ldots, t-1$ ist eine Ziffernfolge der Mantissenlänge t ($\in \mathbb{N}$). t wird rechnerabhängig fixiert. Rechnerabhängig wird gefordert

$$a_0 \neq 0 \qquad\qquad \text{(bei unseren Rechnern)}$$

$$\text{oder} \qquad a_0 = 0 \quad \text{und} \quad a_1 \neq 0\,.$$

Im Rechner wird zur Ein– und Ausgabe von Ergebnissen überlicherweise das Dezimalsystem ($g = 10$) benutzt. Zahlen, die nicht in diese Darstellungsform passen (Zahlen mit größerer Mantissenlänge als t, z.B. Wurzeln und unendliche Dezimalbrüche) werden im allg. betragsmäßig nach folgender Vorschrift gerundet:

Für $|x| = a_0 \,.\, a_1\, a_2 \;\ldots\; a_t\, a_{t+1} \;\ldots\; \cdot\, 10^p$ ist

$$|rd(x)| \quad = \quad \begin{cases} a_0 \,.\, a_1 \;\ldots\; a_{t-1} \cdot 10^p, & \text{falls} \quad a_t < 5 \\ (a_0 \,.\, a_1 \;\ldots\; a_{t-1} + 10^{-(t-1)}) \cdot 10^p, & \text{falls} \quad a_t \ge 5 \end{cases}$$

Das Vorzeichen bleibt ungeändert.

Wir berechnen den maximalen relativen Fehler von $\tilde{rd}(x)$ der Rundung in der Gleitkommadarstellung mit $g = 10$.

Laut Rundungsvorschrift gilt: (falls $|x| \ge m$)

$$\left|\tilde{rd}(x) - x\right| \le 5 \cdot 10^{-t} \cdot 10^p\,,$$

$$|x| \ge a_0 \,.\, a_1\, a_2 \ldots a_{t-1} \cdot 10^p \ge 10^p \quad (\text{beachte } a_0 \neq 0)$$

$$\text{also} \qquad \tfrac{1}{|x|} \le 10^{-p}$$

$$\text{somit} \qquad \underline{\underline{\tfrac{\left|\tilde{rd}(x) - x\right|}{|x|}}} \le 5 \cdot 10^{-t} \cdot 10^p \cdot 10^{-p} = \underline{\underline{\tfrac{1}{2} \cdot 10^{-t+1}}}\,.$$

Auf die gleiche Weise zeigt man für eine beliebige Basis g

$$\frac{|rd(x) - x|}{|x|} \quad \leq \quad \frac{1}{2} \cdot g^{-t+1} \, .$$

Diese Zahl heißt *Maschinengenauigkeit*.

Sie gilt einheitlich für den Gesamtbereich der darstellbaren Zahlen, und hängt bei vorgegebener Basis g nur von der Mantissenlänge t (der Anzahl der verwendeten Ziffern) ab.

Das folgende Beispiel zeigt, daß dies für den absoluten Fehler nicht gilt.

Beispiel: $g = 10, \; t = 10, \; |p| \leq 99.$

$$\left.\begin{array}{l} 1.000000000 \cdot 10^{-99} \\ 1.000000001 \cdot 10^{-99} \end{array}\right\} \quad \text{Differenz} \qquad 1 \cdot 10^{-9} \cdot 10^{-99} \;=\; 10^{-108}$$

$$\left.\begin{array}{l} 9.999999998 \cdot 10^{99} \\ 9.999999999 \cdot 10^{99} \end{array}\right\} \quad \text{Differenz} \qquad 1 \cdot 10^{-9} \cdot 10^{99} \;=\; 10^{90}$$

Die absolute Differenz benachbarter Zahlen, und somit auch der Rundungsfehler, wächst mit dem Betrag der dargestellten Zahl.

Fehler bei der Rechnung

Fehler beim Rechnen mit Computern entstehen aus folgenden Gründen:

1) Die Eingabedaten sind fehlerhaft

   a) weil nur eine begrenzte Anzahl von Ziffern (im Dezimalsystem) für die Eingabe zur Verfügung stehen (1. Rundung),

   b) weil diese Ziffern im Rahmen der numerischen Verarbeitung in das Dualsystem (oder Hexadezimalsystem) übertragen werden (2. Rundung),

2) a) weil sich die Eingangsfehler aus 1) bei der Rechnung fortpflanzen und

   b) weil bei jeder Rechenoperation i. allg. wieder gerundet werden muß, wodurch neue Fehler entstehen, die sich wiederum fortpflanzen,

3) durch mangelhafte Programmierung, die den Gegebenheiten aus 1) und 2) nicht Rechnung trägt.

Wir illustrieren dies durch einige Beispiele.

a) Für die Subtraktion

$$5.000000000001 - 5$$

liefert Matlab Version 6.5 im long-Format das Ergebnis

$$1.000089...e - 15$$

statt

$$1.00000000000000\, e - 15\,.$$

Dies entspricht einem absoluten Fehler von $8.9 \cdot 10^{-20}$ und einem relativen Fehler von $8.9 \cdot 10^{-5}$.

Erklärung: Ganze Zahlen x mit $|x| < g^t$ sind üblicherweise Maschinenzahlen, also ist 5 eine Maschinenzahl, nicht aber 5.000000000001. Hier tritt zunächst ein Rundungsfehler auf beim Konvertieren ins Dualsystem.

Er allein kann aber nicht die Größe des relativen Fehlers erklären, der um mehrere Zehnerpotenzen größer ist als der der Ausgangsdaten. Er beruht auf dem Phänomen der sog. „Auslöschung", die bei der Subtraktion annähernd gleich großer Größen auftritt (fehlerverstärkende Operation). Wir gehen darauf gleich noch ein.

b) Für jedes $x \neq 0$ gilt $\left(\left(\frac{1}{x}\right)/10 + 1\right) * x - x \;=\; 0.1\,.$

Eine Abfrage auf Gleichheit, wie im folgenden Programmausschnitt, (in Matlab-Formulierung), trifft fast immer Fehlentscheidungen.

```
for   x = 1 : 10
      y = ((1/x)/10 + 1) * x - x ;
      if  y == 0.1
         disp('gut gegangen')
      else
        disp('schief gegangen')
      end
end
```

Das Programm wird fast immer „schief gegangen" ausdrucken, da auf Grund von Rundungsfehlern i. allg. $y \neq 0.1$ ist.

c) Bezeichnet maxint die größte ganze Zahl, die der Rechner darstellen kann, so wird (sollte) das Programm

```
n = maxint - 10;
for  i = 1 : 20
   disp(n)
   n = n + 1;
end
```

wegen „Zahlbereichsüberlauf" aussteigen.

Es gibt unzählige weitere Fehler, die man machen kann. Wie kann man die schlimmsten verhüten?

Hilfe gegen rundungsbedingte Rechenfehler
ist nur begrenzt möglich, aber nicht zu vernachlässigen. Man kann zunächst

eine Fehleranalyse der Grundrechenarten
durchführen (um zu wissen, welche Operationen wie gefährlich sind), die angibt wie sich Eingabefehler in jedem Schritt auswirken.

Es bezeichne ε_x den relativen Fehler bei der Darstellung der Zahl x im Gleitpunktsystem der Rechenanlage

$$\varepsilon_x = \frac{rd(x) - x}{x} \qquad \text{oder} \qquad rd(x) = x\,(1 + \varepsilon_x) \ .$$

Addition:
Wir berechnen den relativen Fehler ε_{x+y} der Summe $x + y$.

$$\varepsilon_{x+y} \approx \frac{rd(x) + rd(y) - (x + y)}{(x + y)}$$

Man kann nur „$\approx$“ schreiben, denn $rd(x)+rd(y)$ muß keine Maschinenzahl sein. Durch Ausrechnen folgt

$$\begin{aligned} \varepsilon_{x+y} &\approx \frac{rd(x) - x}{x + y} + \frac{rd(y) - y}{x + y} \\ &= \frac{x}{x + y}\,\varepsilon_x + \frac{y}{x + y}\,\varepsilon_y \ . \end{aligned}$$

FAZIT: Bei der Addition von Zahlen gleichen Vorzeichens addiert sich der relative Fehler der Eingangsdaten höchstens, da $\frac{x}{x+y},\ \frac{y}{x+y} \le 1$.

Subtraktion:
Man ersetze in obiger Rechnung y durch $-y$ (also $x > 0,\ -y > 0$), dann erhält man

$$\varepsilon_{x-y} \approx \frac{x}{x - y}\,\varepsilon_x - \frac{y}{x - y}\,\varepsilon_y \ .$$

Hier kann ein Unglück passieren, wenn x und y fast gleich groß sind (vgl. das Beispiel a)). Ist zum Beispiel y eine Maschinenzahl, also $\varepsilon_y = 0,\ x - y \approx 10^{-12},\ x \approx 5$, so folgt

$$\varepsilon_{x-y} \approx 5 \cdot 10^{12}\,\varepsilon_x \ .$$

Dieses Phänomen bezeichnet man als „*Auslöschung*“ (richtiger Ziffern).

Multiplikation:

$$\begin{aligned}
\varepsilon_{x\cdot y} &\approx \frac{rd(x)\,rd(y) - x\cdot y}{x\cdot y} &&= \frac{x(1+\varepsilon_x)\cdot y(1+\varepsilon_y) - xy}{xy} \\
&= \frac{xy\,\varepsilon_x + xy\,\varepsilon_y + xy\,\varepsilon_x\,\varepsilon_y}{xy} &&= \varepsilon_x + \varepsilon_y + \varepsilon_x\,\varepsilon_y \;\approx\; \varepsilon_x + \varepsilon_y
\end{aligned}$$

also $\qquad \varepsilon_{xy} \;\approx\; \varepsilon_x + \varepsilon_y \,,\quad$ also „relativ ungefährlich“ .

Division: Analog zu oben erhält man

$$\varepsilon_{x/y} \;=\; \varepsilon_x - \varepsilon_y \,.$$

Gefährlich ist also vor allem die Subtraktion annähernd gleich großer Zahlen. Wir werden in den Übungen Beispiele für ihre Auswirkung und ihre Vermeidung kennenlernen.

Natürlich sollte man Fehleranalysen wie oben für ganze mathematische Verfahren durchführen. Dies ist sehr aufwendig und im Rahmen dieser Veranstaltung nicht möglich (vgl. dazu etwa: Wilkinson: Rundungsfehler).

Wir müssen uns hier mit Hinweisen bzgl. der einzelnen zu behandelnden Verfahren zufrieden geben.

Allgemein kann man nur empfehlen:

- Vermeide Rechenoperationen, die Fehler verstärken (z.B. Auslöschung).
- Vermeide Verfahren, die eine zu genaue Angabe von Eingabedaten oder Zwischenergebnissen (vgl. etwa Beispiel b)) verlangen.
- Vermeide Eingabewerte, für die das Problem sich kaum von einem unlösbaren Problem unterscheidet, oder die in der Nähe von Werten liegen, für das sich das Verfahren unvorhersehbar verhält.
- Vermeide rundungsfehleranfällige Verfahren (Beispiele werden wir kennenlernen).

§ 2 Polynome

Definition 2.1
Seien $a_0, a_1, a_2, \ldots, a_n$ reelle oder komplexe Zahlen ($a_i \in \mathbb{R}$, $i = 0, \ldots, n$ oder $a_i \in \mathbb{C}$, $i = 0, \ldots, n$). Dann versteht man unter einem *Polynom* p

$$p(x) \;=\; a_0 + a_1\,x + a_2\,x^2 + \;\ldots\; + a_n\,x^n \;=\; \sum_{i=0}^{n} a_i\,x^i, \qquad x^0 \;:=\; 1\,,$$

eine Abbildung

$$\begin{aligned} p \;:\; \mathbb{C} &\longrightarrow \mathbb{C} \\ x &\longmapsto p(x)\,. \end{aligned}$$

Ist $a_n \neq 0$ (Koeffizient der höchsten auftretenden x–Potenz), so heißt das Polynom p vom *Grad* n.
Die Menge der Polynome vom Höchstgrad n (n eine natürliche Zahl: $n \in \mathbb{N}$) bezeichnen wir mit

$$\Pi_n \quad = \quad \left\{ p \;:\; p(x) \;=\; \sum_{i=0}^{n} a_i\,x^i, \quad a_i \in \mathbb{C} \right\}\;.$$

Mathematisch von Interesse sind die beiden Probleme: Sei $p \in \Pi_n$ gegeben:

a) berechne $p(\hat{x})$ für ein vorgegebenes $\hat{x}$ (Polynomwertberechnung),

b) bestimme ein $\tilde{x}$ derart, daß $p(\tilde{x}) = 0$ (Nullstellenproblem).

Für beide Probleme führen wir ein Bespiel an, welches das Vorkommen von Polynomen demonstriert.

Beispiel (Zinseszinsrechnung)
Ein Unternehmen plant die Anschaffung einer Maschine zum Preis von 20.000,00 Euro. Bis zu ihrer Verschrottung nach 5 Betriebsjahren werden von dieser Investition die aus der folgenden Tabelle ersichtlichen Beiträge zum Gewinn erwartet.

Sei
a_5 = Kaufpreis der Maschine,
n = Anzahl der Jahre bis zur Verschrottung der Maschine,
a_n = Gewinn nach $5 - n$ Jahren,
a_0 = Gewinn im 5. Jahr einschließlich Verschrottungskosten oder Verschrottungserlös der Maschine.

n	4	3	2	1	0
a_n	2200	6000	6900	6900	4300

a) Eine Bank bietet für den Kauf der Maschine einen Kredit zu 10 % Jahreszins an, und sowohl Zinsen als auch Tilgung sollen von dem mit dieser Maschine erwirtschafteten Gewinn bezahlt werden. Lohnt sich diese Art der Finanzierung für das Unternehmen?

b) Wie hoch muß der Zinssatz auf dem Kapitalmarkt sein, damit es für das Unternehmen lohnender ist, seine eigenen Mittel in fest verzinsliche Wertpapiere statt in die neue Maschine zu investieren?

Fallen bei einem Geschäft Zahlungen zu verschiedenen Zeitpunkten an, dann müssen die Beträge auf einen einheitlichen Termin auf– oder abgezinst werden. Wir wählen als Bezugspunkt das Ende des 5. Jahres nach Anschaffung der Maschine. Bei einem Zinssatz von z % haben Zahlungen, die k Jahre zuvor geleistet wurden, den $(1+\frac{z}{100})^k$–fachen Wert des ursprünglich gezahlten Betrages. Die Gewährung des Kredits stellen wir uns so vor, daß das Unternehmen beim Maschinenkauf sein Konto entsprechend überzieht und alle ein– oder ausgezahlten Beträge mit dem Zinssatz des Kredits zugunsten bzw. zu Lasten der Firma verzinst werden. Dann sind die folgenden Buchungen zu berücksichtigen (wobei $a_5 = 20.000$ die Kaufkosten bezeichnet):

Gegenstand	Betrag	Wert am Ende des 5. Jahres
Maschinenkauf	$-a_5$	$-a_5\left(1+\frac{z}{100}\right)^5$
Einzahlung aus Gewinn	a_4	$a_4\left(1+\frac{z}{100}\right)^4$
– " –	a_3	$a_3\left(1+\frac{z}{100}\right)^3$
– " –	a_2	$a_2\left(1+\frac{z}{100}\right)^2$
– " –	a_1	$a_1\left(1+\frac{z}{100}\right)$
Gewinn und Liquidationserlös	a_0	a_0

Setzt man $x = \left(1+\frac{z}{100}\right)$, so wird der Gesamtwert aller Transaktionen beschrieben durch die Funktion

$$p(x) \quad = \quad a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 - a_5 x^5 \,.$$

In Teil a) unseres Problems muß also für gegebene a_i und gegebenes x der Wert $p(x)$ des Polynoms bestimmt werden (*Polynomwertberechnung*).
In Teil b) ist ein $\hat{x}$ $\left(= 1+\frac{\hat{z}}{100}\right)$ gesucht, das eine spezielle Eigenschaft haben soll: Mit dem Zinssatz $\hat{z}$ aufgezinst, sollen die Gewinne aus der Investition der Maschine genau so viel wert sein, wie die Anlage des Geldes auf dem Kapitalmarkt zu eben diesem Zinssatz. (Dieser Zinssatz heißt interner Zinsfuß der Investition. Für jeden höheren Zinssatz ist dann die Maschine weniger rentabel als die Geldanlage auf dem Kapitalmarkt.) Für das $\hat{x}$ muß also gelten

$$a_0 + a_1 \hat{x} + a_2 \hat{x}^2 + a_3 \hat{x}^3 + a_4 \hat{x}^4 \quad = \quad a_5 \hat{x}^5$$

bzw. $$p(\hat{x}) \quad = \quad 0 \qquad\qquad (\textit{Nullstellenberechnung})\,.$$

Wir befassen uns zunächst mit einer geeigneten Methode zur Polynomwertberechnung und betrachten zunächst die Holzhammermethode zur Polynomwertberechnung (wehe, wer diese programmiert).

$$\text{Berechne}\quad \left.\begin{array}{ll} a_n\, x^n & (n\ \text{ Multiplikationen}) \\ a_{n-1}\, x^{n-1} & (n-1\ \text{ Multipl.}) \\ \vdots & \\ a_1\, x & (1\ \text{ Multipl.}) \\ a_0 & \end{array}\right\} \quad \begin{array}{l} 1+2+\ldots+n \;=\; \frac{(n+1)n}{2}\quad \text{Multipl.} \\ n\quad \text{Additionen} \end{array}$$

und addiere.

Etwas Aufwand sparen kann man, indem x^{i-1} gespeichert wird und daraus x^i mit einer zusätzlichen Multiplikation errechnet wird.

Die elegante Methode wird zunächst an einem Beispiel demonstriert:

$$\begin{array}{llllllllllll}
n=4: & p(x) & = & 1 & + & 2x & + & 3x^2 & + & 4x^3 & + & 5x^4 \\
 & & = & 1 & + & x(2 & + & 3x & + & 4x^2 & + & 5x^3\,) \\
 & & = & 1 & + & x(2 & + & x(3 & + & 4x & + & 5x^2\,)) \\
 & & = & 1 & + & x(2 & + & x(3 & + & x(4 & + & 5x\,))) \\
 & & = & a_0 & + & x(a_1 & + & x(a_2 & + & x(a_3 & + & x\,a_4\,))) \\
 & & = & a_{n-4} & + & x(a_{n-3} & + & x(a_{n-2} & + & x(a_{n-1} & + & \underbrace{x\,a_n}_{x\,b_{n-1}}\,)))
\end{array}$$

$$x(a_{n-2} + \underbrace{x \cdot b_{n-2}}\,)$$

$$= a_{n-4} + x(a_{n-3} + \underbrace{x \cdot b_{n-3}}\,)$$

$$= \underbrace{a_{n-4} + x \cdot b_{n-4}}_{b_{n-5}}$$

Diese Auswertung benötigt nur 4 Multiplikationen und 4 Additionen.

Wir fassen diese Klammer– und Multiplikationsmethode im sog. Horner–Schema zusammen:

Horner–Schema zur Berechnung von $p(\hat{x})$.

$$\begin{array}{ccccccc}
a_n & a_{n-1} & a_{n-2} & a_{n-3} & \ldots & a_1 & a_0 \\
\downarrow + & \downarrow + & \downarrow + & \downarrow + & & \downarrow + & \downarrow + \\
0 & \nearrow\ \hat{x}\, b_{n-1} & \nearrow\ \hat{x}\, b_{n-2} & \nearrow\ \hat{x}\, b_{n-3} & \nearrow\ \ldots & \hat{x}\, b_1 & \nearrow\ \hat{x}\, b_0 \\
\hline
b_{n-1} & b_{n-2} & b_{n-3} & b_{n-4} & & b_0 & b_{-1} \\
|| & & & & & & || \\
a_n & & & & & & p(\hat{x})
\end{array}$$

Es gilt also laut Beispiel

$$\boxed{\begin{array}{rcll} b_{n-1} & = & a_n & \\ b_{i-1} & = & a_i + \hat{x}\, b_i, & i = n-1, n-2, \ldots, 1 \\ b_{-1} & = & a_0 + \hat{x}\, b_0 & = \; p(\hat{x}) \end{array}} \qquad \text{Horner–Schema} \quad (2.1)$$

Um zu sehen, was diese Umrechnungsformeln allgemein leisten, tragen wir sie in unser gegebenes Polynom $p(x)$ ein.

Sei also: $a_i = b_{i-1} - \hat{x}\, b_i, \quad i = n-1, n-2, \ldots, 0, \quad a_n = b_{n-1}$, dann folgt:

$$\begin{aligned}
p(x) \;=\; \sum_{i=0}^{n} a_i\, x^i \;&=\; \sum_{i=0}^{n-1} a_i\, x^i \;+\; a_n\, x^n \\
&=\; \sum_{i=0}^{n-1} (b_{i-1} - \hat{x}\, b_i)\, x^i \;+\; b_{n-1}\, x^n \\
&=\; \sum_{i=0}^{n-1} b_{i-1}\, x^i \;-\; \sum_{i=0}^{n-1} \hat{x}\, b_i\, x^i \;+\; b_{n-1}\, x^n \\
&=\; b_{-1} \;+\; \sum_{i=1}^{n-1} b_{i-1}\, x^i \;-\; \sum_{i=0}^{n-1} \hat{x}\, b_i\, x^i \;+\; b_{n-1}\, x^n \\
&=\; b_{-1} \;+\; \sum_{i=1}^{n} b_{i-1}\, x^i \;-\; \hat{x} \sum_{i=0}^{n-1} b_i\, x^i \\
&=\; b_{-1} \;+\; x \sum_{i=1}^{n} b_{i-1}\, x^{i-1} \;-\; \hat{x} \sum_{i=0}^{n-1} b_i\, x^i \\
&=\; b_{-1} \;+\; x \sum_{i=0}^{n-1} b_i\, x^i \;-\; \hat{x} \sum_{i=0}^{n-1} b_i\, x^i \\
&=\; b_{-1} \;+\; (x - \hat{x}) \sum_{i=0}^{n-1} b_i\, x^i \,.
\end{aligned}$$

Setzt man nun $x = \hat{x}$, so folgt $p(\hat{x}) = b_{-1}$, also die Behauptung der 3. Zeile von (2.1), und wir erhalten

$$p(x) \quad = \quad p(\hat{x}) \;+\; (x - \hat{x}) \sum_{i=0}^{n-1} b_i\, x^i \,.$$

Wir fassen das Ergebnis in folgendem Satz zusammen.

Satz 2.2
Sei $p_n(x) = \sum\limits_{i=0}^{n} a_i\, x^i$ ein beliebiges Polynom n–ten Grades $(n \in \mathbb{N})$ und $\hat{x} \in \mathbb{C}$ beliebig aber fest. Dann ist das Polynom $p_{n-1}(x) := \sum\limits_{i=0}^{n-1} b_i\, x^i$, dessen Koeffizienten b_i mit Hilfe des Horner–Schemas aus $p_n(x)$ und $\hat{x}$ berechnet werden, vom Grad $n-1$, und es gilt

$$p_n(x) \quad = \quad p_n(\hat{x}) \,+\, (x - \hat{x})\, p_{n-1}(x)\,. \tag{2.2}$$

Durch Differenzieren von (2.2) folgt

$$\begin{aligned} p_n'(x) \quad &= \quad p_{n-1}(x) \,+\, (x-\hat{x})\, p_{n-1}'(x)\,, \\ \text{also} \qquad p_n'(\hat{x}) \quad &= \quad p_{n-1}(\hat{x})\,. \end{aligned}$$

Dies liefert die

Folgerung 2.3
Die Ableitung $p_n'(\hat{x})$ eines beliebigen Polynoms $p_n \in \Pi_n$ an einer vorgegebenen Stelle $\hat{x}$ ist als Polynomwert von $p_{n-1} \in \Pi_{n-1}$ berechenbar, wobei p_{n-1} aus p_n mittels Horner–Schema berechnet wurde.

Dies kann mit Hilfe des doppelten Horner–Schemas ausgeführt werden.

Beispiel: $p(x) = 1 - 2x + 3x^2 - 4x^3 + 5x^4, \qquad \hat{x} = 2$

$$\begin{array}{c|ccccccc}
 & 5 & -4 & 3 & -2 & 1 & & \\
 & (a_4) & (a_3) & (a_2) & (a_1) & (a_0) & & \\
 & \downarrow + & \downarrow + & \downarrow + & \downarrow + & \downarrow + & & \\
\hat{x} = 2 & & 2\cdot 5 & 2\cdot 6 & 2\cdot 15 & 2\cdot 28 & & \\
\hline
 & 5 & 6 & 15 & 28 & 57 & = & p(2) \\
 & (b_3) & (b_2) & (b_1) & (b_0) & (b_{-1}) & & \\
 & & & & & & & \\
 & & 2\cdot 5 & 2\cdot 16 & 2\cdot 47 & & & \\
\hline
 & 5 & 16 & 47 & 122 & = & p'(2) &
\end{array}$$

Bemerkung: Das Schema kann zur Berechnung höherer Ableitungen erweitert werden. Vorsicht: Hierbei treten noch Faktoren auf (vgl. z.B. Opfer, Korollar 2.5).

Ist $\hat{x}$ eine Nullstelle von $p_n(x)$, so zeigt Satz 2.2

$$p_n(x) \quad = \quad (x - \hat{x})\, p_{n-1}(x)\,. \tag{2.3}$$

Man kann mit dem Horner–Schema also einen Linearfaktor $(x - \hat{x})$ abspalten.

Definition 2.4
Hat ein Polynom $p \in \Pi_n$ eine Darstellung

$$p(x) \;=\; (x - \hat{x})^k\, q(x)\,, \qquad k \;\in\; \mathbb{N}\,,$$

mit einem Polynom $q \in \Pi_{n-k}$ mit $q(\hat{x}) \neq 0$, so heißt $\hat{x}$ *k–fache Nullstelle* von p, und k heißt *algebraische Vielfachheit* der Nullstelle $\hat{x}$.

Unter Benutzung dieser Begriffsbildung erhält man eine weitere Folgerung aus Satz 2.2.

Lemma 2.5
Ein Polynom $p_n \in \Pi_n$ mit $n+1$ Nullstellen (wobei mehrfache Nullstellen zugelassen sind: eine k–fache Nullstelle zählt dann als k Nullstellen) verschwindet identisch, d.h. $p_n(x) = 0$ für alle $x \in \mathbb{C}$. (Oder äquivalent: $p(x) = \sum\limits_{i=0}^{n} a_i\, x^i$ und $a_i = 0$ für alle i)
Andere Formulierung: $p_n \in \Pi_n$ hat höchstens n Nullstellen.

Beweis durch vollständige Induktion (zunächst für paarweise verschiedene Nullstellen):
Induktionsanfang: $n = 0$, also $p_0(x) = a_0$, ist $p_0(\hat{x}) = 0$ für ein $\hat{x} \in \mathbb{C}$ so folgt $a_0 = 0$.

Induktionsvoraussetzung: Die Behauptung sei richtig für $n - 1$.

Induktionsschritt: $p_n \in \Pi_n$ habe die $n + 1$ (paarweise verschiedenen) Nullstellen $x_1, \dots, x_{n+1}$; dann gilt nach Satz 2.2

$$p_n(x) \;=\; p_n(x_{n+1}) + (x - x_{n+1})\, p_{n-1}(x) \;=\; (x - x_{n+1})\, p_{n-1}(x),\;\; p_{n-1} \in \Pi_{n-1}$$

p_{n-1} hat die Nullstellen $x_1, \dots, x_n$, also ist $p_{n-1} \equiv 0$ nach Induktionsvoraussetzung.

Hat p_n mehrfache Nullstellen, so hat es nach Def. 2.4 eine Darstellung

$$p(x) \;=\; \left(\prod_{j=1}^{\ell} (x - x_j)^{k_j} \right) q_{n-m}(x), \quad m \;=\; \sum_{j=1}^{\ell} k_j, \quad k_j \;>\; 1, \quad q_{n-m} \;\in\; \Pi_{n-m}$$

mit einem Polynom q_{n-m}, das keine mehrfachen Nullstellen besitzt und auf das der vorige Beweis angewendet werden kann. ■

Auf das Problem der *Nullstellenberechnung von Polynomen* gehen wir an dieser Stelle nur beispielhaft ein. Wir werden später Methoden zur Berechnung von Nullstellen von Funktionen in allgemeinerem Rahmen untersuchen.

Das Newton–Verfahren

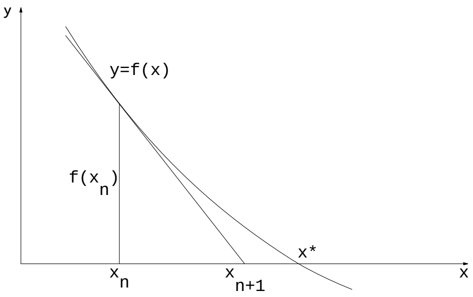
Sei $f(x)$ eine reell– wertige Funktion einer reellen Variablen (z.B. ein Polynom).

Für einen Ausgangspunkt $x_0 \in \mathbb{R}$

berechne man die Folge $(x_n)_{n\in\mathbb{N}}$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad x_0 \in \mathbb{R}.$$

(Iterationsverfahren)



Die Folge ist erklärt, falls $f'(x_n) \neq 0$ ist für alle n.

Wird $y = f(x) = p_n(x)$ durch ein Polynom gegeben, so können Nenner und Zähler der Iterationsvorschrift mit Hilfe des Horner–Schemas berechnet werden.

Wir werden (später) zeigen (was obige Zeichnung vermuten läßt):
Sei $y = f(x)$ eine stetig differenzierbare Funktion mit einer Nullstelle x^*. Ist dann $f'(x^*) \neq 0$ und wird x_0 hinreichend nahe bei x^* gewählt, so konvergiert die durch das Newton–Verfahren gelieferte Folge gegen x^*.

Problematisch ist immer das Finden einer geeigneten Ausgangsnäherung x_0 für x^*. Für Polynome wird das erleichtert durch

Satz 2.6 *Einschließungssatz für Polynomnullstellen.*

Hat das Polynom $p(x) = \sum_{i=0}^{n} a_i\, x^i$ mit $a_n = 1$ eine Nullstelle $x^* : p(x^*) = 0$, so gilt

a) $|x^*| \leq \max\left(1, \sum_{i=0}^{n-1} |a_i|\right)$

b) $|x^*| \leq \max_{i=1,\dots,n-1} (|a_0|,\ 1 + |a_i|)$

Beweis:

Wir zeigen: Außerhalb dieser Schranken gilt immer $|p(x)| > 0$.

Hilfsmittel: Dreiecksungleichung (z.B. Königsberger 1, §2.2)

$$||x|-|y|| \overset{(1)}{\leq} |x-y| \overset{(2)}{\leq} |x|+|y| \qquad \forall\, x,y \in \mathbb{R}$$

$$\left|\sum_{i=0}^{n} x_i\right| \overset{(3)}{\leq} \sum_{i=0}^{n} |x_i| \qquad \forall\, x_i \in \mathbb{R}$$

Wir zitieren diese Formeln unter den angegebenen Nummern (1)–(3).

a) Sei $|x| > \max\left(1, \sum_{i=0}^{n-1} |a_i|\right)$. Es gilt wegen $a_n = 1$:

$$\begin{aligned}
|p(x)| = \left|x^n + \sum_{i=0}^{n-1} a_i\, x^i\right| &\overset{(1)}{\geq} |x^n| - \left|\sum_{i=0}^{n-1} a_i\, x^i\right| \\
&\overset{(3)}{\geq} |x^n| - \sum_{i=0}^{n-1} |a_i\, x^i| = |x^n| - \sum_{i=0}^{n-1} |a_i|\,|x^i| \\
&\geq |x^n| - \sum_{i=0}^{n-1} |a_i|\,|x^{n-1}|, \quad \text{denn} \quad |x| > 1\,; \\
&= |x^{n-1}| \left(|x| - \sum_{i=0}^{n-1} |a_i|\right) > 0\,,
\end{aligned}$$

denn $|x^{n-1}| > 1$ wegen $|x| > 1$ und $|x| - \sum_{i=0}^{n-1} |a_i| > 0$ nach obiger Voraussetzung.

b) Sei $|x| > \max_{i=1,\ldots,n-1} (|a_0|, 1 + |a_i|)$. Unter Beweis a) wurde gezeigt:

$$\begin{aligned}
|p(x)| &\geq |x^n| - \sum_{i=0}^{n-1} |a_i|\,|x^i| \\
&= |x^{n-1}| \underbrace{(|x| - |a_{n-1}|)}_{>1 \quad \text{nach Voraussetzung}} - \sum_{i=0}^{n-2} |a_i|\,|x^i| \\
&> |x^{n-1}| - \sum_{i=0}^{n-2} |a_i|\,|x^i| = |x^{n-2}| \underbrace{(|x| - |a_{n-2}|)}_{>1} - \sum_{i=0}^{n-3} |a_i|\,|x^i| \\
&> |x^{n-2}| - \sum_{i=0}^{n-3} |a_i|\,|x^i| = |x^{n-3}| \underbrace{(|x| - |a_{n-3}|)}_{>1} - \sum_{i=0}^{n-4} |a_i|\,|x^i| \\
&\vdots \\
&> |x^1| - \sum_{i=0}^{0} |a_i|\,|x^i| = |x| - |a_0| > 0 \quad \text{nach Voraussetzung.}
\end{aligned}$$

■

§ 3 Interpolation

Motivation und Beispiele:

a) Zeichne die Höhenlinie eines Landschaftsquerschnitts, nachdem die Höhe in verschiedenen Punkten gemessen wurde.

b) Bei einer physikalischen oder chemischen Versuchsreihe sollen Meßpunkte durch eine glatte Kurve verbunden (interpoliert) werden.

c) Eine Funktion f, von der man nur einzelne Punkte kennt, soll durch eine glatte Kurve dargestellt oder angenähert werden.

Es ist mathematisch naheliegend, die Meßpunkte durch ein Polynom (einfachste mathematische Kurve) zu verbinden. Das führt zu folgender Problemstellung:

Interpolationsproblem:
Gegeben seien Punkte (x_j, f_j), $j = 0, 1, \ldots, n$, $x_j, f_j \in \mathbb{C}$, $n \in \mathbb{N}$, mit paarweise verschiedenen Stützstellen $x_i \neq x_j$ für $i \neq j$.

Gesucht ist ein Polynom $p(x)$ möglichst niedrigen Grades, das diese Punkte verbindet (interpoliert), d.h.

$$p(x_j) = f_j, \qquad j = 0, 1, 2, \ldots, n\,. \tag{3.1}$$

Fragen:

1. Gibt es so ein „Interpolationspolynom“ (Existenzfrage) und wenn ja, von welchem Grad ist es?

2. Gibt es verschiedene, ggf. wieviele Polynome, welche der Interpolationsbedingung (3.1) genügen? (Eindeutigkeitsfrage)

Es ist ein naheliegender Ansatz, die Koeffizienten a_j eines Polynoms

$$p(x) = \sum_{j=0}^{k} a_j\, x^j$$

von noch unbekanntem Grad k durch ein Gleichungssystem zu bestimmen, in dem die Koeffizienten als Unbekannte auftreten.

$$\begin{array}{lcl} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n & = & f_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n & = & f_1 \\ \vdots & & \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n & = & f_n \end{array} \tag{3.2}$$

Dieser Ansatz läßt vermuten, daß ein Polynom n–ten Grades die Interpolationsbedingung (3.1) erfüllt ($n+1$ Gleichungen für die $n+1$ Unbekannten a_j, $j = 0, 1, \ldots, n$), falls das Gleichungssystem lösbar ist.

Wir werden später sehen, daß das System (3.2) lösbar ist und man auf diese Weise ein Polynom mit Höchstgrad n als Lösung des Interpolationsproblems berechnen kann.

Diese Vorgehensweise ist jedoch für eine Rechnung mit Computer *nicht empfehlenswert*, da solche Gleichungssysteme oft eine „schlechte Kondition" besitzen, d.h. die Lösung mit dem Rechner kann sehr rundungsfehleranfällig sein. Beispiele hierfür werden wir noch kennenlernen.

Die Existenzfrage kann auch ohne Rechnung gelöst werden. Lagrange hat folgende Lösung angegeben.

Satz 3.1 (Lagrange)
Das Interpolationsproblem

„Bestimme $p_n \in \Pi_n$ so, daß $p_n(x_j) = f_j$ für $j = 0, 1, \ldots, n$"

wird gelöst durch

$$p_n(x) = \sum_{j=0}^{n} \ell_j(x)\, f_j \tag{3.3}$$

wobei

$$\begin{aligned} \ell_j(x) &= \frac{(x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_{j-1})\,(x - x_{j+1}) \cdot \ldots \cdot (x - x_n)}{(x_j - x_0)\,(x_j - x_1) \cdot \ldots \cdot (x_j - x_{j-1})\,(x_j - x_{j+1}) \cdot \ldots \cdot (x_j - x_n)} \\ &= \prod_{\substack{\nu=0 \\ \nu \neq j}}^{n} \frac{(x - x_\nu)}{(x_j - x_\nu)} \end{aligned} \tag{3.4}$$

(Interpolationspolynom nach Lagrange)

Beweis:
Durch Ausmultiplizieren von (3.4) erkennt man $\ell_j \in \Pi_n \ \forall j$, also auch $p_n \in \Pi_n$.

Da gemäß (3.4)

$$\ell_i(x_j) = \delta_{ij} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases},$$

folgt aus (3.3): $p_n(x_j) = f_j \quad \forall j$. ■

Beachte:
Im Rechner ist die Interpolationsgenauigkeit von p_n an den Stützstellen x_j gleich der Genauigkeit, mit der die Werte f_j im Rechner dargestellt werden können, denn 0 und 1 sind immer Maschinenzahlen.

Als nächstes stellt sich die Frage nach der Eindeutigkeit der Lösung. Sie wird beantwortet durch

Satz 3.2
Das Interpolationsproblem aus Satz 3.1 hat genau eine Lösung.

Beweis (indirekt):
Angenommen, es gebe zwei verschiedene Polynome

$$p_n^{(1)},\, p_n^{(2)} \;\in\; \Pi_n \qquad \text{mit} \quad p_n^{(1)}(x_j) \;=\; p_n^{(2)}(x_j) \;=\; f_j, \qquad j = 0, 1, \ldots, n$$

Dann ist $p(x) := p_n^{(1)}(x) - p_n^{(2)}(x)$ ebenfalls ein Polynom vom Höchstgrad n und besitzt $n+1$ Nullstellen: $p(x_j) = 0,\ j = 0, 1, \ldots, n$. Laut Lemma 2.5 gilt dann $p(x) \equiv 0\ \forall x$, also $p_n^{(1)}(x) = p_n^{(2)}(x)$. ∎

Bemerkung:
Diese beiden Sätze besagen auch, daß das Gleichungssystem (3.2) genau eine Lösung hat.

Für die Berechnung von Funktionswerten ist die Benutzung des Interpolationspolynoms in Langrange–Form zu aufwendig (zu viele Multiplikationen und Divisionen). Eine für numerische Zwecke bessere Darstellung wurde von Newton gegeben.

Satz 3.3 (Newton)
Das Interpolationsproblem aus Satz 3.1 wird auch gelöst durch das *Newtonsche Interpolationspolynom*

$$\begin{aligned} p_n(x) &= c_0 + c_1(x - x_0) + c_2(x - x_0)\,(x - x_1) + \ldots + \\ &\quad\; c_n(x - x_0)\,(x - x_1) \cdot \ldots \cdot (x - x_{n-1}) \\ &= \sum_{j=0}^{n} c_j \prod_{\nu=0}^{j-1} (x - x_\nu) \end{aligned} \tag{3.5}$$

mit geeigneten Konstanten $c_i \in \mathbb{C}$.

$$\left(\text{Konvention:} \quad \prod_{\nu=0}^{j-1} (x - x_\nu) \;:=\; 1 \;\text{ falls }\; j - 1 < 0 \right)$$

Beweis:
$p_n \in \Pi_n$ folgt direkt aus der Darstellung (3.5). Die Berechenbarkeit der Konstanten c_j ergibt sich unmittelbar, indem man die Interpolationsbedingungen als Gleichungssystem aufschreibt:

$$
\begin{aligned}
p_n(x_0) &= c_0 && = f_0 \\
p_n(x_1) &= c_0 + c_1(x_1 - x_0) && = f_1 \\
p_n(x_2) &= c_0 + c_1(x_2 - x_0) + c_2(x_2 - x_0)\,(x_2 - x_1) && = f_2 \\
&\ \vdots && \ \vdots \\
p_n(x_n) &= c_0 + c_1(x_n - x_0) + c_2(x_n - x_0)\,(x_n - x_1) + \ldots + c_n \prod_{\nu=0}^{n-1} (x_n - x_\nu) && = f_n
\end{aligned}
\tag{3.6}
$$

Die Konstanten c_j sind „von oben nach unten" berechenbar. Damit sind die Interpolationsbedingungen einschließlich der Gradforderung erfüllt. Die Eindeutigkeitsfrage ist bereits durch Satz 3.2 geklärt. ■

Aus der Darstellung (3.5) und dem zugehörigen Gleichungssystem (3.6) kann man folgende Eigenschaften ablesen:

1) Für jedes $j \in \{0, 1, \ldots, n\}$ wird der Koeffizient c_j nur aus den ersten $(j+1)$ Gleichungen $p_n(x_i) = f_i,\ i = 0, 1, \ldots, j,$ berechnet, ist also von den Interpolationsbedingungen für $i > j$ unabhängig.

   Wir schreiben:

$$
\begin{aligned}
c_0 &= f[x_0] = f_0 \\
c_i &= f[x_0, \ldots, x_i], \qquad i = 1, \ldots, n\,.
\end{aligned}
$$

   Aus dieser Eigenschaft folgt insbesondere:

   Ist $p_n \in \Pi_n$ Lösung der Interpolationsaufgabe für die Punkte $(x_i, f_i),\ i = 0, \ldots, n,$ und nimmt man einen weiteren Punkt (x_{n+1}, f_{n+1}) hinzu, so wird die Lösung $p_{n+1} \in \Pi_{n+1}$ des erweiterten Interpolationsproblems gemäß (3.5) gegeben durch

$$
p_{n+1}(x) = p_n(x) + c_{n+1} \prod_{j=0}^{n} (x - x_j) \tag{3.7}
$$

   und c_{n+1} errechnet sich aus der Gleichung

$$
c_{n+1} = \frac{f_{n+1} - p_n(x_{n+1})}{\prod\limits_{j=0}^{n} (x_{n+1} - x_j)}
$$

   Man kann also die Lösung des Interpolationsproblems für $(n+1)$ Punkte „ausbauen" zur Lösung des Problems für $(n+2)$ Punkte, ein Vorteil, den weder die Lösung des Gleichungssystems (3.2) noch die Lagrange–Form bietet.

2) Vergleicht man das Newton–Polynom mit der üblichen Polynomdarstellung

$$p_n(x) \quad = \quad \sum_{j=0}^{n} c_j \prod_{\nu=0}^{j-1} (x - x_\nu) \quad = \quad \sum_{j=0}^{n} a_j \, x^j \, ,$$

so gilt für den Koeffizienten a_n der höchsten x–Potenz

$$a_n \;=\; c_n \qquad \text{(Beweis durch Ausmultiplizieren)}$$

Da Interpolationpolynome eindeutig sind (Satz 3.2), folgt hieraus, daß $a_n = c_n = f[x_0, \ldots, x_n]$ unabhängig ist von der Reihenfolge der Punkte (x_i, f_i), $i \leq j$.

3) Weiter zeigt (3.6): Ist a_j der Koeffizient der höchsten x–Potenz des Interpolationspolynoms für die Punkte (x_i, f_i), $i = 0, 1, \ldots, j$, so folgt wie in 2) $a_j = c_j = f[x_0, \ldots, x_j]$ und man erhält analog:

Alle $c_j = f[x_0, \ldots, x_j]$, $j = 0, 1, \ldots, n$, $(c_0 = f[x_0]$ falls $j = 0)$, sind unabhängig von der Reihenfolge der Punkte (x_i, f_i).

4) Sind die c_i einmal bekannt, so erfordert die Auswertung des Polynoms $p_n(x)$ an einer Stelle x wesentlich weniger Rechenoperationen als bei Benutzung der Lagrange–Form, insbesondere auch, weil man (3.5) durch ein „Horner–ähnliches“ Schema programmieren kann (Aufgabe!).

Will man ein Interpolationspolynom nur an einer (oder nur an sehr wenigen) Stelle(n) auswerten, so empfiehlt sich dafür die Neville–Formel, die diese Auswertung ohne Berechnung der c_i leistet.

Satz 3.4 (Rekursionsformel für Polynome nach Neville)
Bei gegebenen Punkten (x_j, f_j), $j = 0, \ldots, n$, bezeichne $P_{i,i+1,\ldots,i+k}$ das Interpolationspolynom $\in \Pi_k$ zu den Punkten (x_j, f_j), $j = i, i+1, \ldots, i+k$. Dann gilt die Rekursionsformel

$$\begin{aligned} P_i(x) &= f_i \, , \\ P_{i,\ldots,i+k}(x) &= \frac{(x - x_i)\, P_{i+1,\ldots,i+k}(x) - (x - x_{i+k})\, P_{i,\ldots,i+k-1}(x)}{x_{i+k} - x_i} \, , \quad k > 0 \, . \end{aligned} \tag{3.8}$$

Beweis:
Aus der Berechnungsformel (3.8) folgt $P_{i,\ldots,i+k} \in \Pi_k$ und durch Einsetzen der Werte (x_j, f_j), $j = i, i+1, \ldots, i+k$, erhält man direkt

$$P_{i,\ldots,i+k}(x_j) \;=\; f_j \, , \qquad j = i, \ldots, i+k \, ,$$

also die Interpolationseigenschaft. Die Eindeutigkeit ist durch Satz 3.2 gesichert. ∎

Die Berechnung eines einzelnen Polynomwertes unter Benutzung von Satz 3.4 gestaltet sich übersichtlich nach dem Neville–Schema, das wir für $n = 3$ angeben

Neville–Schema (für $n = 3$)

$$\begin{array}{c|lllllll}
x & P_i(x) & & P_{i,i+1}(x) & & P_{i,i+1,i+2}(x) & & P_{i,i+1,i+2,i+3}(x) \; = \; p_3(x) \\
\hline
x_0 & P_0(x) = f_0 & \searrow & & & & & \\
 & & \nearrow & P_{0,1}(x) & \searrow & & & \\
x_1 & P_1(x) = f_1 & \searrow & & \nearrow & P_{0,1,2}(x) & \searrow & \\
 & & \nearrow & P_{1,2}(x) & \searrow & & \nearrow & P_{0,1,2,3}(x) \\
x_2 & P_2(x) = f_2 & \searrow & & \nearrow & P_{1,2,3}(x) & & \\
 & & \nearrow & P_{2,3}(x) & & & & \\
x_3 & P_3(x) = f_3 & & & & & &
\end{array}$$

Für die Programmierung schreibt man die Neville–Formel am besten in der Gestalt

$$P_{i,\ldots,i+k}(x) \;=\; P_{i+1,\ldots,i+k}(x) \;+\; \frac{(P_{i+1,\ldots,i+k}(x) - P_{i,\ldots,i+k-1}(x))\,(x - x_{i+k})}{x_{i+k} - x_i}\,, \qquad (3.8')$$

die weniger Multiplikationen als (3.8) benötigt.

Will man das Interpolationspolynom an mehreren Stellen auswerten, so ist es günstiger (weniger Rechenoperationen), die Newtonkoeffizienten $c_i = f[x_0, \ldots, x_i]$ (*dividierte Differenzen*) zu berechnen und dann (3.5) nach einem Horner–ähnlichen Schema auszuwerten.

Satz 3.5 (Rekursive Berechnung der dividierten Differenzen)
Die *dividierten Differenzen* genügen der Rekursion

$$\begin{aligned}
f[x_j] &= f_j, \qquad j = i, i+1, \ldots, i+k \\
f[x_i, x_{i+1}, \ldots, x_{i+k}] &= \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}
\end{aligned} \qquad (3.9)$$

Bemerkung:
Dieser Satz erklärt den Namen „dividierte Differenzen“.

Beweis:
Wir beachten, daß $f[x_{i+1}, \ldots, x_{i+k}]$ bzw. $f[x_{i,\ldots,i+k-1}]$ (nach Abschnitt 2 (vor Satz 3.4)) die Koeffizienten der höchsten x–Potenz von $P_{i+1,\ldots,i+k}(x)$ bzw. $P_{i,\ldots,i+k-1}(x)$ sind. Dann liefert der Koeffizientenvergleich von x^k in der Formel (3.8) die Behauptung (3.9). ■

Auch die Berechnung der dividierten Differenzen gemäß (3.9) kann man übersichtlich in einem Schema anordnen.

Schema der dividierten Differenzen (hier für $n = 3$)

x	$f[\ ]$	$f[\ ,\ ]$	$f[\ ,\ ,\ ]$	$f[\ ,\ ,\ ,\ ]$
x_0	f_0			
		$f[x_0, x_1]$		
x_1	f_1		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
x_2	f_2		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	f_3			

In der oberen Diagonalen stehen die Koeffizienten c_i des Newtonpolynoms

$$c_i \quad = \quad f[x_0, \ldots, x_i]\,.$$

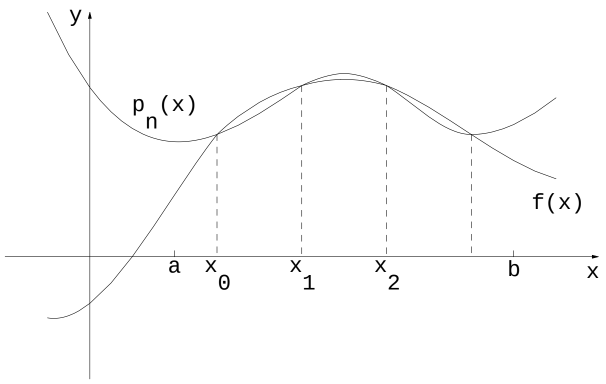
Zur Herstellung des Schemas benötigt man insgesamt $n + (n-1) + \ldots + 1 = n(n+1)/2$ Divisionen und doppelt soviele Subtraktionen. Dies ist ein konkurrenzlos geringer Aufwand im Vergleich zur Auflösung des Gleichungssystems (3.6) oder der durch (3.7) nahegelegten Methode, ganz zu schweigen vom Aufwand, den die Lagrange–Formel bei der numerischen Auswertung benötigt. U.a. darum ist dieses Verfahren auch weniger rundungsfehleranfällig.

Der Interpolationsfehler

Sind bei einer Interpolationsaufgabe nur Stützstellen x_j und Stützwerte f_j bekannt, so ist mit dem Aufstellen des zugehörigen Interpolationspolynoms die Aufgabe erschöpfend behandelt. Sind jedoch die f_i Funktionswerte einer Funktion in den Werten x_i aus einem Intervall $[a, b]$, so kann man fragen, wie gut das Interpolationspolynom die Funktion f zwischen den Stützwerten approximiert, d.h. man möchte wissen, wie groß der Approximationsfehler ist:

$$\varepsilon(x) \;:=\; f(x) \,-\, p_n(x)\,, \qquad x \,\in\, [a, b]\,.$$

Daß diese Fragestellung sinnvoll ist, belegt etwa folgendes Beispiel: Die Funktion $f(x) = \sin\, x$ wird durch eine unendliche Reihe definiert (vgl. Königsberger §10.2), die zur Auswertung in einem Rechner ungeeignet ist. Praktischerweise wird $\sin\, x$ im Rechner durch ein Polynom approximiert (das einfach und schnell auszuwerten ist), dessen Abweichung von $f(x) = \sin\, x$ so klein ist, daß sie im Rahmen der Rechnergenauigkeit vernachlässigt werden kann.



Um eine Aussage über den Approximationsfehler machen zu können, benötigt man natürlich weitere Informationen über f.

Wir untersuchen also folgendes

Problem:

$p_n \in \Pi_n$ interpoliere die Funktion f in den Stützstellen $x_i,\ i = 0, 1, \ldots, n,$ die in einem Intervall $[a, b] \subset \mathbb{R}$ liegen mögen. Gesucht ist der Interpolationsfehler $\varepsilon(z)$ zwischen den Stützstellen

$$\varepsilon(z) \;=\; f(z) \,-\, p_n(z), \qquad z \in [a, b], \qquad z \neq x_i, \quad i = 0, 1, \ldots, n\,.$$

Trick: Betrachte z $(=: x_{n+1})$ als zusätzliche, beliebige aber feste (natürlich unbekannte) Stützstelle. $p_{n+1} \in \Pi_{n+1}$ interpoliere f in den Stellen $x_0, x_1, \ldots, x_n, z$. Insbesondere ist also $f(z) = p_{n+1}(z)$.

Berücksichtigt man die Eigenschaft (3.7) des Newtonschen Interpolationspolynoms, so gilt für den Fehler

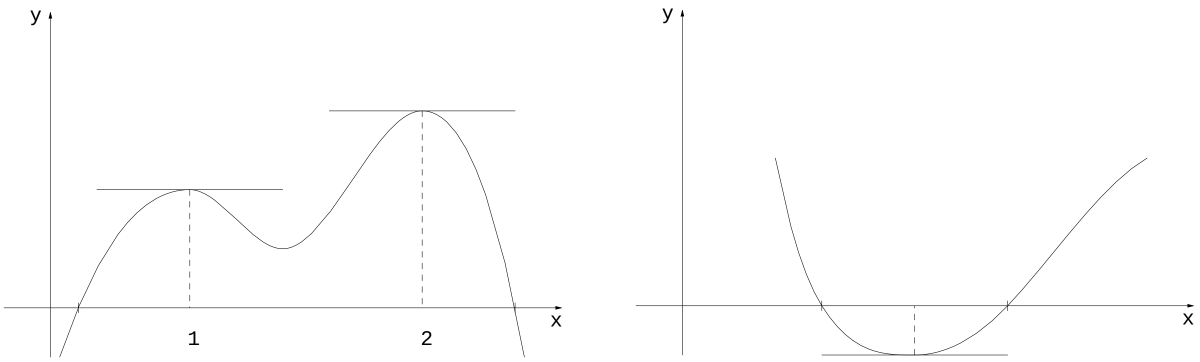
$$\begin{aligned} \varepsilon(z) &= f(z) \,-\, p_n(z) \\ &= p_{n+1}(z) \,-\, p_n(z) \\ &= c_{n+1} \prod_{i=0}^{n} (z - x_i) \\ &= f[x_0, x_1, \ldots, x_n, z] \prod_{i=0}^{n} (z - x_i) \end{aligned} \tag{3.10}$$

Wir zeigen nun, wie die dividierte Differenz $f[x_0, x_1, \ldots, x_n, z]$ mit der Funktion f zusammenhängt. Dazu benötigen wir den Satz von Rolle (Königsberger §9.4), der in der Analysis bewiesen wird.

Satz 3.6 (Rolle)

Sei $f : [\alpha, \beta] \to \mathbb{R}$ eine auf dem Intervall $[\alpha, \beta],\ -\infty < \alpha < \beta < \infty,$ stetig differenzierbare Funktion mit $f(\alpha) = 0 = f(\beta)$. Dann gibt es (mindestens) ein $\xi \in (\alpha, \beta)$ mit $f'(\xi) = 0$.

Veranschaulichung



Vorsicht:
Dieser Satz ist für komplexwertige Funktionen falsch!

Beispiel: $f(x) \; = \; e^{ix} - 1 \; = \; \cos x + i \sin x - 1, \;\; [\alpha, \beta] \; = \; [0, 2\pi].$

Die folgenden Überlegungen sind also nur für reellwertige Funktionen richtig.

Wir setzen also voraus

$f : [a,b] \;\to\; \mathbb{R},$	(reellwertig)
$f \;\in\; C^{n+1}[a,b]$	(d.h. f gehört zur Menge der Funktionen, die auf dem Intervall $[a,b]$ $(n+1)$–mal stetig differenzierbar sind)

und wenden den Satz von Rolle an auf die Differenz $r(x)$ (vgl. (3.10))

$$\begin{aligned} r(x) \;&:=\; \varepsilon(x) \;-\; f[x_0, \ldots, x_n, z] \prod_{i=0}^{n} (x - x_i)\,, \quad z \text{ beliebig aber fest} \\ &=\; f(x) \;-\; p_n(x) \;-\; f[x_0, x_1, \ldots, x_n, z] \prod_{i=0}^{n} (x - x_i) \end{aligned} \tag{3.11}$$

und die Ableitungen von r. Beachte, daß $r \in C^{n+1}[a,b]$.

$r(x)$ hat $n+2$ Nullstellen	:	$x_0, x_1, \ldots, x_n, z$. Zwischen je zwei Nullstellen von r liegt nach Satz 3.6 eine Nullstelle von $r'(x)$, also folgt
$r'(x)$ hat $n+1$ Nullstellen ,		und analog
$r''(x)$ hat n Nullstellen		
$\vdots$		
$r^{(n+1)}(x)$ hat eine Nullstelle ξ	:	$r^{(n+1)}(\xi) \;=\; 0, \quad \xi \;\in\; [a,b]\,.$

Nun ist $f \;\in\; C^{n+1}[a,b]$, $p_n \in \Pi_n$ also $p_n^{(n+1)} \;\equiv 0$, $f[x_0, x_1, \ldots, x_n, z]$ eine Zahl und $\prod_{i=0}^{n} (x - x_i)$ ein Polynom, dessen $(n+1)$ste Ableitung $= (n+1)!$ ist. Deshalb folgt aus (3.11) und $r^{(n+1)}(\xi) = 0$

$$r^{(n+1)}(\xi) \;=\; 0 \;=\; f^{(n+1)}(\xi) \;-\; f[x_0, x_1, \ldots, x_n, z] \,\cdot\, (n+1)!$$

bzw. (setze $z = x_{n+1}$)

$$f[x_0, x_1, \ldots, x_n, x_{n+1}] \;=\; \frac{1}{(n+1)!}\, f^{(n+1)}(\xi), \qquad \xi \;\in\; [a,b]\,. \tag{3.12}$$

Wird dies in $\varepsilon(z)$ eingesetzt (vgl. (3.10)), so folgt

$$\varepsilon(z) \;=\; f(z) - p_n(z) \;=\; \frac{1}{(n+1)!}\, f^{(n+1)}(\xi) \cdot \prod_{i=0}^{n} (z - x_i) \tag{3.13}$$

mit einer von z abhängigen Stelle $\xi \in [a,b]$.

Die Herleitung gilt zunächst für $z \neq x_i$, doch ist (3.13) trivialerweise auch für $z = x_i$ richtig. Beachte ferner, daß bei festen Knoten x_i, $i = 0,\dots,n$ die Zwischenstelle ξ von z abhängt. Zusammengefaßt gilt also:

Satz 3.7

$p_n \in \Pi_n$ interpoliere die Funktion $f : [a,b] \to \mathbb{R}$ an den Stützstellen $x_i \in [a,b]$, $i = 0,1,\dots,n$, $-\infty < a, b < \infty$.

a) Dann gilt für den Interpolationsfehler (vgl. (3.10))

(3.10) $$f(z) - p_n(z) \;=\; f[x_0, x_1, \dots, x_n, z] \prod_{i=0}^{n} (z - x_i), \quad z \in [a,b],\ z \neq x_i,\ \forall i$$

b) Ist zusätzlich $f \in C^{n+1}[a,b]$, so gibt es ein $\xi \in [a,b]$ mit

(3.12) $$f[x_0, x_1, \dots, x_{n+1}] \;=\; \tfrac{1}{(1+n)!}\, f^{(n+1)}(\xi),\ x_i \in [a,b], \quad i = 0,\dots,n+1, \quad x_i \neq x_j \text{ für } i \neq j$$

und der Interpolationsfehler kann dargestellt werden durch (3.13)

(3.13) $$f(x) - p_n(x) \;=\; \tfrac{1}{(1+n)!}\, f^{(n+1)}(\xi_x) \cdot \prod_{i=0}^{n} (x - x_i),$$
$$\forall\, x \in [a,b], \quad \xi_x \in [a,b], \quad \text{abhängig von } x$$

bzw.

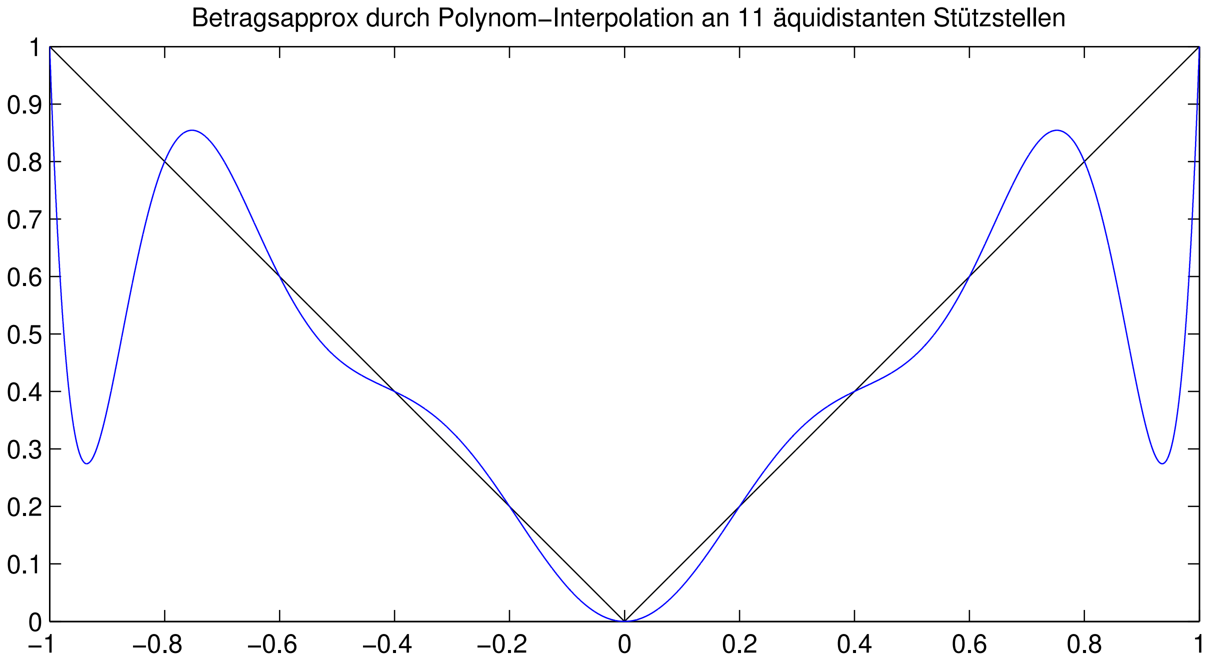
(3.14) $$|f(x) - p_n(x)| \;\leq\; \frac{1}{(1+n)!} \max_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right| \cdot \left| \prod_{i=0}^{n} (x - x_i) \right|.$$

Bemerkungen:

1) Daß $\max_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right|$ unter den Voraussetzungen unseres Satzes existiert, wird in der Analysis bewiesen. Für $f(x) = \sqrt{x}$, $x \in [0,1]$ ist diese Aussage z.B. falsch (da $f \notin C^{n+1}[0,1]$, sondern nur $\in C^{n+1}(0,1)$)

2) Da der Existenzbeweis für die Zwischenstelle ξ nicht konstruktiv ist, ist die betragsmäßige Abschätzung des Interpolationsfehlers für praktische Zwecke günstiger.

Als Beispiel interpolieren wir die Betragsfunktion im Intervall $[-1, 1]$ an 11 äquidistanten Stützstellen.

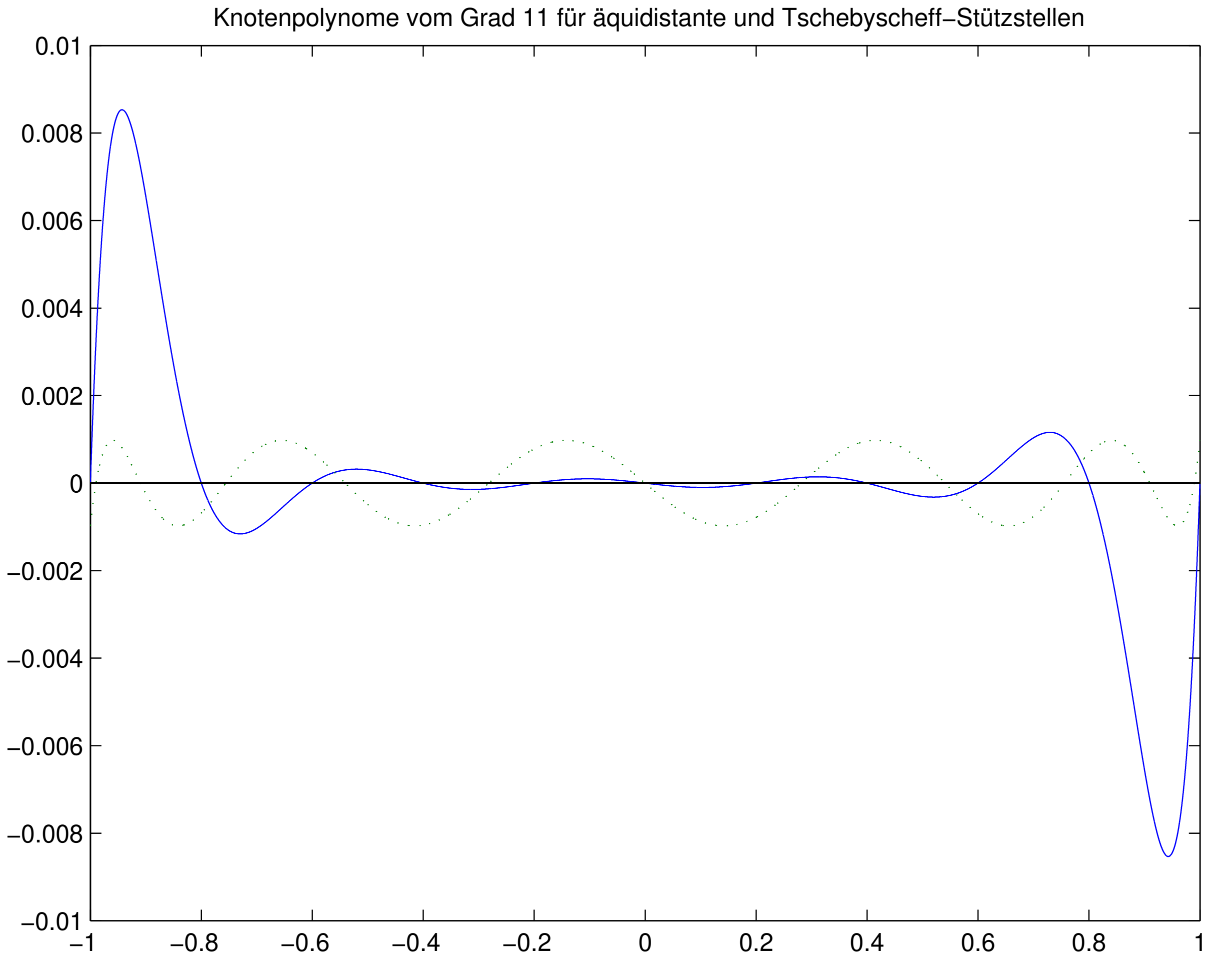


Auffällig sind die großen Fehler in der Nähe der Intervallränder. Die Fehlerabschätzung gibt Anlaß zu überlegen, ob und wie man den Approximationsfehler verkleinern kann. Den Ausdruck $f^{(n+1)}(\xi)/(n+1)!$ kann man nicht beeinflussen, da ξ nicht konstruktiv ist, wohl aber das

$$\textit{Knotenpolynom} \qquad w(x) \;:=\; \prod_{i=0}^{n} (x - x_i)\,, \tag{3.15}$$

denn über die Wahl der x_i wurde bisher noch nicht verfügt.

Nun zeigen numerische Beispiele, daß bei äquidistanter Knotenwahl die Ausschläge von $w(x)$ an den Intervallenden besonders groß werden. Die maximalen Ausschläge werden kleiner, wenn man die Stützstellen mehr auf den Rand der Intervalle konzentriert. Dafür werden die Ausschläge in der Intervallmitte etwas größer. Typisch hierfür ist folgendes Bild, das den Funktionsverlauf von $w(x)$ in $[-1, 1]$ für $n = 10$ (d.h. 11 äquidistant verteilte Stützstellen in $[-1, 1]$) zeigt (durchgezogene Linie) und gestrichelt den Verlauf, wenn man als Stützstellen die sogenannen Tschebyscheff–Stellen wählt (natürlich ebenfalls 11 Stück, vgl. dazu den folgenden Satz).



Es stellt sich also folgende

Frage: Wie kann man die Stützstellen (Knoten) $x_i \in [a, b]$, $i = 0, 1, \ldots, n$, wählen, damit der maximale Funktionswert $\max\limits_{x\in[a,b]} \left|\prod\limits_{i=0}^{n} (x - x_i)\right|$ möglichst klein wird? D.h. gesucht wird

$$\min_{x_0,\ldots,x_n} \max_{x\in[a,b]} \left|\prod_{i=0}^{n} (x - x_i)\right|$$

(sofern das Minimum existiert, was a priori nicht klar ist).

Zunächst sollte jedoch im Vorwege geklärt werden, ob man dieses Min. (falls existent) für jedes Intervall gesondert suchen muß oder ob man sich auf ein Referenzintervall beschränken kann. Letzteres ist tatsächlich der Fall. Wir zeigen zunächst:

Werden das Intervall $[a,b]$, $(-\infty < a < b < \infty)$, und die Stützstellen $x_i \in [a,b]$ durch die umkehrbar eindeutige lineare Transformation

$$y \quad = \quad c + \frac{x-a}{b-a}\,(d-c) \tag{3.16}$$

auf das Intervall $[c,d]$, $(-\infty < c < d < \infty)$, und die Werte y_i abgebildet, so werden auch die Extrema von $w(x) = \prod_{i=0}^{n}(x-x_i)$ auf die Extrema von $\tilde{w}(y) = \prod_{i=0}^{n}(y-y_i)$ abgebildet.

Die Behauptung ergibt sich aus der Beziehung

$$\begin{aligned}\tilde{w}(y) \;&=\; \prod_{i=0}^{n}(y-y_i) \;=\; \prod_{i=0}^{n}\left(\left[c+\frac{x-a}{b-a}\,(d-c)\right] \;-\; \left[c+\frac{x_i-a}{b-a}\,(d-c)\right]\right)\\ &=\; \left(\frac{d-c}{b-a}\right)^{n+1}\prod_{i=0}^{n}(x-x_i) \;=\; \left(\frac{d-c}{b-a}\right)^{n+1} w(x)\end{aligned} \tag{3.17}$$

und

$$\frac{d\tilde{w}(y(x))}{dx} \quad = \quad \tilde{w}'(y)\cdot\frac{dy(x)}{dx} \quad = \quad \left(\frac{d-c}{b-a}\right)^{n+1} w'(x)\,.$$

Es ist also $w'(x) = 0$ genau dann, wenn $\tilde{w}'(y) = 0$ mit $y = y(x)$ gem. (3.16), denn $\frac{dy}{dx} = \left(\frac{d-c}{b-a}\right) \neq 0$. ■

Wir können für die weiteren Untersuchungen also das Referenzintervall $[-1,1]$ $(c=-1,$ $d=+1)$ wählen, geeignete Stützstellen $y_i \in [-1,1]$ für $\tilde{w}(y)$ so bestimmen, daß $\max\limits_{y\in[-1,1]} |\prod_{i=0}^{n}(y-y_i)|$ minimal wird und danach die Stützstellen $x_i \in [a,b]$ durch die Rücktransformation

$$x_i \quad = \quad a + \frac{y_i-c}{d-c}\,(b-a) \quad = \quad a + \frac{y_i+1}{2}\,(b-a)$$

bestimmen.

Dann wird unsere vorseitige Frage durch die folgenden beiden Sätze beantwortet.

Satz 3.8 (Tschebyscheff–Polynome)

a) Die Funktionen $T_n(y)$, die definiert werden durch

$$T_n(y) \;=\; \cos\,(n \arccos\, y), \qquad n = 0,1,2,\ldots, \quad y \in [-1,1] \tag{3.18}$$

genügen der Rekursionsformel

$$T_{n+1}(y) \;=\; 2y\,T_n(y) \;-\; T_{n-1}(y), \qquad T_0(y) \;=\; 1, \quad T_1(y) = y, \quad n \in \mathbb{N}, \tag{3.19}$$

sind also Polynome vom Grad n ($T_n \in \Pi_n$: Tschebyscheff–Polynome).

b) T_n hat die n verschiedenen (Tschebyscheff–) Nullstellen

$$y_j \;=\; \cos\left(\frac{2j+1}{2n}\,\pi\right) \;\in\; (-1,1), \qquad j = 0,1,\ldots,n-1 \tag{3.20}$$

und nimmt im Intervall $[-1,1]$ seine Extrema an in den $n+1$ Extremalstellen

$$y_j^{(e)} \;=\; \cos\left(\frac{j\,\pi}{n}\right) \;\in\; [-1,1], \qquad j = 0,1,\ldots,n\,, \tag{3.21}$$

mit den Extremalwerten

$$T_n\left(y_j^{(e)}\right) \;=\; (-1)^j, \qquad j = 0,1,\ldots,n\,. \tag{3.22}$$

c) T_n besitzt die Darstellung

$$T_n(y) \;=\; 2^{n-1}\,\prod_{j=0}^{n-1}\,(y - y_j)\,, \quad n \in \mathbb{N}\,. \tag{3.23}$$

Unsere Frage nach der bestmöglichen Fehlerabschätzung für die Interpolation wird dann beantwortet durch

Satz 3.9

Der maximale Extremalwert von $|w(y)| \;=\; \left|\prod_{j=0}^{n}(y - y_j)\right|$ für $y \in [-1,1]$ wird minimiert, wenn man als y_j die T–Nullstellen von T_{n+1} wählt, d.h. für alle $q(y) := \prod_{j=0}^{n}(y - \xi_j)$ mit paarweise verschiedenen Werten $\xi_j \in \mathbb{R}$ gilt

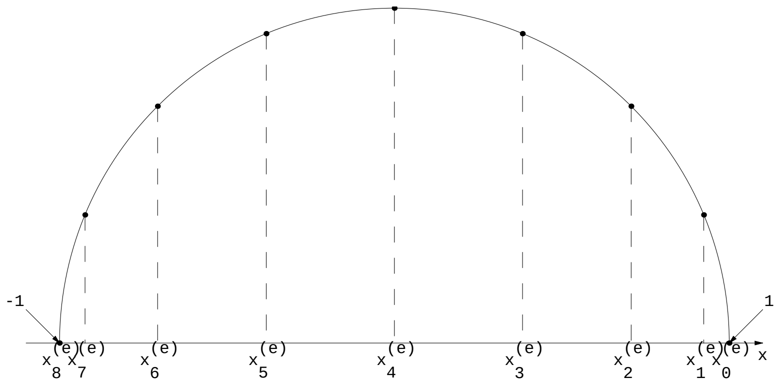
$$\max_{y\in[-1,1]}\;|w(y)| \;\;\le\;\; \max_{y\in[-1,1]}\;|q(y)| \tag{3.24}$$

Vor den Beweis dieser Sätze stellen wir noch einige

Bemerkungen

1) Durch (3.18) sind die T_n nur für das Intervall $[-1, 1]$ definiert, sie können aber natürlich mittels (3.19) auf $\mathbb{R}$ fortgesetzt werden.

