Mathematik IV für Elektrotechnik
Mathematik III für Informatik

Vorlesungsskriptum

Stefan Ulbrich

Fachbereich Mathematik
Technische Universität Darmstadt

Sommersemester 2011

Inhaltsverzeichnis

Numerische Mathematik

Kapitel 1

Einführung

Viele Problemstellungen aus den Ingenieur- und Naturwissenschaften lassen sich durch mathematische Modelle beschreiben, in denen häufig lineare oder nichtlineare Gleichungssysteme, Integrale, Eigenwertprobleme, gewöhnliche oder partielle Differentialgleichungen auftreten. In nahezu allen praxisrelevanten Fällen läßt das mathematische Modell keine analytische Lösung zu. Vielmehr muss die Lösung durch geeignete Verfahren auf einem Rechner näherungsweise bestimmt werden. Hierbei ist es wichtig, dass das verwendete Verfahren robust, genau und möglichst schnell ist. Die Entwicklung derartiger Verfahren ist Gegenstand der Numerischen Mathematik, einem inzwischen sehr bedeutenden Gebiet der Angewandten Mathematik. Die Numerische Mathematik entwickelt effiziente rechnergestützte Verfahren zur Lösung mathematischer Problemstellungen, unter anderem der oben genannten. Die Vorlesung gibt eine Einführung in die numerische Behandlung der folgenden Problemstellungen

- Interpolation
- Lineare Gleichungssysteme
- Nichtlineare Gleichungssysteme
- Eigenwertprobleme
- Numerische Integration
- Anfangswertprobleme für gewöhnliche Differentialgleichungen
- Partielle Differentialgleichungen (gegebenenfalls ganz kurz)

Kapitel 2

Interpolation

Häufig liegen von einem funktionalen Zusammenhang $y = f(x)$, $f : [a, b] \to \mathbb{R}$ nur eine begrenzte Zahl von Werten $y_i = f(x_i)$, $i = 0, \ldots, n$, vor, man möchte jedoch $f(x)$ für beliebiges $x \in [a, b]$ näherungsweise berechnen, plotten, etc.. Dies führt auf das

Interpolationsproblem: Suche eine einfache Ersatzfunktion $\Phi(x)$ mit

$$\Phi(x_i) = y_i, \quad i = 0, \ldots, n.$$

Wunsch: Der Fehler $f(x) - \Phi(x)$ sollte auf $[a, b]$ klein sein.

Beispiele:

1. Die Funktion $f(x)$ ist aufwändig zu berechnen (z.B. $\sin(x)$, $\exp(x)$, $\ln(x)$, $\Gamma(x)$, etc.) und es sind nur die Werte $y_i = f(x_i)$, $i = 0, \ldots, n$, bekannt.
   Gesucht: Genaue Approximation $\Phi(x)$ für $f(x)$, oder $\Phi'(x)$ für $f'(x)$.

2. Ein Experiment (oder eine numerische Berechnung) beschreibt einen unbekannten funktionalen Zusammenhang $y = f(x)$ und liefert zu Eingangsparametern x_i die Werte y_i.
   Gesucht: Gutes Modell $\Phi(x)$ für das unbekannte $f(x)$.

3. Ein digitales Audiosignal (CD, MP3-Player, DVD, ...) liefert zum Zeitpunkt t_i, $i = 0, \ldots, n$, die Amplitude y_i.
   Gesucht: Wie sieht das zugehörige analoge Audiosignal $y(t)$ aus?

4. Ein digitales Audiosignal (t_i, y_i), $i = 0, \ldots, n$, zur Abtastrate 44,1 kHz (CD) soll umgesampelt werden auf die Abtastrate 48 kHz (DAT, DVD-Video).
   Gesucht: $(\tilde{t}_j, y(\tilde{t}_j))$ für die 48 kHz-Abtastzeiten $\tilde{t}_j$.

5. 2D-Beispiel: Durch Datenpunkte (x_i, y_i, z_i) soll eine glatte Fläche $(x, y, z(x, y))$ gelegt werden (CAD, Computergrafik, Laserscanner, etc.).

Formale Aufgabenstellung

Gegeben sei eine Ansatzfunktion $\Phi(x; a_0, \ldots, a_n)$, $x \in \mathbb{R}$, die von Parametern $a_0, \ldots, a_n \in \mathbb{R}$ abhängt. In diesem Kapitel beschäftigen wir uns mit der folgenden

Interpolationsaufgabe: Zu gegebenen Paaren

$$(x_i, y_i), \quad i = 0, \ldots, n \ \text{ mit } x_i, y_i \in \mathbb{R}, \quad x_i \neq x_j \text{ für } i \neq j$$

sollen die Parameter $a_0, \ldots, a_n$ so bestimmt werden, dass die Interpolationsbedingungen gelten

$$\Phi(x_i; a_0, \ldots, a_n) = y_i, \quad i = 0, \ldots, n.$$

Die Paare (x_i, y_i) werden als *Stützpunkte* bezeichnet.

2.1 Polynominterpolation

Sehr verbreitet ist die Polynominterpolation. Hier verwendet man als Ansatzfunktion Polynome vom Grad $\leq n$, also

$$p_n(x) = \Phi(x; a_0, \ldots, a_n) = a_0 + a_1 x + \ldots + a_n x^n.$$

Die Interpolationsaufgabe lautet dann: Finde ein Polynom $p_n(x)$ vom Grad $\leq n$, das die Interpolationsbedingungen erfüllt

$$p_n(x_i) = y_i, \quad i = 0, \ldots, n. \tag{2.1}$$

Naiver Lösungsansatz: Ein naheliegender, aber in der Praxis untauglicher Ansatz ist folgender: (2.1) liefert die $n+1$ linearen Gleichungen

$$a_0 + x_i a_1 + x_i^2 a_2 + \ldots + x_i^n a_n = y_i, \quad i = 0, \ldots, n,$$

für die $n+1$ Koeffizienten $a_0, \ldots, a_n$. In Matrixform lautet das Gleichungssystem

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdot & x_0^n \\ 1 & x_1 & x_1^2 & \cdot & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdot & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{2.2}$$

Gründe für Unbrauchbarkeit des Verfahrens:

- Das Auflösen des Gleichungssystems (2.2) ist mit $O(n^3)$ elementaren Rechenoperationen im Vergleich zu den nachfolgenden $O(n^2)$-Verfahren sehr teuer.

- Die Koeffizientenmatrix in (2.2) (Vandermonde-Matrix) ist zwar invertierbar, aber für größere n *extrem schlecht konditioniert*. Daher kann das Gleichungssystem (2.2) auf einem Computer nicht genau gelöst werden, da Rundungsfehler wegen der schlechten Kondition dramatisch verstärkt werden (siehe Kapitel 3).

2.1.1 Interpolationsformel von Lagrange

Als numerisch stabile und effiziente Lösung der Interpolationsaufgabe bietet sich folgendes Vorgehen an: Wir betrachten das

Lagrangesche Interpolationspolynom

$$p_n(x) = \sum_{k=0}^{n} y_k L_{k,n}(x) \quad \text{mit} \quad L_{k,n}(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}. \tag{2.3}$$

Die Lagrange-Polynome sind gerade so gewählt, dass gilt

$$L_{k,n}(x_i) = \begin{cases} 1 & \text{falls } k = i, \\ 0 & \text{sonst.} \end{cases} =: \delta_{ki}.$$

δ_{ki} ist das Kronecker-Symbol. p_n in (2.3) erfüllt die Interpolationsbedingungen (2.1), denn

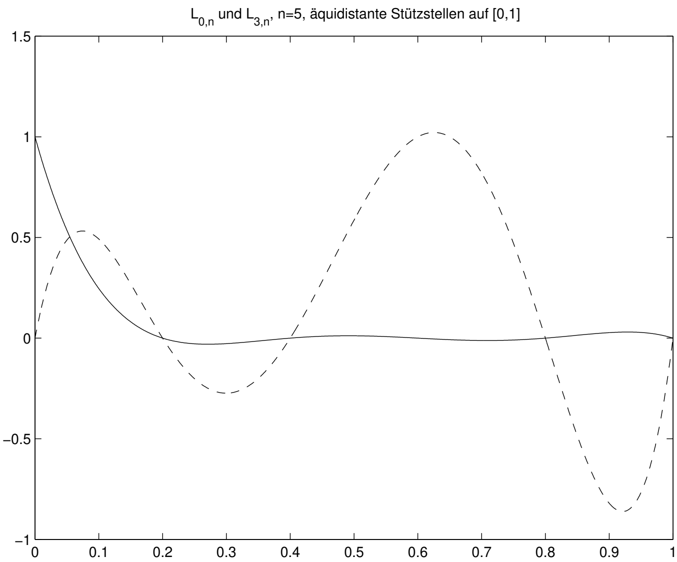


Abbildung 2.1: $L_{0,5}$ und $L_{3,5}$ für äquidistante Stützstellen auf $[0, 1]$.

$$p_n(x_i) = \sum_{k=0}^{n} y_k L_{k,n}(x_i) = \sum_{k=0}^{n} y_k \delta_{ki} = y_i.$$

Tatsächlich ist dies die einzige Lösung der Interpolationsaufgabe:

Satz 2.1.1 *Es gibt genau ein Polynom $p_n(x)$ vom Grad $\leq n$, das die Interpolationsbedingungen* (2.1) *erfüllt, nämlich* (2.3).

Beweis: Das Polynom (2.3) hat Grad $\leq n$ und erfüllt (2.1). Gäbe es eine weitere Lösung $\tilde{p}_n(x)$, dann ist $p_n(x) - \tilde{p}_n(x)$ ein Polynom vom Grad $\leq n$ mit $n+1$ verschiedenen Nullstellen $x_0, \ldots, x_n$, muss also identisch 0 sein. $\square$

Bemerkung: (2.3) zeigt, dass p_n linear von y_k abhängt. $\square$

Die Darstellung (2.3) von Lagrange ist für theoretische Zwecke sehr nützlich und wird auch in der Praxis oft angewendet.

Vorteile:

- Der Rechenaufwand beträgt:
  Koeffizientenberechnung (Nenner in (2.3)): $O(n^2)$
  Auswertung von $p_n(x)$: $O(n)$

- Intuitive, bequeme Darstellung.

Beispiel: Polynominterpolant von $f(x) = \sin(\pi x)$ auf $[0, 2]$ für $n = 5$ und äquidistante Stützstellen $x_i = \frac{2i}{5}$.

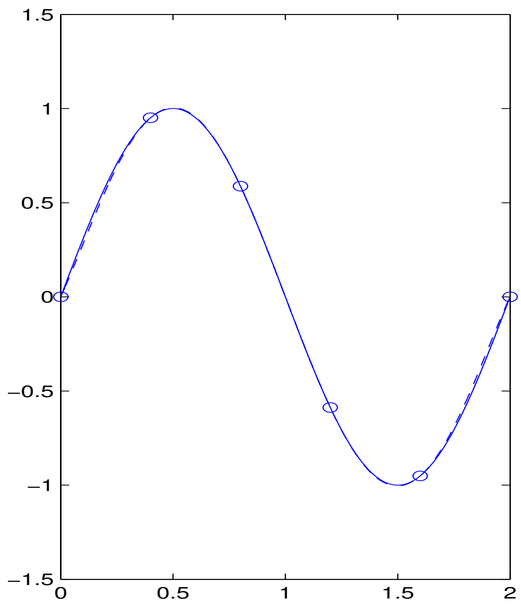
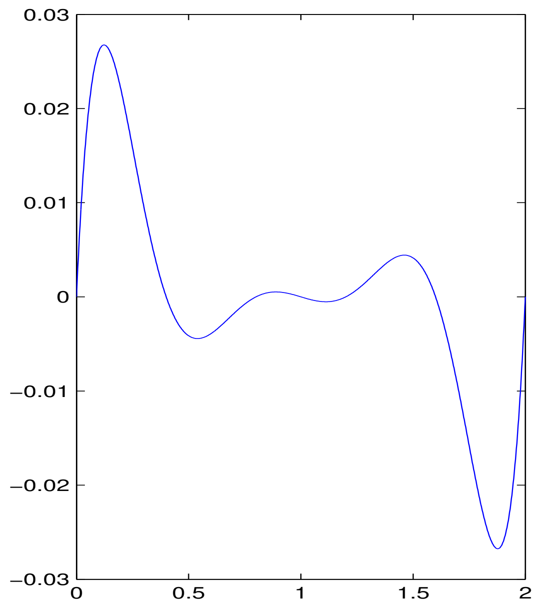


Abbildung 2.2: $\sin(\pi x)$ und $p_5(x)$ (gestrichelt)



Fehler $\sin(\pi x) - p_5(x)$.

In der Praxis, insbesondere wenn die effiziente Hinzunahme weiterer Stützstellen möglich sein soll, ist die folgende *Newtonsche Interpolationsformel* angenehmer.

2.1.2 Newtonsche Interpolationsformel

Wir wählen als Ansatz die *Newtonsche Darstellung*

$$p_n(x) = \gamma_0 + \gamma_1(x - x_0) + \gamma_2(x - x_0)(x - x_1) + \ldots + \gamma_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}).$$

Einsetzen in (2.1) liefert nun

$$\begin{aligned}
&& \gamma_0 &= y_0 \\
\gamma_0 + \gamma_1(x_1 - x_0) = y_1 &\Longrightarrow & \gamma_1 &= \frac{y_1 - y_0}{x_1 - x_0} \\
\gamma_0 + \gamma_1(x_2 - x_0) + \gamma_2(x_2 - x_0)(x_2 - x_1) = y_2 &\Longrightarrow & \gamma_2 &= \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0} \\
&& &\vdots
\end{aligned}$$

Man bezeichnet $f_{[x_0,\ldots,x_i]} := \gamma_i$ als die i-te *dividierte Differenz* zu den Stützstellen $x_0, \ldots, x_i$, wobei $f_{[x_0]} = \gamma_0 = y_0$.

Allgemein berechnen sich die dividierten Differenzen zu den Stützstellen $x_j, \ldots, x_{j+k}$ über die Rekursion

$$\begin{aligned}
j = 0, \ldots, n: &\quad f_{[x_j]} = y_j \\
k = 1, \ldots, n: \quad j = 0, \ldots, n - k: &\quad f_{[x_j,\ldots,x_{j+k}]} = \frac{f_{[x_{j+1},\ldots,x_{j+k}]} - f_{[x_j,\ldots,x_{j+k-1}]}}{x_{j+k} - x_j}.
\end{aligned} \tag{2.4}$$

Man erhält das

Newtonsche Interpolationspolynom

$$p_n(x) = \gamma_0 + \sum_{i=1}^{n} \gamma_i(x - x_0)\cdots(x - x_{i-1}), \quad \gamma_i = f_{[x_0,\ldots,x_i]} \tag{2.5}$$

mit den dividierten Differenzen $f_{[x_0,\ldots,x_i]}$ aus (2.4).

Begründung: Für $n = 0$ ist die Darstellung klar. Sind $p_{1,\ldots,i+1}$ und $p_{0,\ldots,i}$ die Interpolanten in $x_1, \ldots, x_{i+1}$ bzw. $x_0, \ldots, x_i$ vom Grad $\le i$, dann gilt

$$\begin{aligned}
p_{i+1}(x) &= \frac{(x - x_0)p_{1,\ldots,i+1}(x) + (x_{i+1} - x)p_{0,\ldots,i}(x)}{x_{i+1} - x_0} \\
&= \frac{f_{[x_1,\ldots,x_{i+1}]} - f_{[x_0,\ldots,x_i]}}{x_{i+1} - x_0}(x - x_0)\cdots(x - x_i) + \underbrace{\text{Polynom vom Grad } i}_{:=q_i(x)}.
\end{aligned}$$

Da der erste Summand in $x_0, \ldots, x_i$ verschwindet, gilt $q_i(x) = p_i(x)$ wegen (2.1). Vergleich mit (2.5) liefert (2.4). $\square$

Wir erhalten aus (2.4) folgende Vorschrift zur Berechnung der Koeffizienten $\gamma_i = f_{[x_0,\dots,x_i]}$:

Berechnung der dividierten Differenzen:

Setze $f_{[x_j]} = y_j$, $j = 0, \dots, n$.

Berechne für $k = 1, \dots, n$ *und* $j = 0, \dots, n-k$*:*

$$f_{[x_j,\dots,x_{j+k}]} = \frac{f_{[x_{j+1},\dots,x_{j+k}]} - f_{[x_j,\dots,x_{j+k-1}]}}{x_{j+k} - x_j}.$$

Wir erhalten also das Schema

$$\begin{array}{c|lll}
x_0 & f_{[x_0]} = y_0 \searrow & & \\
 & & f_{[x_0,x_1]} \searrow & \\
x_1 & f_{[x_1]} = y_1 \begin{smallmatrix}\nearrow\\ \searrow\end{smallmatrix} & & f_{[x_0,x_1,x_2]} \\
 & & f_{[x_1,x_2]} \begin{smallmatrix}\nearrow\\ \searrow\end{smallmatrix} & \\
x_2 & f_{[x_2]} = y_2 \nearrow & & \\
\vdots & & &
\end{array}$$

Vorteile:

- Der Rechenaufwand beträgt:
  Berechnung der dividierten Differenzen: $O(n^2)$
  Auswertung von $p_n(x)$: $O(n)$
- Hinzunahme einer neuen Stützstelle erfordert nur die Berechnung von n zusätzlichen dividierten Differenzen.

2.1.3 Fehlerabschätzungen

Nimmt man an, dass die Stützwerte von einer Funktion $f : [a, b] \to \mathbb{R}$ herrühren, also

$$y_i = f(x_i), \quad i = 0, \dots, n,$$

dann erhebt sich die Frage, wie gut das Interpolationspolynom p_n auf $[a, b]$ mit f übereinstimmt. Es gilt der folgende Satz:

Satz 2.1.2 *Sei f $(n+1)$-mal stetig differenzierbar, kurz $f \in C^{n+1}([a, b])$. Seien $x_0, \dots, x_n \in [a, b]$ verschiedene Punkte und sei p_n das eindeutige Interpolationspolynom vom Grad $\le n$ zu den Stützwerten $(x_i, f(x_i))$, $i = 0, \dots, n$. Dann existiert zu jedem $x \in [a, b]$ ein $\xi_x \in [a, b]$ mit*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x - x_0) \cdots (x - x_n).$$

Das Restglied der Interpolation hat also zwei Faktoren: Das sogenannte *Knotenpolynom*

$$\omega(x) = \prod_{i=0}^{n} (x - x_i)$$

und den Faktor $\frac{f^{(n+1)}(\xi_x)}{(n+1)!}$. Durch Abschätzung beider Terme ergibt sich zum Beispiel folgende Fehlerabschätzung.

Korollar 2.1.3 *Unter den Voraussetzungen von Satz 2.1.2 gilt*

$$\max_{x\in[a,b]} |f(x) - p_n(x)| \le \max_{x\in[a,b]} \frac{|f^{(n+1)}(x)|}{(n+1)!} \max_{x\in[a,b]} |\omega(x)| \le \max_{x\in[a,b]} \frac{|f^{(n+1)}(x)|}{(n+1)!} (b-a)^{n+1}.$$

Achtung: Bei äquidistanter Wahl der Stützpunkte, also $x_i = a + ih$, $h = (b-a)/n$, ist nicht immer gewährleistet, dass gilt

$$\lim_{n\to\infty} f(x) - p_n(x) = 0 \quad \text{für alle } x \in [a,b].$$

Zum Beispiel ist dies für $f(x) = \frac{1}{1+x^2}$ auf $[a,b] = [-5,5]$ der Fall. □

Als Ausweg kann man x_i als die sog. *Tschebyscheff-Abszissen* wählen, für die $\max_{x\in[a,b]} |\omega(x)|$ minimal wird: Wahl der

Tschebyscheffabszissen

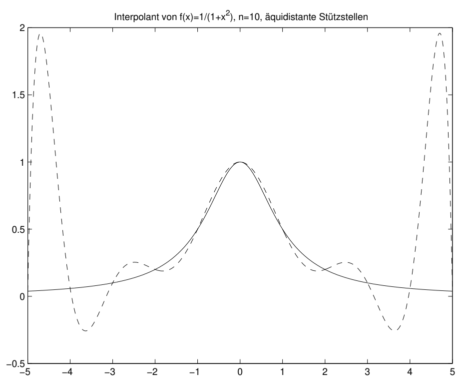
$$x_i = \frac{b-a}{2} \cos\left(\frac{2i+1}{n+1}\frac{\pi}{2}\right) + \frac{b+a}{2}, \quad i = 0, \dots, n. \tag{2.6}$$

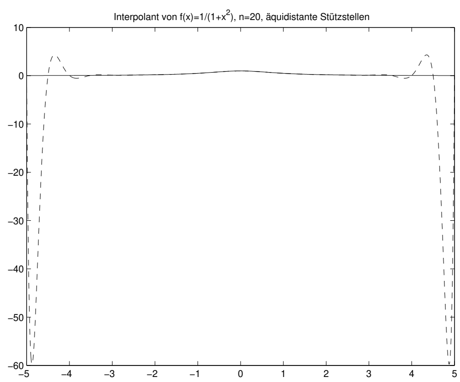
liefert den minimalen Wert für $\max_{x\in[a,b]} |\omega(x)|$, nämlich

$$\max_{x\in[a,b]} |\omega(x)| = \left(\frac{b-a}{2}\right)^{n+1} 2^{-n}.$$

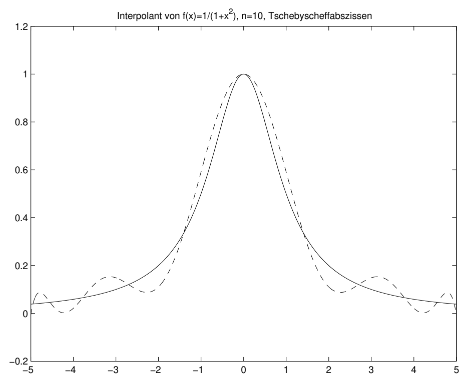
Beipiel: $f(x) = \frac{1}{1+x^2}$ auf $[a,b] = [-5,5]$. Wie bereits erwähnt, geht bei äquidistanten Stützstellen der Fehler $f(x) - p_n(x)$ für $n \to \infty$ nicht an allen Stellen $x \in [a,b]$ gegen 0.

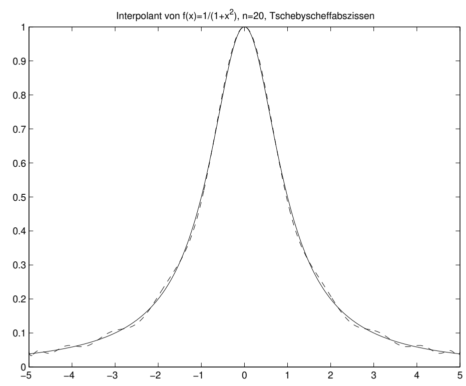
Interpolant bei äquidistanten Stützstellen:





Interpolant bei Tschebyscheffstützstellen:





Allgemein sollte man in der Praxis nicht n sehr groß wählen, sondern besser stückweise in kleinen Intervallen vorgehen, siehe 2.2.

2.1.4 Anwendungen der Polynominterpolation

Wir geben eine Auswahl von Anwendungen für die Polynominterpolation an:

1. **Approximation einer Funktion auf einem Intervall:** Wir haben gesehen, dass hierzu nicht äquidistante Stützstellen sondern die Tschbyscheffabszissen gewählt werden sollten.

2. **Inverse Interpolation:** Sei $f : [a, b] \to \mathbb{R}$ bijektiv, also $f'(x) \neq 0$ auf $[a, b]$. Sind dann (x_i, y_i), $y_i = f(x_i)$, Stützpunkte von f, dann sind (y_i, x_i) wegen $x_i = f^{-1}(y_i)$ Stützpunkte für f^{-1} und eine Approximation von f^{-1} kann durch Interpolation der Stützpunkte (y_i, x_i) gewonnen werden.

3. **Numerische Integration:** (Kapitel 3)
   Zur näherungsweisen Berechnung des Integrals einer Funktion kann man zunächst ein Interpolationspolynom bestimmen, das anschließend einfach integriert werden kann: $\displaystyle\int_a^b f(x)\,dx \approx \int_a^b p_n(x)\,dx.$

4. **Numerische Differentiation:** Mit einem Interpolationspolynom p_n von f ist p_n' eine Approximation von f'.

2.2 Spline-Interpolation

Bei der Polynominterpolation wird die Funktion f auf dem Intervall $[a, b]$ durch *ein* Polynom vom Grad n interpoliert. Wir hatten festgestellt, dass große Genauigkeit nicht immer durch die Wahl vieler Stützstellen sichergestellt werden kann.

Als Ausweg kann man stückweise Interpolation verwenden. Hierbei zerlegt man das Ausgangsintervall $[a, b]$ in kleine Teilintervalle und verwendet auf jedem Teilintervall ein interpolierendes Polynom fester Ordnung. An den Intervallgrenzen sorgt man dafür, dass die Polynome k-mal stetig differenzierbar ineinander übergehen, wobei k fest ist, und die Welligkeit des Interpolanten möglichst klein ist. Dieses Konzept führt auf die Spline-Interpolation.

2.2.1 Grundlagen

Sei

$$\Delta = \{x_i \,:\, a = x_0 < x_1 < \ldots < x_n = b\}$$

eine Zerlegung des Intervalls $[a, b]$. Aus historischen Gründen nennt man die x_i *Knoten*.

Definition 2.2.1 *Eine* Splinefunktion der Ordnung l zur Zerlegung Δ *ist eine Funktion* $s : [a, b] \to \mathbb{R}$ *mit folgenden Eigenschaften*

- *Es gilt* $s \in C^{l-1}([a, b])$*,* s *ist also stetig und* $l - 1$*-mal stetig differenzierbar.*
- s *stimmt auf jedem Intervall* $[x_i, x_{i+1}]$ *mit einem Polynom* s_i *vom Grad* $\leq l$ *überein.*

Die Menge dieser Splinefunktionen bezeichnen wir mit $S_{\Delta,l}$*.*

Im Folgenden betrachten wir nur den Fall $l = 1$ (*lineare Splines*) und $l = 3$ (*kubische Splines*).

Wir wollen nun Splines zur Interpolation verwenden und betrachten folgende Aufgabenstellung:

Spline-Interpolation:

Zu einer Zerlegung $\Delta = \{x_i \,:\, a = x_0 < x_1 < \ldots < x_n = b\}$ und Werten $y_i \in \mathbb{R}$, $i = 0, \ldots, n$ bestimme $s \in S_{\Delta,l}$ mit

$$s(x_i) = y_i, \quad i = 0, \ldots, n. \tag{2.7}$$

2.2.2 Interpolation mit linearen Splines

Ein linearer Spline $s \in S_{\Delta,1}$ ist stetig und auf jedem Intervall $[x_i, x_{i+1}]$ ein Polynom s_i vom Grad ≤ 1. Die Interpolationsbedingungen (2.7) erfordern daher $s_i(x_i) = y_i$, $s_i(x_{i+1}) = y_{i+1}$ und legen s_i eindeutig fest zu

$$s(x) = s_i(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} y_i + \frac{x - x_i}{x_{i+1} - x_i} y_{i+1} \quad \forall\, x \in [x_i, x_{i+1}]. \tag{2.8}$$

Definieren wir die ”Dachfunktionen”

$$\varphi_i(x) = \begin{cases} 0 & \text{falls } x < x_{i-1}, \\ \frac{x-x_{i-1}}{x_i-x_{i-1}} & \text{falls } x \in [x_{i-1}, x_i], \\ \frac{x_{i+1}-x}{x_{i+1}-x_i} & \text{falls } x \in [x_i, x_{i+1}], \\ 0 & \text{falls } x > x_{i+1}. \end{cases}$$

mit beliebigen Hilfsknoten $x_{-1} < a$ und $x_{n+1} > b$, dann erhalten wir für $s(x)$ auf $[a, b]$ die bequeme Darstellung

$$s(x) = \sum_{i=0}^{n} y_i \varphi_i(x), \quad x \in [a, b].$$

Satz 2.2.2 *Zu einer Zerlegung $\Delta = \{x_i : a = x_0 < x_1 < \ldots < x_n = b\}$ von $[a, b]$ und Werten y_i, $i = 0, \ldots, n$, existiert genau ein interpolierender linearer Spline.*

Ferner gilt folgende Fehlerabschätzung.

Satz 2.2.3 *Sei $f \in C^2([a, b])$. Dann gilt für jede Zerlegung $\Delta = \{x_i \,;\, a = x_0 < x_1 < \ldots < x_n = b\}$ von $[a, b]$ und den zugehörigen interpolierenden linearen Spline $s \in S_{\Delta,1}$ von f*

$$\max_{x\in[a,b]} |f(x) - s(x)| \leq \frac{1}{8} \max_{x\in[a,b]} |f''(x)| h_{max}^2 \quad \textit{mit } h_{max} = \max_{i=0,\ldots,n-1} x_{i+1} - x_i.$$

Beweis: Auf jedem Intervall $[x_i, x_{i+1}]$ ist s ein interpolierendes Polynom vom Grad ≤ 1. Daher gilt nach Satz 2.1.2

$$|f(x) - s(x)| = \frac{|f''(\xi)|}{2!}(x_{i+1} - x)(x - x_i) \leq \frac{|f''(\xi)|}{2!} \frac{h_{max}^2}{4} \quad \forall\, x \in [x_i, x_{i+1}]$$

mit einem $\xi \in [x_i, x_{i+1}]$. Daraus folgt unmittelbar die Behauptung. □

2.2.3 Interpolation mit kubischen Splines

Kubische Splines sind zweimal stetig differenzierbar aus kubischen Polynomen zusammengesetzt. Wir werden sehen, dass die Interpolation mit kubischen Splines es gestattet, gegebene Punkte durch eine Funktion minimaler Krümmung zu interpolieren.

Berechnung kubischer Spline-Interpolanten

Ist $s \in S_{\Delta,3}$ ein kubischer Spline, dann ist s'' offensichtlich stetig und stückweise linear, also $s'' \in S_{\Delta,1}$. Es bietet sich daher an, s_i durch Integration von s_i'' zu bestimmen.