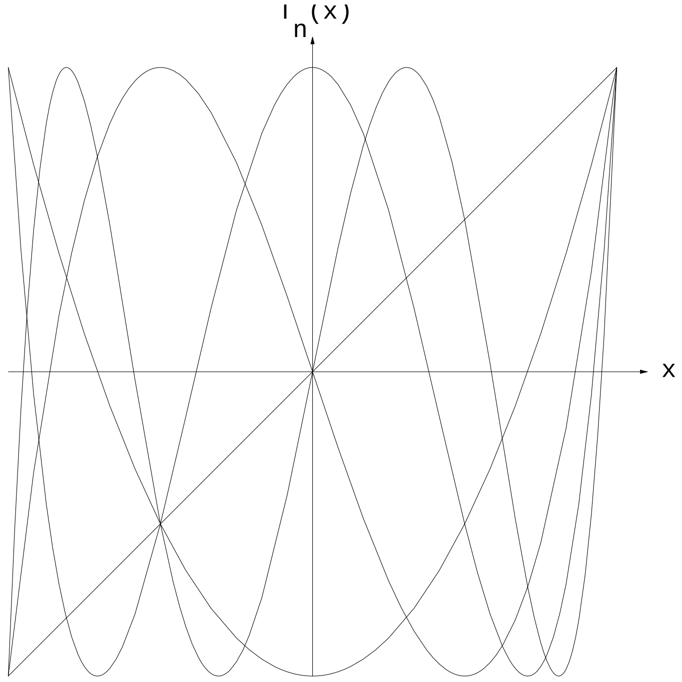
2) **Alle** Nullstellen der T–Polynome liegen in $(-1, 1)$, sie sind, ebenso wie die Extremalstellen $y_j^{(e)}$, symmetrisch zum Nullpunkt verteilt. Man kann sie sich geometrisch vorstellen als Projektionen von regelmäßig auf dem Halbkreis verteilten Punkten, z.B.



Extremalstellen von $T_8(x)$

3) Die Nullstellen und Extremalstellen in (3.20), (3.21) sind in absteigender Reihenfolge geordnet: $y_{j+1} < y_j$.

4) Zwei der Extremalstellen $\in [-1, 1]$ sind keine Waagepunkte von T_n sondern Randmaxima bzw. Randminima. Vergleiche dazu auch



Tschebyscheff–Polynome $T_n(x)$, $n = 1(1)5$.

5) Aus (3.22) und (3.23) folgt

$$(3.25) \qquad \left|\prod_{j=0}^{n-1} (y-y_j)\right| \;=\; 2^{-(n-1)}\,|T_n(y)| \;\le\; 2^{-(n-1)} \quad \text{in} \quad [-1,1]\,.$$

Das Gleichheitszeichen wird in den Extremalstellen (3.21) angenommen.

Beweis Satz 3.8

a) Grundlage des Beweises ist die trigonometrische Identität

$$(3.26) \qquad \cos[(n+1)z] \;+\; \cos[(n-1)z] \;=\; 2\cos z\,\cos(nz), \qquad n \in \mathbb{N}\,.$$

Man bestätigt sie sofort mit Hilfe der Additionstheoreme

$$\cos(\alpha \pm \beta) \quad = \quad \cos\alpha\,\cos\beta \mp \sin\alpha\,\sin\beta$$

indem man auf der linken Seite von (3.26) $\alpha = nz,\ \beta = z$ setzt.

Setzt man in (3.26) $z = \operatorname{arc\,cos} y$, so folgt

$$T_{n+1}(y) \;+\; T_{n-1}(y) \quad = \quad 2y\,T_n(y)\,.$$

Die Anfangswerte $T_0(y) = 1,\ T_1(y) = y$ erhält man sofort aus (3.18). Damit ist das Bildungsgesetz (3.19) für die Funktionen (3.18) bewiesen. Es zeigt — mittels vollständiger Induktion — sofort $T_n \in \Pi_n$.

b) Die cos–Nullstellen sind bekannt. Aus (vgl. (3.18))

$$\cos(n\,\operatorname{arc\,cos} y) \;=\; 0 \quad \text{folgt}$$

$$n\,\operatorname{arc\,cos} y \;=\; (2k+1)\,\tfrac{\pi}{2}\,, \quad k \in \mathbb{Z}, \quad \text{bzw.}$$

$$y_j \;=\; \cos\left(\tfrac{2j+1}{2n}\,\pi\right)\,, \qquad j = 0,1,\dots,n-1, \quad n \ge 1, \quad \text{also (3.20)}\,.$$

Man kann sich auf $j = 0,1,\dots,n-1$ beschränken, danach wiederholen sich die Nullstellen auf Grund der Periodizität von cos.

Die Extremalstellen der cos–Funktion sind bekannt. Aus $|\cos(n\,\operatorname{arc\,cos} y)| = 1$ folgt $n\,\operatorname{arc\,cos} y = k\,\pi,\ k \in \mathbb{Z}$ also $y_j^{(e)} = \cos\frac{j\,\pi}{n},\ j = 0,\dots,n$ (Periodizität) und $T(y_j^{(e)}) = (-1)^j$ durch Einsetzen in (3.18), also sind (3.19) – (3.22) bewiesen.

c) Da $T_n \in \Pi_n$ die n Nullstellen y_j besitzt, hat es die Form $T_n(y) = c_n \prod\limits_{j=0}^{n-1} (y-y_j)$ mit einer Konstanten c_n. Dies folgt aus (2.3) (Horner–Schema).

Aus dem Bildungsgesetz (3.19) für T_n folgt, daß der Koeffizient c_n von y^n gleich 2^{n-1} ist, also gilt (3.23). ■

Beweis Satz 3.9 (indirekt):
Wir nehmen an (vgl. (3.24)): Es gebe ξ_j, sodaß

$$\max_{y\in[-1,1]} |q(y)| \;=\; \max_{y\in[-1,1]} \left|\prod_{j=0}^{n} (y-\xi_j)\right| \;<\; \max_{y\in[-1,1]} |w(y)| \;=\; \max_{y\in[-1,1]} \left|\prod_{j=0}^{n} (y-y_j)\right| .$$

Wir leiten nun mit Hilfe von Satz 3.8 (für „$n+1$" statt „n", da Satz 3.9 Aussagen über T_{n+1} macht) einen Widerspruch her.
Beachte dazu: Die Extremalstellen sind absteigend nummeriert und ihre Funktionswerte haben gleichen Betrag aber alternierendes Vorzeichen (vgl. dazu Satz 10.6). Aus der Annahme folgt für die Funktionswerte $q(y_j^{(e)})$ in den Extremalstellen von T_{n+1} (vgl. (3.21)–(3.23) für $n+1$ statt n)

$$\begin{aligned}
q\left(y_0^{(e)}\right) &\le \max_{y\in[-1,1]} |q(y)| < \max_{y\in[-1,1]} \left|\prod_{j=0}^{n} (y-y_j)\right| = 2^{-n} = w\left(y_0^{(e)}\right) ,\\
q\left(y_1^{(e)}\right) &\ge -\max_{y\in[-1,1]} |q(y)| > -\max_{y\in[-1,1]} \left|\prod_{j=0}^{n} (y-y_j)\right| = -2^{-n} = w\left(y_1^{(e)}\right) ,\\
q\left(y_2^{(e)}\right) &< 2^{-n} = w\left(y_2^{(e)}\right)\\
&\vdots
\end{aligned}$$

allgemein also

$$q\left(y_j^{(e)}\right) \quad \begin{cases} < w\left(y_j^{(e)}\right) , & \text{falls } j \text{ gerade} \\ > w\left(y_j^{(e)}\right) , & \text{falls } j \text{ ungerade}, \quad j = 0,1,\ldots,n+1 . \end{cases}$$

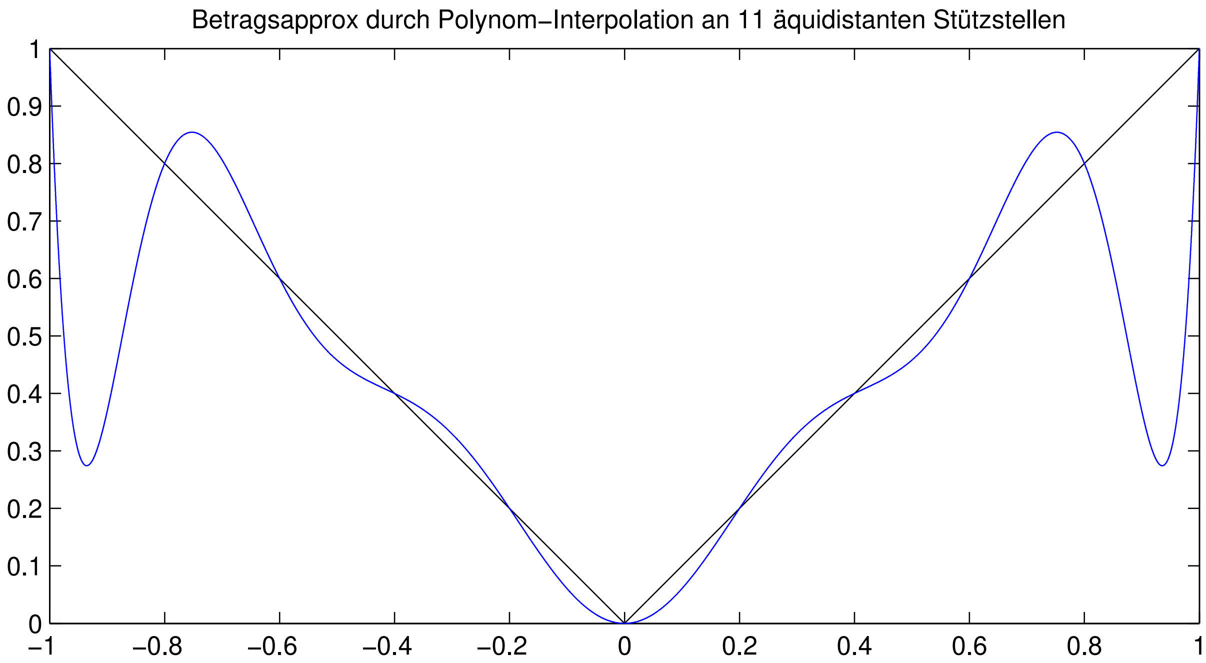
Hieraus folgt für das Differenzpolynom $p(y) := w(y) - q(y)$

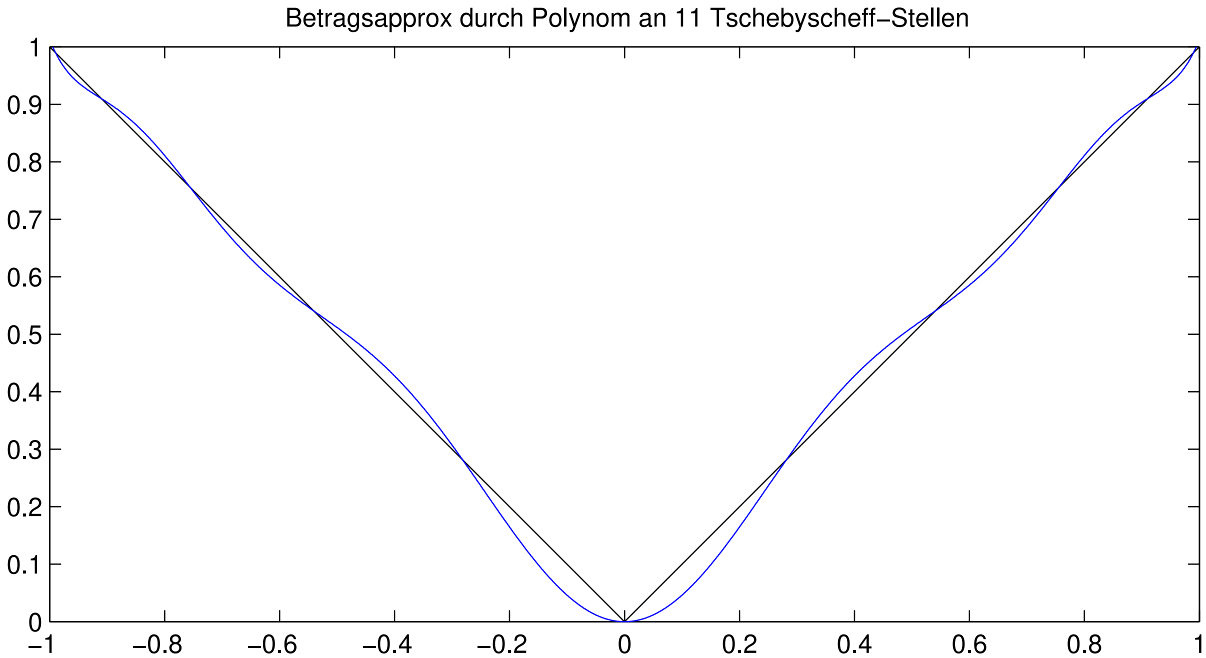
$$p\left(y_j^{(e)}\right) = w\left(y_j^{(e)}\right) - q\left(y_j^{(e)}\right) \quad \begin{cases} > 0, & \text{falls } j \text{ gerade} \\ < 0, & \text{falls } j \text{ ungerade}, \quad j = 0,1,\ldots,n+1 . \end{cases}$$

$p(y)$ hat also in $[-1,1]$ $n+1$ Vorzeichenwechsel, also $n+1$ Nullstellen. Nun ist aber $p \in \Pi_n$, denn sowohl bei $w(y)$ als auch bei $q(y)$ hat y^{n+1} den Koeffizienten 1, also verschwindet y^{n+1} bei der Differenzbildung. Damit folgt aus Lemma 2.5, daß $p \equiv 0$, also $w(y) = q(y)$. Dies ist aber ein Widerspruch zur Annahme. ■

Nachbemerkung: Wesentlich für den Beweis ist die Tatsache, daß alle Extremalwerte von T_{n+1} den **gleichen** Betrag und alternierendes Vorzeichen haben (vgl. dazu Satz 10.6).

Um die Auswirkung der Wahl der Tschebyscheff-Nullstellen auf die Approximationsgenaigkeit des Interpolationsverfahrens zu demonstrieren, greifen wir nochmals das Beispiel der Betragsapproximation auf und stellen zum Vergleich die Ergebnisse der Interpolation mit äquidistanten- und Tschbebyscheff-Nullstellen nebeneinander.


Betragsapprox durch Polynom−Interpolation an 11 äquidistanten Stützstellen
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1


Betragsapprox durch Polynom an 11 Tschebyscheff−Stellen
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

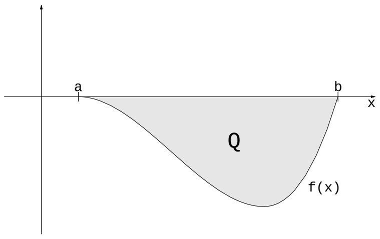
§ 4 Numerische Integration

Im Rahmen dieser Veranstaltung beschränken wir uns auf die numerische Integration (auch numerische Quadratur genannt) von reellwertigen Funktionen f einer reellen Variablen. Berechnet werden soll

$$I(f) \quad = \quad \int\limits_a^b f(x)\,dx\,.$$

Warum numerisch?

Beispiel 1)
Um die Wassermenge zu bestimmen, die ein Fluß pro Zeiteinheit transportiert, ist es nötig, den Flußquerschnitt $Q = \int\limits_a^b -f(x)\,dx$ zu bestimmen. Dazu wird die Flußtiefe $f(x)$ an mehreren Stellen gemessen. Da sie damit nur an einzelnen Punkten bekannt ist, kann das Integral nicht mit Hilfe einer Stammfunktion integriert werden.



Beispiel 2)
Bei vielen Integrationsaufgaben in der Mathematik (z.B. beim Lösen von Differentialgleichungen oder der Längenbestimmung von Kurven) ist die zu integrierende Funktion zwar bekannt, aber nicht geschlossen integrierbar, d.h. die Stammfunktion kann nicht explizit angegeben werden. Einfache Beispiele hierfür sind etwa $f(x) = e^{-x^2}$, $f(x) = \sin x^2$, $f(x) = \frac{\sin x}{\sqrt{x}}$. Auch hier muß man numerisch vorgehen.

Interpolatorische Quadratur
Die naheliegende Vorgehensweise besteht darin, $f(x)$ durch ein Interpolationspolynom $p_n(x)$ zu ersetzen und $\int\limits_a^b p_n(x)\,dx$ statt $\int\limits_a^b f(x)\,dx$ zu berechnen. Ist die Funktion f gutmütig, z.B. $(n+1)$ mal stetig differenzierbar, so existiert eine Fehlerabschätzung für den Interpolationsfehler (vgl. Satz 3.7) und durch Integration über den Interpolationsfehler kann man dann sogar eine Fehlerabschätzung für den Integrationsfehler erhalten. Diese Ideen wollen wir zunächst verfolgen. Ist also $f \in C^{n+1}[a,b]$, so gilt (vgl. Satz 3.7)

$$\begin{aligned} \left| \int\limits_a^b f(x)\,dx - \int\limits_a^b p_n(x)\,dx \right| \quad &= \quad \left| \int\limits_a^b \left(f(x) - p_n(x)\right)\,dx \right| \quad \leq \\ \leq \quad \int\limits_a^b |f(x) - p_n(x)|\,dx \quad &\leq \quad \frac{1}{(n+1)!} \max_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right| \int\limits_a^b \left| \prod_{j=0}^n (x - x_j) \right| dx\,. \end{aligned} \tag{4.1}$$

Die Integration der Ungleichung (3.14) ist erlaubt, da das Integral eine monotone Funktion ist (Analysis!).

Wir bezeichnen mit

$$I_n(f) \quad = \quad \int\limits_a^b p_n(x)\,dx$$

eine Integrationsformel mit einem Interpolationspolynom n-ter Ordnung für das exakte Integral

$$I(f) \quad = \quad \int\limits_a^b f(x)\,dx.$$

Mit dieser Bezeichnung besagt (4.1):

$I_n(f)$ integriert Polynome bis zum Höchstgrad n exakt.

Benutzt man die Lagrange–Form des Interpolationspolynoms (vgl. Satz 3.1)

$$p_n(x) \;=\; \sum_{j=0}^{n} \ell_j(x)\, f(x_j)\,, \qquad \ell_j(x) \;=\; \prod_{\substack{\nu=0\\ \nu\neq j}}^{n} \frac{(x-x_\nu)}{(x_j-x_\nu)}\,,$$

so ist durch Integration unmittelbar einsichtig, daß Integrationsformeln die folgende Gestalt haben

$$I_n(f) \quad = \quad \int\limits_a^b p_n(x)\,dx \quad = \quad \sum_{j=0}^{n} A_j\, f(x_j)$$

$$\text{mit von den } x_j \text{ abhängigen Gewichten} \tag{4.2}$$

$$A_j \quad = \quad \int\limits_a^b \ell_j(x)\,dx\,.$$

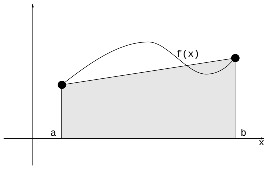
Wählt man die Interpolationsknoten x_j äquidistant, so heißen die Integrationsformeln (4.2) **Newton–Cotes–Formeln**.

Am gebräuchlichsten sind die Formeln für $n = 1$ (Trapezregel) und für $n = 2$ (Simpsonregel), die wir zunächst herleiten.

n = 1:

$$x_0 \;=\; a, \quad x_1 \;=\; b, \quad p_1(x) \;=\; \frac{x-b}{a-b}\,f(a) \;+\; \frac{x-a}{b-a}\,f(b)\,,$$

$$A_0 \;=\; \int\limits_a^b \frac{x-b}{a-b}\,dx \;=\; \frac{b-a}{2}\,, \qquad A_1 \;=\; \int\limits_a^b \frac{x-a}{b-a}\,dx \;=\; \frac{b-a}{2}\,,$$



also

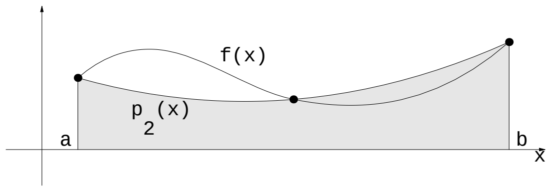
$$\boxed{I_1(f) \;=\; \frac{b-a}{2}\,\big(f(a)+f(b)\big) \qquad \textbf{Trapezregel}} \tag{4.3}$$

n = 2:

$$x_0 \;=\; a, \quad x_1 \;=\; \frac{a+b}{2}, \quad x_2 \;= b\,, \quad A_0 \;=\; \int\limits_a^b \ell_0(x)\,dx \;=\; \frac{b-a}{6}\,,$$

$$A_1 \;=\; \int\limits_a^b \ell_1(x)\,dx \;=\; \frac{2}{3}\,(b-a)\,,$$

$$A_2 \;=\; \int\limits_a^b \ell_2(x)\,dx \;=\; \frac{b-a}{6}\,,$$



und damit

$$\boxed{I_2(f) = \int\limits_a^b p_2(x)\,dx \;=\; \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) \quad \textbf{Simpsonregel}} \qquad (4.4)$$

Schon die beiden Zeichnungen lassen vermuten, daß die angegebenen Formeln (insbesondere bei längeren Intervallen) für praktische Zwecke zu ungenau sind.

Die nächstliegende Idee wäre, unter Benutzung von mehr Stützstellen, Interpolationspolynome höheren Grades zu verwenden. Dies ist nicht empfehlenswert, da sich praktisch zeigt, und auch theoretisch untermauert werden kann, daß die Näherungen $\int\limits_a^b p_n(x)\,dx$ für $\int\limits_a^b f(x)\,dx$ mit wachsendem n nicht in wünschenswertem Maß besser werden.

Erfolgreicher ist die Idee, das Intervall $[a, b]$ in Teilintervalle zu zerlegen und auf jedes Teilintervall die Trapez– oder Simpsonregel anzuwenden.

Zerlegt man $[a, b]$ in n gleich große Teilintervalle durch

$$a \;=\; x_0 < x_1 < \ldots < x_{n-1} < x_n \;=\; b\,,$$

so liest man aus (4.3) sofort ab die

Zusammengesetzte Trapezregel

$$\begin{aligned} T(h)\,(f) \;&=\; \frac{b-a}{2n}\left[f(x_0) \,+\, 2\sum_{j=1}^{n-1} f(x_j) \,+\, f(x_n)\right], \\ &=\; \frac{h}{2}\left[f(x_0) \,+\, 2\sum_{j=1}^{n-1} f(x_0 + j\,h) \,+\, f(x_n)\right] \\ &\text{mit} \qquad h \;=\; \frac{b-a}{n} \;=\; x_j - x_{j-1},\;\; j = 1, \ldots, n \end{aligned} \qquad (4.5)$$

Der Stützstellenabstand h heißt auch *Maschenweite*.

Bei der zusammengesetzten Simpsonregel müssen die Intervallmittelpunkte der Teilintervalle mitnumeriert werden, also

$$a \quad = \quad x_0 \, < \, x_1 \, < \, x_2 \, < \, \ldots \, < \, x_n \quad = \quad b, \qquad \boldsymbol{n} \ \textbf{gerade}\,.$$

Die Teilintervalle für die Simpsonregel sind also $[x_{2j}, x_{2j+2}]$, $j = 0, \ldots, \frac{n}{2} - 1$, und die Maschenweite $h = \frac{b-a}{n} = x_j - x_{j-1}$, $j = 1, 2, \ldots, n$. Damit folgt aus (4.4) die

Zusammengesetzte Simpsonformel

$$\begin{aligned} S(h)\,(f) \;&=\; \frac{b-a}{3n} \left[f(x_0) \;+\; 2 \sum_{j=1}^{\frac{n}{2}-1} f(x_{2j}) \;+\; 4 \sum_{j=0}^{\frac{n}{2}-1} f(x_{2j+1}) \;+\; f(x_n) \right] \\ &=\; \frac{h}{3} \left[f(a) \;+\; 2 \sum_{j=1}^{\frac{n}{2}-1} f(a+2jh) \;+\; 4 \sum_{j=0}^{\frac{n}{2}-1} f(a+(2j+1)h) \;+\; f(b) \right], \\ &\text{mit} \qquad h \;=\; \frac{b-a}{n} \;=\; x_j - x_{j-1}, \quad j = 1, 2, \ldots, n, \quad n \ \text{gerade}\,. \end{aligned} \tag{4.6}$$

Quadraturfehler

Wir erhalten Abschätzungen für den Quadraturfehler durch Integration über den Interpolationsfehler (vgl. (4.1)), falls f die nötigen Differenzierbarkeitseigenschaften besitzt.

Im Fall $n = 1$ (Trapezregel) gilt

$$\int_a^b \left| \prod_{j=0}^{1} (x - x_j) \right| dx \;=\; \int_a^b |x-a| \,\cdot\, |x-b| \, dx \;=\; \int_a^b (x-a)\,(b-x)\, dx \;=\; \frac{(b-a)^3}{6}\,.$$

Damit folgt aus (4.1) für den

Quadraturfehler der Trapezregel

$$|I(f) - I_1(f)| \;\leq\; \frac{(b-a)^3}{12} \max_{\xi \in [a,b]} |f''(\xi)| \tag{4.7}$$

Bemerkung

Auf Grund ihrer Herleitung integriert die Trapezregel Funktionen f, die Polynome vom Grad ≤ 1 sind, exakt (lineare Interpolation). Dies spiegelt sich auch in der Fehlerabschätzung wieder. Für Polynome 1. Grades ist $f'' = 0$, der Quadraturfehler also $= 0$.

Entsprechend wird man erwarten, daß die Simpsonregel Polynome vom Grad ≤ 2 exakt integriert. In der Tat integriert sie jedoch sogar Polynome vom 3. Grad exakt, **sofern**

man die Stützstellen äquidistant wählt. Dies beruht auf folgendem

Lemma 4.1
Sei p_2 das Interpolationspolynom für f mit den Stützstellen $x_0 = a, x_1 = \frac{a+b}{2}, x_2 = b$, und p_3 das Interpolationspolynom für f mit den Stützstellen $x_0 = a, x_1 = \frac{a+b}{2}, x_2 = b$ und $z \in (a,b),\ z \neq x_j,\ j = 0,1,2$. Dann gilt

$$\int_a^b p_2(x)\,dx \quad = \quad \int_a^b p_3(x)\,dx$$

Beweis:
Für das Interpolationspolynom p_3 in der Newton–Form gilt (vgl. (3.7)) mit $z = x_3$, $c_3 = f[x_0, x_1, x_2, x_3]$

$$p_3(x) \quad = \quad p_2(x) \ + \ f[x_0, x_1, x_2, x_3] \prod_{j=0}^{2} (x - x_j)\,, \quad \text{also}$$

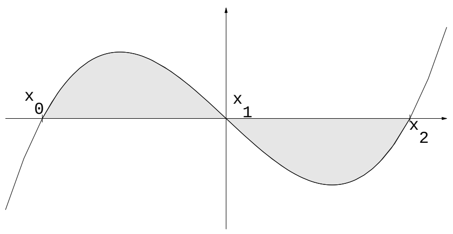
$$\int_a^b p_3(x)\,dx \quad = \quad \int_a^b p_2(x)\,dx \ + \ f[x_0, \ldots, x_3] \int_a^b \prod_{j=0}^{2} (x - x_j)\,dx\,.$$

Nun ist aber

$$\int_a^b \prod_{j=0}^{2} (x - x_j)\,dx \quad = \quad \int_a^b (x-a)\left(x - \frac{a+b}{2}\right)(x-b)\,dx \quad = \quad 0\,,$$

wie man durch Ausrechnen feststellt, womit die Behauptung gezeigt ist. ∎

Dieses Ergebnis ist auch geometrisch einzusehen, denn auf Grund der Äquidistanz der Nullstellen ist das Polynom $\prod\limits_{j=0}^{2} (x - x_j)$ punktsymmetrisch zu x_1; die beiden schraffierten Flächen sind bis auf das Vorzeichen also gleich.



Dieses Ergebnis hat zur Folge, daß man für die Abschätzung des Quadraturfehlers der Simpsonregel den Interpolationsfehler für Interpolationspolynome 3. Grades benutzen kann, also

$$\left| \int_a^b f(x)\,dx - \int_a^b p_2(x)\,dx \right| \ = \ \left| \int_a^b f(x)\,dx - \int_a^b p_3(x)\,dx \right| \ \leq \ \int_a^b |f(x) - p_3(x)|\,dx$$

$$\leq \ \frac{1}{4!} \max_{\xi \in [a,b]} \left| f^{(4)}(\xi) \right| \int_a^b \left| \prod_{j=0}^{3} (x - x_j) \right| dx\,, \qquad (x_3 := z)\,,$$

also

$$\left| \int_a^b f(x)\,dx - \int_a^b p_2(x)\,dx \right| \le \frac{1}{4!} \max_{\xi\in[a,b]} \left|f^{(4)}(\xi)\right| \int_a^b \left| \prod_{j=0}^{3} (x - x_j) \right| dx.$$

Diese Abschätzung gilt zunächst für alle $z \neq x_j$, $j = 0, 1, 2$. Die linke Seite dieser Ungleichung ist unabhängig von $x_3 = z$ und die rechte Seite ist eine stetige Funktion von x_3. Man kann also den Grenzübergang $x_3 \to x_1$ durchführen, ohne die Gültigkeit der Gleichung zu verletzen und erhält

$$\lim_{x_3\to x_1} \int_a^b \left| \prod_{j=0}^{3} (x - x_j) \right| dx = \int_a^b (x-a)(x - \frac{a+b}{2})^2(b-x)\,dx = \frac{(b-a)^5}{120}.$$

Damit erhalten wir den

Quadraturfehler der Simpsonregel

$$|I(f) - I_2(f)| \;\le\; \frac{(b-a)^5}{2880} \max_{\xi\in[a,b]} \left|f^{(4)}(\xi)\right| \tag{4.8}$$

Zusammenfassend erhalten wir also folgenden

Satz 4.2

a) Ist $f \in C^2[a,b]$, so gilt für

$$I_1(f) \;=\; \int_a^b p_1(x)\,dx \;=\; \frac{b-a}{2}\,[f(a) + f(b)]\,, \quad \textit{Trapezregel}, \tag{4.3}$$

die Fehlerabschätzung

$$\left| \int_a^b f(x)\,dx - I_1(f) \right| \;\le\; \frac{(b-a)^3}{12} \max_{\xi\in[a,b]} \left|f^{(2)}(\xi)\right| \tag{4.7}$$

b) Ist $f \in C^4[a,b]$, so gilt für

$$I_2(f) \;=\; \int_a^b p_2(x)\,dx \;=\; \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right], \quad \textit{Simpson–Regel}, \tag{4.4}$$

mit äquidistanten Stützstellen die Fehlerabschätzung

$$\left| \int_a^b f(x)\,dx - I_2(f) \right| \;\le\; \frac{(b-a)^5}{2880} \max_{\xi\in[a,b]} \left|f^{(4)}(\xi)\right|\,. \tag{4.8}$$

Ausgehend von diesem Satz ist es einfach, auch Abschätzungen für die Quadraturfehler der zusammengesetzten Formeln abzuleiten.

Bei der *zusammengesetzten Trapezformel* (4.5) ist (4.7) auf jedes Teilintervall $[x_j, x_{j+1}]$ anzuwenden. Bezeichnet I_1^j die Trapezregel in $[x_j, x_{j+1}]$, so folgt

$$\begin{aligned}
\left| \int_{x_0}^{x_n} f(x)\,dx - T(h)(f) \right| &= \left| \sum_{j=0}^{n-1} \left(\int_{x_j}^{x_{j+1}} f(x)\,dx - I_1^j(f) \right) \right| \leq \sum_{j=0}^{n-1} \left| \int_{x_j}^{x_{j+1}} f(x)\,dx - I_1^j(f) \right| \\
&\leq \sum_{j=0}^{n-1} \frac{(x_{j+1}-x_j)^3}{12} \max_{\xi \in [x_j, x_{j+1}]} \left| f''(\xi) \right| \\
&\leq \frac{n}{12} \cdot \left(\frac{b-a}{n} \right)^3 \max_{\xi \in [a,b]} \left| f''(\xi) \right| \\
&= \frac{(b-a)^3}{12 n^2} \max_{\xi \in [a,b]} \left| f''(\xi) \right| , \qquad \text{und mit} \quad h = \frac{b-a}{n} \\
&= \frac{h^2}{12} (b-a) \max_{\xi \in [a,b]} \left| f''(\xi) \right| . \qquad (4.9)
\end{aligned}$$

Analog erhält man für die *zusammengesetzte Simpsonregel* (4.6) durch Anwendung von (4.8) auf die Teilintervalle $[x_{2j}, x_{2j+2}]$, $j = 0, 1, \ldots, \frac{n}{2} - 1$, wobei I_2^{2j} die Simpsonregel im Intervall $[x_{2j}, x_{2j+2}]$ bezeichnet:

$$\begin{aligned}
\left| \int_{x_0}^{x_n} f(x)\,dx - S(h)(f) \right| &= \left| \sum_{j=0}^{\frac{n}{2}-1} \left(\int_{x_{2j}}^{x_{2j+2}} f(x)\,dx - I_2^{2j}(f) \right) \right| \\
&\leq \sum_{j=0}^{\frac{n}{2}-1} \left| \int_{x_{2j}}^{x_{2j+2}} f(x)\,dx - I_2^{2j}(f) \right| \\
&\leq \sum_{j=0}^{\frac{n}{2}-1} \frac{1}{2880} \left(\frac{2(b-a)}{n} \right)^5 \max_{\xi \in [x_{2j}, x_{2j+2}]} \left| f^{(4)}(\xi) \right| \\
&\leq \frac{n}{2} \cdot \frac{2^5}{2880} \left(\frac{b-a}{n} \right)^5 \max_{\xi \in [x_0, x_n]} \left| f^{(4)}(\xi) \right| \\
&= \left(\frac{b-a}{n} \right)^4 \frac{(b-a)}{180} \max_{\xi \in [x_0, x_n]} \left| f^{(4)}(\xi) \right| , \quad \text{und mit } h = \frac{b-a}{n} \\
&= \frac{h^4}{180} (b-a) \max_{\xi \in [a,b]} \left| f^{(4)}(\xi) \right| . \qquad (4.10)
\end{aligned}$$

Zusammenfassend gilt also

Satz 4.3

a) Ist $f \in C^2[a,b]$, so gilt für die *zusammengesetzte Trapezregel*

$$T(h)\,(f) \;=\; \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(a+jh) + f(b)\right], \qquad h \;=\; \frac{b-a}{n} \tag{4.5}$$

die Fehlerabschätzung

$$\left|\int_a^b f(x)\,dx - T(h)\,(f)\right| \;\leq\; \frac{h^2}{12}\,(b-a)\,\max_{\xi\in[a,b]} |f''(\xi)|\;. \tag{4.9}$$

b) Ist $f \in C^4[a,b]$ so gilt für die *zusammengesetzte Simpsonregel*

$$S(h)\,(f) = \tfrac{h}{3}\left[f(a) + 2\sum_{j=1}^{\frac{n}{2}-1} f(a+2jh) + 4\sum_{j=0}^{\frac{n}{2}-1} f(a+(2j+1)h) + f(b)\right], \qquad h \;=\; \frac{b-a}{n} \tag{4.6}$$

die Fehlerabschätzung

$$\left|\int_a^b f(x)\,dx - S(h)\,(f)\right| \;\leq\; \frac{h^4}{180}\,(b-a)\,\max_{\xi\in[a,b]} \left|f^{(4)}(\xi)\right|\;. \tag{4.10}$$

Bemerkung:

Die Potenz von h in den Fehlerabschätzungen nennt man die *Ordnung der Verfahren.*

Konvergenz der Integrationsformeln

Bei der reinen Interpolationsquadratur ist es schwierig Konvergenzaussagen zu machen, da z.B die hohen Ableitungen $f^{n+1}(\xi)$ entweder nicht existieren oder nur sehr schlecht auswertbar sind. Im Fall der zusammengesetzten Integrationsformeln ist die Konvergenzfrage sehr einfach zu klären. Man benötigt nur Ableitungen niedriger Ordnung und kann beliebige kleine Fehler garantieren. Aus Satz 4.3 folgt sofort die

Folgerung 4.4

Unter den Voraussetzungen von Satz 4.3 konvergieren sowohl die zusammengesetzte Trapezregel als auch die zusammengesetzte Simpsonformel für $h = \frac{b-a}{n} \to 0$, d.h.

$$\lim_{h\to 0} \left| \int_a^b f(x)\,dx \;-\; T(h)\,(f) \right| \;=\; 0, \quad \text{falls} \quad f \in C^2[a,b],$$

$$\lim_{h\to 0} \left| \int_a^b f(x)\,dx \;-\; S(h)\,(f) \right| \;=\; 0, \quad \text{falls} \quad f \in C^4[a,b].$$

Der Beweis ist unmittelbar aus (4.9) bzw. (4.10) abzulesen, da alle Größen der Fehlerabschätzungen bis auf h Konstanten sind, die unabhängig von h sind. ■

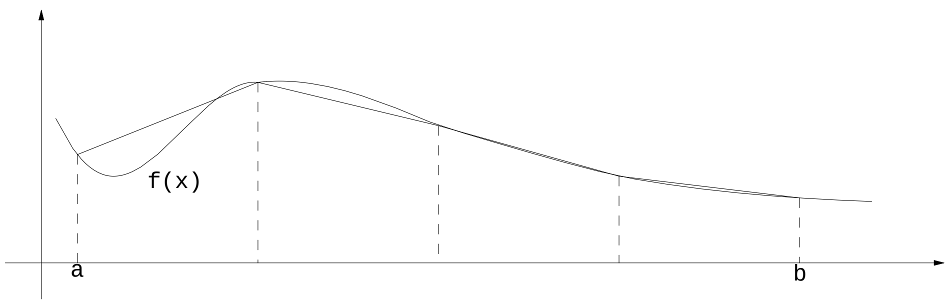
Bemerkung:

Die Fehlerabschätzungen bedeuten anschaulich: Wird die Schrittweite halbiert (n verdoppelt), so sinkt bei der zusammengesetzten Trapezregel die Fehlerschranke um den Faktor $\left(\frac{1}{2}\right)^2$, bei der zusammengesetzten Simpsonformel sogar um den Faktor $\left(\frac{1}{2}\right)^4$.

Beachte: Der absolute Fehler hängt natürlich von den Größen $|f''(\xi)|$ bzw. $\left|f^{(4)}(\xi)\right|$ in den einzelnen Teilintervallen ab. Wenn diese Größen in $[a,b]$ nicht allzusehr variieren, gilt obige Überlegung auch „in etwa“ für den absoluten Fehler.

Adaptive Quadraturformeln

Fehlerschranken bieten die Möglichkeit, vor Beginn der Rechnung abzuschätzen, wie groß die Schrittweite h gewählt werden muß, damit der Fehler eine vorgegebene Schranke nicht übersteigt, vorausgesetzt natürlich, man kennt Schranken für die betreffenden Ableitungen. Dies ist oft nicht der Fall oder mit großen Mühen verbunden. Außerdem sind Formeln mit äquidistanten Stützstellen oft mit zu großem Rechenaufwand verbunden, weil sie auch in Regionen, in denen es nicht nötig ist, die Schrittweite verkleinern, wie man schon an folgendem einfachen Beispiel für die zusammengesetzte Trapezregel erkennt.



Es ist deshalb sinnvoll, nach Verfahren zu suchen, welche die Schrittweite selbst an die Eigenschaften der Funktion f anpassen (adaptieren), also kleine Schrittweiten wählen, wenn sich die Funktion (oder das Integral) stark ändern, große, wenn dies nicht der Fall ist.

ZIEL: Entwickle ein Verfahren mit Schrittweitenanpassung, das für eine vorgegebene Funktion $\int_a^b f(x)\,dx$ bis auf einen vorgegebenen Fehler $\varepsilon > 0$ genau berechnet.

Ausgangspunkt der Überlegungen ist 1), daß man weiß, wie „in etwa" sich der Fehler bei Intervallhalbierungen verhält (vgl. die obige Bemerkung) und 2), daß die numerische Differenz der Integrationswerte bei Intervallhalbierung ja numerisch anfällt. Beide Fakten kann man benutzen, um den wirklichen Integrationsfehler zu **schätzen** und daraus eine **Schätzung** für eine geeignete Schrittweite abzuleiten. Wir wollen dies (in einer Rohform – Verfeinerungen sind möglich –) für die zusammengesetzte Simpsonregel demonstrieren.

Sei also L die (vorläufig noch unbekannte) Zahl von Teilintervallen $I_j = [x_{j-1}, x_j]$ von $[a, b]$ der (vorläufig noch unbekannten) Länge $h_j = x_j - x_{j-1}$. Wir bezeichnen den exakten Integralwert im Intervall $[x_{j-1}, x_j]$ durch

$$I^j(f) \quad = \quad \int_{x_{j-1}}^{x_j} f(x)\,dx\,,$$

die Simpsonformel mit der Maschenweite $\frac{h_j}{2}$ in I^j durch (vgl. (4.4))

$$S^j(f) \quad = \quad \frac{h_j}{6}\left[f(x_{j-1}) \;+\; 4f\left(x_{j-1} + \frac{h_j}{2}\right) \;+\; f(x_j)\right],$$

und die Simpsonformel mit halbierter Maschenweite $\frac{h_j}{4}$ durch

$$Q^j(f) = \frac{1}{6}\,\frac{h_j}{2}\left[f(x_{j-1}) + 4f\left(x_{j-1} + \frac{h_j}{4}\right) + 2f\left(x_{j-1} + \frac{h_j}{2}\right) + 4f\left(x_{j-1} + \frac{3h_j}{4}\right) + f(x_j)\right].$$

Aus der Fehlerabschätzung (4.8) folgt die Existenz einer Konstanten c_j, so daß gilt

$$\left|I^j(f) - S^j(f)\right| \quad = \quad c_j\, h_j^5 \tag{4.11}$$

Halbiert man das Intervall $[x_{j-1}, x_j]$ und wendet die Fehlerabschätzung auf beide Hälften an, so existieren Konstanten c_{j_1} und c_{j_2}, so daß

$$\left|I^j(f) - Q^j(f)\right| \quad = \quad (c_{j_1} + c_{j_2})\left(\frac{h_j}{2}\right)^5. \tag{4.12}$$

Wir nehmen nun an, daß sich c_{j_1} und c_{j_2} nicht wesentlich von c_j unterscheiden. Dies ist um so eher erfüllt, je weniger sich $f^{(4)}$ ändert, d.h. u.a. je kleiner die Teilintervalle werden. Setzt man also $c_{j_1} \approx c_{j_2} \approx c_j$, so geht (4.12) über in

$$\left|I^j(f) - Q^j(f)\right| \quad \approx \quad 2c_j\left(\frac{h_j}{2}\right)^5 \qquad \left(= \frac{1}{16}\,c_j\,h_j^5\right). \tag{4.13}$$

Dies ermöglicht einen Vergleich mit (4.11):

$$\left|I^j(f) - Q^j(f)\right| \;\approx\; \frac{1}{16}\left|I^j(f) - S^j(f)\right| \;\leq\; \frac{1}{16}\left(\left|I^j(f) - Q^j(f)\right| + \left|Q^j(f) - S^j(f)\right|\right),$$

bzw.

$$\left|I^j(f) - Q^j(f)\right| \;\lesssim\; \frac{1}{15}\left|Q^j(f) - S^j(f)\right| .$$

Hieraus folgt für den Gesamtfehler

$$\begin{aligned} \left|I(f) - \sum_{j=1}^{L} Q^j(f)\right| &= \left|\sum_{j=1}^{L} \left(I^j(f) - Q^j(f)\right)\right| \\ &\leq \sum_{j=1}^{L} \left|I^j(f) - Q^j(f)\right| \\ &\lesssim \tfrac{1}{15} \sum_{j=1}^{L} \left|Q^j(f) - S^j(f)\right| . \end{aligned} \tag{4.14}$$

Die Werte $|Q^j(f) - S^j(f)|$ sind nach der Intervallhalbierung und der numerischen Integration ja bekannt. Genügt nun h_j der **Forderung**

$$\left|Q^j(f) - S^j(f)\right| \;\leq\; \frac{15 h_j \cdot \varepsilon}{b-a}, \tag{4.15}$$

so folgt wegen $\sum_{j=1}^{L} h_j = b - a$ aus (4.15) und (4.14)

$$\left|I(f) - \sum_{j=1}^{L} Q^j(f)\right| \;\lesssim\; \varepsilon . \tag{4.16}$$

Umsetzung in ein Verfahren

START: $x_0 := a$, $x_1 := b$, berechne S^1, Q^1 und prüfe (4.15) für $h_1 = (b-a)$. Ist (4.15) erfüllt $\longrightarrow$ Ende, andernfalls

$$x_0 = a, \quad x_1 = \frac{a+b}{2}, \quad x_2 := b, \quad I_1 = \left[a, \frac{a+b}{2}\right], \quad I_2 = \left[\frac{a+b}{2}, b\right], \quad h_1 = h_2 = \frac{b-a}{2}.$$

Prüfe (4.15) für beide Teilintervalle.

Für das (die) Teilintervall(e), in dem (denen) (4.15) nicht erfüllt ist, wird die Schrittweite halbiert, usw.

Ist (4.15) in allen Teilintervallen erfüllt, gilt (4.16).

Daß das beschriebene (sehr einfache) Verfahren ein Schätzverfahren ist (wie alle raffinierteren adaptiven Verfahren auch), belegt folgendes einfache

Beispiel:

Berechne für $f(x) = \cos 4x$ das Integral $\int\limits_0^{2\pi} \cos 4x \, dx$. Nun ist $\cos 4x = 1$ für $x = j\,\frac{\pi}{2}$, $j = 0, 1, 2, 3, 4$. Also gilt $S^1(f) = 2\pi$, $Q^1(f) = 2\pi$, das Verfahren bricht ab, weil (4.15) erfüllt ist, und liefert den Wert 2π statt 0.

Natürlich passiert dieser Zusammenbruch nicht, wenn man gleich mit einer höheren Zahl von Stützstellen beginnt, doch ist immer Vorsicht geboten.

Von anderen möglichen und wichtigen Integrationsmethoden wollen wir hier nur die

Gauß–Quadraturformeln

erwähnen. Sie sind wie die bisherigen Formeln von der Gestalt

$$G(f) \quad = \quad \sum_{j=0}^{n} A_j \, f(x_j) \, .$$

Verlangt man als Güteforderung, daß durch solche Formeln Polynome möglichst hohen Grades exakt integriert werden, so liegt es nahe zu versuchen, sowohl die Stützstellen als auch die Gewichte A_j gemäß dieser Güteforderung zu bestimmen. Bei vorgegebenem n hat man dann $2n+2$ Unbekannte A_j, x_j. Setzt man zu ihrer Bestimmung die Forderung an

$$\sum_{j=0}^{n} A_j \, (x_j)^k \quad = \quad \int\limits_a^b x^k \, dx \quad = \quad \frac{1}{k+1} \left(b^{k+1} - a^{k+1} \right) , \qquad k = 0, 1, 2, \ldots, 2n+1$$

(d.h. $2n+2$ Gleichungen für $2n+2$ Unbekannte), so besteht die Hoffnung, Polynome bis zum Grad $(2n+1)$ exakt integrieren zu können, falls dieses (nicht lineare) Gleichungssystem eine Lösung besitzt.

Es ist in der Tat möglich, solche Quadraturformeln zu konstruieren, allerdings mit einem aufwendigeren mathematischen Apparat, als er uns bisher zur Verfügung steht (Hermite Interpolation, orthogonale Polynome u.a.). In manchen Taschenrechnern sind solche Gaußformeln implementiert. Beispielsweise erhält man für $n = 1$ (also 2 Stützstellen) die Fehlerabschätzung

$$I(f) \, - \, G_1(f) \quad = \quad \frac{(b-a)^5}{4320} \, f^{(4)}(\xi) \, , \qquad \xi \, \in \, [a, b] \, ,$$

welche ca. um den Faktor $1,5$ kleiner ist als der Fehler der Simpsonformel, die 3 Stützstellen benötigt.

Vom Rechenaufwand her (Zahl der nötigen Funktionsauswertungen) sind diese Formeln konkurrenzlos günstig. Nachteilig ist:

1) Für jedes n müssen die Stützstellen (Gauß–Knoten) neu berechnet werden.

2) Teilergebnisse für den Fall n können für den Fall $n+1$ nicht weiterverwendet werden.

3) Es ist i.a. unmöglich, von vorneherein festzustellen, welches n die gewünschte Genauigkeit garantiert.

Man muß also n schrittweise erhöhen und hat als Abbruchkriterium nur aufzuhören, wenn 2 aufeinanderfolgende Näherungen im Rahmen der geforderten Genauigkeit übereinstimmen.

Deshalb gehen mit wachsendem n die theoretischen Vorteile der Gauß–Integration bald verloren.

§ 5 Lineare Gleichungssysteme

Wir beginnen mit einem einfachen

Beispiel: Produktionsmodell von Leontief.

In der Volkswirtschaftslehre muß man typischerweise eine ganze Reihe von Objekten untersuchen, die in wechselseitigen Abhängigkeiten stehen. Um das Prinzip deutlich zu machen, mit dem eine Behandlung dieser Abhängigkeit versucht werden kann, werden wir die Situation stark vereinfachen und die Wirtschaft nur in die 3 Sektoren

- Landwirtschaft
- produzierendes Gewerbe
- Transportwesen

aufteilen. Eine stärkere Differenzierung bringt zwar stärker realitätsbezogene Ergebnisse, macht aber unsere mathematische Diskussion zunächst unübersichtlicher.

Wir nehmen folgende Abhängigkeiten an:

a) Bei der Produktion landwirtschaftlicher Güter (Lebensmittel) im Werte von 1000,00 DM werden landwirtschaftliche Güter im Wert von 300,00 DM (Getreide, ...), Transportleistungen im Wert von 100,00 DM und industrielle Güter im Werte von 200,00 DM (Landmaschinen, ...) gebraucht.

b) Die Herstellung von Industrieprodukten im Wert von 1000,00 DM benötigt Lebensmittel im Wert von 200,00 DM, andere Industrieprodukte im Wert von 400,00 DM und Transportleistungen im Wert von 100,00 DM.

c) Zur Erbringung von Transportleistungen im Wert von 1000,00 DM sind Lebensmittel im Wert von 100,00 DM, industriell gefertigte Güter (Fahrzeuge, Treibstoff,...) im Wert von 200,00 DM und Transportleistungen (z.B. Treibstoffversorgung) im Wert von 100,00 DM erforderlich.

Welche Mengen müssen die einzelnen Sektoren produzieren, damit die Gesamtwirtschaft folgende Überschüsse erwirtschaftet

Landwirtschaft	20.000	DM
Industrie	40.000	DM
Transportwesen	0	DM

x_L gebe an, wieviele Mengeneinheiten im Wert von je 1000,00 DM die Landwirtschaft produziert. x_I und x_T seien die entsprechenden Werte für Industrie und Transportwesen. Die gesuchten Mengen müssen also folgende Bedingungen erfüllen:

$$\begin{aligned} 0.7\,x_L - 0.2\,x_I - 0.1\,x_T &= 20 \\ -0.2\,x_L + 0.6\,x_I - 0.1\,x_T &= 40 \\ -0.1\,x_L - 0.2\,x_I + 0.9\,x_T &= 0 \end{aligned}$$

Diese Aufgabenstellung verlangt also — etwas allgemeiner formuliert — das Lösen eines Systems linearer Gleichungen

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \ldots & + & a_{1n}\,x_n & = & b_1 \\ a_{21}\,x_1 & + & a_{22}\,x_2 & + & \ldots & + & a_{2n}\,x_n & = & b_2 \\ \vdots & & & & & & & & \vdots \\ a_{m1}\,x_1 & + & a_{m2}\,x_2 & + & \ldots & + & a_{mn}\,x_n & = & b_m \end{array} \tag{5.1}$$

mit gegebenen Zahlen a_{ij} und b_i und gesuchten Größen x_j, $i = 1, \ldots, m$, $j = 1, \ldots, n$. (m Gleichungen, n Unbekannte).

In der Linearen Algebra wird untersucht, wann ein solches System lösbar ist und welche Struktur (Vektorraum, lineare Mannigfaltigkeit,...) die Menge der Lösungen hat.

Wir wollen im folgenden untersuchen, wie man im Spezialfall, daß das System genau eine Lösung hat, diese möglichst schnell und genau berechnen kann.

Wir werden sehen (haben schon gesehen §3,(3.2)), daß Lösen von Systemen der Art (5.1) auch als Teilprobleme zu anderen mathematischen Problemen auftreten.

Wir nehmen jetzt also an, daß das System (5.1) ebenso viele Gleichungen wie Unbekannte hat $(m = n)$, und führen eine schematische Darstellung ein, die sowohl der besseren Übersicht als auch der bequemeren Verarbeitung im Computer dient. Wir fassen die Koeffizienten a_{jk}, die rechten Seiten b_j und die Unbekannten x_k zu rechteckigen Zahlenfeldern — Matrizen genannt — zusammen

$$\boldsymbol{A} := \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}, \qquad \boldsymbol{b} := \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}, \qquad \boldsymbol{x} := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

und schreiben

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \tag{5.1'}$$

oder auch kurz

$$\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b} \tag{5.1''}$$

$\boldsymbol{A}$ heißt $(n \times n)$–Matrix (n Zeilen, n Spalten), die $(n \times 1)$–Matrizen $\boldsymbol{b}$ und $\boldsymbol{x}$ heißen (Spalten)–Vektoren.

$$\text{Bezeichnung:} \quad \boldsymbol{A} \in \mathbb{R}^{n \times n}, \qquad \boldsymbol{b}, \boldsymbol{x} \in \mathbb{R}^{n \times 1} = \mathbb{R}^n$$

Diese Bezeichnung deutet an, daß die Einträge der Matrizen reelle Zahlen sind. Natürlich ist auch z.B. $A \in \mathbb{C}^{n \times n}$ möglich. Der 1. obere Index bezeichnet immer die Zeilenzahl, der 2.te die Spaltenzahl.

Den Ausdruck $\boldsymbol{A}\,\boldsymbol{x}$ kann man als Matrixprodukt verstehen, bzw. als Produkt Matrix · Vektor, wenn $\boldsymbol{x} \in \mathbb{R}^n$. Das Bildungsgesetz für das Produkt ergibt sich aus dem Vergleich von (5.1) (für $n = m$) und (5.1'). (vgl. dazu auch Fischer, §2.4)

Das Gaußsche Eliminationsverfahren (GEV)

Es macht keine Schwierigkeiten, ein lineares Gleichungssystem zu lösen, wenn es folgende Gestalt hat.

$$\begin{array}{ccccccccc} a_{11}\,x_1 & + & a_{12}\,x_2 & + & \dots & + & a_{1n}\,x_n & = & b_1 \\ & & a_{22}\,x_2 & + & \dots & + & a_{2n}\,x_n & = & b_2 \\ & & \ddots & & & & \vdots & & \\ & & & & \ddots & & \vdots & & \\ & & & & & & a_{nn}\,x_n & = & b_n \end{array} \tag{5.2}$$

Man sagt dann, die Matrix $\boldsymbol{A} = (a_{ij})_{i,j=1,\dots,n}$ hat *obere Dreiecksgestalt*: d.h. $a_{ij} = 0$ für alle $i > j$.

$$\boldsymbol{A} \quad := \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} & & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & & \dots & a_{2n} \\ 0 & 0 & \ddots & & & \vdots \\ \vdots & & \ddots & a_{ii} & & \vdots \\ & & & \ddots & \ddots & \\ 0 & \dots & \dots & \dots & 0 & a_{nn} \end{pmatrix}, \qquad \text{obere Dreiecksmatrix}\,. \tag{5.3}$$

Wenn alle *Diagonalelemente* $a_{ii} \neq 0,\ i = 1,\dots,n$ sind, erhält man die eindeutige Lösung von (5.2) durch **Rückwärtseinsetzen**:

$$\begin{array}{lclcl} x_n & = & b_n/a_{nn} & & \\ x_{n-1} & = & (b_{n-1} & - & a_{n-1,n}\,x_n)/a_{n-1,n-1} \\ \vdots & & \vdots & & \vdots \\ x_2 & = & (b_2 & - & a_{2,n}\,x_n \;-\; \dots \;-\; a_{2,3}\,x_3)/a_{22} \\ x_1 & = & (b_1 & - & a_{1,n}\,x_n \;-\; \dots \;-\; a_{1,3}\,x_3 - a_{1,2}\,x_2)/a_{11} \end{array} \tag{5.4}$$

oder in der Summenschreibweise

$$x_i \;=\; \left(b_i \;-\; \sum_{j=i+1}^{n} a_{ij}\,x_j\right) \Big/ \, a_{ii}\,, \qquad i = n,\ n-1,\dots,1\,. \tag{5.4'}$$

$\left(\text{Beachte, daß } \sum\limits_{j=h}^{\ell} \dots \;=\; 0, \text{ falls } \ell < h\,.\right)$

Erfreulicherweise lassen sich alle eindeutig lösbaren linearen Gleichungssysteme auf die Form (5.2) bringen. Dies gelingt mit Hilfe der folgenden elementaren Umformungen, welche die Lösungsmenge des Systems, das sind alle Vektoren $\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, die dem System (5.1) für $m = n$ genügen, nicht ändern:

1) Man kann zu einer Gleichung (d.h. einer Zeile von $\boldsymbol{A}$ und der zugehörigen Komponente von $\boldsymbol{b}$, vgl. (5.1')) ein Vielfaches einer anderen Gleichung addieren.

2) Die Lösungsmenge ändert sich nicht, wenn man die Reihenfolge der Gleichungen vertauscht.

Wir benutzen diese beiden Eigenschaften, um das Gleichungssystem $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ so umzuformen, daß alle Elemente a_{ij} unter der Hauptdiagonalen (also $i > j$) „zu Null gemacht" (eliminiert) werden.

Wir bezeichnen das System in der Ausgangsform mit der vollbesetzten Matrix $\boldsymbol{A}$ mit

$$\boldsymbol{A}^{(0)}\,\boldsymbol{x} \quad = \quad \boldsymbol{b}^{(0)}\,.$$

Im 1. Schritt subtrahieren wir für $i = 2, \ldots, n$ von der i–ten Gleichung das $a_{i1}^{(0)} \big/ a_{11}^{(0)}$–fache der 1. Gleichung und nennen das dadurch entstehende Gleichungssystem

$$\boldsymbol{A}^{(1)}\,\boldsymbol{x} \quad = \quad \boldsymbol{b}^{(1)}\,.$$

Das Bildungsgesetz lautet also für $i = 2, \ldots, n$:

$$\begin{aligned} \ell_{i1} \quad &:= \quad a_{i1}^{(0)} \big/ a_{11}^{(0)}\,; \\ a_{ij}^{(1)} \quad &:= \quad a_{ij}^{(0)} \;-\; \ell_{i1} \cdot a_{1j}^{(0)}\,, \qquad j = 1, \ldots, n \\ b_i^{(1)} \quad &:= \quad b_i^{(0)} \;-\; \ell_{i1} \cdot b_1^{(0)}\,, \end{aligned} \tag{5.5}$$

die 1. Zeile bleibt ungeändert.

Bemerkung:
Die Elemente $a_{i1}^{(1)}$ für $i > 1$ werden, falls sie später noch gebraucht werden sollten, nicht programmiert (gerechnet), sondern gesetzt, $a_{i1}^{(1)} = 0$, zur Vermeidung von unnötigen Rundungsfehlern.

Nach dem 1. Schritt hat das System also die Gestalt $\boldsymbol{A}^{(1)}\,\boldsymbol{x} = \boldsymbol{b}^{(1)}$ mit

$$\boldsymbol{A}^{(1)} \;=\; \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \ldots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ \vdots & a_{32}^{(1)} & \ldots & a_{3n}^{(1)} \\ \vdots & \vdots & & \\ 0 & a_{n2}^{(1)} & & a_{nn}^{(1)} \end{pmatrix}, \qquad \boldsymbol{b}^{(1)} \;=\; \begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ \vdots \\ \vdots \\ b_n^{(1)} \end{pmatrix}, \tag{5.6}$$

Im 2. Schritt subtrahieren wir für $i = 3, \ldots, n$ von der (i)ten Gleichung das $a_{i2}^{(1)} \,\big/\, a_{22}^{(1)}$ fache der 2. Gleichung, d.h. wir annulieren nun alle Elemente unter $a_{22}^{(1)}$. Das entstehende System nennen wir

$$\boldsymbol{A}^{(2)}\,\boldsymbol{x} \quad = \quad \boldsymbol{b}^{(2)}\,.$$

Bildungsgesetz: Für $i = 3, \ldots, n$ sei

$$\begin{array}{rrcl} & \ell_{i2} & := & a_{i2}^{(1)} \,/\, a_{22}^{(1)}\,, \\ & a_{ij}^{(2)} & := & a_{ij}^{(1)} \,-\, \ell_{i2}\, a_{2j}^{(1)}\,, \quad j = 3, \ldots, n \\ \text{setze} & a_{i2}^{(2)} & := & 0 \quad \text{für} \quad i \,>\, 2\,, \\ & b_i^{(2)} & := & b_i^{(1)} \,-\, \ell_{i2}\, b_2^{(1)}\,, \\ \text{und natürlich:} & & & \text{die ersten beiden Zeilen bleiben ungeändert.} \end{array} \tag{5.7}$$

Dann hat $\boldsymbol{A}^{(2)}$ die Gestalt

$$\boldsymbol{A}^{(2)} \quad = \quad \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \ldots & \ldots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & \ldots & \ldots & a_{2n}^{(1)} \\ \vdots & 0 & a_{33}^{(2)} & \ldots & a_{3n}^{(2)} \\ \vdots & \vdots & a_{43}^{(2)} & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \ldots & a_{nn}^{(2)} \end{pmatrix} \tag{5.8}$$

Dieses Verfahren führt man fort. Im k–ten Schritt subtrahiert man für $i = k+1, \ldots, n$ von der i–ten Gleichung das $a_{ik}^{(k-1)} \,/\, a_{kk}^{(k-1)}$–fache der k–ten Gleichung, um die Elemente unter $a_{kk}^{(k-1)}$ zu annulieren, wir haben also das

Bildungsgesetz: Für $i = k+1, \ldots, n$ sei

$$\begin{array}{rrcl} \textbf{k–ter Schritt} & \ell_{ik} & := & a_{ik}^{(k-1)} \,/\, a_{kk}^{(k-1)}\,, \\ & a_{ij}^{(k)} & := & a_{ij}^{(k-1)} \,-\, \ell_{ik}\, a_{kj}^{(k-1)}\,, \quad j = k, \ldots, n \\ & b_i^{(k)} & := & b_i^{(k-1)} \,-\, \ell_{ik}\, b_k^{(k-1)}\,, \\ & & & \text{die Zeilen } 1, \ldots, k \text{ bleiben unverändert.} \end{array} \tag{5.9}$$

Nach $(n-1)$ Schritten erhalten wir schließlich $\boldsymbol{A}^{(n-1)}\,\boldsymbol{x} = \boldsymbol{b}^{(n-1)}$ mit der oberen Dreiecksmatrix

$$\boldsymbol{A}^{(n-1)} = \begin{pmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \dots & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & a_{nn}^{(n-1)} \end{pmatrix} \tag{5.10}$$

Das System $\boldsymbol{A}^{(n-1)}\,\boldsymbol{x} = \boldsymbol{b}^{(n-1)}$ hat die gleiche Lösungsmenge wie das Ausgangssystem. $\boldsymbol{A}^{(n-1)}\,\boldsymbol{x} = \boldsymbol{b}^{(n-1)}$ kann durch Rückwärtseinsetzen gelöst werden.

Mit Hilfe des Matrixkalküls kann man den k–ten Schritt des GEV wie folgt beschreiben:

$$\boldsymbol{A}^{(k)} = \boldsymbol{L}_{k-1}\,\boldsymbol{A}^{(k-1)} \tag{5.11}$$

mit der Matrix

$$\boldsymbol{L}_{k-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -\ell_{k+1,k} & \ddots & \\ & & \vdots & & \ddots \\ & & -\ell_{n,k} & & & 1 \end{pmatrix} \tag{5.12}$$

mit ℓ_{ik} gemäß (5.9), also

$$\ell_{ik} = a_{ik}^{(k-1)} \,/\, a_{kk}^{(k-1)} \quad \text{für} \quad i = k+1, \dots, n$$

Man erhält somit insgesamt

$$\boldsymbol{A}^{(n-1)} = \boldsymbol{L}_{n-2}\,\boldsymbol{A}^{(n-2)} = \boldsymbol{L}_{n-2}\,\boldsymbol{L}_{n-3}\,\dots\,\boldsymbol{L}_0\,\boldsymbol{A}^{(0)}\,. \tag{5.13}$$

Es ist einfach nachzurechnen, daß das Produkt $\Lambda_{k-1}\,L_{k-1} = I$ (Einheitsmatrix) ist mit

$$\boldsymbol{\Lambda}_{k-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{k+1,k} & \ddots & \\ & & \vdots & & \ddots \\ & & \ell_{n,k} & & & 1 \end{pmatrix} \tag{5.14}$$

Etwas mühsamer ist es festzustellen, daß sich das Produkt

$$\boldsymbol{L} = \boldsymbol{\Lambda}_0\,\boldsymbol{\Lambda}_1\,\dots\,\boldsymbol{\Lambda}_{n-2} = \begin{pmatrix} 1 & & & & \\ \ell_{21} & 1 & & & \\ \vdots & \ell_{32} & \ddots & & \\ \ell_{n-1,1} & \vdots & \ddots & 1 & \\ \ell_{n,1} & \ell_{n,2} & \dots & \ell_{n,n-1} & 1 \end{pmatrix} \tag{5.15}$$

in der angegebenen einfachen Form als linke Dreiecksmatrix ausrechnen läßt.

Multiplizieren wir die Gleichung (5.13) der Reihe nach von links mit $\boldsymbol{\Lambda}_{n-2}, \boldsymbol{\Lambda}_{n-3}, \ldots, \boldsymbol{\Lambda}_0$, so erhalten wir wegen $\boldsymbol{\Lambda}_k \boldsymbol{L}_k = I$ und (5.15)

$$\boldsymbol{L}\,\boldsymbol{A}^{(n-1)} \quad = \quad \boldsymbol{A}^{(0)} \quad = \quad \boldsymbol{A}\,. \tag{5.16}$$

Die schließlich ausgerechnete Matrix $\boldsymbol{R} = \boldsymbol{A}^{(n-1)}$ ist eine (rechte) obere Dreiecksmatrix (vgl. (5.10)). D.h. wir haben mit dem GEV die Ausgangsmatrix $\boldsymbol{A}$ in ein Produkt

$$\boldsymbol{A} \quad = \quad \boldsymbol{L} \cdot \boldsymbol{R} \tag{5.16'}$$

einer linken mit einer rechten Dreiecksmatrix zerlegt. Die Darstellung (5.16') heißt daher auch **LR–Zerlegung von $\boldsymbol{A}$**.

Damit läßt sich das Gaußverfahren prinzipiell in 3 Lösungsschritte aufteilen

$$\left\{\begin{array}{llll} 1. & \boldsymbol{A} = \boldsymbol{L}\,\boldsymbol{R} & & (\text{Zerlegung von } \boldsymbol{A}) \\ 2. & \boldsymbol{L}\,\boldsymbol{c} = \boldsymbol{b} & & (\text{Vorwärtseinsetzen} \;\Rightarrow\; \boldsymbol{c} = \boldsymbol{L}^{-1}\boldsymbol{b} = \boldsymbol{b}^{(n-1)} = \boldsymbol{R}\boldsymbol{x}) \\ 3. & \boldsymbol{R}\boldsymbol{x} = \boldsymbol{c} & & (\text{Rückwärtseinsetzen} \;\Rightarrow\; \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}) \end{array}\right. \tag{5.17}$$

Bemerkungen

1) Im k–ten Schritt ändern sich die ersten k Zeilen von $\boldsymbol{A}^{(k-1)}$ und $\boldsymbol{b}^{(k-1)}$ nicht mehr.

2) Sollte im k–ten Schritt $a_{kk}^{(k-1)} = 0$ sein (das muß abgeprüft werden), so vertauschen wir die k–te Gleichung mit einer späteren s–ten Gleichung $(s > k)$, für welche $a_{sk}^{(k-1)} \neq 0$ gilt. Gibt es kein solches s, so sind die ersten k Spalten von $\boldsymbol{A}$ (als Vektoren aufgefaßt) linear abhängig und in der Linearen Algebra wird gezeigt, daß das Gleichungssystem dann nicht eindeutig lösbar ist.

3) Natürlich lassen sich mit dem GEV auch $(m \times n)$–Gleichungssysteme mit $n \geq m$ behandeln (vgl. Lineare Algebra).

4) Soll für eine quadratische Matrix $\boldsymbol{A}$ die Inverse $\boldsymbol{X}$ berechnet werden, $\boldsymbol{A}\,\boldsymbol{X} = \boldsymbol{I}$, so erhält man die Spalten $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n$ von $\boldsymbol{X}$ durch Lösung der Systeme

$$\boldsymbol{A}\,\boldsymbol{x}^k \;=\; \boldsymbol{e}^k, \qquad k = 1, \ldots, n, \qquad \boldsymbol{e}^k \;=\; k\text{–ter Einheitsvektor}\,.$$

Zu ihrer Lösung wird man **nur einmal** $\boldsymbol{L}$ und $\boldsymbol{R}$ gemäß (5.17) berechnen und danach $\boldsymbol{L}\,\boldsymbol{c}^k = \boldsymbol{e}^k$ und $\boldsymbol{R}\,\boldsymbol{x}^k = \boldsymbol{c}^k$ lösen.

Beispiel zu 2):

Das System $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ ist ohne Zeilenvertauschung nicht lösbar.

Numerische Schwierigkeiten, Pivot–Suche

Die Bemerkung 2) s.o. ist oft nur von theoretischem Wert, da der Rechner auf Grund von Zahldarstellungsschwierigkeiten und Rundungsfehlern das Ergebnis Null einer Rechnung nur selten als Null erkennt, sondern z.B. $a_{kk}^{(k-1)} = 3.5 \cdot 10^{-25}$ erhält. Wird mit diesem Wert weiter gerechnet, so kann nur Unsinn herauskommen.

Wir zeigen dies an einem **1. Beispiel**, das wir der Einfachheit halber mit 6–stelliger Arithmetik rechnen (genauer: jeder Rechenschritt wird zunächst mit höherer Genauigkeit ausgeführt, dann wird auf sechs Dezimalstellen gerundet; dabei wird die Berechnung von $a + b \cdot c$ als ein Rechenschritt angesehen):

$$\boldsymbol{A}^{(0)} = \begin{pmatrix} 11 & 44 & 1 \\ 0.1 & 0.4 & 3 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \boldsymbol{b}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

1. Eliminationsschritt

$$\boldsymbol{A}^{(1)} = \begin{pmatrix} 11 & 44 & 1 \\ 0 & -4{\cdot}10^{-8} & 2.99091 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \boldsymbol{b}^{(1)} = \begin{pmatrix} 1 \\ 0.990909 \\ 1 \end{pmatrix}$$

2. Eliminationsschritt

$$\boldsymbol{A}^{(2)} = \begin{pmatrix} 11 & 44 & 1 \\ 0 & -4{\cdot}10^{-8} & 2.99091 \\ 0 & 0 & 7.47727 \cdot 10^{7} \end{pmatrix}, \quad \boldsymbol{b}^{(2)} = \begin{pmatrix} 1 \\ 0.990909 \\ 2.47727 \cdot 10^{7} \end{pmatrix}$$

Lösung: $\mathbf{x}^T = (-41.8765,\ 10.4843,\ 0.331307)$

Zum Vergleich die exakte Lösung

$$\boldsymbol{x}^T = (-5.26444,\ 1.33131,\ 0.331307)$$

GRUND:

Im 1. Eliminationsschritt ist $a_{22}^{(1)} = 0.4 - \frac{0.1}{11} \cdot 44$. Nun ist $\frac{0.1}{11}$ keine Maschinenzahl, weshalb der Rechner $a_{22}^{(1)} = 0$ nicht erhält. Das Ergebnis ist katastrophal.

Selbst wenn $a_{kk}^{(k-1)}$ tatsächlich $\neq 0$ ist, aber sehr klein im Vergleich zu anderen Matrixeinträgen, können sich erhebliche Schwierigkeiten ergeben, wie folgendes **2. Beispiel** zeigt (wieder mit 6–stelliger Arithmetik):

$$A^{(0)} = \begin{pmatrix} 0.001 & 1 & 1 \\ -1 & 0.004 & 0.004 \\ -1000 & 0.004 & 0.000004 \end{pmatrix}, \quad b^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

1. Eliminationsschritt

$$A^{(1)} = \begin{pmatrix} 0.001 & 1 & 1 \\ 0 & 1000 & 1000 \\ 0 & 1\cdot 10^6 & 1\cdot 10^6 \end{pmatrix}, \qquad b^{(1)} = \begin{pmatrix} 1 \\ 1001 \\ 1\cdot 10^6 \end{pmatrix}.$$

2. Eliminationsschritt

$$A^{(2)} = \begin{pmatrix} 0.001 & 1 & 1 \\ 0 & 1000 & 1000 \\ 0 & 0 & 0 \end{pmatrix}, \qquad b^{(2)} = \begin{pmatrix} 1 \\ 1001 \\ -1000 \end{pmatrix}.$$

Gleichungssystem unlösbar.

GRUND:
Das Unglück passiert schon im 1. Eliminationsschritt, nach welchem die Zeilen 2 und 3 von $A^{(1)}$ linear abhängig sind. Der Rechner erhält

$$\begin{aligned} a_{22}^{(1)} &= a_{23}^{(1)} = 0.004 - \tfrac{(-1)}{0.001} = 0.004 + 1000 && \approx 1000 \\ a_{32}^{(1)} &= 0.004 - \tfrac{(-1000)}{0.001} = 0.004 + 10^6 && \approx 10^6 \\ a_{33}^{(1)} &= 0.000004 + 10^6 && \approx 10^6 \end{aligned}$$

Er kann im Rahmen seiner Genauigkeit (6 Dezimalen) die Zahlen $10^3 + 0.004$ und 10^3 bzw. $10^6 + 0.004$ und 10^6 usw. bei der Differenz– bzw. Summenbildung nicht mehr unterscheiden. Dies liegt an der absoluten Größe der Faktoren

$$\begin{aligned} \ell_{21} &= a_{21}^{(0)} \,/\, a_{11}^{(0)} = -10^3 \\ \ell_{31} &= a_{31}^{(0)} \,/\, a_{11}^{(0)} = -10^6 \end{aligned}$$

Folge: Die Einträge der 2. und 3. Zeile „gehen unter" in $A^{(1)}$.

Die folgenden beiden Regeln sollen helfen, das Auftreten dieser Phänomene einzuschränken.

1. einfache Pivotsuche (Spaltenpivotsuche)
Vor dem k–ten Eliminationsschritt suche man die betragsgrößte der Zahlen $a_{ik}^{(k-1)}$, $i \geq k$, also etwa $a_{i_0 k}^{(k-1)}$, und vertausche dann in $A^{(k-1)}$ die i_0–te Zeile mit der k–ten Zeile und entsprechend $b_{i_0}^{(k-1)}$ mit $b_k^{(k-1)}$. $a_{i_0 k}^{(k-1)}$ heißt dann Pivotelement. (Man macht also durch geeignete Vertauschung der Gleichungen die Faktoren $\ell_{ik} = a_{ik}^{(k-1)} \,/\, a_{kk}^{(k-1)}$ betragsmäßig möglichst klein.)

2. totale Pivotsuche

Vor dem k–ten Eliminationsschritt suche man die betragsgrößte der Zahlen $a_{ij}^{(k-1)}$, $i \geq k$, $j \geq k$, diese sei $a_{rs}^{(k-1)}$. Dann vertausche man die k–te Zeile mit der r–ten Zeile — natürlich auch $b_k^{(k-1)}$ und $b_r^{(k-1)}$ — und die k–te Spalte mit der s–ten Spalte. $\left(a_{rs}^{(k-1)} \text{ heißt Pivotelement}\right)$.

Dadurch bekommen die Faktoren ℓ_{ik} betragsmäßig die kleinst möglichen Werte. Dadurch wird gesichert, daß bei der Differenzbildung (vgl. 5.9))

$$a_{ij}^{(k)} \;:=\; a_{ij}^{(k-1)} - \ell_{ik}\, a_{kj}^{(k-1)}\,, \qquad j = k, \ldots, n$$

in der neuen i–ten Zeile möglichst viel Information der alten i–ten Zeile erhalten bleibt.

Die totale Pivotsuche gilt als die stabilste Variante des Gaußschen Eliminationsverfahrens (allerdings auch als die aufwendigste, da sehr viele Vergleiche nötig sind bis man das betragsgrößte Element gefunden hat). Zumindest auf die einfache Pivotsuche sollte man **keinesfalls verzichten**.

Schon mit der einfachen Pivotsuche liefern unsere Beispiele jetzt vernünftige Lösungen (wieder mit 6–stelliger Arithmetik).

Beispiel 1:

$A^{(0)}$ und $A^{(1)}$ wie auf S. 50, dann Vertauschung von Zeile 2 und 3 und 2. Eliminationsschritt

$$A^{(2)} \;=\; \begin{pmatrix} 11 & 44 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 2.99091 \end{pmatrix}, \qquad b^{(2)} \;=\; \begin{pmatrix} 1 \\ 1 \\ 0.990909 \end{pmatrix}$$

und die vernünftige Lösung

$$x^T \;=\; (-5.26444,\; 1.33131,\; 0.331037)$$

Beispiel 2:

$A^{(0)}$, $b^{(0)}$ wie auf S. 51, Tausch von 1. und 3. Gleichung und 1. Eliminationsschritt liefern

$$A^{(1)} \;=\; \begin{pmatrix} -1000 & 0.004 & 0.000004 \\ 0 & 0.003996 & 0.004 \\ 0 & 1 & 1 \end{pmatrix}, \qquad b^{(1)} \;=\; \begin{pmatrix} 1 \\ 0.999 \\ 1 \end{pmatrix}$$

Tausch von 2. und 3. Gleichung und 2. Eliminationsschritt

$$A^{(2)} \;=\; \begin{pmatrix} -1000 & 0.004 & 0.000004 \\ 0 & 1 & 1 \\ 0 & 0 & 4\cdot 10^{-6} \end{pmatrix}, \qquad b^{(2)} \;=\; \begin{pmatrix} 1 \\ 1 \\ 0.995004 \end{pmatrix}$$

Lösung: $x^T \;=\; (-0.995005,\; -2.48750 \cdot 10^5,\; 2.48751 \cdot 10^5)$.

Bemerkung zur Programmierung der Pivotsuche

Die Vertauschung von Gleichungen (einfache Pivotsuche) hat keinen Einfluß auf die Lösungsmenge des Gleichungssystems, die Vertauschung von Spalten ebenfalls nicht, sie entspricht nur einer Änderung der Reihenfolge der Komponenten des Lösungsvektors, worüber man allerdings Buch führen muß. Wird das Verfahren programmiert, so muß man die Vertauschung von Zeilen und Spalten, die sehr aufwendig ist, nicht tatsächlich durchführen. Es ist einfacher, sich die Vertauschungen mit Hilfe von Merkvektoren zu merken.

Vor Beginn des Verfahrens definiert man für die Zeilen und Spalten von $\boldsymbol{A}$ Merkvektoren $z[i],\ s[j],\ i,j = 1,\dots,n,$ die zunächst die natürliche Reihenfolge der Zeilen und Spalten speichern, d.h. $z[i] = s[i] = i,\ \ i = 1,\dots,n.$

Die Matrixelemente a_{ij}, die rechten Seiten b_i und die Komponenten x_j des Lösungsvektors werden aufgerufen durch $a[z[i], s[j]]$, $b[z[i]]$ und $x[s[j]]$.

Soll beispielsweise die 3. und 7. Gleichung vertauscht werden, so setzt man mit einer Hilfsgröße h:

$$\begin{aligned} h &:= z[3]\,; \\ z[3] &:= z[7]\,; \\ z[7] &:= h\,; \end{aligned}$$

dann wird die neue 3. Zeile durch $a[z[3], s[k]],\ k = 1,\dots,n,$ die neue rechte Seite durch $b[z[3]]$ aufgerufen. Entsprechend verfährt man beim Spaltentausch.

Es gibt einen Typ von Matrizen, die sog. *positiv definiten Matrizen*, bei denen man auf Zeilen– und Spaltenvertauschung verzichten kann und wobei man eine „sparsamere“ (Zahl der Rechenoperationen) Variante des GEV anwenden kann, das Cholesky–Verfahren. Auf seine Beschreibung verzichten wir an dieser Stelle auf Grund mangelnden mathematischen Handwerkzeugs.

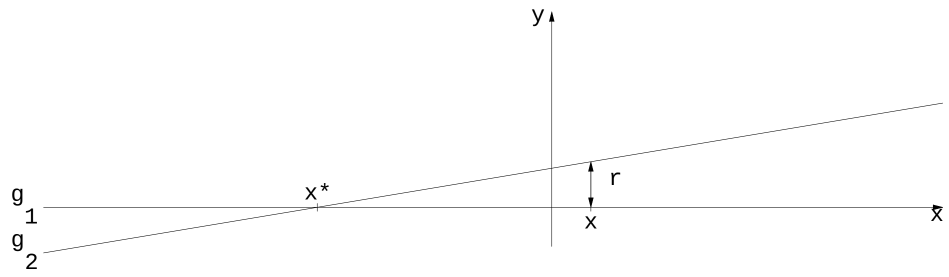
Mit der Pivotsuche sind noch nicht alle auftretenden numerischen Probleme gelöst. Auf 2 Punkte wollen wir noch hinweisen:

a) Ein singuläres Gleichungssystem (d.h. die Zeilen von $\boldsymbol{A}$ sind linear abhängig) wird (jedenfalls theoretisch) vom GEV daran erkannt, daß alle Pivotsuchen erfolglos verlaufen, d.h. kein nicht verschwindendes Pivotelement finden. Durch Rundungsfehler könnte es aber passieren (Beispiel 1)), daß das System trotzdem als lösbar erscheint. Eine hieraus berechnete Lösung hat wenig Sinn. Man behilft sich in der Praxis überlicherweise so, daß man ein Gleichungssystem für nicht behandelbar erklärt, wenn ein Pivotelement betragsmäßig kleiner ist als ein vorgegebenes $\varepsilon > 0$ (z.B. $\varepsilon = 10^{-8}$). Dies Vorgehen erfolgt aber ausschließlich aus praktischen Gründen. Es kann durchaus vorkommen, daß man auf diese Weise ein problemlos lösbares Gleichungssystem für nicht behandelbar hält bzw., wenn man das Pivotelement und damit (bei totaler Pivotsuche) alle restlichen Elemente $= 0$ setzt, man damit numerisch linear abhängige Zeilen findet, die in Wirklichkeit gar nicht linear abhängig sind. Dies kann sich, zum Beispiel bei der späteren numerischen Behandlung von linearen Optimierungsaufgaben, als problematisch erweisen.

b) Auf Grund von Rechenungenauigkeiten wird in der Regel die berechnete Lösung $\hat{\boldsymbol{x}}$ nicht mit der exakten Lösung $\boldsymbol{x}^*$ übereinstimmen, es wird nur $\boldsymbol{A}\,\hat{\boldsymbol{x}} \approx \boldsymbol{b}$ gelten. Man muß sich deshalb fragen, ob die Größe des *Defekts* $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{A}\,\hat{\boldsymbol{x}}$ (auch *Residuum* genannt) eine Aussage über die Größe des unbekannten Fehlers $\boldsymbol{\epsilon} = \boldsymbol{x}^* - \hat{\boldsymbol{x}}$ zuläßt.

Die in b) angeschnittene Frage läßt sich nur bedingt mit ja beantworten. Wie eine Fehleranalyse des GEV (die wir hier und jetzt noch nicht durchführen können) zeigt, kann bei fast linear abhängigen Zeilen das Residuum sehr klein ausfallen, der Fehler $\boldsymbol{\epsilon} = \boldsymbol{x}^* - \hat{\boldsymbol{x}}$ jedoch sehr groß sein.

Man mache sich das geometrisch deutlich im Fall $n = 2$. Dann stellen die beiden Gleichungen 2 Geraden dar, ihre Lösung den Schnittpunkt der Geraden. Sind die Geraden fast linear abhängig, so ist $\boldsymbol{r}$ klein und $\boldsymbol{\epsilon}$ groß (vgl. das Beispiel: $g_1 = x$–Achse)



Wir nennen ein Gleichungssystem *schlecht konditioniert*, wenn kleine Änderungen der Eingabewerte (a_{ij}, b_i) große Änderungen der Lösung zur Folge haben. (Eine präzise Definition liegt im Augenblick noch außerhalb unserer mathematischen Fähigkeiten.)

Man kann zeigen, daß dieses Phänomen bei linearen Fast–Abhängigkeiten auftritt (Aufgabe: Man zeige dies an einem Beispiel für $n = 2$).

Variable rechte Seiten

Gelegentlich muß man dasselbe Gleichungssystem mit verschiedenen rechten Seiten lösen. Z.B. wollen wir für unser Modell der Volkswirtschaft berechnen, welche Mengen die Sektoren jeweils produzieren müssen, um verschiedene auswärtige Nachfragen befriedigen zu können. Man wird dann nicht jedesmal das Gleichungssystem von neuem lösen. Man kann diesen Arbeitsaufwand reduzieren:

- Für die Matrix $\boldsymbol{A}$ wird einmal das Eliminationsverfahren vollständig durchgeführt.
- Im Rahmen einer Laufanweisung, die sich über alle rechten Seiten erstreckt, werden diese gemäß (5.9), letzte Zeile, umgeformt und danach wird, innerhalb derselben Laufanweisung, das Rückwärtseinsetzen für jede Seite durchgeführt. (Natürlich müssen dabei etwaige Zeilen– und Spaltenvertauschungen berücksichtigt werden.)

Formuliert man das im Matrixkalkül, so bedeutet das, daß einmal der 1. Schritt aus (5.17) ausgeführt wird und danach für jede rechte Seite die Schritte 2 und 3. Etwaige Zeilen– und Spaltenvertauschungen müssen durch Permutationsmatrizen berücksichtigt werden (vgl. Literatur).

Bemerkungen:

a) Variable rechte Seiten treten auch auf, wenn man die Inverse einer Matrix berechnen will.

b) Die Berechnung von Determinanten (zumindest für $n > 5$) wird mit Hilfe des GEV ausgeführt. Beachte dazu, daß die Determinanten der Umformungsmatrizen alle $=1$ sind. Die Determinante von $\mathbf{A}$ ist dann gleich dem Produkt der Diagonalelemente von $\mathbb{R}$.

c) Das GEV ist nicht das einzige Verfahren zur Lösung von Gleichungssystemen. Man kann dies (insbesondere bei großen und dünn besetzten Matrizen) auch mit Hilfe iterativer Verfahren tun. Wir kommen darauf zu einem späteren Zeitpunkt zurück.

§ 6 Lineare Optimierung

Um die Aufgabenstellung deutlich zu machen, beginnen wir mit einem (natürlich sehr vereinfachten) Beispiel:

Produktionsplan einer (zugegebenermaßen sehr kleinen) Schuhfabrik. Hergestellt werden sollen Damen– und Herrenschuhe, und zwar jeweils nur ein Modell. Die Produktionsbedingungen ergeben sich aus der folgenden Tabelle.

		Damenschuh	Herrenschuh	verfügbar
Herstellungszeit	$[h]$	20	10	8000
Maschinenbearbeitung	$[h]$	4	5	2000
Lederbedarf	$[dm^2]$	6	15	4500
Reingewinn	$[Euro]$	16	32	

Unter der Annahme, daß keine Absatzschwierigkeiten entstehen, soll berechnet werden, wieviele Damen– und Herrenschuhe hergestellt werden müssen, damit der Gewinn optimal wird, natürlich unter Einhaltung obiger Restriktionen.

Mathematische Formulierung:

$$\text{Sei} \quad \begin{array}{ll} x_1 & \text{die Zahl der produzierten Damenschuhe}\,, \\ x_2 & \text{die Zahl der produzierten Herrenschuhe}\,. \end{array}$$

Dann lauten die Produktionsbedingungen:

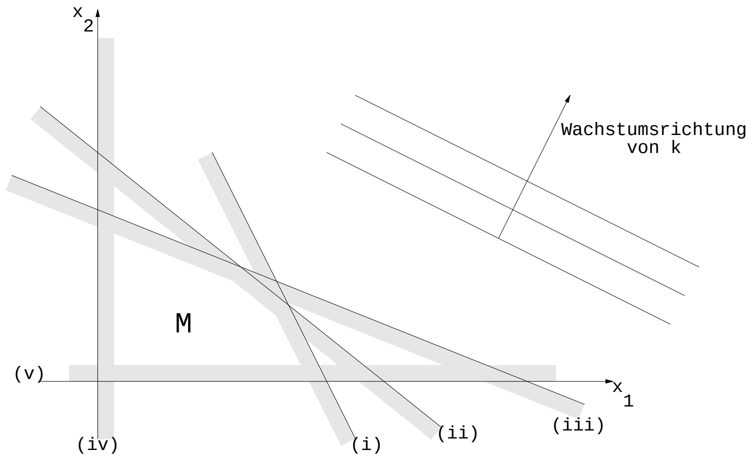
$$\left\{ \begin{array}{rcrcll} 20x_1 & + & 10x_2 & \leq & 8000 & \text{(i)} \\ 4x_1 & + & 5x_2 & \leq & 2000 & \text{(ii)} \\ 6x_1 & + & 15x_2 & \leq & 4500 & \text{(iii)} \\ & & x_1 & \geq & 0 & \text{(iv)} \\ & & x_2 & \geq & 0 & \text{(v)} \end{array} \right. \tag{6.1}$$

Gesucht sind Zahlen (x_1, x_2), die diesen Ungleichungen genügen und den Gewinn maximieren, also

$$\text{Gewinn:} \qquad f(x_1, x_2) \quad = \quad 16x_1 \, + \, 32x_2 \quad \stackrel{!}{=} \quad \max\,. \tag{6.2}$$

Die Funktion f aus (6.2) heißt *Zielfunktion*, die Ungleichungen (6.1) heißen *Nebenbedingungen*. In diesem Beispiel kann man sich die Lösung geometrisch veranschaulichen. Gleichzeitig legt es die Lösungsidee für den allgemeinen Fall nahe.

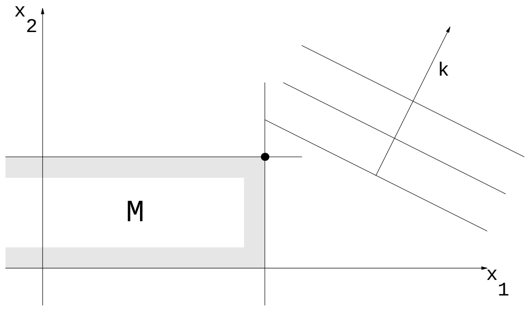
Die Ungleichungen (6.1) beschreiben Halbebenen in $\mathbb{R}^2$, die durch Geraden begrenzt werden (vgl. Zeichnung). Der Durchschnitt M aller dieser Halbebenen beschreibt den *zulässigen Bereich* aller der Punktepaare (x_1, x_2), die den Ungleichungen (6.1) genügen. $f(x_1, x_2) = 16x_1 + 32x_2 = k$, mit variablem k, beschreibt eine Geradenschar. Es ist ein maximales k_m derart zu bestimmen, daß (mindestens) ein Punkt des zulässigen Bereichs M auf der Geraden $f(x_1, x_2) = k_m$ liegt.

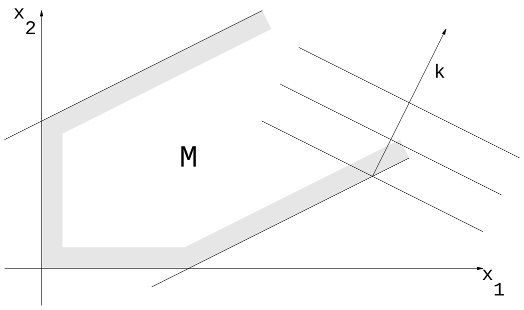


Die geometrische Lösung ist nun einfach. Man verschiebt die Geradenschar so lange im Sinne fallender k, bis zum ersten Mal ein Punkt aus M auf einer Geraden liegt oder, falls die Geradenschar die gleiche Steigung hat wie eine Begrenzungsstrecke von M, sogar eine ganze Strecke. Dieser Punkt oder alle Punkte der genannten Strecke optimieren die Zielfunktion.

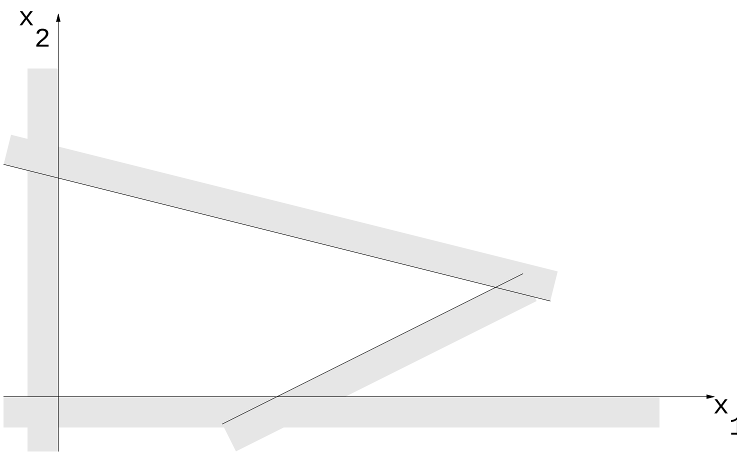
Wir wollen uns an weiteren Beispielen dieser Art klar machen, was je nach Beschaffenheit von M (zulässiger Bereich) und f (Zielfunktion) passieren kann.

1) falls M nicht beschränkt ist, kann f ein Maximum annehmen, es muß dies aber nicht.

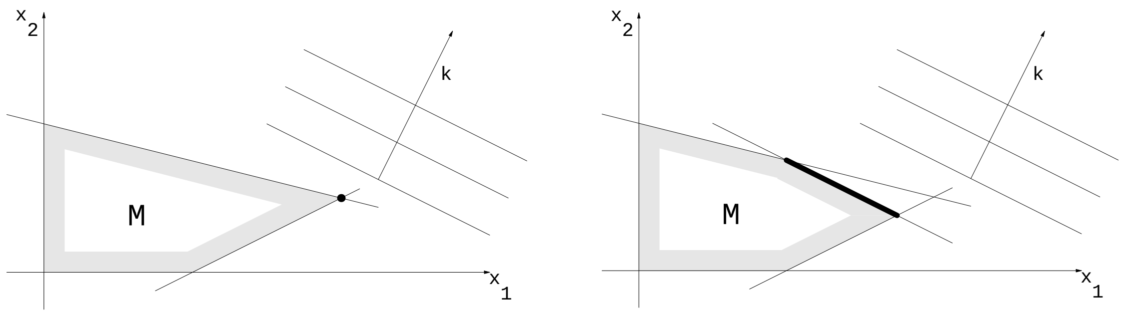




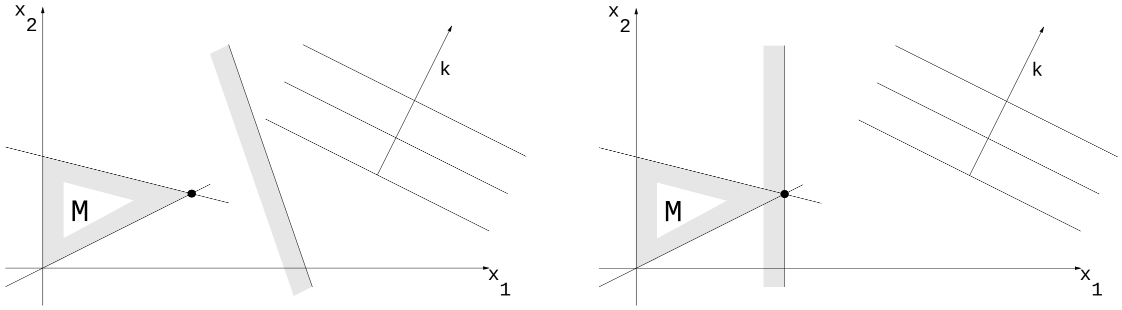
2) $M = \emptyset$ ist möglich.



3) Die optimale Lösung kann, muß aber nicht eindeutig sein.



4) Es gibt überflüssige Nebenbedingungen.



Fazit:
Alle Beispiele zeigen: Wenn eine optimale Lösung existiert, dann wird sie (vielleicht nicht nur, aber auch) in einer Ecke des zulässigen Bereichs angenommen.

Im allgemeinen Fall wird ein lineares Optimierungsproblem die folgende Gestalt haben

$$\begin{array}{rcccccccl}
c_1\,x_1 & + & c_2\,x_2 & + & \dots & + & c_n\,x_n & \overset{!}{=} & \max \\
a_{11}\,x_1 & + & a_{12}\,x_2 & + & \dots & + & a_{1n}\,x_n & \leq & b_1 \\
a_{21}\,x_1 & + & & \dots & & + & a_{2n}\,x_n & \leq & b_2 \\
\vdots & & & & & & & & \\
a_{m1}\,x_1 & + & & & & + & a_{mn}\,x_n & \leq & b_m \\
 & & & & & & x_1 & \geq & 0 \\
 & & & & & & \vdots & & \\
 & & & & & & x_n & \geq & 0
\end{array}$$

Wir schreiben das kurz in Matrixschreibweise. Mit $\boldsymbol{c}^T = (c_1, \dots, c_n) \in \mathbb{R}^n$, $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{m \times n}$, $\boldsymbol{x}^T = (x_1, \dots, x_n) \in \mathbb{R}^n$, $\boldsymbol{b}^T = (b_1, \dots, b_m) \in \mathbb{R}^m$ folgt

$$\left\{ \begin{array}{rcl} \boldsymbol{c}^T \boldsymbol{x} & \overset{!}{=} & \max \\ \boldsymbol{A}\,\boldsymbol{x} & \leq & \boldsymbol{b} \\ \boldsymbol{x} & \geq & \boldsymbol{0} \end{array} \right. \tag{L'}$$

$\boldsymbol{c}^T \boldsymbol{x}$ heißt *Zielfunktion*, $M = \{\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{A}\,\boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$ *zulässiger Bereich.* $\boldsymbol{A}\,\boldsymbol{x} \leq \boldsymbol{b}$ und $\boldsymbol{x} \geq \boldsymbol{0}$ sind im Sinne obiger Ungleichungen komponentenweise zu verstehen. Die Elemente $\in M$ heißen *zulässige Punkte* (zulässige Lösungen) und ein zulässiges $\boldsymbol{x} \in M$ heißt *optimal*, wenn für alle zulässigen Vektoren $\boldsymbol{y} \in M$ gilt $\boldsymbol{c}^T \boldsymbol{x} \geq \boldsymbol{c}^T \boldsymbol{y}$.

Natürlich sind auch andere Formulierungen von (L') möglich und gebräuchlich. Dies hängt von der jeweiligen Aufgabenstellung ab. Wir stellen sie hier kurz zusammen und zeigen, wie sie sich ineinander überführen lassen (um einen möglichst allgemeinen Typ von Optimierungsaufgaben behandeln zu können).

a) Eine Maximierungsaufgabe wird zu einer Minimierungsaufgabe durch Übergang zum Negativen der Zielfunktion.

$$\boldsymbol{c}^T \boldsymbol{x} \;=\; \max \qquad \Longleftrightarrow \qquad -\boldsymbol{c}^T \boldsymbol{x} \;=\; \min$$

b) Eine Ungleichung

$$a_{i1}\,x_1 + \dots + a_{in}\,x_n \;\leq\; b_i$$

kann durch Einführen einer *Schlupfvariablen* $y_i \geq 0$ in eine Gleichung

$$a_{i1}\,x_1 + \dots + a_{in}\,x_n + y_i \;=\; b_i$$

überführt werden.

c) Tritt eine Gleichung als Nebenbedingung auf,

$$a_{j1}\,x_1 + \dots + a_{jn}\,x_n \;=\; b_j\,,$$

so kann sie durch 2 Ungleichungen ersetzt werden:

$$\begin{array}{rcccccl} a_{j1}\,x_1 & + & \dots & + & a_{jn}\,x_n & \leq & b_j \\ -a_{j1}\,x_1 & - & \dots & - & a_{jn}\,x_n & \leq & -b_j \end{array}$$

d) Eine Komponente x_i von $\boldsymbol{x}$, für die keine Vorzeichenbedingung besteht, kann ersetzt werden durch den Ausdruck $x_i = x_{i+} - x_{i-}$ und die Forderung $x_{i+} \geq 0,\ x_{i-} \geq 0$.

Vorsicht: Diese Zerlegung muß auch in der Zielfunktion berücksichtigt werden.

e) Jede „$\leq$“–Ungleichung kann durch Multiplikation mit -1 in eine „$\geq$“–Ungleichung überführt werden (und umgekehrt).

Insbesondere kann also jede Optimierungsaufgabe mit linearer Zielfunktion und linearen Nebenbedingungen in die folgende Form gebracht werden

$$\left\{\begin{array}{rcll} \boldsymbol{c}^T\boldsymbol{x} & \overset{!}{=} & \min & \\ \boldsymbol{A}\,\boldsymbol{x} & = & \boldsymbol{b} & \\ \boldsymbol{x} & \geq & 0\,, & \text{wobei } \boldsymbol{c}\in \mathbb{R}^n, \\ & & & \boldsymbol{A}\in\mathbb{R}^{m\times n},\ \ \boldsymbol{b}\in\mathbb{R}^m \quad \text{gegeben sind}\,. \end{array}\right. \tag{L}$$

Wir verdeutlichen an einem Beispiel die Überführung einer vorgelegten Aufgabe in die Form (L). Man beachte dabei, daß sich bei Einführung von Schlupfvariablen die Zahl der Nebenbedingungen vergrößert und daß insbesondere bei Ersetzung einer nicht vorzeichenbeschränkten Variablen durch 2 Ungleichungen (im Beispiel $x_{1+}\geq 0,\ x_{1-}\geq 0$) sich auch i.a. die Variablenanzahl mit Koeffizienten $\neq 0$ in der Zielfunktion vergrößert.

Beispiel: Die Aufgabe

$$\begin{array}{rcl} -\,2x_1+3x_2 & \overset{!}{=} & \max \\ x_1+\ \ x_2 & \geq & 5 \\ -x_1+\ \ x_2 & \leq & 7 \\ x_1 \qquad & \leq & 10 \\ x_1 & \multicolumn{2}{l}{\text{nicht vorzeichenbeschränkt}} \\ x_2\geq & 0 & \end{array}$$

wird so zu

$$\begin{array}{lrcr} & 2x_{1+}-2x_{1-}-3x_2+0y_1+0y_2+0y_3 & \overset{!}{=} & \min \\ \text{unter} & -x_{1+}+\ x_{1-}+\ x_2+\ y_1 \qquad\qquad & = & -5 \\ & -x_{1+}+\ x_{1-}+\ x_2 \qquad +\ y_2 \qquad & = & +7 \\ & +x_{1+}-\ x_{1-} \qquad\qquad\qquad +\ y_3 & = & +10 \\ & x_{1+} & \geq & 0 \\ & x_{1-} & \geq & 0 \\ & x_2 & \geq & 0 \\ & y_1 & \geq & 0 \\ & y_2 & \geq & 0 \\ & y_3 & \geq & 0 \end{array}$$

$$\begin{array}{lll} \boldsymbol{c}^T\boldsymbol{x}\overset{!}{=}\min, & \boldsymbol{x}=(x_{1+},x_{1-},x_2,y_1,y_2,y_3)^T, & \boldsymbol{c}=(2,-2,-3,0,0,0)^T\in\mathbb{R}^6 \\ \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}, & \boldsymbol{A}=(\boldsymbol{a}^1,-\boldsymbol{a}^1,\boldsymbol{a}^2,\boldsymbol{I}_m)\in\mathbb{R}^{m\times n}, & \boldsymbol{a}^1=(-1,-1,1)^T,\boldsymbol{a}^2=(1,1,0)^T \\ \boldsymbol{x}\geq 0, & \multicolumn{2}{l}{m=3=\text{Zahl der Gleichungen}\ <\ n=6=\text{Zahl der Variablen}} \end{array}$$

Bemerkungen zur Form (L)**:**
Das Fazit aus den Beispielen 1) – 4) läßt vermuten, daß „Ecken“ bei der Lösung der Optimierungsprobleme eine wesentliche Rolle spielen werden. In $\mathbb{R}^2$ und $\mathbb{R}^3$ lassen sich Ecken als Schnittpunkte von Geraden oder Ebenen, also durch Gleichungen beschreiben. Obwohl in den meisten Fällen Nebenbedingungen durch Ungleichungen gegeben werden („$\boldsymbol{A}\,\boldsymbol{x} \leq \boldsymbol{b}$“), ist es deshalb günstiger, sie in Form von Gleichungen und leichter zu behandelnden Vorzeichenrestriktionen darzustellen.

Um zu einer Lösungsidee für das Problem (L) zu kommen, kehren wir nochmals zu den anfangs betrachteten Beispielen 1) – 4) zurück.

Da die optimale Lösung auch immer in einer Ecke angenommen wird, liegt es nahe, alle Ecken zu bestimmen, in ihnen den Zielfunktionswert zu berechnen und die Ecke mit dem optimalen Zielfunktionswert auszuwählen.

Dieses Vorgehen ist problematisch, weil es einerseits, insbesondere bei vielen Ungleichungsrestriktionen, schwierig sein wird, alle Ecken zu bestimmen, und andererseits dabei viele Ecken umsonst berechnet werden.

Erfolgversprechend ist die nächste Idee: Man beginne mit einer Ecke (wie findet man die?), bestimme in dieser Ecke den Zielfunktionswert und gehe dann längs einer Kante, längs der der Zielfunktionswert abnimmt, zu einer Nachbarecke über und bestimme deren Zielfunktionswert usw., bis man am Minimum angelangt ist. (Wie erkennt man das?) Dabei können natürlich alle die Situationen eintreten, die wir in den Beispielen 1) – 4) kennengelernt haben. Insbesondere hoffen wir aber, daß man auf diese Weise nicht alle Ecken durchlaufen muß, bis man am Minimum angekommen ist.

Man mache sich diese Problematik auch an Beispielen im $\mathbb{R}^3$ klar.

Wir haben in der Tat mit der obigen Vorgehensweise die Grundidee des nun zu besprechenden Lösungsalgorithmus beschrieben. Bei der Umsetzung dieser geometrischen Lösungsidee in einen Lösungsalgorithmus für die Aufgabenstellung (L) treten eine Reihe von Problemen auf, wie etwa

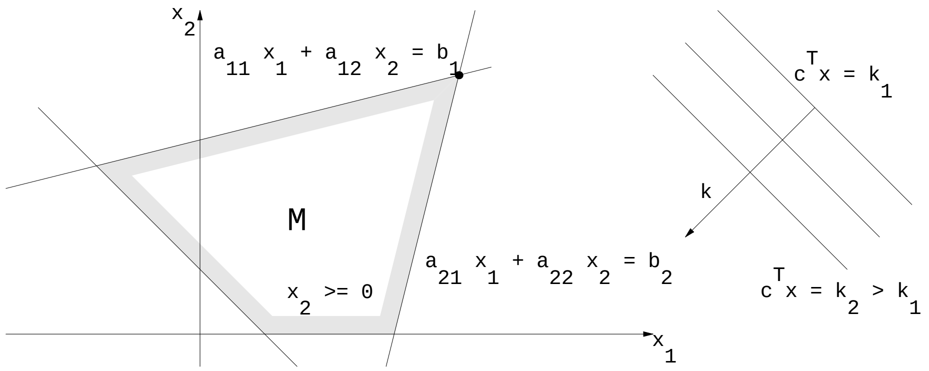
- Wie beschreibt man Ecken im $\mathbb{R}^n$?
- Wie beschreibt man Kanten?
- Was heißt „Laufen auf einer Kante“?
- Wird der optimale Wert wirklich in einer Ecke angenommen?
- Was bedeuten die Beobachtungen aus den Beispielen 1) – 4) im allgemeinen Fall?

Diese und weitere auftauchende Fragen wollen wir im folgenden zu lösen versuchen. Dabei werden die Lösungsideen aus der anschaulichen Problemstellung (L') kommen (, die wir der Anschauung halber nochmals, samt ihrer geometrischen Interpretation skizzieren,) und werden dann in die algebraische und algorithmische Sprache für das Problem (L) übersetzt werden müssen.

Lineare Optimierungsaufgaben: Formulierungen

$$\left\{\begin{array}{lllll} & & & \text{Matrixschreibweise} & \\ \sum\limits_{j=1}^{n} c_j\, x_j \overset{!}{=} \min\,, & \boldsymbol{c} = (c_1, \ldots, c_n)^T \in \mathbb{R}^n & & \boldsymbol{c}^T \boldsymbol{x} \overset{!}{=} \min & \\ \sum\limits_{j=1}^{n} a_{ij}\, x_j \le b_i, & i = 1, \ldots, m\,, & & \boldsymbol{A}\,\boldsymbol{x} \le \boldsymbol{b} & \\ & \boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{m\times n} & & & \\ & \boldsymbol{b} = (b_1, \ldots, b_m)^T \in \mathbb{R}^m & & & \\ x_i \ge 0,\ i \in J \subset \{1, \ldots, n\} & & & x_i \ge 0, \quad i \in J & \end{array}\right. \qquad (L')$$

Im $\mathbb{R}^2$ (auch $\mathbb{R}^3$) hat man die Veranschaulichung



$M = \{\boldsymbol{x} \in \mathbb{R}^n;\, \boldsymbol{A}\,\boldsymbol{x} \le \boldsymbol{b},\, x_i \ge 0,\, i \in J\}$ heißt *zulässiger Bereich*, $\boldsymbol{c}^T\boldsymbol{x}$ heißt *Zielfunktion*.

Problem (L)

$$\left\{\begin{array}{l} \boldsymbol{c}^T\,\boldsymbol{x} \overset{!}{=} \min, \qquad \boldsymbol{c} \in \mathbb{R}^n \\ \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{m\times n}, \quad \boldsymbol{b} \in \mathbb{R}^m \\ \boldsymbol{x} \ge \boldsymbol{0} \end{array}\right. \qquad \text{(L)}$$

$M = \{\boldsymbol{x} \in \mathbb{R}^n;\ \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \ge \boldsymbol{0}\}$ heißt *zulässiger Bereich*, $\boldsymbol{c}^T\boldsymbol{x}$ heißt *Zielfunktion*, und $\boldsymbol{x}^*$ heißt *optimal*, falls $\boldsymbol{x}^* \in M$ und $\boldsymbol{c}^T\,\boldsymbol{x}^* \le \boldsymbol{c}^T\,\boldsymbol{x}\ \forall\, \boldsymbol{x} \in M$.

Beachte: Die Vektoren $\boldsymbol{c}, \boldsymbol{b}$ und die Matrix $\boldsymbol{A}$ sind gemäß den Umformungen nicht die gleichen wie in (L').

Beschreibung von Ecken

Der zulässige Bereich eines linearen Optimierungsproblems wird im $\mathbb{R}^2$ durch Geraden begrenzt, im $\mathbb{R}^3$ durch Ebenen, im $\mathbb{R}^n$ durch sog. Hyperebenen.

Geometrisch anschaulich ist eine Ecke eines solchen Bereichs M ein Punkt $\in M$, der nicht auf der Verbindungsgeraden zweier Punkte liegt, die auch $\in M$ sind.

Mathematische Fassung dieses Sachverhalts:

Definition 6.1

1) Eine Menge $K \subset \mathbb{R}^n$ heißt *konvex*

$$\overset{\text{(Def.)}}{\Longleftrightarrow} \quad \begin{cases} \forall\, \boldsymbol{x}^1, \boldsymbol{x}^2 \;\in\; K \quad \text{gilt} \\ \boldsymbol{x} \;=\; \lambda\, \boldsymbol{x}^1 + (1-\lambda)\, \boldsymbol{x}^2 \;\in\; K, \quad 0 \;\le\; \lambda \;\le\; 1 \end{cases}$$

$\boldsymbol{x}$ heißt *Konvexkombination* von $\boldsymbol{x}^1$ und $\boldsymbol{x}^2$.

2) Eine Konvexkombination heißt *echt*

$$\overset{\text{(Def.)}}{\Longleftrightarrow} \qquad \lambda \;\neq\; 0,\, 1$$

3) Sei $K \subset \mathbb{R}^n$ eine Menge, die durch Hyperebenen begrenzt wird.
$\boldsymbol{x} \in K$ heißt *Ecke von* K

$\overset{\text{(Def.)}}{\Longleftrightarrow}$ $\boldsymbol{x}$ hat keine Darstellung als echte Konvex–kombination 2er verschiedener Punkte von K

Sätzchen 6.2:
Die zulässige Menge M eines linearen Optimierungsproblems (in der Normalform)

$$M \;=\; \{\boldsymbol{x} \in \mathbb{R}^n,\; \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b},\; \boldsymbol{x} \ge 0\} \quad \text{wo} \quad \boldsymbol{A} \;\in\; \mathbb{R}^{m\times n},\; \boldsymbol{b} \;\in\; \mathbb{R}^m,\; m \;<\; n\,,$$

ist konvex.

Beweis:
Seien $\boldsymbol{x}^1, \boldsymbol{x}^2 \in M$, $\boldsymbol{x}^1 \neq \boldsymbol{x}^2$ und $\boldsymbol{x} = \lambda\, \boldsymbol{x}^1 + (1-\lambda)\, \boldsymbol{x}^2$, $\quad \lambda \in [0,1]$, so folgt

$$\begin{aligned} \boldsymbol{A}\,\boldsymbol{x} \;&=\; \boldsymbol{A}(\lambda\, \boldsymbol{x}^1 + (1-\lambda)\, \boldsymbol{x}^2) \;=\; \lambda\, \boldsymbol{A}\,\boldsymbol{x}^1 + (1-\lambda)\, \boldsymbol{A}\,\boldsymbol{x}^2 \\ &=\; \lambda\, \boldsymbol{b} \quad + (1-\lambda)\, \boldsymbol{b} \;=\; \boldsymbol{b}\,. \end{aligned}$$

Aus $\boldsymbol{x} = \underbrace{\lambda}_{\ge 0}\; \underbrace{\boldsymbol{x}^1}_{\ge 0} + \underbrace{(1-\lambda)}_{\ge 0}\; \underbrace{\boldsymbol{x}^2}_{\ge 0}$ folgt $\boldsymbol{x} \ge \boldsymbol{0}$, also $\boldsymbol{x} \in M$. ■

Charakterisierung von Ecken
Sei $\boldsymbol{A} = (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n)$, d.h. $\boldsymbol{a}^i$ seien die Spalten von $\boldsymbol{A}$. Für $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ kann $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ geschrieben werden als $\sum\limits_{i=1}^{n} \boldsymbol{a}^i\, x_i = \boldsymbol{b}$.

Sei $M = \{\boldsymbol{x} \in \mathbb{R}^n\,;\ \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$ der zulässige Bereich eines linearen Optimierungsproblems, dann kann $\boldsymbol{x} \in M$ (Œ bis auf Umnumerierung) geschrieben werden als

$$\boldsymbol{x} = (x_1, \ldots, x_p, 0, \ldots, 0)^T, \qquad x_1, \ldots, x_p > 0, \qquad 0 \leq p \leq n\,.$$

Dann gilt folgende Charakterisierung von Ecken:

Satz 6.3

Sei $\overline{\boldsymbol{x}} \in M = \{\boldsymbol{x} \in \mathbb{R}^n\,;\ \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0}\}$, so daß

$$\sum_{i=1}^{n} \boldsymbol{a}^i\,\overline{x}_i = \boldsymbol{b},$$

$$\overline{x}_i > 0 \quad \text{für} \quad i = 1, \ldots, p, \qquad \overline{x}_i = 0 \quad \text{für} \quad i = p+1, \ldots, n, \quad p \geq 0.$$

Dann gilt

a) $\overline{\boldsymbol{x}} \equiv \boldsymbol{0}$ ist Ecke von M $\iff$ $\boldsymbol{b} = \boldsymbol{0}\,.$ $(p = 0)$

b) $\overline{\boldsymbol{x}} \not\equiv \boldsymbol{0}$ ist Ecke von M $\iff$ Die Spalten $\boldsymbol{a}^i$, die zu positiven Komponenten von $\overline{\boldsymbol{x}}$ gehören, sind linear unabhängig. $(p > 0)$

Beweis:

a) offensichtlich, denn $\boldsymbol{0} = \lambda \boldsymbol{x}^1 + (1-\lambda)\boldsymbol{x}^2,\ \boldsymbol{x}^1, \boldsymbol{x}^2 \geq 0,\ \lambda > 0 \implies \boldsymbol{x}^1 = \boldsymbol{x}^2 = \boldsymbol{0}$.

b) „$\Rightarrow$“ (indirekt): Annahme: $\boldsymbol{a}^1, \ldots, \boldsymbol{a}^p$ seien linear abhängig, d.h. $\exists (d_1, \ldots, d_p)^T \in \mathbb{R}^p$, nicht alle $d_i = 0$, mit $\sum\limits_{i=1}^{p} \boldsymbol{a}^i\,d_i = \boldsymbol{0}$.

Setze

$$\left.\begin{array}{ll} \boldsymbol{d} := (d_1, \ldots, d_p, 0, \ldots, 0)^T \in \mathbb{R}^n \Rightarrow & \boldsymbol{A}\,\boldsymbol{d} = \boldsymbol{0} \\ \text{für } \overline{\boldsymbol{x}} \in M \text{ gilt} & \left\{\begin{array}{l} \boldsymbol{A}\,\overline{\boldsymbol{x}} = \boldsymbol{b} \\ \overline{\boldsymbol{x}} \geq \boldsymbol{0} \end{array}\right. \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \boldsymbol{A}(\overline{\boldsymbol{x}} + \delta\,\boldsymbol{d}) = \boldsymbol{b} \quad \forall\,\delta \in \mathbb{R} \\ \overline{\boldsymbol{x}} + \delta\,\boldsymbol{d} \geq 0 \text{ für kleine } |\delta| \in \mathbb{R}\,, \\ \text{da } \overline{x}_i > 0 \text{ für } i = 1, \ldots, p\,, \end{array}\right.$$

d.h. $\exists\,\delta > 0$, so daß $\overline{\boldsymbol{x}} + \delta\,\boldsymbol{d} = \boldsymbol{x}^1 \in M$, $\overline{\boldsymbol{x}} - \delta\,\boldsymbol{d} = \boldsymbol{x}^2 \in M$ und $\boldsymbol{x}^1 \neq \boldsymbol{x}^2$

$$\Rightarrow \qquad \overline{\boldsymbol{x}} = \frac{1}{2}(\boldsymbol{x}^1 + \boldsymbol{x}^2) \qquad \overset{\text{(Def.)}}{\Longrightarrow} \qquad \overline{\boldsymbol{x}} \text{ ist keine Ecke}\,.$$

Beweis:

b) „$\Leftarrow$“ (indirekt) Annahme: $\overline{\boldsymbol{x}}$ ist keine Ecke, d.h.

$$\exists\, \boldsymbol{x}^1, \boldsymbol{x}^2 \in M,\ \ \boldsymbol{x}^1 \neq \boldsymbol{x}^2, \ \text{ so daß }\ \overline{\boldsymbol{x}} = \underbrace{\lambda}_{>0}\ \underbrace{\boldsymbol{x}^1}_{\geq \mathbf{0}} + \underbrace{(1-\lambda)}_{>0}\ \underbrace{\boldsymbol{x}^2}_{\geq \mathbf{0}},\ \ \lambda \in (0,1)$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\left\{ \begin{array}{l} \boldsymbol{A}\,\boldsymbol{x}^1 = \boldsymbol{b} \\ \boldsymbol{A}\,\boldsymbol{x}^2 = \boldsymbol{b} \end{array} \right\} \qquad \left\{ \begin{array}{l} x_i^1 = x_i^2 = 0 \text{ für } i > p \\ \text{da } \boldsymbol{x}^1, \boldsymbol{x}^2 \geq \mathbf{0} \text{ und } \bar{x}_i = 0 \text{ für } i > p \end{array} \right\}$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$\left.\begin{array}{l} \boldsymbol{A}(\boldsymbol{x}^1 - \boldsymbol{x}^2) \ = \ \mathbf{0} \qquad \Rightarrow \ \sum\limits_{i=1}^{p} \boldsymbol{a}^i (x_i^1 - x_i^2) = \mathbf{0} \ \overset{\boldsymbol{a}^i \ \text{lin.}}{\underset{\text{unabh.}}{\Longrightarrow}} \ \left\{ \begin{array}{l} x_i^1 = x_i^2 \\ \text{für } i \leq p \end{array} \right\} \end{array}\right\} \Rightarrow \ \boldsymbol{x}^1 = \boldsymbol{x}^2 .$$

■

Korollar 6.4
Eine Ecke von M hat maximal m positive Komponenten, wo $m = \text{Rang}\ \boldsymbol{A}$.

Eine Ecke $\boldsymbol{x}^*$ heißt *entartet*, wenn sie weniger als m positive Komponenten besitzt. Man kann dann die zu positiven Komponenten von $\boldsymbol{x}^*$ gehörenden Spaltenvektoren von $\boldsymbol{A}$ durch Hinzunahme weiterer Spalten von $\boldsymbol{A}$ zu einer Basis des $\mathbb{R}^m$ ergänzen und kommt somit zum Begriff der Basislösung.

Definition 6.5
Gegeben sei das Problem $\boldsymbol{c}^T \boldsymbol{x} \ \overset{!}{=} \ \min, \quad \boldsymbol{x} \in M = \{\boldsymbol{x} \in \mathbb{R}^n;\ \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \mathbf{0}\}$,
$\text{Rg}\,\boldsymbol{A} = m$.

$$\left.\begin{array}{l} \boldsymbol{x} \in M \text{ heißt } \textit{Basislösung} \\ \text{(Basisvektor) zur Indexmenge } J \end{array}\right\} \ \overset{\text{(Def.)}}{\Longleftrightarrow} \ \left\{ \begin{array}{l} \exists\, m \text{ linear unabhängige Spalten} \\ \qquad \boldsymbol{a}^i \text{ von } \boldsymbol{A}\,, \\ i \in J,\ |J| = m\,, \quad \text{so daß} \\ x_i = 0 \quad \text{für} \quad i \notin J \quad \text{und} \\ \sum\limits_{i \in J} \boldsymbol{a}^i\, x_i = \boldsymbol{b} \end{array} \right.$$

Beachte:

Es wird nicht gefordert $x_i > 0$ für $i \in J$. Jede Ecke $\boldsymbol{x} \in M$ definiert also eine Basislösung zu einer (im entarteten Fall nicht eindeutigen) Basis des $\mathbb{R}^m$ aus Spaltenvektoren von $\boldsymbol{A}$. Umgekehrt beschreibt jede Basislösung eindeutig eine Ecke.

Beispiel einer entarteten Ecke bzw. einer entarteten Basislösung: Wird im $\mathbb{R}^3$ M durch eine 4– seitige Pyramide, die auf der $(x_1, x_2)-$Ebene steht, dargestellt, so ist ihre Spitze eine entartete Ecke. (Formulierung als (L) im $\mathbb{R}^7$: x_1, x_2, x_3, 4 Schlupfvariable y_i, $\mathrm{Rg}(\boldsymbol{A}) = 4$; $y_i = 0$, $i = 1...4$, beschreibt die Pyramidenspitze, das ist eine Restriktion zu viel.)

Als nächstes müssen wir uns fragen: Hat das Problem (L) überhaupt Ecken (im $\mathbb{R}^2$ wird z.B. durch $x_2 \geq 0$, $x_2 \leq 2$ ein Bereich ohne Ecken beschrieben), und wird, falls ja, ein optimaler Zielfunktionswert (falls einer existiert, vgl. Beispiel 1) auch in einer Ecke angenommen? Antwort gibt

Satz 6.6

Für (L) gilt:

a) $M \neq \emptyset \;\Rightarrow\; \exists$ eine Basislösung

(d.h. wenn zulässige Punkte existieren, so gibt es auch eine Basislösung).

b) $\exists\, \boldsymbol{x}^* \in M$ optimal $\;\Rightarrow\; \exists$ eine optimale Basislösung.

Beweis a):

Geometrische Idee für a).

Ist $\boldsymbol{x}^* \in M$ keine Ecke, so existiert eine Gerade durch $\boldsymbol{x}^*$, auf der man in M in beiden Richtungen ein Stück laufen kann (vgl. Def. von Ecke), bis man zu einem Punkt $\overline{\boldsymbol{x}}$ einer Kante (bzw. einer Seite) kommt. $\overline{\boldsymbol{x}}$ erfüllt eine Gleichheitsrestriktion mehr als $\boldsymbol{x}^*$ (die Schlupfvariable der zugehörigen Ungleichheitsrestriktion ist $= 0$). $\overline{\boldsymbol{x}}$ hat also mehr Nullkomponenten als $\boldsymbol{x}^*$. Ist $\overline{\boldsymbol{x}}$ keine Ecke, so kann man obigen Prozeß wiederholen.

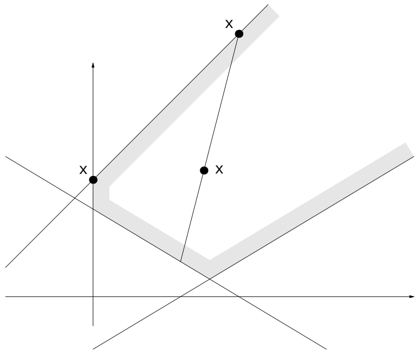


Abb. 6.2

Mathematische Umsetzung:

$\boldsymbol{x}^*$ **ist keine Basislösung**, d.h. $\boldsymbol{x}^* = (x_1^*, \ldots, x_p^*, 0, \ldots, 0)^T$ (geeignete Numerierung)

$$\begin{aligned} x_i^* &> 0, \qquad i = 1, \ldots, p\,, \\ \sum_{i=1}^{p} \boldsymbol{a}^i\, x_i^* &= \boldsymbol{b}\,, \\ \boldsymbol{a}^i, \qquad i &= 1, \ldots, p \quad \text{linear abhängig, d.h.} \end{aligned}$$

$$\exists y_i \in \mathbb{R}, i = 1, \ldots, p \quad : \quad \sum_{i=1}^{p} \boldsymbol{a}^i\, y_i = \boldsymbol{0}, \quad \text{nicht alle} \quad y_i = 0. \tag{6.3}$$

Laufen auf einer Geraden durch $\boldsymbol{x}^*$:

$$\boldsymbol{x} \;=\; \boldsymbol{x}^* + \varepsilon\,\tilde{\boldsymbol{y}}, \qquad \varepsilon \in \mathbb{R}, \qquad \tilde{\boldsymbol{y}} \in \mathbb{R}^n \text{ fest}.$$

Laufen in M, d.h. $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \geq 0$ bzw.

$$\boldsymbol{A}\,(\boldsymbol{x}^* + \varepsilon\,\tilde{\boldsymbol{y}}) \;=\; \boldsymbol{b}\,, \tag{6.4}$$

$$\boldsymbol{x}^* + \varepsilon\,\tilde{\boldsymbol{y}} \;\geq\; \mathbf{0}\,. \tag{6.5}$$

Aus (6.4) folgt wegen $\boldsymbol{A}\,\boldsymbol{x}^* \;=\; \boldsymbol{b}$

$$\boldsymbol{A}\,\tilde{\boldsymbol{y}} \;=\; \mathbf{0}\,. \tag{6.6}$$

Nur $\tilde{\boldsymbol{y}}$, die diese Bedingungen erfüllen, sind zulässig (man verdeutliche sich das für den Fall, daß $\boldsymbol{x}^*$ selbst schon auf einer Kante liegt). Solche Vektoren $\tilde{\boldsymbol{y}}$ existieren aber nach (6.3), weil $\boldsymbol{x}^*$ keine Ecke ist.

$$\tilde{\boldsymbol{y}} \;:=\; (y_1, \ldots, y_p, 0, \ldots, 0)^T\,.$$

Weiter muß gelten (6.5), d.h.

$$\begin{aligned} x_i^* + \varepsilon\, y_i &\;\geq\; 0 \quad i = 1, \ldots, p \quad (\text{erfüllbar für kleine } |\varepsilon|, \text{ da } x_i^* > 0)\,, \\ x_\nu^* + \varepsilon\, y_\nu &\;=\; 0 \quad \text{für ein } \nu \in \{1, \ldots, p\} \quad (\text{man möchte eine weitere} \\ &\qquad\qquad \text{Nullkomponente})\,. \end{aligned} \tag{6.7}$$

Welche Bedingungen resultieren aus (6.7) für ε? (beachte $x_i^* > 0 \;\; \forall\, i = 1, \ldots, p$)

$$\left.\begin{aligned} &\text{Für} \;\; \varepsilon > 0 \;\; \text{folgt} \;\; \frac{x_i^*}{\varepsilon} \;\geq\; -y_i \\ &\text{für} \;\; \varepsilon < 0 \;\; \text{folgt} \;\; \frac{x_i^*}{\varepsilon} \;\leq\; -y_i \end{aligned}\right\} \quad \text{für} \;\; i = 1, \ldots, p\,,$$

bzw.

$$\left.\begin{aligned} &\frac{1}{\varepsilon} \geq -\frac{y_i}{x_i^*} \qquad \text{falls} \;\; \varepsilon > 0 \\ &\frac{1}{\varepsilon} \leq -\frac{y_i}{x_i^*} \qquad \text{falls} \;\; \varepsilon < 0 \end{aligned}\right\} \quad \text{für} \;\; i = 1, \ldots, p\,.$$

Wir unterscheiden die Fälle:

1) $\varepsilon > 0, \; \exists\, y_i < 0 \quad \Rightarrow \quad \frac{1}{\varepsilon} \geq \max\limits_{y_i < 0} \left(-\frac{y_i}{x_i^*}\right) =: \frac{1}{\varepsilon_1} \quad \Longleftrightarrow \quad \varepsilon \leq \varepsilon_1$

2) $\varepsilon > 0, \; y_i > 0 \;\forall\, i \quad \Rightarrow \quad \varepsilon > 0$ beliebig wählbar

3) $\varepsilon < 0, \; \exists\, y_i > 0 \quad \Rightarrow \quad \frac{1}{\varepsilon} \leq \min\limits_{y_i > 0} \left(-\frac{y_i}{x_i^*}\right) =: \frac{1}{\varepsilon_2} \quad \Longleftrightarrow \quad \varepsilon \geq \varepsilon_2$

4) $\varepsilon < 0, \; y_i < 0 \;\forall\, i \quad \Rightarrow \quad \varepsilon < 0$ beliebig wählbar

Fazit:

ε ist mindestens einseitig beschränkt; für $\varepsilon = \varepsilon_1$ oder $\varepsilon = \varepsilon_2$ hat der Punkt $\overline{\boldsymbol{x}} := \boldsymbol{x}^* + \varepsilon\, \tilde{\boldsymbol{y}}$ mindestens eine Nullkomponente mehr als $\boldsymbol{x}^*$.

Beweis b):

Geometrische Idee für b): Ist $\boldsymbol{x}^*$ bereits optimal, z.B. wenn $\boldsymbol{x}^*$ auf einer „Seite“ von M liegt (wie etwa $\overline{\boldsymbol{x}}$: vgl. Abb.6.2), so sind alle Punkte dieser Seite, da sie zulässig sind, auch optimal. Der Punkt $\hat{\boldsymbol{x}}$, der wie in Teil a) konstruiert wird, muß dann den gleichen Zielfunktionswert haben wie $\overline{\boldsymbol{x}}$ und mindestens eine Nullkomponente mehr. D.h. alle nach Teil a) konstruierten Punkte sind ebenfalls optimal.

Mathematische Umsetzung

$\boldsymbol{x}^*$ optimal $\iff$ $\boldsymbol{c}^T \boldsymbol{x}^* \le \boldsymbol{c}^T \boldsymbol{x} \;\forall\, \boldsymbol{x} \in M$

Œ sei $|\varepsilon_1| \le |\varepsilon_2|$, dann gilt insbesondere

$$\left.\begin{array}{ll} \boldsymbol{c}^T \boldsymbol{x}^* & \le \boldsymbol{c}^T\,(\boldsymbol{x}^* + \varepsilon_1\,\tilde{\boldsymbol{y}}) \\ \boldsymbol{c}^T \boldsymbol{x}^* & \le \boldsymbol{c}^T\,(\boldsymbol{x}^* - \varepsilon_1\,\tilde{\boldsymbol{y}}) \end{array}\right\} \quad \Rightarrow \quad \boldsymbol{c}^T \boldsymbol{y} \;=\; \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{c}^T \boldsymbol{x}^* \;=\; \boldsymbol{c}^T\,(\boldsymbol{x}^* + \varepsilon_1\,\tilde{\boldsymbol{y}})$$

■

Das Simplex–Verfahren

Wir beziehen uns auf die Aufgabenstellung: Bestimme $\boldsymbol{x} \in \mathbb{R}^n$, so daß

$$\begin{cases} \boldsymbol{c}^T \boldsymbol{x} \;\stackrel{!}{=}\; \min, \qquad \boldsymbol{c} \in \mathbb{R}^n\,, \\ \boldsymbol{A}\,\boldsymbol{x} \;=\; \boldsymbol{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{m\times n}, \quad \text{Rang } \boldsymbol{A} \;=\; m \;<\; n, \qquad \boldsymbol{b} \in \mathbb{R}^m\,, \\ \boldsymbol{x} \;\ge\; \boldsymbol{0}\,. \end{cases} \tag{L}$$

Wir wissen: Wenn eine optimale Lösung existiert, so wird sie auch in einer Ecke des zulässigen Bereichs $M = \{\boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \ge \boldsymbol{0}\}$ angenommen.

Bemerkung:

Rang $\boldsymbol{A} = m$ sichert die Lösbarkeit des Gleichungssystems, bedeutet aber, daß linear abhängige Gleichungen ausgeschieden werden müssen. Dies kann numerisch schwierig sein.

Idee des Verfahrens

1) Man geht aus von einer Ecke von M (wie man die findet, wird später untersucht) und bestimmt für diese Ecke den Zielfunktionswert.

2) Man untersucht, ob es „benachbarte“ Ecken gibt, die einen kleineren Zielfunktionswert haben und geht gegebenenfalls zu einer solchen Ecke über (Eckentausch). Diesen Eckentausch führt man so lange fort, bis keine benachbarte Ecke mit kleinerem Zielfunktionswert mehr zu finden ist.

1. Unterproblem: Wie vergleicht man formal (man sucht ja einen Rechenalgorithmus) am geeignetsten die Zielfunktionswerte zweier Ecken (bzw. einer Ecke und eines anderen zulässigen Punktes)?

2. Unterproblem: In welcher (formalen) Beziehung steht eine Ecke $\boldsymbol{x}^*$ zu einem anderen zulässigen Punkt $\overline{\boldsymbol{x}}$?

Sei also $\overline{\boldsymbol{x}} \in M$ und $\boldsymbol{x}^* \in M$ eine **Basislösung**, d.h. $\exists\, J \subset \{1,\ldots,n\}$, $|J| = m$, so daß

$$\begin{cases} \boldsymbol{a}^i, \quad i \in J \quad \text{linear unabhängig}, \\ x_i^* = 0, \quad i \notin J, \\ \sum\limits_{i\in J} x_i^* \boldsymbol{a}^i = \boldsymbol{b}, \quad \boldsymbol{x}^* \geq \boldsymbol{0}. \end{cases}$$

Vergleich von $\boldsymbol{x}^*$ und $\overline{\boldsymbol{x}}$:
$\boldsymbol{x}^*, \overline{\boldsymbol{x}} \in M$ liefert

$$\sum_{i\in J} x_i^* \boldsymbol{a}^i = \boldsymbol{b} = \sum_{j=1}^n \overline{\boldsymbol{x}}_j \boldsymbol{a}^j . \tag{6.8}$$

Idee:
Die $\boldsymbol{a}^i$, $i \in J$ bilden eine Basis für alle Spalten $\boldsymbol{a}^j$. Setzt man diese Basisdarstellung in obige Gleichung ein, so erhält man eine Beziehung zwischen x_i^* und $\overline{x}_i$.

Basisdarstellung: $\forall$ Spalten $\boldsymbol{a}^j$ von $\boldsymbol{A}$ $\exists\, d_{ij} \in \mathbb{R}$, $i \in J$, $j = 1,\ldots,n$ mit

$$\boldsymbol{a}^j = \sum_{i\in J} d_{ij}\, \boldsymbol{a}^i, \quad j = 1,\ldots,n, \quad \text{wobei für alle} \quad j \in J: \; d_{ij} = \begin{Bmatrix} 1 \text{ für } i = j \\ 0 \text{ für } i \neq j \end{Bmatrix} . \tag{6.9}$$

Einsetzen von (6.9) in (6.8) ergibt

$$\sum_{i\in J} x_i^* \boldsymbol{a}^i = \sum_{j=1}^n \overline{x}_j \left(\sum_{i\in J} d_{ij}\, \boldsymbol{a}^i \right) = \sum_{i\in J} \left(\sum_{j=1}^n d_{ij}\, \overline{x}_j \right) \boldsymbol{a}^i .$$

Koeffizientenvergleich der $\boldsymbol{a}^i$, $i \in J$, ergibt für $i \in J$:

$$\begin{aligned} x_i^* &= \sum_{j=1}^n d_{ij}\, \overline{x}_j \\ &= \sum_{j\in J} d_{ij}\, \overline{x}_j + \sum_{j\notin J} d_{ij}\, \overline{x}_j \\ &= \overline{x}_i + \sum_{j\notin J} d_{ij}\, \overline{x}_j \qquad \text{gemäß (6.9)}, \end{aligned}$$

also folgt

$$\boxed{\overline{x}_i = x_i^* - \sum_{j\notin J} d_{ij}\, \overline{x}_j , \qquad i \in J} \tag{6.10}$$

Vergleich von $\boldsymbol{c}^T \boldsymbol{x}^*$ und $\boldsymbol{c}^T \overline{\boldsymbol{x}}$:
Zum Vergleich wird (6.10) in die Zielfunktion eingesetzt:

$$\begin{aligned} \boldsymbol{c}^T \overline{\boldsymbol{x}} &= \sum_{i\in J} c_i \overline{x}_i + \sum_{i\notin J} c_i \overline{x}_i = \sum_{i\in J} c_i \left(x_i^* - \sum_{j\notin J} d_{ij} \overline{x}_j \right) + \sum_{j\notin J} c_j \overline{x}_j \\ &= \sum_{i\in J} c_i x_i^* - \sum_{j\notin J} \sum_{i\in J} c_i d_{ij} \overline{x}_j + \sum_{j\notin J} c_j \overline{x}_j , \end{aligned}$$

also

$$\boxed{\boldsymbol{c}^T \overline{\boldsymbol{x}} = \boldsymbol{c}^T \boldsymbol{x}^* + \sum_{j\notin J} \left(c_j - \sum_{i\in J} c_i d_{ij} \right) \overline{x}_j} \tag{6.11}$$

Hieraus liest man ab (beachte: die d_{ij} sind unabhängig von $\overline{\boldsymbol{x}}$):

Satz 6.7: Optimalitätskriterium
$c_j - \sum_{i\in J} c_i d_{ij} \geq 0 \ \forall j \notin J \quad \Rightarrow \quad \boldsymbol{x}^*$ optimal

bzw. ist $\boldsymbol{x}^*$ nicht optimal, dann gilt

$$\exists\, r \notin J : \quad c_r - \sum_{i\in J} c_i d_{ir} < 0 . \tag{6.12}$$

Unter dieser Voraussetzung arbeiten wir weiter.

Herleitung des Austauschschrittes
Ist $\boldsymbol{x}^*$ nicht optimal, so sucht man eine neue (benachbarte) Ecke $\overline{\boldsymbol{x}}$ mit

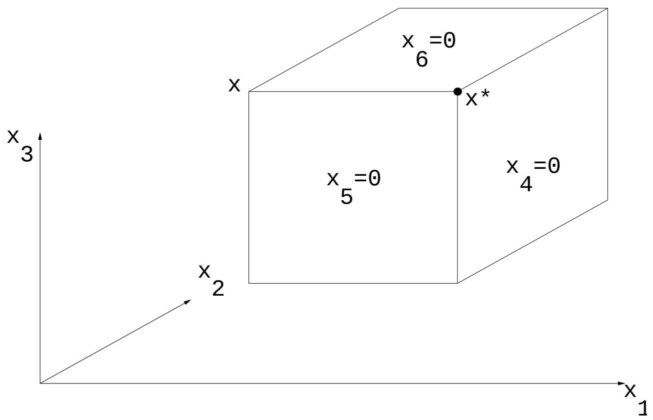
$$\boldsymbol{c}^T \overline{\boldsymbol{x}} < \boldsymbol{c}^T \boldsymbol{x}^* ,$$

gegen die $\boldsymbol{x}^*$ ausgetauscht werden soll. Wir verdeutlichen das Vorgehen zunächst an einem Beispiel.

Beispiel: Eckentausch
Sei M etwa ein Würfel im $\mathbb{R}^3$, dessen 6 Seiten durch die Ungleichungen

$$\sum_{j=1}^{3} a_{ij} x_j \leq b_i, \qquad i = 1, \ldots, 6 \tag{$*$}$$

beschrieben werden.



Nach Einführung von Schlupfvariablen $x_4, \ldots, x_9$ entsprechen $(*)$ die Gleichheitsrestriktionen

$$\sum_{j=1}^{3} a_{ij}\, x_j \; + \; x_{i+3} \;\; = \;\; b_i, \qquad i = 1, \ldots, 6$$

und die Vorzeichenrestriktionen

$$x_i \;\; \geq \;\; 0, \qquad i = 4, \ldots, 9\,.$$

Die Ecke $\boldsymbol{x}^*$ wird bestimmt durch die Restriktionen

$$x_j \;\; = \;\; 0, \qquad j = 4, 5, 6\,.$$

Will man zur Ecke $\overline{\boldsymbol{x}}$, so wird die Restriktion $x_4 = 0$ aufgegeben.

Eine Ecke $\boldsymbol{x}^*$ hat die Eigenschaft, daß sie mindestens $n-m$ Nullkomponenten hat (vgl. Korollar 6.4), d.h., daß sie mindestens $n-m$ Vorzeichenrestriktionen als Gleichungsrestriktionen erfüllt, d.h. daß sie auf mindestens $n-m$ Hyperebenen $x_\nu = 0$ liegt $(\nu \notin J)$ (vgl. Beispiel: Eckentausch). Man geht zu einer Nachbarecke $\overline{\boldsymbol{x}}$, indem man eine dieser Restriktionen aufgibt: d.h. für ein $r \notin J$ läßt man $\overline{x}_r = \delta > 0$ positiv werden, behält aber die anderen Restriktionen $\overline{x}_\nu = x^*_\nu = 0,\ \nu \notin J,\ \nu \neq r$ bei (im Beispiel sind das die Ebenen $x_5 = 0,\ x_6 = 0$). Wenn man sich auf der Kante von $\boldsymbol{x}^*$ nach $\overline{\boldsymbol{x}}$ bewegt, so ändern sich die Komponenten $x_i,\ i \in J$. Man macht für $\overline{\boldsymbol{x}}$ also zunächst den Ansatz

$$\overline{\boldsymbol{x}} \;=\; \begin{cases} \overline{x}_r = \delta > 0, \;\text{ für ein }\; r \notin J\,, \\ \overline{x}_j = x^*_j - \alpha_j\,,\;\; j \in J,\;\; \alpha_j = \;\text{ die Änderungen}\,, \\ \overline{x}_j = 0,\;\; \forall j \notin J,\;\; j \neq r \\ \qquad\qquad (\text{ die anderen Nullkomponenten will man beibehalten})\,. \end{cases} \tag{6.13}$$

r, δ und die Änderungen α_j müssen wir noch untersuchen.

Das $r \notin J$ wird man so wählen wollen, daß längs der ausgesuchten Kante der Zielfunktionswert möglichst schnell fällt. Setzt man $\overline{\boldsymbol{x}}$ in (6.11) ein, so folgt

$$\boldsymbol{c}^T\, \overline{\boldsymbol{x}} \;\;=\;\; \boldsymbol{c}^T\, \boldsymbol{x}^* + \left(c_r \;-\; \sum_{i \in J} c_i\, d_{ir} \right) \;\; \underbrace{\overline{x}_r}_{>\,0}\,. \tag{6.14}$$

Man wird also ein r aus (6.12) so wählen, daß gilt

$$t_r \quad := \quad c_r - \sum_{i\in J} c_i\, d_{ir} \quad = \quad \min_{j\notin J} \left(c_j - \sum_{i\in J} c_i\, d_{ij} \right) \; < \; 0\,, \tag{6.12a}$$

d.h. man wählt die Kante aus längs der der Zielfunktionswert am schnellsten abnimmt, denn t_r ist die Ableitung der Zielfunktion nach $\overline{x}_r$.
Bewegt man sich auf der Kante, so bedeutet das, daß die Änderungen α_i so beschaffen sein müssen, daß $\overline{\boldsymbol{x}}$ zulässig bleibt, d.h. $\boldsymbol{A}\,\overline{\boldsymbol{x}} = \boldsymbol{b}$, $\overline{\boldsymbol{x}} \geq 0$. Es muß also gelten

$$\begin{aligned} \boldsymbol{A}\,\overline{\boldsymbol{x}} \quad &= \quad \sum_{j\in J} \left(x_j^* - \alpha_j\right) \boldsymbol{a}^j \; + \; \delta\,\boldsymbol{a}^r \quad \stackrel{!}{=} \quad \boldsymbol{b} \quad \text{(Zulässigkeitsbedingung)} \\ &= \quad \underbrace{\sum_{j\in J} x_j^*\, \boldsymbol{a}^j}_{\substack{=\ \boldsymbol{b} \\ \text{da } \boldsymbol{x}^* \in M}} \quad - \; \sum_{j\in J} \alpha_j\, \boldsymbol{a}^j \; + \; \delta\,\boldsymbol{a}^r \quad \stackrel{!}{=} \quad \boldsymbol{b}\,, \end{aligned}$$

und somit

$$\delta\,\boldsymbol{a}^r \; - \; \sum_{j\in J} \alpha_j\, \boldsymbol{a}^j \quad = \quad \boldsymbol{0}\,. \tag{6.15}$$

Benutzt man für $\boldsymbol{a}^r$ die Basisdarstellung (6.9), so folgt aus (6.15)

$$\boldsymbol{0} \quad = \quad \delta \sum_{j\in J} d_{jr}\, \boldsymbol{a}^j \; - \; \sum_{j\in J} \alpha_j\, \boldsymbol{a}^j \quad = \quad \sum_{j\in J} \left(\delta\, d_{jr} - \alpha_j\right) \boldsymbol{a}^j$$

und daraus wegen der Basiseigenschaft der $\boldsymbol{a}^j$: $\delta\, d_{jr} - \alpha_j = 0$ bzw.

$$\alpha_j \quad = \quad \delta\, d_{jr}\,.$$

Unser verbesserter Ansatz für $\overline{\boldsymbol{x}}$ muß also lauten:

$$\overline{\boldsymbol{x}} \; = \; \overline{\boldsymbol{x}}(\delta) \; = \; \begin{cases} \overline{x}_r \; = \; \delta \; > \; 0\,, & r \; \notin \; J, \quad r \text{ gemäß (6.12a)} \\ \overline{x}_j \; = \; x_j^* - \delta\, d_{jr}\,, & j \; \in \; J \\ \overline{\boldsymbol{x}}_j \; = \; 0, \;\; \forall j \notin J, \;\; j \neq r & \end{cases} \tag{6.16}$$

$$\delta \; > \; 0 \text{ nur so groß, daß } x_j^* - \delta\, d_{jr} \quad \begin{cases} \geq \; 0 \;\; \forall\, j \in J, \\ = \; 0 \;\text{ für ein }\; \nu \in J \;\; \text{(falls möglich)}, \end{cases}$$

(man möchte ja eine neue Ecke haben,
braucht also eine neue Nullkomponente).

Was passiert, falls $d_{jr} \leq 0\; \forall\, j \in J$? Dann ist $\overline{\boldsymbol{x}}(\delta) \geq 0\; \forall\, \delta \in \mathbb{R}^+$ und wegen $\boldsymbol{A}\,\overline{\boldsymbol{x}}(\delta) = \boldsymbol{b}$ (so wurde der Ansatz konstruiert), gilt dann $\overline{\boldsymbol{x}}(\delta) \in M \;\; \forall\, \delta \in \mathbb{R}^+$, d.h. M ist nicht beschränkt, und wegen (6.12) und nach (6.14) ist auch die Zielfunktion nicht nach unten beschränkt, d.h. es gilt der

Satz 6.8: Abbruchkriterium
Gilt für ein

$$\begin{aligned} r \notin J: \;\; c_r - \sum_{j\in J} c_j\, d_{jr} \quad &< \quad 0 \quad \text{und} \\ d_{jr} \quad &\leq \quad 0\;\; \forall\, j \in J\,, \end{aligned}$$

so ist die Zielfunktion nicht nach unten beschränkt (d.h. es gibt keine optimale Lösung).

Wir setzen im folgenden also voraus

$$\exists\, j \in J \;:\; d_{jr} \;>\; 0, \qquad r \text{ gemäß (6.12a)}\,. \tag{6.17}$$

Dann muß für den Ansatz (6.16) die Zulässigkeitsforderung $\overline{\boldsymbol{x}}(\delta) \geq 0$ gesichert werden, d.h.

$$\begin{array}{rll} & x_j^* - d_{jr}\,\delta \geq 0 & \forall\, j \in J \text{ mit } d_{jr} > 0,\; r \text{ gemäß (6.12a)} \\ \Longleftrightarrow & \delta \leq \frac{x_j^*}{d_{jr}} & \forall\, j \in J \text{ mit } d_{jr} > 0,\; r \text{ gemäß (6.12a)}\,. \end{array}$$

Man wählt also in (6.16) (um eine neue Nullkomponente zu erhalten)

$$\delta \quad = \quad \min_{\substack{j \in J \\ d_{jr} > 0 \\ r \text{ nach (6.12a)}}} \frac{x_j^*}{d_{jr}} \quad =: \quad \frac{x_\nu^*}{d_{\nu\,r}} \quad (\nu \text{ nicht notwendig eindeutig})\,. \tag{6.18}$$

d.h. wegen $x_\nu^* - d_{\nu r}\,\delta = 0$ wird $\boldsymbol{a}^\nu$ aus der Basis ausgeschieden und durch $\boldsymbol{a}^r$ ersetzt.

Beachte:
$\delta = 0$ ist leider möglich, wenn $x_\nu^* = 0$ (vgl. Bemerkung zu Satz 6.9).
Anschaulich müßte $\overline{\boldsymbol{x}}(\delta)$ gemäß (6.16), (6.18) wieder eine Ecke sein.

Dies ist in der Tat richtig, wie folgender Satz zeigt.