Seien $M_i = s_i''(x_i)$. Man nennt M_i *Momente*. Dann gilt nach (2.8)

$$s_i''(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i} M_i + \frac{x - x_i}{x_{i+1} - x_i} M_{i+1}.$$

Zweifache Integration ergibt dann den Ansatz

$$s_i(x) = \frac{1}{6} \left(\frac{(x_{i+1} - x)^3}{x_{i+1} - x_i} M_i + \frac{(x - x_i)^3}{x_{i+1} - x_i} M_{i+1} \right) + c_i(x - x_i) + d_i$$

mit Konstanten $c_i, d_i \in \mathbb{R}$. Wir berechnen c_i und d_i aus den Bedingungen

$$s_i(x_i) = y_i, \quad s_i(x_{i+1}) = y_{i+1}.$$

Mit

$$h_i = x_{i+1} - x_i$$

liefert dies

$$d_i = y_i - \frac{h_i^2}{6} M_i, \quad c_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{6}(M_{i+1} - M_i).$$

Einsetzen in die Gleichungen $s_i'(x_i) = s_{i-1}'(x_i)$ ergibt schließlich folgende Gleichungen für die Momente M_i:

$$\text{(2.9)} \quad \frac{h_{i-1}}{6} M_{i-1} + \frac{h_{i-1} + h_i}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}, \quad i = 1, \dots, n-1.$$

Dies sind $n - 1$ Gleichungen für $n + 1$ Unbekannte. Der Spline-Interpolant wird eindeutig durch zwei zusätzlich Randbedingungen:

Wichtige Randbedingungen für kubische Splines:

a) Natürliche Randbedingungen: $s''(a) = s''(b) = 0$, also $M_0 = M_n = 0$

b) Hermite-Randbedingungen: $s'(a) = f'(a)$, $s'(b) = f'(b)$, also

$$\frac{h_0}{3} M_0 + \frac{h_0}{6} M_1 = \frac{y_1 - y_0}{h_0} - f'(a), \quad \frac{h_{n-1}}{3} M_n + \frac{h_{n-1}}{6} M_{n-1} = f'(b) - \frac{y_n - y_{n-1}}{h_{n-1}}.$$

Für jeden der Fälle a)-b) ergibt sich zusammen mit (2.9) eine eindeutige Lösung für $M_0, \dots, M_n$.

Für a) und b) erhält man ein strikt diagonaldominantes tridiagonales Gleichungssystem der Form

$$\text{(2.10)} \quad \begin{pmatrix} \mu_0 & \lambda_0 & & & & \\ \frac{h_0}{6} & \frac{h_0+h_1}{3} & \frac{h_1}{6} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \frac{h_{i-1}}{6} & \frac{h_{i-1}+h_i}{3} & \frac{h_i}{6} & \\ & & & \ddots & \ddots & \ddots \\ & & & & \lambda_n & \mu_n \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_n \end{pmatrix} = \begin{pmatrix} b_0 \\ \frac{y_2-y_1}{h_1} - \frac{y_1-y_0}{h_0} \\ \vdots \\ \frac{y_{i+1}-y_i}{h_i} - \frac{y_i-y_{i-1}}{h_{i-1}} \\ \vdots \\ b_n \end{pmatrix}.$$

Für a) kann man zum Beispiel $b_0 = b_n = \lambda_0 = \lambda_n = 0$ und $\mu_0 = \mu_n = 1$ wählen. Wegen der strikten Diagonaldominanz ist nach dem Satz von Gershgorin 0 kein Eigenwert und daher ist die Koeffizientenmatrix invertierbar.

Minimaleigenschaften kubischer Splines

Es zeigt sich, dass der kubische Spline-Interpolant mit Randbedingung a) oder b) unter allen zweimal stetig differenzierbaren minimale Krümmung im folgenden Sinne hat:

Satz 2.2.4 *Gegeben sei eine beliebige Funktion $f \in C^2([a,b])$ und eine Unterteilung Δ von $[a,b]$. Dann gilt für den kubischen Spline-Interpolanten $s \in S_{\Delta,3}$ mit Randbedingungen a) oder b)*

$$\int_a^b f''(x)^2\,dx = \int_a^b s''(x)^2\,dx + \int_a^b (f''(x) - s''(x))^2\,dx \geq \int_a^b s''(x)^2\,dx.$$

Beweis: Siehe zum Beispiel [St94], [Pl00]. □

Fehlerabschätzung für kubische Spline-Interpolation

Unter Verwendung der Tatsache, dass die Momente $\hat{M}_i = f''(x_i)$ das Gleichungssystem (2.10) auf $O(h_{max}^3)$ mit $h_{max} = \max_{0 \leq i < n} h_i$ erfüllen und die Norm der Inversen der Koeffizientenmatrix in (2.10) von der Ordnung $O(1/h_{min})$ ist mit $h_{min} = \min_{0 \leq i < n} h_i$, kann man folgendes Resultat zeigen.

Satz 2.2.5 *Sei $f \in C^4([a,b])$ mit $f''(a) = f''(b) = 0$. Dann gilt für jede Unterteilung Δ mit dem kubischen Spline-Interpolanten $s \in S_{\Delta,3}$ zu Randbedingungen a)*

$$|f(x) - s(x)| \leq \frac{h_{max}}{h_{min}} \sup_{\xi \in [a,b]} |f^{(4)}(\xi)| h_{max}^4,$$

$$|f^{(k)}(x) - s^{(k)}(x)| \leq \frac{2h_{max}}{h_{min}} \sup_{\xi \in [a,b]} |f^{(4)}(\xi)| h_{max}^{4-k}, \quad k = 1, 2.$$

Beweis: Siehe zum Beispiel [Pl00]. □

Für Hermite-Randbedingungen lässt sich der Satz verschärfen:

Satz 2.2.6 *Sei $f \in C^4([a,b])$. Dann gilt für jede Unterteilung Δ mit dem kubischen Spline-Interpolanten $s \in S_{\Delta,3}$ zu Randbedingungen b)*

$$|f(x) - s(x)| \leq \frac{5}{384} \sup_{\xi \in [a,b]} |f^{(4)}(\xi)| h_{max}^4,$$

$$|f^{(k)}(x) - s^{(k)}(x)| \leq \frac{2h_{max}}{h_{min}} \sup_{\xi \in [a,b]} |f^{(4)}(\xi)| h_{max}^{4-k}, \quad k = 1, 2.$$

Beweis: Siehe zum Beispiel [DB02, Pl00, TS90]. □

Kapitel 3

Numerische Integration

In diesem Kapitel stellen wir einige wichtige Verfahren zur näherungsweisen Berechnung bestimmter Integrale $\int_a^b f(x)\,dx$ vor.

Integrationsaufgabe:

Zu gegebenem integrierbarem $f : [a,b] \to \mathbb{R}$ berechne

$$I(f) = \int_a^b f(x)\,dx.$$

Schon für relativ einfache Funktionen läßt sich die Stammfunktion nicht analytisch angeben, etwa $\frac{\sin x}{x}$ und e^{-x^2}. Man ist dann auf numerische Integrationsverfahren angewiesen.

Wichtige numerische Integrationsverfahren beruhen darauf, f mit Hilfe einiger Stützpunkte $(x_i, f(x_i))$, $x_i \in [a,b]$ durch ein Polynom p_n zu interpolieren und dann dieses zu integrieren. Diese Vorgehensweise liefert die Integralapproximation

$$I_n(x) = \int_a^b p_n(x)\,dx$$

und wird als *interpolatorische Quadratur* bezeichnet.

3.1 Newton-Cotes-Quadratur

3.1.1 Geschlossene Newton-Cotes-Quadratur

Wir wählen für $n \in \mathbb{N}$ die äquidistanten Stützstellen

$$x_i = a + ih, \quad i = 0, \ldots, n, \quad \text{mit } h = \frac{b-a}{n}.$$

Dann lautet das Interpolationspolynom p_n in Lagrange-Darstellung

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_{i,n}(x), \quad L_{i,n}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}.$$

Wir erhalten nun die numerische Quadraturformel

$$I_n(f) = \int_a^b p_n(x)\,dx = \sum_{i=0}^{n} f(x_i) \int_a^b L_{i,n}(x)\,dx.$$

Mit der Substitution $x = a + sh$ und $s \in [0, n]$ ergibt sich die

Geschlossene Newton-Cotes Formel:

$$I_n(f) = h \sum_{i=0}^{n} \alpha_{i,n} f(x_i),$$

$$\text{mit} \quad \alpha_{i,n} = \int_0^n \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{s - j}{i - j}\,ds. \tag{3.1}$$

Die Zahlen $\alpha_{0,n}, \ldots, \alpha_{n,n}$ heißen *Gewichte*. Sie sind *unabhängig* von f und $[a, b]$ und somit tabellierbar. Es gilt stets

$$\sum_{i=0}^{n} h\alpha_{i,n} = b - a.$$

Definition 3.1.1 *Eine Integrationsformel* $J(f) = \sum_{i=0}^{n} \beta_i f(x_i)$ *heißt* exakt vom Grad n, *falls sie alle Polynome bis mindestens vom Grad* n *exakt integriert.*

Die geschlossene Newton-Cotes Formel $I_n(f)$ ist nach Konstruktion exakt vom Grad n.

Es ist wichtig, eine Abschätzung für den Fehler

$$E_n(f) := I(f) - I_n(f)$$

zur Verfügung zu haben. Nach Korollar 2.1.3 gilt

$$|f(x) - p_n(x)| \leq \frac{|f^{(n+1)}(\xi)|}{(n+1)!}(b - a)^{n+1}$$

mit einem $\xi \in [a, b]$. Dies ergibt mit einem (unter Umständen anderen) $\xi \in [a, b]$

$$\left| \int_a^b f(x)\,dx - \int_a^b p_n(x)\,dx \right| \leq \int_a^b |f(x) - p_n(x)|\,dx \leq \frac{|f^{(n+1)}(\xi)|}{(n+1)!}(b - a)^{n+2}.$$

Eine verfeinerte Restgliedabschätzung ergibt sich durch Taylorentwicklung. Dies liefert die folgende Tabelle.

n	$\alpha_{i,n}$					Fehler $E_n(f)$	Name
1	$\frac{1}{2}$	$\frac{1}{2}$				$-\frac{f^{(2)}(\xi)}{12}h^3$	Trapezregel
2	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$			$-\frac{f^{(4)}(\xi)}{90}h^5$	Simpson-Regel
3	$\frac{3}{8}$	$\frac{9}{8}$	$\frac{9}{8}$	$\frac{3}{8}$		$-\frac{3f^{(4)}(\xi)}{80}h^5$	3/8-Regel
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{14}{45}$	$-\frac{8f^{(6)}(\xi)}{945}h^7$	Milne-Regel

Für $n \geq 7$ treten leider negative Gewichte auf und die Formeln werden dadurch zunehmend numerisch instabil, da Auslöschung auftritt.

3.1.2 Offene Newton-Cotes-Quadratur

Hier wählen wir für $n \in \mathbb{N} \cup \{0\}$ die in $]a, b[$ liegenden äquidistanten Stützstellen

$$x_i = a + ih, \quad i = 1, \ldots, n+1, \quad \text{mit } h = \frac{b-a}{n+2}.$$

Geht man völlig analog vor, dann erhält man wiederum interpolatorische Interpolationsformeln, die

Offene Newton-Cotes Formel:

$$\tilde{I}_n(f) = h \sum_{i=1}^{n+1} \tilde{\alpha}_{i,n} f(x_i), \quad \tilde{\alpha}_{i,n} = \int_0^{n+2} \prod_{\substack{j=1 \\ j \neq i}}^{n+1} \frac{s-j}{i-j}\, ds.$$

Für den Quadraturfehler ergeben sich ähnliche Formeln wie im geschlossenen Fall:

n	$\tilde{\alpha}_{i,n}$			Fehler $\tilde{E}_n(f)$	Name
0	2			$\frac{f^{(2)}(\xi)}{3}h^3$	Rechteck-Regel
1	$\frac{3}{2}$	$\frac{3}{2}$		$\frac{3f^{(2)}(\xi)}{4}h^3$	
2	$\frac{8}{3}$	$-\frac{4}{3}$	$\frac{8}{3}$	$\frac{28f^{(4)}(\xi)}{90}h^5$	

3.2 Die summierten Newton-Cotes-Formeln

Die Newton-Cotes-Formeln liefern nur genaue Ergebnisse, solange das Integrationsintervall klein und die Zahl der Knoten nicht zu groß ist. Es bietet sich wieder an, das Intervall $[a, b]$ in kleinere Intervalle zu zerlegen und auf diesen jeweils mit einer Newton-Cotes-Formel zu arbeiten.

Wir zerlegen dazu das Intervall $[a, b]$ in m Teilintervalle der Länge $H = \frac{b-a}{m}$, wenden die Newton-Cotes-Formeln vom Grad n einzeln auf diese Teilintervalle an und summieren die

Teilintegrale auf: Mit

$$\begin{aligned} N &= m \cdot n, \quad H = \frac{b-a}{m}, \quad h = \frac{H}{n} = \frac{b-a}{N} \\ x_i &= a + ih, \quad i = 0, \dots, N, \\ y_j &= a + jH, \quad j = 0, \dots, m \end{aligned}$$

ergibt sich wegen

$$I(f) = \sum_{j=0}^{m-1} \int_{y_j}^{y_{j+1}} f(x)\, dx$$

die

Summierte geschlossene Newton-Cotes-Formel

$$S_N^{(n)}(f) = h \sum_{j=0}^{m-1} \sum_{i=0}^{n} \alpha_{i,n} f(x_{jn+i}).$$

Die Gewichte $\alpha_{i,n}$ ergeben sich wieder aus (3.1). Der Quadraturfehler

$$R_N^{(n)}(f) = I(f) - S_N^{(n)}(f)$$

ergibt sich durch Summation der Fehler auf den Teilintervallen.

Satz 3.2.1 *Sei $f \in C^{n+2}([a,b])$. Dann existiert eine Zwischenstelle $\xi \in]a,b[$ mit*

$$R_N^{(n)}(f) = \begin{cases} C(n) f^{(n+2)}(\xi)(b-a)h^{n+2} & \text{für gerades } n, \\ C(n) f^{(n+1)}(\xi)(b-a)h^{n+1} & \text{für ungerades } n. \end{cases}$$

Hierbei ist $C(n)$ eine nur von n abhängige Konstante.

Wir geben noch die gebräuchlichsten summierten Formeln mit Quadraturfehler an:

Summierte Trapezregel: (geschlossen, $n = 1$, $h = \frac{b-a}{m}$)

$$S_N^{(1)}(f) = \frac{h}{2} \sum_{j=0}^{m-1} (f(x_j) + f(x_{j+1})), \quad x_j = a + jh.$$

Fehler: $R_N^{(1)}(f) = -\dfrac{f''(\xi)}{12}(b-a)h^2$.

Summierte Simpson-Regel: (geschlossen, $n = 2$, $h = \frac{b-a}{2m}$)

$$S_N^{(2)}(f) = \frac{h}{3} \sum_{j=0}^{m-1} (f(x_{2j}) + 4f(x_{2j+1}) + f(x_{2j+2})), \quad x_j = a + jh.$$

Fehler: $R_N^{(2)}(f) = -\dfrac{f^{(4)}(\xi)}{180}(b-a)h^4$

Summierte Rechteck-Regel: (offen, $n = 0$, $2m = N$, $h = \frac{b-a}{N}$)

$$\tilde{S}_N^{(0)}(f) = 2h \sum_{j=1}^{m} f(x_{2j-1}), \quad x_j = a + jh.$$

Fehler: $\tilde{R}_N^{(0)}(f) = \dfrac{f''(\xi)}{6}(b-a)h^2$

Kapitel 4

Lineare Gleichungssysteme

4.1 Problemstellung und Einführung

In diesem Kapitel betrachten wir direkte Verfahren zur Lösung von linearen Gleichungssystemen.

Lineares Gleichungssystem: Gesucht ist eine Lösung $x \in \mathbb{R}^n$ von

$$Ax = b. \tag{4.1}$$

mit

(4.2)

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n,n}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n.$$

Die hier besprochenen direkten Methoden liefern– rundungsfehlerfreie Rechnung vorausgesetzt– die Lösung von (4.1) in endlich vielen Rechenschritten. Bekanntlich ist (4.1) die Matrixschreibweise für

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i, \quad i = 1, \ldots, n.$$

Lineare Gleichungssysteme treten in der Praxis als Hilfsproblem bei einer Vielzahl von Problemstellungen auf, z.B. bei der Lösung von Rand- und Randanfangswertaufgaben für gewöhnliche und partielle Differentialgleichungen (Schaltkreissimulation, elektromagnetische Felder, ...), in der Bildverarbeitung, usw. . Schätzungen besagen, dass etwa 75% der Rechenzeit im technisch-wissenschaftlichen Bereich auf die Lösung von linearen Gleichungssystemen entfällt.

Wir erinnern zunächst an folgenden Sachverhalt.

Proposition 4.1.1 *Das lineare Gleichungssystem* (4.1) *hat eine Lösung genau dann, wenn gilt*

$$\textit{rang}(A) = \textit{rang}(A, b).$$

Hierbei ist bekanntlich für eine Matrix $B \in \mathbb{R}^{n,m}$ der Rang definiert durch

$$\begin{aligned} \mathrm{Rang}(B) &= \textit{Maximalzahl } r \textit{ der linear unabhängigen Zeilenvektoren} \\ &= \textit{Maximalzahl } r \textit{ der linear unabhängigen Spaltenvektoren}. \end{aligned}$$

Das lineare Gleichungssystem (4.1) *hat eine eindeutige Lösung genau dann, wenn A invertierbar ist (oder gleichbedeutend:* $\det(A) \neq 0$ *). Die eindeutige Lösung lautet dann*

$$x = A^{-1}b.$$

4.2 Das Gaußsche Eliminationsverfahren, Dreieckszerlegung einer Matrix

Das grundsätzliche Vorgehen der Gauß-Elimination ist aus der Linearen Algebra bekannt. Wir werden das Verfahren kurz wiederholen und zeigen, wie man daraus eine Dreieckszerlegung einer Matrix erhält. Zudem werden wir uns klarmachen, welchen Einfluss Rundungsfehler haben können und wie dieser Einfluss wirksam bekämpft werden kann.

Die Grundidee des Gaußschen Eliminationsverfahrens besteht darin, das Gleichungssystem (4.1) durch die elementaren Operationen

- Addition eines Vielfachen einer Gleichung zu einer anderen,
- Zeilenvertauschungen, d.h. Vertauschen von Gleichungen
- Spaltenvertauschungen, die einer Umnummerierung der Unbekannten entsprechen,

in ein Gleichungssystem der Form

$$Ry = c, \quad y_{\sigma_i} = x_i, \quad i = 1, \ldots, n,$$

mit der durchgeführten Spaltenpermutation $(\sigma_1, \ldots, \sigma_n)$ und einer oberen Dreiecksmatrix

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

zu überführen, das dieselben Lösungen wie (4.1) besitzt. (4.3) ist ein sogenanntes *gestaffeltes Gleichungssystem*, das man leicht durch Rückwärtssubstitution lösen kann, solange R invertierbar ist. Werden keine Spaltenvertauschungen durchgeführt, dann gilt $x = y$.

4.2.1 Lösung gestaffelter Gleichungssysteme

Gestaffelte Gleichungssysteme

$$Ry = c \tag{4.3}$$

mit einer oberen Dreiecksmatrix

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}, \tag{4.4}$$

sowie

$$Lz = d \tag{4.5}$$

mit einer unteren Dreiecksmatrix

$$L = \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix},$$

lassen sich offensichtlich leicht durch Rückwärts- bzw. Vorwärtssubstitution lösen:

Satz 4.2.1 *Seien $R = (r_{ij}) \in \mathbb{R}^{n,n}$ und $L = (l_{ij}) \in \mathbb{R}^{n,n}$ invertierbare obere bzw. untere Dreiecksmatrizen und $c = (c_1, \ldots, c_n)^T$, $d = (d_1, \ldots, d_n)^T$ Spaltenvektoren. Dann lassen sich die Lösungen von* (4.3) *bzw.* (4.5) *folgendermaßen berechnen:*

a) **Rückwärtssubstitution für obere Dreieckssysteme** (4.3)**:**

$$\boxed{y_i = \frac{c_i - \sum_{j=i+1}^{n} r_{ij} y_j}{r_{ii}}, \quad i = n, n-1, \ldots, 1.}$$

b) **Vorwärtssubstitution für untere Dreieckssysteme** (4.5)**:**

$$\boxed{z_i = \frac{d_i - \sum_{j=1}^{i-1} l_{ij} z_j}{l_{ii}}, \quad i = 1, 2, \ldots, n.}$$

Beweis: zu a): Da R invertierbar ist, gilt

$$\det(R) = r_{11} r_{22} \cdots r_{nn} \neq 0,$$

also $r_{ii} \neq 0$, $i = 1, \ldots, n$. Somit ergibt sich

$$\begin{aligned} y_n &= \frac{c_n}{r_{nn}} \\ y_{n-1} &= \frac{c_{n-1} - r_{n-1,n} y_n}{r_{n-1,n-1}} \\ &\vdots \end{aligned}$$

und somit induktiv a).

zu b): Wegen $\det(L) = l_{11}l_{22}\cdots l_{nn} \neq 0$ gilt $l_{ii} \neq 0$, $i = 1, \ldots, n$. Somit ergibt sich

$$\begin{aligned} z_1 &= \frac{d_1}{l_{11}} \\ z_2 &= \frac{d_2 - l_{2,1}z_1}{l_{22}} \\ &\vdots \end{aligned}$$

und wir erhalten induktiv b). □

Bemerkung: Der Aufwand für die Rückwärtssubstitution ist $O(n^2)$ an elementaren Rechenoperationen, falls nicht zusätzlich eine spezielle Besetztheitsstruktur vorliegt (Dünnbesetztheit, Bandstruktur). □

4.2.2 Das Gaußsche Eliminationsverfahrens

Wir erklären nun (die grundsätzliche Vorgehensweise sollte aus der Linearen Algebra bekannt sein), wie man mit dem Gaußschen Eliminationsverfahren ein gestaffeltes Gleichungssystem erhält. Statt mit den Gleichungen (4.1) zu arbeiten, ist es bequemer, die Operationen an der um die rechte Seite erweiterten Koeffizientenmatrix

$$(A, b) = \left(\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & b_n \end{array}\right)$$

durchzuführen.

Beim Gaußschen Eliminationsverfahren geht man nun wie folgt vor:

Grundkonzept des Gaußschen Eliminationsverfahrens:

0. **Initialisierung:** $(A^{(1)}, b^{(1)}) = \left(\begin{array}{ccc|c} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right) := (A, b).$

1. **Pivotsuche:** Suche eine Gleichung r, die von x_1 abhängt, also mit $a_{r1}^{(1)} \neq 0$ und

vertausche sie mit der ersten Gleichung:

$$(A^{(1)}, b^{(1)}) = \left(\begin{array}{ccc|c} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{r1}^{(1)} & \cdots & a_{rn}^{(1)} & b_r^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right) \quad \rightsquigarrow \quad \left(\begin{array}{ccc|c} a_{r1}^{(1)} & \cdots & a_{rn}^{(1)} & b_r^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{11}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right)$$

$$=: \left(\begin{array}{ccc|c} \tilde{a}_{11}^{(1)} & \cdots & \tilde{a}_{1n}^{(1)} & \tilde{b}_1^{(1)} \\ \vdots & & \vdots & \vdots \\ \tilde{a}_{n1}^{(1)} & \cdots & \tilde{a}_{nn}^{(1)} & \tilde{b}_n^{(1)} \end{array}\right) = (\tilde{A}^{(1)}, \tilde{b}^{(1)}).$$

Ist A invertierbar, dann existiert immer ein solches r, da wegen der Invertierbarkeit von A die erste Spalte nicht verschwinden kann.

2. **Elimination:** Subtrahiere geeignete Vielfache der ersten Gleichung von den übrigen Gleichungen derart, dass die Koeffizienten von x_1 in diesen Gleichungen verschwinden. Offensichtlich muss man hierzu jeweils das l_{i1}-fache mit

$$l_{i1} = \frac{\tilde{a}_{i1}^{(1)}}{\tilde{a}_{11}^{(1)}}$$

der ersten Gleichung von der i-ten Gleichung subtrahieren:

$$(\tilde{A}^{(1)}, \tilde{b}^{(1)}) \quad \rightsquigarrow \quad (A^{(2)}, b^{(2)}) = \left(\begin{array}{cccc|c} \tilde{a}_{11}^{(1)} & \tilde{a}_{12}^{(1)} & \cdots & \tilde{a}_{1n}^{(1)} & \tilde{b}_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right)$$

$$=: \left(\begin{array}{ccc|c} \tilde{a}_{11}^{(1)} & \cdots & \tilde{a}_{1n}^{(1)} & \tilde{b}_1^{(1)} \\ \hline 0 & & & \\ \vdots & \hat{A}^{(2)} & & \hat{b}^{(2)} \\ 0 & & & \end{array}\right).$$

3. **Iteration:** Wende für $k = 2, \ldots, n-1$ Schritt 1. und 2. auf $(\hat{A}^{(k)}, \hat{b}^{(k)})$ an:

   1_k. Wähle ein Pivotelement $a_{rk}^{(k)} \neq 0$, $k \leq r \leq n$, vertausche Zeile k und r $\rightsquigarrow \ (\tilde{A}^{(k)}, \tilde{b}^{(k)})$

   2_k. Subtrahiere das l_{ik}-fache mit

$$l_{ik} = \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}}$$

   der k-ten Gleichung von der i-ten Gleichung, $i = k+1, \ldots, n$.
   $\rightsquigarrow \ (A^{(k+1)}, b^{(k+1)})$

Nach k Eliminationsschritten

$$(A,b) =: (A^{(1)}, b^{(1)}) \to (A^{(2)}, b^{(2)}) \to \ldots \to (A^{(k+1)}, b^{(k+1)})$$

erhalten wir also eine Zwischenmatrix der Form

$$(A^{(k+1)}, b^{(k+1)}) = \left(\begin{array}{ccc|ccc|c} \tilde{a}_{11}^{(1)} & \cdots & \tilde{a}_{1k}^{(1)} & \cdots & & \tilde{a}_{1n}^{(1)} & \tilde{b}_1^{(1)} \\ 0 & \ddots & & & & \vdots & \vdots \\ & & \tilde{a}_{kk}^{(k)} & \cdots & & \tilde{a}_{kn}^{(k)} & \tilde{b}_k^{(k)} \\ \hline & & 0 & & & & \\ & & \vdots & \hat{A}^{(k+1)} & & & \hat{b}^{(k+1)} \\ & & 0 & & & & \end{array}\right).$$

Nach $n-1$ Eliminationsschritten liegt somit ein gestaffeltes Gleichungssystem (4.3)

$$Rx = c, \quad \text{mit} \quad R = A^{(n)}, \quad c = b^{(n)}$$

vor.

Pivotstrategie

Das Element $a_{rk}^{(k)}$, das in Schritt 1_k bestimmt wird, heißt *Pivotelement*. Theoretisch kann man bei der Pivotsuche jedes $a_{rk}^{(k)} \neq 0$ als Pivotelement wählen. Die Wahl kleiner Pivotelemente kann aber zu einer dramatischen Verstärkung von Rundungsfehlern führen. Gewöhnlich trifft man daher die Wahl von $a_{rk}^{(k)}$ durch

Spaltenpivotsuche: Wähle $k \leq r \leq n$ mit $\lvert a_{rk}^{(k)}\rvert = \max_{k \leq i \leq n} \lvert a_{ik}^{(k)}\rvert$.

Hierbei sollten die Zeilen von A "equilibriert" sein, also ihre Normen dieselbe Größenordnung haben.

Beispiel 4.2.1 Betrachte das Beispiel

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & -2 & 4 \\ 2 & 1 & -2 \end{pmatrix} x = \begin{pmatrix} 2 \\ 10 \\ -2 \end{pmatrix}$$

Dies liefert

$$\left(\begin{array}{rrr|r} 1 & 2 & -1 & 2 \\ 2 & -2 & 4 & 10 \\ 2 & 1 & -2 & -2 \end{array}\right) \quad \overset{\text{Spaltenpivotsuche}}{\longrightarrow} \quad \left(\begin{array}{rrr|r} 2 & -2 & 4 & 10 \\ 1 & 2 & -1 & 2 \\ 2 & 1 & -2 & -2 \end{array}\right)$$

$$\overset{\text{Elimination}}{\longrightarrow} \quad \begin{array}{l} -(\underbrace{1/2}_{=l_{21}}) \cdot \text{Zeile 1} \\ -(\underbrace{1}_{=l_{31}}) \cdot \text{Zeile 1} \end{array} \left(\begin{array}{rrr|r} 2 & -2 & 4 & 10 \\ 0 & 3 & -3 & -3 \\ 0 & 3 & -6 & -12 \end{array}\right)$$

$$\overset{\text{Spaltenpivotsuche+Elimination}}{\longrightarrow} \quad -\underbrace{1}_{=l_{32}} \cdot \text{Zeile 2} \left(\begin{array}{rrr|r} 2 & -2 & 4 & 10 \\ 0 & 3 & -3 & -3 \\ 0 & 0 & -3 & -9 \end{array}\right)$$

4.2.3 Praktische Implementierung des Gauß-Verfahrens

Bei der Realisierung auf einem Rechner speichert man in der Regel auch die verwendeten Multiplikatoren l_{ik}. Wir werden sehen, dass das Gaußsche Eliminationsverfahren dann ”gratis” eine Dreieckszerlegung (oder LR-Zerlegung) von A der Form

$$LR = PA \tag{4.6}$$

liefert. Hierbei ist $R \in \mathbb{R}^{n,n}$ eine obere Dreiecksmatrix (4.4), $L \in \mathbb{R}^{n,n}$ eine untere Dreiecksmatrix der Form

$$L = \begin{pmatrix} 1 & & 0 & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & & \ddots & \ddots & \\ l_{n1} & & \cdots & l_{n,n-1} & 1 \end{pmatrix}, \tag{4.7}$$

und P eine *Permutationsmatrix*, die lediglich die Zeilen von A permutiert.