Satz 6.9: Austauschschritt
Ist $\boldsymbol{x}^*$ Basislösung zur Indexmenge J, und existiert ein $r \notin J$ mit

$$c_r \;-\; \sum_{j \in J} d_{jr}\, c_j \;<\; 0$$

und ein $\mu \in J$ mit $d_{\mu\,r} > 0$, so ist für

$$\delta \quad = \quad \min_{\substack{\mu \in J \\ d_{\mu\,r} > 0}} \frac{x_\mu^*}{d_{\mu\,r}} \quad =: \quad \frac{x_\nu^*}{d_{\nu\,r}}$$

$$\overline{\boldsymbol{x}} \quad = \quad \left\{ \begin{array}{lcl} \overline{x}_r & = & \delta \\ \overline{x}_j & = & x_j^* - d_{jr}\,\delta,\;\; j \in J \\ \overline{x}_j & = & 0,\;\; \forall j \notin J,\;\; j \neq r \end{array} \right.$$

eine Basislösung mit $\boldsymbol{c}^T\,\overline{\boldsymbol{x}} \leq \boldsymbol{c}^T\,\boldsymbol{x}^*$ zur Indexmenge $\overline{J} \;=\; (J \setminus \{\nu\}\,) \cup \{r\}$.

Beweis:
$\overline{\boldsymbol{x}} \in M$ gilt laut Konstruktion. Für die Zielfunktion gilt nur (vgl.(6.14))

$$\boldsymbol{c}^T\,\overline{\boldsymbol{x}} \quad \leq \quad \boldsymbol{c}^T\,\boldsymbol{x}^* \;+\; \left(c_r \;-\; \sum_{i \in J} c_i\, d_{ir} \right) \quad \underbrace{\overline{x}_r}_{\geq\, 0}\,.$$

denn $x_r = \delta = 0$ ist möglich.

Zu zeigen bleibt: $a^j,\; j \in J,\; j \neq \nu$ und a^r sind linear unabhängig.

Annahme: Sie seien linear abhängig. Dann folgt: $\exists\, \lambda_i \in \mathbb{R},\; i \in J,\; i \neq \nu,\; \lambda_r,$ so daß

$$\sum_{\substack{i\in J\\ i\neq\nu}} \lambda_i\, \boldsymbol{a}^i \;+\; \lambda_r\, \boldsymbol{a}^r \;=\; \boldsymbol{0}, \text{ nicht alle } \lambda_i = 0, \text{ insbesondere } \lambda_r \neq 0\,, \tag{6.19}$$

denn aus $\lambda_r = 0$ folgte $\lambda_i = 0\; \forall\, i \in J,\; i \neq \nu$ (Basiseigenschaft der a^i). Wir setzen in (6.19) für $\boldsymbol{a}^r$ die Basisdarstellung $\boldsymbol{a}^r = \sum\limits_{i\in J} d_{ir}\, \boldsymbol{a}^i$ ein $\Rightarrow$

$$\boldsymbol{0} \;=\; \sum_{\substack{i\in J\\ i\neq\nu}} \lambda_i\, \boldsymbol{a}^i \;+\; \lambda_r \sum_{i\in J} d_{ir}\, \boldsymbol{a}^i \;=\; \sum_{\substack{i\in J\\ i\neq\nu}} (\lambda_i + \lambda_r\, d_{ir})\; \boldsymbol{a}^i \;+\; \lambda_r\, d_{\nu r}\, \boldsymbol{a}^\nu\,.$$

Da die $\boldsymbol{a}^i,\; i \in J$ eine Basis bilden, müssen alle Koeffizienten der $\boldsymbol{a}^i$ verschwinden, insbesondere $\lambda_r\, d_{\nu r} = 0$. Dies ist ein Widerspruch wegen $\lambda_r \neq 0,\; d_{\nu r} > 0$. ■

Bemerkung zu Satz 6.9:

1) Ist die Ecke $\boldsymbol{x}^*$ nicht entartet (d.h. $x_j^* > 0\; \forall\, j \in J$), so liefert der Austauschschritt (Satz 6.9) eine Ecke $\overline{\boldsymbol{x}}$ mit $\boldsymbol{c}^T\, \overline{\boldsymbol{x}} < \boldsymbol{c}^T\, \boldsymbol{x}^*$.

2) $\delta = 0$ kann vorkommen, wenn

$$x_j^* \;>\; 0 \quad \forall\, j \in J$$

nicht erfüllt ist. Dann ist die *Ecke* $\boldsymbol{x}^*$ *entartet*. Dann ist zwar $\overline{\boldsymbol{x}}$ wieder eine Basislösung mit der Indexmenge $\overline{J} \neq \boldsymbol{J}$. Aber $\boldsymbol{x}^*$ und $\overline{\boldsymbol{x}}$ beschreiben dieselbe Ecke. Es können Zyklen entstehen, bei denen in jedem Schritt eine neue Basislösung gefunden wird, die aber immer dieselbe Ecke beschreibt. Man kann solche Zyklen vermeiden, (vgl. Collatz/Wetterling), das Verfahren ist dann jedoch sehr aufwendig.

3) Ist $\boldsymbol{x}^*$ nicht entartete Ecke und lokales Minimum, so wird in $\boldsymbol{x}^*$ auch das globale Minimum angenommen (Übung!).

4) $\overline{\boldsymbol{x}}$ kann entartet sein (ohne daß $\boldsymbol{x}^*$ entartet ist), wenn δ im Satz 6.9 für mehr als einen Index ν angenommen wird.

Bestimmung einer Ausgangsbasislösung

Manchen Optimierungsaufgaben kann man eine Ausgangsecke ansehen. Sind die Nebenbedingungen ursprünglich in der Form $\boldsymbol{A}\,\boldsymbol{x} \leq \boldsymbol{b},\; \boldsymbol{x} \geq 0$ mit einem Vektor $\boldsymbol{b} \geq 0$ (komponentenweise) gegeben und $\boldsymbol{A} \in \mathbb{R}^{p\times n}$, so kann man durch Einführung von p Schlupfvariablen $y_i \geq 0,\; i = 1,\ldots,p$ die Nebenbedingungen auf Normalform bringen:

$$\left(\begin{array}{cccc} & 1 & & \\ \boldsymbol{A} & & 1 & \\ & & & \ddots \\ \underbrace{\quad}_{n} & & & \underbrace{\qquad\qquad 1}_{p} \end{array}\right) \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ y_1 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_p \end{pmatrix},$$

$$\boldsymbol{x} \;\geq\; 0, \qquad \boldsymbol{y} \;=\; (y_1, \ldots, y_p)^T \;\geq\; \boldsymbol{0}\,.$$

Dann ist Rang $(\boldsymbol{A}\,\boldsymbol{I}_p) = p$, $\boldsymbol{I}_p = p \times p$–Einheitsmatrix, und $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{pmatrix}$ ist wegen $\boldsymbol{b} \geq \boldsymbol{0}$ eine Ausgangsbasislösung bzw. Ecke des Problems (vgl. Satz 6.3, Def. 6.5).

Kann man für die Aufgabe in der Normalform

$$\left\{ \begin{array}{rcll} \boldsymbol{c}^T \boldsymbol{x} & \overset{!}{=} & \min\,, & \boldsymbol{c} \in \mathrm{I\!R}^n \\ \boldsymbol{A}\,\boldsymbol{x} & = & \boldsymbol{b}\,, & \boldsymbol{A} \in \mathrm{I\!R}^{m \times n}, \quad \boldsymbol{b} \in \mathrm{I\!R}^m\,, \quad \mathrm{Rg}\,\boldsymbol{A} = m \\ \boldsymbol{x} & \geq & \boldsymbol{0} & \end{array} \right. \tag{L}$$

keine Basislösung (Ecke) finden, so kann das Lösen eines Hilfsproblems, zu dem eine Ausgangsbasislösung bekannt ist, mit Hilfe des beschriebenen Simplex–Verfahrens Abhilfe schaffen. Es gilt

Satz 6.10

Für das Problem (L) mit $\boldsymbol{b} \geq 0$ (das ist keine Einschränkung) definieren wir mit $\boldsymbol{e} = (1, \ldots, 1)^T \in \mathrm{I\!R}^m$ das Hilfsproblem

$$\left\{ \begin{array}{rcll} \boldsymbol{e}^T \boldsymbol{y} & \overset{!}{=} & \min & \\ \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{y} & = & \boldsymbol{b}\,, & \boldsymbol{y} = (y_1, \ldots, y_m)^T \\ \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} & \geq & \boldsymbol{0} & \end{array} \right. \tag{$*$}$$

a) Der Vektor $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$, $\boldsymbol{x} = \boldsymbol{0}$, $\boldsymbol{y} = \boldsymbol{b}$ ist eine Basislösung von $(*)$, und $(*)$ hat eine optimale Basislösung $\begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}$.

b) Ist $\boldsymbol{y}^* \neq \boldsymbol{0}$, so hat (L) keine zulässigen Punkte, also auch keine Lösung.

c) Ist $\boldsymbol{y}^* = \boldsymbol{0}$, so wird durch $\boldsymbol{x}^*$ eine Ecke von (L) gegeben.

Beweis

a) $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{pmatrix}$ ist eine Ecke (wie oben), also besitzt $(*)$ zulässige Punkte und da die Zielfunktion nach unten beschränkt ist $(\boldsymbol{y} \geq \boldsymbol{0})$, existiert auch eine Lösung und damit auch eine Basislösung (Satz 6.6).

b) Hätte (L) einen zulässigen Punkt $\hat{\boldsymbol{x}}$, so wäre $\begin{pmatrix} \hat{\boldsymbol{x}} \\ \boldsymbol{0} \end{pmatrix} \in \mathrm{I\!R}^{n+m}$ ein zulässiger Punkt von $(*)$ mit Zielfunktionswert Null im Widerspruch zur Voraussetzung b).

c) Ist $\begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{0} \end{pmatrix}$ Lösung von $(*)$, so sind die zu positiven Komponenten von $\boldsymbol{x}^*$ gehörenden Spaltenvektoren von $\boldsymbol{A}$ linear unabhängig. Also ist $\boldsymbol{x}^*$ eine (möglicherweise entartete) Ecke von (L). ■

Eine weitere Möglichkeit zur Beschaffung einer Ausgangsecke und zur gleichzeitigen Lösung von (L) bietet

Satz 6.11

Für das Problem (L) mit $\boldsymbol{b} \geq \boldsymbol{0}$ definieren wir mit $\boldsymbol{e} = (1, \ldots, 1)^T \in \mathbb{R}^m$ und einem $S > 0$, $S \in \mathbb{R}$, das Hilfsproblem

$$\left\{ \begin{array}{lcl} \boldsymbol{c}^T \boldsymbol{x} + S\, \boldsymbol{e}^T \boldsymbol{y} & \stackrel{!}{=} & \min \\ \boldsymbol{A}\boldsymbol{x} + \boldsymbol{y} & = & \boldsymbol{b} \\ \boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} & \geq & \boldsymbol{0} \end{array} \right. \qquad (**)$$

a) Der Vektor $\boldsymbol{z} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{pmatrix}$ ist eine Basislösung von $(**)$ zur Indexmenge $\{n+1,\, n+2,\, \ldots,\, n+m\}$.

b) Ist $S > 0$ hinreichend groß, so gilt: Hat $(**)$ eine Lösung $\begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}$ mit $\boldsymbol{e}^T \boldsymbol{y}^* > 0$, so hat das Ausgangsproblem (L) keine zulässigen Vektoren, ist also unlösbar.

c) Ist $\begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}$ eine Lösung von $(**)$ mit $\boldsymbol{e}^T \boldsymbol{y}^* = 0$, so ist $\boldsymbol{x}^*$ auch Lösung von (L).

Beweis:

a) $\boldsymbol{z}$ ist Basislösung nach Def. 6.5.

b) Dies Aussage kann hier nicht bewiesen werden, da hierzu zusätzliche Kenntnisse notwendig sind.

c) Laut Voraussetzung ist $\boldsymbol{c}^T \boldsymbol{x}^* \leq \boldsymbol{c}^T \tilde{\boldsymbol{x}} + S\, \boldsymbol{e}^T \tilde{\boldsymbol{y}}$ für alle für $(**)$ zulässigen Punkte $\tilde{\boldsymbol{z}} = \begin{pmatrix} \tilde{\boldsymbol{x}} \\ \tilde{\boldsymbol{y}} \end{pmatrix}$. Ist $\overline{\boldsymbol{x}}$ zulässig für (L), so ist $\overline{\boldsymbol{z}} = \begin{pmatrix} \overline{\boldsymbol{x}} \\ \boldsymbol{0} \end{pmatrix}$ zulässig für $(**)$, d.h. $\boldsymbol{c}^T \boldsymbol{x}^* \leq \boldsymbol{c}^T \overline{\boldsymbol{x}}$, d.h. $\boldsymbol{x}^*$ löst (L). ■

Bemerkung:

Das in c) ausgerechnete $\boldsymbol{x}^*$ ist zur erhaltenen Indexmenge nicht notwendig ein Basisvektor für (L), wohl aber eine Ecke (vgl. dazu Collatz–Wetterling [71, p. 36–38]), denn $\begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}$ ist Basislösung von $(**)$ zu einer Indexmenge J^{**}, welche Indizes $> n$ enthalten kann.

Praktische Durchführung

Ein Iterationsschritt läßt sich jetzt folgendermaßen darstellen:

Sei eine m–elementige Indexmenge J, die zu einem Basisvektor $\boldsymbol{x}$ gehört, bekannt. Den Basisvektor $\boldsymbol{x}$ braucht man an dieser Stelle noch nicht zu kennen. Wir definieren zu dieser Indexmenge J die $(m \times m)$–Matrix

$$\boldsymbol{B} \;=\; (\boldsymbol{a}^i), \qquad i \;\in\; J, \qquad \boldsymbol{a}^i \;=\; i\text{–te Spalte von } \boldsymbol{A}\,, \tag{6.20}$$

die definitionsgemäß regulär ist. Sei $\boldsymbol{D} = (d_{ij}),\ i \in J,\ j = 1, 2, \ldots, n$ die *Tableaumatrix*, die früher (vgl. (6.9)) durch

$$\boldsymbol{a}^j \;=\; \sum_{i \in J} d_{ij}\, \boldsymbol{a}^i, \quad j \;=\; 1, 2, \ldots, n\,, \quad \text{mit} \quad d_{ij} \;=\; \delta_{ij} \text{ für } j \;\in\; J \tag{6.21}$$

definiert wurde. Mit Hilfe von $\boldsymbol{B}$ aus (6.20) lautet (6.21)

$$\boldsymbol{A} \;=\; \boldsymbol{B}\,\boldsymbol{D} \quad \Longleftrightarrow \quad \boldsymbol{A}^T \;=\; \boldsymbol{D}^T\,\boldsymbol{B}^T\,. \tag{6.22}$$

Wir setzen in Übereinstimmung mit (6.11)

$$\boldsymbol{s} = (s_1, s_2, \ldots, s_n)^T\,, \quad s_j = \sum_{i \in J} d_{ij}\, c_i\,, \quad j = 1, 2, \ldots, n, \quad \boldsymbol{c}^J = (c_i), i \in J\,. \tag{6.23}$$

Dann ist wegen (6.23)

$$\boldsymbol{s} \quad = \quad \boldsymbol{D}^T\, \boldsymbol{c}^J\,. \tag{6.24}$$

Wir benutzen zum Rechnen jetzt nur die in (6.20) vorhandene Information, nämlich J. Der Vektor $\boldsymbol{s}$ kann aus (6.22) und (6.24) ohne Benutzung von $\boldsymbol{D}$ ausgerechnet werden: Multiplizieren wir den rechten Teil von (6.22) mit einem beliebigen Vektor $\boldsymbol{y} \in \mathbb{R}^m$, so erhalten wir $\boldsymbol{A}^T\,\boldsymbol{y} = \boldsymbol{D}^T\,\boldsymbol{B}^T\,\boldsymbol{y}$. Bestimmen wir jetzt $\boldsymbol{y}$ so, daß $\boldsymbol{B}^T\,\boldsymbol{y} = \boldsymbol{c}^J$ ist, so ist $\boldsymbol{A}^T\,\boldsymbol{y} = \boldsymbol{D}^T\,\boldsymbol{c}^J = \boldsymbol{s}$, d.h. wir erhalten für $\boldsymbol{s}$ die Berechnungsvorschrift

$$\boldsymbol{B}^T\,\boldsymbol{y} \;=\; \boldsymbol{c}^J \qquad \Longleftrightarrow \qquad \boldsymbol{s} \;=\; \boldsymbol{A}^T\,\boldsymbol{y}\,. \tag{6.25}$$

Die *Pivotspalte* r erhält man daraus, sofern

$$t_r \;=\; c_r \;-\; s_r \;=\; \min_{j \notin J}\,(c_j - s_j) \;<\; 0\,; \tag{6.26}$$

man vgl. dazu (6.12a) und Satz 6.7. Die r–te Spalte $\boldsymbol{d}^r$ von $\boldsymbol{D}$, die in (6.18) bzw. Satz 6.9 benötigt wird, ergibt sich aus dem linken Teil von (6.22), nämlich

$$\boldsymbol{B}\,\boldsymbol{d}^r \quad = \quad \boldsymbol{a}^r\,, \tag{6.27}$$

wobei $\boldsymbol{a}^r$, wie üblich, die r–te Spalte von $\boldsymbol{A}$ bezeichnet. Danach kann Satz 6.8 abgeprüft werden. Gleichzeitig kann man den entsprechenden Basisvektor $\boldsymbol{x}$ aus

$$\boldsymbol{B}\,\boldsymbol{x} \quad = \quad \boldsymbol{b} \tag{6.28}$$

ausrechnen. Sind $\boldsymbol{d}^r = (d_{ir})$ und $\boldsymbol{x}$ bekannt, so kann nach der Formel (6.18) die *Pivotzeile* ν bestimmt werden und damit auch die neue Indexmenge

$$\overline{J} \quad = \quad (J \cup \{r\}) \setminus \{\nu\}\,. \tag{6.29}$$

Die fehlenden Komponenten von $\boldsymbol{x}$ müssen am Schluß durch Null ergänzt werden.

Dieses Verfahren ist für große m aufwendig, für kleine m (etwa ≤ 10) durchaus passabel. Eine Verringerung des Rechenaufwandes kann erreicht werden, wenn nicht dreimal das entsprechende Gleichungssystem pro Schritt neu gelöst wird, sondern einmalig in jedem Schritt eine Zerlegung von $\boldsymbol{B}$ (etwa nach dem Gaußschen Eliminationsverfahren) hergestellt wird. Haben wir eine Zerlegung der Form

$$\boldsymbol{F}\,\boldsymbol{B} \;=\; \boldsymbol{R}, \qquad \boldsymbol{R} \text{ ist rechte Dreiecksmatrix,} \qquad \boldsymbol{F} \text{ regulär} \tag{6.30}$$

hergestellt, so kann man die auftretenden Gleichungssysteme vom Typ

$$\text{a)} \quad \boldsymbol{B}^T\boldsymbol{y} \;=\; \boldsymbol{c}^J\,, \qquad\qquad \text{b)} \quad \boldsymbol{B}\,\boldsymbol{x} \;=\; \boldsymbol{b} \tag{6.31}$$

folgendermaßen lösen:

$$\text{a)} \quad \boldsymbol{R}^T\boldsymbol{z} \;=\; \boldsymbol{c}^J\,, \quad \boldsymbol{y} \;=\; \boldsymbol{F}^T\boldsymbol{z} \qquad \text{b)} \quad \boldsymbol{R}\boldsymbol{x} \;=\; \boldsymbol{F}\,\boldsymbol{b}\,. \tag{6.32}$$

Fall a) kann also durch Vorwärtseinsetzen, b) durch Rückwärtseinsetzen gelöst werden. In jedem Fall kommt eine Multiplikation einer Matrix mit einem Vektor hinzu.
Man beachte, daß $\mathbf{F}$ keine Dreiecksmatrix sein muß, wenn das GEV Zeilenvertauschungen enthält, insbesondere also bei der Durchführung mit Spaltenpivotsuche. Ist $\mathbf{I}_{ik}$ die Matrix, die aus der Einheitsmatrix $\mathbf{I}$ entsteht durch Vertauschen der i-ten und k-ten Spalte, so sind in $\mathbf{I}_{ik}\mathbf{A}$ in $\mathbf{A}$ die Zeilen i und k vertauscht. Durch $\mathbf{I}_{ik}$ enthält $\mathbf{F}$ Elemente unterhalb und oberhalb der Diagonalen.
Hat man speziell die Zerlegung $\boldsymbol{B} = \boldsymbol{L}\,\boldsymbol{R}$, $\boldsymbol{L} =$ linke Dreiecksmatrix, so lauten a) und b)

$$\text{a)} \quad \boldsymbol{R}^T\boldsymbol{z} \;=\; \boldsymbol{c}^J\,, \quad \boldsymbol{L}^T\boldsymbol{y} \;=\; \boldsymbol{z} \qquad \text{b)} \quad \boldsymbol{L}\,\boldsymbol{z} \;=\; \boldsymbol{b}\,, \quad \boldsymbol{R}\boldsymbol{x} \;=\; \boldsymbol{z}\,. \tag{6.33}$$

Eine weitere Verringerung der Anzahl der Operationen kann erreicht werden, wenn man berücksichtigt, daß zwei aufeinanderfolgende Matrizen $\boldsymbol{B}$, $\overline{\boldsymbol{B}}$, nämlich

$$\boldsymbol{B} \;=\; (\boldsymbol{a}^i)\,, \quad i \in J\,, \qquad \overline{\boldsymbol{B}} \;=\; (\boldsymbol{a}^j)\,, \quad j \in \overline{J} \;=\; (J \cup \{r\}) \setminus \{\nu\} \tag{6.34}$$

sich nur in einer Spalte unterscheiden, so daß auch die entsprechenden Zerlegungen $\boldsymbol{F}\boldsymbol{B} = \boldsymbol{R}$, $\overline{\boldsymbol{F}}\,\overline{\boldsymbol{B}} = \overline{\boldsymbol{R}}$ relativ leicht auseinander ausgerechnet werden können. Techniken zur Zerlegung von $\overline{\boldsymbol{B}}$ unter Benutzung der Zerlegung von $\boldsymbol{B}$ werden *Modifikationstechniken* (Glashoff–Gustafson [78, p. 165]) oder *Updatingstechniken* genannt.

Updating von Matrixzerlegungen

Sei $\boldsymbol{B} = (\boldsymbol{b}^1, \dots, \boldsymbol{b}^m)$ eine $m \times m$–Matrix mit den Spalten $\boldsymbol{b}^1, \dots, \boldsymbol{b}^m$. Bekannt sei die Zerlegung

$$\boldsymbol{F}\boldsymbol{B} \; = \; \boldsymbol{R}, \qquad \boldsymbol{R} \text{ rechte Dreiecksmatrix}, \qquad \boldsymbol{F} \text{ regulär}\,. \tag{6.35}$$

(Eine solche Zerlegung kann man z.B. durch Gauß–Elimination oder mit dem Householder–Verfahren gewinnen). Sei nun $\boldsymbol{b}^*$ ein weiterer vorgegebener Vektor und $\overline{\boldsymbol{B}}$ die Matrix, die aus $\boldsymbol{B}$ entsteht durch Streichen von $\boldsymbol{b}^p$ und Hinzufügen von $\boldsymbol{b}^*$ als letzte Spalte ($p \in \{1, \dots, m\}$):

$$\overline{\boldsymbol{B}} \; = \; \left(\boldsymbol{b}^1, \dots, \boldsymbol{b}^{p-1},\; \boldsymbol{b}^{p+1}, \dots, \boldsymbol{b}^m,\; \boldsymbol{b}^*\right)\,.$$

Gesucht ist eine (6.35) entsprechende Zerlegung für $\overline{\boldsymbol{B}}$, also

$$\overline{\boldsymbol{F}}\,\overline{\boldsymbol{B}} \; = \; \overline{\boldsymbol{R}}, \qquad \overline{\boldsymbol{R}} \text{ rechte Dreiecksmatrix}, \qquad \overline{\boldsymbol{F}} \text{ regulär}\,, \tag{6.36}$$

und zwar möglichst billig durch „Aufdatieren“ von (6.35).

Sehen wir uns zunächst die Matrix $\hat{\boldsymbol{R}} := \boldsymbol{F}\,\overline{\boldsymbol{B}}$ an! Es ist

$$\begin{aligned}
\hat{\boldsymbol{R}} \quad &= \quad \left(\boldsymbol{F}\boldsymbol{b}^1, \dots, \boldsymbol{F}\boldsymbol{b}^{p-1},\; \boldsymbol{F}\boldsymbol{b}^{p+1}, \dots, \boldsymbol{F}\boldsymbol{b}^m,\; \boldsymbol{F}\boldsymbol{b}^*\right) \\
&= \quad \begin{array}{c}
\begin{pmatrix}
\times & \times & \dots & \times & \times & \dots & & & \times \\
& \times & \dots & \times & \times & \dots & & & \times \\
& & \ddots & \vdots & \vdots & & & & \vdots \\
& & & \times & \times & & & & \vdots \\
& & & & \times & & & & \vdots \\
& & & & \times & \times & & & \vdots \\
& 0 & & & & \times & \ddots & & \vdots \\
& & & & & & \ddots & & \vdots \\
& & & & & & & \times & \times
\end{pmatrix} \\
\uparrow \\
p\text{–te Spalte}
\end{array}\,.
\end{aligned}$$

Die Matrix $\hat{\boldsymbol{R}}$ enthält als letzte Spalte den Vektor $\boldsymbol{F}\boldsymbol{b}^*$, davor die Spalten von $\boldsymbol{R}$ mit Ausnahme der p–ten Spalte. Die $m - p$ i.a. von 0 verschiedenen Matrixelemente unterhalb der Diagonalen von $\hat{\boldsymbol{R}}$ können nun leicht durch Gauß–Eliminationsschritte mit Spaltenpivotsuche zu 0 gemacht werden; dabei braucht man nur eine Vertauschung *benachbarter* Zeilen in Betracht zu ziehen. Einem solchen Eliminationsschritt entspricht eine Multiplikation von links mit einer Matrix $\boldsymbol{G}_i$, die entweder von der Form

$$
\begin{array}{c}
\begin{pmatrix}
1 & & & & & & & 0 \\
& \ddots & & & & & & \\
& & 1 & & & & & \\
& & & 1 & 0 & & & \\
& & & -\ell_{i+1} & 1 & & & \\
& & & & & 1 & & \\
& & & & & & \ddots & \\
0 & & & & & & & 1
\end{pmatrix}
\end{array}
\begin{array}{l}
\\ \\ \\ \leftarrow \quad i \\ \leftarrow \quad i+1 \\ \\ \\ \\
\end{array}
$$

(falls kein Zeilentausch stattfindet) oder von der Form

$$
\begin{array}{c}
\begin{pmatrix}
1 & & & & & & & 0 \\
& \ddots & & & & & & \\
& & 1 & & & & & \\
& & & 0 & 1 & & & \\
& & & 1 & -\ell_{i+1} & & & \\
& & & & & 1 & & \\
& & & & & & \ddots & \\
0 & & & & & & & 1
\end{pmatrix}
\end{array}
\begin{array}{l}
\\ \\ \\ \leftarrow \quad i \\ \leftarrow \quad i+1 \\ \\ \\ \\
\end{array}
$$

(falls ein Zeilentausch stattfindet) ist. Aufgrund der Pivotsuche ist $|\ell_{i+1}| \leq 1$. Nach $m-p$ Eliminationsschritten hat man

$$\boldsymbol{G}_{m-1} \ldots \boldsymbol{G}_p \hat{\boldsymbol{R}} = \overline{\boldsymbol{R}} \tag{6.37}$$

mit einer rechten Dreiecksmatrix $\overline{\boldsymbol{R}}$. Setzt man hierin $\hat{\boldsymbol{R}} = \boldsymbol{F}\overline{\boldsymbol{B}}$ ein, so erhält man mit

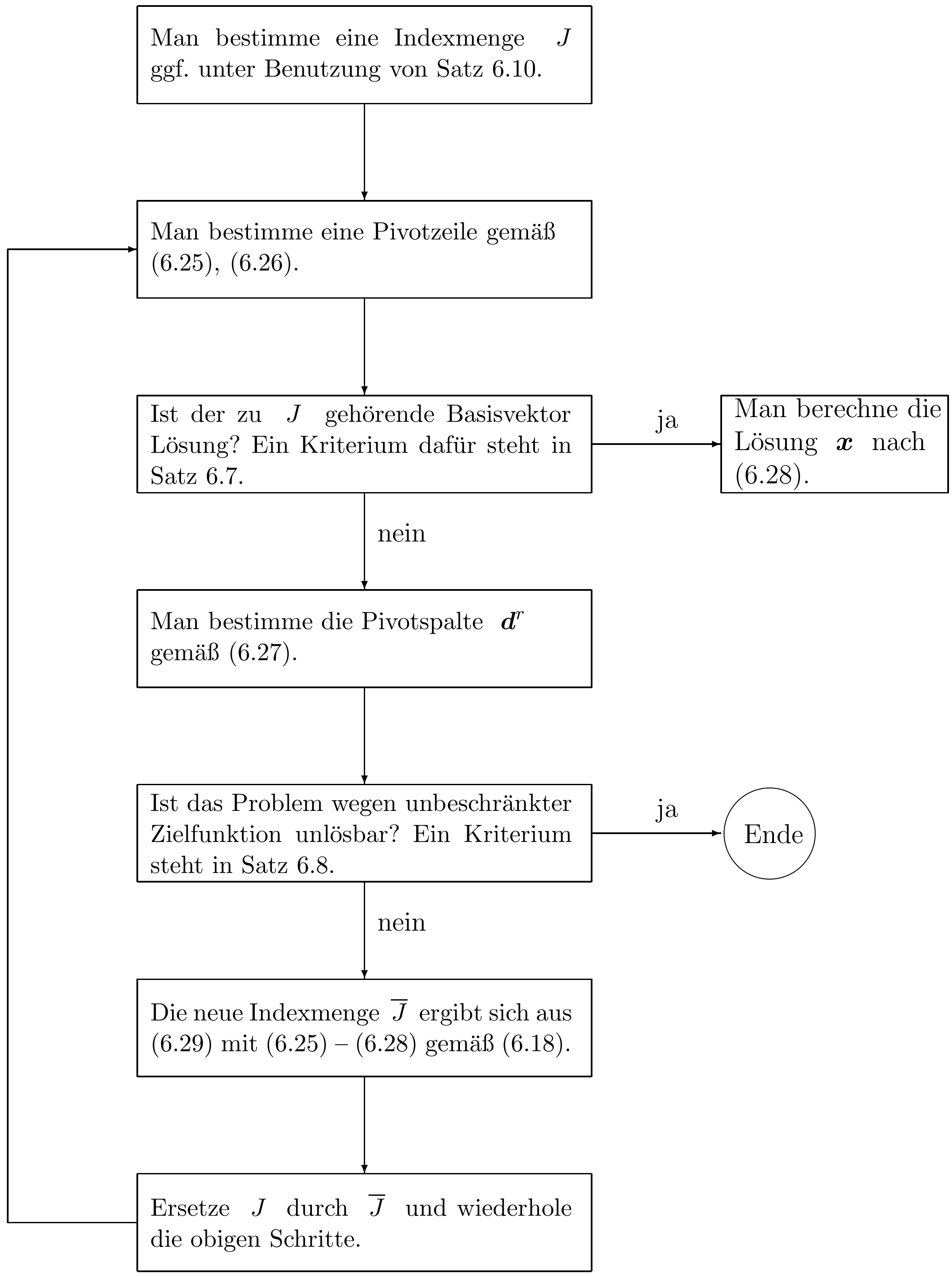
$$\boldsymbol{G}_{m-1} \ldots \boldsymbol{G}_p \boldsymbol{F} =: \overline{\boldsymbol{F}} \tag{6.38}$$

eine Zerlegung der gewünschten Form (6.36).

Für die numerische Rechnung kann man aus (6.37) und (6.38) ablesen, daß man $\overline{\boldsymbol{R}}$ und $\overline{\boldsymbol{F}}$ dadurch erhält, daß man die Eliminationsschritte *simultan* auf $\hat{\boldsymbol{R}}$ und $\boldsymbol{F}$ anwendet. Während das direkte Ausrechnen einer Zerlegung (6.36) durch Gauß–Elimination ca. $c \cdot m^3$ Multiplikationen und Divisionen kostet, verbraucht das beschriebene Aufdatieren von (6.35) lediglich ca. $(m-p)^2$ Operationen.

Bemerkung:
Nach Abschnitt “Bestimmung einer Ausgangsbasislösung” ist die Ausgangsmatrix B in (6.35) oft die Einheitsmatrix, wodurch die Zerlegung (6.35) trivialerweise gegeben ist.

Schematisch lassen sich die auszuführenden Schritte so zusammenfassen:



Übersicht: Rechenschritte für das Simplexverfahren

Beispiel: Wir lösen $\boldsymbol{c}^T\boldsymbol{x} \overset{!}{=} \min$, $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$, $\boldsymbol{x} \geq \boldsymbol{0}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 30 & 60 & 0 & 1 & 0 \\ 2 & 10 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 1200 \\ 42000 \\ 5200 \end{pmatrix}, \quad \boldsymbol{c} = \begin{pmatrix} -120 \\ -360 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

nach dem angegebenen Verfahren.

Erster Schritt: $J = \{3,4,5\}$, $\boldsymbol{B} = \boldsymbol{I} =$ Einheitsmatrix, $\boldsymbol{c}^J = (c_3, c_4, c_5)^T = (0,0,0)^T$. Also hat nach (6.25) das System $\boldsymbol{B}^T\boldsymbol{y} = \boldsymbol{c}^J$ die Lösung $\boldsymbol{y} = \boldsymbol{0}$ und somit $\boldsymbol{s} = \boldsymbol{0}$ und $t_2 = \min\limits_{j=1,2}(c_j - 0) = -360$, d.h. $r = 2$. Aus (6.27) folgt $\boldsymbol{B}\,\boldsymbol{d} = \boldsymbol{a}^2 = (1,60,10)^T$, d.h. $\boldsymbol{d} = \boldsymbol{a}^2$ und aus (6.28) folgt $\boldsymbol{x} = \boldsymbol{b} = (1200, 42000, 5200)^T$. Nach (6.18) erhält man $\delta = \min\limits_{j\in J}\left\{\frac{x_j}{d_{j2}} : d_{j2} > 0\right\} = \frac{x_5}{d_{52}} = 520$, und somit $\nu = 5$.

Zweiter Schritt: $J = \{3,4,5\} \cup \{r\} \backslash \{\nu\} = \{3,4,2\}$, $\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 60 \\ 0 & 0 & 10 \end{pmatrix}$,
$\boldsymbol{c}^J = (0,0,-360)^T$, $\boldsymbol{B}^T\boldsymbol{y} = \boldsymbol{c}^J$ hat die Lösung $\boldsymbol{y} = (0,0,-36)^T$ und daraus folgt $\boldsymbol{s} = \boldsymbol{A}^T\boldsymbol{y} = (-72,-360,0,0,-36)^T$ und $t_1 = \min\limits_{j=1,5}(c_j - s_j) = -48$, $r = 1$. Aus $\boldsymbol{B}\,\boldsymbol{d} = \boldsymbol{a}^1 = (1,30,2)^T$ folgt $\boldsymbol{d} = (4/5, 18, 1/5)^T$, und aus $\boldsymbol{B}\,\boldsymbol{x} = \boldsymbol{b}$ folgt $\boldsymbol{x} = (680, 10800, 520)^T$ und $\delta = \min\limits_{j\in J}\left\{\frac{x_j}{d_{j1}} : d_{j1} > 0\right\} = \frac{x_4}{d_{41}} = 600$, also $\nu = 4$.

Dritter Schritt: $J = \{3,1,2\}$, $\boldsymbol{B} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 30 & 60 \\ 0 & 2 & 10 \end{pmatrix}$, $\boldsymbol{c}^J = (0,-120,-360)^T$,
$\boldsymbol{B}^T\boldsymbol{y} = \boldsymbol{c}^J$ hat die Lösung $\boldsymbol{y} = (0,-8/3,-20)^T$ und daraus folgt $\boldsymbol{s} = \boldsymbol{A}^T\boldsymbol{y} = (-120,-360,0,-8/3,-20)^T$ und alle $t_j \geq 0$, $j = 4,5$, d.h., wir sind bei der Lösung angekommen, die sich aus $\boldsymbol{B}\,\boldsymbol{x} = \boldsymbol{b}$ zu $\boldsymbol{x} = (x_3, x_1, x_2)^T = (200, 600, 400)^T$ berechnet. Die endgültige Lösung ist damit $\tilde{\boldsymbol{x}} = (600, 400, 200, 0, 0)^T$ und $\boldsymbol{c}^T\tilde{\boldsymbol{x}} = -216000$. Man beachte, daß die Indexmengen J zweckmäßigerweise nicht nach der Größe der Indizes geordnet werden sollten (bei der Updating–Methode ist entsprechend der Aufdatierung von (6.35) vorzugehen, d. h. ν entfernen, die Elemente aufrücken lassen und an letzter Stelle r hinzufügen).

§ 7 Spline–Interpolation

Wir kehren nochmals zur Interpolationsaufgabe aus § 3 zurück. Eine reellwertige Funktion f soll durch eine (mindestens einmal stetig differenzierbare) Funktion, welche f in den Punkten $(x_j, f_j), f_j = f(x_j)$, $j = 0, 1, \ldots, n$ interpoliert, approximiert werden. Die Untersuchungen des § 3 zur Fehlerabschätzung haben gezeigt, daß bei Polynominterpolation — insbesondere bei vielen Stützstellen — die Fehlerfunktion starke Ausschläge besitzt. Insbesondere garantiert die Fehlerabschätzung (3.14) nicht notwendig die Konvergenz der approximierenden Funktion gegen f, wenn die Anzahl der Stützstellen gegen Unendlich strebt.

Daher liegt die Idee nahe — analog zur numerischen Integration mit Hilfe zusammengesetzter Formeln —, eine Interpolationsfunktion $S(x)$ zu konstruieren durch Zusammenstückeln von Polynomen niedrigeren Grades in den einzelnen Teilintervallen. Man erhält so eine „Spline–Funktion“ (genaue Definition folgt noch). Insbesondere kubische Splines (auf die wir uns hier beschränken wollen), die durch Zusammensetzung von kubischen Polynomen entstehen, haben sich als nützlich erwiesen.

Seien also die Punkte (x_j, f_j), und eine Unterteilung Δ_n gegeben:

$$\Delta_n \quad : \quad a \;=\; x_0 \;<\; x_1 \;<\; x_2 \;<\; \ldots \;<\; x_{n-1} \;<\; x_n \;=\; b\,. \tag{7.1}$$

Dadurch werden Teilintervalle I_j der Länge h_j definiert.

$$I_j \;=\; [x_{j-1}, x_j]\,, \qquad h_j \;=\; x_j - x_{j-1}\,, \qquad j = 1, \ldots, n\,. \tag{7.2}$$

In I_j wählen wir den Ansatz

$$s_j(x) \;=\; a_j + b_j(x - x_{j-1}) + c_j(x - x_{j-1})^2 + d_j(x - x_{j-1})^3\,, \quad j = 1, \ldots, n\,. \tag{7.3}$$

Bei n Teilintervallen haben wir also $4n$ Unbekannte a_j, b_j, c_j, d_j zu bestimmen. Wir überlegen zunächst, welche Stetigkeits– und Differenzierbarkeitsbedingungen wir an die durch die s_j zusammengesetzte Funktion $S(x)$ stellen können.

Die Interpolationsbedingungen für S

$$s_j(x_{j-1}) \;=\; f_{j-1}, \qquad s_j(x_j) \;=\; f_j, \qquad j = 1, \ldots, n\,, \tag{7.4}$$

liefern $2n$ Gleichungen und garantieren die Stetigkeit von S.

Die Stetigkeit der Ableitung an den inneren Punkten, also $S \in C^1[a, b]$,

$$s'_j(x_j) \;=\; s'_{j+1}(x_j)\,, \qquad j = 1, \ldots, n-1\,, \tag{7.5}$$

liefert weitere $n-1$ Gleichungen. Also kann man sogar die Stetigkeit der zweiten Ableitung — $S \in C^2[a, b]$ — fordern.

$$s''_j(x_j) \;=\; s''_{j+1}(x_j)\,, \qquad j = 1, \ldots, n-1\,. \tag{7.6}$$

Wir haben also noch $4n - (2n + 2(n-1)) = 2$ Bedingungen frei. Sie werden üblicherweise benutzt, um entweder die ersten oder zweiten Ableitungen von S an den Intervallenden a, b vorzuschreiben.

Definition 7.1

a) Ein *kubischer Spline* zur Unterteilung Δ_n (vgl. (7.1)) ist eine reelle Funktion $S : [a, b] \to \mathbb{R}$ mit den Eigenschaften

   (i) $S \in C^2[a, b]$,

   (ii) auf jedem Teilintervall $[x_{j-1}, x_j]$, $j = 1, \dots, n$ stimmt S mit einem Polynom s_j 3. Grades überein.

b) Ein Spline, der (7.4) erfüllt, heißt *interpolierender Spline*.

c) Interpolierende Splines, die je eine der folgenden Zusatzbedingungen erfüllen, haben besondere Namen:

(7.7) *natürlicher Spline* falls $S''(a) = S''(b) = 0$,

(7.8) *periodischer Spline* falls $S^{(i)}(a) = S^{(i)}(b), \quad i = 0, 1, 2$,

(7.9) *allgemeiner Spline* falls (z.B.) $S'(a) = f'_0, S'(b) = f'_n$ für $f'_0, f'_n \in \mathbb{R}$.

Beachte:

(7.8) ist nur sinnvoll, wenn die Periodizitätsbedingung $S(a) = S(b)$, d.h. $f_0 = f_n$, bereits in (7.4) enthalten ist.

Korrekterweise müßte man einen die Funktion f interpolierenden kubischen Spline zur Unterteilung Δ_n zum Beispiel mit $S_3(\Delta_n, f; x)$ bezeichnen. Wir bleiben der Einfachheit halber, da keine Mißverständnisse zu befürchten sind, bei der einfacheren Bezeichnung $S(x)$.

Bemerkung:

Natürliche Splines lassen sich aus der folgenden physikalischen Motivation, die aus dem Schiffbau stammt, herleiten (vgl. Schwarz [88, § 3.7.1]):

Durch Punkte (x_i, f_i), $i = 0, \dots, n$, werde eine dünne homogene, elastische Latte gelegt, die in den Stützpunkten gelenkig gelagert sei und dort keinen äußeren Kräften unterliegt. Die Biegelinie $S(x)$ dieser Latte (dieses Splines) bestimmt sich so, daß die Deformationsenergie der Latte durch die angenommene Form minimiert wird. Unter leicht vereinfachenden Annahmen kann man zeigen, daß diese Biegelinie ein natürlicher interpolierender Spline ist. Eine etwas andere Formulierung dieses Extremalprinzips (der natürliche Spline als interpolierende Kurve mit minimaler Krümmung) findet man in Stoer [89, § 2.4.1].

Berechnung kubischer Splines: Die Momentenmethode

Natürlich könnte man die Konstanten a_j, b_j, c_j, d_j aus (7.3) bestimmen durch die Lösung des z.B. durch die Bedingungen (7.4)–(7.7) gegebenen Gleichungssystems für die $4n$ Unbekannten. Praktischer ist es jedoch, die Konstanten a_j, b_j, c_j, d_j durch die sogenannten

$$\text{Momente} \quad M_j \;=\; S''(x_j)\,, \qquad j = 0, \ldots, n\,,$$

auszudrücken, die zunächst natürlich noch unbekannt sind, und danach ein Gleichungssystem für die $n+1$ Unbekannten M_j zu lösen. Dies scheint auf den ersten Blick verwunderlich, ergibt sich aber von selbst, wenn man überlegt, daß S'' in jedem Teilintervall eine Gerade ist, die dargestellt werden kann durch

$$\begin{aligned} s_j''(x) \;&=\; M_{j-1} + \frac{M_j - M_{j-1}}{h_j}\,(x - x_{j-1})\,, \\ h_j \;&=\; x_j - x_{j-1}, \quad x \in [x_{j-1}, x_j]\,, \quad j = 1, \ldots, n\,. \end{aligned} \tag{7.10}$$

Dieser Ansatz genügt bereits den Bedingungen (7.6).

Wir werden diese Gleichungen hochintegrieren, die Integrationskonstanten mittels der Interpolationsbedingungen (7.4) bestimmen und schließlich ein Gleichungssystem zur Berechnung der M_j aufstellen mit Hilfe der Stetigkeitsforderung (7.5) für die ersten Ableitungen.

Durch Integration von (7.10) erhält man mit Integrationskonstanten A_j, B_j

$$s_j'(x) \;=\; B_j \;+\; M_{j-1}(x - x_{j-1}) \;+\; \frac{M_j - M_{j-1}}{2h_j}\,(x - x_{j-1})^2\,, \tag{7.11}$$

$$s_j(x) \;=\; A_j \;+\; B_j(x - x_{j-1}) \;+\; \frac{M_{j-1}}{2}\,(x - x_{j-1})^2 \;+\; \frac{M_j - M_{j-1}}{6h_j}\,(x - x_{j-1})^3\,. \tag{7.12}$$

Wegen $s_j(x_{j-1}) = f_{j-1}$, $s_j(x_j) = f_j$, vgl. (7.4), erhält man für A_j und B_j die Gleichungen

$$\begin{aligned} A_j \;&=\; f_{j-1}\,, \\ B_j \;&=\; \frac{f_j - f_{j-1}}{h_j} \;-\; \frac{h_j}{6}\,(M_j + 2M_{j-1})\,. \end{aligned} \tag{7.13}$$

Setzt man diese Werte in (7.12) ein, so erhält man für s_j die Darstellung

$$\left\{ \begin{aligned} s_j(x) \;&=\; a_j \;+\; b_j(x - x_{j-1}) \;+\; c_j(x - x_{j-1})^2 \;+\; d_j(x - x_{j-1})^3\,, \\ a_j \;&=\; f_{j-1}\,, \\ b_j \;&=\; \frac{f_j - f_{j-1}}{h_j} \;-\; \frac{2M_{j-1} + M_j}{6}\,h_j\,, \\ c_j \;&=\; \frac{M_{j-1}}{2}\,, \\ d_j \;&=\; \frac{M_j - M_{j-1}}{6h_j}\,, \qquad x \;\in\; [x_{j-1}, x_j]\,, \quad j = 1, \ldots, n\,. \end{aligned} \right. \tag{7.14}$$

Die Stetigkeit von S' in den inneren Punkten: $s'_j(x_j) = s'_{j+1}(x_j)$, vgl. (7.5), bedeutet gemäß (7.11)

$$B_j + \frac{M_j + M_{j-1}}{2} h_j = B_{j+1}, \qquad j = 1, \ldots, n-1 .$$

Durch Einsetzen von B_j, B_{j+1} gemäß (7.13) erhält man

$$\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{f_{j+1} - f_j}{h_{j+1}} - \frac{f_j - f_{j-1}}{h_j}, \quad j = 1, \ldots, n-1 . \tag{7.15}$$

Dies sind $n-1$ Gleichungen für die $n+1$ Unbekannten $M_0, \ldots, M_n$. Je zwei weitere Gleichungen liefert jede der Bedingungen (7.7) bis (7.9). (Beachte zu (7.8): $S(a) = S(b)$ muß schon in (7.4) enthalten sein.)

Multipliziert man (7.15) mit $\frac{6}{h_j + h_{j+1}}$, so erhält man mit den Abkürzungen

$$\left\{ \begin{array}{l} \lambda_j = \dfrac{h_{j+1}}{h_j + h_{j+1}}, \qquad \mu_j = 1 - \lambda_j = \dfrac{h_j}{h_j + h_{j+1}}, \\[2ex] d_j = \dfrac{6}{h_j + h_{j+1}} \left\{ \dfrac{f_{j+1} - f_j}{h_{j+1}} - \dfrac{f_j - f_{j-1}}{h_j} \right\}, \qquad j = 1, \ldots, n-1, \end{array} \right. \tag{7.16}$$

das Gleichungssystem

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j, \qquad j = 1, \ldots, n-1, \tag{7.17}$$

Definiert man zusätzlich im Fall (7.7) für $j = 0$ und $j = n$

$$\lambda_0 = 0, \quad d_0 = 0, \quad \mu_n = 0, \quad d_n = 0, \tag{7.18}$$

bzw. im Fall (7.9)

$$\begin{array}{l} \lambda_0 = 1, \quad d_0 = \dfrac{6}{h_1} \left(\dfrac{f_1 - f_0}{h_1} - f'_0 \right), \quad \mu_n = 1, \\[2ex] d_n = \dfrac{6}{h_n} \left(f'_n - \dfrac{f_n - f_{n-1}}{h_n} \right), \end{array} \tag{7.19}$$

so sind (7.7) bzw. (7.9) jeweils äquivalent zu den beiden Gleichungen

$$\begin{aligned} 2M_0 + \lambda_0 M_1 &= d_0, \\ \mu_n M_{n-1} + 2M_n &= d_n . \end{aligned}$$

Für die Fälle (7.7), (7.9) erhält man insgesamt also ein Gleichungssystem für die M_j mit einer sogenannten *Tridiagonalmatrix*

$$\begin{pmatrix} 2 & \lambda_0 & & & \\ \mu_1 & 2 & \lambda_1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & 2 & \lambda_{n-1} \\ & & & \mu_n & 2 \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ \\ M_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ \\ d_n \end{pmatrix} . \tag{7.20}$$

Im periodischen Fall (7.8) lautet die 1. Gleichung aus (7.17) wegen $M_0 = M_n$ (aus $S''(a) = S''(b)$)

$$2M_1 \; + \; \lambda_1 \, M_2 \; + \; \mu_1 \, M_n \;\; = \;\; d_1 \, .$$

Weiter wird $f_n = f_0$ vorausgesetzt, es bleibt also $s_1'(x_0) = s_n'(x_n)$ einzuarbeiten. Nun folgt aus (7.14) wegen $f_0 = f_n, \; M_0 = M_n$

$$\begin{aligned} s_1'(x_0) \;\; &= \;\; \frac{f_1 - f_n}{h_1} \quad - \quad \frac{2M_n + M_1}{6} \, h_1 \, , \\ s_n'(x_n) \;\; &= \;\; \frac{f_n - f_{n-1}}{h_n} \quad + \quad \frac{2M_n + M_{n-1}}{6} \, h_n \, . \end{aligned}$$

Gleichsetzen liefert (nach etwas Rechnung)

$$\frac{h_1}{h_1 + h_n} \, M_1 \; + \; \frac{h_n}{h_1 + h_n} \, M_{n-1} \; + \; 2M_n \;\; = \;\; \frac{6}{h_1 + h_n} \left(\frac{f_1 - f_n}{h_1} \; - \; \frac{f_n - f_{n-1}}{h_n} \right) \, .$$

Definiert man also im periodischen Fall zusätzlich

$$\begin{aligned} \lambda_n \;\; &= \;\; \frac{h_1}{h_1 + h_n} \, , \quad \mu_n \;\; = \;\; 1 - \lambda_n \;\; = \;\; \frac{h_n}{h_1 + h_n} \, , \\ d_n \;\; &= \;\; \frac{6}{h_1 + h_n} \left(\frac{f_1 - f_n}{h_1} \; - \; \frac{f_n - f_{n-1}}{h_n} \right) \, , \end{aligned} \tag{7.21}$$

so kann man obige Gleichung als weitere Bestimmungsgleichung zu (7.17) für $j = n$ hinzunehmen und erhält damit das lineare Gleichungssystem für M_1 bis M_n:

$$\begin{pmatrix} 2 & \lambda_1 & & & \mu_1 \\ \mu_2 & 2 & \lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & \lambda_{n-1} \\ \lambda_n & & & \mu_n & 2 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ \\ M_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ \\ d_n \end{pmatrix} . \tag{7.22}$$

Wenn wir nun noch zeigen, daß die Gleichungssysteme (7.20), (7.22) eindeutig lösbar sind, sind wir fertig. Dazu beachten wir, daß die λ_j, μ_j nur von der Zerlegung Δ_n abhängen und nicht von den f_j, und daß für alle j gilt

$$\lambda_j \;\geq\; 0, \qquad \mu_j \;\geq\; 0, \qquad \lambda_j + \mu_j \;=\; 1 \quad \text{oder} \quad = \; 0 \, .$$

Lemma 7.2
Für jede Unterteilung Δ_n gemäß (7.1) sind die Koeffizientenmatrizen von (7.20) und (7.22) nicht singulär.

Beweis:
Wir beschränken uns auf den Fall (7.20). Der Fall (7.22) wird analog gezeigt.

Bezeichne $\boldsymbol{A}$ die Koeffizientenmatrix aus (7.20), so zeigen wir

$$\boldsymbol{A}\boldsymbol{x} \;=\; \boldsymbol{0} \quad\Longrightarrow\quad \boldsymbol{x} \;=\; \boldsymbol{0} \quad \forall\, \boldsymbol{x}\,.$$

(Ist ein homogenes Gleichungssystem nur trivial lösbar, so ist die Koeffizientenmatrix regulär.)

Insbesondere für die betragsmaximale Komponente x_r von $\boldsymbol{x}$, $|x_r| = \max\limits_{j=0,\ldots,n} |x_j|$, gilt

$$|\mu_r\, x_{r-1} + 2x_r + \lambda_r\, x_{r+1}| \;=\; 0 \qquad (\mu_0 \;=\; 0, \quad \lambda_n \;=\; 0)\,.$$

Mit der Dreiecksungleichung folgt (beachte $\mu_j, \lambda_j \geq 0$)

$$0 \;\geq\; 2|x_r| \;-\; \mu_r|x_{r-1}| \;-\; \lambda_r|x_{r+1}| \quad \text{und da} \quad |x_{r-1}|,\, |x_{r+1}| \;\leq\; |x_r|$$

$$0 \;\geq\; 2|x_r| \;-\; (\mu_r + \lambda_r)|x_r| \;=\; (2 - \underbrace{(\mu_r + \lambda_r)}_{0 \text{ oder } 1})|x_r| \;\geq\; |x_r|$$

und damit $|x_j| = 0 \;\forall\, j$. ■

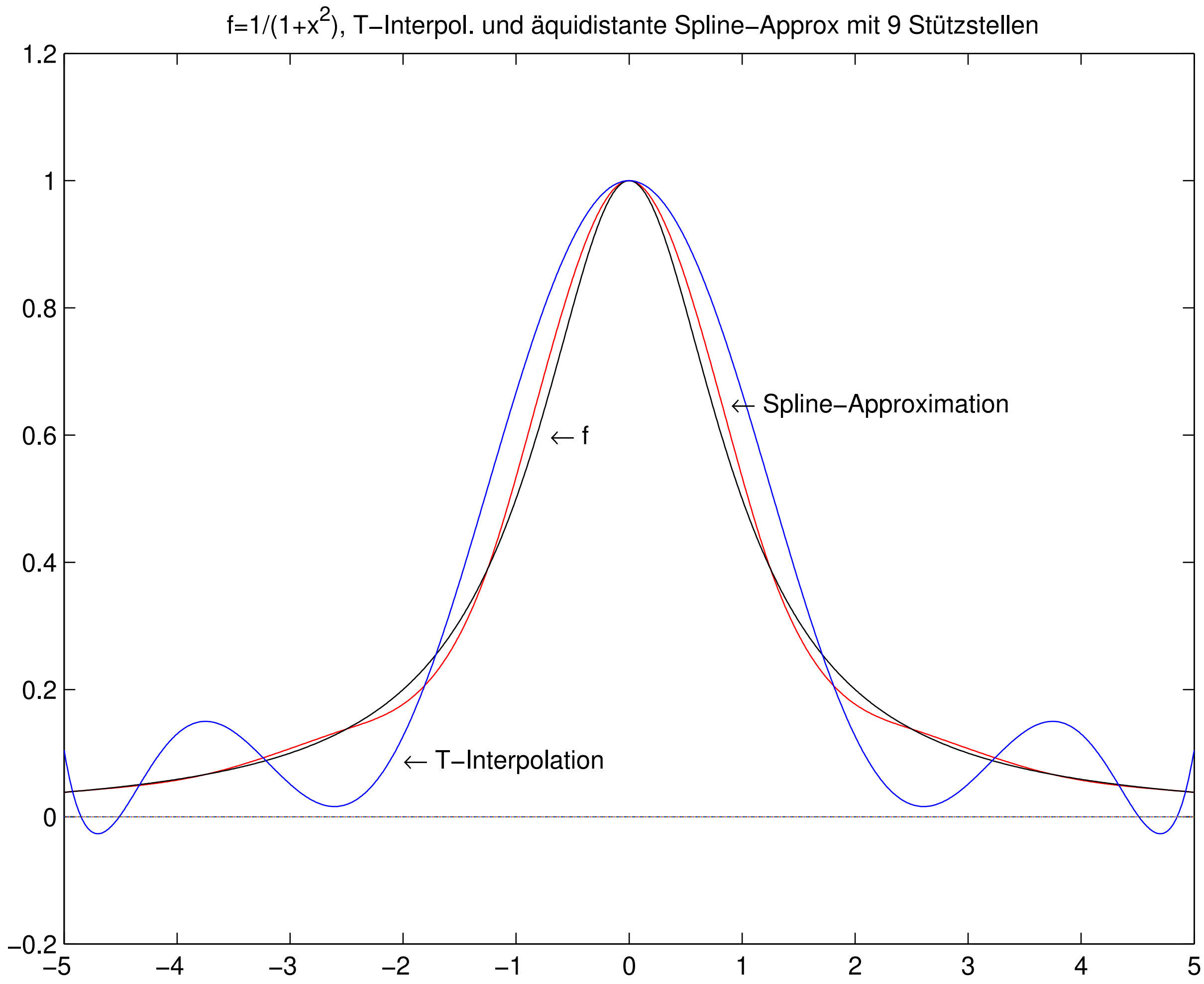
Bemerkungen:
Für Funktionen $f \in C^4[a,b]$ genügt der interpolierende Spline $S(x)$ mit der Zusatzbedingung (7.9) bei äquidistanten Unterteilungen Δ_n (d.h. $h = x_{i+1} - x_i \;\forall\, i$) folgender Fehlerabschätzung

$$|f(x) - S(x)| \quad\leq\quad \left(\frac{5}{384}\right) \cdot\, h^4 \max_{x\in[a,b]} \left|f^{(4)}(x)\right|\,,$$

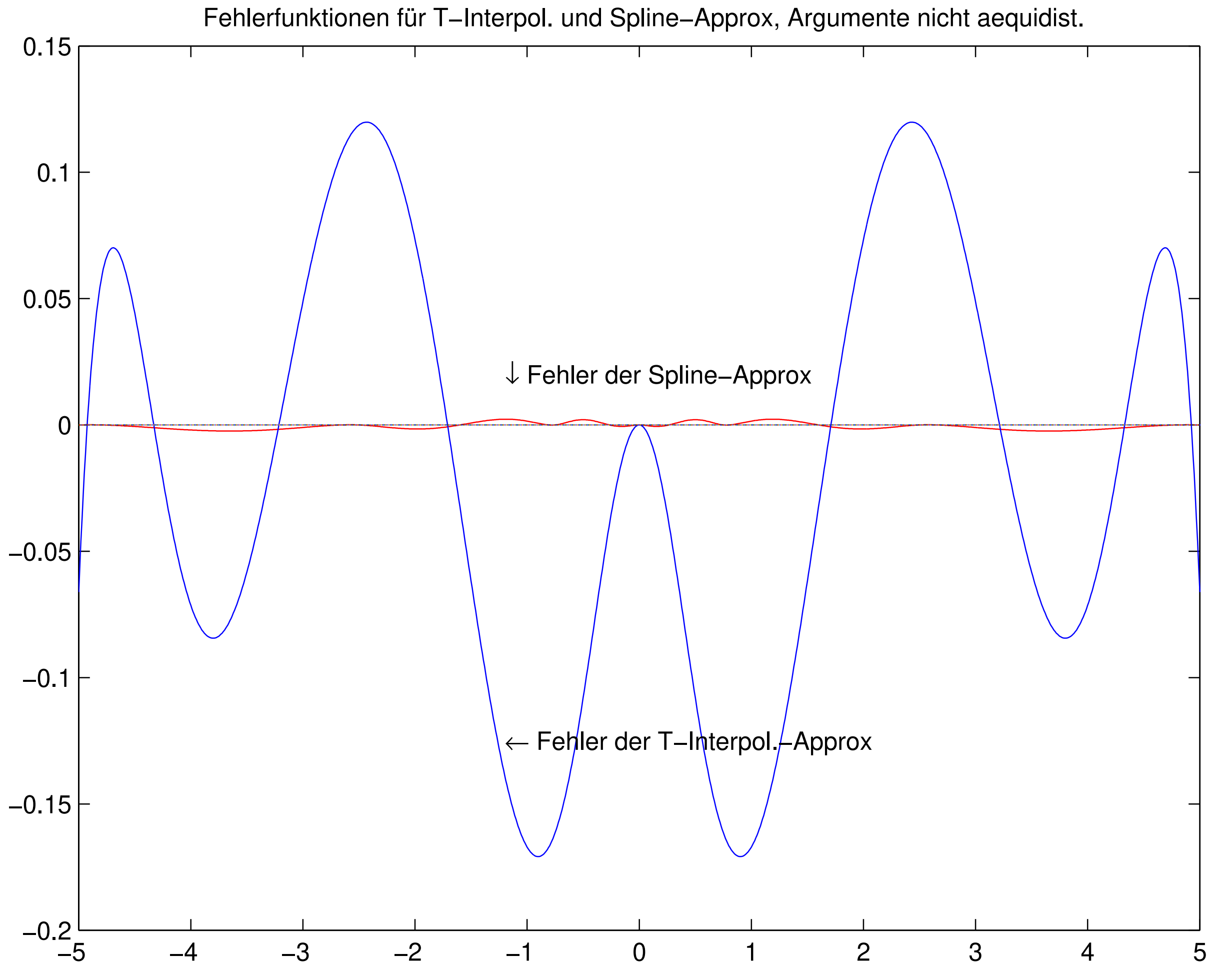
vgl. Stoer [89, § 2.4.3]. Hieraus folgt, daß die Splineinterpolierende bei Gitterverfeinerung von 4. Ordnung gegen die Funktion f konvergiert.

Durch Umformungen kann man statt (7.20) und (7.22) auch Gleichungssysteme mit symmetrischen Matrizen erhalten, die man mit einer Variante des Gaußschen Eliminationsverfahrens lösen kann (Cholesky–Verfahren). Diese Matrizen haben i. allg. eine schlechtere Kondition als die Matrizen aus (7.20), (7.22), sind also beim Lösen der Gleichungssysteme rundungsfehleranfälliger (vgl. Werner 1, Kap. 3.3.3, Bemerkung 3.7).

Zum Abschluß vergleichen wir die Approximationsgüte der kubischen Spline-Approximation mit der der Tschebyschev-Interpolation am Beispiel der Funktion $f(x) = 1/(1+x^2)$, zunächst für äquidistante Spline-Stützstellen.



Werden die Stützstellen optimal gewählt (z.B. mit Hilfe der Matlab-Prozedur *fminsearch*), so kann man in der Zeichnung Funktion und Spline kaum mehr unterscheiden. Der Fehler der Spline-Approximation im gesamten Intervall ist dann $< 2.2 * 10^{-3}$.
In einem weiteren Bild vergleichen wir den Fehler der T-Interpolation mit dem der optimalen Spline-Approximation.


Fehlerfunktionen für T−Interpol. und Spline−Approx, Argumente nicht aequidist.
↓ Fehler der Spline−Approx
← Fehler der T−Interpol.−Approx
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
−5
−4
−3
−2
−1
0
1
2
3
4
5

§ 8 Normen und Skalarprodukte

Dieser Paragraph dient vor allem der Beschaffung weiteren mathematischen Handwerkszeuges. Wir wollen die Notwendigkeit hierfür an 4 ersten Beispielen, die in späteren Kapiteln wieder aufgegriffen werden, demonstrieren.

Beispiel 1: Man berechne die Schnittpunkte zweier Ellipsen

$$\begin{aligned} a_1 x^2 + b_1 xy + c_1 y^2 + d_1 x + e_1 y &= f_1 , \\ a_2 x^2 + b_2 xy + c_2 y^2 + d_2 x + e_2 y &= f_2 , \quad a_i , b_i , c_i , d_i , f_i, e_i \in \mathbb{R} . \end{aligned}$$

Diese Gleichungen sind nicht elementar nach x, y auflösbar. Man wird Iterationsverfahren konstruieren müssen, die eine Folge $\{(x_n, y_n)\}_{n\in\mathbb{N}}$ von Näherungslösungen liefern. Natürlich möchte man wissen, „wie gut“ diese Näherungslösungen sind und ob sie gegen die exakte Lösung konvergieren. D.h. man benötigt einen Abstandsbegriff für Punkte im $\mathbb{R}^2$ (allgemeiner im $\mathbb{R}^n$, $\mathbb{C}^n$) ($\rightarrow$ Iterationsverfahren).

Beispiel 2: Wir haben schon in §3 untersucht, wie gut ein Interpolationspolynom p_n eine Funktion f in einem Intervall approximiert. Zur weiteren Untersuchung dieses Problems braucht man einen Abstandsbegriff für Funktionen, der besagt, wie weit p_n von f weg ist.

Die Schwierigkeiten dieser Beispiele bewältigen wir durch die Einführung eines komfortablen Abstandbegriffs für Elemente x, y eines linearen Raumes, nämlich der *Norm* ihrer Differenz. Bezeichnung: $\|x - y\|$.

Definition 8.1
Sei X ein linearer Raum (Vektorraum) über K ($K = \mathbb{R}$ oder $K = \mathbb{C}$). Eine Abbildung $\| \ \| : X \to \mathbb{R}$ heißt *Norm auf* X (*Vektornorm*), wenn sie folgende Eigenschaften hat:

(i) $\|x\| \geq 0 \ \forall\, x \in X, \quad \|x\| = 0 \ \Leftrightarrow \ x = 0,$ (*Definitheit*),

(ii) $\|\alpha\, x\| = |\alpha|\, \|x\| \ \forall\, \alpha \in K,$ (*schwache Homogenität*),

(iii) $\|x + y\| \leq \|x\| + \|y\|,$ (*Dreiecksungleichung*).

Ein linearer Raum X, in dem eine Norm definiert ist, heißt *normierter* Raum.
Bezeichnung: $(X, \| \ \|)$.

Bemerkung:
Anschaulich beschreibt $\|x\|$ die Entfernung eines Punktes x vom Nullpunkt und $\|a-b\|$ die Entfernung der Punkte a und b. $d(x, y) := \|x-y\|$ ist eine Metrik (vgl. Forster II, § 1).

Beispiele von Normen:

$$X = \mathbb{R}^1 \ : \qquad \|x\| = |x|\,,$$

$$X = \mathbb{R}^n \text{ oder } \mathbb{C}^n \ : \quad \text{für} \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$$

heißt

$$\|\boldsymbol{x}\|_\infty = \max_{j=1,\ldots,n} |x_j| \qquad \textit{Maximumnorm}, \tag{8.1}$$

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^{n} |x_j|^2} \qquad \textit{Euklidische Norm},$$

$$\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{j=1}^{n} |x_j|^p}, \quad p \in \mathbb{N} \qquad \textit{Hölder–Norm, } p\textit{–Norm.} \tag{8.2}$$

$$X = C[a,b] = \Big\{ f; \ f : [a,b] \overset{\text{stetig}}{\longrightarrow} K, \quad K = \mathbb{R} \text{ oder } \mathbb{C} \Big\}\,,$$

dann heißt

$$\|f\|_\infty = \max_{x \in [a,b]} |f(x)| \qquad \textit{Maximumnorm}, \tag{8.3}$$

$$\|f\|_{L_p} = \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}, \qquad L_p\textit{–Norm, Hölder–Norm.} \tag{8.4}$$

Nachweis der Normeigenschaften

Der Nachweis der Eigenschaften (i), (ii) ist für alle angeführten Beispiele offensichtlich.

Aufgabe:

Man beweise die Eigenschaft (iii) (Dreiecksungleichung) für die Maximumnorm in $\mathbb{R}^n$ und $C[a,b]$.

Die Dreiecksungleichung (Eigenschaft (iii)) für die Hölder–Norm, bzw. für die L_p–Norm, ist unter dem Namen *Minkowski'sche Ungleichung* bekannt. Sie wird in der Analysis bewiesen (vgl. Königsberger 1, § 9.8, Forster I, § 16). Für den Spezialfall $p = 2$ wird der Beweis in Satz 8.7 nachgetragen.

In normierten Räumen kann man Konvergenz und Stetigkeit erklären.

Definition 8.2. $(X, \|\ \|_x)$ und $(Y, \|\ \|_y)$ seien normierte Räume.

a) Eine Funktion $f : X \to Y$ heißt *stetig an der Stelle* $\hat{x} \in X$, falls gilt:

$$\forall \varepsilon > 0 \ \exists \delta = \delta(\varepsilon, \hat{x}) > 0 : \quad \|x - \hat{x}\|_x < \delta \quad \Rightarrow \quad \|f(x) - f(\hat{x})\|_y < \varepsilon$$

b) Eine Folge $\{x_n\} \subset X$ heißt *konvergent gegen ein* $\hat{x} \in X$ falls gilt:

$$\forall \varepsilon > 0 \ \exists N \in \mathbb{N} : \quad \|x_n - \hat{x}\|_x < \varepsilon \quad \forall n \geq N\,.$$

Man vergleiche zu diesen Begriffen Königsberger 1, § 5, Forster II, § 2.

Mit diesen Begriffen können wir weitere Normeigenschaften zeigen.

Sätzchen 8.3 In $(X, \| \ \|)$ gilt:
a) $\|x - y\| \ \geq \ | \, \|x\| - \|y\| \, | \quad \forall \, x, y \in X.$

b) $f(x) \ := \ \|x\|$ ist eine stetige Funktion von $X \to \mathbb{R}$.

Beweis:

a) Aus der Dreiecksungleichung folgt

$$\|x + (y - x)\| \ \leq \ \|x\| + \|y - x\| \quad \text{also} \quad \|y - x\| \ \geq \ \|y\| - \|x\| \, .$$

Vertauschung von x und y liefert $\|y - x\| \ \geq \ \|x\| - \|y\|$ und damit a).

b) Aus a) folgt

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

Man braucht also nur $\delta := \varepsilon$ zu setzen, um die Stetigkeitsdefinition 8.2a) zu erhalten. ■

Definition 8.2 wirft die Frage auf, ob Stetigkeit und Konvergenz normabhängige Eigenschaften sind. (Angenommen in X sind zwei verschiedene Normen erklärt und eine Folge oder Funktion ist bzgl. der einen Norm konvergent oder stetig. Ist sie das auch bzgl. der anderen Norm?) Eine solche Norm–Abhängigkeit besteht zum Teil tatsächlich. Um das einzusehen, erklären wir

Definition 8.4
Sei X ein linearer Raum mit zwei Normen $\| \cdot \|_\alpha, \| \cdot \|_\beta$.
a) $\| \cdot \|_\beta$ heißt stärker als $\| \cdot \|_\alpha \ \overset{\text{def.}}{\Longleftrightarrow} \ \exists \, M > 0 : \ \|x\|_\alpha \leq M \|x\|_\beta \ \forall \, x \in X.$

b) $\| \cdot \|_\alpha, \| \cdot \|_\beta$ heißen äquivalent $\overset{\text{def.}}{\Longleftrightarrow} \exists \, m, M > 0 : m\|x\|_\beta \leq \|x\|_\alpha \leq M\|x\|_\beta \forall \, x \in X.$

Aufgabe:
Zeige, daß im $\mathbb{R}^n$ die 1–Norm, die 2–Norm und die Maximumnorm äquivalent sind (Angabe von m, M).

Aufgabe:

Sei X ein linearer Raum mit den Normen $\| \|_\alpha$ und $\| \|_\beta$ und $f : (X, \| \|_\alpha) \to (Y, \| \|_y)$. $\| \|_\alpha$ sei stärker als $\| \|_\beta$. Man zeige: Konvergiert eine Folge $\{x_n\} \subset X$ bzgl. $\| \|_\alpha$, bzw. ist f bzgl. $\| \|_\beta$ stetig, so liegen Konvergenz bzw. Stetigkeit auch bzgl. $\| \|_\beta$ bzw. $\| \|_\alpha$ vor.

Aufgabe:

In $C[0,1]$ sei die Folge $f_n(x) = x^n$, $n \in \mathbb{N}$ gegeben. Man zeige

a) daß sie bzgl. $\| \|_{L_2}$ konvergiert, nicht aber bzgl. $\| \|_\infty$,

b) daß $\| \|_\infty$ stärker ist als $\| \|_{L_2}$.

Die letzte Aufgabe zeigt, daß Stetigkeit und Konvergenz in normierten Räumen tatsächlich normabhängig sein können. Eine Sonderrolle spielen die normierten Räume endlicher Dimension. Wir zeigen dies für das Beispiel des $\mathbb{R}^n$.

Satz 8.5 Im $\mathbb{R}^n$ sind alle Normen äquivalent.

Beweis:

Wir zeigen: Jede beliebige Norm $\| \|$ im $(\mathbb{R}^n, \| \|_\infty)$ ist äquivalent zu $\| \|_\infty$.
Beweisstruktur: 1) Zeige: in $(\mathbb{R}^n, \| \|_\infty)$ ist die Menge

$$S_\infty \quad = \quad \{\boldsymbol{x} \in \mathbb{R}^n;\ \|\boldsymbol{x}\|_\infty = 1\}$$

kompakt, 2) Jede Norm im $\mathbb{R}^n$ ist stetig bzgl. $\| \|_\infty$, nimmt auf S_∞ also Maximum und Minimuman, woraus direkt die Behauptung folgt.

S_∞ ist bzgl. $\| \cdot \|_\infty$ *beschränkt* d.h. $\exists\, k > 0 : \ \|\boldsymbol{x}\|_\infty < k \quad \forall\, \boldsymbol{x} \in S_\infty$ und *abgeschlossen*, denn ist $\{\boldsymbol{x}^n\} \subset S_\infty$ eine konvergente Folge: $\lim_{n\to\infty} \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_\infty = 0$, so folgt aus $|\, \|\boldsymbol{x}^n\|_\infty - \|\boldsymbol{x}^*\|_\infty \,| \le \|\boldsymbol{x}^n - \boldsymbol{x}^*\|_\infty$, daß $\|\boldsymbol{x}^*\|_\infty = 1$, also $\boldsymbol{x}^* \in S_\infty$.

S_∞ ist also *kompakt* (Satz von Heine–Borel, vgl. z.B. Forster II, § 3, Satz 5).
Jede beliebige Norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ im $\mathbb{R}^n$ ist eine stetige Funktion bzgl. $\| \|_\infty$, denn

$$\begin{aligned} |\, \|\boldsymbol{x}\| - \|\boldsymbol{y}\| \,| \ &\le\ \|\boldsymbol{x} - \boldsymbol{y}\| \ = \ \left\| \sum_{i=1}^n (x_i - y_i)\, \boldsymbol{e}^i \right\| \qquad (\boldsymbol{e}^i \mathrel{\widehat{=}} \text{Einheitsvektoren}) \\ &\le\ \sum_{i=1}^n \|(x_i - y_i)\, \boldsymbol{e}^i\| \ = \ \sum_{i=1}^n |x_i - y_i|\, \|\boldsymbol{e}^i\| \\ &\le\ \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \underbrace{\sum_{i=1}^n \|\boldsymbol{e}^i\|}_{=:\,C} \quad = \quad C \|\boldsymbol{x} - \boldsymbol{y}\|_\infty\,. \end{aligned}$$

Also nimmt die stetige Funktion $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ ihr Minimum m und ihr Maximum M auf S_∞ an (Königsberger 1, § 7.5, Königsberger 2, Forster II, § 3). Da $\|\boldsymbol{x}\| > 0 \quad \forall\, \boldsymbol{x} \in S_\infty$ (wegen Def. 8.1, (i)), ist $m > 0$. Also gilt

$$0 \ < \ m \ \le \ \|\boldsymbol{x}\| \ \le \ M \qquad \forall\, \boldsymbol{x} \in S_\infty\,. \tag{8.5}$$

Da insbesondere $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_\infty} \in S_\infty \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x} \neq \boldsymbol{0}$ liefert (8.5)

$$0 < m \le \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_\infty} \right\| \le M \quad \text{bzw.}$$
$$m\,\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\| \le M\,\|\boldsymbol{x}\|_\infty \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x} \neq \boldsymbol{0}\,.$$

Diese Ungleichung gilt auch für $\boldsymbol{x} = \boldsymbol{0}$, also sind $\|\ \|$ und $\|\ \|_\infty$ äquivalent. ■

Bemerkung:
Daß Normen äquivalent sind, bedeutet nicht, daß sie numerisch gleichwertig sind.