Wir erhalten die folgende Implementierung des Gauß-Verfahrens mit Spaltenpivotsuche:

Algorithmus 1 Gaußsches Eliminationsverfahren mit Spaltenpivotsuche

Setze $(A^{(1)}, b^{(1)}) = (A, b)$ *und* $L^{(1)} = 0 \in \mathbb{R}^{n,n}$.

Für $k = 1, 2, \dots, n-1$*:*

1. **Spaltenpivotsuche:** *Bestimme* $k \le r \le n$ *mit*

$$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|.$$

Falls $a_{rk}^{(k)} = 0$: STOP, A ist singulär.

Vertausche die Zeilen r und k von $(A^{(k)}, b^{(k)})$ und von $L^{(k)}$. Das Ergebnis sei formal mit $(\tilde{A}^{(k)}, \tilde{b}^{(k)})$, $\tilde{L}^{(k)}$ bezeichnet.

2. **Elimination:** *Subtrahiere für $i = k+1, \ldots, n$ das l_{ik}-fache, $l_{ik} = \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}}$, der k-ten Zeile von $(\tilde{A}^{(k)}, \tilde{b}^{(k)})$ von der i-ten Zeile und füge die Multiplikatoren l_{ik} in $\tilde{L}^{(k)}$ ein. Das Ergebnis sei formal mit $(A^{(k+1)}, b^{(k+1)})$ und $L^{(k+1)}$ bezeichnet.*

   Im Detail: *Initialisiere $(A^{(k+1)}, b^{(k+1)}) := (\tilde{A}^{(k)}, \tilde{b}^{(k)})$, $\quad L^{(k+1)} := \tilde{L}^{(k)}$.*

   Für $i = k+1, \ldots, n$;

$$
\begin{aligned}
l_{ik} &= \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}}, \\
b_i^{(k+1)} &= \tilde{b}_i^{(k)} - l_{ik}\tilde{b}_k^{(k)}, \\
a_{ik}^{(k+1)} &= 0, \\
l_{ik}^{(k+1)} &= l_{ik} \quad \textit{(Multiplikator speichern).}
\end{aligned}
$$

   Für $j = k+1, \ldots, n$:

$$a_{ij}^{(k+1)} = \tilde{a}_{ij}^{(k)} - l_{ik}\tilde{a}_{kj}^{(k)}$$

Ergebnis: *$R := A^{(n)}$, $c := b^{(n)}$, $L := I + L^{(n)}$ mit der Einheitsmatrix $I \in \mathbb{R}^{n,n}$.*

Bei einer Implementierung auf dem Rechner kann man für die Speicherung aller $A^{(k)}$, $b^{(k)}$, $\tilde{A}^{(k)}$, $\tilde{b}^{(k)}$ die Felder verwenden, in denen A und b gespeichert waren. $L^{(k)}$ kann man anstelle der enstehenden Nullen im strikten unteren Dreieck platzsparend speichern.

Bemerkung: Der Rechenaufwand ist $n^3/3 - n/3$ an elementaren Rechenoperationen, falls nicht zusätzlich eine spezielle Besetztheitsstruktur vorliegt. □

Vollständige Pivotsuche

Anstelle der Spaltenpivotsuche kann man auch *vollständige Pivotsuche* durchführen, bei der man die Pivotsuche nicht auf die erste Spalte beschränkt. Schritt 1 in Algorithmus 1 ist dann wie folgt zu modifizieren:

Algorithmus 2 Gaußsches Eliminationsverfahren mit vollständiger Pivotsuche *Algorithmus 1 mit folgender Modifikation von Schritt 1:*

$1.'$ **Vollständige Pivotsuche:** *Bestimme $k \le r \le n$, $k \le s \le n$ mit*

$$|a_{rs}^{(k)}| = \max_{k \le i,j \le n} |a_{ij}^{(k)}|.$$

Falls $a_{rs}^{(k)} = 0$: STOP, A ist singulär.

Vertausche die Zeilen r und k sowie die Spalten s und k von $(A^{(k)}, b^{(k)})$ und von $L^{(k)}$. Das Ergebnis sei formal mit $(\tilde{A}^{(k)}, \tilde{b}^{(k)})$, $\tilde{L}^{(k)}$ bezeichnet.

Achtung: Bei jeder Spaltenvertauschung müssen die Komponenten von x entsprechend umnumeriert werden, d.h. nach Lösen von (4.3) müssen die Komponenten des Ergebnisvektors x zurückgetauscht werden.

In der Regel wird vollständige Pivotsuche nur bei "fast singulären" Matrizen angewandt, um den Rundungsfehlerinfluß minimal zu halten. □

4.2.4 Gewinnung einer Dreieckszerlegung

Wir zeigen nun, dass das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche (Algorithmus 1) eine Dreieckszerlegung (oder LR-Zerlegung) von A der Form

$$LR = PA \tag{4.6}$$

liefert. Hierbei ist $R \in \mathbb{R}^{n,n}$ bzw. $L \in \mathbb{R}^{n,n}$ die von Algorithmus 1 gelieferte obere Dreiecksmatrix der Form (4.4) bzw. untere Dreiecksmatrix der Form (4.7) und P ist eine *Permutationsmatrix* für die durchgeführten Zeilenvertauschungen.

Das Gaußsche Eliminationsverfahren mit vollständiger Pivotsuche (Algorithmus 2) liefert eine Dreieckszerlegung

$$LR = PAQ, \tag{4.8}$$

wobei P bzw. Q Permutationsmatrizen für die durchgeführten Zeilen- bzw. Spaltenvertauschungen sind.

Wir betrachten im Folgenden nur die Spaltenpivotsuche genauer, die vollständige Pivotsuche kann ähnlich behandelt werden.

Bemerkung: Eine Dreieckszerlegung (4.6) oder (4.8) ist sehr nützlich, wenn man (4.1) für mehrere rechte Seiten lösen will. Tatsächlich gilt

$$Ax = b \quad \Longleftrightarrow \quad PAQy = Pb, \;\; x = Qy \quad \Longleftrightarrow \quad L\underbrace{Ry}_{=:z} = Pb, \;\; x = Qy,$$

wobei $Q = I$ bei Spaltenpivotsuche. Man erhält nun x durch folgende Schritte:

Vorwärts-Rückwärtssubstitution für Dreieckszerlegung:

Löse $Lz = Pb$ nach z durch Vorwärtssubstitution gemäß Satz 4.2.1.

Löse $Ry = z$ nach y durch Rückwärtssubstitution gemäß Satz 4.2.1.

Lösung: $x = Qy$.

Liegt also die Dreieckszerlegung vor, dann kann (4.1) für jede rechte Seite in $O(n^2)$ Operationen berechnet werden. □

Wir überlegen uns nun, dass das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche eine Dreieckszerlegung (4.6) liefert. Die vollständige Pivotsuche kann analog untersucht werden.

Matrixdarstellung der Eliminationsschritte

Wir betrachten das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche (Algorithmus 1). Formal läßt sich der Übergang $(A^{(k)}, b^{(k)}) \to (\tilde{A}^{(k)}, \tilde{b}^{(k)}) \to (A^{(k+1)}, b^{(k+1)})$ durch Multiplikation mit Matrizen darstellen. Tatsächlich gilt

$$\begin{aligned} (\tilde{A}^{(k)}, \tilde{b}^{(k)}) &= P_k(A^{(k)}, b^{(k)}) \quad (Zeilenvertauschung) \\ (A^{(k+1)}, b^{(k+1)}) = L_k(\tilde{A}^{(k)}, \tilde{b}^{(k)}) &= L_k P_k(A^{(k)}, b^{(k)}) \quad (Elimination) \end{aligned}$$

mit der *elementaren Permutationsmatrix* (vertausche Zeile k und r der Einheitsmatrix)

$$P_k = \begin{array}{r} \\ \\ \\ k \to \\ \\ \\ \\ r \to \\ \\ \\ \end{array} \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & & & & 1 & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & 1 & & & & 0 & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \end{pmatrix} \tag{4.9}$$

und der *elementaren Eliminationsmatrix*

$$L_k = \begin{pmatrix} 1 & & & 0 & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{k+1,k} & 1 & & \\ 0 & & \vdots & & \ddots & \\ & & -l_{nk} & & 0 & 1 \end{pmatrix} \tag{4.10}$$

Man prüft leicht, dass P_k, L_k invertierbar sind mit

$$(4.11) \qquad P_k^{-1} = P_k, \quad L_k^{-1} = \begin{pmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ & & l_{k+1,k} & 1 & & \\ 0 & & \vdots & & \ddots & \\ & & l_{nk} & & 0 & 1 \end{pmatrix}$$

Sei nun die Ausgangsmatrix $A = A^{(1)}$ invertierbar. Dann ist also nach jedem Schritt (Pivotsuche + Elimination) das Ergebnis $A^{(k+1)} = L_k P_k A^{(k)}$ wieder invertierbar und wir erhalten, wie wir uns schon überlegt haben, die Struktur

$$(4.12) \quad (A^{(k+1)}, b^{(k+1)}) = L_k P_k (A^{(k)}, b^{(k)}) = \left(\begin{array}{ccc|cc|c} r_{11} & \cdots & r_{1k} & \cdots & r_{1n} & c_1 \\ 0 & \ddots & & & \vdots & \vdots \\ & & r_{kk} & \cdots & r_{kn} & c_k \\ \hline & & 0 & & & \\ & & \vdots & \hat{A}^{(k+1)} & & \hat{b}^{(k+1)} \\ & & 0 & & & \end{array}\right).$$

Nach den $n-1$ Schritten des Gaußschen Algorithmus erhalten wir somit

$$R = A^{(n)} = L_{n-1} P_{n-1} \cdots L_1 P_1 A.$$

Multiplikation mit

$$(L_{n-1} P_{n-1} \cdots L_1 P_1)^{-1} = P_1^{-1} L_1^{-1} \cdots P_{n-1}^{-1} L_{n-1}^{-1}$$

ergibt wegen $P_k^{-1} = P_k$

$$A = P_1 L_1^{-1} \cdots P_{n-1} L_{n-1}^{-1} R.$$

Sind im Eliminationsverfahren keine Zeilenvertauschungen nötig, dann erhalten wir

$$A = L_1^{-1} \cdots L_{n-1}^{-1} R =: LR.$$

und man rechnet mit (4.11) leicht nach, dass gilt

$$L = L_1^{-1} \cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & 0 & & \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \vdots & & \ddots & \ddots & \\ l_{n1} & & \cdots & l_{n,n-1} & 1 \end{pmatrix} = I + L^{(n)}.$$

Mit den vom Gauß-Verfahren gelieferten Matrizen L und R gilt also ohne Pivotsuche

$$A = LR.$$

Allgemein gilt der folgende Satz.

Satz 4.2.2 *Es sei $A \in \mathbb{R}^{n,n}$ nichtsingulär. Dann gilt:*

i) Das Gaußsche Eliminationsverfahren aus Algorithmus 1 liefert eine untere Dreiecksmatrix L der Form (4.7) *und eine obere Dreiecksmatrix R mit*

$$LR = PA.$$

Hierbei ist $P = P_{n-1} \cdots P_1$ eine Permutationsmatrix, wobei jeweils P_k die Permutationsmatrix für die Zeilenvertauschung im k-ten Schritt ist.

ii) Algorithmus 2 liefert eine Dreieckszerlegung

$$LR = PAQ$$

Hierbei ist P wie eben und $Q = Q_1 \cdots Q_{n-1}$, wobei jeweils Q_k die Permutationsmatrix für die Spaltenvertauschung im k-ten Schritt ist.

Beweis: Finden keine Zeilenvertauschungen statt, dann haben wir die Behauptung bereits gezeigt.

Für Interessierte: Der allgemeine Fall ist etwas technisch. Setze $C_k = L_k^{-1} - I$. Man prüft leicht die Gültigkeit der Formel

$$L_k^{-1} P_i = P_i (P_i C_k + I), \quad i \geq k.$$

Dies ergibt

$$P_1 L_1^{-1} \cdots P_{n-1} L_{n-1}^{-1} = P_1 \cdots P_{n-1} (P_{n-1} \cdots P_2 C_1 + I) \cdots (P_{n-1} C_{n-2} + I) L_{n-1}^{-1} = P^{-1} L,$$

wobei die Faktoren

$$\tilde{L}_k = P_{n-1} \cdots P_{k+1} C_k + I$$

aus L_k^{-1} durch Permutation der Einträge l_{ik} entstehen, die sich aus den nachfolgenden Pivotschritten ergeben. Dieselben Einträge l_{ik} werden vom Gauß-Verfahren in $L^{(k+1}$ eingetragen und bis zum Ergebnis $L^{(n)}$ genauso vertauscht.

Algorithmus 2 ist nichts anderes als Algorithmus 1 angewendet auf AQ. $\square$

Es gibt einige wichtige Teilklassen von Matrizen, bei denen auf die Pivotsuche verzichtet werden kann:

Matrizenklassen, die keine Pivotsuche erfordern

- $A = A^T$ ist symmetrisch posititv definit, also

$$x^T A x > 0 \quad \forall\, x \in \mathbb{R}^n \setminus \{0\}.$$

  Wir gehen hierauf noch ein.

- A ist strikt diagonaldominant, d.h.

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

  siehe Übung.

- A ist M-Matrix, d.h. es gilt

$$\begin{aligned}
&a_{ii} > 0, \;\; i = 1, \ldots, n,\\
&a_{ij} \leq 0, \;\; i \neq j\\
&D^{-1}(A - D), \;\; D = \operatorname{diag}(a_{11}, \ldots, a_{nn}) \;\text{ hat lauter Eigenwerte vom Betrag} < 1.
\end{aligned}$$

Nach Satz 4.2.2 gibt es für eine invertierbare Matrix A immer eine Permutationsmatrix P, so dass eine Dreieckszerlegung

$$LR = PA$$

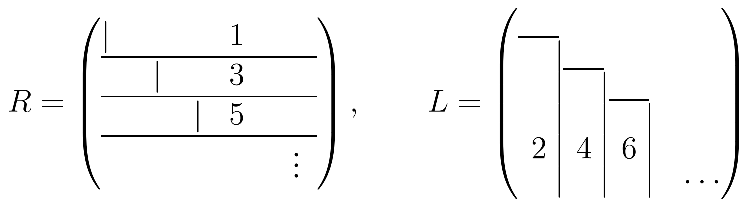
mit L, R der Form (4.7), (4.4) existiert. Tatsächlich sind L, R eindeutig bestimmt:

Satz 4.2.3 *Sei $A \in \mathbb{R}^{n,n}$ invertierbar und sei $P \in \mathbb{R}^{n,n}$ eine Permutationsmatrix, so dass eine Dreieckszerlegung* (4.6) *mit L, R der Form* (4.7), (4.4) *existiert. Dann sind L und R eindeutig bestimmt.*

Beweis: **Für Interessierte:** Die Beziehung $LR = B$ liefert die folgenden n^2 Gleichungen für die n^2 unbekannten Einträge in L und R

$$b_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik} r_{kj}, \quad (l_{ii} = 1).$$

Hieraus lassen sich l_{ik} und r_{kj} zum Beispiel in folgender Reihenfolge berechnen:



□

4.2.5 Das Cholesky-Verfahren

Für allgemeine invertierbare Matrizen kann das Gauß-Verfahren ohne Pivotsuche zusammenbrechen und wir werden sehen, dass auch aus Gründen der numerischen Stabilität eine Pivotsuche ratsam ist. Für die wichtige Klasse der positiv definiten Matrizen ist jedoch das Gauß-Verfahren immer ohne Pivotsuche numerisch stabil durchführbar.