Sucht man bei einer Approximationsaufgabe eine beste Approximation, so wird die Behandlung dieser Aufgabe besonders einfach, wenn man einen „Senkrecht–Begriff“ zur Verfügung hat.

Beispiel 3: Gesucht ist im $\mathbb{R}^3$ der Punkt $\hat{\boldsymbol{x}}$ einer Ebene H, der zu einem gegebenen Punkt $\boldsymbol{x}^*$ außerhalb der Ebene einen minimalen Abstand hat. Die Anschauung zeigt (und auch die Mathematik): $\hat{\boldsymbol{x}}$ ist der Fußpunkt des Lots von $\boldsymbol{x}^*$ auf die Ebene H.

Senkrechtbeziehungen werden durch Skalarprodukte beschrieben. Diese Skalarprodukte liefern auch besonders schöne Normen (→ Lineare Approximation, Householder-Verfahren).

Definition 8.6
Sei X ein linearer Raum über K ($K = \mathbb{R}$ oder $K = \mathbb{C}$).
Eine Abbildung $\varphi : X \times X \to K$ heißt *Skalarprodukt* (*inneres Produkt*), wenn sie für alle $\lambda \in K$, $x, x_j, y \in X$ folgende Eigenschaften hat:

α) $\quad \varphi(x,x) \ge 0,\quad \varphi(x,x) = 0 \Leftrightarrow x = 0 \qquad$ (*Definitheit*)

β) $\quad \left.\begin{array}{lcl} \varphi(x_1 + x_2, y) &=& \varphi(x_1,y) + \varphi(x_2,y) \\ \varphi(\lambda\, x, y) &=& \lambda\,\varphi(x,y) \end{array}\right\} \qquad$ (*Linearität*)

γ) $\quad \varphi(x,y) = \overline{\varphi(y,x)} \qquad$ (*Schiefsymmetrie*)

$\left(\overline{\varphi(y,x)}\right.$ bezeichnet den konjugiert komplexen Wert von $\left.\varphi(y,x).\right)$

Ein linearer Raum X, in dem ein Skalarprodukt erklärt ist, heißt *unitär*, im Spezialfall $K = \mathbb{R}$ heißt er auch *euklidisch*.
Bemerkung: Überlicherweise benutzt man eine der Bezeichnungen

$$\varphi(x,y) = (x,y) \quad \text{oder} \quad \varphi(x,y) = \langle x, y\rangle\,.$$

Ein unitärer Raum X wird mit $(X, (\ ,\ ))$ bezeichnet.

Standardbeispiele:

1) $\mathbb{R}^n \ : \ (\boldsymbol{x}, \boldsymbol{y}) \ = \ \sum_{j=1}^{n} x_j \, y_j \ = \ \boldsymbol{x}^T \boldsymbol{y},$

1a) $\mathbb{C}^n \ : \ (\boldsymbol{x}, \boldsymbol{y}) \ = \ \sum_{j=1}^{n} x_j \, \overline{y_j} \ = \ \boldsymbol{x}^T \overline{\boldsymbol{y}},$

2) $C[a,b]$ (reellwertige Funktion) : $(f,g) \ = \ \int_a^b f(x)\, g(x)\, dx\,,$

2a) $C[a,b]$ (komplexwertige Funktion) : $(f,g) \ = \ \int_a^b f(x)\, \overline{g(x)}\, dx\,.$

Daßdiese Beispiele die Eigenschaften α)–γ) erfüllen, ist offensichtlich.

In jedem unitären Raum ist auch eine Norm definiert, wie der nächste Satz zeigt.

Satz 8.7
In $(X, (\ ,\ ))$ erfüllt die durch

$$\|x\| \quad := \quad \sqrt{(x,x)} \tag{8.6}$$

definierte Abbildung von X nach $\mathbb{R}$ folgende Eigenschaften:

a) $\|x\| \geq 0$,
$\|x\| = 0 \ \Leftrightarrow \ x = 0.$ *(Definitheit)*

b) $\|\lambda\, x\| = |\lambda|\, \|x\|$ *(schwache Homogenität)*

c) $|(x,y)| \leq \|x\|\, \|y\|$ (*Cauchy–Schwarz'sche Ungleichung*)
Das Gleichheitszeichen gilt genau dann, wenn x, y linear abhängig sind.

d) $\|x + y\| \leq \|x\| + \|y\|$ (Dreiecksungleichung)
Das Gleichheitszeichen gilt, falls $y = \lambda\, x$ oder $x = \lambda\, y$ mit $\lambda \geq 0$.

Also ist durch (8.6) eine Norm definiert (vgl. Def. 8.1).

Bemerkungen:
Die genannten Standardbeispiele für Skalarprodukte liefern die Hölder–Normen für $p = 2$ (vgl.(8.2.), (8.4)). Für sie wird durch d) der Beweis der Dreiecksungleichung nachgetragen. Diese Normen haben den Vorteil differenzierbar zu sein (im Gegensatz zur Maximumnorm).

Beweis:
Die Eigenschaften a), b) sind offensichtlich.
c): Für $y = 0$ gilt c), wir können also $y \neq 0$ annehmen. Dann folgt für $\lambda \in \mathbb{C}$ aus den Skarlarprodukteigenschaften:

$$0 \ \leq \ \|x + \lambda\, y\|^2 \ = \ (x + \lambda\, y, x + \lambda\, y) \ = \ (x,x) + \lambda(y,x) + \overline{\lambda}(x,y) + \lambda\, \overline{\lambda}(y,y)\,. \tag{8.7}$$

Setzt man speziell $\lambda = -\frac{(x,y)}{\|y\|^2}$, so folgt mit Def. 8.6, β), γ)

$$(8.8) \qquad 0 \le \|x\|^2 - \frac{|(x,y)|^2}{\|y\|^2} \quad \text{bzw.}$$

$$(8.9) \qquad |(x,y)| \le \|x\|\,\|y\|\,.$$

Das „=" in (8.7), und damit in (8.9) gilt genau dann, wenn $x = -\lambda\, y$ wegen Def. 8.6 α). Für diesen Fall rechnet man direkt nach:

$$|(x,y)| = |(-\lambda)\,(y,y)| = |\lambda|\,\|y\|^2 = \|\lambda\, y\|\,\|y\| = \|x\|\,\|y\|\,.$$

Vgl. dazu auch die geometrische Bedeutung in Lemma 8.8.

d): Setze in (8.7) $\lambda = 1 \quad \Rightarrow$

$$\begin{aligned} \|x+y\|^2 &= \|x\|^2 + 2\mathrm{Re}\,(x,y) + \|y\|^2 \\ &\le \|x\|^2 + 2|(x,y)| + \|y\|^2 \quad \text{und mit c)} \\ &\le \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 \\ &= (\,\|x\| + \|y\|)^2\,, \quad \text{also d).} \end{aligned}$$

Setzt man in beiden Seiten $y = \lambda\, x$, so sieht man, daß d) mit „=" gilt, falls $\lambda \ge 0$. Man mache sich im $\mathbb{R}^n$ die geometrische Bedeutung von $\lambda \ge 0$ klar. ■

Eine geometrische Deutung des Skalarprodukts sowie die Möglichkeit, mit seiner Hilfe Winkel zu messen, wird aufgezeigt durch

Lemma 8.8

Im $\mathbb{R}^n$ gilt mit $(\boldsymbol{x},\boldsymbol{y}) = \sum_{j=1}^{n} x_j\, y_j$

$$(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2 \cos\alpha\,, \qquad \alpha = \sphericalangle\,(\boldsymbol{x},\boldsymbol{y})\,. \tag{8.10}$$

Beweis:

Zwei Vektoren $\boldsymbol{x}, \boldsymbol{y}$ spannen ein Dreieck auf mit den Seiten $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}$

Laut den Eigenschaften des Skalarprodukts gilt

$$\begin{aligned} (\boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}) &= (\boldsymbol{x},\boldsymbol{x}) + (\boldsymbol{y},\boldsymbol{y}) - (\boldsymbol{x},\boldsymbol{y}) - (\boldsymbol{y},\boldsymbol{x}) \quad \text{bzw., im } \mathbb{R}^n \qquad (8.11) \\ \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 &= \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2(\boldsymbol{x},\boldsymbol{y})\,. \end{aligned}$$

Der cos–Satz der Trigonometrie besagt

$$\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \;=\; \|\boldsymbol{x}\|_2^2 \,+\, \|\boldsymbol{y}\|_2^2 \,-\, 2\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2 \cos\alpha$$

Die Behauptung folgt aus dem Vergleich der beiden Gleichungen. ■

Offensichtlich gilt

$$\boldsymbol{x},\boldsymbol{y}\in \mathrm{I\!R}^n \text{ sind orthogonal} \quad\Longleftrightarrow\quad \alpha = \frac{\pi}{2} \text{ oder } \frac{3\pi}{2} \quad\Longleftrightarrow\quad (\boldsymbol{x},\boldsymbol{y}) \;=\; 0\,.$$

Diese Eigenschaft nimmt man zum Anlaß, allgemein zu definieren:

Definition 8.9
In einem unitären Raum $(X,(\ ,\ ))$ heißen Elemente x,y *orthogonal zueinander* (Bezeichnung $x\perp y$), wenn gilt $(x,y)=0$.

Mit dieser Definition liest man aus Gleichung (8.11), die nichts $\mathrm{I\!R}^n$–Spezifisches enthält, ab:

Sätzchen 8.10
In einem unitären Raum $(X,(\ ,\ ))$ gilt für $\|x\|=\sqrt{(x,x)}$:

$$x\;\perp\;y \quad\Longleftrightarrow\quad \|x-y\|^2 \;=\; \|x\|^2+\|y\|^2 \qquad (\textit{Pythagoras})$$

Bemerkung:
Dieses Sätzchen hilft uns, geometrische Beweisideen aus dem $\mathrm{I\!R}^3$, die sich auf den Pythagoras beziehen, auch auf abstrakte unitäre Räume zu übertragen (vgl. etwa den Beweis des Projektionssatzes in der linearen Approximation, Satz 9.1).

Ähnlich wie die Größe eines Elements durch seine Norm beschrieben wird, kann man die Größe einer linearen Abbildung durch ihren maximalen Streckungsfaktor beschreiben. Warum das von Interesse ist, zeigt das

Beispiel 4: Bei der Untersuchung linearer Gleichungssysteme möchte man gerne wissen, wie weit die Lösung $\boldsymbol{x}^*$ von $\boldsymbol{A}\,\boldsymbol{x}=\boldsymbol{b}$ entfernt ist von der Lösung des Systems $\boldsymbol{A}\,\boldsymbol{x}=\hat{\boldsymbol{b}}$, wo $\hat{\boldsymbol{b}}$ eine Rundung von $\boldsymbol{b}$ ist.

Dazu untersucht man mit Hilfe der inversen Matrix $\boldsymbol{A}^{-1}$

$$\begin{aligned}
& \boldsymbol{A}\,\boldsymbol{x}^* - \boldsymbol{A}\,\hat{\boldsymbol{x}} &&= \boldsymbol{A}\,(\boldsymbol{x}^*-\hat{\boldsymbol{x}}) \;=\; \boldsymbol{b}-\hat{\boldsymbol{b}}\,,\\
\text{bzw.}\quad & \boldsymbol{x}^* - \hat{\boldsymbol{x}} &&= \boldsymbol{A}^{-1}\left(\boldsymbol{b}-\hat{\boldsymbol{b}}\right)\,,\\
\text{bzw.}\quad & \|\boldsymbol{x}^*-\hat{\boldsymbol{x}}\| &&= \left\|\boldsymbol{A}^{-1}(\boldsymbol{b}-\hat{\boldsymbol{b}})\right\|\,.
\end{aligned}$$

Man möchte wissen, um wieviel die Norm des Bildelements $\|\boldsymbol{x}^* - \hat{\boldsymbol{x}}\|$ unter der Matrix $\boldsymbol{A}^{-1}$ größer ist als die Norm $\|\boldsymbol{b} - \hat{\boldsymbol{b}}\|$ des Urbildelements, d.h. man sucht ein (möglichst kleines) $K > 0$ mit

$$\|\boldsymbol{x}^* - \hat{\boldsymbol{x}}\| \quad \leq \quad K \left\|\boldsymbol{b} - \hat{\boldsymbol{b}}\right\| ,$$

d.h. wir suchen den maximalen Streckungsfaktor K von $\boldsymbol{A}^{-1}$.
($\rightarrow$ Matrixnormen, Kondition, Fehlerabschätzungen)

Um dieses Problem anzugehen, benötigen wir Hilfsmittel aus der Linearen Algebra und der Analysis.

Aus der Linearen Algebra ist bekannt, daß jede lineare Abbildung $\boldsymbol{A}$ des $\mathbb{R}^n$ in sich durch eine Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ charakterisiert wird. Man verwendet deshalb für die Matrix und die Abbildung dieselbe Bezeichnung A, obwohl es sich mathematisch um verschiedene Begriffe handelt (vgl. Fischer § 2.5).

Als weiteres Hilfsmittel aus der Analysis benötigen wir

Satz 8.11
Jede lineare Abbildung A des normierten $\mathbb{R}^n$ in sich ist stetig.

Beweis:
Sei

$$\begin{array}{rcl} \boldsymbol{A}: \ \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ \boldsymbol{x} & \longrightarrow & \boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x} \end{array} \quad , \qquad \boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} .$$

Dann gilt für die 1–Norm

$$\begin{aligned} \|\boldsymbol{A}\,\boldsymbol{x}\|_1 = \|\boldsymbol{y}\|_1 = \sum_{i=1}^n |y_i| \ &= \ \sum_{i=1}^n \left| \sum_{j=1}^n a_{ij}\,x_j \right| \\ &\leq \ \sum_{i=1}^n \sum_{j=1}^n |a_{ij}|\,|x_j| \ = \ \sum_{j=1}^n |x_j| \sum_{i=1}^n |a_{ij}| \\ &\leq \ \sum_{j=1}^n |x_j| \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}| \end{aligned}$$

also

$$\|\boldsymbol{A}\,\boldsymbol{x}\|_1 \ \leq \ K_1\|\boldsymbol{x}\|_1 \quad \text{mit} \quad K_1 \ = \ \max_{j=1,\dots,n} \sum_{i=1}^n |a_{ij}| \tag{8.12}$$

K_1 ist von $\boldsymbol{x}$ unabhängig. Wegen

$$\|\boldsymbol{A}\,\boldsymbol{x}^1 - \boldsymbol{A}\,\boldsymbol{x}^2\|_1 \quad = \quad \|\boldsymbol{A}(\boldsymbol{x}^1 - \boldsymbol{x}^2)\|_1 \quad \leq \quad K_1\|\boldsymbol{x}^1 - \boldsymbol{x}^2\|_1 \,,$$

folgt hieraus die Stetigkeit von $\boldsymbol{A}$ bzgl. $\|\cdot\|_1$ (setze $\delta = \frac{\varepsilon}{K_1}$ in Def. 8.2 a)). Da im $\mathbb{R}^n$ alle Normen äquivalent sind (Satz 8.5), folgt nach der 2. Aufgabe nach Def. 8.4, daß $\boldsymbol{A}$ bzgl. aller Normen stetig ist. ∎

Bezeichnung:
Die Menge der linearen Abbildungen des $\mathbb{R}^n$ in sich wird mit $L(\mathbb{R}^n)$ bezeichnet.

Mit den eingeführten Hilfsmitteln zeigen wir:

Satz 8.12
Sei $\boldsymbol{A} \in L(\mathbb{R}^n)$ und $\|\cdot\|$ eine Norm im $\mathbb{R}^n$. Dann gilt: Der maximale Streckungsfaktor $\|\boldsymbol{A}\|$ von $\boldsymbol{A}$

$$\|\boldsymbol{A}\| \;:=\; \inf\left\{K \geq 0 : \;\|\boldsymbol{A}\,\boldsymbol{x}\| \leq K\|\boldsymbol{x}\| \;\forall\, \boldsymbol{x} \in \mathbb{R}^n\right\} \tag{8.13}$$

ist wohldefiniert, erfüllt die Normaxiome (Def. 8.1), und es gilt

$$\|\boldsymbol{A}\,\boldsymbol{B}\| \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{B}\| \quad \forall\, \boldsymbol{A}, \boldsymbol{B} \in L(\mathbb{R}^n)\,. \quad (\textit{Submultiplikativität}) \tag{8.14}$$

Beweis:
In (8.13) kann man ohne Einschränkung $\boldsymbol{x} \neq \boldsymbol{0}$ annehmen. Deshalb kann man $\|\boldsymbol{A}\|$ äquivalent definieren durch

$$\|\boldsymbol{A}\| \quad := \quad \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\,\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

Damit folgt

$$\|\boldsymbol{A}\| \quad := \quad \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\,\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \quad = \quad \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\left\|\frac{\boldsymbol{A}\,\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\|}{\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\|} \quad = \quad \sup_{\|\boldsymbol{y}\|=1} \|\boldsymbol{A}\,\boldsymbol{y}\| \tag{8.15}$$

Nun ist $f(\boldsymbol{y}) := \|\boldsymbol{A}\,\boldsymbol{y}\|$ eine stetige Funktion, denn $\boldsymbol{A}$ ist stetig (Satz 8.11), die Norm ist stetig (Satz 8.3) und die Hintereinanderausführung stetiger Funktionen ist stetig (Analysis!). Also nimmt $f(\boldsymbol{y})$ auf dem Kompaktum $\|\boldsymbol{y}\| = 1$ sein Maximum an (vgl. Beweis von Satz 8.5). Wir können in (8.15) also max statt sup schreiben und $\|\boldsymbol{A}\|$ ist wohldefiniert. Die Normeigenschaften (i) und (ii) folgen direkt aus (8.15), ebenso die Eigenschaft

$$\|\boldsymbol{A}\,\boldsymbol{x}\| \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{x}\|\,, \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^n\,. \tag{8.16}$$

Die Dreiecksungleichung (iii) ist erfüllt wegen

$$\begin{aligned} \|\boldsymbol{A} + \boldsymbol{B}\| \;&=\; \max_{\|\boldsymbol{x}\|=1} \|(\boldsymbol{A} + \boldsymbol{B})\,(\boldsymbol{x})\| \;\leq\; \max_{\|\boldsymbol{x}\|=1} \left(\|\boldsymbol{A}\,\boldsymbol{x}\| + \|\boldsymbol{B}\,\boldsymbol{x}\|\right) \\ &\leq\; \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\,\boldsymbol{x}\| \;+\; \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{B}\,\boldsymbol{x}\| \;=\; \|\boldsymbol{A}\| + \|\boldsymbol{B}\|\,. \end{aligned} \tag{8.17}$$

Schließlich gilt mit (8.16)

$$\|\boldsymbol{A}\,\boldsymbol{B}\,\boldsymbol{x}\| \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{B}\,\boldsymbol{x}\| \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|\,\|\boldsymbol{x}\| \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^n\,, \tag{8.18}$$

dies impiliziert nach (8.13): $\|\boldsymbol{A}\,\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$. ∎

Auf Grund dieses Satzes definieren wir

Definition 8.13
Eine Norm im linearen Raum der $(n \times n)$–Matrizen heißt *Matrixnorm*, falls für alle $(n \times n)$–Matrizen gilt

$$\left.\begin{array}{ll} \text{(i)} & \|\boldsymbol{A}\| \geq 0, \\ & \|\boldsymbol{A}\| = 0 \;\Leftrightarrow\; \boldsymbol{A} = \Theta \\ \text{(ii)} & \|\alpha\,\boldsymbol{A}\| = |\alpha|\,\|\boldsymbol{A}\| \quad \forall\,\alpha \in \mathbb{C} \\ \text{(iii)} & \|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\| \end{array}\right\} \text{Normaxiome}$$

$$\text{(iv)} \quad \|\boldsymbol{A}\,\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|\,. \qquad \text{(Submultiplikativität)}$$

Vorsicht:
Der Begriff Matrixnorm wird in der Literatur nicht einheitlich gleich verwendet. Da die $(n \times n)$–Matrizen einen Vektorraum (linearen Raum) bilden (lin. Algebra!), nennen einige Autoren jede Abbildung der $(n \times n)$–Matrizen nach $\mathbb{R}$, die (i)–(iii) erfüllt, Matrixnorm.

Beispiel:
$\|\boldsymbol{A}\| := \max\limits_{i,j=1,\ldots,n} |a_{ij}|$ erfüllt (i)–(iii), aber nicht (iv).

GemäßSatz 8.12 wird durch (8.13) bzw. (8.15) eine Matrixnorm definiert.

Definition 8.14
Sei $\|\cdot\|_V$ eine Vektornorm im $\mathbb{R}^n$, dann heißt

$$\|\boldsymbol{A}\|_M \quad := \quad \max_{\|\boldsymbol{x}\|_V=1} \|\boldsymbol{A}\,\boldsymbol{x}\|_V$$

die der Vektornorm $\|\cdot\|_V$ *zugeordnete Matrixnorm von A* (*natürliche Matrixnorm, Operatornorm*) .

Die zugeordnete Matrixnorm ist also **an eine Vektornorm gekoppelt**, was auch durch (8.16) zum Ausdruck kommt. Eine solche Koppelung ist in Def. 8.13 nicht enthalten. Deshalb treffen wir die

Definition 8.15
Eine Matrixnorm $\|\cdot\|_M$ heißt *passend* zu (*verträglich* mit) einer Vektornorm $\|\cdot\|_V$, falls für beliebige Matrizen $\boldsymbol{A}$ gilt

$$\|\boldsymbol{A}\,\boldsymbol{x}\|_V \;\leq\; \|\boldsymbol{A}\|_M\,\|\boldsymbol{x}\|_V \quad \forall\,\boldsymbol{x} \in \mathbb{R}^n\,.$$

Mit diesen Begriffen erhält man sofort die

Folgerung 8.16

Eine Matrixnorm $\|\cdot\|_M$ ist einer Vektornorm $\|\cdot\|_V$ genau dann zugeordnet, wenn

1) $\|\cdot\|_M$ passend zu $\|\cdot\|_V$,

2) $\forall$ Matrix $\boldsymbol{A}\ \exists\, \boldsymbol{x} \neq \boldsymbol{0}:\ \|\boldsymbol{A}\,\boldsymbol{x}\|_V = \|\boldsymbol{A}\|_M\, \|\boldsymbol{x}\|_V$.

Beweis:

Die Äquivalenz ergibt sich aus (8.13), (8.15), (8.16), weil in (8.15) das Supremum angenommen wird. ■

Bemerkung:

Da Vektor– und Matrixnorm einander zugeordnet oder miteinander verträglich sind, müßte man eigentlich zur Unterscheidung der verschiedenen Matrix– und Vektornormen Indizes anbringen. Dies wird überlicherweise aus Bequemlichkeit unterlassen. „Die richtige Bedeutung ergibt sich aus dem Zusammenhang.“

Den verschiedenen Vektornormen entsprechend, erhält man aus Satz 8.12 verschiedene Matrixnormen.

Satz 8.17 Beispiele natürlicher Matrixnormen

Für $\boldsymbol{x} \in \mathbb{R}^n$ und $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ sind einander zugeordnet

a) $\|\boldsymbol{x}\|_1$ und die *Spaltensummennorm* $\|\boldsymbol{A}\|_1 = \max\limits_j \sum\limits_i |a_{ij}|$,

b) $\|\boldsymbol{x}\|_2$ und die *Spektralnorm* $\|\boldsymbol{A}\|_2 = \sqrt{\rho(\boldsymbol{A}\,\boldsymbol{A}^*)}$, (vgl. Bem.)

c) $\|\boldsymbol{x}\|_\infty$ und die *Zeilensummennorm* $\|\boldsymbol{A}\|_\infty = \max\limits_i \sum\limits_j |a_{ij}|$.

Bemerkung:

μ heißt Eigenwert einer Matrix $\boldsymbol{B} \in \mathbb{C}^{n\times n}$, falls ein Vektor $\boldsymbol{x} \neq \boldsymbol{0}$ existiert mit $\boldsymbol{B}\,\boldsymbol{x} = \mu\,\boldsymbol{x}$ (vgl. Fischer: Lineare Algebra § 5). Für $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ ist $\boldsymbol{A}^* := \overline{\boldsymbol{A}}^T$. Man kann zeigen: Alle Eigenwerte μ_j $(j = 1, \dots, n)$ von $\boldsymbol{A}\,\boldsymbol{A}^*$ sind ≥ 0. Dann wird definiert $\rho(\boldsymbol{A}\,\boldsymbol{A}^*) = \max \mu_j$, und es ist dann $\|\boldsymbol{A}\|_2 = \max\limits_j \sqrt{\mu_j}$ eine Matrixnorm.

Beachte: Bezeichnet λ_j die Eigenwerte einer Matrix $\boldsymbol{A}$, so ist im allgemeinen $\max\limits_j |\lambda_j|$ **keine** Matrixnorm.

Beispiel: $\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\max\limits_j |\lambda_j| = 0$ aber $\boldsymbol{A} \neq \boldsymbol{0}$.

Beweis zu Satz 8.17:
Wir beweisen nur a) und c). Bzgl. b) verweisen wir auf die Literatur (z.B. Schwarz § 1.2.1, Opfer § 7.1).

Wir beweisen jeweils die Eigenschaften 1), 2) aus Folgerung 8.16.

a) 1) In (8.12) wurde schon gezeigt (vgl. Def. 8.15)

$$\|\boldsymbol{A}\,\boldsymbol{x}\|_1 \quad \leq \quad \max_j \sum_i |a_{ij}| \cdot \|\boldsymbol{x}\|_1 \,.$$

2) Für den Index j_0 sei $\|\boldsymbol{A}\|_1 = \sum\limits_i |a_{ij_0}|$. Dann gilt für den Einheitsvektor $\boldsymbol{x} = \boldsymbol{e}^{j_0}$:

$$\boldsymbol{A}\,\boldsymbol{e}^{j_0} \;=\; \begin{pmatrix} a_{1j_0} \\ \vdots \\ a_{nj_0} \end{pmatrix}, \qquad \left\|\boldsymbol{A}\,\boldsymbol{e}^{j_0}\right\|_1 \;=\; \sum_i |a_{ij_0}| \;=\; \|\boldsymbol{A}\|_1 \cdot \underbrace{\left\|\boldsymbol{e}^{j_0}\right\|_1}_{=\,1} .$$

c) 1)
$$\begin{aligned} \|\boldsymbol{A}\,\boldsymbol{x}\|_\infty \;=\; \max_i \left|\sum_j a_{ij}\,x_j\right| \;&\leq\; \max_i \sum_j |a_{ij}|\,|x_j| \\ &\leq\; \max_i \sum_j |a_{ij}|\,\|\boldsymbol{x}\|_\infty \;=\; \|\boldsymbol{A}\|_\infty\,\|\boldsymbol{x}\|_\infty \,. \end{aligned}$$

2) Für den Index i_0 sei $\|\boldsymbol{A}\|_\infty = \sum\limits_j |a_{i_0 j}|$.

Wähle $\boldsymbol{x} \in \mathbb{R}^n$ mit $\boldsymbol{x}_j \;=\; \begin{cases} 0 & \text{falls} \quad a_{i_0 j} = 0\,, \\ \frac{a_{i_0 j}}{|a_{i_0 j}|} & \text{sonst}\,. \end{cases}$

Dann ist $\|\boldsymbol{x}\|_\infty = 1$ und

$$\begin{aligned} \|\boldsymbol{A}\,\boldsymbol{x}\|_\infty \;&=\; \max_i \left|\sum_j a_{ij}\,x_j\right| \;\geq\; \left|\sum_j a_{i_0 j}\,x_j\right| \\ &=\; \sum_j |a_{i_0 j}| \;=\; \|\boldsymbol{A}\|_\infty \; \underbrace{\|\boldsymbol{x}\|_\infty}_{=\,1} . \end{aligned}$$

■

Der folgende Satz zeigt, daß es Matrixnormen gibt, die keiner Vektornorm zugeordnet sind und, daß nicht jede Matrixnorm mit jeder Vektornorm verträglich sein muß. Er rechtfertigt also die Definition von „passend“ und „zugeordnet“.

Satz 8.18

a) $\|\boldsymbol{A}\|_F = \sqrt{\sum\limits_{i,j} |a_{ij}|^2}$ ist eine Matrixnorm, die zu $\|\boldsymbol{x}\|_2$ paßt. (Sie wird unter den Namen *Frobenius–Norm*, *Schur–Norm*, *Erhard–Schmidt–Norm* gehandelt.)

b) $\|\boldsymbol{A}\|_F$ ist keiner Vektornorm zugeordnet, falls $n > 1$.

c) $\|\boldsymbol{A}\|_\infty$ paßt nicht zu $\|\boldsymbol{x}\|_1$ und $\|\boldsymbol{A}\|_1$ nicht zu $\|\boldsymbol{x}\|_\infty$.

Bemerkung:
Die Schur–Norm hat numerische Bedeutung, da man sie gelegentlich als Abschätzung (obere Schranke) für die (numerisch oft schwer berechenbare) Spektralnorm benutzen kann (vgl. Def. 8.14).

Beweis:

a) Die Normeigenschaften (i)–(iii) aus Def. 8.13 sind trivialerweise erfüllt, da $\|\boldsymbol{A}\|_F$ nichts anderes ist als die Vektornorm $\|\cdot\|_2$ im $\mathbb{R}^{n^2}$.

$$\text{(iv)} \qquad \|\boldsymbol{AB}\|_F^2 = \sum_{j,k=1}^{n} \left| \sum_{\mu=1}^{n} a_{j\mu}\, b_{\mu\, k} \right|^2 \overset{\text{CSU}}{\leq} \sum_{j,k=1}^{n} \left\{ \left(\sum_{\mu=1}^{n} |a_{j\mu}|^2 \right) \left(\sum_{\mu=1}^{n} |b_{\mu\, k}|^2 \right) \right\}$$

$$= \left(\sum_{\mu,j=1}^{n} |a_{j\,\mu}|^2 \right) \left(\sum_{\mu,k=1}^{n} |b_{\mu\, k}|^2 \right) = \|\boldsymbol{A}\|_F^2\, \|\boldsymbol{B}\|_F^2 \, .$$

Verträglichkeit zu $\|\cdot\|_2$:

$$\|\boldsymbol{A}\,\boldsymbol{x}\|_2 = \left\{ \sum_{j=1}^{n} \left(\sum_{k=1}^{n} a_{jk}\, x_k \right)^2 \right\}^{\frac{1}{2}} \overset{\text{CSU}}{\leq} \left(\sum_{j=1}^{n} \left\{ \left(\sum_{k=1}^{n} |a_{jk}|^2 \right) \left(\sum_{k=1}^{n} |x_k|^2 \right) \right\} \right)^{\frac{1}{2}}$$

$$= \left(\sum_{j,k=1}^{n} |a_{jk}|^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{n} |x_k|^2 \right)^{\frac{1}{2}} = \|\boldsymbol{A}\|_F\, \|\boldsymbol{x}\|_2 \, .$$

b) Aus Def. 8.14 folgt: Für jede einer Vektornorm $\|\cdot\|_V$ zugeordnete Matrixnorm $\|\cdot\|_M$ gilt für die Einheitsmatrix $\boldsymbol{E}$: $\|\boldsymbol{E}\|_M = 1$. Ist $\|\boldsymbol{E}\|_M > 1$, so folgt $\|\boldsymbol{x}\|_V = \|\boldsymbol{E}\,\boldsymbol{x}\|_V < \|\boldsymbol{E}\|_M \cdot \|\boldsymbol{x}\|_V \ \forall \boldsymbol{x}$. Dann kann $\|\cdot\|_M$ nach Folgerung 8.16 keiner Vektornorm zugeordnet sein. Für die Frobeniusnorm gilt $\|\boldsymbol{E}\|_F = \sqrt{n}$.

c) Beweis durch Angabe von Beispielen. Aufgabe! ■

Abschließend zeigen wir, daß es neben den angegebenen Vektor– und Matrixnormen noch ∞ viele andere gibt.

Satz 8.19
Sei $\boldsymbol{H}$ eine nichtsinguläre Matrix, dann gilt:

a) Ist $\|\boldsymbol{x}\|$ eine Vektornorm, so ist

$$\text{(8.19)} \qquad \|\boldsymbol{x}\|_H := \left\|\boldsymbol{H}^{-1}\,\boldsymbol{x}\right\|$$

wieder eine Vektornorm (*transformierte Vektornorm*).

b) Ist $\|\boldsymbol{A}\|$ eine Matrixnorm, so ist

$$\text{(8.20)} \qquad \|\boldsymbol{A}\|_H := \left\|\boldsymbol{H}^{-1}\,\boldsymbol{A}\,\boldsymbol{H}\right\|$$

wieder eine Matrixnorm (*transformierte Matrixnorm*).

c) Beim Übergang von den Normen zu den transformierten Normen bleiben die Eigenschaften *passend* und *zugeordnet* erhalten.

Beweis: Übungsaufgabe. ■

Als erste unmittelbare Anwendung von Vektor– und Matrixnormen kommen wir auf das motivierende Beispiel 4 zurück.

Seien $\hat{\boldsymbol{b}} \in \mathbb{R}^n$ eine Rundung von $\boldsymbol{b} \in \mathbb{R}^n$ und $\boldsymbol{x}$ bzw. $\hat{\boldsymbol{x}}$ die Lösungen, der Gleichungssysteme $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ bzw. $\boldsymbol{A}\,\hat{\boldsymbol{x}} = \hat{\boldsymbol{b}}$. Wir wollen den relativen Fehler von $\hat{\boldsymbol{x}}$ abschätzen, falls $\boldsymbol{A}^{-1}$ existiert. Es ist (vgl. Beispiel 4)

$$\begin{aligned}
\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| &= \|\boldsymbol{A}^{-1}(\boldsymbol{b} - \hat{\boldsymbol{b}})\| \;\leq\; \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{b} - \hat{\boldsymbol{b}}\| \quad \text{und} \\
\|\boldsymbol{b}\| &= \|\boldsymbol{A}\,\boldsymbol{x}\| \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{x}\| \quad \text{bzw.} \\
\tfrac{1}{\|\boldsymbol{x}\|} &\leq \tfrac{\|\boldsymbol{A}\|}{\|\boldsymbol{b}\|}\,.
\end{aligned}$$

Also folgt

$$\frac{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|\,\frac{\|\boldsymbol{b} - \hat{\boldsymbol{b}}\|}{\|\boldsymbol{b}\|} \tag{8.21}$$

Hierdurch wird beschrieben, wie sich der relative Fehler der rechten Seite des Gleichungssystems auf den relativen Fehler der Lösung auswirkt.

Definition 8.20

Die Zahl

$$\kappa\,(\boldsymbol{A}) \;:=\; \mathrm{Kond}\,(\boldsymbol{A}) \;:=\; \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\| \tag{8.22}$$

heißt *Kondition der* (regulären) *Matrix* $\boldsymbol{A}$.

Die Kondition beschreibt die Rundungsfehleranfälligkeit der Lösung eines linearen Gleichungssystems. Wegen (vgl. Beweis Satz 8.18b))

$$1 \;\leq\; \|\boldsymbol{E}\| \;=\; \|\boldsymbol{A}\,\boldsymbol{A}^{-1}\| \;\leq\; \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|$$

fällt sie immer ≥ 1 aus.

Man kann, noch allgemeiner als in (8.21), zeigen (vgl. Schwarz § 1.2.2): Vergleicht man die Lösungen von $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ und $(\boldsymbol{A} + \Delta\,\boldsymbol{A})\,\boldsymbol{x} = (\boldsymbol{b} + \Delta\,\boldsymbol{b})$, wobei $\Delta\,\boldsymbol{A}$ und $\Delta\,\boldsymbol{b}$ Störungen sind, die z.B. durch Rundungsfehler hervorgerufen werden, so gilt für den relativen Fehler die Abschätzung

$$\frac{\|\Delta\,\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \;\leq\; \frac{\mathrm{Kond}\,(\boldsymbol{A})}{1 - \mathrm{Kond}\,(\boldsymbol{A})\,\frac{\|\Delta\,\boldsymbol{A}\|}{\|\boldsymbol{A}\|}} \left\{ \frac{\|\Delta\,\boldsymbol{A}\|}{\|\boldsymbol{A}\|} + \frac{\|\Delta\,\boldsymbol{b}\|}{\|\boldsymbol{b}\|} \right\}. \tag{8.23}$$

Man sieht auch hier, daß die Schranke für den relativen Fehler von $\boldsymbol{x}$ mit der Kondition von $\boldsymbol{A}$ monoton wächst und daß sich (8.23) auf (8.21) reduziert, falls $\|\Delta\,\boldsymbol{A}\| = 0$.

Was bedeutet dies für die Lösung des linearen Gleichungssystems $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$? Wenn $\boldsymbol{A}$ problembedingt eine schlechte Kondition hat, so ist das i.allg. nicht zu ändern. Man kann dann nur darauf achten, daß das Verfahren, das zur Lösung von $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ verwendet wird, die Kondition nicht verschlechtert. Dies ist beim GEV leider der Fall, denn „gelöst“

wird nicht das Problem $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$, sondern das umgeformte Problem $\boldsymbol{A}^{(n-1)}\,\boldsymbol{x} = \boldsymbol{b}^{(n-1)}$ mit der oberen Dreiecksmatrix $\boldsymbol{A}^{(n-1)} = \boldsymbol{L}_{n-2}\,\ldots\,\boldsymbol{L}_1\,\boldsymbol{L}_0\,\boldsymbol{A}$ (vgl. (5.13)), und man kann ausrechnen, daß die Anwendung jedes einzelnen $\boldsymbol{L}_j$ die Kondition verschlechtert, d.h. die Kondition von $\boldsymbol{A}^{(n-1)}$ ist schlechter als die von $\boldsymbol{A}$, was besonders dann kritisch ist, wenn die Kondition von $\boldsymbol{A}$ ohnehin schon schlecht ist.

Wir kommen auf dieses Problem zurück bei der Ausgleichsrechnung (vgl. § 9, Householder–Verfahren).

§ 9 Lineare Ausgleichsrechnung, Überbestimmte Gleichungs–systeme

Wir beginnen wieder mit einem Beispiel, das die zu besprechende Problematik erläutert:

Beispiel:
Es sollen die Atomgewichte von Stickstoff und Sauerstoff bestimmt werden. Meßbar sind die Molekulargewichte folgender Stickoxyde.

	NO	N_2O	NO_2	N_2O_3	N_2O_4	N_2O_5
Molekulargew.	30.006	44.013	46.006	76.012	92.011	108.010

Bezeichnet man mit x_1 bzw. x_2 das Atomgewicht von Stickstoff und Sauerstoff, so erhält man aus der Tabelle für x_1, x_2 folgende Gleichungen

$$\begin{aligned} x_1 + \ \ x_2 &= \ \ 30.006 \\ 2x_1 + \ \ x_2 &= \ \ 44.013 \\ x_1 + 2x_2 &= \ \ 46.006 \\ 2x_1 + 3x_2 &= \ \ 76.012 \\ 2x_1 + 4x_2 &= \ \ 92.011 \\ 2x_1 + 5x_2 &= 108.010 \end{aligned} \tag{9.1}$$

Dieses Gleichungssystem $\boldsymbol{Ax} = \boldsymbol{b}$ ist überbestimmt (mehr Gleichungen als Unbekannte). Da alle Meßergebnisse in der Tabelle (z. Teil durch die Meßanordnungen bedingt) fehlerbehaftet sind, wird es üblicherweise keine Lösung besitzen. Deshalb wird man auch durch Auswahl von je 2 Gleichungen und deren Lösung unterschiedliche und üblicherweise auch falsche Ergebnisse erhalten. Man möchte daher die einzelnen Meßergebnisse gerne *ausgleichen*, indem man alle Messungen zur Bestimmung von x_1, x_2 benutzt, d.h. man möchte $\hat{x}_1, \hat{x}_2$ so bestimmen, daß die Abweichungen von den Meßwerten b_i (das sind die rechten Seiten der Gleichungen) von den Werten $\hat{b_i} = (\boldsymbol{Ax})_i$ „möglichst gering" ausfallen.

Schon von Gauß wurde zur Bearbeitung des Problems die sogenannte *„Methode der kleinsten Quadrate"* vorgeschlagen, die wir an Hand des Beispiels schildern wollen. Diese Methode präzisiert „möglichst gering" durch die Forderung (die unter gewissen Annahmen über die (statistische) Natur der Meßfehler auch sinnvoll erscheint)

$$\sum_{i=1}^{6} (\hat{b_i} - b_i)^2 \quad \stackrel{!}{=} \quad \min\ . \tag{9.2}$$

Bezeichnet man das System (9.1) durch

$$\boldsymbol{A}\,\boldsymbol{x} \ = \ \boldsymbol{b}, \qquad \boldsymbol{x} \ = \ (x_1, x_2)^T, \qquad \boldsymbol{b} \ = \ (b_1, \ldots, b_6)^T\,, \tag{9.1a}$$

so erfüllt die gesuchte Lösung $\hat{\boldsymbol{x}}$ die Gleichung $\boldsymbol{A}\,\hat{\boldsymbol{x}} = \hat{\boldsymbol{b}}$. Damit schreibt sich (9.2) als

$$\|\hat{\boldsymbol{b}} - \boldsymbol{b}\|_2^2 \ = \ \|\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 \ \stackrel{!}{=} \ \min\,, \qquad (\|\cdot\|_2 \mathrel{\widehat{=}} \text{Euklidische Norm}) \tag{9.3}$$

Damit wird (9.2) zu einer Forderung an die zu bestimmende „Näherungslösung $\hat{\boldsymbol{x}}$“ von (9.1). Dieses Beispiel führt uns also auf folgende

Lineare Ausgleichsaufgabe

Gegeben seien eine $m \times n$–Matrix $\boldsymbol{A}$ mit $m > n$ und $\boldsymbol{b} \in \mathbb{R}^m$.
Gesucht wird ein $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ mit (9.4)

$$\|\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2 \;\leq\; \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b}\|_2 \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^n\,.$$

Beachtet man, daß

$$V \quad = \quad \{\boldsymbol{y} \in \mathbb{R}^m\,;\;\; \boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x},\;\; \boldsymbol{x} \in \mathbb{R}^n\}$$

ein linearer Teilraum des $\mathbb{R}^m$ ist (vgl. Fischer, Satz 3.2.2), so kann man dieses Ausgleichsproblem auch schreiben als

Lineare Approximationsaufgabe

Sei V ein linearer Teilraum des $\mathbb{R}^m$ und $\boldsymbol{b} \in \mathbb{R}^m$.
Gesucht wird ein $\hat{\boldsymbol{y}} \in V$ mit (9.5)

$$\|\hat{\boldsymbol{y}} - \boldsymbol{b}\|_2 \;\leq\; \|\boldsymbol{y} - \boldsymbol{b}\|_2 \qquad \forall\, \boldsymbol{y} \in V\,.$$

Die Approximationsaufgabe heißt linear, weil die zur Konkurrenz zugelassenen Vektoren $\boldsymbol{y}$ in einem linearen Teilraum liegen. Diese Darstellung zeigt, daß die Ausgleichsrechnung ein Teilgebiet der Approximation darstellt.

Beachte:

Die Aufgabe (9.5) ist genau das Problem aus § 8, Beispiel 3, und wir erwarten, daß es auch die im Beispiel 3 vorgeschlagene Lösung besitzt. Daß dies tatsächlich so ist, beweist — gleich in etwas allgemeinerem Rahmen — der folgende Projektionssatz.

Satz 9.1 Projektionssatz

Sei $(X, (\ ,\ ))$ ein unitärer Raum, $V \subset X$ ein linearer Teilraum und $b \in X \backslash V$. Dann gilt

a) $\hat{v}$ ist beste Approximation aus V für b genau dann, wenn gilt $b - \hat{v} \perp V$, d.h.

(9.6) $$\|\hat{v} - b\| \;\leq\; \|v - b\| \quad \forall\, v \in V \quad \Longleftrightarrow \quad (\hat{v} - b, v) \;=\; 0 \quad \forall\, v \in V$$

b) Falls eine beste Approximation $\hat{v}$ existiert, ist sie eindeutig.

c) Sei $\dim\, V = n < \infty$ und $u^1, \ldots, u^n$ eine Basis von V, dann existiert eine beste Approximation $\hat{v} = \sum\limits_{i=1}^{n} v_i\, u^i$. Sie ist Lösung des Gleichungssystems

(9.7) $$(\hat{v}, u^j) \;=\; \sum_{i=1}^{n} v_i (u^i, u^j) \;=\; (b, u^j), \quad j = 1, \ldots, n \quad \textit{Normalgleichungen}$$

Bemerkungen

1) Das Gleichungssystem (9.7) ist nichts anderes als die Charakterisierung $b - \hat{v} \perp V$ aus a) für den endlichdimensionalen Fall. Sie reicht also, falls V endlich dimensional ist, aus, um die Existenz einer besten Approximation $\hat{v}$ zu garantieren.

2) $\hat{v}$ ist die orthogonale Projektion von $b \notin V$ auf $V : \hat{v} = P(b)$, daher der Name *Projektionssatz*. Sie ist, falls sie existiert, eindeutig gemäß b). Daß dies so ist, beruht wesentlich auf der Tatsache, daß die Norm durch ein Skalarprodukt gegeben wird: $\|x\| = \sqrt{(x,x)}$, man also einen Senkrechtbegriff zur Verfügung hat.

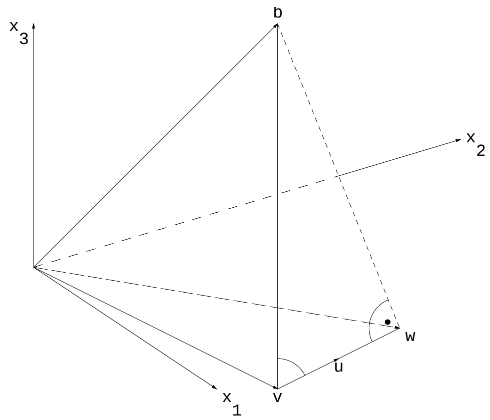
3) Das Gleichungssystem (9.7) hat die Koeffizientenmatrix $((u^i, u^j))_{i,j=1,\ldots,n}$. Sie heißt *Gram'sche Matrix*.

Beweis von Satz 9.1:

a) „$\Rightarrow$" (indirekt)

Annahme: $\exists\, u \in V : (b - \hat{v}, u) =: k \neq 0$.

Wir erläutern die Beweisidee zunächst geometrisch anschaulich am Beispiel: $X = \mathbb{R}^3$, $V = (x_1, x_2)$–Ebene, $b \notin V$.



Die Annahme lautet im Beispiel

$$\gamma \quad = \quad \sphericalangle\,(b - \hat{v}, u) \quad \neq \quad 90^\circ,\ 270^\circ\,.$$

Wir können also das Lot von b auf die Gerade $g(t) = \hat{v} + t\,u$, $t \in \mathbb{R}$, fällen und finden als Lotpunkt

(9.8) $$w \quad := \quad \hat{v} + \tilde{t}\,u\,.$$

Nun besagt der Pythagoras

(9.9) $$\|b - \hat{v}\|_2^2 \;=\; \|b - w\|_2^2 + \underbrace{\|w - \hat{v}\|_2^2}_{>\,0} > \|b - w\|_2^2\,,$$

d.h. $\hat{v}$ kann nicht Bestapproximation sein, weil w besser ist.

Um diesen Beweis auf den allgemeinen Fall zu übertragen, berechnen wir das w aus (9.8) (zunächst immer noch für unser Beispiel). Es gilt

$$\cos\gamma = \frac{\|w-\hat{v}\|_2}{\|b-\hat{v}\|_2} \quad \text{und} \quad k = (b-\hat{v},u) = \|b-\hat{v}\|_2\,\|u\|_2 \cos\gamma\,. \quad \text{(vgl. (8.10))}$$

Elimination von $\cos\gamma$ liefert

$$\frac{\|w-\hat{v}\|_2}{\|b-\hat{v}\|_2} = \frac{k}{\|b-\hat{v}\|_2\,\|u\|_2} \quad \text{bzw.} \quad \|w-\hat{v}\|_2 = \frac{k}{\|u\|_2}\,.$$

Somit folgt

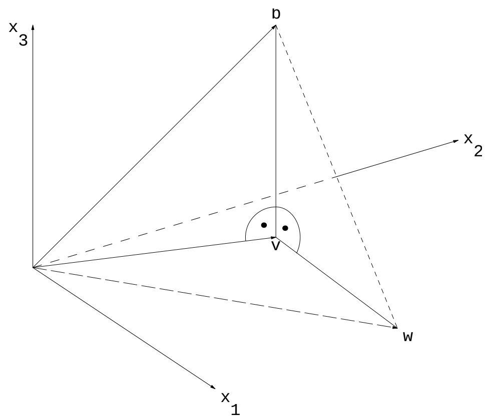
(9.8a)
$$w = \hat{v} + \|w-\hat{v}\|_2 \cdot \frac{u}{\|u\|_2} = \hat{v} + \frac{k}{\|u\|_2^2}\cdot u\,.$$

Im *allgemeinen Fall* führt die Annahme zum Widerspruch, wenn wir zeigen: $w = \hat{v} + \frac{k}{\|u\|_2^2}\,u$ ist eine bessere Approximation für b als $\hat{v}$.

Dies rechnen wir einfach nach.

$$\begin{aligned}
\|b-w\|^2 = (b-w,b-w) &= \left(b-\hat{v}-\frac{k}{\|u\|^2}\,u,\quad b-\hat{v}-\frac{k}{\|u\|^2}\,u\right)\\
&= (b-\hat{v},b-\hat{v}) - \left(b-\hat{v},\frac{k}{\|u\|^2}\,u\right) - \left(\frac{k}{\|u\|^2}\,u, b-\hat{v}\right) + \left(\frac{k}{\|u\|^2}\,u, \frac{k}{\|u\|^2}\,u\right)\\
&= \|b-\hat{v}\|^2 - \frac{\overline{k}}{\|u\|^2}\,\underbrace{(b-\hat{v},u)}_{=\,k} - \frac{k}{\|u\|^2}\,\underbrace{(u,b-\hat{v})}_{=\,\overline{k}} + \frac{k\,\overline{k}}{\|u\|^2}\\
&= \|b-\hat{v}\|^2 - \frac{|k|^2}{\|u\|^2} < \|b-\hat{v}\|^2\,.
\end{aligned}$$

a) „$\Leftarrow$“ Der Beweis für den allgemeinen Fall ist geometrisch unmittelbar anschaulich.



Erfüllt $\hat{v} : (b-\hat{v},w) = 0 \quad \forall\, w \in V$ (ist also orthogonale Projektion von b auf V), so gilt insbesondere

$$(b-\hat{v}, w-\hat{v}) = 0 \quad \forall\, w \in V\,.$$

Dann besagt der Pythagoras (vgl. Sätzchen 8.10)

$$\|b-w\|^2 \;=\; \|b-\hat{v}-(w-\hat{v})\|^2 \;=\; \|b-\hat{v}\|^2 + \underbrace{\|w-\hat{v}\|^2}_{>0 \text{ falls } w\neq\hat{v}} \;>\; \|b-\hat{v}\|^2\,,$$

d.h. $\hat{v}$ ist die beste Approximation an b und zugleich:

b) Jede beliebige Approximation $w \in V$ mit $w \neq \hat{v}$ ist schlechtere Approximation als $\hat{v}$. Also ist $\hat{v}$ eindeutig.

c) Jedes $w \in V$ hat eine Darstellung $w = \sum\limits_{i=1}^{n} w_i\, u^i$. Aus

$$(b-\hat{v}, w) \;=\; \left(b-\hat{v}, \sum_{i=1}^{n} w_i\, u^i\right) \;=\; \sum_{i=1}^{n} \overline{w}_i (b-\hat{v}, u^i)$$

erkennt man

(9.10) $$(b-\hat{v}, w) \;=\; 0 \quad \forall\, w \in V \quad \Longleftrightarrow \quad (b-\hat{v}, u^i) \;=\; 0, \quad i = 1, \ldots, n\,.$$

Die Normalgleichungen sind also nichts anderes als eine Charakterisierung der besten Approximation im endlich dimensionalen Fall. Macht man für $\hat{v}$ den Ansatz $\hat{v} = \sum\limits_{i=1}^{n} v_i\, u^i$, so geht die rechte Seite von (9.10) über in das Gleichungssystem

$$\sum_{i=1}^{n} v_i (u^i, u^j) \;=\; (b, u^j), \qquad j = 1, \ldots, n$$

mit der Koeffizientenmatrix (*Gram'sche Matrix*)

$$A \;=\; \big((u^i, u^j)\big)_{i,j=1,\ldots,n}\,.$$

Diese Matrix ist regulär (Beweis als Aufgabe). ■

Wir wenden den Projektionssatz an zur

Lösung der Ausgleichsaufgabe

Nach dem bisher Bewiesenen gelten folgende Äquivalenzen für $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m > n$, $\boldsymbol{b} \in \mathbb{R}^m$:

$$
\begin{array}{rl}
& \hat{\boldsymbol{x}} \;\; \text{löst die Ausgleichsaufgabe (9.4) :} \quad \|\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2 \;\stackrel{!}{=}\; \min \;. \\
\Longleftrightarrow & \hat{\boldsymbol{x}} \;\; \text{löst das Approximationsproblem (9.5):} \\
& \|\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}\|_2 \;\le\; \|\boldsymbol{y} - \boldsymbol{b}\|_2 \quad \forall\, \boldsymbol{y} \in \boldsymbol{A}(\mathbb{R}^n) \\
\Longleftrightarrow & (\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}, \boldsymbol{A}\,\boldsymbol{x}) \;=\; 0 \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n \quad \text{(gem. Projektionssatz)}
\end{array}
$$

Umformulierung mit Matrixschreibweise

$$
\begin{aligned}
(\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}, \boldsymbol{A}\,\boldsymbol{x}) &= (\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b})^T \boldsymbol{A}\,\boldsymbol{x} \\
&= (\boldsymbol{A}^T(\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{b}))^T\,\boldsymbol{x} \\
&= (\boldsymbol{A}^T\,\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{A}^T\,\boldsymbol{b}, \boldsymbol{x})
\end{aligned}
$$

liefert

$$
\begin{array}{rl}
\Longleftrightarrow & (\boldsymbol{A}^T\,\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{A}^T\,\boldsymbol{b}, \boldsymbol{x}) \;=\; 0 \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n \quad (\text{setze}\;\; \boldsymbol{x} = \boldsymbol{A}^T\,\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{A}^T\,\boldsymbol{b}) \\
\Longleftrightarrow & \boldsymbol{A}^T\,\boldsymbol{A}\,\hat{\boldsymbol{x}} - \boldsymbol{A}^T\,\boldsymbol{b} \;=\; \boldsymbol{0} \\
\Longleftrightarrow & \boxed{\boldsymbol{A}^T\,\boldsymbol{A}\,\hat{\boldsymbol{x}} \;=\; \boldsymbol{A}^T\,\boldsymbol{b} \quad \textit{Normalgleichungen der Ausgleichsrechnung.}}
\end{array}
$$

Satz 9.2
Sei $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $m > n$ und Rang $\boldsymbol{A} = n$.
Dann sind die *Normalgleichungen der Ausgleichsrechnung*

$$
\boldsymbol{A}^T\,\boldsymbol{A}\,\hat{\boldsymbol{x}} \quad = \quad \boldsymbol{A}^T\,\boldsymbol{b}
$$

eindeutig lösbar und liefern die Lösung des Ausgleichsproblems (9.4).

Beweis:
Wegen Rang $\boldsymbol{A} = n$ sind die Spalten $\boldsymbol{a}^k$ von $\boldsymbol{A}$: $\boldsymbol{a}^k = \boldsymbol{A}\,\boldsymbol{e}^k$, $k = 1, \ldots, n$, $\boldsymbol{e}^k \mathrel{\widehat{=}}$ Einheitsvektoren des $\mathbb{R}^n$, linear unabhängig, bilden also eine Basis für den Raum $V = \boldsymbol{A}(\mathbb{R}^n)$. Dann ist $(\boldsymbol{A}^T\,\boldsymbol{A}) = \big((\boldsymbol{a}^i, \boldsymbol{a}^k)\big)_{i,k=1,\ldots,n}$ eine *Gram'sche Matrix*, also nicht singulär (vgl. Beweis c) des Projektionssatzes). ■

Mit dem Nachweis der Existenz einer eindeutigen Lösung des Ausgleichsproblems ist — numerisch betrachtet — die Aufgabe leider noch lange nicht gelöst, denn die numerischen Ergebnisse, die man durch Lösen der Normalgleichungen erhält, sind oft herzhaft schlecht. Dies hat mehrere Gründe.

Grund 1 ist „naturbedingt“ durch die Überbestimmtheit des Gleichunggsystems, was oft zur Folge hat, daß jede quadratische Teilmatrix von A auf Grund von „linearen Fastabhängigkeiten“ schlecht konditioniert ist.

Grund 2 liegt in der Transformation des Ausgleichsproblems in das System der Normalgleichungen. Diese Transformation verschlechtert die Lösungseigenschaften der Aufgabe. Um dies zu verdeutlichen (eine detaillierte Analyse findet man in Schwarz § 7.2), betrachten wir den Grenzfall, daß $\boldsymbol{A}$ schon eine quadratische Matrix ist, das Ausgleichsproblem also durch Lösung der Gleichungen $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$, statt $\boldsymbol{A}^T\,\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{A}^T\,\boldsymbol{b}$ gefunden werden könnte. Nun kann man zeigen — mit Einsatz von ein wenig Eigenwerttheorie — daß für die Spektralnorm gilt: Kond $(\boldsymbol{A}^T\boldsymbol{A}) = (\text{Kond }(\boldsymbol{A}))^2$. Ist also A ohnehin schon schlecht konditioniert (Kond $(\boldsymbol{A}) \gg 1$), so gilt das in umsogrößerem Maße für $\boldsymbol{A}^T\boldsymbol{A}$.

Grund 3 liegt darin, daß das Überführen von $\boldsymbol{A}^T\boldsymbol{A}$ in obere Dreiecksgestalt durch das Gauß'sche Eliminationsverfahren die Kondition nochmals verschlechtert (vgl. § 8, Ende).

Zur besseren numerischen Lösung des Ausgleichsproblems wird man also nach einem Verfahren suchen, welches das Ausgleichsproblem in eine lösbare Gestalt transformiert, die nicht schlechter konditioniert ist als das Ausgangsproblem. Dies leisten orthogonale Transformationen, auf denen das folgende Verfahren basiert.

Wir gehen aus von der Formulierung (9.5) als Approximationsproblem:

$$\text{Bestimme} \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2, \;\; \boldsymbol{A}\in\mathbb{R}^{m\times n}, \;\; \boldsymbol{b}\in\mathbb{R}^m, \;\; m>n\,.$$

Die zu besprechende Methode wird durch folgende Fakten und Fragen motiviert.

1) Ist $\boldsymbol{A}$ eine obere Dreiecksmatrix mit Rang $\boldsymbol{A} = n$, also

$$\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \quad = \quad \underbrace{\begin{pmatrix} & \widehat{\boldsymbol{A}} \\ \mathbf{O} & \\ \hline & \mathbf{O} & \end{pmatrix}}_{n} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hline \overline{\boldsymbol{b}} \end{pmatrix}$$

so gilt gemäß der Definition von $\|\cdot\|_2$:

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 \quad = \quad \|\widehat{\boldsymbol{A}}\boldsymbol{x}-\hat{\boldsymbol{b}}\|_2^2 + \|\overline{\boldsymbol{b}}\|_2^2\,.$$

Dann wird $\min \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2$ in der Lösung von $\widehat{\boldsymbol{A}}\boldsymbol{x} = \hat{\boldsymbol{b}}$ angenommen.

2) Ist es möglich, $\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}$ durch Linksmultiplikation mit einer $m\times m$–Matrix $\boldsymbol{F}$ konditionsneutral so umzuformen, daß $\boldsymbol{F}\boldsymbol{A}$ Dreiecksgestalt besitzt **und**

$$(9.11) \qquad \|\boldsymbol{F}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\|_2 \;=\; \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2 \quad \forall\, \boldsymbol{x}\in\mathbb{R}^n \;?$$

Dann nämlich würden $\|\boldsymbol{F}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\|_2$ und $\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2$ ihr Minimum im selben Punkt $\hat{\boldsymbol{x}}$ annehmen und $\hat{\boldsymbol{x}}$ könnte wie unter 1) berechnet werden.

Matrizen $\boldsymbol{F}$, die (9.11) gewährleisten, heißen *längentreu* (oder *orthogonal*, vgl. Satz 9.3). Solche Matrizen existieren offensichtlich (Drehungen, Spiegelungen). Sie haben zudem die Eigenschaft, daß für jede quadratische Matrix $\boldsymbol{B}$ in der Spektralnorm gilt

Kond $(\boldsymbol{B})$ = Kond $(\boldsymbol{F}\,\boldsymbol{B})$, vgl. Bemerkung am Schluß des §, die Transformation ist also konditionsneutral (im Gegensatz zu den Matrizen des GEV).

Householder hat 1958 gezeigt, daß man mit Spiegelungen $\boldsymbol{A}$ auf obere Dreiecksgestalt transformieren kann (*Verfahren von Housholder*). Givens hat, ebenfalls 1958, gezeigt, daß dies auch mit Drehungen möglich ist (*Givens Verfahren*). Da das Verfahren von Givens in etwa einen doppelt so großen Rechenaufwand benötigt wie das Householder–Verfahren, werden wir das letztere besprechen. Wir untersuchen also zunächst

Orthogonale Matrizen, Spiegelungen

Definition und Satz 9.3

Eine Matrix $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ heißt *orthogonal*, falls eine der folgenden äquivalenten Eigenschaften gilt.

(9.12) $\quad \boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{E}$ (*die Spalten sind orthogonale Einheitsvektoren*)

(9.13) $\quad (\boldsymbol{F}\,\boldsymbol{x}, \boldsymbol{F}\,\boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{y}) \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ (*winkeltreu*)

(9.14) $\quad \|\boldsymbol{F}\,\boldsymbol{y}\|_2 = \|\boldsymbol{y}\|_2 \quad \forall\, \boldsymbol{y} \in \mathbb{R}^n$ (*längentreu*)

Beweis:

(9.12) ⇒ (9.13): $(\boldsymbol{F}\,\boldsymbol{x}, \boldsymbol{F}\,\boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{F}^T\boldsymbol{F}\,\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{y})$.

(9.13) ⇒ (9.14): Setze oben $\boldsymbol{x} = \boldsymbol{y}$, dann ist $\|\boldsymbol{F}\,\boldsymbol{y}\|_2^2 = (\boldsymbol{F}\,\boldsymbol{y}, \boldsymbol{F}\,\boldsymbol{y}) = \|\boldsymbol{y}\|_2^2$.

(9.14) ⇒ (9.12): $\|\boldsymbol{F}\,\boldsymbol{y}\|_2^2 = (\boldsymbol{F}\,\boldsymbol{y}, \boldsymbol{F}\,\boldsymbol{y}) = \boldsymbol{y}^T\boldsymbol{F}^T\boldsymbol{F}\,\boldsymbol{y} = (\boldsymbol{y}, \boldsymbol{y}) \quad \forall\, \boldsymbol{y} \in \mathbb{R}^n$ laut (9.14).

Setze in die letzte Gleichung insbesondere die Vektoren

$$\boldsymbol{y} = (0, \ldots, 0, \underset{\substack{\uparrow\\ i}}{1}, 0, \ldots, 0, \underset{\substack{\uparrow\\ k}}{1}, 0, \ldots, 0)^T, \qquad \forall\, i, k \in \{1, \ldots, n\}$$

ein. Daraus folgt für die symmetrische Matrix $\boldsymbol{A} = \boldsymbol{F}^T\boldsymbol{F}$ für $i = k$: $a_{i,i} = 1$ und damit für $i \neq k : a_{ik} = 0$. ■

Leicht einzusehen sind folgende

Eigenschaften orthogonaler Matrizen

(9.15) $\quad \boldsymbol{F}$ orthogonal ⇒ $\boldsymbol{F}^{-1}$ orthogonal

(9.16) $\quad \boldsymbol{F}$ orthogonal ⇒ $\|\boldsymbol{F}\|_2 = \|\boldsymbol{F}^{-1}\|_2 = 1$ (Spektralnorm)

(9.17) $\quad$ Das Produkt orthogonaler Matrizen ist orthogonal.

Beweis:

(9.15): $\|\boldsymbol{y}\|_2 = \|\boldsymbol{F}(\boldsymbol{F}^{-1}\,\boldsymbol{y})\|_2 = \|\boldsymbol{F}^{-1}\,\boldsymbol{y}\|_2$ nach (9.14) .

(9.16): $\|\boldsymbol{F}\|_2 := \sup\limits_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{F}\,\boldsymbol{y}\|_2 = \sup\limits_{\|\boldsymbol{y}\|_2=1} \|\boldsymbol{y}\|_2 = 1,$ ebenso für $\boldsymbol{F}^{-1}$.

(9.17): Sei $\boldsymbol{A}^T\,\boldsymbol{A} = \boldsymbol{E},\ \boldsymbol{B}^T\,\boldsymbol{B} = \boldsymbol{E} \ \Rightarrow\ (\boldsymbol{A}\,\boldsymbol{B})^T(\boldsymbol{A}\,\boldsymbol{B}) = \boldsymbol{B}^T\,\boldsymbol{A}^T\,\boldsymbol{A}\,\boldsymbol{B} = \boldsymbol{E}.$ ■

Wir leiten nun anschaulich her, wie eine Matrix $\boldsymbol{H}$ aussieht, die eine Spiegelung beschreibt, und zeigen danach, daß die so erhaltene Matrix tatsächlich orthogonal ist.

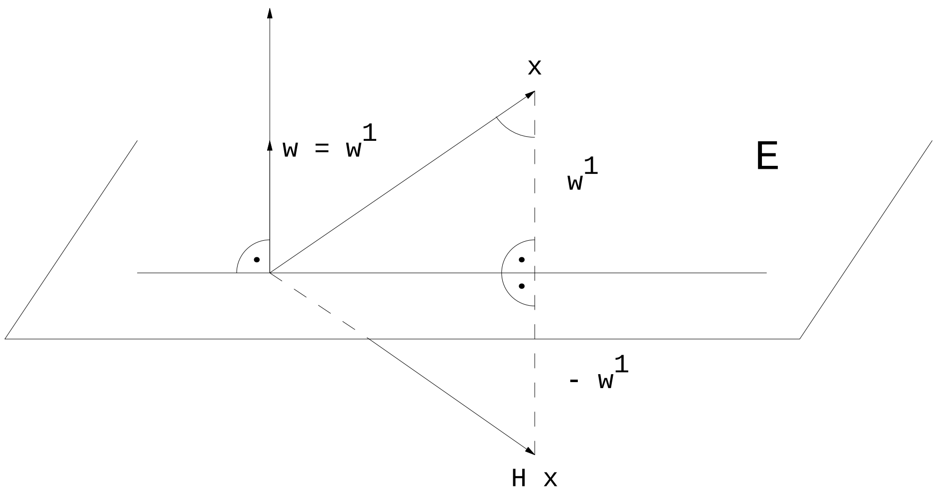
Eine Hyperebene E kann beschrieben werden durch

$$E \;=\; \{\boldsymbol{x} \in \mathbb{R}^n;\ \boldsymbol{w}^T\,\boldsymbol{x} = 0\} \;=:\; \boldsymbol{w}^\perp \quad \text{für ein } \boldsymbol{w} \in \mathbb{R}^n \text{ mit } \|\boldsymbol{w}\|_2 = 1\,.$$

Es existiert eine Orthonormalbasis $\{\boldsymbol{w}^1, \ldots, \boldsymbol{w}^n\}$ des $\mathbb{R}^n$ mit $\boldsymbol{w}^1 = \boldsymbol{w}$ (vgl. Lineare Algebra: Orthogonalisierungsverfahren von Erhard Schmidt oder eine Drehung des $\mathbb{R}^n$, die $\boldsymbol{e}^1$ nach $\boldsymbol{w}$ dreht). Bezüglich dieser Basis hat $\boldsymbol{x} \in \mathbb{R}^n$ eine Darstellung

$$\boldsymbol{x} \;=\; \sum_{j=1}^{n} \alpha_j\,\boldsymbol{w}^j\,,$$

und die Abbildungsvorschrift für die Spiegelung von $\boldsymbol{x}$ an der Hyperebene E ergibt sich aus der Zeichnung



$$\begin{array}{rcl} \boldsymbol{H} : \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ \boldsymbol{x} = \sum\limits_{j=1}^{n} \alpha_j\,\boldsymbol{w}^j & \longrightarrow & \boldsymbol{H}\,\boldsymbol{x} = -\alpha_1\,\boldsymbol{w}^1 + \sum\limits_{j=2}^{n} \alpha_j\,\boldsymbol{w}^j \\ & & = \boldsymbol{x} - 2\alpha_1\,\boldsymbol{w}^1 \end{array}$$

Wir berechnen α_1:

$$\left.\begin{array}{rcll} \cos\gamma & = & \dfrac{\|\alpha_1\,\boldsymbol{w}\|_2}{\|\boldsymbol{x}\|_2} & \text{(vgl. Zeichnung)} \\ \cos\gamma & = & \dfrac{(-\boldsymbol{w}, -\boldsymbol{x})}{\|\boldsymbol{w}\|_2\,\|\boldsymbol{x}\|_2} & \text{(Lemma 8.8)} \end{array}\right\} \overset{\|w\|_2=1}{\Longrightarrow} \quad \alpha_1 \;=\; (\boldsymbol{w}, \boldsymbol{x}) \quad (\text{Œ}\,\alpha_1 > 0)$$

Also folgt

$$\begin{aligned}
\boldsymbol{H}\,\boldsymbol{x} &= \boldsymbol{x} - 2\alpha_1\,\boldsymbol{w} = \boldsymbol{x} - 2(\boldsymbol{w},\boldsymbol{x})\,\boldsymbol{w} = \boldsymbol{x} - 2\,\boldsymbol{w}(\boldsymbol{w},\boldsymbol{x}) \\
&= \boldsymbol{x} - 2\boldsymbol{w}\,\boldsymbol{w}^T\,\boldsymbol{x} = (\boldsymbol{E} - 2\boldsymbol{w}\,\boldsymbol{w}^T)\,\boldsymbol{x}\,.
\end{aligned}$$

Definition und Lemma 9.4

$$\boldsymbol{H} = \boldsymbol{E} - 2\boldsymbol{w}\,\boldsymbol{w}^T \quad \text{mit} \quad \|\boldsymbol{w}\|_2 = 1,\ \boldsymbol{w} \in \mathbb{R}^n \tag{9.18}$$

heißt *Householdermatrix.*
Die Matrix

$$\boldsymbol{w}\,\boldsymbol{w}^T = \begin{pmatrix} w_1\,w_1 & \dots & w_1\,w_n \\ \vdots & & \vdots \\ w_n\,w_1 & \dots & w_n\,w_n \end{pmatrix} \tag{9.19}$$

heißt *Dyade* oder *dyadisches* Produkt von $\boldsymbol{w}$ und $\boldsymbol{w}^T$. Householdermatrizen sind symmetrisch und orthogonal.

Beweis:
Die Symmetrie ist offensichtlich. Die Orthogonalität folgt gemäß Satz 9.3 aus

$$\begin{aligned}
\boldsymbol{H}^T\,\boldsymbol{H} &= (\boldsymbol{E} - 2\boldsymbol{w}\,\boldsymbol{w}^T)\,(\boldsymbol{E} - 2\boldsymbol{w}\,\boldsymbol{w}^T) \\
&= \boldsymbol{E} - 4\boldsymbol{w}\,\boldsymbol{w}^T + 4\boldsymbol{w}\,\underbrace{\boldsymbol{w}^T\,\boldsymbol{w}}_{=\,1}\,\boldsymbol{w}^T = \boldsymbol{E}\,.
\end{aligned}$$

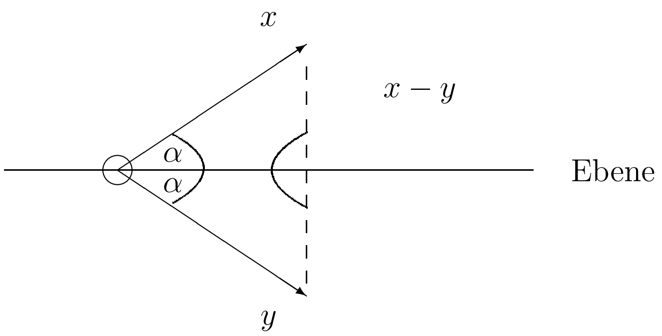
■

Es ist anschaulich klar, daß zwei gleich lange Vektoren durch eine Spiegelung ineinander abgebildet werden können. Wir formulieren dies als

Lemma 9.5
Zu $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{x} \neq \boldsymbol{y}$, $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2$ gibt es eine Householdermatrix $\boldsymbol{H} \in \mathbb{R}^{m\times m}$ mit $\boldsymbol{H}\,\boldsymbol{x} = \boldsymbol{y}$.
Es ist

$$\boldsymbol{H} = \boldsymbol{E}_m - 2\,\boldsymbol{w}\,\boldsymbol{w}^T \quad \text{mit} \quad \boldsymbol{w} = \pm\frac{\boldsymbol{x}-\boldsymbol{y}}{\|\boldsymbol{x}-\boldsymbol{y}\|_2}\,, \qquad \boldsymbol{E}_m \mathrel{\widehat{=}} \text{Einheitsmatrix im } \mathbb{R}^m$$

Beweis: Die Gestalt von $\boldsymbol{w}$ erhält man aus der folgenden Zeichnung. Daß dieses $\boldsymbol{w}$ die Behauptung erfüllt, rechnen wir nach.



Wir rechnen das Ergebnis nach.

$$\boldsymbol{H}\,\boldsymbol{x} = \boldsymbol{y} \iff (\boldsymbol{E} - 2\boldsymbol{w}\,\boldsymbol{w}^T)\,\boldsymbol{x} \;=\; \boldsymbol{y}$$

$$\iff \boldsymbol{x} - 2\,\frac{(\boldsymbol{x}-\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|_2} \cdot \frac{(\boldsymbol{x}-\boldsymbol{y})^T}{\|\boldsymbol{x}-\boldsymbol{y}\|_2}\,\boldsymbol{x} \;=\; \boldsymbol{y}$$

$$\iff \boldsymbol{x} - (\boldsymbol{x}-\boldsymbol{y})\,\frac{2(\boldsymbol{x}-\boldsymbol{y})^T\,\boldsymbol{x}}{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2} \;=\; \boldsymbol{y}$$

Nun ist

$$\begin{aligned}
2(\boldsymbol{x}-\boldsymbol{y})^T\,\boldsymbol{x} \;&=\; 2(\boldsymbol{x}-\boldsymbol{y},\boldsymbol{x}) \;=\; 2(\boldsymbol{x},\boldsymbol{x}) - 2(\boldsymbol{x},\boldsymbol{y}) \quad \text{und da} \quad \|\boldsymbol{x}\|_2 \;=\; \|\boldsymbol{y}\|_2 \\
&=\; (\boldsymbol{x},\boldsymbol{x}) \;-\; 2(\boldsymbol{x},\boldsymbol{y}) \;+\; (\boldsymbol{y},\boldsymbol{y}) \;=\; (\boldsymbol{x}-\boldsymbol{y},\boldsymbol{x}-\boldsymbol{y}) \\
&=\; \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\,.
\end{aligned}$$

■

Nach Bereitstellung der mathematischen Hilfsmittel beschreiben wir nun das

Householder–Verfahren

Gezeigt wird, daß eine Matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m > n$, durch eine Folge von n Spiegelungen $\boldsymbol{F} = \boldsymbol{H}_n \cdot \ldots \cdot \boldsymbol{H}_1$ auf obere Dreiecksgestalt transformiert werden kann:

$$\boldsymbol{F}\,\boldsymbol{A} \;=\; \boldsymbol{H}_n \cdot \ldots \cdot \boldsymbol{H}_1\,\boldsymbol{A} \;=\; \boldsymbol{A}^{(n)}\,, \quad \boldsymbol{A}^{(n)} \;=\; \begin{pmatrix} \widehat{\boldsymbol{A}}^{(n)} \\ \\ \\ \mathbf{O} \end{pmatrix}, \quad \widehat{\boldsymbol{A}}^{(n)} \in \mathbb{R}^{n\times n}\,.$$

Schritt 1: Gesucht wird eine Matrix $\boldsymbol{H}_1 \in \mathbb{R}^{m\times m}$ derart, daß in $\boldsymbol{H}_1 \boldsymbol{A} = \boldsymbol{A}^{(1)}$ die Elemente „unter $a_{11}^{(1)}$" gleich Null sind, oder, anders formuliert, derart, daß die 1. Spalte $\boldsymbol{a}^1$ von $\boldsymbol{A}$ durch $\boldsymbol{H}_1$ auf ein Vielfaches des ersten Einheitsvektors $\boldsymbol{e}^1$ abgebildet wird.

Dazu wenden wir Lemma 9.5 an auf den Spezialfall $\boldsymbol{x} = \boldsymbol{a}^1$ ($\widehat{=}$ 1. Spalte von A) und $\boldsymbol{y} = \pm \|\boldsymbol{x}\|_2 \, \boldsymbol{e}^1$. Dann wird

$$\boldsymbol{w} \quad = \quad \frac{\boldsymbol{x} \mp \|\boldsymbol{x}\|_2 \, \boldsymbol{e}^1}{\|\boldsymbol{x} \mp \|\boldsymbol{x}\|_2 \, \boldsymbol{e}^1\|_2} \quad =: \quad \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} \, .$$

Um numerische Auslöschungen zu vermeiden, wählen wir das Vorzeichen in $\boldsymbol{u}$ gemäß

$$\boldsymbol{u} \quad = \quad \boldsymbol{x} + \operatorname{sgn}(x_1) \, \|\boldsymbol{x}\|_2 \, \boldsymbol{e}^1 \quad = \quad (x_1 + \operatorname{sgn}(x_1) \, \|\boldsymbol{x}\|_2 \, , \; x_2, \ldots, x_n)^T$$

(mit $\operatorname{sgn}(x_1) := 1$, falls $x_1 = 0$) und $\boldsymbol{H}$ erhält die Gestalt

$$\boldsymbol{H} \quad = \quad \boldsymbol{E}_m - \frac{2 \, \boldsymbol{u} \, \boldsymbol{u}^T}{\|\boldsymbol{u}\|_2^2} \quad = \quad \boldsymbol{E}_m - \frac{\boldsymbol{u} \, \boldsymbol{u}^T}{c}$$

mit

$$\begin{aligned} c \quad &= \quad \frac{1}{2} \, \|\boldsymbol{u}\|_2^2 \quad = \quad \frac{1}{2} \left(x_1^2 + 2x_1 \operatorname{sgn}(x_1) \, \|\boldsymbol{x}\|_2 + \|\boldsymbol{x}\|_2^2 \; + \; \sum_{j=2}^{n} x_j^2 \right) \\ &= \quad \|\boldsymbol{x}\|_2^2 \; + \; |x_1| \;\; \|\boldsymbol{x}\|_2 \, . \end{aligned}$$

Insgesamt also

Lemma 9.5'

$\boldsymbol{x} = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$ wird numerisch stabil auf ein Vielfaches des Einheitsvektors $\boldsymbol{e}^1$ abgebildet durch die Householdermatrix

$$\begin{aligned} \boldsymbol{H} \quad &= \quad \boldsymbol{E}_m \; - \; \tfrac{\boldsymbol{u}\,\boldsymbol{u}^T}{c} \\ \boldsymbol{u} \quad &= \quad \boldsymbol{x} \; + \; \operatorname{sgn}(x_1) \, \|\boldsymbol{x}\|_2 \, \boldsymbol{e}^1 \\ c \quad &= \quad \|\boldsymbol{x}\|_2^2 \; + \; |x_1| \;\; \|\boldsymbol{x}\|_2 \, . \end{aligned} \tag{9.20}$$

Schritt 1 ist durchgeführt, wenn wir $\boldsymbol{H}_1$ wählen gemäß Lemma 9.5' mit $\boldsymbol{x} = \boldsymbol{a}^1$ (1. Spalte von $\boldsymbol{A}$). Da $\boldsymbol{H}_1$ regulär ist, gilt für $\boldsymbol{A}^{(1)} = \boldsymbol{H}_1 \, \boldsymbol{A}$

$$\text{Rang } \boldsymbol{A}^{(1)} \quad = \quad \text{Rang } \boldsymbol{A} \, .$$

Im 2. Schritt wenden wir das gleiche Verfahren an auf die Matrix, die durch Streichen der 1. Zeile und 1. Spalte von $\boldsymbol{A}^{(1)} = \boldsymbol{H}_1 \, \boldsymbol{A}$ entsteht, usw.

Schritt $\boldsymbol{k}$+1 Nach k Schritten hat $\boldsymbol{A}^{(k)} = \boldsymbol{H}_k \cdot \ldots \cdot \boldsymbol{H}_1 \boldsymbol{A}$ die Gestalt

$$\boldsymbol{A}^{(k)} = \underbrace{\left(\begin{array}{ccc|c} * & \cdots & \cdots & \cdots \quad \cdots \; * \\ & \ddots & & \vdots \\ & & * & \cdots \quad \cdots \; * \\ \hline & \mathbf{O} & & \widetilde{\boldsymbol{A}}^{(k)} \end{array}\right)}_{k} \begin{matrix} \left.\vphantom{\begin{matrix} a\\a\\a \end{matrix}}\right\} k \\ \\ \end{matrix} \in \mathrm{I\!R}^{m\times n}, \quad \tilde{\boldsymbol{A}}^{(k)} \in \mathrm{I\!R}^{(m-k)\times(n-k)}$$

Wir wenden das Vorgehen aus Schritt 1 an auf die Matrix $\tilde{\boldsymbol{A}}^{(k)}$, die aus $\boldsymbol{A}^{(k)}$ entsteht durch Streichen der ersten k Zeilen und Spalten. Wir realisieren dies durch

$$\boldsymbol{H}_{k+1} = \left(\begin{array}{c|c} \underbrace{\begin{matrix} 1 & & \\ & \ddots & \\ & & 1 \end{matrix}}_{k} & \mathbf{O} \\ \hline \mathbf{O} & \widetilde{\boldsymbol{H}}_{k+1} \end{array}\right) \in \mathrm{I\!R}^{m\times m}.$$

Dann bleiben durch Anwenden von $\boldsymbol{H}_{k+1}$ auf $\boldsymbol{A}^{(k)}$ die ersten k Zeilen und Spalten unverändert. $\widetilde{\boldsymbol{H}}_{k+1}$ wählt man entsprechend (9.20):

$$\begin{aligned}
\widetilde{\boldsymbol{H}}_{k+1} &= \boldsymbol{E}_{m-k} - \frac{\tilde{\boldsymbol{u}}^{k+1}\,\tilde{\boldsymbol{u}}^{k+1^T}}{c_{k+1}} \\
\tilde{\boldsymbol{a}}^{k+1} &= \left(a^{(k)}_{k+1,k+1}, \ldots, a^{(k)}_{m,k+1}\right)^T = \text{1. Spalte von } \widetilde{\boldsymbol{A}}^{(k)} \\
\tilde{\boldsymbol{u}}^{k+1} &= \tilde{\boldsymbol{a}}^{k+1} + \operatorname{sgn}\left(a^{(k)}_{k+1,k+1}\right) \left\|\tilde{\boldsymbol{a}}^{k+1}\right\|_2 \tilde{\boldsymbol{e}}^1 \\
& \qquad \tilde{\boldsymbol{e}}^1 = \text{erster Einheitsvektor} \in \mathrm{I\!R}^{m-k} \\
c_{k+1} &= \left\|\tilde{\boldsymbol{a}}^{k+1}\right\|_2^2 + \left|a^{(k)}_{k+1,k+1}\right| \left\|\tilde{\boldsymbol{a}}^{k+1}\right\|_2
\end{aligned}$$

und wieder gilt Rang $\boldsymbol{A}^{(k+1)} = $ Rang $\boldsymbol{A}^{(k)} = $ Rang $\boldsymbol{A}$.

Beachte: Rechentechnisch wird natürlich nur die „Restmatrix“ $\tilde{A}^{(k)}$ umgeformt.

Zusammenfassend erhält man also folgenden

Satz 9.6 Householder–Verfahren
Vorgelegt sei die Ausgleichsaufgabe:

$$\text{Bestimme} \quad \min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\,\boldsymbol{x}-\boldsymbol{b}\|_2, \;\; \boldsymbol{A}\in\mathbb{R}^{m\times n}, \;\; m>n, \;\; \boldsymbol{b}\in\mathbb{R}^m, \;\; \text{Rang } \boldsymbol{A}=n\,.$$

Dann gibt es Householdermatrizen $\boldsymbol{H}_1,\ldots,\boldsymbol{H}_n$ gemäß den Schritten $1,\ldots,n$ derart, daß mit $\boldsymbol{F}=\boldsymbol{H}_n\cdot\ldots\cdot\boldsymbol{H}_1$ gilt

$$\|\boldsymbol{F}(\boldsymbol{A}\,\boldsymbol{x}-\boldsymbol{b})\|_2 \;=\; \|\boldsymbol{F}\,\boldsymbol{A}\,\boldsymbol{x}-\boldsymbol{F}\,\boldsymbol{b}\|_2 \;=\; \|\boldsymbol{A}^{(n)}\boldsymbol{x}-\boldsymbol{b}^n\|_2 \;=\; \|\boldsymbol{A}\,\boldsymbol{x}-\boldsymbol{b}\|_2$$

mit

$$\boldsymbol{F}\,\boldsymbol{A}\,\boldsymbol{x}-\boldsymbol{F}\,\boldsymbol{b} \;=\; \begin{pmatrix} \widetilde{\boldsymbol{A}}^{(n)} \\ \\ \\ \mathbf{O} \end{pmatrix} \boldsymbol{x} \;-\; \begin{pmatrix} \tilde{\boldsymbol{b}}^n \\ \\ \\ \hat{\boldsymbol{b}}^n \end{pmatrix}, \qquad \widetilde{\boldsymbol{A}}^{(n)}\in\mathbb{R}^{n\times n}, \;\; \tilde{\boldsymbol{b}}^n\in\mathbb{R}^n\,.$$

Die Lösung der Ausgleichsaufgabe erhält man als Lösung des Gleichungssystems

$$\widetilde{\boldsymbol{A}}^{(n)}\boldsymbol{x} \;=\; \tilde{\boldsymbol{b}}^n\,.$$

Folgerung und Definition 9.7
Das Householder–Verfahren liefert für eine Matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$, $m\geq n$ eine Zerlegung

$$\boldsymbol{A} \;=\; \boldsymbol{Q}\,\boldsymbol{R} \qquad (\boldsymbol{Q}\boldsymbol{R}\text{–}\textit{Zerlegung})$$

mit einer rechten oberen Dreiecksmatrix $\boldsymbol{R}=\boldsymbol{F}\,\boldsymbol{A}$ (vgl. Satz 9.6) und einer orthogonalen Matrix $\boldsymbol{Q}=\boldsymbol{F}^{-1}$ (vgl. (9.15)).

Bemerkung:
Natürlich läßt sich das Householder–Verfahren auch anwenden zur Lösung eines linearen Gleichungssystems $\boldsymbol{A}\,\boldsymbol{x}=\boldsymbol{b}$, $\boldsymbol{A}\in\mathbb{R}^{n\times n}$, $\boldsymbol{b}\in\mathbb{R}^n$.

Wegen

$$\|\boldsymbol{F}\,\boldsymbol{A}\|_2 \;=\; \sup_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{F}\,\boldsymbol{A}\,\boldsymbol{x}\|_2 \;=\; \sup_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{A}\,\boldsymbol{x}\|_2 \;=\; \|\boldsymbol{A}\|_2\,,$$

$$\|(\boldsymbol{F}\,\boldsymbol{A})^{-1}\|_2 \;=\; \|\boldsymbol{A}^{-1}\,\boldsymbol{F}^{-1}\|_2 \;=\; \sup_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{A}^{-1}\,\boldsymbol{F}^{-1}\,\boldsymbol{x}\|_2 \;=\; \sup_{\|\boldsymbol{y}\|_2=1}\|\boldsymbol{A}^{-1}\,\boldsymbol{y}\|_2 \;=\; \|\boldsymbol{A}^{-1}\|_2$$

ist Kond $(\boldsymbol{A}) = $ Kond $(\boldsymbol{F}\,\boldsymbol{A})$ (vgl. Def. 8.20).

Das Householder–Verfahren verschlechtert also nicht die Kondition von A, im Gegensatz zum GEV. Dies muß allerdings mit einem größeren Rechenaufwand bezahlt werden. Man kann zeigen, daß das Householder–Verfahren doppelt so viele Multiplikationen benötigt wie das GEV.

§ 10 Approximation von Funktionen

Wir wiederholen zunächst die schon aus § 9, (9.5) bekannte Formulierung einer Approximationsaufgabe in etwas allgemeinerem Rahmen.

Definition 10.1
Sei $(X, \| \ \|)$ ein normierter Raum (über $\mathbb{R}$ oder $\mathbb{C}$), $f \in X$, $V \subset X$, $V \neq \emptyset$ eine Teilmenge. Jedes Element $\hat{v} \in V$ mit

$$\|f - \hat{v}\| \ \leq \ \|f - v\| \qquad \forall\, v \in V \tag{10.1}$$

heißt *beste Approximation für f (Minimallösung).*

$$\operatorname{dist}(f, V) \quad = \quad \inf_{v \in V} \|f - v\| \quad =: \quad \rho_V(f) \tag{10.2}$$

heißt *Minimalabstand* (*Minimalabweichung*).
Das Problem, die beste Approximation für f zu finden, heißt *Approximationsaufgabe.*
Die Approximationsaufgabe heißt *linear*, falls V ein linearer Teilraum von X ist.

Wir belegen zunächst durch Beispiele, daß die Lösung der Aufgabe — und damit die gesamte Approximationstheorie — normabhängig ist.

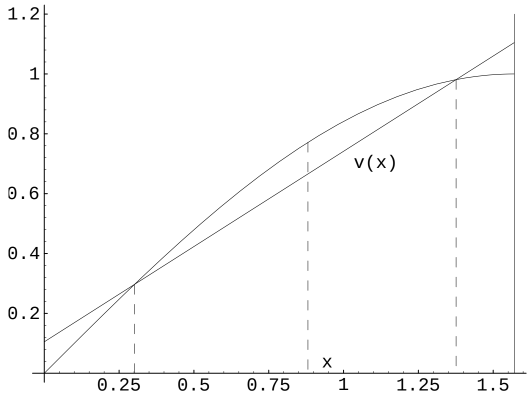
Beispiel 1:
Die Darstellung von Funktionen im Rechner (z.B. trigonometrische Funktionen, Exponentialfunktionen, Wurzelfunktionen) verlangt selbstverständlich die Maximumnorm

$$\|f\|_\infty \ = \ \max_{x \in [a,b]} |f(x)|\,, \qquad \text{(auch } \textit{Tschebyscheff–Norm}, \ \textit{T–Norm} \text{ genannt)}$$

da man sicherstellen muß, daß über den gesamten Approximationsbereich ein Maximalfehler der Funktionswerte nicht überschritten wird (*gleichmäßige Approximation*).

Schon ein einfaches Unterbeispiel zeigt, daß die Bezeichnung *lineare* Approximationsaufgabe insofern irreführend sein kann, als sie sich keineswegs durch lauter lineare Aufgaben lösen läßt.

Beispiel 1a)



Bestimme für $f(x) = \sin x \quad \left(f \in C\left[0, \frac{\pi}{2}\right]\right)$ die beste Approximation

$$\hat{v}(x) = ax + b \quad \left(\hat{v} \in \Pi_1 \left[0, \frac{\pi}{2}\right]\right)$$

bzgl. der T–Norm.

Es müssen a, b so bestimmt werden, daß gilt

$$\left\{ \begin{array}{l} \hat{v}(0) - \sin 0 \;=\; \sin \overline{x} - \hat{v}(\overline{x}) \;=\; \hat{v}\left(\frac{\pi}{2}\right) - \sin \frac{\pi}{2}\,, \\[2ex] \overline{x} \quad \text{so, daß} \quad \max\limits_{x \in [x_1, x_2]} |\hat{v}(x) - \sin x| \;=\; \sin \overline{x} - \hat{v}(\overline{x})\,, \end{array} \right. \tag{10.3}$$

denn jede andere Wahl der Geraden würde eines der Extrema verschlechtern.

Durch Einsetzen von $\hat{v}(x)$ für $x = 0$ und $x = \frac{\pi}{2}$ folgt zunächst

$$b \;=\; a\frac{\pi}{2} + b - 1\,, \quad \text{also} \quad a \;=\; \frac{2}{\pi}\,.$$

Eine notwendige Bedingung für $\overline{x}$ (Max–Bedingung) ist

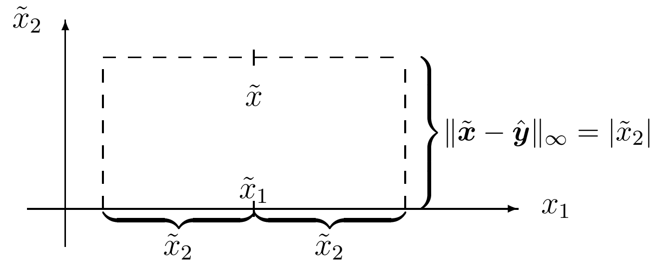
$$(\hat{v}(\overline{x}) - \sin \overline{x})' \;=\; \frac{2}{\pi} - \cos \overline{x} \;=\; 0\,.$$

Dies ist eine *nichtlineare* Gleichung zur Bestimmung von $\overline{x}$. Nach Bestimmung eines geeigneten $\overline{x}$, kann b bestimmt werden gemäß(10.3) aus

$$\sin \overline{x} - \hat{v}(\overline{x}) \;=\; \hat{v}\left(\frac{\pi}{2}\right) - 1\,.$$

Beispiel 1b)

Ein weiteres simples Beispiel zeigt, daß die beste Approximation bzgl. der T–Norm keineswegs eindeutig sein muß.



Bestimme für ein

$$\tilde{\boldsymbol{x}} \;=\; (\tilde{x}_1, \tilde{x}_2)^T \in \mathbb{R}^2, \quad \tilde{x}_2 \neq 0$$

die beste Approximation $\hat{\boldsymbol{y}} \in \mathbb{R}$ bzgl. $\|\ \|_\infty$.

Offensichtlich gilt für alle $\hat{y}_1 \in [\tilde{x}_1 - \tilde{x}_2, \tilde{x}_1 + \tilde{x}_2]$

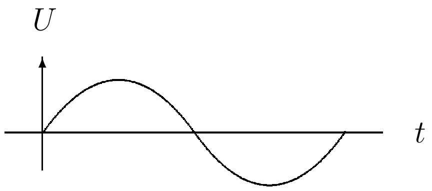
$$\left\|\tilde{\boldsymbol{x}} - (\hat{y}_1, 0)^T\right\|_\infty \;=\; |\tilde{x}_2| \;=:\; \rho(\tilde{\boldsymbol{x}})\,.$$

Daß die T–Norm nicht allein seligmachend sein kann, haben wir schon bei der Lösung der Ausgleichsaufgabe in § 9 gesehen. Wir führen eine weitere Approximationsaufgabe an, welche deutlich macht, daß eine Behandlung mit der T–Norm nicht angemessen sein kann.

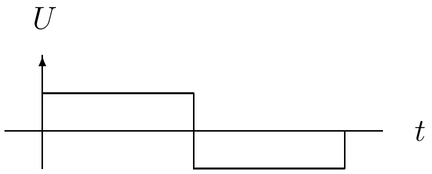
Beispiel 2:

In einem Synthesizer werden elektrische Schwingungen (periodische Veränderungen der Spannung U über der Zeit t) erzeugt, die der Benutzer dann nach eigener Wahl überlagern, modulieren oder auf andere Weise bearbeiten kann. Der Benutzer hat dabei virtuell 4 Schwingungsformen zur Verfügung:

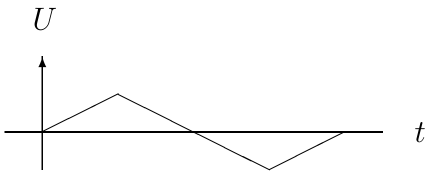
a) die Sinusschwingung



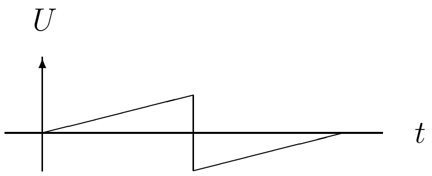
b) die Rechteckschwingung



c) die Dreieckschwingung



d) die Sägezahnschwingung



Tatsächlich werden jedoch die letzten drei Formen durch die Überlagerung von Sinusschwingungen erzeugt. Die Anzahl der Sinusgeneratoren , die z.B. für einen Rechteckgenerator verwendet werden, hängt dabei vermutlich von der Preisklasse des Synthesizers ab (ist aber sicher stets endlich).

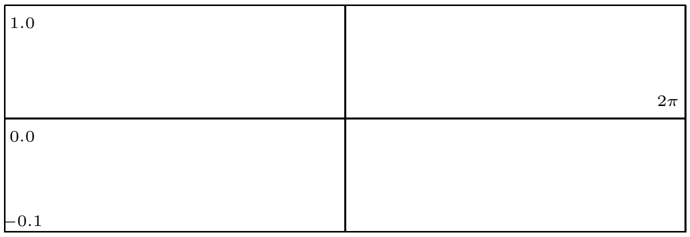
Wenn wir die Schwingungen als 2π–periodisch ansehen, können wir die Aufgabe, durch Sinusgeneratoren einen Rechteckgenerator zu simulieren, mathematisch so formulieren:
Gegeben ist eine Funktion $r \in C^{-1}[0, 2\pi]$ (= Menge der integrierbaren, aber nicht notwendig stetigen Funktionen auf dem Intervall $[0, 2\pi]$). Gesucht ist eine Funktion $s(x) = a_1 \sin(b_1 x) + \ldots + a_p \sin(b_p x)$ ($p \in \mathbb{N}$ gegeben), die sich von $r(x)$ möglichst wenig unterscheiden. Die Menge der als Linearkombination zur Verfügung stehenden Sinusfunktionen bezeichnen wir im weiteren mit $S[0, 2\pi]$.

Welches ist nun die für dieses Beispiel „angemessene“ Norm?

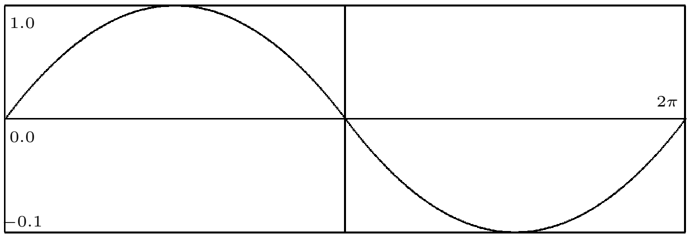
Wählen wir etwa f als Rechteckschwingung mit der Amplitude 1 und als Norm die *Maximumnorm*, dann gilt für jede stetige Funktion g:

$$\max_{0 \le t \le 2\pi} |f(t) - g(t)| \;\ge\; 1 \qquad \text{(Warum?)}\,.$$

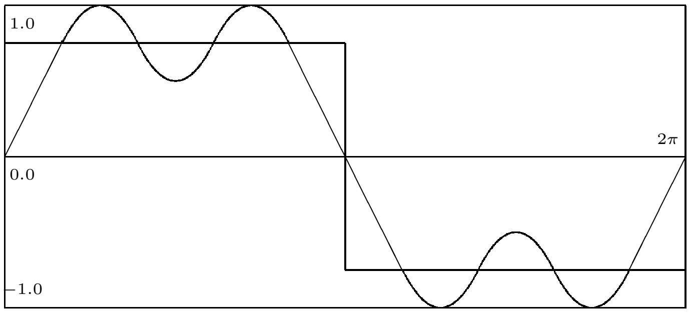
Jede Funktion $\overline{s} \in S[0, 2\pi]$ mit $\|f - \overline{s}\|_\infty = 1$ ist also eine beste Approximation. Einige Beispiele solcher Funktionen sind im folgenden angegeben. Obwohl alle diese Funktionen im Sinne unserer Definition beste Approximationen sind, ist wohl klar, daß wir sie von unserer ursprünglichen Fragestellung her nicht als gleichwertige Lösungsmöglichkeiten ansehen können:



$$\overline{s}_1(t) \;\equiv\; 0$$



$$\overline{s}_2(t) \;=\; \sin\, t$$



$$\overline{s}_3(t) \;=\; \tfrac{4}{\pi}\,\sin\, t + \tfrac{4}{3\pi}\,\sin\, 3t$$

Dieser unbefriedigende Zustand spricht dafür, eine andere Norm zu wählen. Insbesondere das 3. Beispiel legt nahe, eine Norm zu wählen, die den Betrag der Fläche zwischen Rechtecksfunktion und Approximation mißt. In Frage kommen u.a. die

$$L_1\text{–Norm} \qquad \|u\|_1 \;:=\; \int_0^{2\pi} |u(t)|\,dt$$

oder die

$$L_2\text{–Norm} \qquad \|u\|_2 \;:=\; \left(\int_0^{2\pi} |u(t)|^2\,dt\right)^{1/2} .$$

Die L_2–Norm ergibt sich aus dem Skalarprodukt

$$(u,v) \quad := \quad \int_0^{2\pi} u(t)\,v(t)\,dt$$

und bietet insbesondere den Vorteil, daß der Projektionssatz 9.1 angewandt werden kann. Deshalb beschäftigen wir uns zunächst mit der

Funktionsapproximation in unitären Räumen

Vorgelegt sei ein unitärer Funktionenraum $(X, (\ ,\ ))$. Die Norm wird durch das Skalarprodukt gegeben: $\|f\| = \sqrt{(f,f)}$. Eine Funktion f wird durch eine Linearkombination *endlich vieler* Funktionen u^j, $j = 1, \ldots, n$, welche einen linearen Teilraum V_n von X aufspannen, approximiert: $\hat{v}(x) = \sum_{j=1}^{n} v_j\, u^j(x)$, $v_j \in \mathbb{R}$.

Die Existenz und Eindeutigkeit der besten Approximation $\hat{v}$ für f ist durch Satz 9.1 (Projektionssatz) bereits gesichert. Er liefert auch die Berechnungsvorschrift für $\hat{v}$ in Form der *Normalgleichungen*

$$(\hat{v}, u^j) \;=\; \sum_{i=1}^{n} v_i (u^i, u^j) \;=\; (f, u^j)\,, \qquad j = 1, \ldots, n\,. \tag{10.4}$$

Ausschlaggebend für die numerische Qualität ihrer Lösung ist die Kondition der Koeffizientenmatrix $\left((u^i, u^j)\right)_{i,j=1,\ldots,n}$ (*Gram'sche Matrix*). Sie wird bestimmt durch die Basiselemente u^j von V_n. Besonders einfach wird die Lösung von (10.4), wenn die Basiselemente u^j ein *Orthogonal*system bilden (d.h. $(u^i, u^j) = 0$ für $i \neq j$). Dann wird die Gram'sche Matrix zu einer Diagonalmatrix und man erhält die Lösung von (10.4) in der Form

$$v_i \;=\; \frac{1}{(u^i, u^i)}\,(f, u^i), \qquad i = 1, \ldots, n\,. \tag{10.5}$$

Noch bequemer ist es, wenn die Vektoren u^j ein *Orthonormal*system bilden (d.h. $(u^i, u^j) = \delta_{ij}$, *Kroneckersymbol*), was sich leicht durch die Normierung

$$\tilde{u}^j \;=\; \frac{u^j}{\sqrt{(u^j, u^j)}}\,, \qquad j = 1, \ldots, n$$

erreichen läßt.

Dann ist die Gram'sche Matrix gleich der Einheitsmatrix, welche bzgl. jeder einer Vektornorm zugeordneten Matrixnorm die Norm 1 hat und damit auch die bestmögliche Kondition 1. Beispiele für diesen Fall liefert die

Trigonometrische Approximation

Das prominenteste Beispiel der Funktionsapproximation in unitären Räumen ist die *Fourier–Approximation*. Sie wird besonders gerne benutzt, um periodische Funktionen p durch eine Linearkombination von sin– und cos–Schwingungen zu approximieren (vgl. Bsp. 2). Als Raum wählt man $X = C^{-1}[-\pi, \pi]$ (= Raum der auf $[-\pi, \pi]$ integriebaren, nicht notwendig stetigen Funkionen). Als Approximationsraum V_{2n+1} der Dimension $2n+1$ wird der von den Funktionen

$$1,\ \cos x,\ \sin x,\ \cos 2x,\ \sin 2x,\ \ldots,\ \cos nx,\ \sin nx \tag{10.6}$$

aufgespannte Raum gewählt. Das Skalarprodukt wird gegeben durch

$$(u, v) \;=\; \int_{-\pi}^{\pi} u(x)\, v(x)\, dx \tag{10.7}$$

Satz 10.2
Die trigonometrischen Funktionen (10.6) bilden für das Intervall $[-\pi, \pi]$ bzgl. des Skalarprodukts (10.7) ein Orthogonalsystem. Es gilt für $j, k \in \mathbb{N}_0$

$$\int_{-\pi}^{\pi} \cos(jx)\ \cos(kx)\, dx \;=\; \begin{cases} 0 & \text{für alle} \quad j \neq k \\ 2\pi & \text{für} \quad j = k = 0 \\ \pi & \text{für} \quad j = k > 0 \end{cases}$$

$$\int_{-\pi}^{\pi} \sin(jx)\ \sin(kx)\, dx \;=\; \begin{cases} 0 & \text{für alle} \quad j \neq k \\ \pi & \text{für} \quad j = k > 0 \end{cases}$$

$$\int_{-\pi}^{\pi} \cos(jx)\ \sin(kx)\, dx \;=\; 0 \quad \text{für alle} \quad j,\, k$$

Beweis: Nachrechnen oder Schwarz § 4.1

Wegen Satz 10.2 setzt man die beste Approximation $\hat{v}$ für f an in der Gestalt

$$\hat{v}(x) \;=\; \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx) \quad (\textit{Fourierpolynom}) \tag{10.8}$$

und erhält die *Fourierkoeffizienten* a_k, b_k aus (10.5) und Satz 10.2

$$\boxed{\begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\ \cos(kx)\, dx, \qquad k = 0, 1, \ldots, n \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\ \sin(kx)\, dx, \qquad k = 1, 2, \ldots, n \end{aligned}} \tag{10.9}$$

Sie müssen üblicherweise durch numerische Integration berechnet werden.

Bemerkung
Ob man das Intervall $[-\pi, \pi]$ oder $[0, 2\pi]$ oder ein beliebiges anderes Intervall der Länge 2π wählt, ist auf Grund der Periodizität der Funktionen (10.6) gleichgültig. Hat f ein anderes Periodenintervall der Länge 2ℓ, so benutzt man statt (10.6) die Funktionen

$$1, \quad \sin \frac{k\,\pi}{\ell}\, x, \quad \cos \frac{k\,\pi}{\ell}\, x, \quad k = 1, \ldots, n$$

welche die Periode 2ℓ haben. Betrachtet man das Intervall $[-\ell, \ell]$, so ergeben sich die

Fourierkoeffizienten zu

$$\boxed{\begin{aligned} a_k &= \frac{1}{\ell}\int\limits_{-\ell}^{\ell} f(x)\,\cos\frac{k\pi}{\ell}x\,dx, \qquad k=0,1,\ldots,n \\ b_k &= \frac{1}{\ell}\int\limits_{-\ell}^{\ell} f(x)\,\sin\frac{k\,\pi}{\ell}x\,dx, \qquad k=1,2,\ldots,n \end{aligned}}$$

Aufgabe:
Zeige: Das Fourierpolynom für Beispiel 2) ergibt sich als

$$\begin{aligned} \hat{v}_n(x) &= \frac{4}{\pi}\sum_{k=1}^{\left[\frac{n+1}{2}\right]} \frac{\sin(2k-1)x}{2k-1}, \\ \left[\frac{n+1}{2}\right] &= \text{größte ganze Zahl} \quad \le \frac{n+1}{2}. \end{aligned}$$

Mit den bisher behandelten Punkten ist der Komplex der Fourierapproximation noch keineswegs erschöpfend behandelt. Wir erwähnen einige der ausstehenden Fragen, auf die wir in dieser Vorlesung jedoch nicht eingehen.

1) Konvergenz des Fourierpolynoms für $n \to \infty$ gegen f bzgl. der Norm $\|u\|_{L_2} = \left(\int\limits_{-\pi}^{\pi} u^2(x)\,dx\right)^{1/2}$ (*Konvergenz im quadratischen Mittel*)?

2) Inwieweit impliziert die Konvergenz des Fourierpolynoms bzgl. der L_2–Norm die punktweise Konvergenz?

3) Wie kann man die Fourierkoeffizienten schnell und effizient berechnen?

4) Anwendung der Fourierapproximation zur numerischen Integration.

Approximation durch orthogonale Polynome.
Daß die Normalgleichungen nicht notwendig so schöne Eigenschaften haben wie bei der Fourierapproximation, zeigt das

Beispiel:
Approximiere $f \in C[0,1]$ durch die Funktionen

$$1,\ x,\ x^2,\ \ldots,\ x^n, \tag{10.10}$$

(die einen linearen $(n+1)$–dimensionalen Teilraum von $C[0,1]$ bilden) bzgl. der Norm, die durch $(u,v) = \int\limits_0^1 u(x)\,v(x)dx$ induziert wird.

Die Berechnung der Gram'schen Matrix (Aufgabe) liefert die „Horrormatrix“ (bzgl. der Kondition)

$$H = \left(\frac{1}{i+k-1} \right)_{i,k=1,\ldots,n+1} \qquad \textit{Hilbertmatrix}\,.$$

Abhilfe

Zur Lösung des Beispiels berechne man eine Basis aus orthogonalen Polynomen (Orthogonalisierungsverfahren von Erhard Schmidt) und löse die Approximationsaufgabe bzgl. dieser neuen Basisvektoren.

Beispiele für orthogonale Polynomräume

Da die Tschebyscheff–Polynome (vgl. § 3) durch trigonometrische Funktionen definiert wurden, ist zu erwarten, daß auch sie Orthogonalitätseigenschaften besitzen. In der Tat gilt

Satz 10.3

Die durch

$$T_n(x) = \cos(n \arccos x)\,, \qquad x \in [-1,1]$$

definierten T–Polynome erfüllen bzgl. des Skalarprodukts

$$(u,v) = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\, u(x)\, v(x)\, dx$$

mit der *Gewichtsfunktion*

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

die Orthogonalitätseigenschaften

$$(T_k(x)\,, T_j(x)) = \left\{ \begin{array}{lll} 0\,, & \text{falls} & k \neq j \\ \frac{1}{2}\pi\,, & \text{falls} & k = j > 0 \\ \pi\,, & \text{falls} & k = j = 0 \end{array} \right\}, \qquad k,j \in \mathbb{N}_0\,.$$

Beweis: Schwarz § 4.3

Orthogonalisiert man die Polynome $1,\, x,\, x^2,\, \ldots$ im Intervall $[-1,1]$ bzgl. $(u,v) = \int_{-1}^{1} u(x)\, v(x)\, dx$, so erhält man die

$$\textit{Legendre–Polynome} \qquad P_n(x) = \frac{1}{2^n\, n!}\, \frac{d^n}{d\,x^n} \left[(x^2-1)^n \right], \qquad n \in \mathbb{N}_0$$

mit der Orthogonalitätsrelation

$$\int_{-1}^{1} P_m(x)\, P_n(x)\, dx \quad = \quad \left\{ \begin{array}{ll} 0 & \text{falls} \quad m \neq n \\ \frac{2}{2n+1} & \text{falls} \quad m = n \end{array} \right\} \quad m, n \in \mathbb{N}_0 \,.$$

Beweis: Schwarz § 4.3.3.

Bemerkung:
Die Legendre–Polynome werden benutzt zur Konstruktion von Gauß–Quadraturformeln (vgl. Werner § 4, Opfer § 4, Stoer § 3.5).

Gleichmäßige Funktionsapproximation

Unter „gleichmäßig" versteht man üblicherweise die Approximation bzgl. der Maximumnorm, allgemeiner aber auch die Approximation bzgl. einer Norm, die nicht durch ein Skalarprodukt induziert wird. Deshalb ist in diesem Fall die Existenz einer Minimallösung noch nicht gesichert, ebensowenig wie die Eindeutigkeit (vgl. die Beispiele). Wir zeigen zunächst die Existenz.

Satz 10.4 Existenz einer Minimallösung
Sei $(X, \|\,\|)$ ein normierter Raum über $\mathbb{R}$ (oder $\mathbb{C}$) und $V \subset X$ ein n–dimensionaler Teilraum. Dann folgt

$$\forall\, f \in X \;\exists \text{ (mindestens) eine } \textit{Minimallösung } \hat{v} \in V, \quad \text{d.h.}$$

$$\|f - \hat{v}\| \;\leq\; \|f - v\| \quad \forall\, v \in V\,.$$

$$\rho_V(f) \;=\; \|f - \hat{v}\| \quad \text{heißt } \textit{Minimalabstand}.$$

Beweis: (für $\mathbb{R}$, er verläuft analog für $\mathbb{C}$).
Jedes $v \in V$ hat eine Darstellung als Linearkombination der Basiselemente v^j von V:

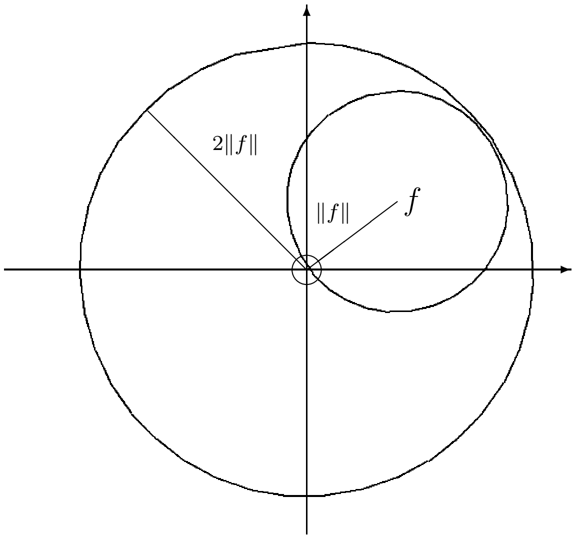
$$v \;=\; \sum_{j=1}^{n} \alpha_j\, v^j\,, \qquad \alpha_j \in \mathbb{R}\,.$$

Die Null ist eine Approximation für f und

$$\left\{ \begin{array}{l} \text{jedes } \; v \in V \;\; (\text{sogar } \; v \in X) \;\; \text{mit } \; \|v\| > 2\|f\| \;\; \text{ist} \\ \text{eine schlechtere Approximation für } \; f \;\; \text{als die Null,} \end{array} \right. \tag{10.11}$$

denn

$$\|f - v\| \;\geq\; |\, \|v\| - \|f\|\, | \;\geq\; \|v\| - \|f\| \;>\; 2\|f\| - \|f\| \;=\; \|f\|\,.$$



Wir zeigen:

$$\exists \text{ Kugel } \quad K = \{\alpha \in \mathbb{R}^n\,;\ \|\alpha\|_2 \le k\} \ : $$
$$\|v\| \ = \ \left\| \sum_{j=1}^{n} \alpha_j \, v^j \right\| \ > \ 2\|f\| \quad \forall\, \alpha = (\alpha_1, \ldots, \alpha_n)^T \notin K\,. \tag{10.12}$$

$\psi(\beta) = \left\| \sum\limits_{j=1}^{n} \beta_j \, v^j \right\|$ ist stetig auf dem Kompaktum $\{\beta \in \mathbb{R}^n;\ \|\beta\|_2 = 1\}$ (vgl. Satz 8.3), nimmt dort also sein Minimum m an. Da die v^j linear unabhängig sind, ist $m > 0$. Das bedeutet insbesondere

$$\left\| \sum_{j=1}^{n} \frac{\alpha_j}{\|\alpha\|_2} \, v^j \right\| \ \ge \ m \qquad \forall\, \alpha \in \mathbb{R}^n,\ \alpha \ne 0$$

oder

$$\|v\| \ = \ \left\| \sum_{j=1}^{n} \alpha_j \, v^j \right\| \ \ge \ m\|\alpha\|_2\,. \tag{10.13}$$

Wählt man in (10.12) $k = \frac{2\|f\|}{m}$, so folgt aus $\alpha \notin K$ mit (10.13) $\|v\| > 2\|f\|$. Nach (10.11) kann dann v für $\alpha \notin K$ keine beste Approximation enthalten.

Wegen (10.12) genügt es, $\min\limits_{\alpha \in K} \left\| f - \sum\limits_{j=1}^{n} \alpha_j \, v^j \right\|$ zu suchen.

Da $\varphi(\alpha_1, \ldots, \alpha_n) = \left\| f - \sum\limits_{j=1}^{n} \alpha_j \, v^j \right\|$ eine stetige Funktion ist, nimmt sie auf der kompakten Menge K ihr Minimum an. ■

Im weiteren beschränken wir uns auf die

Approximation von Funktionen $f \in C[a,b]$ durch Polynome
bzgl. der Maximumnorm $\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$.

Im unitären Fall war eine Optimalitätsbedingung — die Senkrechtbeziehung (9.6) — charakteristisch für die Minimallösung $\hat{v}$ und lieferte darüberhinaus eine Berechnungsvorschrift für $\hat{v}$ (die Normalgleichungen). Im normierten Raum $(C[a,b], \|\ \|_\infty)$ steht kein Skalarprodukt zur Verfügung. Wir müssen also ein anderes Optimalitätskriterium für diesen Fall entwickeln.

Wir gehen aus von der Situation:
Sei $f \in C[a,b]$ und $p^* \in \Pi_{n-1}[a,b]$ (das ist ein n–dimensionaler Unterraum von $C[a,b]$) eine Näherung. Die Fehlerfunktion bezeichnen wir mit $d^*(x) = f(x) - p^*(x)$.

Dann gelten folgende Äquivalenzen (geschlossene Implikationskette):

$$p^* \quad \text{ist keine Minimallösung für} \quad f \tag{10.14}$$

$$\Rightarrow \quad \begin{cases} \exists \quad \text{Korrekturpolynom} \quad p \in \Pi_{n-1} \quad \text{mit} \\ |d^*(x) - p(x)| \ < \ |d^*(x)| \quad \forall x \in M \ = \ \{x \in [a,b];\ |d^*(x)| = \|f - p^*\|_\infty\} \\ \qquad\qquad\qquad\qquad\qquad\qquad (\textit{Extremalmenge}) \end{cases}$$

Dies gilt auf Grund der Definition von $\|\ \|_\infty$.

$$\Rightarrow \qquad \exists\, p \in \Pi_{n-1} : \quad 0 \neq \operatorname{sgn} d^*(x) \ = \ \operatorname{sgn} p(x) \quad (\text{bzw.} \quad d^*(x)\, p(x) \ > \ 0) \quad \forall\, x \in M$$

Bew. „$\Rightarrow$“: Wäre $\operatorname{sgn} d^*(\tilde{x}) = -\operatorname{sgn} p(\tilde{x})$ für ein $\tilde{x} \in M$,
so folgte $|d^*(\tilde{x}) - p(\tilde{x})| \ = \ |d^*(\tilde{x})| + |p(\tilde{x})| \ \geq \ |d^*(\tilde{x})|$. $\boldsymbol{W}$!

$\Rightarrow$ (10.14) Der Beweis wird in Lemma 10.9 nachgetragen.

Damit liefert die vorletzte Folgerung folgendes

Optimalitätskriterium
$\nexists\, p \in \Pi_{n-1} \ : \quad (f(x) - p^*(x))\ p(x) > 0 \quad \forall\, x \in M \quad \Longleftrightarrow \quad p^*$ ist optimal für f.

Wir müssen nur noch untersuchen, was dieses Kriterium für Polynome bedeutet. Die Eigenschaften der Polynome wurden bisher noch nicht ausgenutzt. Es gelten folgende äquivalente Umformungen:

$$\nexists\, p \in \Pi_{n-1} : \quad d^*(x)\, p(x) \; > \; 0 \quad \forall\, x \in M\,. \tag{10.15}$$

$$\iff \quad \begin{cases} \forall\, p \in \Pi_{n-1} \;\text{ hat }\; d^*(x) \;\text{ in }\; M \;\text{ mehr Vorzeichenwechsel als }\; p, \;\text{ wobei }\; M \\ \text{im Sinne wachsender }\; x \;\text{ durchlaufen wird.} \end{cases}$$

Bemerkung: (10.15) kann nur gelten, wenn d^* in M Vorzeichenwechsel hat.

Bew. „$\Leftarrow$“: klar, da p den Vorzeichenwechseln von d^* nicht folgen kann.
Bew. „$\Rightarrow$“ (indirekt): Hat d^* nicht mehr als $m \le n-1$ (= maximale Nullstellenanzahl von p) Vorzeichenwechsel, so gibt es Punkte $\xi_1, \ldots, \xi_m \in (a,b)\backslash M$, welche die Vorzeichenbereiche von M trennen (beachte: d^* ist stetig), d.h.

$$x_a \in M \cap [\xi_{i-1}, \xi_i], \;\; x_b \in M \cap [\xi_i, \xi_{i+1}] \;\Rightarrow\; \operatorname{sgn} d^*(x_a) = -\operatorname{sgn} d^*(x_b)\,.$$

Dann erfüllt $p(x) \; = \; \varepsilon \prod\limits_{j=1}^{m} (x - \xi_j)$ für $\varepsilon \; = \; 1$ oder -1 die Bedingung $\exists\, p \in \Pi_{n-1} : \quad d^*(x)\, p(x) \; > \; 0 \quad \forall\, x \in M.$

$$\iff \quad \begin{cases} \text{Es gibt }\; n+1 \;\text{ Punkte }\; t_j \in M : \;\; a \le t_0 < t_1 < \ldots < t_n \le b, \;\text{ so daß} \\ d^*(t_j) = f(t_j) - p^*(t_j), \;\; j = 0, 1, \ldots, n \;\text{ alternierendes Vorzeichen hat.} \end{cases}$$

denn d^* hat mindestens n Vorzeichenwechsel, d.h. mindestens einen mehr als die Maximalnullstellenzahl von $p \in \Pi_{n-1}$.

Definition 10.5

Eine Menge von $n+1$ Punkten $t_i : \; a \le t_0 < t_1 < \ldots < t_n \le b$ heißt *Alternante* (der Länge $n+1$) für $f \in C[a,b]$ und $p^* \in \Pi_{n-1}\,[a,b]$, falls

$$\operatorname{sgn}\left(f(t_j) - p^*(t_j)\right) \; = \; \varepsilon(-1)^j, \qquad j = 0, 1, \ldots, n, \quad \varepsilon = 1 \;\text{ oder }\; -1\,.$$

Die letzte der gezeigten Äquivalenzen, zusammen mit dem (noch nicht bewiesenen) Lemma 10.9, formulieren wir als

Satz 10.6 Alternantensatz

Sei $f \in C[a,b]$. Dann ist eine Näherung $p^* \in \Pi_{n-1}\,[a,b]$ für f bzgl. der Maximumnorm $\|\;\|_\infty$ genau dann Minimallösung, wenn eine Alternante der Länge $n+1$ für f und p^* existiert, so daß

für $a \; \le \; t_0 \; < \; t_1 \; < \; \ldots \; < \; t_n \; \le \; b$ gilt

$$f(t_j) - p^*(t_j) \; = \; \varepsilon(-1)^j\, \|f - p^*\|_\infty, \qquad j = 0, 1, \ldots, n, \quad \varepsilon = +1 \;\text{ oder }\; -1\,.$$

Beachte
Die Alternante muß nicht eindeutig sein (Beispiel?). Es kann aber die Eindeutigkeit der Minimallösung gezeigt werden. Sie hängt wesentlich ab von den Polynomeigenschaften, also den Eigenschaften der Funktionen aus dem Approximationsraum, und kann aus dem Optimalitätskriterium, dem Alternantensatz, gefolgert werden.

> **Folgerung 10.7**
> Die Minimallösung $p^* \in \Pi_{n-1}[a,b]$ für $f \in C[a,b]$ bzgl. der Maximumnorm ist eindeutig.

Beweis (indirekt)
Annahme: Es gibt 2 Minimallösungen $v_1, v_2 \in \Pi_{n-1}$, $v_1 \neq v_2$, d.h.

$$\rho(f) \quad = \quad \|f - v_1\|_\infty \quad = \quad \|f - v_2\|_\infty \,.$$

Dann ist auch

$$h \quad = \quad \frac{1}{2}\,(v_1 + v_2)$$

eine Minimallösung (allgemeiner: die Menge der Minimallösungen ist konvex), denn

$$\begin{aligned} \|f - h\|_\infty \quad &= \quad \left\|f - \tfrac{1}{2}\,(v_1 + v_2)\right\|_\infty \quad = \quad \left\|\tfrac{1}{2}\,(f - v_1) + \tfrac{1}{2}\,(f - v_2)\right\|_\infty \\ &\leq \quad \tfrac{1}{2}\,\|f - v_1\|_\infty \; + \; \tfrac{1}{2}\,\|f - v_2\|_\infty \quad = \quad \rho(f)\,. \end{aligned}$$

Gemäß Satz 10.6 hat h eine Alternante $t_0, t_1, \ldots, t_n$ so daß

$$f(t_j) - h(t_j) \;=\; \frac{1}{2}\,(f - v_1)\,(t_j) + \frac{1}{2}\,(f - v_2)\,(t_j) \;=\; \varepsilon(-1)^j\,\rho(f), \quad \forall\, j, \quad \varepsilon = 1 \text{ oder } -1\,,$$

insbesondere also

$$(*) \qquad \rho(f) = \left|\frac{1}{2}\,(f - v_1)\,(t_j) \;+\; \frac{1}{2}\,(f - v_2)\,(t_j)\right|.$$

Wegen

$$|(f - v_i)\,(t_j)| \;\leq\; \|f - v_i\|_\infty \;=\; \rho(f), \quad i = 1,2,\; j = 0,1,\ldots,n$$

folgt mit $(*)$ $\quad \dfrac{1}{2}|(f - v_1)\,(t_j)| + \dfrac{1}{2}|(f - v_2)\,(t_j)| \;=\; \rho(f),$

d.h. die t_j müssen für v_1 und v_2 aus der Extremalmenge sein und es ist

$$|(f - v_1)\,(t_j)| = |(f - v_2 1)\,(t_j)| \quad \forall j.$$

Wegen $(*)$ folgt daraus

$$(f - v_1)\,(t_j) \quad = \quad (f - v_2)\,(t_j)\,,$$

bzw.

$$v_1(t_j) \;=\; v_2(t_j) \qquad j = 0,1,\ldots,n\,,$$

und nach dem Lemma 2.5 ist wegen $v_i \in \Pi_{n-1}$: $v_1 - v_2 \equiv 0$. $W!$ ∎

Die Minimalabweichung $\rho(f) = \|f - p^*\|_\infty$ im Alternantensatz ist im allgemeinen nicht bekannt. Daher ist folgende Abschwächung des Satzes interessant, die sowohl untere und obere Schranken für den Minimalabstand als auch die Grundlage zu einem numerischen Verfahren zur Bestimmung der Minimallösung liefert.

Satz 10.8

a) Für $f \in C[a,b]$, $p \in \Pi_{n-1}[a,b]$, sei $a \le t_0 < t_1 < \ldots < t_n \le b$ eine Alternante, d.h.

$$(f-p)(t_j)\,, \qquad j = 0,1,\ldots,n \quad \text{ist alternierend}\,.$$

Dann folgt

$$\min_{j=0,\ldots,n} |(f-p)(t_j)| \quad \le \quad \rho(f) \;:=\; \min_{q \in \Pi_{n-1}[a,b]} \|f-q\|_\infty \;\le\; \|f-p\|_\infty\,. \tag{10.16}$$

b) Zu jeder Unterteilung $a \le x_0 < x_1 < \ldots < x_n \le b$, (also insbesondere zur Alternante aus a)) existieren genau ein $\mu \in \mathbb{R}$ und ein $\tilde{p} \in \Pi_{n-1}[a,b]$: $\tilde{p}(x) = \sum\limits_{\nu=0}^{n-1} a_\nu\, x^\nu$, so daß gilt

$$(f-\tilde{p})(x_j) \;=\; (-1)^j\,\mu\,, \qquad j = 0,1,\ldots,n\,, \tag{10.17}$$

$$|\mu| \;=\; \min_{j=0,\ldots,n} |(f-\tilde{p})(x_j)| \;\le\; \rho(f)\,. \tag{10.18}$$

Bilden die x_j auch eine Alternante für p und f, so gilt

$$\min_{j=0,\ldots,n} |(f-p)(x_j)| \;\le\; |\mu|\,. \tag{10.19}$$

Bedeutung

1. (10.17), (10.18) sichern die Erfüllbarkeit der Voraussetzung aus a), falls $\mu \neq 0$.
2. (10.19) zeigt, daß das gemäß (10.17) berechnete Polynom $\tilde{p}$ bzgl. einer festen Alternante $\{x_j\}$ die bestmögliche untere Schranke für die Minimalabweichung liefert. $\|f-\tilde{p}\|_\infty$ liefert eine obere Schranke für $\rho(f)$.

Beweis: Wir zeigen zunächst Teil b). Dann folgt die erste Ungleichung von (10.16) aus (10.18) und (10.19). Die 2. Ungleichung (10.16) ist trivial.

Beweis (10.17). Dieses System für die Unbekannten $a_0, \ldots, a_{n-1}, \mu$ hat eine eindeutige Lösung, falls gezeigt wird, daß das homogene System

$$\tilde{p}(x_j) \;+\; (-1)^j\,\mu \;=\; 0, \qquad j = 0,1,\ldots,n$$

nur die Nullösung hat.

Ist $\mu = 0$, so hat $\tilde{p}$ $n+1$ Nullstellen, also ist $\tilde{p} \equiv 0$.
Ist $\mu \neq 0$, so hat $\tilde{p}$ in jedem Intervall $[x_j, x_{j+1}]$, $j = 0,\ldots,n-1$ mindestens eine Nullstelle, insgesamt also n Stück, also $\tilde{p} \equiv 0$ und damit auch $\mu = 0$, also $\boldsymbol{W}$!

Beweis (10.18) (indirekt). Sei p^* Minimallösung für f.

Annahme: $\min\limits_{j=0,\dots,n} |(f-\tilde{p})(x_j)| > \|f-p^*\|_\infty = \rho(f)$.

$$\begin{aligned}
&\text{Es ist} && (p^*-\tilde{p})(x_j) = (f-\tilde{p})(x_j) - (f-p^*)(x_j) \quad \forall j \\
&\overset{\text{Ann.}}{\Longrightarrow} && \operatorname{sgn}(p^*-\tilde{p})(x_j) = \operatorname{sgn}(f-\tilde{p})(x_j) \neq 0, \qquad j=0,\dots,n \\
&\Longrightarrow && p^*-\tilde{p} \in \Pi_{n-1} \quad \text{hat } n \text{ Nullstellen} \Rightarrow p^* = \tilde{p} \\
& && \text{im Widerspruch zur Annahme.}
\end{aligned}$$

Beweis (10.19) wird wie eben geführt: Ersetze $\tilde{p}$ durch p und p^* durch $\tilde{p}$. Dann folgt aus der Annahme

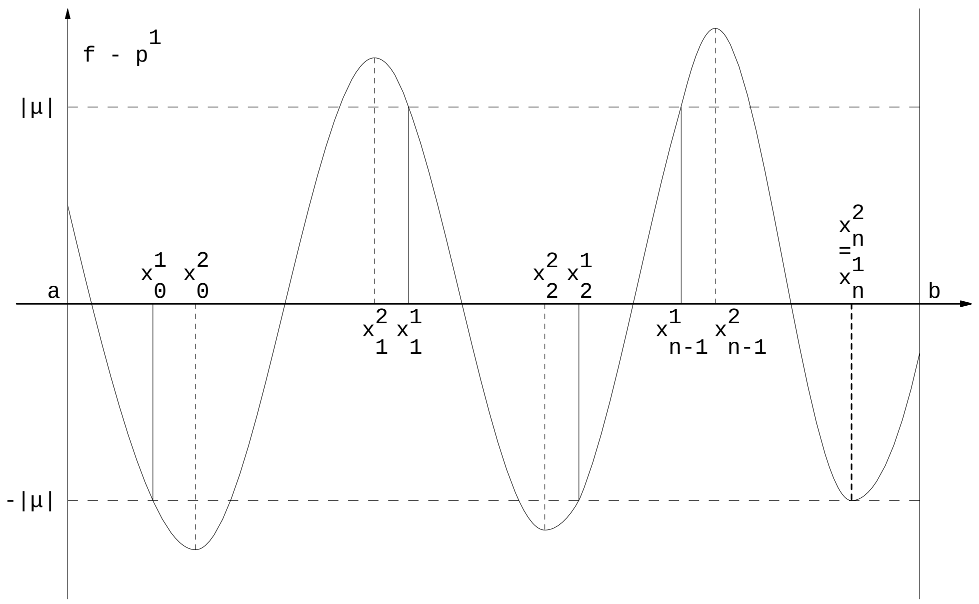
$$\min_{j=0,\dots,n} |(f-p)(x_j)| > |(f-\tilde{p})(x_i)| = |\mu| \quad \forall i$$

wie eben $p = \tilde{p}$. ■

Der Remez–Algorithmus

Der vorige Satz erlaubt die Konstruktion eines Verfahrens zur iterativen Bestimmung der Minimallösung. Er beruht auf der iterativen Bestimmung einer Alternante, welche den Bedingungen des Alternantensatzes genügt. Wir beschreiben einen Iterationsschritt.

Gehe aus von einer „Anfangsalternante“ $a \le x_0^{(1)} < x_1^{(1)} < \dots < x_n^{(1)} \le b$. Die beste untere Schranke μ_1 für die Minimalabweichung bzgl. der Alternante $\{x_j^{(1)}\}$ erhält man durch Berechnung eines Polynoms $p_1 \in \Pi_{n-1}$ gemäß Satz 10.8 b). Dies liefert qualitativ folgende Situation:



Fallen die Punkte $\{x_j^{(1)}\}$ mit den Extremalstellen von $f - p_1$ zusammen, so ist p_1 nach dem Alternantensatz Minimallösung. Andernfalls wählt man als neue Alternante $a \le x_0^{(2)} < x_1^{(2)} < \ldots < x_n^{(2)} \le b$ für f und p_1 die Extremalstellen von $f - p_1$ (man will die Extrema nivellieren) und bestimmt gemäß Satz 10.8 b) das Polynom $p_2 \in \Pi_{n-1}$, das die beste untere Schranke μ_2 für die Minimalabweichung bzgl. der Punktmenge $\{x_j^{(2)}\}$ liefert. Man setzt dieses Verfahren fort, nun mit p_2, $\{x_j^{(2)}\}$ an Stelle von p_1, $\{x_j^{(1)}\}$.

Man kann zeigen, daß die Folge $\{|\mu_j|\}$ der unteren Schranken strikt monoton wachsend gegen die Minimalabweichung $\rho(f)$ konvergiert und die Folge p_j der Polynome gegen das eindeutig bestimmte Minimalpolynom konvergiert. Das Verfahren wird abgebrochen, wenn die Differenz $\|f - p_j\|_\infty - |\mu_j|$ zwischen der oberen und unteren Schranke für die Minimalabweichung (die das Verfahren ja mitliefert) klein genug ausfällt.

Für verfahrenstechnische Einzelheiten verweisen wir auf die entsprechende Literatur über Approximationstheorie (z.B. Powell, M.J.D.: Approximation Theory and Methods, Cambridge University Press 1981).

Beschaffung einer Ausgangsalternante.

Natürlich kann man mit einer beliebig vorgegebenen Punktmenge $a \le x_0^{(1)} < x_1^{(1)} < \ldots < x_n^{(1)} \le b$ beginnen und das zugehörige Polynom p_1 nach Satz 10.8 b) berechnen.

Es ist jedoch zu vermuten, daß eine Alternante, die zu einem Polynom $\overline{p}$ gehört, das schon eine gute Näherung für f ist, „besser“ ausfallen wird. Die Ergebnisse aus § 3 legen es nahe, als Polynom $\overline{p} \in \Pi_{n-1}$ das Interpolationspolynom $\overline{p} \in \Pi_{n-1}$ zu wählen, das als Stützstellen die Nullstellen von T_n besitzt ($\overline{p}$ muß nicht berechnet werden) und als Alternante (vgl. (3.14)) die Extremalstellen von T_n

$$x_j = \frac{1}{2}\left\{(a+b) - (b-a)\cos\frac{j\,\pi}{n}\right\}, \qquad j = 0, 1, \ldots, n\,.$$

(vgl. dazu (3.21) aus Satz 3.8 und die inverse Transformation von (3.16), d.h. die letzte Gleichung auf S. 27)

Es steht noch aus die Vervollständigung einer Äquivalenzaussage, die zum Beweis des Optimalitätskriteriums und damit zum Alternantensatz führte.

Lemma 10.9

Sei $f \in C[a,b]$ und $p^* \in \Pi_{n-1}[a,b]$ eine Näherung. Es existiere ein $p \in \Pi_{n-1}$ mit

$$d^*(x)\,p(x) := [f(x) - p^*(x)]\,p(x) > 0 \quad \forall\, x \in M = \{x \in [a,b]; |d^*(x)| = \|f - p^*\|_\infty\}\,.$$

Dann gilt:

$$\exists\,\delta > 0: \quad \max_{x\in[a,b]} |d^*(x) - \delta\,p(x)| < \max_{x\in[a,b]} |d^*(x)| \tag{10.20}$$

(d.h. $p^* + \delta\,p$ ist eine bessere Näherung für f als p^*).

Beweis:
Ohne Einschränkung sei

$$|p(x)| \;\leq\; 1 \qquad \forall\, x \;\in\; [a,b]\,, \tag{10.21}$$

denn da $p(x)$ im abgeschlossenen Intervall $[a,b]$ durch eine Konstante $K>0$ beschränkt wird, kann man, falls $K\geq 1$, in der Behauptung δ durch $\frac{\delta}{K}$ ersetzen.

Sei nun

$$M' \;:=\; \{x\in[a,b];\; d^*(x)\,p(x)\leq 0\}\,.$$

Laut Voraussetzung gilt $d^*(x)\,p(x)>0\ \forall\, x\in M$.

Also ist $M\cap M'=\emptyset$ und falls $M'\neq\emptyset$

$$\begin{aligned} d &:= \max_{x\in M'}|d^*(x)| \;<\; \max_{x\in M}|d^*(x)|\,,\\ d &:= 0 \quad \text{falls} \quad M' \;=\; \emptyset\,. \end{aligned} \tag{10.22}$$

Das Maximum d wird angenommen, da d^*p stetig, also M' als Urbild einer abgeschlossenen Menge abgeschlossen (und beschränkt) ist.

Wir setzen

$$\delta \;:=\; \frac{1}{2}\left(\max_{x\in[a,b]}|d^*(x)|-d\right)\,, \qquad (>\;0 \text{ wegen } (10.22)) \tag{10.23}$$

betrachten

$$\xi\in[a,b]\;:\quad |d^*(\xi)-\delta\,p(\xi)| \;=\; \max_{x\in[a,b]}|d^*(x)-\delta\,p(x)|\,, \tag{10.24}$$

und beweisen (10.20) für die Fälle $\xi\in M'$ und $\xi\notin M'$.

Sei $\xi\notin M'$:
Dieser Fall beinhaltet auch $M'=\emptyset$.
Dann ist $d^*(\xi)\,p(\xi)\;>\;0$, d.h. daß $d^*(\xi)$ und $p(\xi)$ gleiches Vorzeichen haben und wir erhalten

$$\begin{aligned} \max_{x\in[a,b]}|d^*(x)-\delta\,p(x)| &\overset{(10.24)}{=} |d^*(\xi)\;-\;\delta\,p(\xi)|\\ &< \max\,(\,|d^*(\xi)|,\;|\delta\,p(\xi)|\,)\,, \quad \text{denn } d^*(\xi)\,p(\xi)>0\\ &\leq \max_{x\in[a,b]}|d^*(x)|\,. \end{aligned}$$

Sei $\xi\in M'$:
Dann gilt die Abschätzung

$$
\begin{aligned}
\max_{x\in[a,b]} |d^*(x) - \delta\, p(x)| &\overset{(10.24)}{\leq} |d^*(\xi)| + \delta |p(\xi)| \leq d + \delta \\
&\qquad \text{wegen} \quad \xi \in M',\ \ (10.22),\ (10.21) \\
&= d + \frac{1}{2}\left(\max_{x\in[a,b]} |d^*(x)| - d\right) \\
&= \frac{1}{2}\left(\max_{x\in[a,b]} |d^*(x)| + d\right) \\
&\overset{(10.22)}{<} \frac{1}{2}\left(\max_{x\in[a,b]} |d^*(x)| + \max_{x\in[a,b]} |d^*(x)|\right) \\
&= \max_{x\in[a,b]} |d^*(x)|\,.
\end{aligned}
$$

In beiden Fällen wurde also eine bessere Näherung gefunden. ■

§ 11 Iterative Lösung linearer und nichtlinearer Gleichungen

Der Fixpunktsatz

Iterative Verfahren zur Lösung von Problemen sind immer dann angesagt, wenn eine analytische Lösung nicht möglich ist (vgl. Beispiel 1 aus § 8) oder wenn sie zu lange dauert oder numerisch zu anfällig ist (vgl. GEV). Kepler beschäftigte sich schon im 15. Jahrhundert im Rahmen seiner astronomischen Untersuchungen mit der nach ihm benannten Gleichung

$$x \;=\; a + b \sin x\,, \qquad a, b \in \mathbb{R} \text{ gegeben;} \qquad \textit{Kepler Gleichung} \tag{11.1}$$

Er entwickelte zu ihrer Lösung das Verfahren der *sukzessiven Approximation*, mit dem wir uns in Satz 11.1 beschäftigen werden. Gleichung (11.1) beschreibt ein

Fixpunktproblem: Sei $D \subseteq \mathbb{R}^n$ und $\boldsymbol{g} : D \to \mathbb{R}^n$. Bestimme einen Fixpunkt $\boldsymbol{x}^*$ von $\boldsymbol{g}$:

$$\boldsymbol{x}^* \;=\; \boldsymbol{g}(\boldsymbol{x}^*)\,. \tag{11.2}$$

Dieses Problem ist äquivalent zu einem

Nullstellenproblem:
Sei $D \subseteq \mathbb{R}^n$ und $\boldsymbol{f} : D \to \mathbb{R}^n$. Bestimme eine Nullstelle $\boldsymbol{x}^*$ von $\boldsymbol{f}$:

$$\boldsymbol{f}(\boldsymbol{x}^*) \;=\; \boldsymbol{0}\,. \tag{11.3}$$

Beweis:
Für $\boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})$ geht (11.2) in (11.3) über. Für ein $k \in \mathbb{R}$, $k \neq 0$, oder noch allgemeiner: $\boldsymbol{k} \in \mathbb{R}^{n \times n}$, $\det \boldsymbol{k} \neq 0$ und $\boldsymbol{g}(\boldsymbol{x}) := \boldsymbol{x} - \boldsymbol{k}\,\boldsymbol{f}(\boldsymbol{x})$ geht (11.3) in (11.2) über. (Eine geeignete Wahl von k kann helfen die Eigenschaften eines Lösungsverfahrens zu beeinflussen.)

Kepler hat seine Fixpunktgleichung (11.1) gelöst durch das

Verfahren der sukzessiven Approximation.
Für einen Ausgangspunkt $\boldsymbol{x}^0 \in D$ wird eine Folge $\{\boldsymbol{x}^\nu\} \subset \mathbb{R}^n$, $\nu \in \mathbb{N}$, bestimmt gemäß

$$\boldsymbol{x}^{\nu+1} \;=\; \boldsymbol{g}\left(\boldsymbol{x}^\nu\right)\,.$$

Unter geeigneten Voraussetzungen konvergiert diese Folge gegen einen Fixpunkt. Für $\boldsymbol{g} \in C(\mathbb{R})$ hat man die geometrische Deutung:

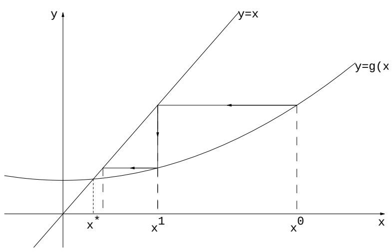


Abb. 1

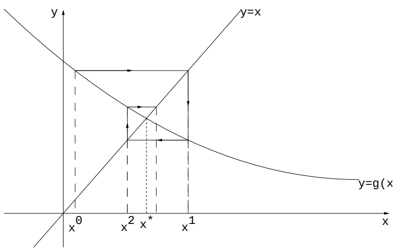


Abb. 2

An Hand der folgenden Zeichnungen verdeutliche man sich, welchen Voraussetzungen ein $g \in C(\mathbb{R})$ genügen muß, damit die Folge $\{x^n\}$ gegen einen Fixpunkt $x^* = g(x^*)$ konvergiert.

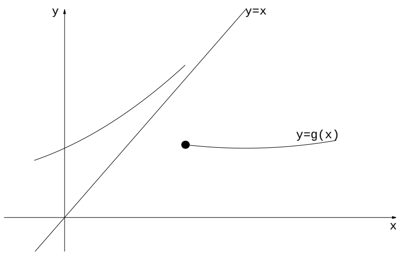


Abb. 3

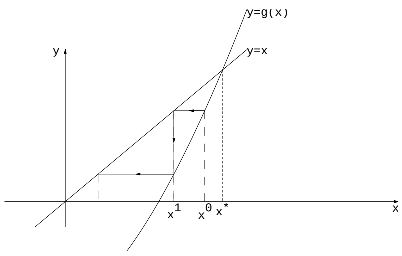


Abb. 4

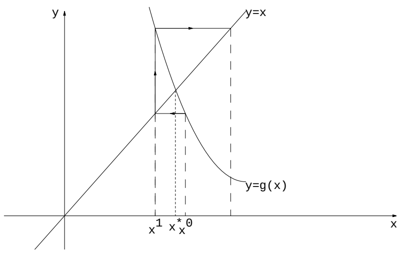


Abb. 5

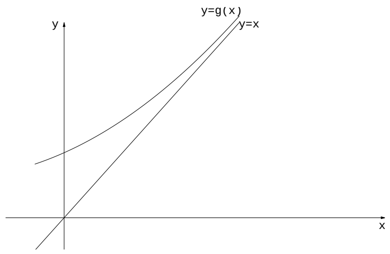


Abb. 6

Folgende Voraussetzungen sind unmittelbar zu ersehen:

> g muß einen Bereich D (welchen?) in sich abbilden: $g(D) \subseteq D$,
> sonst kann die Folge gar nicht berechnet werden.

Weiter erkennt man aus den Abbildungen:

Abb. 3 $\Rightarrow$ g muß stetig sein in D, sonst existiert vielleicht gar kein Fixpunkt.

Abb. 1, 2, 4, 5 $\Rightarrow$ Der Betrag der Steigung von g muß kleiner sein als 1 (= Steigung von $y = x$), sonst divergiert das Verfahren. Ableitungsfrei formuliert bedeutet das: $\left|\frac{g(x)-g(y)}{x-y}\right| < 1$ bzw., damit „Nenner = 0“ nicht stört: $|g(x) - g(y)| < |x - y|$.

Abb. 6 $\Rightarrow$ Die vorige Voraussetzung muß verschärft werden zu $|g(x)-g(y)| \leq L|x - y|$ für ein L: $0 < L < 1$, sonst existiert kein Fixpunkt.

Abb. 1, 2 $\Rightarrow$ Das Verfahren liefert nicht nur monotone Folgen. Da der Fixpunkt *a priori* nicht bekannt ist, kommt als einziges Konvergenzkriterium das Cauchy–Kriterium in Frage (Vollständigkeit des Raumes). Damit eine Cauchyfolge in einer Menge D einen Grenzwert hat, muß diese abgeschlossen sein.

Damit haben wir alle Voraussetzungen zusammen, unter denen Banach für den allgemeineren Fall eines normierten Raumes den folgenden Fixpunktsatz bewiesen hat.

Satz 11.1 Fixpunktsatz von Banach
Sei $(X, \| \ \|)$ ein *Banachraum* (d.h. linear, normiert, vollständig). $D \subset X$ sei eine abgeschlossene Teilmenge und $g : D \to X$ eine Abbildung mit folgenden Eigenschaften:

a) $g(D) \subseteq D$ (*Selbstabbildung*)

b) g ist *Lipschitz–stetig in* D d.h.

$$\begin{aligned} &\exists\, L > 0 \ : \ \|g(x) - g(y)\| \ \leq \ L\|x - y\| \quad \forall\, x, y \in D \quad \text{und} \\ &L < 1 \quad (\textit{Kontraktion})\,. \end{aligned} \tag{11.4}$$

Dann gilt

1) Für jeden Startwert $x^0 \in D$ ist die *sukzessive Iteration*

$$x^{n+1} \ = \ g(x^n), \qquad n \ \in \ \mathbb{N}_0$$

durchführbar.

2) Die Folge $\{x^n\}$ konvergiert gegen ein $x^* \in D$.

3) $g(x^*) = x^*$ (*Fixpunkteigenschaft*).

4) Es gibt nur einen Fixpunkt in D (*Eindeutigkeit*).

5) Es gilt die Fehlerabschätzung

$$\begin{aligned} \|x^* - x^n\| \ &\leq \ \frac{L}{1-L}\,\|x^n - x^{n-1}\| \qquad (\textit{a posteriori–Abschätzung}) \\ &\leq \ \frac{L^n}{1-L}\,\|x^1 - x^0\| \qquad (\textit{a priori–Abschätzung})\,. \end{aligned} \tag{11.5}$$

Beweis:

1) Die Durchführbarkeit des Verfahrens folgt aus der Selbstabbildungseigenschaft.

2) Da g kontrahierend ist, gilt für beliebiges i:

$$\begin{aligned} \|x^{i+1} - x^i\| \ &= \ \|g(x^i) - g(x^{i-1})\| \\ &\leq \ L\|x^i - x^{i-1}\| \\ &\leq \ L^2\|x^{i-1} - x^{i-2}\| \\ &\leq \ \cdots \\ &\leq \ L^i\|x^1 - x^0\|\,. \end{aligned} \tag{11.6}$$

Für beliebiges $j > i$ ist dann

$$\begin{aligned}
\|x^j - x^i\| &= \|x^j - x^{j-1} + x^{j-1} - x^{j-2} + \ldots + x^{i+1} - x^i\| \\
&= \left\| \sum_{k=i}^{j-1} \left(x^{k+1} - x^k\right) \right\| \\
&\le \sum_{k=i}^{j-1} \left\|x^{k+1} - x^k\right\| \\
&\overset{(11.6)}{\le} \sum_{k=i}^{j-1} L^k \|x^1 - x^0\| \qquad (11.7)\\
&= L^i \|x^1 - x^0\| \sum_{k=i}^{j-1} L^{k-i} \\
&\le L^i \|x^1 - x^0\| \sum_{k=0}^{\infty} L^k \qquad \text{(Grenzwert der} \\
&\le \frac{L^i}{1-L} \|x^1 - x^0\| \qquad \text{geometrischen Reihe)}\,.
\end{aligned}$$

Also ist $\{x^i\}$ eine Cauchy–Folge in D. Da X vollständig ist, konvergiert sie gegen ein $x^* \in D$, denn D ist abgeschlossen.