Definition 4.2.4 *Eine reelle Matrix $A \in \mathbb{R}^{n,n}$ heißt positiv definit, falls gilt*

$$A = A^T, \quad x^T Ax > 0 \quad \forall\, x \in \mathbb{R}^n \setminus \{0\}.$$

und positiv semi-definit, falls gilt

$$A = A^T, \quad x^T Ax \geq 0 \quad \forall\, x \in \mathbb{R}^n.$$

Allgemeiner heißt eine komplexe Matrix $A \in \mathbb{C}^{n,n}$ positiv definit, falls gilt

$$A = A^H, \quad x^H Ax > 0 \quad \forall\, x \in \mathbb{C}^n \setminus \{0\}.$$

und positiv semi-definit, falls gilt

$$A = A^H, \quad x^H Ax \geq 0 \quad \forall\, x \in \mathbb{C}^n.$$

Hierbei ist $A^H = (\bar{a}_{ji})_{1 \leq i \leq n, 1 \leq j \leq n}$, wobei $\bar{a}_{ji}$ komplexe Konjugation bezeichnet.

Positiv definite Matrizen treten sehr oft in Anwendungen auf, etwa bei der numerischen Lösung von elliptischen (z.B. Laplace-Gleichung) und parabolischen (z.B. Wärmeleitungsgleichung) partiellen Differentialgleichungen.

Satz 4.2.5 *Für eine positiv definite Matrix A existiert A^{-1} und ist wieder positiv definit. Zudem gilt: Alle Eigenwerte sind positiv, alle Hauptuntermatrizen $A_{kl} = (a_{ij})_{k \leq i,j \leq l}$, $1 \leq k \leq l \leq n$ sind wieder positiv definit und alle Hauptminoren $\det(A_{kl})$ sind positiv.*

Beweis: Elementare Übung aus der linearen Algebra. Siehe z.B. Stoer [St94]. □

Eine effiziente Variante des Gaußschen Verfahrens für Gleichungssysteme mit positiv definiter Matrix wurde von Cholesky angegeben. Das Cholesky-Verfahren beruht auf der folgenden Beobachtung

Satz 4.2.6 *Es sei $A \in \mathbb{R}^{n,n}$ positiv definit. Dann gibt es genau eine untere Dreieckmatrix L mit positiven Diagonaleinträgen $l_{ii} > 0$, so dass*

$$LL^T = A \quad \textit{(Cholesky-Zerlegung)}.$$

Ferner besitzt A eine eindeutige Dreieckszerlegung

$$\tilde{L}\tilde{R} = A,$$

wobei $\tilde{L} = LD^{-1}$, $\tilde{R} = DL^T$ mit $D = \operatorname{diag}(l_{11}, \ldots, l_{nn})$. Sie wird vom Gauß-Verfahren ohne Pivotsuche geliefert.

Der Beweis kann durch vollständige Induktion nach n erfolgen, wir wollen ihn aber nicht ausführen.

Die Cholesky-Zerlegung $LL^T = A$ berechnet man durch Auflösen der $\frac{n(n+1)}{2}$ Gleichungen (aus Symmetriegründen muss nur das untere Dreieck mit Diagonale betrachtet werden)

(4.13) $$a_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk}, \quad \text{für } j \le i, \; i = 1, \dots, n.$$

Man kann hieraus die Elemente von L spaltenweise in der Reihenfolge

$$l_{11}, \dots, l_{n1}, l_{22}, \dots, l_{n2}, \dots, l_{nn}$$

berechnen. Für die erste Spalte von L (setze $j = 1$) ergibt sich

$$\begin{aligned} a_{11} &= l_{11}^2, \text{also } \; l_{11} = \sqrt{a_{11}} \\ a_{i1} &= l_{i1} l_{11}, \text{also } \; l_{i1} = a_{i1}/l_{11}. \end{aligned}$$

Sukzessives Auflösen nach l_{ij}, $i = j, \dots, n$ liefert den folgenden Algorithmus.

Algorithmus 3 Cholesky-Verfahren zur Berechnung der Zerlegung $LL^T = A$

Für $j = 1, \dots, n$

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$$

Für $i = j + 1, \dots, n$*:*

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}}$$

Bemerkung: Das Cholesky-Verfahren hat einige schöne Eigenschaften:

- Da das Cholesky-Verfahren die Symmetrie ausnutzt, benötigt es neben n Quadratwurzeln nur noch ca. $n^3/6$ Operationen. Dies ist etwa die Hälfte der beim Gauß-Verfahren benötigten Operationen.
- Aus (4.13) folgt
  $$|l_{ij}| \le \sqrt{a_{ii}}, \quad j \le i, \; i = 1, \dots, n.$$
  Die Elemente der Matrix L können daher nicht zu groß werden. Dies ist ein wesentlicher Grund für die numerische Stabilität des Cholesky-Verfahrens.
- Das Cholesky-Verfahren ist die effizienteste allgemeine Testmethode auf positive Definitheit. Man muss hierbei Algorithmus 3 nur wie folgt erweitern:
  $$a = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2. \qquad \textit{Falls } a \le 0\textit{: STOP, } A \textit{ nicht positiv definit.}$$
  Sonst setze $l_{jj} = \sqrt{a}$.

4.3 Fehlerabschätzungen und Rundungsfehlereinfluß

Bei der Beschreibung der direkten Verfahren zur Lösung von Gleichungssystemen sind wir bisher davon ausgegangen, dass alle Ausgangsdaten exakt vorliegen und die Rechnung ohne Rundungsfehler durchgeführt wird. Dies ist jedoch unrealistisch, denn insbesondere bei großen Systemen können Rundungsfehler die Rechnung erheblich beeinflussen.

4.3.1 Fehlerabschätzungen für gestörte Gleichungssysteme

Wir studieren zunächst, wie stark sich die Lösung eines linearen Gleichungssystems bei Störung von Matrix und rechter Seite ändert. Vorgelegt sei ein lineares Gleichungssystem

$$Ax = b$$

und ein gestörtes System

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

mit ΔA und Δb ”klein”.

Frage: Wie klein ist $x - \tilde{x}$?

Diese Fragestellung ist von größter praktischer Bedeutung:

- Man kann abschätzen, wie sensitiv die Lösung bezüglich Störungen von Matrix und rechter Seite ist.
- Eine berechnete Näherungslösung (z.B. mit einer Implementierung des Gauß-Verfahrens) $\tilde{x}$ von $Ax = b$ ist exakte Lösung des Systems

  $$A\tilde{x} = b + \Delta b, \quad \text{mit dem Residuum } \Delta b = A\tilde{x} - b.$$

  Man kann nun aus dem leicht berechenbaren Residuum $\Delta b = A\tilde{x} - b$ Schranken an den unbekannten Fehler $\|x - \tilde{x}\|$ ableiten.

Es stellt sich heraus, dass die sogenannte Kondition einer Matrix diesen Störeinfluss beschreibt.

Zur Messung von $x - \tilde{x}$, Δb und ΔA benötigen wir einen ”Längenbegriff” für Vektoren und Matrizen.

Definition 4.3.1 *Eine* Vektornorm *auf* $\mathbb{R}^n$ *ist eine Abbildung* $x \in \mathbb{R}^n \mapsto \|x\| \in [0, \infty[$ *mit folgenden Eigenschaften:*

a) $\|x\| = 0$ *nur für* $x = 0$.

b) $\|\alpha x\| = |\alpha|\,\|x\|$ *für alle* $\alpha \in \mathbb{R}$ *und alle* $x \in \mathbb{R}^n$.

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

Nun sollen auch *Matrix-Normen* eingeführt werden. Sei hierzu $\|\cdot\|$ eine beliebige Norm auf $\mathbb{R}^n$. Dann können wir auf $\mathbb{R}^{n,n}$ eine zugehörige Matrix-Norm definieren durch

$$\|A\| := \sup_{\|x\|=1} \|Ax\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{4.14}$$

für $A \in \mathbb{R}^{n,n}$. Sie heißt die durch die *Vektornorm* $\|\cdot\|$ *indizierte Matrix-Norm.*

Sie hat wiederum die Eigenschaften

$\|A\| = 0$ nur für $A = 0$.

$\|\alpha A\| = |\alpha|\,\|A\|$ für alle $\alpha \in \mathbb{R}$ und alle $A \in \mathbb{R}^{n,n}$.

$\|A + B\| \leq \|A\| + \|B\|$ für alle $A, B \in \mathbb{R}^{n,n}$ (Dreiecksungleichung).

Zusätzlich sichert (4.14) die nützlichen Ungleichungen

$\|Ax\| \leq \|A\|\,\|x\|$ für alle $x \in \mathbb{R}^n$ und alle $A \in \mathbb{R}^{n,n}$ (Verträglichkeitsbedingung)

$\|AB\| \leq \|A\|\,\|B\|$ für alle $A, B \in \mathbb{R}^{n,n}$ (Submultiplikativität).

Beispiele:

$$\|x\|_2 = \sqrt{x^T x} \text{ induziert } \|A\|_2 = \sqrt{\lambda_{max}(A^T A)}$$

$$\|x\|_1 = \sum_{i=1}^n |x_i| \text{ induziert } \|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}| \quad \text{(Spaltensummennorm)}$$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i| \text{ induziert } \|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}| \quad \text{(Zeilensummennorm)}$$

Wir sind nun in der Lage, die bereits erwähnte Kondition einer Matrix einzuführen.

Definition 4.3.2 *Sei* $A \in \mathbb{R}^{n,n}$ *invertierbar und sei* $\|\cdot\|$ *eine induzierte Matrixnorm. Dann heißt die Zahl*

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

die Kondition *von* A *bezüglich der Matrixnorm.*

Man kann nun folgendes zeigen.

Satz 4.3.3 (Störeinfluss von Matrix und rechter Seite)
Sei $A \in \mathbb{R}^{n,n}$ invertierbar, $b, \Delta b \in \mathbb{R}^n$, $b \neq 0$ und $\Delta A \in \mathbb{R}^{n,n}$ mit $\|\Delta A\| < 1/\|A^{-1}\|$ mit einer beliebigen durch eine Norm $\|\cdot\|$ auf $\mathbb{R}^n$ induzierten Matrixnorm $\|\cdot\|$. Ist x die Lösung von

$$Ax = b$$

und $\tilde{x}$ die Lösung von

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

dann gilt

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

Beweis: Wir betrachten der Einfachheit halber nur den Fall $\Delta A = 0$. Dann liefert Subtraktion der gestörten und ungestörten Gleichung

$$A(\tilde{x} - x) = \Delta b,$$

also

$$\|\tilde{x} - x\| = \|A^{-1}\Delta b\| \leq \|A^{-1}\| \|\Delta b\|.$$

Wegen $\|b\| = \|Ax\| \leq \|A\| \|x\|$ folgt $1/\|x\| \leq \|A\|/\|b\|$ und somit

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta b\|}{\|b\|}.$$

□

Die Kondition bestimmt also die Sensitivität bezüglich Störungen von Matrix und rechter Seite.

4.3.2 Ergänzung: Rundungsfehlereinfluß beim Gauß-Verfahren

Auf einem Rechner wird eine Zahl $\neq 0$ nach IEEE-Standard dargestellt in der Form

$$\pm 1, \alpha_1\, \alpha_2\, \ldots\, \alpha_{t-1} \cdot 2^b, \quad \alpha_i \in \{0, 1\}, b \in \{b_-, \ldots, b_+\},$$

z.B. bei der heute üblichen doppelten Genauigkeit

$$t = 53 \text{ (ca. 15 Dezimalstellen)},\ b_- = -1022,\ b_+ = 1023.$$

Alle elementaren Rechenoperationen sind nach IEEE-Standard so zu implementieren, dass das Ergebnis der Operation das gerundete exakte Ergebnis ist, ausser bei Exponenten-Über- oder Unterlauf. Bezeichnen wir mit $+_g$, $-_g$, usw. die Rechenoperationen auf einem Rechner, dann gilt also z.B.

$$x +_g y = \mathrm{rd}(x + y).$$

Hierbei rundet rd zur nächstgelegenen Gleitpunktzahl. Es gilt für den relativen Fehler

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \le 2^{-t} =: \text{eps}, \quad \text{eps: Maschinengenauigkeit.}$$

Somit gilt bei jeder Rechenoperation $\circ_g \in \{+_g, -_g, *_g, /_g\}$

$$x \circ_g y = (x \circ y)(1 + \varepsilon), \quad |\varepsilon| \le \text{eps}.$$

Rundungsfehleranalyse für das Gauß-Verfahren

Durch eine elementare aber aufwendige Abschätzung der beim Gauß-Verfahren auftretenden Rundungsfehlerverstärkung erhält man folgendes Resultat.

Satz 4.3.4 *Sei $A \in \mathbb{R}^{n,n}$ invertierbar. Wendet man das Gauß-Verfahren auf einem Rechner mit Maschinengenauigkeit eps mit einer Pivot-Technik an, die $|l_{ij}| \le 1$ sicherstellt (z.B. Spaltenpivotsuche oder totale Pivotsuche), dann errechnet man $\bar{L}$, $\bar{R}$ mit*

$$\bar{L}\bar{R} = PAQ + F, \quad |f_{ij}| \le 2j\bar{a}\frac{eps}{1 - eps}.$$

Hierbei sind P, Q die aus der Pivotsuche resultierenden Permutationen und

$$\bar{a} = \max_k \bar{a}_k, \quad \bar{a}_k = \max_{i,j} |a_{ij}^{(k)}|. \tag{4.15}$$

Berechnet man mit Hilfe von $\bar{L}$, $\bar{R}$ durch Vorwärts- und Rückwärtssubstitution eine Näherungslösung $\bar{x}$ von $Ax = b$, dann existiert eine Matrix E mit

$$(A + E)\bar{x} = b, \quad |e_{ij}| \le \frac{2(n+1)eps}{1 - neps}(|\bar{L}||\bar{R}|)_{ij} \le \frac{2(n+1)eps}{1 - neps} n\bar{a}.$$

Hierbei bezeichnet $|\bar{L}| = (|\bar{l}_{ij}|)$, $|\bar{R}| = (|\bar{r}_{ij}|)$.

Beweis: Siehe Stoer [St94]. □

Bemerkung: Mit Satz 4.3.3 kann man nun auch den relativen Fehler der Näherungslösung $\bar{x}$ abschätzen. □

Einfluß der Pivot-Strategie

Die Größe von $\bar{a}$ in (4.15) hängt von der Pivotstrategie ab. Man kann folgendes zeigen:

- **Spaltenpivotsuche:** $\bar{a}_k \le 2^k \max_{i,j} |a_{ij}|$.

  Diese Schranke kann erreicht werden, ist aber in der Regel viel zu pessimistisch. In der Praxis tritt fast immer $\bar{a}_k \le 10 \max_{i,j} |a_{ij}|$ auf.

- **Spaltenpivotsuche bei Tridiagonalmatrizen:** $\bar{a}_k \leq 2 \max_{i,j} |a_{ij}|$.
- **Vollständige Pivotsuche:** $\bar{a}_k \leq f(k) \max_{i,j} |a_{ij}|, \quad f(k) = k^{1/2}(2^1 3^{1/2} \cdots k^{1/(k-1)})^{1/2}$.

  $f(n)$ wächst recht langsam. Es ist bislang kein Beispiel mit $\bar{a}_k \geq (k+1) \max_{i,j} |a_{ij}|$ entdeckt worden.

Kapitel 5

Nichtlineare Gleichungssysteme

5.1 Einführung

Wir betrachten in diesem Kapitel Verfahren zur Lösung von nichtlinearen Gleichungssystemen.

Nichtlineares Gleichungssystem: Gesucht ist eine Lösung $x \in D$ von

$$F(x) = 0$$

mit einer gegebenen Abbildung

$$F = \begin{pmatrix} F_1 \\ \vdots \\ F_n \end{pmatrix} : D \subset \mathbb{R}^n \to \mathbb{R}^n,$$

$D \subset \mathbb{R}^n$ nichtleer und abgeschlossen.

Viele praxisrelevante Probleme, insbesondere im Hochtechnologiebereich, sind nichtlinear und erfordern die Lösung nichtlinearer Gleichungssysteme. So führt zum Beispiel die Schaltkreissimulation und die Diskretisierung nichtlinearer partieller Differentialgleichungen (Wetter- und Klimamodelle, strukturmechanische Berechnungen, Umformprozesse aus der Produktionstechnik...) auf große nichtlineare Gleichungssysteme.

Im Gegensatz zu linearen Gleichungssystemen, bei denen nur genau eine Lösung, keine Lösung oder ein ganzer affiner Unterraum als Lösung auftreten kann, sind bei nichtlinearen Gleichungen auch mehrere oder unendlich viele isolierte Lösungen möglich.

Beispiel 5.1.1

1. $n = 1$, $D = \mathbb{R}$, $F(x) = x^2 - a$, $a > 0$.
   Es gibt zwei reelle Lösungen $x = \pm\sqrt{a}$.

2. $n = 1$, $D = \mathbb{R}$, $F(x) = x^2 + a$, $a > 0$.

   Es existiert keine reelle Lösung.

3. $n = 1$, $D = \mathbb{R}$, $F(x) = x\sin(x)$.

   Es gibt unendlich viele Lösungen $x = k\pi$, $k \in \mathbb{Z}$.

4. Schnittpunkte des Einheitskreises mit der Geraden $G : x_2 = ax_1 + b$, $a, b \in \mathbb{R}$: $n = 2$, $D = \mathbb{R}^2$, $F(x) = \binom{x_1^2+x_2^2-1}{x_2-ax_1-b}$.

   Je nach Wahl von a, b gibt es zwei, eine oder keine relle Lösung.

Sehr oft ist die Funktion F stetig differenzierbar, d.h. die partiellen Ableitungen $\frac{\partial F_i}{\partial x_j}$, $1 \le i, j \le n$ existieren und sind stetig. In diesem Fall gilt (Taylorentwicklung erster Ordnung)

$$F(x + s) = F(x) + F'(x)s + R(x; s)$$

mit der Jacobi-Matrix

$$F'(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \cdots & \frac{\partial F_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial F_n}{\partial x_1}(x) & \cdots & \frac{\partial F_n}{\partial x_n}(x) \end{pmatrix}.$$

und einem Restglied $R(x; s)$, wobei

$$\lim_{s\to 0} \frac{\|R(x; s)\|}{\|s\|} = 0, \quad \text{kurz:} \quad R(x; s) = o(\|s\|).$$

Dies ist wesentlich für die Entwicklung schneller Lösungsverfahren.

5.2 Das Newton-Verfahren

Das Newton-Verfahren ist eines der wichtigsten Verfahren zur Lösung nichtlinearer Gleichungssysteme, da es nahe der Lösung sehr schnell konvergiert. Der Einfachheit halber nehmen wir im folgenden den Fall $D = \mathbb{R}^n$ an.

Wir betrachten das Newton-Verfahren zur Lösung eines nichtlinearen Gleichungssystems

$$F(x) = 0 \tag{5.1}$$

mit $F : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar.

5.2.1 Herleitung des Verfahrens

Anschauliche Herleitung im eindimensionalen Fall

Sei zunächst $n = 1$. Dann ist $F(x)$ eine reelle Funktion. Sei $x^{(k)}$ eine Näherung einer Lösung $\bar{x}$ von (5.1). Die Idee des Newton-Verfahrens besteht darin, F in $x^{(k)}$ durch die Tangente an $(x, F(x))$ im Punkt $x^{(k)}$ zu approximieren und als nächste Iterierte $x^{(k+1)}$ die Nullstelle der Tangente zu wählen.

Die Tangentengleichung lautet

$$y = F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)})$$

und $x^{(k+1)}$ ist die Lösung von

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

Im Falle $F'(x^{(k)}) \neq 0$ ergibt sich

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

Es gilt also

$$x^{(k+1)} = x^{(k)} + s^{(k)},$$

wobei $s^{(k)}$ die Lösung der Gleichung ist

$$F'(x^{(k)}) s^{(k)} = -F(x^{(k)}).$$

Beispiel: Für $F(x) = x^2 - a$, $a > 0$ ergibt sich

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

Der allgemeine Fall

Zur allgemeinen Motivation des Newton-Verfahrens für (5.1) sei $x^{(k)} \in \mathbb{R}^n$ ein gegebener Punkt. Dann ist $\bar{x}$ eine Lösung von (5.1) genau dann, wenn $\bar{x} = x^{(k)} + s$ gilt mit einer Lösung s von

$$F(x^{(k)} + s) = 0. \tag{5.2}$$

Die Idee des Newton-Verfahrens besteht darin, $F(x^{(k)} + s)$ durch die Taylorentwicklung erster Ordnung zu ersetzen: Es gilt

$$F(x^{(k)} + s) = F(x^{(k)}) + F'(x^{(k)})s + o(\|s\|)$$

mit der Jacobi-Matrix $F'(x^{(k)})$ von F in $x^{(k)}$ und das Restglied wird für kurze s klein.

Bei der k-ten Iteration des Newton-Verfahrens ersetzt man daher (5.2) durch die linearisierte Gleichung

$$F(x^{(k)}) + F'(x^{(k)})s = 0.$$

Dies ergibt

Algorithmus 4 Lokales Newton-Verfahren für Gleichungssysteme

Wähle einen Startpunkt $x^{(0)} \in \mathbb{R}^n$.

Für $k = 0, 1, \dots$:

1. *Falls* $F(x^{(k)}) = 0$*: STOP mit Ergebnis* $x^{(k)}$.
2. *Berechne den Newton-Schritt* $s^{(k)} \in \mathbb{R}^n$ *durch Lösen der* Newton-Gleichung
$$F'(x^{(k)})s^{(k)} = -F(x^{(k)}).$$
3. *Setze* $x^{(k+1)} = x^{(k)} + s^{(k)}$.

5.2.2 Superlineare und quadratische lokale Konvergenz des Newton-Verfahrens

Wir werden sehen, dass unter geeigneten Voraussetzungen die schnelle lokale Konvergenz des Newton-Verfahrens gezeigt werden kann.

Wir verwenden im folgenden der Einfachheit halber immer die euklidische Norm $\|\cdot\|_2$ mit induzierter Matrix-Norm $\|\cdot\|_2$, obwohl wir genausogut jede andere Norm verwenden könnten.

Der folgende Satz zeigt die superlineare bzw. quadratische lokale Konvergenz des Newton-Verfahrens.

Satz 5.2.1 (Schnelle lokale Konvergenz des Newton-Verfahrens)
Sei $F : \mathbb{R}^n \to \mathbb{R}^n$ *stetig differenzierbar und sei* $\bar{x} \in \mathbb{R}^n$ *ein Punkt mit* $F(\bar{x}) = 0$ *und* $F'(\bar{x})$ *nichtsingulär. Dann gibt es* $\delta > 0$*, so dass gilt:*

i) $\bar{x}$ *ist die einzige Nullstelle von* F *auf* $B_\delta(\bar{x})$

ii) Für alle $x^{(0)} \in B_\delta(\bar{x})$ *mit der* δ*-Kugel*
$$B_\delta(\bar{x}) := \{x \in \mathbb{R}^n \ : \ \|x - \bar{x}\|_2 < \delta\}$$
terminiert Algorithmus 4 entweder mit $x^{(k)} = \bar{x}$ *oder erzeugt eine Folge* $(x^{(k)}) \subset B_\delta(\bar{x})$*, die superlinear gegen* $\bar{x}$ *konvergiert, d.h.*
$$\lim_{k\to\infty} x_k = \bar{x}, \quad \textit{wobei } \|x_{k+1} - \bar{x}\|_2 \le \nu_k \|x_k - \bar{x}\|_2$$
mit einer Nullfolge $\nu_k \searrow 0$.

iii) Ist F' Lipschitz-stetig auf $B_\delta(\bar{x})$ mit Konstante L, gilt also

$$\|F'(x) - F'(y)\|_2 \leq L\|x - y\|_2 \quad \forall\, x, y \in B_\delta(\bar{x}),$$

dann konvergiert $(x^{(k)})$ sogar quadratisch gegen $\bar{x}$, d.h.

$$\lim_{k\to\infty} x_k = \bar{x}, \quad \textit{wobei } \|x_{k+1} - \bar{x}\|_2 \leq C\|x_k - \bar{x}\|_2^2,$$

wobei für $\delta > 0$ klein genug $C = \|F'(\bar{x})^{-1}\|_2 L$ gewählt werden kann.

Hinweis: *F' ist automatisch Lipschitz-stetig auf $B_\delta(\bar{x})$, falls F zweimal stetig differenzierbar auf der abgeschlossenen Kugel $\overline{B_\delta(\bar{x})}$ ist.*

Leider konvergiert das Newton-Verfahren aus Algorithmus 4 in der Regel nur für Startpunkte, die nahe genug an einer Lösung $\bar{x}$ liegen.

Beispiel 5.2.1 Betrachte $F(x) = \frac{x}{\sqrt{1+x^2}}$. F hat die eindeutige Nullstelle $\bar{x}$ und ist stetig differenzierbar mit $F'(x) > 0$. Trotzdem konvergiert das Newton-Verfahren für jeden Startpunkt mit $|x^{(0)}| > 1$ nicht. Siehe Übung.

Um Konvergenz für beliebige Startpunkte erzielen zu können, muss man das Newton-Verfahren geeignet globalisieren.

5.2.3 Globalisierung des Newton-Verfahrens

In diesem Abschnitt beschreiben wir eine Modifikation des Newton-Verfahrens, die für eine große Klasse von Funktionen F globale Konvergenz, d.h. Konvergenz von einem beliebigen Startpunkt aus, sicherstellt.

Den Ausgangspunkt bildet die Beobachtung, dass jede Lösung $\bar{x}$ von (5.1) ein globales Minimum des Minimierungsproblems

$$\min_{x\in\mathbb{R}^n} \|F(x)\|_2^2$$

ist.

Wir wenden nun folgende Strategie an:

- Wir verwenden den Newton-Schritt $s^{(k)}$ mit einer Schrittweite $\sigma_k \in]0, 1]$, wählen also als Ansatz für die neue Iterierte

$$x^{(k+1)} = x^{(k)} + \sigma_k s^{(k)}.$$

- Wir bestimmen die Schrittweite σ_k so, dass gilt

  (5.3) $$\|F(x^{(k+1)})\|_2 < \|F(x^{(k)})\|_2,$$

  und die Abnahme "ausreichend groß" ist.

Durch Taylorentwicklung der Funktion

$$\phi(\sigma) := \|F(x^{(k)} + \sigma s^{(k)})\|_2^2$$

in $\sigma = 0$ erhält man

$$\phi(\sigma) = \phi(0) + \phi'(0)\sigma + o(\sigma) = \|F(x^{(k)})\|_2^2 + 2\sigma F(x^{(k)})^T F'(x^{(k)}) s^k + o(\sigma)$$

und Einsetzen der Newton-Gleichung $F'(x^{(k)})s^k = -F(x^{(k)})$ liefert

$$\|F(x^{(k)} + \sigma s^{(k)})\|_2^2 = \|F(x^{(k)})\|_2^2 - 2\sigma \|F(x^{(k)})\|_2^2 + o(\sigma).$$

Ist $\delta \in]0,1[$ fest, dann gilt im Fall $F(x^{(k)}) \neq 0$ also für σ klein genug

$$\|F(x^{(k)} + \sigma s^{(k)})\|_2^2 \le \|F(x^{(k)})\|_2^2 - 2\delta\sigma \|F(x^{(k)})\|_2^2.$$

Dies zeigt, dass die folgende Schrittweitenwahl nach Armijo Sinn macht:

Schrittweitenwahl nach Armijo:

Sei $\delta \in]0, 1/2[$ (gute Wahl z.B. $\delta = 10^{-3}$) fest gegeben. Wähle das größte $\sigma_k \in \{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ mit

(5.4) $$\|F(x^{(k)} + \sigma_k s^{(k)})\|_2^2 \le \|F(x^{(k)})\|_2^2 - 2\delta\sigma_k \|F(x^{(k)})\|_2^2.$$

Wir erhalten insgesamt folgendes Verfahren:

Algorithmus 5 Globalisiertes Newton-Verfahren für Gleichungssysteme

Wähle einen Startpunkt $x^{(0)} \in \mathbb{R}^n$.

Für $k = 0, 1, \ldots$:

1. *Falls $F(x^{(k)}) = 0$: STOP mit Ergebnis $x^{(k)}$.*
2. *Berechne den Newton-Schritt $s^{(k)} \in \mathbb{R}^n$ durch Lösen der* Newton-Gleichung

   $$F'(x^{(k)})s^{(k)} = -F(x^{(k)}).$$

3. *Bestimme σ_k nach der Armijo-Regel* (5.4).

4. Setze $x^{(k+1)} = x^{(k)} + \sigma_k s^{(k)}$.

Es gilt folgender Konvergenzsatz.