3) Wir schreiben (11.7) in der Form

$$\|g(x^{j-1}) - x^i\| \quad \le \quad \frac{L^i}{1-L} \|x^1 - x^0\|\,.$$

g ist stetig (Voraussetzung b)), die Norm ebenfalls, wir können also den Grenzübergang $i, j \to \infty$ ausführen und erhalten

$$g(x^*) \ - \ x^* \ = \ 0 \qquad \text{(Fixpunkteigenschaft)}\,.$$

4) Angenommen, es gäbe einen zweiten Fixpunkt $\overline{x} \in D$, dann wäre

$$\|\overline{x} - x^*\| \ = \ \|g(\overline{x}) - g(x^*)\| \ \overset{(11.4)}{<} \ \|\overline{x} - x^*\| \qquad \text{Widerspruch!}$$

5) Läßt man in (11.7) $j \to \infty$ gehen, so folgt (Stetigkeit der Norm)

$$\|x^* - x^i\| \quad \le \quad \frac{L^i}{1-L} \|x^1 - x^0\|\,.$$

Betrachtet man die Näherung x^{n-1} als Ausgangsnäherung (setze $x^{n-1} = x^0$), so folgt hieraus

$$\|x^* - x^n\| \ \le \ \frac{L}{1-L} \|x^n - x^{n-1}\| \quad \text{und mit (11.6)} \qquad (11.8)$$

$$\le \ \frac{L^n}{1-L} \|x^1 - x^0\|\,. \qquad (11.9)$$

■

Bemerkung:
Zur Fehlerabschätzung benutzt man lieber (11.8) als (11.9), da man sonst zu viel „verschenkt“.

Frage:
Wie können die Voraussetzungen des Fixpunktsatzes in der Anwendung erfüllt werden?

Die beiden Voraussetzungen a), b) können nicht getrennt voneinander untersucht werden, da die Kontraktionseigenschaft von g natürlich von den Eigenschaften von g in D abhängt. Insbesondere zeigen die Abbildungen 4 und 5: Ist die Kontraktionseigenschaft nicht gegeben, so wird auch eine Umgebung des Fixpunktes x^* nicht in sich selbst abgebildet. Wir prüfen also zuerst die Kontraktionsbedingung. Eine hinreichende Bedingung für differenzierbare, reellwertige Funktionen einer reellen Variablen ist unmittelbar einsichtig.

Sei $g \in C^1(I)$, $I \subset \mathbb{R}$ ein Intervall, für ein $L: \ 0 < L < 1$ gelte

$$|g'(x)| \ \leq \ L \ < \ 1 \qquad \forall\, x \in I\,.$$

Dann ist g auf I kontraktiv.

Der Beweis folgt unmittelbar aus dem Mittelwertsatz für ein $\hat{\xi} \in I$:

$$|g(x) - g(y)| \ = \ |g'(\hat{\xi})|\,|x-y| \ \leq \ \max_{\xi \in I} |g'(\xi)|\,|x-y| \quad \forall\, x, y \in I\,.$$

Eine analoge Kontraktionsbedingung kann man — ebenfalls mit Hilfe des Mittelwertsatzes — für vektorwertige Funktionen aufstellen.

Sei $\boldsymbol{g} = (g_1, \ldots, g_n)^T \in C^1(D)$, $D \subset \mathbb{R}^n$ konvex, kompakt, dann gilt mit der Jacobi–Matrix $\boldsymbol{g}'(\boldsymbol{x}) = \left(\frac{\partial g_i(\boldsymbol{x})}{\partial x_j}\right)_{i,j=1,\ldots,n}$ (für passende Vektor– und Matrixnormen)

$$\begin{aligned} \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\| \ &\leq \ \max_{t \in [0,1]} \|\boldsymbol{g}'(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))\|\,\|\boldsymbol{x} - \boldsymbol{y}\| \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in D \\ &\leq \ \max_{\boldsymbol{\xi} \in D} \|\boldsymbol{g}'(\boldsymbol{\xi})\|\,\|\boldsymbol{x} - \boldsymbol{y}\|\,; \end{aligned}$$

zum Beweis vgl. Forster II, Satz 5 und Corollar.

Weiter ist zu prüfen, wie man ein $D \subset X$ findet, das durch die Funktion g in sich abgebildet wird und in dem die Kontraktionsbedingung erfüllt ist. Unter den Voraussetzungen gilt (vgl. (11.7)):

$$\|x^j - x^1\| \ \leq \ \frac{L}{1-L}\,\|x^1 - x^0\| \qquad \forall\, j \ \geq \ 1\,,$$

d.h. alle Iterierten liegen in der Kugel $K = \left\{x \in X;\ \|x - x^1\| \leq \frac{L}{1-L}\,\|x^1 - x^0\|\right\}$. Deshalb liegt die Vermutung nahe, daß diese Kugel in sich selbst abgebildet wird, wenn die Kontraktionseigenschaft in dieser Kugel erfüllt ist.

In der Tat gilt

Korollar 11.2
Sei $(X, \| \ \|)$ ein Banachraum, $g : D_g \to X$ mit $D_g \subseteq X$. Für $x^0, x^1 = g(x^0) \in D_g$ und ein $L \in (0,1)$ liege die Kugel K in D_g:

$$K \ := \ \left\{ x \in X; \ \|x - x^1\| \ \leq \ \frac{L}{1-L} \, \|x^1 - x^0\| \right\} \subseteq D_g \quad (\textit{Kugelbedingung}) \, .$$

g sei auf K kontraktiv mit der Kontraktionskonstanten L.
Dann gilt

$$g(K) \quad \subseteq \quad K \, ,$$

und die Aussagen des Fixpunktsatzes gelten für $D = K$.

Beweis:
Für $x \in K$ gilt

$$\begin{aligned} \|g(x) - x^1\| \ &\leq \ \|g(x) - g(x^1)\| \ + \ \|g(x^1) - g(x^0)\| \quad (\text{wegen } \ x^1 = g(x^0)) \\ &\leq \ L\|x - x^1\| \ + \ L\,\|x^1 - x^0\| \\ &\leq \ L \, \frac{L}{1-L} \, \|x^1 - x^0\| \ + \ L\|x^1 - x^0\| \\ &= \ \frac{L}{1-L} \, \|x^1 - x^0\| \, . \end{aligned}$$

■

Bemerkungen

1) Ein Vorteil der sukzessiven Approximation ist, daß sie ableitungsfrei arbeitet.
2) Üblicherweise wird es schwierig sein, im voraus die Kugel K (vgl. Korollar 11.2) zu bestimmen. Sinnvoll ist deshalb folgende

   Vorgehensweise: Man besorge sich (wie?) eine Anfangsnäherung x^0, führe das Iterationsverfahren durch und wenn es aussieht, als konvergiere die Folge, benutze man die letzten errechneten Näherungen als x^1 und x^0 und versuche damit, die Voraussetzungen des Korollars 11.2 zu erfüllen.

Wir demonstrieren dies an einem

Beispiel:
Die Eintauchtiefe h eines schwimmenden Baumstammes (Dichte $= \rho$, Radius $= r$) berechnet sich aus der Formel

$$h \quad = \quad r\left(1 \, - \, \cos \frac{\alpha}{2}\right) \, ,$$

wobei α Lösung der folgenden Fixpunktgleichung ist

$$\alpha \;=\; \sin\alpha \,+\, 2\pi\,\rho \;=:\; g(\alpha)\,.$$

Es ist $\rho = 0.66$ für Buchenholz. Wählt man $\alpha_0 = 0$ als Anfangswert für die Iteration

$$\alpha_{k+1} \;=\; \sin\alpha_k \,+\, 2\pi\,\rho\,, \qquad \alpha_0 \;=\; 0\,,$$

so ist $\alpha_1 = 4.1469$. Wir versuchen für α_0 und α_1 die Kugelbedingung zu erfüllen. Es ist $\pi < \alpha_1 < 3\pi/2$. Dort ist $g'(x) = \cos(x)$, also g monoton wachsend. Deshalb ist in jeder Kugel um α_1 die Lipschitzkonstante $L \geq |g'(\alpha_1)| = |\cos\alpha_1| \geq 0.53$, also $\frac{L}{1-L} \geq 1.12$, also $\frac{L}{1-L}\,|\alpha_1 - \alpha_0| \geq 4.6$. In $|\alpha - \alpha_1| \leq 4.6$ kann keine Kontraktion gelten, da π in dieser Kugel liegt und $|g'(\pi)| = 1$. Für α_1, α_0 kann man also die Voraussetzungen von Korollar 11.2 nicht erfüllen.

Wir iterieren trotzdem und erhalten

k	0	1	5	20	59	60
α_k	0.00	4.1469	3.8924	3.6282	3.65553	3.65529

Die Folge scheint, wenn auch sehr langsam, zu konvergieren.

Wir versuchen, die Voraussetzungen von Korollar 11.2 zu erfüllen, indem wir setzen:

$$x_0 \;=\; \alpha_{59} \;=\; 3.65553 \qquad x_1 \;=\; \alpha_{60} \;=\; 3.65529\,.$$

Nun gilt in $I = [3.5,\, 3.8]$: $L = \max_I |g'(x)| = |g'(3.5)| = 0.94$ und damit $\frac{L}{1-L}\,|x_1 - x_0| = 0.00376$ und tatsächlich liegt $K = \{x;\; |x - x_1| \leq 0.00376\}$ in I, d.h. in K ist g kontraktiv.

Die Voraussetzungen des Korollars sind erfüllt, also auch die des Fixpunktsatzes. Es existiert also eine Lösung $\alpha^* \in \{\alpha \in \mathbb{R}\,;\; |\alpha - \alpha_{60}| \leq 0.00376\}$ und die Fehlerabschätzung liefert

$$|\alpha^* - \alpha_{60}| \;\leq\; \frac{L}{1-L}\,|\alpha_{60} - \alpha_{59}| \;=\; 0.00376\,.$$

Gelegentlich kann man die Voraussetzungen des Fixpunktsatzes (ohne Benutzung der Kugelbedingung) auch mit Hilfe von Monotoniebetrachtungen erfüllen, wie folgt:

Im abgeschlossenen Intervall $I = [3.5,\, 3.8]$ ist g kontraktiv (vgl. oben) und $g' < 0$. Also fällt g monoton in I. Aus $g(3.5) = 3.796$, $g(3.8) = 3.53$ folgt, daß $g(I) \subset I$ und die Voraussetzungen des Fixpunktsatzes sind erfüllt

Das Beispiel zeigt als typische Eigenschaft der Fixpunktiteration: Das Verfahren **konvergiert** unter Umständen **sehr langsam**. Man wird also nach schnelleren Verfahren suchen (vgl. Abschnitt nichtlineare Gleichungen).

Als eine wichtige Anwendung des Verfahrens untersuchen wir aber die

Iterative Lösung von Linearen Gleichungssystemen

Wir haben gesehen, daß die direkte Lösung linearer Gleichungssysteme mit dem GEV (vgl. § 5) unter Umständen sehr rundungsfehleranfällig sein kann. Zudem steigt der Rechenaufwand für die Lösung eines $n \times n$–Systems beim Verfahren mit n^3 (Zahl der Multiplikationen). Dies ist besonders gravierend, wenn große lineare Gleichungssysteme zu lösen sind, wie sie bei der diskreten Lösung partieller Differentialgleichungen auftreten. Dies legt den Gedanken nahe, n ach iterativen Verfahren zu suchen, die in jedem Iterationsschritt eine Näherung — also einen ohnehin fehlerbehafteten Wert — verbessern. Es ist zu erwarten, daß solche Verfahren die Rundungsfehler abbauen. Solche Verfahren könnten an Stelle des GEV eingesetzt werden (vor allem wenn sie wenig Aufwand pro Iterationsschritt verlangen), sie könnten aber auch zur Verbesserung einer „ungenauen GEV–Lösung“ benutzt werden.

Es ist in der Tat möglich, solche Verfahren zu konstruieren und ihre Konvergenz, unter gewissen Voraussetzungen, zu beweisen. Dazu wollen wir nun das Gleichungssystem

$$\boldsymbol{A}\boldsymbol{x} \;=\; \boldsymbol{b}, \qquad \boldsymbol{A} \in \mathbb{R}^{n\times n} \quad \text{regulär}, \tag{11.10}$$

in eine Fixpunktaufgabe umformen, und zwar so, daß die Iteration möglichst einfach wird. Dazu zerlegen wir $\boldsymbol{A}$ additiv

$$\boldsymbol{A} \;=\; \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{R}$$

mit

$$\boldsymbol{L} \;=\; \begin{pmatrix} 0 & \dots & & 0 \\ a_{21} & 0 & & \vdots \\ \vdots & \ddots & \ddots & \\ a_{n1} & & a_{nn-1} & 0 \end{pmatrix}, \qquad \boldsymbol{R} \;=\; \begin{pmatrix} 0 & a_{12} & \dots & a_{nn} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1n} \\ 0 & \dots & \dots & 0 \end{pmatrix},$$

$$\boldsymbol{D} \;=\; \operatorname{diag}(\boldsymbol{A}) \;=\; \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{nn} \end{pmatrix}.$$

Wir setzen voraus, daß A regulär ist. Dann können wir voraussetzen, daß $a_{ii} \neq 0,\ i = 1,\dots,n,$ d.h. $\boldsymbol{D}$ ist invertierbar. Ist diese Voraussetzung nicht erfüllt, so kann man sie durch geeignete Zeilen– oder Spaltenvertauschungen immer erreichen, wenn $\boldsymbol{A}$ regulär ist (d.h. $\det \boldsymbol{A} \neq 0$). Dies folgt aus der Determinantendarstellungsformel (Fischer, Satz 4.2.3)

$$\det \boldsymbol{A} \;=\; \sum_{\sigma\in S_n} \operatorname{sign}(\sigma)\, a_{1\sigma(1)} \cdot \ldots \cdot a_{n\sigma(n)},$$

d.h. es gibt mindestens ein Produkt $a_{1\sigma(1)} \cdot \ldots \cdot a_{n\sigma(n)} \neq 0$. Durch Spaltenvertauschungen kann man erreichen, daß $a_{1\sigma(1)},\dots,a_{n\sigma(n)}$ zu den Diagonalelementen werden. (Natürlich muß bei Zeilenvertauschungen die rechte Seite von (11.10) mitvertauscht werden.) Wir schreiben (11.10) in der Form

$$\boldsymbol{A}\boldsymbol{x} \;=\; \boldsymbol{L}\boldsymbol{x} + \boldsymbol{D}\boldsymbol{x} + \boldsymbol{R}\boldsymbol{x} \;=\; \boldsymbol{b},$$

und, da eine Diagonalmatrix einfach zu invertieren ist (für $\boldsymbol{D} = \operatorname{diag}(a_{ii})$ ist $\boldsymbol{D}^{-1} = \operatorname{diag}\left(\frac{1}{a_{ii}}\right)$), wählen wir die folgende Form

$$\boldsymbol{D}\,\boldsymbol{x} \quad = \quad -(\boldsymbol{L}+\boldsymbol{R})\,\boldsymbol{x} \;+\; \boldsymbol{b}\,, \qquad \text{bzw.}$$

$$\boldsymbol{x} \quad = \quad -\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\,\boldsymbol{x} \;+\; \boldsymbol{D}^{-1}\,\boldsymbol{b} \quad =: \quad \boldsymbol{G}_G(\boldsymbol{x})\,. \tag{11.11}$$

Damit erhalten wir die Iterationsvorschrift

$$\begin{array}{c} \boldsymbol{x}^{k+1} \;=\; -\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\,\boldsymbol{x}^k + \boldsymbol{D}^{-1}\,\boldsymbol{b} \;=\; \boldsymbol{G}_G(\boldsymbol{x}^k) \\ \textit{Gesamtschrittverfahren von Jacobi} \end{array} \tag{11.12}$$

$-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})$ heißt *Iterationsmatrix des Gesamtschrittverfahrens.*

Bei der praktischen Durchführung wird man zu Beginn das System $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ so normieren, daß die Diagonalelemente von $\boldsymbol{A}$ zu 1 werden. Dies entspricht der Multiplikation mit $\boldsymbol{D}^{-1}$ in (11.12). Zur Verdeutlichung schreiben wir das Rechenschema nochmals vollständig auf.

Gesamtschrittverfahren (Jacobi–Verfahren)

1. Durch Zeilen– oder Spaltenvertauschungen sichere man, daß $a_{ii} \neq 0, \;\; i = 1, \ldots, n.$

2. Man normiere die Diagonalelemente in $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ zu 1, d.h.

$$\begin{aligned} \widetilde{\boldsymbol{A}} &= (\tilde{a}_{ij})\,, \quad \tilde{a}_{ij} = \frac{a_{ij}}{a_{ii}}\,, \\ \tilde{b}_i &= \frac{b_i}{a_{ii}}\,, \qquad i,j = 1,\ldots,n \end{aligned} \tag{11.13}$$

3. Für ein Anfangselement $\boldsymbol{x}^0 \in \mathbb{R}^n$ iteriere man gemäß

$$\begin{array}{lllll} x_1^{k+1} = & -\,(\tilde{a}_{1,2}\,x_2^k + \tilde{a}_{1,3}\,x_3^k & & + \ldots + \tilde{a}_{1,n}\,x_n^k) & +\,\tilde{b}_1 \\ x_2^{k+1} = -(\tilde{a}_{2,1}\,x_1^k & & + \tilde{a}_{2,3}\,x_3^k & + \ldots + \tilde{a}_{2,n}\,x_n^k) & +\,\tilde{b}_2 \\ x_3^{k+1} = -(\tilde{a}_{3,1}\,x_1^k & + \tilde{a}_{3,2}\,x_2^k & & + \tilde{a}_{3,n}\,x_n^k) & +\,\tilde{b}_3 \\ \vdots \qquad \vdots & & & \vdots & \vdots \\ x_{n-1}^{k+1} = -(\tilde{a}_{n-1,1}\,x_1^k + \ldots & & & + \tilde{a}_{n-1,n}\,x_n^k) & +\,\tilde{b}_{n-1} \\ x_n^{k+1} = -(\tilde{a}_{n,1}\,x_1^k & + \ldots & + \tilde{a}_{n,n-1}\,x_{n-1}^k) & & +\,\tilde{b}_n \end{array} \tag{11.14}$$

Bemerkung: Unter Matlab läuft $\boldsymbol{x}^{k+1} = -(\widetilde{\mathbf{L}} + \widetilde{\mathbf{R}})\boldsymbol{x}^k + \tilde{\boldsymbol{b}}$ sehr schnell.

Hat man aus der 1. Gleichung die Größe x_1^{k+1} berechnet, so wird im Falle der Konvergenz des Verfahrens der Wert x_1^{k+1} brauchbarer sein als der Wert x_1^k. Man könnte also in der 2. Gleichung statt x_1^k gleich den neuen Wert x_1^{k+1} einsetzen und in die 3. Gleichung die aus den ersten beiden Gleichungen ermittelten Werte x_1^{k+1} und x_2^{k+1} usw. Man erhält so aus (11.14) das

Einzelschrittverfahren (Gauß–Seidel–Verfahren)

$$
\begin{array}{llllll}
x_1^{k+1} = & & -(\tilde a_{1,2}\,x_2^k & +\,\tilde a_{1,3}\,x_3^k & +\ldots+\tilde a_{1,n}\,x_n^k) & +\,\tilde b_1 \\
x_2^{k+1} = & -(\tilde a_{2,1}\,x_1^{k+1} & & +\,\tilde a_{2,3}\,x_3^k & +\ldots+\tilde a_{2,n}\,x_n^k) & +\,\tilde b_2 \\
x_3^{k+1} = & -(\tilde a_{3,1}\,x_1^{k+1} & +\,\tilde a_{3,2}\,x_2^{k+1} & & +\,\tilde a_{3,n}\,x_n^k) & +\,\tilde b_3 \\
\vdots & \vdots & & & \vdots & \vdots \\
x_{n-1}^{k+1} = & -(\tilde a_{n-1,1}\,x_1^{k+1} & +\ldots & & +\,\tilde a_{n-1,n}\,x_n^k) & +\,\tilde b_{n-1} \\
x_n^{k+1} = & -(\tilde a_{n,1}\,x_1^{k+1} & +\ldots & +\,\tilde a_{n,n-1}\,x_{n-1}^{k+1}) & & +\,\tilde b_n
\end{array}
\tag{11.15}
$$

Die Begründung für die Bezeichnungen *Gesamtschritt–* und *Einzelschrittverfahren* ergeben sich direkt aus den Iterationsvorschriften (11.14) und (11.15).

Für theoretische Zwecke kann man (11.15) in Matrixform schreiben. Mit

$$
\begin{aligned}
\widetilde{\boldsymbol{A}} &= \left(\frac{a_{ij}}{a_{ii}}\right), \quad \widetilde{\boldsymbol{L}} = \boldsymbol{D}^{-1}\boldsymbol{L}, \quad \widetilde{\boldsymbol{R}} = \boldsymbol{D}^{-1}\boldsymbol{R}, \\
\tilde{\boldsymbol{b}} &= \boldsymbol{D}^{-1}\boldsymbol{b}, \quad \boldsymbol{I} = \text{Identität},
\end{aligned}
\tag{11.16}
$$

erhalten wir aus (11.15)

$$
\begin{aligned}
(\boldsymbol{I}+\widetilde{\boldsymbol{L}})\,\boldsymbol{x}^{k+1} &= -\widetilde{\boldsymbol{R}}\,\boldsymbol{x}^k + \tilde{\boldsymbol{b}} \\
\boldsymbol{x}^{k+1} &= -(\boldsymbol{I}+\widetilde{\boldsymbol{L}})^{-1}\,\widetilde{\boldsymbol{R}}\,\boldsymbol{x}^k + (\boldsymbol{I}+\widetilde{\boldsymbol{L}})^{-1}\,\tilde{\boldsymbol{b}} \\
&= -(\boldsymbol{D}+\boldsymbol{L})^{-1}\,\boldsymbol{R}\,\boldsymbol{x}^k + (\boldsymbol{D}+\boldsymbol{L})^{-1}\,\boldsymbol{b} =: \boldsymbol{G}_E(\boldsymbol{x}^k)\,.
\end{aligned}
\tag{11.17}
$$

Iteriert wird jedoch nach (11.15) ! Man vermeidet dadurch die Berechnung von Inversen.

$-(\boldsymbol{D}+\boldsymbol{L})^{-1}\,\boldsymbol{R}$ heißt *Iterationsmatrix des Einzelschrittverfahrens*.

Bemerkung: Zu weiteren Iterationsverfahren vgl. die Stichworte Richardson-Iteration, Iteration mit Präkonditionierern, insbesondere SOR-Verfahren, und mit einem anderen Zugang das Verfahren der konjugierten Gradienten.

Um Konvergenzaussagen für die obigen Verfahren zu erhalten, prüfen wir die Voraussetzungen des Fixpunktsatzes 11.1.

$\boldsymbol{G}_G$ und $\boldsymbol{G}_E$ (vgl. (11.12) und (11.17)) sind Abbildungen des $\mathbb{R}^n$ in sich. Für die Kontraktionseigenschaft (mit einander zugeordneten oder zumindest zueinander passenden Matrix– und Vektornormen) prüfen wir

$$
\begin{aligned}
\|\boldsymbol{G}_G(\boldsymbol{x})-\boldsymbol{G}_G(\boldsymbol{y})\| &= \|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\,(\boldsymbol{x}-\boldsymbol{y})\| \\
&\le \|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\|\,\|\boldsymbol{x}-\boldsymbol{y}\| \quad \forall\,\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n\,,
\end{aligned}
\tag{11.18}
$$

und entsprechend

$$
\begin{aligned}
\|\boldsymbol{G}_E(\boldsymbol{x})-\boldsymbol{G}_E(\boldsymbol{y})\| &\le \|(\boldsymbol{D}+\boldsymbol{L})^{-1}\,\boldsymbol{R}\|\,\|\boldsymbol{x}-\boldsymbol{y}\| \\
&= \|(\boldsymbol{I}+\widetilde{\boldsymbol{L}})^{-1}\,\widetilde{\boldsymbol{R}}\|\,\|\boldsymbol{x}-\boldsymbol{y}\| \quad \forall\,\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n\,.
\end{aligned}
\tag{11.19}
$$

Man beachte, daß die Lipschitzkonstanten für die Matrixfunktionen $\boldsymbol{G}_G$ und $\boldsymbol{G}_E$ *unabhängig von* $\boldsymbol{x},\boldsymbol{y}$ sind, im Falle der Kontraktion also für den ganzen $\mathbb{R}^n$ gelten. Man kann in Satz 11.1 also $X = D = \mathbb{R}^n$ wählen.

Kontraktion, und damit auch Konvergenz für beliebige Anfangsvektoren $\boldsymbol{x}^0 \in \mathbb{R}^n$, sind gesichert, wenn man eine Matrixnorm (die zu einer Vektornorm paßt) findet, derart daß für die Iterationsmatrizen gilt

$$\|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\| \;<\; 1 \quad \text{(Gesamtschrittverfahren)}\,, \tag{11.20}$$

$$\|(\boldsymbol{D}+\boldsymbol{L})^{-1}\,\boldsymbol{R}\| \;=\; \|(\boldsymbol{I}+\widetilde{\boldsymbol{L}})^{-1}\,\widetilde{\boldsymbol{R}}\| \;<\; 1 \quad \text{(Einzelschrittverfahren)}\,. \tag{11.21}$$

Dies ist besonders einfach nachzuprüfen für das Gesamtschrittverfahren. Man braucht nur die Matrixnormen aus den Normpaaren der Sätze 8.17 a), c) und 8.18 a) in (11.10) eintragen und erhält dann zusammenfassend

Satz 11.3

Die Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ erfülle (ggf. nach Zeilen– oder Spaltenvertauschungen) eines der Kriterien

Zeilensummenkriterium

$$\|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\|_\infty \;=\; \max_{i=1,\ldots,n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \;<\; 1\,, \tag{11.22}$$

Spaltensummenkriterium

$$\|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\|_1 \;=\; \max_{j=1,\ldots,n} \sum_{\substack{i=1\\ i\neq j}}^{n} \frac{|a_{ij}|}{|a_{jj}|} \;<\; 1\,, \tag{11.23}$$

Quadratsummenkriterium

$$\|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\|_F \;=\; \sqrt{\sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left(\frac{|a_{ij}|}{|a_{ii}|}\right)^2} \;<\; 1\,. \tag{11.24}$$

Dann hat die Gleichung $\boldsymbol{A}\,\boldsymbol{x}=\boldsymbol{b}$ eine eindeutige Lösung $\boldsymbol{x}^*$, gegen welche das **Gesamtschrittverfahren** bei *beliebiger Anfangsnäherung* $\boldsymbol{x}^0 \in \mathbb{R}^n$ konvergiert (**globale Konvergenz**). Es gilt die Fehlerabschätzung (11.5) für die Vektornorm, zu der eine passende Matrixnorm eines der obigen Kriterien erfüllt.

Für das Einzelschrittverfahren zeigen wir nur

Satz 11.4

Erfüllt die Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ (ggf. nach Zeilen– oder Spaltenvertauschungen) das *Zeilensummenkriterium* (11.22), so konvergiert das **Einzelschrittverfahren** bei beliebiger Anfangsnäherung gegen die eindeutige Lösung $\boldsymbol{x}^*$ von $\boldsymbol{A}\,\boldsymbol{x}=\boldsymbol{b}$, und für die Maximumsnorm gilt die Fehlerabschätzung (11.5).

Beweis:

Wir zeigen (vgl. (11.22), (11.19), (11.16)):

$$\|\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\|_\infty \;=\; \|\widetilde{\boldsymbol{L}}+\widetilde{\boldsymbol{R}}\|_\infty \;<\; 1 \quad\Rightarrow\quad \|(\boldsymbol{I}+\widetilde{\boldsymbol{L}})^{-1}\,\widetilde{\boldsymbol{R}}\|_\infty \;<\; 1\,.$$

Wir bezeichnen die Iterationsmatrix mit

$$\begin{aligned}
& \boldsymbol{M} &:=& \ (\boldsymbol{I}+\widetilde{\boldsymbol{L}})^{-1}\,\widetilde{\boldsymbol{R}} \\
\Rightarrow\ & \widetilde{\boldsymbol{R}} &=& \ (\boldsymbol{I}+\widetilde{\boldsymbol{L}})\,\boldsymbol{M} \ = \ \boldsymbol{M}+\widetilde{\boldsymbol{L}}\,\boldsymbol{M} \\
\Rightarrow\ & \boldsymbol{M} &=& \ \widetilde{\boldsymbol{R}} - \widetilde{\boldsymbol{L}}\,\boldsymbol{M}\,.
\end{aligned}$$

Für eine Matrix $\boldsymbol{F}=(f_{ik})$ bezeichne $|\boldsymbol{F}|$ die Matrix mit den Elementen $|f_{ik}|$. Dann gilt

$$\begin{aligned}
\|\boldsymbol{M}\|_\infty &= \|\widetilde{\boldsymbol{R}}-\widetilde{\boldsymbol{L}}\,\boldsymbol{M}\|_\infty \ \le\ \|\,|\widetilde{\boldsymbol{R}}|+|\widetilde{\boldsymbol{L}}\,\boldsymbol{M}|\,\|_\infty \\
&= \max_i \sum_{j=1}^n \left(|\tilde r_{ij}| + \left|\sum_{k=1}^n \tilde\ell_{ik}\, m_{kj}\right| \right) \\
&\le \max_i \left\{ \sum_{j=1}^n |\tilde r_{ij}| + \sum_{k=1}^n |\tilde\ell_{ik}| \underbrace{\sum_{j=1}^n |m_{kj}|}_{\le \|\boldsymbol{M}\|_\infty} \right\} \\
&\le \Big\|\, |\widetilde{\boldsymbol{R}}| + |\widetilde{\boldsymbol{L}}|\ \|\boldsymbol{M}\|_\infty \Big\|_\infty \\
&\le \max\,(1,\|\boldsymbol{M}\|_\infty)\cdot \Big\|\, |\widetilde{\boldsymbol{R}}| + |\widetilde{\boldsymbol{L}}| \Big\|_\infty\,.
\end{aligned}$$

Auf Grund der Dreiecksgestalt von $\widetilde{\boldsymbol{L}}$ und $\widetilde{\boldsymbol{R}}$ gilt für jedes Indexpaar i,j stets: $\tilde\ell_{ij}=0$ oder $\tilde r_{ij}=0$ und falls $i=j$ sogar $\tilde\ell_{ii}=\tilde r_{ii}=0$. Deshalb folgt $|\widetilde{\boldsymbol{R}}|+|\widetilde{\boldsymbol{L}}| = |\widetilde{\boldsymbol{L}}+\widetilde{\boldsymbol{R}}|$ und damit

$$\|\boldsymbol{M}\|_\infty \ \le\ \max\,(1,\|\boldsymbol{M}\|_\infty)\cdot \underbrace{\|\widetilde{\boldsymbol{R}}+\widetilde{\boldsymbol{L}}\|_\infty}_{<1 \text{ nach Voraussetzung}}\,,$$

also

$$\|\boldsymbol{M}\|_\infty \ <\ \max\,(1,\|\boldsymbol{M}\|_\infty) \quad \text{und damit} \quad \|\boldsymbol{M}\|_\infty \ <\ 1\,.$$

■

Hinweise:

1. Die Konvergenzkriterien der beiden vorhergehenden Sätze sind *hinreichend*, nicht notwendig. Für das Einzelschrittverfahren gibt es schärfere, auch hinreichende, Konvergenzkriterien (Sassenfeld–Kriterium). Es gibt auch noch andere Iterationsverfahren zur Lösung linearer Gleichungssysteme (SOR–Verfahren, Verfahren der konjugierten Gradienten u.a.).

2. Es kann vorkommen, daß das Einzelschrittverfahren konvergiert und das Gesamtschrittverfahren nicht und umgekehrt.

3. Im allgemeinen jedoch konvergiert das Einzelschrittverfahren „öfter“ und „schneller“ als das Gesamtschrittverfahren (vgl. z.B. Schaback/Werner: Numerische Mathematik, Springer–Lehrbuch).

Wir belegen letzteres durch ein

Beispiel:
Lösung von $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 0.7 & -0.2 & -0.1 \\ -0.1 & 0.6 & -0.2 \\ -0.1 & -0.1 & 0.9 \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} 20 \\ 40 \\ 0 \end{pmatrix}$$

mit dem Gesamtschrittverfahren (GSV) und dem Einzelschrittverfahren (ESV). Die Näherungswerte $\boldsymbol{x}^k$ sowie die Fehler $\varepsilon^k = \|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_\infty$ entnehmen wir der folgenden Tabelle.

	$k=$	0	1	2	4	8	Lösung
GSV	x_1^k	0	28.571	47.619	52.784	53.712	53.731
	x_2^k	0	66.667	71.429	79.491	80.578	80.597
	x_3^k	0	0	10.582	14.291	14.914	14.925
	ε^k		67	19	2.3	0.03	—
ESV	x_1^k	0	28.571	50.567	53.637	53.732	53.731
	x_2^k	0	71.429	78.798	80.549	80.597	80.597
	x_3^k	0	11.111	14.374	14.910	14.925	14.925
	ε^k		71	22	0.5	0.0004	—

Bisher haben wir „Konvergenzgeschwindigkeit“ mehr intuitiv verstanden. Wir wollen sie nun durch eine Definition auf eine feste Grundlage stellen.

Definition 11.5
Sei $(X, \|\ \|)$ ein normierter Raum und $\{x^k\} \subset X$ eine gegen $x^* \in X$ konvergente Folge.
Die Folge $\{x^k\}$ hat mindestens die *Konvergenzordnung* p, wenn es eine Konstante $C > 0$ und ein $k_0 \in \mathbb{N}$ gibt, so daß

$$\|x^{k+1} - x^*\| \;\leq\; C\|x^k - x^*\|^p \quad \forall\, k \geq k_0\,,$$

wobei $C < 1$, falls $p = 1$.
Ein Verfahren hat die Konvergenzordnung p (in einem Gebiet D), falls alle von ihm gelieferten Folgen (mit Anfangselement $x^0 \in D$) die Konvergenzordnung p besitzen.

Bemerkung:

C und p bestimmen die Konvergenzgeschwindigkeit der Folge: Je kleiner C und je größer p desto schneller konvergiert $\{x^k\}$. Dabei bestimmt die Ordnung p das Verhalten in wesentlich stärkerem Maße als die Konstante C. Letztere ist wichtig beim Vergleich linear konvergenter Verfahren.

Man spricht von	linearer Konvergenz, wenn	$p = 1$,
	superlinearer Konvergenz, wenn	$p > 1$,
	quadratischer Konvergenz, wenn	$p = 2$.

Man erkennt aus der Kontraktionsbedingung sofort, daß das Fixpunktverfahren zumindest linear konvergiert, denn

$$\|x^{k+1} - x^*\| \;=\; \|g(x^k) - g(x^*)\| \;\leq\; L\|x^k - x^*\| \quad \text{mit} \quad 0 \;<\; L \;<\; 1\,.$$

Die Aussage: „Das Einzelschrittverfahren konvergiert schneller als das Gesamtschrittverfahren“ kann man nun dadurch beschreiben, daß die Konstante C aus Definition 11.5 für das Einzelschrittverfahren kleiner ausfällt als für das Gesamtschrittverfahren.

Unbefriedigend an den Konvergenzsätzen 11.3 und 11.4 ist, daß die Konvergenzbedingung von der geschickten Wahl einer Norm abhängt (und davon gibt es reichlich viele, vgl. Satz 8.19). Wir beheben diese Manko durch

Satz 11.6

a) Für jede einer Vektornorm (über $\mathbb{C}^n$) zugeordneten Matrixnorm gilt

$$\|\boldsymbol{A}\| \geq r(\boldsymbol{A}) \;\forall\, \boldsymbol{A} \in \mathbb{R}^{n\times n} \quad r(\boldsymbol{A}) = \max_i \{|\lambda_i| \;:\; \lambda_i = \text{ Eigenwert von } \boldsymbol{A}\}\,.$$

b) $\forall\, \varepsilon > 0 \;\wedge\; \forall\, \boldsymbol{A} \in \mathbb{R}^{n\times n}\; \exists$ eine Vektornorm $\|\cdot\|_V$ und eine zugeordnete Matrixnorm $\|\cdot\|_M$ mit $\|\boldsymbol{A}\|_M \leq r(\boldsymbol{A}) + \varepsilon$.

Beweis a)

Sei $\boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x},\; \boldsymbol{x} \neq \boldsymbol{0} \;\Rightarrow\; \|\boldsymbol{A}\,\boldsymbol{x}\|_V = |\lambda|\,\|\boldsymbol{x}\|_V$

$$\Rightarrow \quad |\lambda| \;=\; \frac{\|\boldsymbol{A}\,\boldsymbol{x}\|_V}{\|\boldsymbol{x}\|_V} \;=\; \left\|\boldsymbol{A}\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_V}\right)\right\|_V \;\leq\; \sup_{\|\boldsymbol{y}\|_V=1} \|\boldsymbol{A}\,\boldsymbol{y}\|_V \;=\; \|\boldsymbol{A}\|_M\,.$$

Beweis b)

In Satz 8.19 wurde insbesondere gezeigt: Sei $\|\;\|_\infty$ die Maximumnorm, $\|\;\|_\infty$ die zugeordnete Matrixnorm, $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ eine nichtsinguläre Matrix, so ist

$$\|\boldsymbol{x}\|_T \;=\; \|\boldsymbol{H}\,\boldsymbol{x}\|_\infty \quad \text{eine Vektornorm und}$$
$$\|\boldsymbol{A}\|_T \;=\; \|\boldsymbol{H}\,\boldsymbol{A}\,\boldsymbol{H}^{-1}\|_\infty \quad \text{die zugeordnete Matrixnorm}\,.$$

Beweisidee b): Wir benutzen diese Aussage, indem wir $\boldsymbol{H} = \boldsymbol{D} \cdot \boldsymbol{T}$ für geeignete nichtsinguläre Matrizen $\boldsymbol{D}\,\boldsymbol{T}$ wählen.

$\boldsymbol{T}$ beschreibe die Ähnlichkeitstransformation, welche $\boldsymbol{A}$ auf Jordan–Normalform bringt:

$$\boldsymbol{J}_A \;:=\; \boldsymbol{T}\,\boldsymbol{A}\,\boldsymbol{T}^{-1} \;=\; \begin{pmatrix} \lambda_1 & t_{1,2} & & 0 \\ & \ddots & \ddots & \\ & & \ddots & t_{n-1,n} \\ 0 & & & \lambda_n \end{pmatrix}, \qquad \begin{array}{l} \lambda_i \text{ die Eigenwerte von } A\,, \\ t_{i,i+1} \;=\; 0 \text{ oder } 1. \end{array}$$

Wir transformieren (ähnlich) $\boldsymbol{J}_A$ mit einer Diagonalmatrix

$$\boldsymbol{D} \;=\; \begin{pmatrix} \varepsilon_1 & & 0 \\ & \ddots & \\ 0 & & \varepsilon_n \end{pmatrix}, \quad \varepsilon_i \;>\; 0 \quad (\text{geeignet})\,.$$

$\Rightarrow$

$\boldsymbol{D}\,\boldsymbol{J}_A$:	i–te Zeile von $\boldsymbol{J}_A$ wird mit ε_i multipliziert .
$(\boldsymbol{D}\,\boldsymbol{J}_A)\,\boldsymbol{D}^{-1}$:	k–te Spalte von $(\boldsymbol{D}\,\boldsymbol{J}_A)$ wird mit $\frac{1}{\varepsilon_k}$ multipliziert .

Mit der Bezeichung

$$(d_{ij}) \;=\; (\boldsymbol{D}\,\boldsymbol{J}_A\,\boldsymbol{D}^{-1}) \quad \text{gilt} \quad \begin{cases} d_{ii} \;=\; \lambda_i\,, \\ d_{i,i+1} \;=\; t_{i,i+1}\,\frac{\varepsilon_i}{\varepsilon_{i+1}}\,, \\ d_{ij} \;=\; 0 \quad \text{sonst}\,. \end{cases}$$

Wähle $\varepsilon_i,\;\; i = 1, \ldots, n,$ so, daß $\frac{\varepsilon_i}{\varepsilon_{i+1}} = \varepsilon \;\forall\, i = 1, \ldots, n-1,$ dann folgt

$$\|\boldsymbol{D}\,\boldsymbol{T}\,\boldsymbol{A}\,\boldsymbol{T}^{-1}\,\boldsymbol{D}^{-1}\|_\infty = \|\boldsymbol{D}\,\boldsymbol{J}_A\,\boldsymbol{D}^{-1}\|_\infty = \max_i\,(|\lambda_i + t_{i,i+1}\,\varepsilon|) \;\leq\; \max_i\,(|\lambda_i| + \varepsilon) = r(\boldsymbol{A}) + \varepsilon\,.$$

■

Hinweise und Bemerkungen

1) Die Kontraktionseigenschaft ist *normabhängig.*

   Da im $\mathbb{R}^n$ (und somit auch im $\mathbb{R}^{n\times n}$) alle Normen äquivalent sind (Satz 8.5), sind die Lipschitzstetigkeit und die Konvergenz im $\mathbb{R}^n$ *unabhängig* von der Norm.

2) Ist also der betragsmaximale Eigenwert der Iterationsmatrix $\boldsymbol{M}$ des Einzelschritt– bzw. Gesamtschrittverfahrens betragsmäßig < 1, so gibt es Vektornormen und zugeordnete Matrixnormen, so daß die Verfahren konvergieren. Da diese Normen im allgemeinen aber unbekannt sind (die Transformation auf Jordan–Normalform ist nicht konstruktiv), muß man auf die Fehlerabschätzung verzichten.

3) Im Paragraphen über Eigenwertaufgaben werden wir das (numerisch relativ einfache) von Mises–Verfahren zur Bestimmung des betragsgrößten Eigenwerts einer Matrix kennenlernen, dessen Voraussetzungen in vielen Fällen erfüllt sind.

4) „$r(\boldsymbol{M}) < 1$“ für die Iterationsmatrix $\boldsymbol{M}$ eines Iterationsverfahrens

$$\boldsymbol{x}^{k+1} \;=\; \boldsymbol{M}\,\boldsymbol{x}^k + \tilde{\boldsymbol{b}}$$

zur Lösung von $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ ist auch eine *notwendige* Konvergenzbedingung.

Beweis:
Jede Lösung $\boldsymbol{x}^*$ von $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ genügt laut Konstruktion der Verfahren der Gleichung $\boldsymbol{x}^* = \boldsymbol{M}\,\boldsymbol{x}^* + \tilde{\boldsymbol{b}}$. Also gilt für den Fehlervektor des Iterationsverfahrens

$$\boldsymbol{x}^{k+1} - \boldsymbol{x}^* \;=\; \boldsymbol{M}(\boldsymbol{x}^k - \boldsymbol{x}^*) \;=\; \boldsymbol{M}^2(\boldsymbol{x}^{k-1} - \boldsymbol{x}^*) \;=\; \ldots \;=\; \boldsymbol{M}^{k+1}(\boldsymbol{x}^0 - \boldsymbol{x}^*)\,.$$

Im Falle der Konvergenz muß $\boldsymbol{x}^{k+1} - \boldsymbol{x}^* \xrightarrow{k} \boldsymbol{0}$ gelten. Wählt man $\boldsymbol{x}^0$ so, daß $\boldsymbol{x}^0 - \boldsymbol{x}^*$ Eigenvektor zu einem Eigenwert λ von $\boldsymbol{M}$ ist, so folgt

$$\begin{aligned} \boldsymbol{x}^{k+1} - \boldsymbol{x}^* &= \lambda^{k+1}\,(\boldsymbol{x}^0 - \boldsymbol{x}^*)\,, \\ \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\| &= |\lambda|^{k+1}\,\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|\,. \end{aligned}$$

Der Fehlervektor konvergiert also nur dann immer gegen Null, wenn $|\lambda| < 1$ ist für alle Eigenwerte λ von A. ■

Iterative Lösung nichtlinearer Gleichungen und Gleichungs–systeme

Natürlich kann man durch direkte Anwendung des Verfahrens der sukzessiven Iteration auch nichtlineare Gleichungen und Gleichungssysteme behandeln (sofern die Kontraktionsbedingung nachgewiesen werden kann, die in vielen Fällen nicht erfüllt ist). Auf Grund der sehr langsamen Konvergenzgeschwindigkeit ist dies im allgemeinen nicht empfehlenswert. Als Standardverfahren empfiehlt sich hier

Das Newton–Verfahren (NV)

Das NV zur Bestimmung einer Nullstelle x^* einer Funktion $f \in C^1(\mathbb{R})$ ist samt seiner geometrischen Veranschaulichung schon aus § 2 bekannt. Wir wollen es nun so verallgemeinern, daß es zur Lösung nichtlinearer Gleichungssysteme

$$\boldsymbol{f}(\boldsymbol{x}) \;=\; \boldsymbol{\theta}\,, \quad \boldsymbol{f}(\boldsymbol{x}) \;=\; (f_1(\boldsymbol{x}),\ldots,f_n(\boldsymbol{x}))^T\,, \quad \boldsymbol{x} \in \mathbb{R}^n \tag{11.25}$$

verwendet werden kann. Die graphische Herleitung des NV ist für den Fall $n > 1$ nicht mehr möglich, wohl aber die algebraische Herleitung, die für die Fälle $n = 1$ und $n > 1$ gleichermaßen gilt. Die Kernidee lautet: In jedem Iterationsschritt wird $\boldsymbol{f}(\boldsymbol{x})$ durch eine lineare Approximation $\tilde{\boldsymbol{f}}(\boldsymbol{x})$ ersetzt. Die Nullstelle von $\tilde{\boldsymbol{f}}(\boldsymbol{x})$ wird bestimmt als Näherung für die Nullstelle $\boldsymbol{x}^*$ von $\boldsymbol{f}(\boldsymbol{x})$. (**Linearisierung**!)

Wir setzen $\boldsymbol{f} \in C^1$ voraus. Dann lautet die Taylorentwicklung von $\boldsymbol{f}$ in einem Startwert $\boldsymbol{x}^0$ (bzw. die Def. der Ableitung $\boldsymbol{f}'(\boldsymbol{x}^0)$)

$$\boldsymbol{f}(\boldsymbol{x}) \;=\; \underbrace{\boldsymbol{f}(\boldsymbol{x}^0) + \boldsymbol{f}'(\boldsymbol{x}^0)\,(\boldsymbol{x} - \boldsymbol{x}^0)}_{\tilde{\boldsymbol{f}}(\boldsymbol{x})} + \sigma\,(\boldsymbol{x} - \boldsymbol{x}^0)\,, \qquad \lim_{\boldsymbol{x}\to\boldsymbol{x}^0} \frac{\sigma\,(\boldsymbol{x} - \boldsymbol{x}^0)}{\|\boldsymbol{x} - \boldsymbol{x}^0\|} \;=\; \boldsymbol{0}\,.$$

Hierbei ist

$$\boldsymbol{f}'(\boldsymbol{x}^0) \quad = \quad \begin{pmatrix} \dfrac{\partial f_1(\boldsymbol{x}^0)}{\partial x_1} & , \dots , & \dfrac{\partial f_1(\boldsymbol{x}^0)}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_n(\boldsymbol{x}^0)}{\partial x_1} & , \dots , & \dfrac{\partial f_n(\boldsymbol{x}^0)}{\partial x_n} \end{pmatrix}$$

die *Jacobi–Matrix* (Ableitung von $\boldsymbol{f}'$) an der Stelle $\boldsymbol{x}^0$ (vgl. Forster II, § 6, Satz 1).

Die Nullstelle $\boldsymbol{x}^1$ von $\tilde{\boldsymbol{f}}(\boldsymbol{x}) = 0$ betrachten wir als Näherung für die Nullstelle $\boldsymbol{x}^*$ von $\boldsymbol{f}(\boldsymbol{x})$. Sofern die Inverse $\boldsymbol{f}'(\boldsymbol{x}^0)^{-1}$ existiert, ist $\tilde{\boldsymbol{f}}(\boldsymbol{x}^1) = 0$ äquivalent mit

$$\boldsymbol{x}^1 \;=\; \boldsymbol{x}^0 \;-\; \boldsymbol{f}'(\boldsymbol{x}^0)^{-1}\, \boldsymbol{f}(\boldsymbol{x}^0)\,.$$

Dies inspiriert das Newton–Verfahren

$$\boldsymbol{x}^{k+1} \;=\; \boldsymbol{x}^k \;-\; \boldsymbol{f}'(\boldsymbol{x}^k)^{-1}\, \boldsymbol{f}(\boldsymbol{x}^k), \qquad k = 0, 1, \ldots \tag{11.26}$$

Zur Durchführung des Verfahrens berechnet man **nicht** etwa die Inverse $\boldsymbol{f}'(\boldsymbol{x}^k)^{-1}$ in jedem Iterationsschritt (das ist zu aufwendig), sondern man berechnet die Newton–Korrektur $\Delta\,\boldsymbol{x} = \boldsymbol{x}^{k+1} - \boldsymbol{x}^k$ als Lösung eines linearen Gleichungssystems. Iteriert wird gemäß

$$\boldsymbol{f}'(\boldsymbol{x}^k)\,\Delta\,\boldsymbol{x} \;=\; -\boldsymbol{f}(\boldsymbol{x}^k)\,, \qquad \boldsymbol{x}^{k+1} \;=\; \boldsymbol{x}^k \;+\; \Delta\,\boldsymbol{x}\,. \tag{11.27}$$

Man kann das NV als Fixpunktverfahren mit der Iterationsfunktion

$$\boldsymbol{g}(x) \;=\; \boldsymbol{x} - \boldsymbol{f}'(\boldsymbol{x})^{-1}\, \boldsymbol{f}(\boldsymbol{x})$$

(vgl. (11.26)) auffassen und durch Anwendung des Fixpunktsatzes 11.1 einen Existenz–, Konvergenz– und Eindeutigkeitssatz beweisen. Wir begnügen uns hier mit einer lokalen Konvergenz– und Eindeutigkeitsaussage, welche die qualitativen Eigenschaften des NV deutlich macht und verzichten auf den Existenzbeweis. Kern der folgenden Aussage ist: Hat die differenzierbare Funktion $\boldsymbol{f}(\boldsymbol{x})$ eine einfache Nullstelle $\boldsymbol{x}^*$ (d.h. $\det \boldsymbol{f}'(\boldsymbol{x}^*) \neq 0$), so konvergiert das NV lokal, quadratisch (d.h. es gibt eine Umgebung U von $\boldsymbol{x}^*$, so daß für jedes Startelement $\boldsymbol{x}^0 \in U$ das NV eine quadratisch gegen $\boldsymbol{x}^*$ konvergente Folge liefert).

Im Anschluß an den Fixpunktsatz haben wir gesehen, daß $\max\limits_{\xi \in D} \|\boldsymbol{g}'(\boldsymbol{\xi})\| < 1$ eine hinreichende Bedingung für die Kontraktionseigenschaft der Iterationsfunktion g war.

Bei der iterativen Lösung linearer Gleichungssysteme hatte die Iterationsfunktion $\boldsymbol{G}(\boldsymbol{x})$ die Gestalt

$$\boldsymbol{G}(\boldsymbol{x}) \;=\; \boldsymbol{M}\,\boldsymbol{x} + \boldsymbol{b}, \quad \boldsymbol{M} \,\in\, \mathbb{R}^{n\times n}, \quad \boldsymbol{b} \,\in\, \mathbb{R}^n\,.$$

Laut Definition der Ableitung ist $\boldsymbol{G}'(\boldsymbol{x}) = \boldsymbol{M}$ und Satz 11.6 zeigte, daß $r(\boldsymbol{M}) = r(\boldsymbol{G}'(\boldsymbol{x})) < 1$ eine notwendige und hinreichende Bedingung für die Konvergenz der Iteration $\boldsymbol{x}^{k+1} = \boldsymbol{G}(\boldsymbol{x}^k)$ war. Es liegt daher nahe, die entsprechende Bedingung für das NV zu untersuchen.

Wir werden unter geeigneten Differenzierbarkeitsvoraussetzungen zeigen, daß für die Iterationsfunktion $\boldsymbol{g}(\boldsymbol{x})$ des NV in einer einfachen Nullstelle $\boldsymbol{x}^*$ von $\boldsymbol{f}(\boldsymbol{x}) = 0$ sogar $\boldsymbol{g}'(\boldsymbol{x}^*) \equiv \boldsymbol{0}$ gilt. Hieraus schließen wir direkt auf die quadratische Konvergenz des NV. Wir führen den Beweis gleich für eine beliebige Raumdimension $n \in \mathbb{N}$.

Satz 11.7 Lokale Konvergenz des NV

Sei $\boldsymbol{f} = (f_1, \dots, f_n)^T \in C^3(D)$, $D \subset \mathbb{R}^n$ offen, und $\boldsymbol{x}^* \in D$ eine einfache Nullstelle von $\boldsymbol{f}$ (d.h. $\det \boldsymbol{f}'(\boldsymbol{x}^*) \neq 0$, bzw. $\boldsymbol{f}'(\boldsymbol{x}^*)$ regulär).
Dann existiert eine Umgebung von $\boldsymbol{x}^*$

$$K(\boldsymbol{x}^*) \;=\; \{\boldsymbol{x} \in \mathbb{R}^n,\; \|\boldsymbol{x} - \boldsymbol{x}^*\| < r\} \;\subset\; D\,,$$

so daß für die Folge der Newton–Iterierten $\{\boldsymbol{x}^k\}$ bei beliebigem Startwert $\boldsymbol{x}^0 \in K(\boldsymbol{x}^*)$ gilt

$$\boldsymbol{x}^k \in K(\boldsymbol{x}^*) \quad \forall\, k \geq 0 \qquad \text{und} \qquad \lim_{k\to\infty} \boldsymbol{x}^k \;=\; \boldsymbol{x}^*\,.$$

Die Folge $\{\boldsymbol{x}^k\}$ konvergiert quadratisch, und $\boldsymbol{x}^*$ ist die einzige Nullstelle von $\boldsymbol{f}$ in $K(\boldsymbol{x}^*)$.

Beweis:

Wegen $\boldsymbol{f} \in C^3(D)$ ist $\det \boldsymbol{f}'(\boldsymbol{x})$ eine stetige Funktion in D. Deshalb folgt aus $\det \boldsymbol{f}'(\boldsymbol{x}^*) \neq 0$, daß eine ganze, abgeschlossene Umgebung U_1 von $\boldsymbol{x}^*$ existiert

$$U_1 \;=\; \{\boldsymbol{x} \in \mathbb{R}^n;\; \|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty \leq r_1\} \;\subseteq\; D \quad \text{mit} \quad \det \boldsymbol{f}'(\boldsymbol{x}) \neq 0 \quad \forall\, \boldsymbol{x} \in U_1\,.$$

Deshalb ist die Iterationsfunktion des Newton–Verfahrens

$$\boldsymbol{g}(\boldsymbol{x}) \;=\; \boldsymbol{x} - \boldsymbol{f}'(\boldsymbol{x})^{-1}\,\boldsymbol{f}(\boldsymbol{x}) \quad \text{definiert} \quad \forall\, \boldsymbol{x} \in U_1 \;\text{ und }\; \boldsymbol{g} \in C^2(U_1)\,.$$

Wir zeigen zunächst $\boldsymbol{g}'(\boldsymbol{x}^*) = \boldsymbol{0}$.

Aus der Definition von $\boldsymbol{g}$ folgt

$$\boldsymbol{f}'(\boldsymbol{x})\,(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{x}) \;+\; \boldsymbol{f}(\boldsymbol{x}) \;=\; \boldsymbol{0}\,. \tag{11.28}$$

Komponentenweise bedeutet dies:

$$\sum_{i=1}^{n} \frac{\partial f_j(\boldsymbol{x})}{\partial x_i}\,(g_i(\boldsymbol{x}) - x_i) \;+\; f_j(\boldsymbol{x}) \;=\; 0\,, \qquad j = 1, \dots, n\,.$$

Differentiation nach x_k (vgl. Forster II, § 6, Satz 3) liefert

$$\sum_{i=1}^{n} \frac{\partial^2 f_j(\boldsymbol{x})}{\partial x_i\,\partial x_k}\,(g_i(\boldsymbol{x}) - x_i) + \sum_{i=1}^{n} \frac{\partial f_j(\boldsymbol{x})}{\partial x_i}\left(\frac{\partial g_i(\boldsymbol{x})}{\partial x_k} - \delta_{ik}\right) + \frac{\partial f_j(\boldsymbol{x})}{\partial x_k} \;=\; 0\,, \quad j = 1, \dots, n\,,$$

bzw. in Matrixschreibweise mit dem Einheitsvektor $\boldsymbol{e}^k$ (wir unterdrücken die Argumente)

$$\begin{pmatrix} \dfrac{\partial^2 f_1}{\partial x_1 \partial x_k} & \cdots & \dfrac{\partial^2 f_1}{\partial x_n \partial x_k} \\ \vdots & & \vdots \\ \dfrac{\partial^2 f_n}{\partial x_1 \partial x_k} & \cdots & \dfrac{\partial^2 f_n}{\partial x_n \partial x_k} \end{pmatrix} \begin{pmatrix} \\ \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{x} \\ \\ \end{pmatrix} + \underbrace{\begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix}}_{\boldsymbol{f}'(\boldsymbol{x})} \begin{pmatrix} \dfrac{\partial g_1}{\partial x_k} \\ \vdots \quad - \boldsymbol{e}^k \\ \dfrac{\partial g_n}{\partial x_k} \end{pmatrix} + \begin{pmatrix} \dfrac{\partial f_1}{\partial x_k} \\ \vdots \\ \dfrac{\partial f_n}{\partial x_k} \end{pmatrix} = \boldsymbol{0}.$$

Vollständige Differentiation von (11.28) nach $\boldsymbol{x}$ liefert also (für jedes k eine Spalte)

$$\boldsymbol{f}''(\boldsymbol{x})\,(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{x}) \;+\; \boldsymbol{f}'(\boldsymbol{x})\,(\boldsymbol{g}'(\boldsymbol{x}) - \boldsymbol{I}) \;+\; \boldsymbol{f}'(\boldsymbol{x}) \;=\; \boldsymbol{0}\,. \tag{11.29}$$

Wegen $\boldsymbol{f}(\boldsymbol{x}^*) = \boldsymbol{0}$ und $\boldsymbol{f}'(\boldsymbol{x}^*)$ regulär, folgt aus (11.28)

$$\boldsymbol{g}(\boldsymbol{x}^*) \quad = \quad \boldsymbol{x}^*\,,$$

und damit aus (11.29) $\boldsymbol{f}'(\boldsymbol{x}^*)\,\boldsymbol{g}'(\boldsymbol{x}^*) = \boldsymbol{0}$, und da homogene Gleichungssysteme mit regulärer Koeffizientenmatrix nur trivial lösbar sind

$$\boldsymbol{g}'(\boldsymbol{x}^*) \quad = \quad \boldsymbol{0}\,.$$

Aus dieser Eigenschaft folgern wir nun direkt die quadratische Konvergenz des NV. Da $\boldsymbol{g} \in C^2(U_1)$, kann man für jede Komponente g_j von $\boldsymbol{g}$ den Taylor'schen Satz anwenden (vgl. Forster II, § 7, Satz 2) (beachte: U_1 ist konvex)

$$\begin{aligned} g_j(\boldsymbol{x}) \;&=\; g_j(\boldsymbol{x}^*) \;+\; g_j'(\boldsymbol{x}^*)\,(\boldsymbol{x} - \boldsymbol{x}^*) \;+\; \frac{1}{2}\,(\boldsymbol{x} - \boldsymbol{x}^*)^T\, g_j''(\boldsymbol{\xi}^j)\,(\boldsymbol{x} - \boldsymbol{x}^*) \\ \text{mit} \quad \boldsymbol{\xi}^j \;&=\; \boldsymbol{x}^* + t_j(\boldsymbol{x} - \boldsymbol{x}^*), \quad t_j \in [0,1]\,. \end{aligned} \tag{11.30}$$

Da $\boldsymbol{g}'(\boldsymbol{x}^*) = \boldsymbol{0}$, folgt aus (11.30)

$$|g_j(\boldsymbol{x}) - g_j(\boldsymbol{x}^*)| \;=\; \frac{1}{2}\,|(\boldsymbol{x} - \boldsymbol{x}^*)^T\, g_j''(\boldsymbol{\xi}^j)\,(\boldsymbol{x} - \boldsymbol{x}^*)|\,. \tag{11.31}$$

Für eine Matrix $\boldsymbol{A} = (a_{ik})$ gilt

$$\begin{aligned} |\boldsymbol{x}^T \boldsymbol{A}\,\boldsymbol{x}| \;&=\; \left| \sum_{i=1}^n x_i \sum_{k=1}^n a_{ik}\,x_k \right| \\ &\leq\; \sum_{i=1}^n |x_i| \sum_{k=1}^n |a_{ik}|\,|x_k| \\ &\leq\; n\,\|\boldsymbol{x}\|_\infty^2 \,\max_i \sum_{k=1}^n |a_{ik}| \\ &=\; n\,\|\boldsymbol{A}\|_\infty\,\|\boldsymbol{x}\|_\infty^2\,. \end{aligned}$$

Also folgt aus (11.31) für alle j

$$\begin{aligned} |g_j(\boldsymbol{x}) - g_j(\boldsymbol{x}^*)| \;&\leq\; \frac{1}{2}\,n \max_j\, \|g_j''(\boldsymbol{\xi}^j)\|_\infty\, \|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty^2\,, \\ &\leq\; \frac{1}{2}\,n \max_j \max_{\boldsymbol{\xi} \in U_1} \|g_j''(\boldsymbol{\xi})\|_\infty\, \|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty^2\,, \end{aligned}$$

denn U_1 ist kompakt.

Also auch

$$
\begin{aligned}
\|\boldsymbol{g}(\boldsymbol{x})-\boldsymbol{g}(\boldsymbol{x}^*)\|_\infty &\le \underbrace{\frac{1}{2}\,n\max_j \max_{\boldsymbol{\xi}\in U_1}\|g_j''(\boldsymbol{\xi})\|_\infty}_{c}\ \|\boldsymbol{x}-\boldsymbol{x}^*\|_\infty^2\,, \\
&= c\,\|\boldsymbol{x}-\boldsymbol{x}^*\|_\infty^2\,.
\end{aligned}
\tag{11.32}
$$

Wählt man eine Ausgangsnäherung $\boldsymbol{x}^0 \in U_1$ mit

$$
c\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty \quad =: \quad \rho\, <\, 1 \tag{11.33}
$$

also

$$
\boldsymbol{x}^0 \in K(\boldsymbol{x}^*) \quad := \quad \left\{ \boldsymbol{x}\in \mathbb{R}^n;\ \|\boldsymbol{x}-\boldsymbol{x}^*\|_\infty \,\le\, r \,:=\, \min\left(\frac{\rho}{c}\,, r_1\right) \right\}\,,
$$

so folgt mit (11.33) und $\boldsymbol{x}^1 = \boldsymbol{g}(\boldsymbol{x}^0)$

$$
\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_\infty \quad\le\quad c\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty \quad\overset{(11.33)}{=}\quad \rho\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty \,<\, \|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty
$$

$$
\begin{aligned}
\|\boldsymbol{x}^2-\boldsymbol{x}^*\|_\infty &\overset{(11.32)}{\le} c\,\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_\infty\,\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_\infty \\
&\le \underbrace{c\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty}_{=\,\rho}\ \underbrace{\|\boldsymbol{x}^1-\boldsymbol{x}^*\|_\infty}_{\le\,\rho\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty} \quad\le\quad \rho^2\,\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty\,,
\end{aligned}
$$

also induktiv

$$
\|\boldsymbol{x}^k-\boldsymbol{x}^*\|_\infty \,\le\, \rho^k\|\boldsymbol{x}^0-\boldsymbol{x}^*\|_\infty\,.
$$

Wegen $\rho < 1$ folgt hieraus $\{\boldsymbol{x}^k\} \subset K(\boldsymbol{x}^*)$ und $\lim\limits_{k\to\infty} \boldsymbol{x}^k = \boldsymbol{x}^*$, und schließlich liefert (11.32)

$$
\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^*\|_\infty \,\le\, c\,\|\boldsymbol{x}^k-\boldsymbol{x}^*\|_\infty^2\,,
$$

also die quadratische Konvergenzgeschwindigkeit.

Eindeutigkeit: Wäre $\hat{\boldsymbol{x}} = \boldsymbol{g}(\hat{\boldsymbol{x}}) \in K(\boldsymbol{x}^*)$ eine weitere Nullstelle von $\boldsymbol{f}$, so würde aus (11.32) folgen

$$
\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|_\infty \,\le\, \underbrace{c\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|_\infty}_{<\,1}\ \|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|_\infty\,,
$$

also ein Widerspruch. ■

Bemerkungen und Hinweise

1) Die Konvergenzaussage erfordert im Gegensatz zum Fixpunktverfahren zwar Differenzierbarkeitseigenschaften aber *keine* Kontraktionsbedingungen. Beispiele, die den Abbildungen 4, 5 zu Beginn des Paragraphen entsprechen, werden mit erfaßt.

2) $\boldsymbol{f} \in C^3(D)$ ist eine Bequemlichkeitsvoraussetzung. Es genügt $\boldsymbol{f} \in C^1(D)$ und $\boldsymbol{f}'(\boldsymbol{x})$ Lipschitzstetig in D (vgl. Deuflhard/Hohmann: Numerische Mathematik, de Gruyter 1991). Der Beweis erfordert dann mehr Aufwand. Man kann sogar auf „$\boldsymbol{f}'(\boldsymbol{x}^*)$ regulär“ verzichten, erhält dann aber keine quadratische Konvergenz mehr (J. Werner: Numerische Mathematik I, Vieweg 1992).

3) In der praktischen Rechnung genügt es im allgemeinen, an Stelle der Jacobi–Matrix eine Näherung zu verwenden (z.B. Ersetzen der Ableitungen durch Differenzenquotienten).

4) Erfahrungsgemäß ist der Konvergenzbereich (= Menge der Startwerte, für die das NV konvergiert) 1. sehr klein und 2. im allgemeinen unbekannt. Man wird also das Verfahren mit einer Anfangsnäherung „auf Verdacht" starten und muß dann während der Durchführung des Verfahrens prüfen, ob Konvergenz vorliegt oder nicht. Es gibt Tests (z.B. *natürlicher Monotonietest*, vgl. Deuflhard/Hohmann), die im Laufe der Rechnung angeben, wann das Verfahren abzubrechen und mit einem — hoffentlich besseren — Startwert neu zu beginnnen ist.

   Gelegentlich kann der Konvergenzbereich erweitert werden durch Verwendung eines *gedämpften NV* (vgl. Deuflhard/Hohmann).

5) Im Beweis des Konvergenzsatzes wurde gezeigt, daß $\boldsymbol{g}'(\boldsymbol{x}^*) = \boldsymbol{\theta}$. Da $\boldsymbol{g}'$ stetig war, folgt hieraus die Existenz einer Kugelumgebung von $\boldsymbol{x}^*$

$$\tilde{K} \;=\; \{\boldsymbol{x} \in \mathbb{R}^n;\; \|\boldsymbol{x} - \boldsymbol{x}^*\| \le \tilde{r}\} \quad \text{mit} \quad \max_{\boldsymbol{\xi} \in \tilde{K}} \|\boldsymbol{g}'(\boldsymbol{\xi})\| \le L < 1\,.$$

   L ist dann eine Kontraktionskonstante für $\boldsymbol{g}$ in $\tilde{K}$ (vgl. Forster II, § 6, Satz 5 + Corollar).

   Wenn das NV konvergiert, also

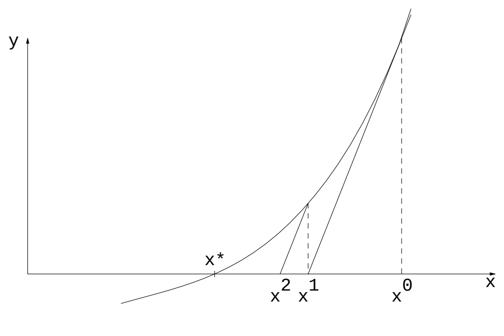
$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \;\xrightarrow{k}\; 0 \quad \text{und} \quad \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \;\xrightarrow{k}\; 0\,,$$

   kann man deshalb für hinreichend großes k im Anwendungsfall die Voraussetzungen des Korollars 11.2 erfüllen und damit auch die Existenz der Nullstelle sichern (vgl. dazu das Vorgehen im Beispiel nach Korollar 11.2) und eine Fehlerabschätzung erhalten. Wegen der notwendigen Berechnung von $\max_{\tilde{K}} \|\boldsymbol{g}'(\boldsymbol{\xi})\|$ ist das mit einigem numerischen Aufwand verbunden.

Wir erwähnen einige Varianten des Verfahrens.

Das vereinfachte Newton–Verfahren
benutzt immer die gleiche Koeffizientenmatrix (Steigung) $\boldsymbol{f}'(\boldsymbol{x}^0)$.

$$\begin{aligned} \boldsymbol{f}'(\boldsymbol{x}^0)\,\Delta\,\boldsymbol{x} \;&=\; -\boldsymbol{f}(\boldsymbol{x}^k)\,, \\ \boldsymbol{x}^{k+1} \;&=\; \boldsymbol{x}^k \,+\, \Delta\,\boldsymbol{x}\,. \end{aligned}$$

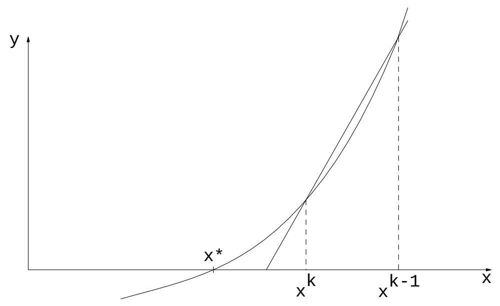


Bei den Gleichungssystemen, die pro Iterationsschritt zu lösen sind, ändert sich nur die rechte Seite. Das Verfahren erfordert also weniger Aufwand. Es konvergiert ebenfalls lokal, aber nur mit *linearer Konvergenzgeschwindigkeit*. Der Beweis kann durch Anwendung des Fixpunktsatzes geführt werden.

Das Sekantenverfahren

arbeitet ableitungsfrei und ist besonders effektiv im Falle einer reellen Funktion einer reellen Variablen. Es entsteht aus dem NV, indem man die Ableitung $f'(x^k)$ durch den Differenzenquotienten $\big(f(x^k) - f(x^{k-1})\big) \,/\, (x^k - x^{k-1})$ ersetzt:

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



Es konvergiert *superlinear* mit der Konvergenzgeschwindigkeit $\frac{1+\sqrt{5}}{2} \approx 1.62$. Die wesentliche Arbeit (Computerzeit) pro Iterationsschritt besteht im Auswerten von Funktionswerten. Beim NV müssen pro Iterationsschritt $f(x^k)$ und $f'(x^k)$ berechnet werden, beim Sekantenverfahren — abgesehen vom 1. Schritt — nur $f(x^k)$, denn $f(x^{k-1})$ ist vom vorigen Iterationsschritt bekannt. Grob gesprochen benötigen also 2 Iterationsschritte Sekantenverfahren den gleichen Aufwand wie ein Schritt NV. Das Sekantenverfahren ist ein sog. 2–stufiges Verfahren: Zur Berechnung von x^{k+1} benögtigt es x^k und x^{k-1}. Es ist also kein Fixpunktverfahren, benötigt also eine andere Beweistechnik.

Wir demonstrieren die Eigenschaften der Verfahren, insbesondere ihre superlineare Konvergenzgeschwindigkeit, indem wir zum Vergleich noch einmal das Beispiel aufgreifen, das wir schon mit der sukzessiven Iteration behandelt hatten im Anschluß an Korollar 11.2

$$\alpha_{k+1} \;=\; \sin\alpha_k \,+\, 2\pi\,\rho\,, \qquad \rho \;=\; 0.66$$

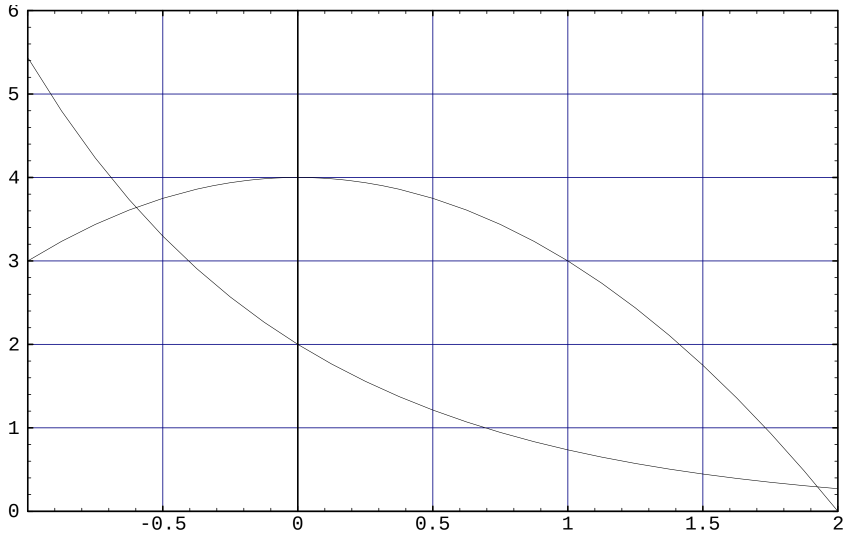
	Newton–Verfahren	Sekantenverfahren
α_0	4.1469	0.0
α_1	3.597148347	4.1469
α_2	3.654994283	3.445403349
α_3	3.655403058	3.671799769
α_4	3.65540308	3.655793574
α_5		3.655440228
α_6		3.65540308
Zahl der Iterationen	4	5
Funktionswert–auswertungen	10	6

Zur Erinnerung: Das Fixpunktverfahren hat diese Genauigkeit nach 60 Iterationen noch nicht erreicht.

Abschließend behandeln wir ein Beispiel zur Lösung zweier nichtlinearer Gleichungen mit dem NV.

Zu bestimmen sei ein Schnittpunkt zweier Funktionen

$$\begin{aligned} x_2 &= 2e^{-x_1}, \\ x_2 &= 4 - x_1^2. \end{aligned}$$



Gesucht ist also eine Nullstelle von

$$\boldsymbol{f}(\boldsymbol{x}) \quad = \quad \begin{pmatrix} x_2\, e^{x_1} - 2 \\ x_1^2 + x_2 - 4 \end{pmatrix} \quad \stackrel{!}{=} \quad \boldsymbol{\theta}\,.$$

Es ist

$$\boldsymbol{f}'(\boldsymbol{x}) \quad = \quad \begin{pmatrix} x_2\, e^{x_1} & e^{x_1} \\ 2x_1 & 1 \end{pmatrix}\,.$$

Aus der Zeichnung lesen wir für die rechte Nullstelle als Startwert den Näherungswert $(x_1^0, x_2^0) = (1.9,\ 0.3)$ ab. Dann ist

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}^0) &= (0.00577,\ -0.09)^T \\ \boldsymbol{f}'(\boldsymbol{x}^0) &= \begin{pmatrix} 2.00577 & 6.68589 \\ 3.8 & 1 \end{pmatrix} \end{aligned}$$

Die Lösung des Systems (11.27): $\boldsymbol{f}'(\boldsymbol{x}^0)\,\Delta\,\boldsymbol{x} = -\boldsymbol{f}(\boldsymbol{x}^0)$ ist $\Delta\,\boldsymbol{x} = (0.026,\ -0.0087)^T$, also $\boldsymbol{x}^1 = (1.926,\ 0.2913)^T$.

Nun werden Funktionswerte und Matrix $\boldsymbol{f}'(\boldsymbol{x}^1)$ neu berechnet zur Durchführung des nächsten Schritts. Die Ergebnisse sind in folgender Tabelle zusammengefaßt:

i	x_1^i	x_2^i	$f_1(\boldsymbol{x}^i)$	$f_2(\boldsymbol{x}^i)$	$\Vert\boldsymbol{x}^i - \boldsymbol{x}^{i-1}\Vert_\infty$
0	1.9	0.3	0.00577	-0.09	—
1	1.925961	0.291349	-0.000839	0.000674	0.02596
2	1.92573714	0.291536495	$-2.4_{10}-7$	$5.0_{10}-8$	0.000224
3	1.92573712	0.291536536	$-5.0_{10}-15$	$8.0_{10}-16$	$4.1_{10}-8$

§ 12 Eigenwertaufgaben für Matrizen

Solche Eigenwertaufgaben sind uns schon in § 11 Satz 11.6 begegnet. Der betragsgrößte Eigenwert gewisser Matrizen lieferte hinreichende und notwendige Konvergenzbedingungen für Iterationsverfahren zur Lösung linearer Gleichungssysteme. Ähnliche Anwendungen tauchen auf bei Stabilitäts– und Konvergenzuntersuchungen bei der Diskretisierung von partiellen Differentialgleichungen.

In den technischen, physikalischen Anwendungen kommen Eigenwertaufgaben für Matrizen oft als Sekundärprobleme vor. Zunächst werden Schwingungsphänomene durch Differentialgleichungen beschrieben, die einen Proportionalitätsfaktor (Parameter) enthalten (Eigenwertaufgaben für Differentialgleichungen). Es zeigt sich, daß diese Gleichungen nur für gewisse Werte dieses Parameters (die sog. Eigenwerte) Lösungen besitzen. Diese Lösungen beschreiben die sog. Eigenschwingungen eines Systems (z.B.: Eigenschwingungen von Brücken, rotierenden Maschinenteilen) und deren Schwingungsformen. Üblicherweise sind diese Differentialgleichungen jedoch nicht elementar auflösbar, obwohl sie eine Lösung besitzten (→ Theorie der gewöhnlichen und partiellen Differentialgleichungen). In diesen Fällen wird die Differentialgleichung *diskretisiert*, d.h. sie wird nur in einer endlichen Anzahl von Punkten betrachtet, wobei die Ableitungen durch Differenzenquotienten ersetzt werden. Man erhält so ein Gleichungssystem (das linear ist, sofern die Differentialgleichung linear war) in dem der o.g. Parameter natürlich ebenfalls wieder auftaucht: eine Eigenwertaufgabe für Matrizen. Lösungen dieser Eigenwertaufgabe liefern Näherungslösungen für die ursprüngliche Eigenwertaufgabe für die Differentialgleichung. Diskretisierungsverfahren sind ein wichtiges, eigenständiges Kapitel in der Numerischen Mathematik, auf das wir hier aus Zeitgründen nur beispielhaft eingehen können.

Eine beeindruckende Sammlung von Eigenwertaufgaben findet man in dem Buch von L. Collatz: „Eigenwertaufgaben mit technischen Anwendungen“, Akademische Verlagsgesellschaft.

Um die oben beschriebene Vorgehensweise zu erläutern, betrachten wir hier das (super–) einfache Beispiel der Eigenschwingungen einer eingespannten Saite. Diese Aufgabe ist sogar elementar lösbar und erlaubt uns daher auch den Vergleich mit den Näherungslösungen, die man aus der diskretisierten Eigenwertaufgabe erhält.

Beispiel: Schwingende Saite

Die Schwingungen $u(x,t)$ einer fest eingespannten Saite der Länge ℓ, die nur in der (x,u)–Ebene schwingt, genügen der Differentialgleichung (vgl. Tychonoff–Samarski: Differentialgleichungen der mathematischen Physik)

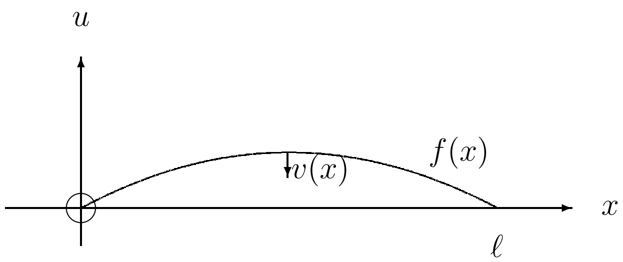
$$\frac{\partial^2 u(x,t)}{\partial t^2} = a^2 \frac{\partial^2 u(x,t)}{\partial x^2} \qquad (a = \text{Materialkonstante}), \tag{12.1}$$

mit den Randbedingungen

$$u(0,t) = u(\ell,t) = 0 \qquad \text{(Saite fest eingespannt)} \tag{12.2}$$

und den Anfangsbedingungen

$$\begin{aligned} u(x,0) &= f(x) \qquad \text{(Anfangsauslenkung)}\,, \\ \frac{\partial\, u(x,0)}{\partial\, t} &= v(x) \qquad \text{(Anfangsgeschwindigkeit)}\,. \end{aligned} \tag{12.3}$$



Wir suchen nach Lösungen in der Gestalt $u(x,t) = y(x)\,z(t)$ (Trennungsansatz). Wird dieser Ansatz in die Differentialgleichung (12.1) eingesetzt, so erhält man

$$\frac{z''(t)}{a^2\,z(t)} = \frac{y''(x)}{y(x)}\,. \tag{12.4}$$

Da die linke Seite nicht von x, die rechte nicht von t abhängt, ist jede Seite eine konstante Größe, die wir mit λ bezeichnen:

$$\frac{z''(t)}{a^2\,z(t)} = \lambda = \frac{y''(x)}{y(x)}\,,$$

bzw.

$$z''(t) = a^2\,\lambda\,z(t) \tag{12.5}$$

$$y''(x) = \lambda\,y(x)\,. \tag{12.6}$$

Wir betrachten zuerst die räumliche Differentialgleichung (12.6). Sie hat mit beliebigen Konstanten α, β die allgemeinen, reellen Lösungen

$$y(x) = \begin{cases} \alpha\,\sin\sqrt{-\lambda}\,x + \beta\,\cos\sqrt{-\lambda}\,x & \text{falls} \quad \lambda < 0\,, \\ \alpha + \beta\,x\,, & \text{falls} \quad \lambda = 0\,, \\ \alpha\,\sinh\sqrt{\lambda}\,x + \beta\,\cosh\sqrt{\lambda}\,x\,, & \text{falls} \quad \lambda > 0\,, \end{cases} \tag{12.7}$$

deren Konstanten α, β wir durch die Anpassung an die Randwerte (12.2) bestimmen wollen. Es soll gelten

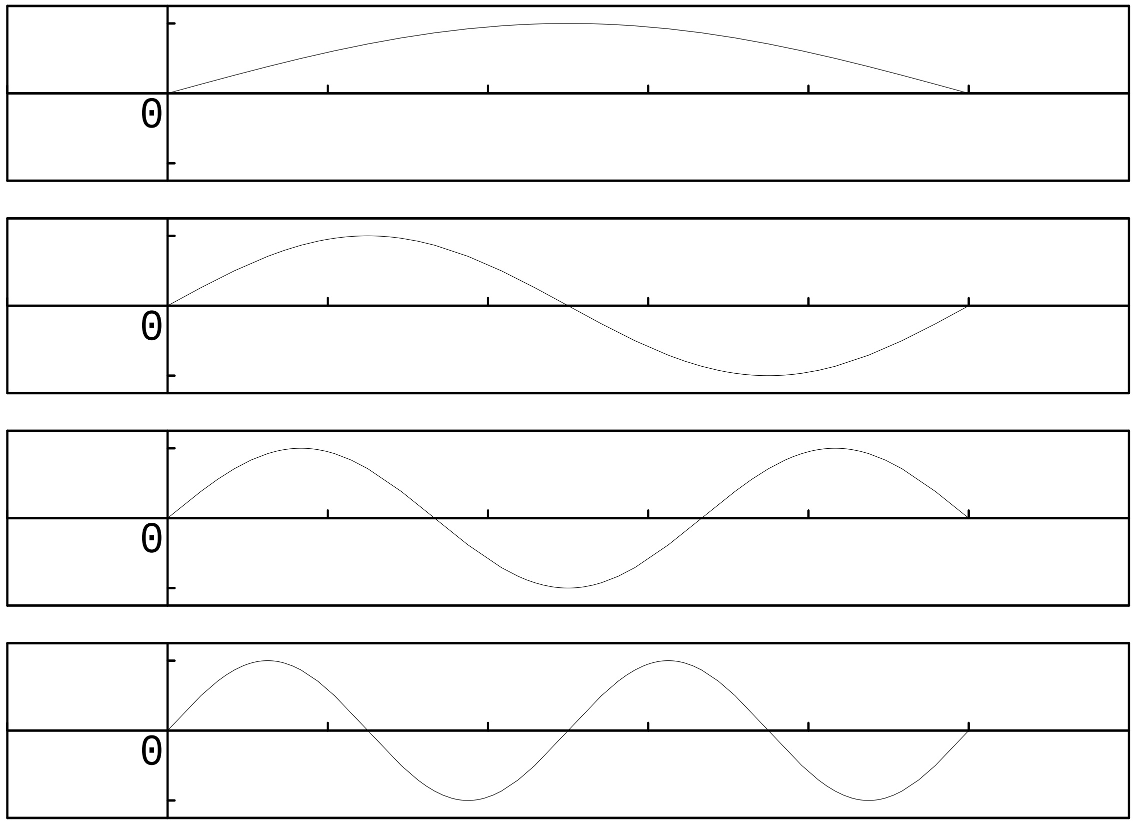
$$y(0) = y(\ell) = 0\,. \tag{12.8}$$

Sowohl für den Fall $\lambda = 0$ als auch für den Fall $\lambda > 0$ folgt aus diesen Gleichungen durch Einsetzen $\alpha = \beta = 0$, also nur die triviale Lösung. Es bleibt also nur der Fall $\lambda < 0$. Aus (12.8) folgt

$$\begin{aligned} \alpha\,\sin 0 + \beta\,\cos 0 &= 0 \;\Rightarrow\; \beta = 0 \quad \text{und hiermit} \\ \alpha\,\sin\sqrt{-\lambda}\,\ell &= 0 \;\Rightarrow\; \sqrt{-\lambda}\,\ell = k\,\pi \ \text{für}\ k \in \mathbb{Z} \\ &\qquad \text{bzw.}\ \lambda_k = -\tfrac{k^2\,\pi^2}{\ell^2}\,,\ k \in \mathbb{Z}\,. \end{aligned}$$

Nur für diese speziellen Werte (Eigenwerte) ist also eine Lösung möglich:

$$y_k(x) \;=\; \alpha \sin\left(\frac{k\,\pi}{\ell}\,x\right), \quad k \in \mathbb{Z},\ \alpha \in \mathbb{R}$$



Mit diesen λ_k integriert man die zeitliche Differentialgleichung (12.5)

$$z''(t) \;=\; \lambda_k\, a^2\, z(t)$$

und erhält analog zu (12.7) die möglichen Lösungen

$$z_k(t) \;=\; \alpha_k \sin\left(a\frac{k\,\pi}{\ell}\,t\right) + \beta_k \cos\left(a\frac{k\,\pi}{\ell}\,t\right), \qquad k \in \mathbb{Z},\ \alpha_k, \beta_k \in \mathbb{R}\,,$$

und damit als mögliche Lösungen für (12.1) gemäß Trennungsansatz

$$u_k(x,t) \;=\; \left(b_k \cos\left(a\frac{k\,\pi}{\ell}\,t\right) + a_k \sin\left(a\frac{k\,\pi}{\ell}\,t\right)\right) \sin\left(\frac{k\,\pi}{\ell}\,x\right), \quad a_k, b_k \in \mathbb{R},\ k \in \mathbb{Z}\,.$$

Man kann die Anfangswerte (12.3) erfüllen und damit die Konstanten a_k, b_k bestimmen durch Überlagerung (Superposition) der partikulären Lösungen $u_k(x,t)$ und Entwicklung von f bzw. v in eine Fourierreihe

$$\begin{aligned}
u(x,0) \quad &= \quad \sum_{k=1}^{\infty} u_k(x,0) \quad &&= \quad \sum_{k=1}^{\infty} b_k \sin\left(\frac{k\,\pi}{\ell}\,x\right) \quad &&\stackrel{!}{=} \quad f(x) \\
\frac{\partial\, u(x,0)}{\partial\, t} \quad &= \quad \sum_{k=1}^{\infty} \frac{\partial}{\partial\, t}\, u_k(x,0) \quad &&= \quad \sum_{k=1}^{\infty} a_k\, \frac{ak\,\pi}{\ell}\, \sin\left(\frac{k\,\pi}{\ell}\,x\right) \quad &&\stackrel{!}{=} \quad v(x)
\end{aligned}$$

Setzt man f und v als ungerade Funktionen auf das Intervall $[-\ell, 0]$ fort, so erhält man (vgl. § 10)

$$b_k \;=\; \frac{2}{\ell}\int\limits_0^{\ell} f(x)\,\sin\frac{k\,\pi\,x}{\ell}\,dx\,, \qquad a_k \;=\; \frac{2}{a\,k\,\pi}\int\limits_0^{\ell} v(x)\,\sin\frac{k\,\pi\,x}{\ell}\,dx\,.$$

Dieses Beispiel zeigt, daß die Eigenwerte von wesentlicher Bedeutung für die Lösung der Aufgabe sind.
Was tut man aber, wenn die Differentialgleichungen (in diesem Beispiel (12.5), (12.6)) nicht lösbar sind? Wie in der Einleitung schon angedeutet, wird *diskretisiert.*

Diskretisierung einer Differentialgleichung

Wir beschränken uns hier beispielhaft auf die Diskretisierung der räumlichen Differentialgleichung (vgl. (12.6))

$$y''(x) \;=\; \lambda\, y(x) \tag{12.9}$$

mit den Randwerten (vgl. (12.8))

$$y(0) \;=\; 0\,, \qquad y(\ell) \;=\; 0 \tag{12.10}$$

Wir teilen das Intervall $[0,\ell]$ in $n+1$ gleichlange Teilintervalle $[x_j, x_{j+1}]$

$$0 \;=\; x_0 \;<\; x_1 \;<\; \ldots \;<\; x_{n+1} \;=\; \ell\,, \qquad x_{j+1} - x_j \;=\; \frac{\ell}{n+1} \;=:\; h\,,$$

und betrachten die Differentialgleichung (12.9) nur in den Punkten x_j:

$$y''(x_j) \;=\; \lambda\, y(x_j)\,, \qquad j = 1, \ldots, n \tag{12.11}$$

(Die Lösung für $j = 0$ und $j = n+1$ ist durch (12.10) schon bekannt.)

Wir wollen zur näherungsweisen Lösung die Ableitungen durch Differenzenquotienten ersetzen. Die 1. Ableitung läßt sich näherungsweise durch den *rückwärtsgenommenen* und den *vorwärtsgenommenen* Differenzenquotienten ersetzen:

$$\overline{y_j^-} \;:=\; \frac{y(x_j) - y(x_{j-1})}{h} \;\approx\; y'(x_j) \;\approx\; \frac{y(x_{j+1}) - y(x_j)}{h} \;=:\; \overline{y_j^+}\,.$$

Aus Symmetriegründen ersetzen wir $y''(x_j)$ durch den *zentralen* Differenzenquotienten 2. Ordnung

$$\begin{aligned}
y''(x_j) \;\approx\; \frac{\overline{y_j^+} - \overline{y_j^-}}{h} \;&=\; \frac{\frac{y(x_{j+1})-y(x_j)}{h} \;-\; \frac{y(x_j)-y(x_{j-1})}{h}}{h} \\
&=\; \frac{y(x_{j+1}) \;-\; 2y(x_j) \;+\; y(x_{j-1})}{h^2} \;=:\; \overline{\overline{y}}_j
\end{aligned} \tag{12.12}$$

Damit geht (12.11) über in

$$\overline{\overline{y_j}} \;\approx\; \lambda\, y(x_j)\,, \qquad j = 1, \ldots, n\,. \tag{12.13}$$

Wir bezeichnen durch $y_j \approx y(x_j)$ Näherungswerte für die Lösung der Aufgabe, setzen diese in (12.13) gemäß (12.12) ein und betrachten die Lösungen y_j des dadurch entstehenden Systems

$$\frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} \;=\; \lambda\, y_j\,, \qquad j = 1, \ldots, n \tag{12.14}$$

als Näherungen für die exakten Werte $y(x_j)$.

Bemerkung

Daß dieses Verfahren gerechtfertigt und wie es ggf. verfeinert werden kann, bedarf einer ausführlichen Theorie (→ Theorie der Diskretisierungsverfahren).

Das Gleichungssystem (12.14) hat die Gestalt

$$\frac{1}{h^2}\begin{pmatrix} -2 & 1 & & \mathbf{O} \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ \mathbf{O} & & 1 & -2 \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \lambda \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}, \qquad h = \frac{\ell}{n+1}\,. \tag{12.15}$$

Die Werte λ, für welche dieses System lösbar ist, heißen *Eigenwerte* der Koeffizientenmatrix aus (12.15). Die zugehörigen Lösungen $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ heißen *Eigenvektoren*, und (12.15) bezeichnet man als *Matrixeigenwertaufgabe*.

In der folgenden Tabelle listen wir die ersten 5 Eigenwerte von (12.15) für die Fälle $\ell = 6$, $n = 5, 10, 100$ auf und vergleichen sie mit den Eigenwerten der Differentialgleichungseigenwertaufgabe (12.6), (12.8).

	EWe der DGL–EWA	EWe der Matrixeigenwertaufgabe		
	$-\lambda_k = \frac{k^2\pi^2}{\ell^2}$	$n = 5$	$n = 10$	$n = 100$
$-\lambda_1$	0.274155	0.267949	0.272297	0.274134
$-\lambda_2$	1.096622	1.0	1.06713	1.09627
$-\lambda_3$	2.467401	2.0	2.3201	2.46561
$-\lambda_4$	4.386490	3.0	3.92971	4.38084
$-\lambda_5$	6.853891	3.732050	5.76555	6.84009

Man erkennt, daß die Eigenwerte der diskretisierten Aufgabe nur sehr langsam mit feiner werdender Diskretisierung gegen die Eigenwerte der Differentialgleichungsaufgabe konvergieren.

Matrixeigenwertaufgaben

Wir untersuchen im folgenden die

Matrixeigenwertaufgabe: Für die Gleichung

$$\boldsymbol{A}\,\boldsymbol{x} \;=\; \lambda\,\boldsymbol{x}\,, \qquad \boldsymbol{A} \in \mathbb{C}^{n\times n} \tag{12.16}$$

sind Lösungen $\lambda \in \mathbb{C}$ und $\boldsymbol{x} \in \mathbb{C}^n$ zu bestimmen.

Dazu erklären wir folgende Begriffe:

Definition 12.1

a) $\lambda \in \mathbb{C}$ heißt *Eigenwert* (EW) von $A \Leftrightarrow \exists\, \boldsymbol{x} \in \mathbb{C}^n,\; \boldsymbol{x} \neq \boldsymbol{0}$ mit $\boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}$.

b) Ein Vektor $\boldsymbol{x}$ mit der Eigenschaft aus a) heißt *Eigenvektor* (EV) zum Eigen– wert λ.

c) Die Menge der Eigenwerte

$$\sigma\,(\boldsymbol{A}) \quad := \quad \{\lambda \in \mathbb{C}\,;\; \exists\, \boldsymbol{x} \neq \boldsymbol{0}\; :\; \boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}\}$$

heißt *Spektrum* von $\boldsymbol{A}$.

d) Die Menge

$$\mathrm{Eig}\,(\boldsymbol{A};\lambda) \quad := \quad \{\boldsymbol{x} \in \mathbb{C}^n\,;\; \boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}\}$$

heißt *Eigenvektorraum* zum Eigenwert λ.

e) Die Menge aller Eigenvektoren von $\boldsymbol{A}$

$$\mathrm{Eig}\,(\boldsymbol{A}) \quad := \quad \{\boldsymbol{x} \in \mathbb{C}^n\,;\; \exists\,\lambda \in \mathbb{C}\; :\; \boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}\}$$

heißt *Eigenraum* von $\boldsymbol{A}$.

Beachte: Wegen der Homogenität der Gleichung (12.16) sind Eigenvektoren nur bis auf einen Faktor $\neq 0$ bestimmt.

Offensichtlich ist (12.16) gleichwertig mit dem homogenen Gleichungssystem

$$(\boldsymbol{A} - \lambda\,\boldsymbol{E})\,\boldsymbol{x} \;=\; \boldsymbol{0} \qquad (\boldsymbol{E} = \text{ Einheitsmatrix})\,. \tag{12.17}$$

Dieses ist genau dann nichttrivial lösbar (gesucht sind Eigenvektoren $\neq 0$), wenn $\boldsymbol{A}-\lambda\,\boldsymbol{E}$ singulär ist, d.h. genau dann wenn

$$\det\,(\boldsymbol{A} - \lambda\,\boldsymbol{E}) \quad = \quad 0\,. \tag{12.18}$$

Durch Entwickeln dieser Determinante erhält man ein Polynom $p(\lambda)$ der Form

$$p(\lambda) \;=\; (-1)^n \left(\lambda^n + \sum_{i=0}^{n-1} a_i\,\lambda^i\right) \;=\; \det\,(\boldsymbol{A} - \lambda\,\boldsymbol{E})\,, \qquad a_i \in \mathbb{R}\,, \tag{12.19}$$

das sog. *charakteristische Polynom* der Matrix $\boldsymbol{A}$ (vgl. Fischer: Lineare Algebra, § 5.2.3).

Hieraus und aus der Determinantentheorie ergeben sich einige grundlegende Eigenschaften, die wir zunächst zusammenstellen.

Die Eigenwerte sind die Nullstellen des charakteristischen Polynoms. (12.20)

Da ein Polynom nach dem Fundamentalsatz der Algebra genau n (ggf. komplexe) Nullstellen besitzt, folgt:

Sind $\lambda_1, \dots, \lambda_k \in \mathbb{C}$ alle verschiedenen Nullstellen des charakteristischen Polynoms $p(\lambda)$, so existiert eine Darstellung

$$p(\lambda) \;=\; (-1)^n \prod_{i=1}^{k} (\lambda - \lambda_i)^{\sigma_i}\,, \quad \sigma_i \in \mathbb{N},\; \sum_{i=1}^{k} \sigma_i = n\,,$$

σ_i heißt *algebraische Vielfachheit* des EW λ_i.

(12.21)

$\boldsymbol{A}$ und $\boldsymbol{A}^T$ haben dasselbe charakteristische Polynom, also auch dieselben EWe inklusive ihrer Vielfachheiten,

$$\text{denn} \quad \det \boldsymbol{B} \;=\; \det \boldsymbol{B}^T \quad \forall\, \boldsymbol{B} \in \mathbb{C}^{n\times n}\,.$$

(12.22)

$\boldsymbol{A}$ ist regulär $\iff$ $\lambda = 0$ ist kein EW von $\boldsymbol{A}$

denn dann hat $\boldsymbol{A}\,\boldsymbol{x} \;=\; (\boldsymbol{A} - 0\boldsymbol{E})\,\boldsymbol{x} = \boldsymbol{0}$ nur die triviale Lösung.

(12.23)

Ist λ komplexer EW einer *reellen* Matrix $\boldsymbol{A}$, so ist der konjugiert komplexe Wert $\overline{\lambda}$ auch Eigenwert von $\boldsymbol{A}$,

denn $\boldsymbol{A} \in \mathbb{R}^{n\times n} \;\Rightarrow\; p(\lambda) \;=\; (-1)^n \left(\lambda^n + \sum\limits_{i=0}^{n-1} a_i\,\lambda^i\right)$

hat nur reelle Koeffizienten, also $a_i = \overline{a_i} \;\Rightarrow$

$$\lambda^n + \sum_{i=0}^{n-1} a_i\,\lambda^i \;=\; 0 \;=\; \overline{\lambda}^n + \sum_{i=0}^{n-1} \overline{a_i}\,\overline{\lambda}^i \;=\; \overline{\lambda}^n + \sum_{i=0}^{n-1} a_i\,\overline{\lambda}^i.$$

(12.24)

Wir haben oben gesehen, daß zu jeder Matrix ein Polynom existiert, dessen Nullstellen die Eigenwerte sind. Auch das Umgekehrte ist richtig. Zu jedem Polynom existiert auch eine Matrix, deren Eigenwerte die Nullstellen des Polynoms sind.

Satz 12.2
Die Matrix

$$A \quad = \quad \begin{pmatrix} 0 & \dots & \dots & \dots & 0 & -a_0 \\ 1 & \ddots & & & \vdots & -a_1 \\ 0 & \ddots & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & \ddots & 0 & -a_{n-2} \\ 0 & \dots & \dots & 0 & 1 & -a_{n-1} \end{pmatrix}$$

hat das charakteristische Polynom

$$\det\left(\boldsymbol{A} - \lambda\,\boldsymbol{E}\right) \quad = \quad (-1)^n \left(\lambda^n + \sum_{i=0}^{n-1} a_i\,\lambda^i\right) .$$

$\boldsymbol{A}$ heißt *Frobenius'sche Begleitmatrix* des Polynoms.

Beweis: Übung in vollständiger Induktion.

Wir sehen also, daß die Aufgabe, alle Nullstellen eines Polynoms zu bestimmen, äquivalent ist zu der, alle Eigenwerte einer Matrix zu bestimmen. Für $n \geq 5$ ist diese Aufgabe also nicht mehr elementar lösbar.

Um einen besseren Überblick zu gewinnen, betrachten wir zunächst einige Beispielmatrizen $\boldsymbol{A}$:

$$\text{a)} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{b)} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \text{c)} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad \text{d)} \begin{bmatrix} a_{11} & \dots & a_{11} \\ & \ddots & \vdots \\ O & & a_{nn} \end{bmatrix}$$

und stellen fest zu:

a) $p(\lambda) = (1-\lambda)^2$, $\lambda_1 = 1$ ist doppelte Nullstelle, also Eigenwert der algebraischen Vielfachheit 2.

$(\boldsymbol{A} - \lambda_1\,\boldsymbol{E})\,\boldsymbol{x} = \boldsymbol{0}$ hat 2 linear unabhängige Lösungen (Eigenvektoren). Ein mehrfacher Eigenwert kann also mehrere linear unabhängige Eigenvektoren haben.

b) $\det \begin{bmatrix} 0-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix} = -\lambda(1-\lambda) = 0, \; \lambda_1 = 0$ und $\lambda_2 = 1$ sind EWe.

Zu $\lambda_1 = 0$ gehört der EV $\boldsymbol{x} = (x_1, 0)^T$ mit $x_1 \neq 0$, denn $\boldsymbol{x}$ ist Lösung von $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{0}$.

Zu $\lambda_2 = 1$ gehört der EV $\boldsymbol{x} = (x_1, x_1)^T$ mit $x_1 \neq 0$, denn $\begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$ hat die Lösung $x_2 = x_1$.

c) det $\begin{bmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix} = (1-\lambda)^2 = 0$, $\lambda_1 = 1$ ist doppelter Eigenwert, aber im Gegensatz zu Beispiel a) existiert nur 1 Eigenvektor, denn $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$ hat nur die Lösung $(x_1, 0)^T$ mit $x_1 \neq 0$.

Ein mehrfacher Eigenwert muß nicht mehrere linear unabhängige Eigenvektoren haben.

d) Der Determinantenentwicklungssatz liefert:

Bei einer oberen (oder unteren) Dreiecksmatrix und damit auch bei einer Diagonalmatrix stehen in der Diagonale die Eigenwerte. Zu jedem Eigenwert gehört mindestens ein EV, denn (12.17) hat dann eine nicht triviale Lösung. Daß mehr nicht ausgesagt werden kann, zeigen insbesondere die Beispiele a)–c). Sie zeigen u.a. auch, daß die algebraische Vielfachheit eines Eigenwerts nicht übereinstimmen muß mit der Zahl der zugehörigen linear unabhängigen Eigenvektoren.

Wir führen deshalb einen weiteren Begriff ein:

geometrische Vielfachheit eines Eigenwerts λ *von* $\boldsymbol{A}$:= dim Eig $(\boldsymbol{A}; \lambda)$
= Anzahl der linear unab–hängigen EVen von λ.

Einen vollständigen Überblick über die Zahl der verschiedenen Eigenwerte einer Matrix $\boldsymbol{A}$, ihre algebraische und geometrische Vielfachheit, liefert die Jordan'sche Normalform ($\rightarrow$ Lineare Algebra u. Analytische Geometrie). Obwohl sie nicht konstruktiv ist, hat sie auch praktische Bedeutung, wie der Beweis von Satz 11.6 gezeigt hat.

Berechnung von Eigenwerten

Eine ganz grobe Lagebestimmung für die Eigenwerte λ einer Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ lieferte schon Satz 11.6: $|\lambda| \leq \|\boldsymbol{A}\|$ für jede Matrixnorm, die einer Vektornorm zugeordnet ist.

Die Tatsache „Bei Diagonalmatrizen stehen die EWe in der Diagonale“ führte zu der Vermutung: „Wenn sich eine Matrix nicht allzusehr von einer Diagonalmatrix unterscheidet, werden sich ihre EWe nicht allzusehr von den Diagonalelementen unterscheiden“. Diese Überlegung wird präzisiert (insbesondere durch Beweisteil c)) durch den

Satz 12.3 Satz von Gerschgorin
Sei $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ $(\mathbb{C}^{n\times n})$, dann gilt:

a) Alle Eigenwerte liegen in der Vereinigung der Kreise

$$Z_i \quad := \quad \left\{ z \in \mathbb{C}\ ;\ |z - a_{ii}| \ \leq \ \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\} .$$

b) Alle Eigenwerte liegen in der Vereinigung der Kreise

$$S_j \quad := \quad \left\{ z \in \mathbb{C}\ ;\ |z - a_{jj}| \ \leq \ \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}| \right\} .$$

c) Jede Zusammenhangskomponente (= maximale zusammenhängende Teilmenge) von $\cup Z_i$ oder $\cup S_j$ enthält genau so viele Eigenwerte wie Kreise an der Komponente beteiligt sind (Eigenwerte und Kreise werden dabei entsprechend ihrer Vielfachheit gezählt).

Beweis:

a) Sei $\lambda \in \sigma(\boldsymbol{A}) \ \Rightarrow\ \exists\, \boldsymbol{x} = (x_1, \ldots, x_n)^T \neq 0\ :\ (\boldsymbol{A} - \lambda\, \boldsymbol{E})\, \boldsymbol{x} = 0$ d.h.

$$\sum_{j=1}^{n} a_{ij}\, x_j \ -\ \lambda\, x_i \ = \ 0, \qquad i = 1, \ldots, n\,,$$

oder

$$(\lambda - a_{ii})\, x_i \ = \ \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}\, x_j, \qquad i = 1, \ldots, n\,.$$

Wir betrachten aus diesem System die μ–te Gleichung, wobei $\|\boldsymbol{x}\|_\infty = |x_\mu|$:

$$(\lambda - a_{\mu\mu})\, x_\mu \quad = \quad \sum_{\substack{j=1 \\ j\neq \mu}}^{n} a_{\mu\, j}\, x_j\ .$$

Division durch x_μ und die Dreiecksungleichung liefern wegen $|x_\mu| \geq |x_i|\ \forall\, i$:

$$|\lambda - a_{\mu\mu}| \ \leq \ \sum_{\substack{j=1 \\ j\neq \mu}}^{n} |a_{\mu\, j}|\, \frac{|x_j|}{|x_\mu|} \ \leq \ \sum_{\substack{j=1 \\ j\neq \mu}}^{n} |a_{\mu\, j}|, \qquad \text{also a)}\,.$$

b) folgt aus a) wegen (12.22).

c) wollen wir hier nur andeuten. (Eine ausführliche Darstellung findet man in J. Werner: Numerische Mathematik 2, § 5.1, Vieweg.) Die Nullstellen eines Polynoms hängen stetig von den Polynomkoeffizienten ab. Da die Determinante einer Matrix stetig von den Matrixelementen abhängt und die Eigenwerte einer Matrix die Nullstellen ihres charakteristischen Polynoms sind, folgt:

$(*)$ Die Eigenwerte einer Matrix hängen stetig von den Matrixkoeffizienten ab.

Diese Aussage benutzen wir. Wir betrachten die Matrizen

$$\boldsymbol{A}(t) \;=\; \boldsymbol{D} + t(\boldsymbol{A} - \boldsymbol{D}) \quad \text{mit} \quad \boldsymbol{D} \;=\; \operatorname{diag}(\boldsymbol{A}) \quad \text{für} \quad t \in [0,1]\,.$$

Für $t = 0$, also $\boldsymbol{A}(0) = \boldsymbol{D}$, ist die Aussage c) richtig. Läßt man nun $t \to 1$ gehen, so folgt mit $(*)$, daß die Eigenwerte nicht „aus der Zusammenhangskomponente der Gerschgorinkreise herausspringen können“. ■

Beachte:

Folgende Aussage ist falsch: „In jedem Gerschgorinkreis liegt mindestens ein EW.“

Gegenbeispiel: $A = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$ hat die EWe $\lambda_{1,2} = \pm\sqrt{2}$, und es gibt keinen EW in $Z_1 = \{z \in \mathbb{C};\; |z - 0| \le 1\}$.

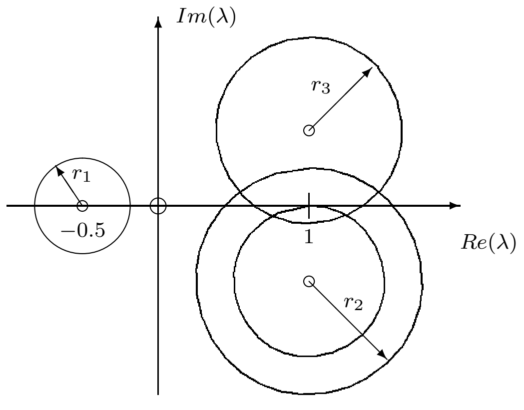
Beispiel zu Satz 12.3

Wir betrachten die Matrix

$$\begin{pmatrix} 1+0.5i & 0.5 & 0.1 \\ 0.3 & 1-0.5i & 0.5 \\ 0.4 & 0 & -0.5 \end{pmatrix}$$

Die Gerschgorin–Radien der Kreise Z_i sind $r_1 = 0.6$, $r_2 = 0.8$, $r_3 = 0.4$, die Radien der Kreise S_i sind $\tilde{r}_1 = 0.7$, $\tilde{r}_2 = 0.5$, $\tilde{r}_3 = 0.6$.

Neben den Kreisen mit den Radien r_1, r_2, r_3 sind mit stark ausgezogenen Rändern die Kreise eingezeichnet, die man erhält, wenn man zur Einschließung der Eigenwerte den Durchschnitt $(\cup Z_i) \cap (\cup S_i)$ der Aussagen a) und b) heranzieht.



Als erste Berechnungsmöglichkeit für die Eigenwerte liegt der Versuch nahe, sie als Nullstellen des charakteristischen Polynoms zu berechnen — zum Beispiel mit dem Newton–Verfahren (oder Sekanten-Verfahren, da ableitungsfrei) — zumal eine Schätzung von Anfangsnäherungen durch den Satz von Gerschgorin erleichtert wird. Als allgemeine Methode ist dies jedoch nicht zu empfehlen, denn einerseits versagt dieses Vorgehen vollständig bei Vorliegen komplexer Eigenwerte (ob solche existieren, weiß man ja im Vorhinein nicht), andereseits ist die Aufstellung des charakteristischen Polynoms i.allg. numerisch aufwendig. Die große Anzahl von Multiplikationen durch die Entwicklung der Determinante bedingt notwendigerweise viele Rundungen. Die Koeffizienten des charakteristischen Polynoms sind also stark rundungsfehlerbehaftet. Außerdem hängen die Nullstellen eines Polynoms *sehr empfindlich* von dessen Koeffizienten ab. Wir belegen dies durch

Beispiel (Wilkinson)

Wird das Polynom

$$p(\lambda) \quad = \quad (\lambda - 1)\,(\lambda - 2) \cdot \;\ldots\; \cdot (\lambda - 20)$$

ausmultipliziert, so ergeben sich Koeffizienten in der Größenordnung zwischen 1 (Koeffizient von λ^{20}) und ca. 10^{20} (der konstante Term ist z.B. 20!). Wir stören den Koeffizienten von λ^{19} (der den Wert 210 hat) um den sehr kleinen Wert $\varepsilon := 2^{-23} \approx 10^{-7}$. In der folgenden Tabelle sind die exakten Nullstellen des gestörten Polynoms

$$\tilde{p}(\lambda) \quad = \quad p(\lambda) \;-\; \varepsilon \cdot \lambda^{19}$$

eingetragen. Trotzt der extrem kleinen Störung sind die Fehler beachtlich. Insbesondere sind fünf Nullstellenpaare komplex.

1.000 000 000	$10.095\,266\,145 \;\pm\; 0.643\,500\,904\,i$
2.000 000 000	$11.793\,633\,881 \;\pm\; 1.652\,329\,728\,i$
3.000 000 000	$13.992\,358\,137 \;\pm\; 2.518\,830\,070\,i$
4.000 000 000	$16.730\,737\,466 \;\pm\; 2.812\,624\,894\,i$
4.999 999 928	$19.502\,439\,400 \;\pm\; 1.940\,330\,347\,i$
6.000 006 944	
6.999 697 234	
8.007 267 603	
8.917 250 249	
20.846 908 101	

Tabelle 5.1: Exakte Nullstellen des Polynoms $\tilde{p}(\lambda)$ für $\varepsilon := 2^{-23}$.

Daher liegt der Gedanke nahe, Matrizen so umzuformen, daß sich 1) die Eigenwerte dabei nicht ändern und 2) das Umformungsergebnis eine der Matrixgestalten aus Beispiel d) liefert, also eine Dreiecks– oder Diagonalmatrix, aus der sich die Eigenwerte leicht ablesen lassen. Welche Umformungen erlaubt sind, zeigt

Satz 12.4
Ist $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, so gilt für jede reguläre Matrix $\boldsymbol{T}$:

a) $\boldsymbol{A}$ und seine *Ähnlichkeitstransformierte* $\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$ haben dieselben Eigenwerte mit denselben algebraischen Vielfachheiten.

b) Ist $\boldsymbol{x}$ EV zum EW λ von A, so ist $\boldsymbol{T}^{-1}\boldsymbol{x}$ EV zum EW λ von $\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$, d.h. insbesondere: die geometrische Vielfachheit von λ bleibt bei Ähnlichkeitstransformationen erhalten.

Beweis:

a) Mit dem Determinantenmultiplikationssatz folgt

$$\begin{aligned}\det(\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T} - \lambda\boldsymbol{E}) &= \det\left\{\boldsymbol{T}^{-1}(\boldsymbol{A}\boldsymbol{T} - \lambda\boldsymbol{T})\right\}\\ &= \det\left\{\boldsymbol{T}^{-1}(\boldsymbol{A} - \lambda\boldsymbol{E})\boldsymbol{T}\right\}\\ &= \det\boldsymbol{T}^{-1}\det(\boldsymbol{A} - \lambda\boldsymbol{E})\det\boldsymbol{T}\\ &= \det(\boldsymbol{A} - \lambda\boldsymbol{E}).\end{aligned}$$

Beide Matrizen haben also das gleiche charakteristische Polynom.

b) Ist $\boldsymbol{x}$ EV zum EW λ, so gilt

$$\begin{aligned}\boldsymbol{A}\boldsymbol{x} &= \lambda\boldsymbol{x}\\ \boldsymbol{A}\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{x} &= \lambda\boldsymbol{x}\\ \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}(\boldsymbol{T}^{-1}\boldsymbol{x}) &= \lambda\boldsymbol{T}^{-1}\boldsymbol{x}.\end{aligned}$$

■

Interessant sind vor allem Matrizen, die sich *diagonalisieren* lassen (d.h. durch Ähnlichkeitstransformation auf Diagonalgestalt bringen lassen). Man nennt diese Matrizen auch *diagonalähnlich.*

Wie wir gleich zeigen werden, spannen die Eigenvektoren dieser Matrizen den ganzen Raum auf. Dies ist eine wesentliche Voraussetzung für die Verfahren der Vektoriteration (von–Mises–Verfahren und Wieland–Verfahren).

Im nächsten Satz stellen wir eine Reihe von Matrizen zusammen, deren Diagonalisierbarkeit aus der „Linearen Algebra“ bekannt ist. Wir zitieren dazu die entsprechenden Paragraphen aus Fischer: Lineare Algebra.

Satz 12.5
Folgende Matrizen sind (durch Ähnlichkeitstransformation) diagonalisierbar:

$$\left.\begin{array}{ll} \textit{unitäre}\text{ Matrizen :} & \boldsymbol{A}^{-1} = \overline{\boldsymbol{A}}^T \;=:\; \boldsymbol{A}^* \\ \textit{orthogonale}\text{ Matrizen (= reell unitär) :} & \boldsymbol{A}^{-1} = \boldsymbol{A}^T \end{array}\right\} \quad \text{§ 6.4.4, 6.1.7}$$

$$\left.\begin{array}{ll} \textit{hermite'sche}\text{ Matrizen :} & \boldsymbol{A} = \boldsymbol{A}^* \\ \textit{symmetrische reelle}\text{ Matrizen :} & \boldsymbol{A} = \boldsymbol{A}^T \end{array}\right\} \quad \text{§ 6.5.3}$$

Die Literaturangaben bezeichnen sich auf Fischer: Lineare Algebra.

Beachte:
Nicht alle Matrizen sind diagonalisierbar (vgl. dazu Jordan'sche Normalform).

Aussagen über die Eigenvektoren liefert

Satz 12.6

a) Sind $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m$ Eigenvektoren zu paarweise verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_m \in \sigma(\boldsymbol{A})$, $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, so sind $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m$ linear unabhängig.

b) Für die Eigenwerte diagonalisierbarer Matrizen sind geometrische und algebraische Vielfachheit gleich. ($\Rightarrow$ Die Eigenvektoren spannen den ganzen Raum auf.)

Beweis:

a) (vgl. Fischer: Lineare Algebra, Lemma 5.1.4)
Beweis durch vollständige Induktion: Der Fall $m = 1$ ist wegen $\boldsymbol{v}^1 \neq \boldsymbol{0}$ klar.
Sei $m \geq 2$ und die Aussage für $m - 1$ schon bewiesen.
Ist

$$\sum_{i=1}^{m} \alpha_i \, \boldsymbol{v}^i \quad = \quad \boldsymbol{0}\,, \tag{$*$}$$

so folgt durch Multiplikation dieser Gleichung mit λ_m, bzw. Anwendung von $\boldsymbol{A}$:

$$\begin{aligned} \boldsymbol{0} = \lambda_m \boldsymbol{0} &= \lambda_m \alpha_1 \boldsymbol{v}^1 + \ldots + \lambda_m \alpha_{m-1} \boldsymbol{v}^{m-1} + \lambda_m \alpha_m \boldsymbol{v}^m\,, \\ \boldsymbol{0} = \boldsymbol{A}\,\boldsymbol{0} &= \lambda_1 \alpha_1 \boldsymbol{v}^1 + \ldots + \lambda_{m-1} \alpha_{m-1} \boldsymbol{v}^{m-1} + \lambda_m \alpha_m \boldsymbol{v}^m\,, \end{aligned}$$

und durch Differenzbildung

$$\boldsymbol{0} \quad = \quad \sum_{i=1}^{m-1} \alpha_i \, (\lambda_m - \lambda_i)\, \boldsymbol{v}^i\,.$$

Da $\lambda_m \neq \lambda_i, i = 1, \ldots, m{-}1$, impliziert die lineare Unabhängigkeit der $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{m-1}$, daß $\alpha_i = 0$, $i = 1, \ldots, m - 1$. Damit liefert ($*$): $\alpha_m = 0$.

b) Ist $\boldsymbol{D}$ eine Diagonalmatrix, so stehen in der Diagonale die Eigenwerte:

$$\boldsymbol{D} \quad = \quad \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

und die Einheitsvektoren $\boldsymbol{e}^i$ $(i = 1, \ldots, n)$ sind Eigenvektoren zu den Eigenwerten λ_i, insbesondere sind also algebraische und geometrische Vielfachheit der λ_i gleich.

Ist $\boldsymbol{A}$ diagonalisierbar, so existiert eine reguläre Matrix $\boldsymbol{T}$, so daß

$$\boldsymbol{T}^{-1} \boldsymbol{A} \boldsymbol{T} \quad = \quad \boldsymbol{D}$$

eine Diagonalmatrix ist. Also ist $\boldsymbol{A} = \boldsymbol{T} \boldsymbol{D} \boldsymbol{T}^{-1}$ und mit Satz 12.4 b) folgt, daß $\boldsymbol{T} \boldsymbol{e}^i$, $i = 1, \ldots, n$ linear unabhängige Eigenvektoren zu den Eigenwerten λ_i von $\boldsymbol{A}$ sind. ■

Vektoriteration

Oft ist nur der betragsgrößte EW einer Matrix von Interesse (vgl. z.B. Satz 11.6). Diesen kann man oft durch ein einfaches Verfahren erhalten. Wir beschränken uns bei der Beschreibung auf reelle Matrizen und einfache, reelle, betragsgrößte EWe. Erweiterungen sind möglich (vgl. die Bemerkungen nach Satz 12.7), und wir schildern zunächst die

Direkte Vektoriteration (von–Mises–Verfahren)
Ist $\boldsymbol{A}$ eine Matrix, deren Eigenvektoren $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^n$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_n$ den ganzen Raum aufspannen, so kann man für einen Anfangsvektor

$$\boldsymbol{y}^0 \quad = \quad \sum_{j=1}^{n} \alpha_j \, \boldsymbol{v}^j$$

folgende Vektoriteration betrachten

$$\boldsymbol{y}^{k+1} \quad = \quad \boldsymbol{A} \, \boldsymbol{y}^k \quad = \quad \boldsymbol{A}^{k+1} \, \boldsymbol{y}^0 \, . \tag{12.25}$$

Wegen $\boldsymbol{A}^k \, \boldsymbol{v}^j = \boldsymbol{A}^{k-1}(\boldsymbol{A} \, v^j) = \boldsymbol{A}^{k-1} \lambda_j \, \boldsymbol{v}^j = \ldots = \lambda_j^k \, \boldsymbol{v}^j$, $j = 1, \ldots, n$ gilt

$$\begin{aligned} \boldsymbol{y}^{k+1} \quad &= \quad \boldsymbol{A}^{k+1} \, \boldsymbol{y}^0 \quad = \quad \boldsymbol{A}^{k+1} \sum_{j=1}^{n} \alpha_j \, \boldsymbol{v}^j \\ &= \quad \sum_{j=1}^{n} \alpha_j \, \lambda_j^{k+1} \, \boldsymbol{v}^j \quad \text{und falls} \quad \alpha_1 \neq 0 \\ &= \quad \lambda_1^{k+1} \left(\alpha_1 \, \boldsymbol{v}^1 \; + \underbrace{\sum_{j=2}^{n} \alpha_j \, \left(\frac{\lambda_j}{\lambda_1} \right)^{k+1} \boldsymbol{v}^j}_{\boldsymbol{r}(k)} \right) \end{aligned} \tag{12.26}$$

$$\lim_{k \to \infty} \boldsymbol{r}(k) \quad = \quad 0 \quad \text{falls} \quad \frac{|\lambda_j|}{|\lambda_1|} \; < \; 1 \quad \forall \, j \neq 1$$

d.h. der betragsgrößte Eigenwert λ_1 „setzt sich durch“ und die Iteration konvergiert „ggf. bis auf ein Vorzeichen (falls $\lambda_1 < 0$)“ gegen ein Vielfaches von $\boldsymbol{v}^1$.

Um zu verhindern, daß der Faktor $\alpha_1 \lambda_1^{k+1}$ zu stark anwächst (falls $|\lambda_1| > 1$) oder zu stark gegen Null geht (falls $|\lambda_1| < 1$), wird in (12.25) $\boldsymbol{y}^{k+1}$ nach jedem Iterationsschritt zu 1 normiert, d.h. wir berechnen eine Folge $\{x^k\}$ nach der Vorschrift des nächsten Satzes und erhalten unter Verwendung der Sammlung der bisherigen Voraussetzungen

Satz 12.7 (direkte Vektoriteration, von–Mises–Verfahren)

1) Sei $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ eine diagonalisierbare Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$ und den zugehörigen normierten Eigenvektoren $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^n$.

2) Der betragsgrößte Eigenwert λ_1 (nach entsprechender Numerierung) sei einfach, d.h. $|\lambda_1| > |\lambda_j| \ \forall\, j > 1$.

3) $\boldsymbol{y}^0 \in \mathbb{R}^n$ sei ein Vektor, in dessen Darstellung

$$\boldsymbol{y}^0 \quad = \quad \sum_{j=1}^{n} \alpha_j \, \boldsymbol{v}^j$$

der Koeffizient $\alpha_1 \neq 0$ ist.

Wir konstruieren eine Folge $\{\boldsymbol{x}^k\}$ gemäß

$$\boldsymbol{x}^0 \ = \ \frac{\boldsymbol{y}^0}{\|\boldsymbol{y}^0\|}\,, \qquad \boldsymbol{y}^{k+1} \ = \ \boldsymbol{A}\,\boldsymbol{x}^k\,, \qquad \boldsymbol{x}^{k+1} \ = \ \frac{\boldsymbol{y}^{k+1}}{\|\boldsymbol{y}^{k+1}\|}\,. \tag{12.27}$$

Dann gilt:

a) Die Folge $\{\boldsymbol{x}^k\}$ konvergiert (ggf. bis aufs Vorzeichen) gegen den zu 1 normierten Eigenvektor $\boldsymbol{v}^1$ von λ_1, genauer:

(12.28) $$\lim_{k\to\infty} (\operatorname{sgn}\lambda_1)^k\, \boldsymbol{x}^k \quad = \quad (\operatorname{sgn}\alpha_1)\, \boldsymbol{v}^1\,.$$

b) λ_1 kann man erhalten als Grenzwert des *Rayleigh–Quotienten* $R(\boldsymbol{x}^k)$

Mit $\quad R(\boldsymbol{x}^k) \ := \ \frac{(\boldsymbol{x}^k)^T \boldsymbol{A}\, \boldsymbol{x}^k}{(\boldsymbol{x}^k)^T\, \boldsymbol{x}^k} \ =: \ \lambda_1^{(k)} \quad$ folgt

(12.29) $$\lim_{k\to\infty} \lambda^{(k)} = \lim_{k\to\infty} \frac{(\boldsymbol{x}^k)^T\, \boldsymbol{A}\, \boldsymbol{x}^k}{(\boldsymbol{x}^k)^T\, \boldsymbol{x}^k} \quad = \quad \lim_{k\to\infty} R(\boldsymbol{x}^k) \quad = \quad \lambda_1\,.$$

Beweis:

Beachte zunächst: Der betragsgrößte Eigenwert λ_1 ist einfach und daher reell (wegen (12.24)). Also hat λ_1 auch einen reellen Eigenvektor.

a) Durch vollständige Induktion erhalten wir sofort

$$(12.31) \qquad \boldsymbol{x}^k = \frac{\boldsymbol{A}^k\,\boldsymbol{y}^0}{\|\boldsymbol{A}^k\,\boldsymbol{y}^0\|}\,, \qquad k = 0,1,2,\ldots$$

Hierbei zeigen wir nur den Induktionsschritt:

$$\boldsymbol{x}^{k+1} = \frac{\boldsymbol{y}^{k+1}}{\|\boldsymbol{y}^{k+1}\|} = \frac{\boldsymbol{A}\,\boldsymbol{x}^k}{\|\boldsymbol{A}\,\boldsymbol{x}^k\|} \overset{\text{Ind.-vor.}}{=} \left(\frac{\boldsymbol{A}^{k+1}\boldsymbol{y}^0}{\|\boldsymbol{A}^k\boldsymbol{y}^0\|}\right) \Big/ \left\|\frac{\boldsymbol{A}^{k+1}\boldsymbol{y}^0}{\|\boldsymbol{A}^k\boldsymbol{y}^0\|}\right\| = \frac{\boldsymbol{A}^{k+1}\boldsymbol{y}^0}{\|\boldsymbol{A}^{k+1}\boldsymbol{y}^0\|}\,.$$

Aus (12.31) folgt mit (12.26) sofort

$$\lim_{k\to\infty} (\operatorname{sgn}\lambda_1)^k\,\boldsymbol{x}^k = \frac{\alpha_1|\lambda_1|^k\,\boldsymbol{v}^1}{\|\alpha_1\,|\lambda_1|^k\,\boldsymbol{v}^1\|} = (\operatorname{sgn}\alpha_1)\,\boldsymbol{v}^1\,.$$

b) Ist $\boldsymbol{x}$ Eigenvektor von $\boldsymbol{A}$ zu λ, so ist $\boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}$, $\boldsymbol{x}^T\,\boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}^T\,\boldsymbol{x}$, also

$$R(\boldsymbol{x}) := \frac{\boldsymbol{x}^T\,\boldsymbol{A}\,\boldsymbol{x}}{\boldsymbol{x}^T\,\boldsymbol{x}} = \lambda\,.$$

$R(\boldsymbol{x})$ ist eine stetige Funktion von $\boldsymbol{x}$. Also gilt

$$\lim_{k\to\infty} R(\boldsymbol{x}^k) = R(\boldsymbol{v}^1) = \lambda_1\,.$$

Beachte: $\boldsymbol{x}^k \to \boldsymbol{v}^1$ gilt laut (12.28) nur bis auf einen Vorzeichenfaktor. Dies ist jedoch unerheblich, da $R(\boldsymbol{x})$ unabhängig vom Vorzeichen von $\boldsymbol{x}$ ist. ∎

Bemerkungen

a) Üblicherweise wählt man $\boldsymbol{y}^0 = (1,1,\ldots,1)^T$ in der Hoffnung, $\alpha_1 \neq 0$ zu haben. Sollte dies nicht der Fall sein (was man ja nicht weiß), so wird gemäß der aufgezeigten Theorie das Verfahren zunächst gegen den „betragszweit"größten EW konvergieren, sofern dieser einfach ist. Auf Grund der Rundungsfehler wird sich jedoch in $\boldsymbol{x}^k$ sehr schnell eine Komponente $\neq 0$ in Richtung $\boldsymbol{v}^1$ einschleichen, und das Verfahren konvergiert schließlich doch gegen $\boldsymbol{v}^1$ (ausprobieren!).

b) Das Verfahren kann auf komplexe Matrizen und mehrfache betragsgrößte Eigenwerte erweitert werden (vgl. etwa Burg/Haf/Wille: Höhere Mathematik für Ingenieure II, § 3.7.10).

c) Es kann sogar auf nichtdiagonalisierbare Matrizen erweitert werden, sofern der dominante (betragsgrößte) EW einfach ist (vgl. Bemerkung in Stoer/Bulirsch, § 6.6.3).

d) Ein Vorteil des Verfahrens ist seine einfache Durchführbarkeit.

e) Die Konvergenz wird (offensichtlich) bestimmt durch den Konvergenzfaktor $\max\limits_{i\neq 1}\left|\frac{\lambda_i}{\lambda_1}\right|$. Ist λ_1 nur schwach dominant, so ist die Konvergenz sehr langsam. Nachteilig ist, daß auch nur λ_1 berechnet werden kann.

Für die Praxis wichtiger, weil diese Nachteile vermieden werden, ist daher die

Inverse Vektoriteration (Wieland 1945)

Ihr Name rührt daher, daß sie formal nichts anderes ist als die Anwendung der direkten Matrixiteration auf eine inverse Matrix.

Angenommen für einen beliebigen einfachen Eigenwert λ_i einer diagonalähnlichen Matrix $\boldsymbol{A}$ sei eine Schätzung $\tilde{\lambda}$ $(\approx \lambda_i)$ bekannt, die so gut ist, daß

$$|\tilde{\lambda} - \lambda_i| < |\tilde{\lambda} - \lambda_j| \;\forall j \neq i \quad \text{bzw.} \quad \frac{1}{|\tilde{\lambda} - \lambda_i|} > \frac{1}{|\tilde{\lambda} - \lambda_j|} \;\forall j \neq i\,. \tag{12.34}$$

Ist $\tilde{\lambda}$ kein Eigenwert von $\boldsymbol{A}$, so existiert $(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})^{-1}$, und es sind äquivalent:

$$\mu_i = \frac{1}{\lambda_i - \tilde{\lambda}} \text{ ist EW von } (\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})^{-1} \text{ zum EV } \boldsymbol{v}^i \iff \lambda_i \text{ ist EW von } \boldsymbol{A} \text{ zum EV } \boldsymbol{v}^i\,.$$

Der Beweis folgt aus den Äquivalenzen

$$\begin{aligned}
(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})^{-1}\,\boldsymbol{v}^i &= \frac{1}{\lambda_i - \tilde{\lambda}}\,\boldsymbol{v}^i \\
\boldsymbol{v}^i &= \frac{1}{\lambda_i - \tilde{\lambda}}\,(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})\,\boldsymbol{v}^i \\
(\lambda_i - \tilde{\lambda})\,\boldsymbol{v}^i &= (\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})\,\boldsymbol{v}^i \\
\lambda_i\,\boldsymbol{v}^i &= \boldsymbol{A}\,\boldsymbol{v}^i\,.
\end{aligned}$$

Gemäß (12.34) ist also $\frac{1}{\lambda_i - \tilde{\lambda}}$ der betragsgrößte EW von $(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})^{-1}$. Man kann also die direkte Matrixiteration mit der inversen Matrix $(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})^{-1}$ durchführen:

$$\boldsymbol{y}^{k+1} = (\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})^{-1}\,\boldsymbol{x}^k \qquad \text{(vgl. (12.27))}.$$

Praktisch wird natürlich nicht die inverse Matrix berechnet, sondern in jedem Iterationsschritt das lineare Gleichungssystem (vgl. (12.27)) gelöst

$$(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})\,\boldsymbol{y}^{k+1} = \boldsymbol{x}^k\,, \qquad \boldsymbol{x}^0 = \frac{\boldsymbol{y}^0}{\|\boldsymbol{y}^0\|}\,, \qquad \boldsymbol{x}^{k+1} = \frac{\boldsymbol{y}^{k+1}}{\|\boldsymbol{y}^{k+1}\|}\,.$$

Die Koeffizientenmatrix ist immer die gleiche. Abgesehen vom 1. Schritt reduziert sich in jedem Folgeschritt die Arbeit auf „Rückwärtseinsetzen“. Es bleibt dem Leser überlassen, den zu Satz 12.7 analogen Satz sowie die zugehörigen Bemerkungen für die inverse Vektoriteration zu formulieren.

Zusätzliche Bemerkungen

1) Ist $\tilde{\lambda}$ eine sehr gute Schätzung für λ_i, so gilt

$$\frac{|\lambda_i - \tilde{\lambda}|}{|\lambda_j - \tilde{\lambda}|} \ll 1 \quad \forall\, j \neq i$$

und das Verfahren konvergiert sehr schnell.

2) Durch geeignete Wahl von $\tilde{\lambda}$ kann man bei beliebigem Startvektor $\boldsymbol{y}^0$ einzelne EWe und EVen herausgreifen.

3) Die Matrix $(\boldsymbol{A} - \tilde{\lambda}\,\boldsymbol{E})$ ist bei gut gewähltem $\tilde{\lambda} = \lambda_i$ fast singulär. Allgemein bedeutet das eine schlechte Kondition. Im vorliegenden Fall entstehen daraus jedoch keine Komplikationen, da wir „nur die Richtung" des Eigenvektors suchen. Diese spezielle Aufgabe ist gut konditioniert (vgl. Deuflhardt/Hohmann: Numerische Mathematik, de Gruyter, Bemerkung 5.6).

4) Ist 0 kein EW der Matrix $\boldsymbol{A}$, so kann man durch inverse Iteration den betragskleinsten Eigenwert berechnen. Dieser ist in vielen Anwendungen der Technik von Bedeutung, da er bei Schwingungsproblemen oft die Grundschwingung (Schwingung mit der kleinsten Frequenz) beschreibt (vgl. dazu auch unser einleitendes Beispiel zu Beginn des §).

Wir beschreiben abschließend das heute am meisten angewandte Verfahren zur numerischen Berechnung aller Eigenwerte von Matrizen.

Der QR–Algorithmus

(zur praktischen Berechnung von EWen)

Die Methode ist zu komplex, um sie in allen Einzelheiten zu schildern. Deshalb müssen wir uns hier auf die Grundidee beschränken.

Aufgabe: Bestimme alle EWe einer Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

Schritt 1 Man transformiert $\boldsymbol{A}$ durch eine Transformation, welche die EWe nicht ändert (also eine Ähnlichkeitstransformation: $\tilde{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$), in eine „einfachere“ Form, die sog. *Hessenberg–Form*.

Definition 12.8

Sei $\boldsymbol{A} = (a_{ik}) \in \mathbb{R}^{n\times n}$:

$\boldsymbol{A}$ heißt *Hessenberg–Matrix* $\iff$ $a_{ik} = 0 \ \forall\, i > k+1$

$\boldsymbol{A}$ heißt *tridiagonal* $\iff$ $a_{ik} = 0 \ \forall\, k+1 < i$ und $i < k-1$

Hessenberg–Form:

$$\widetilde{\boldsymbol{A}} \quad := \quad \begin{pmatrix} * & \dots & \dots & \dots & * \\ * & \ddots & & & \vdots \\ 0 & * & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & * & * \end{pmatrix}$$

Tridiagonale Gestalt:

$$\widetilde{\boldsymbol{A}} \quad := \quad \begin{pmatrix} * & * & 0 & \dots & 0 \\ * & * & * & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & * \\ 0 & \dots & 0 & * & * \end{pmatrix}$$

Satz 12.9

Jedes $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ läßt sich (z.B. durch Householder–Transformationen) in $n-2$ Schritten auf Hessenberg–Form transformieren; ist $\boldsymbol{A}$ symmetrisch, sogar auf Tridiagonalform.

Beweis:
Unter Verwendung der Aussage (vgl. Lemma 9.5)

> **Lemma 9.5"**
> Zu $\boldsymbol{z} \in \mathbb{R}^r$, $\boldsymbol{z} \neq 0$ wähle
>
> $$\boldsymbol{v} \;:=\; \frac{\boldsymbol{z} - \alpha\, \boldsymbol{e}^1}{\|\boldsymbol{z} - \alpha\, \boldsymbol{e}^1\|_2}\,, \qquad \alpha \;=\; -\mathrm{sgn}(z_1)\, \|\boldsymbol{z}\|_2\,.$$
>
> Dann gilt
>
> $$\boldsymbol{H}(\boldsymbol{z}) \quad = \quad (\boldsymbol{E} - 2\boldsymbol{v}\,\boldsymbol{v}^T)\,\boldsymbol{z} \quad = \quad \alpha\,\boldsymbol{e}^1\,.$$

beschreiben wir das Transformationsverfahren, das $\boldsymbol{A}$ schrittweise in Hessenbergform, bzw. Tridiagonalform überführt.

$$\begin{aligned} \boldsymbol{A}_1 &:= \boldsymbol{A} \\ \boldsymbol{A}_j &:= \boldsymbol{H}_j\,\boldsymbol{A}_{j-1}\,\boldsymbol{H}_j^T\,, \qquad j = 2, \ldots, n-1\,. \end{aligned}$$

$$\boldsymbol{H}_j \;:=\; \left(\begin{array}{ccc|c} 1 & & & \\ & \ddots & & \mathbf{O} \\ & & 1 & \\ \hline & \mathbf{O} & & \boldsymbol{T}_j \end{array}\right), \qquad \begin{array}{l} \boldsymbol{T}_j \in \mathbb{R}^{(n-j+1)\times(n-j+1)}\,, \\ \boldsymbol{T}_j \;=\; \boldsymbol{E}_{n-j+1} - 2\boldsymbol{v}^j\,\boldsymbol{v}^{j^T} \end{array}$$

Wie die Vektoren $\boldsymbol{v}^j$ gewählt werden müssen, geht aus der folgenden Beschreibung des Verfahrens hervor. Die $\boldsymbol{H}_j$ sind Householdermatrizen (vgl. Definition und Lemma 9.4) also orthogonal (vgl. (9.12)). Daher gehen die $\boldsymbol{A}_j$ aus $\boldsymbol{A}_1$ durch Ähnlichkeitstransformationen hervor, welche die Eigenwerte unverändert lassen (vgl. Satz 12.6).

Wir geben nun den j–ten Schritt des Verfahrens an. Er zeigt, wie und warum das Verfahren funktioniert und wie die Transformationsmatrizen $\boldsymbol{H}_j$ und die Vektoren $\boldsymbol{v}^j$ gewählt werden müssen. Dabei beschränken wir uns zunächst auf den Fall „$\boldsymbol{A}$ symmetrisch“ und zeichnen einen Transformationsschritt auf.

Empfehlung:
Zur Verständniserleichterung der Transformationsformeln der nächsten Seite schreibe man sich die Transformation für den 1. Schritt $(j = 2)$ auf. Die erste Spalte $\boldsymbol{a}^1$ von $\boldsymbol{A}_1 := \boldsymbol{A}$ hat dann die Bezeichnung

$$\boldsymbol{a}^1 \;=\; \begin{pmatrix} \delta_1 \\ \boldsymbol{z} \end{pmatrix} \quad \text{mit dem Vektor} \quad \boldsymbol{z} \;=\; (\boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)^T\,,$$

und die Matrix $\boldsymbol{H}_2$ wirkt bei Multiplikation von links mittels der $(n-1)\times(n-1)$–Matrix T_2 in der 1. Spalte $\boldsymbol{a}^1$ von $\boldsymbol{A}$ nur auf den Teilvektor $(z_2, z_3, \ldots, z_n)^T$, den sie auf ein Vielfaches von $\boldsymbol{e}^1 \in \mathbb{R}^{n-1}$ abbildet.

$$
\begin{array}{r}
\\ j-2 \left\{ \begin{array}{c} \\ \\ \\ \end{array} \right. \\ j-1 \rightarrow \\ n-j+1 \left\{ \begin{array}{c} \\ \\ \\ \end{array} \right.
\end{array}
\overset{\boldsymbol{H}_j}{\left(\begin{array}{cc|c|c}
1 & \mathbf{0} & 0 & \\
 & \ddots & \vdots & \mathbf{0} \\
\mathbf{0} & 1 & 0 & \\ \hline
0 \;\dots & 0 & 1 & 0 \;\dots\; 0 \\ \hline
 & & 0 & \\
 & \mathbf{0} & \vdots & \boldsymbol{T}_j \\
 & & 0 &
\end{array}\right)}
\overset{\boldsymbol{A}_{j-1}}{\left(\begin{array}{ccc|c|c}
\delta_1 & \gamma_2 & & & \\
\gamma_2 & \ddots & \ddots & & \mathbf{0} \\
\mathbf{0} & \ddots & \delta_{j-2} & \gamma_{j-1} & \\ \hline
0 \;\dots & 0 & \gamma_{j-1} & \delta_{j-1} & \boldsymbol{z}^T \\ \hline
 & \mathbf{0} & & \boldsymbol{z} & \widetilde{\boldsymbol{A}}_{j-1}
\end{array}\right)}
\overset{\boldsymbol{H}_j^T}{\left(\begin{array}{cc|c|c}
1 & \mathbf{0} & 0 & \\
 & \ddots & \vdots & \mathbf{0} \\
\mathbf{0} & 1 & 0 & \\ \hline
0 \;\dots & 0 & 1 & 0 \;\dots\; 0 \\ \hline
 & & 0 & \\
 & \mathbf{0} & \vdots & \boldsymbol{T}_j^T \\
 & & 0 &
\end{array}\right)} =
$$

$$
= \left(\begin{array}{ccc|c|c}
\delta_1 & \gamma_2 & & & \\
\gamma_2 & \ddots & \ddots & & \mathbf{0} \\
 & \ddots & \delta_{j-2} & \gamma_{j-1} & \\ \hline
0 \;\dots & 0 & \gamma_{j-1} & \delta_{j-1} & \boldsymbol{z}^T \, \boldsymbol{T}_j^T \\ \hline
 & \mathbf{0} & & \boldsymbol{T}_j \, \boldsymbol{z} & \boldsymbol{T}_j \, \widetilde{\boldsymbol{A}}_{j-1} \, \boldsymbol{T}_j^T
\end{array}\right) =: \boldsymbol{A}_j
$$

Dabei ist $\boldsymbol{T}_j$ eine Householdermatrix mit

$$
\boldsymbol{T}_j \, \boldsymbol{z} \quad = \quad \left(\boldsymbol{E}_{n-j+1} - 2\boldsymbol{v}^j \, \boldsymbol{v}^{j^T} \right) \boldsymbol{z} \quad = \quad \begin{pmatrix} \gamma_j & 0 \\ \vdots & 0 \end{pmatrix}
$$

$\boldsymbol{v}^j$ wird gemäß Lemma 9.5' aus $\boldsymbol{z}$ berechnet.

Damit ist die Tridiagonalisierung einen Schritt weiter. Man erkennt, daß wenn $\boldsymbol{A}_{j-1}$ symmetrisch ist, auch $\boldsymbol{A}_j$ symmetrisch ist. Ist $\boldsymbol{A}$ nicht symmetrisch (und damit $\boldsymbol{A}_{j-1}$ nicht symmetrisch), so wird dennoch $\boldsymbol{T}_j$ gleich gewählt, jedoch steht dann in $\boldsymbol{A}_{j-1}$ in der „2. Hälfte“ der $(j-1)$–ten Zeile nicht $\boldsymbol{z}^T$, sondern eine Zeile $\tilde{\boldsymbol{z}}^T$. Damit steht in der „2. Hälfte“ von $\boldsymbol{A}_j$ in der $(j-1)$–ten Zeile statt $\boldsymbol{z}^T \, \boldsymbol{T}_j^T$ der Ausdruck $\tilde{\boldsymbol{z}}^T \, \boldsymbol{T}_j^T$, d.h. auch $\boldsymbol{A}_j$ ist nicht symmetrisch, aber man ist einen Schritt weiter in der Umformung zur Hessenberg–Gestalt.

Bemerkung

Man kann das Verfahren auch auf komplexe Matrizen erweitern. Statt „symmetrisch“ muß man dann hermitesch voraussetzen, das komplexe Skalarprodukt benutzen und $\boldsymbol{H} = \boldsymbol{E} - 2\boldsymbol{v} \, \boldsymbol{v}^*$ setzen.

Schritt 2 des Verfahrens — der eigentliche QR–Algorithmus — wird auf eine Hessenberg–Matrix angewandt (also nach Durchführung von Schritt 1).

Wir wissen bereits nach Satz 9.6 und Folgerung 9.7: Zu jeder Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ existiert eine

QR–Zerlegung:

$$\boldsymbol{A} \quad = \quad \boldsymbol{Q}\,\boldsymbol{R}$$

mit einer orthogonalen Matrix $\boldsymbol{Q}$ (z.B. Produkt von Householder–Matrizen oder von sog. Givens-Rotationen – letztere sind angesichts der speziellen Struktur einer Hessenberg-Matrix günstiger) und einer oberen Dreiecksmatrix $\boldsymbol{R}$.

Liegt eine QR–Zerlegung von A vor, so definieren wir die

QR–Transformation:

$$\boldsymbol{A} \;=\; \boldsymbol{Q}\,\boldsymbol{R} \quad \longrightarrow \quad \boldsymbol{A}' \;=\; \boldsymbol{R}\,\boldsymbol{Q}\,.$$

Die QR–Transformation ist eine orthogonale Ähnlichkeitstransformation, denn
$\boldsymbol{A} = \boldsymbol{Q}\,\boldsymbol{R} \Rightarrow \boldsymbol{R} = \boldsymbol{Q}^{-1}\,\boldsymbol{A}$ ($\boldsymbol{Q}^{-1}$ existiert, da $\boldsymbol{Q}$ orthogonal ist), also

$$\boldsymbol{A}' \quad = \quad \boldsymbol{R}\boldsymbol{Q} \quad = \quad \boldsymbol{Q}^{-1}\,\boldsymbol{A}\,\boldsymbol{Q}\,.$$

Also hat $\boldsymbol{A}'$ dieselben EWe wie $\boldsymbol{A}$.

Man kann nun ausrechnen:

$$\begin{array}{llcll} \boldsymbol{A} & \text{hat Hessenberg–Form} & \Rightarrow & \boldsymbol{A}' & \text{hat Hessenberg–Form} \\ \boldsymbol{A} & \text{tridiagonal} & \Rightarrow & \boldsymbol{A}' & \text{tridiagonal}\,. \end{array}$$

(Zu dieser Rechnerei muß $\boldsymbol{Q}$ als Produkt der obigen Housholder–Matrizen dargestellt werden, die sich aus den Spalten von $\boldsymbol{A}$ berechnen.)

Man geht nun aus von $\boldsymbol{A}$ (in Hessenberg–Form) und erzeugt mittels QR–Transformationen eine Folge von zu $\boldsymbol{A}$ orthogonal ähnlichen Matrizen

$$\boldsymbol{A} \quad =: \quad \boldsymbol{A}_0 \;\rightarrow\; \boldsymbol{A}_1 \;\rightarrow\; \boldsymbol{A}_2 \;\rightarrow\; \ldots$$

Ist $\boldsymbol{A}$ reellwertig, so konvergiert diese Folge gegen eine rechte „Quasidreiecksmatrix“

$$\boldsymbol{R} \quad := \quad \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \ldots & \ldots & \boldsymbol{R}_{1m} \\ 0 & \boldsymbol{R}_{22} & \ldots & \ldots & \boldsymbol{R}_{2m} \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & 0 & & 0 & \boldsymbol{R}_{mm} \end{pmatrix}.$$

Die Matrizen $\boldsymbol{R}_{ii}\,(i = 1,\ldots,m)$ sind entweder 1×1–Matrizen oder 2×2–Matrizen, denen man sofort die EWe von $\boldsymbol{R}$ entnehmen kann, da $\det(\boldsymbol{R} - \lambda\,\boldsymbol{E}) = \prod_{i=1}^{m} \det(\boldsymbol{R}_{ii} - \lambda\,\boldsymbol{E}_{ii})$, wobei $\boldsymbol{E}_{ii}$ eine Einheitsmatrix derselben Dimension wie $\boldsymbol{R}_{ii}$ ist.

Bemerkungen:

Da man das Verfahren nach endlich vielen Schritten abbricht, erhält man nur Approximationen für die EWe, die man ggf. mit der inversen Vektoriteration verbessern kann.

Für die praktische Durchführung sind noch einige Zusatzüberlegungen notwendig (die wir hier nicht behandeln können).

Algol–Programme für das Verfahren findet man in Wilkinson/Reinsch: Handbook for automatic computation II, 1971. Ausgehend von den in diesem Band zusammengestellten Algorithmen, ist das Fortran–Programmpaket `Eispack` entwickelt worden. Inzwischen gibt es auch PC–geeignete Ableger in Pascal (z.B. in Mathpak). In Matlab und Mathematica ist das QR–Verfahren ebenfalls implementiert.