Satz 5.2.2 *Sei* $F : \mathbb{R}^n \to \mathbb{R}^n$ *stetig differenzierbar und* $x^{(0)} \in \mathbb{R}^n$ *beliebig. Ist* $F'(x)$ *invertierbar für alle* x *in der Niveaumenge*

$$N_f(x^{(0)}) := \left\{y \,:\, f(y) \le f(x^{(0)})\right\}, \quad f(x) = \|F(x)\|_2^2$$

und ist $N_f(x^{(0)})$ *kompakt (also beschränkt und abgeschlossen), dann terminiert Algorithmus 5 mit Startpunkt* $x^{(0)}$ *entweder endlich oder erzeugt eine Folge* $(x^{(k)}) \subset N_f(x^{(0)})$*, für die gilt:*

i) $(x^{(k)})$ *konvergiert gegen eine Lösung* $\bar{x}$ *von* (5.1).

ii) Es gibt $l \ge 0$ *mit* $\sigma_k = 1$ *für alle* $k \ge l$*. Das Verfahren geht also in das lokale Newton-Verfahren über und konvergiert superlinear bzw. quadratisch gegen* $\bar{x}$*.*

Kapitel 6

Numerische Behandlung von Anfangswertproblemen gewöhnlicher Differentialgleichungen

6.1 Einführung

Viele Anwendungen aus Naturwissenschaft, Technik und Wirtschaft führen auf Anfangswertprobleme für gewöhnliche Differentialgleichungen.

Anfangswertproblem: Gegeben sei eine Funktion $f : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$ und ein Anfangswert $y_0 \in \mathbb{R}^n$. Gesucht ist eine Funktion $y : [a, b] \to \mathbb{R}^n$, deren Ableitung y' eine *gewöhnlichen Differentialgleichung* der Form

$$y'(t) = f(t, y(t)), \quad t \in [a, b]$$

erfüllt und die zudem der *Anfangsbedingung*

$$y(a) = y_0$$

genügt. Also kurz

$$\text{(AWP)} \qquad \begin{aligned} & y'(t) = f(t, y(t)), \quad t \in [a, b] && (6.1) \\ & y(a) = y_0 && (6.2) \end{aligned}$$

In vielen Fällen bezeichnet t die Zeit, was die Bezeichnung Anfangswertproblem rechtfertigt.

Anwendungen: Bewegungsgleichungen (z.B. Fahrdynamik, Planetenbewegung), Reaktionskinetik, Schaltkreissimulation, etc.

Grundlegend für die Existenz und Eindeutigkeit einer Lösung von (AWP) ist der folgende

Satz 6.1.1 (Existenz- und Eindeutigkeitssatz)
$f : [a,b] \times \mathbb{R}^n \to \mathbb{R}^n$ sei stetig. Ferner gebe es eine feste Zahl $L > 0$ mit

$$\|f(t,y) - f(t,z)\| \leq L\|y-z\| \quad \textit{für alle } t \in [a,b] \textit{ und } y,z \in \mathbb{R}^n \quad \textit{(Lipschitz-Bedingung).}$$

Dann gilt:

a) *(Picard/Lindelöf) Zu jedem $y_0 \in \mathbb{R}^n$ besitzt (AWP) genau eine Lösung $y \in C^1([a,b];\mathbb{R}^n)$.*

b) *Sind y, z Lösungen zu den Anfangswerten $y(a) = y_0$ bzw. $z(a) = z_0$, dann gilt*

$$\|y(t) - z(t)\| \leq e^{L(t-a)}\|y_0 - z_0\| \quad \forall\, t \in [a,b]. \tag{6.3}$$

Für einen Beweis siehe z.B. Heuser [Heu89] oder Walter [Wa86]. Teil b) ist eine Folge des Lemmas von Gronwall.

Bemerkung: Teil b) besagt, dass die Lösung *stetig* vom Anfangswert y_0 abhängt.

6.1.1 Grundkonzept numerischer Verfahren

Zur numerischen Lösung von (AWP) zerlegen wir das Intervall $[a,b]$ in Teilintervalle:

$$t_j = a + jh, \quad j = 0,1,\ldots,N, \quad h = \frac{b-a}{N}.$$

Durch Integration von (AWP) erhält man mit der Abkürzung $y_j = y(t_j)$

$$y_{j+1} = y_j + \int_{t_j}^{t_{j+1}} y'(t)\,dt = y_j + \int_{t_j}^{t_{j+1}} f(t,y(t))\,dt. \tag{6.4}$$

Das Integral rechts kann nicht exakt berechnet werden, da $y(t)$ unbekannt ist. Wir approximieren daher das Integral durch interpolatorische Quadratur und erhalten hieraus einen numerischen Algorithmus zur Berechnungen von Näherungen

$$u_j \approx y(t_j), \quad j = 1,\ldots,N, \quad u_0 = y_0.$$

Den Fehler

$$e_j = y(t_j) - u_j$$

bezeichnet man als *Diskretisierungsfehler*.

6.1.2 Einige wichtige Verfahren

Approximiert man das Integral in (6.4) durch die Rechtecksregel, wobei wir das linke Intervallende als Stützpunkt verwenden, also

$$\int_{t_j}^{t_{j+1}} f(t, y(t))\, dt \approx hf(t_j, y_j),$$

dann erhalten wir das

Explizite Euler-Verfahren:

$$\begin{aligned} u_0 &:= y_0 \\ u_{j+1} &:= u_j + hf(t_j, u_j), \quad j = 0, \ldots, N-1. \end{aligned} \tag{6.5}$$

Verwenden wir zur Approximation des Integrals die Rechtecksregel mit dem rechten Randpunkt t_{j+1} als Stützstelle, dann erhalten wir das

Implizite Euler-Verfahren:

$$\begin{aligned} u_0 &:= y_0 \\ u_{j+1} &:= u_j + hf(t_{j+1}, u_{j+1}), \quad j = 0, \ldots, N-1. \end{aligned} \tag{6.6}$$

Hierbei ist zu beachten, dass für jedes j die Gleichung nach u_{j+1} aufgelöst werden muss.

Approximiert man das Integral in (6.4) durch die Trapezregel, dann erhält man

$$u_{j+1} = u_j + \frac{h}{2}\left(f(t_j, u_j) + f(t_{j+1}, u_{j+1})\right).$$

Die rechte Seite hängt von u_{j+1} ab, das Verfahren ist also implizit. Ersetzt man rechts u_{j+1} durch den expliziten Euler-Schritt $u_{j+1} = u_j + hf(t_j, u_j)$, dann ergibt sich das

Verfahren von Heun, erstes Runge-Kutta-Verfahren 2. Ordnung: (Heun, 1900)

$$u_0 = y_0, \quad u_{j+1} = u_j + \frac{h}{2}\left(f(t_j, u_j) + f(t_{j+1}, u_j + hf(t_j, u_j))\right), j = 0, \ldots, N-1.$$

Das Verfahren kann auch geschrieben werden als

$$u_{j+1} = u_j + \frac{h}{2}(k_1 + k_2)$$

mit $k_1 = f(t_j, u_j)$, $k_2 = f(t_{j+1}, u_j + hk_1)$.

Approximieren wir das Integral durch die Mittelpunktsregel und $u_{j+1/2}$ durch den Euler-Schritt $u_j + h/2 f(t_j, u_j)$, dann ergibt sich das

Modifizierte Euler-Verfahren, zweites Runge-Kutta-Verfahren 2. Ordnung: (Runge, 1895)

$$u_0 = y_0, \quad u_{j+1} = u_j + hf(t_j + h/2, u_j + (h/2)\, f(t_j, u_j)), \quad j = 0, \dots, N-1.$$

Das Verfahren kann auch geschrieben werden als

$$u_{j+1} = u_j + hk_2$$

mit $k_1 = f(t_j, u_j)$, $k_2 = f(t_j + h/2, u_j + (h/2)\, k_1)$.

Wenden wir schließlich die Simpson-Regel an und ersetzen $u_{j+1/2}$, u_{j+1} geeignet durch Taylorentwicklungen, dann ergibt sich das sehr genaue und beliebte

Klassische Runge-Kutta-Verfahren 4. Ordnung (RK4)

$$\begin{aligned}
u_0 &= y_0 \\
u_{j+1} &= u_j + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad j = 0, \dots, N-1 \\
\text{mit } \; k_1 &= f(t_j, u_j) \\
k_2 &= f(t_j + h/2, u_j + (h/2)\, k_1) \\
k_3 &= f(t_j + h/2, u_j + (h/2)\, k_2) \\
k_4 &= f(t_{j+1}, u_j + hk_3)
\end{aligned}$$

6.1.3 Konvergenz und Konsistenz

Wir wollen nun die vorgestellten Verfahren auf ihre praktische Brauchbarkeit und Genauigkeit hin untersuchen. Die Verfahren lassen sich in der allgemeinen Form

$$\begin{aligned}
u_0 &= y_0 \\
u_{j+1} &= u_j + h\, \phi(t_j, h; u_j, u_{j+1}), \quad j = 0, \dots, N-1,
\end{aligned} \tag{6.7}$$

schreiben,

Definition 6.1.2 *Die Funktion $\phi(t, h; u, v)$ in* (6.7) *heißt* Verfahrensfunktion. *Hängt ϕ nicht von v ab, dann heißt das Verfahren* explizit, *sonst* implizit.

Die Größe

$$\begin{aligned}
\tau(t, h) &= \frac{1}{h}\left(y(t+h) - y(t) - h\phi(t, h; y(t), y(t+h))\right), \quad h > 0, \;\; t \in [a, b-h], \\
&= 1/h \times \textit{ Defekt bei Einsetzen der Lösung in das Verfahren}
\end{aligned}$$

heißt der lokale Abbruchfehler *oder* Konsistenzfehler *des Verfahrens* (6.7) *für (AWP) an der Stelle t.*

Definition 6.1.3 *Das Verfahren* (6.7) *heißt zu (AWP) konsistent von der Ordnung p, falls es Konstanten $C > 0$ und $\bar{h} > 0$ gibt mit*

$$\|\tau(t,h)\| \le Ch^p \quad \textit{für alle } 0 < h \le \bar{h} \textit{ und alle } t \in [a, b-h].$$

Das Verfahren (6.7) *heißt* stabil, *falls eine Konstante $K > 0$ existiert mit*

$$\|\phi(t,h;u,v) - \phi(t,h;\tilde{u},\tilde{v})\| \le K\left(\|u-\tilde{u}\| + \|v-\tilde{v}\|\right) \quad \textit{für alle } t \in [a,b]\ u,v,\tilde{u},\tilde{v} \in \mathbb{R}^n.$$

Das Verfahren (6.7) *heißt* konvergent von der Ordnung p, *falls mit Konstanten $M > 0, H > 0$ gilt*

$$\|e_j\| = \|y(t_j) - u_j\| \le Mh^p, \quad \textit{für } j = 0,\dots,N \textit{ und alle } h = \tfrac{b-a}{N} \le H.$$

Besipiel: Explizites Euler-Verfahren Das Euler-Verfahren hat Konsistenzordnung 1.

Nachweis: Sei $f \in C^1([a,b] \times \mathbb{R}^n; \mathbb{R}^n)$ und y Lösung von $y' = f(t,y)$. Dann ist $y' \in C^1([a,b];\mathbb{R}^n)$, also $y \in C^2([a,b];\mathbb{R}^n)$ und Taylorentwicklung liefert komponentenweise mit einem $\xi_i \in [0,1]$

$$y(t+h) = y(t) + y'(t)h + \frac{1}{2}(y_i''(t+\xi_i h))_{1\le i\le n}h^2 = y(t) + f(t,y(t))h + \frac{1}{2}(y_i''(t+\xi_i h))_{1\le i\le n}h^2.$$

Also ergibt sich

$$\begin{aligned}\|\tau(t,h)\| &= \left\|\frac{1}{h}(y(t+h) - y(t) - hf(t,y(t)))\right\| = \frac{1}{2}\|(y_i''(t+\xi_i h))_{1\le i\le n}\|h \\ &\le \frac{1}{2}\|(\sup_{s\in[a,b]} |y_i''(s)|)_{1\le i\le n}\|h.\end{aligned}$$

Damit hat das Euler-Verfahren Konsistenzordnung 1. $\square$

Verfahren	Konsistenzordnung
Expl. Euler	1
Impl. Euler	1
Heun	2
Mod. Euler	2
RK4	4

6.1.4 Ein Konvergenzsatz

Wir beweisen nun einen grundlegenden Konvergenzsatz für explizite Einschrittverfahren.

Satz 6.1.4 *Sei $y \in C^1([a,b];\mathbb{R}^n)$ Lösung von (AWP). Das Verfahren* (6.7) *sei konsistent von der Ordnung p und stabil. Dann ist das Verfahren konvergent von der Ordnung p. Genauer gibt es $H > 0$, so dass für den globalen Diskretisierungsfehler gilt*

$$\|e_j\| = \|y(t_j) - u_j\| \le \frac{e^{4K|t_j - a|} - 1}{4K} 2Ch^p \quad \textit{für } j = 0,\dots,N \textit{ und alle } h = \tfrac{b-a}{N} \le H.$$

Beweis: (für Interessierte) Setze

$$y_j = y(t_j), \quad e_j = y_j - u_j, \quad j = 0,\dots,N.$$

Dann gilt für $j = 0,\dots,N-1$ nach Definition des Verfahrens (6.7) und des lokalen Diskretisierungsfehlers

$$\begin{aligned} u_{j+1} &= u_j + h\phi(t_j, h; u_j, u_{j+1}),\\ y_{j+1} &= y_j + h\phi(t_j, h; y_j, y_{j+1}) + h\tau(t_j, h). \end{aligned}$$

Subtraktion der ersten von der zweiten Gleichung ergibt

$$e_{j+1} = e_j + h(\phi(t_j, h; y_j, y_{j+1}) - \phi(t_j, h; u_j, u_{j+1})) + h\tau(t_j, h).$$

Sei nun $0 < h = (b-a)/N \le \bar{h}$. Wegen $t_j \in [a, b-h]$ liefert die Konsistenzbedingung $\|\tau(t_j, h)\| \le Ch^p$. Zusammen mit der Stabilität des Verfahrens erhalten wir daher mit der Dreiecksungleichung

$$\|e_{j+1}\| \le (1 + hK)\|e_j\| + hK\|e_{j+1}\| + hCh^p$$

Wähle nun $0 < H \le \bar{h}$ so klein, dass gilt $HK \le 1/2$. Dann ergibt sich für alle $0 < h = (b-a)/N \le H$

$$\|e_{j+1}\| \le \frac{1 + hK}{1 - hK}\|e_j\| + h2Ch^p \le (1 + h4K)\|e_j\| + h2Ch^p$$

Das nachfolgende Lemma liefert nun wegen $e_0 = 0$

$$\|e_{j+1}\| \le \frac{e^{4K|t_{j+1} - a|} - 1}{4K} 2Ch^p.$$

Damit ist der Satz bewiesen. □

Wir benötigen zur Vervollständigung des Beweises noch das folgende *diskrete Gronwall-Lemma* zur Abschätzung der Fehlerakkumulation.

Lemma 6.1.5 *Für Zahlen $L > 0$, $a_j \ge 0$, $h_j > 0$ und $b \ge 0$ sei*

$$a_{j+1} \le (1 + h_j L)a_j + h_j b, \quad j = 0, 1, \dots, n-1.$$

Dann gilt

$$a_j \le \frac{e^{Lt_j} - 1}{L} b + e^{Lt_j} a_0 \quad \textit{mit } t_j := \sum_{i=0}^{j-1} h_i.$$

Beweis: (für Interessierte) Für $j = 0$ ist die Behauptung klar. Der Induktionsschritt $j \to j+1$ ergibt sich aus

$$
\begin{aligned}
a_{j+1} &\le \underbrace{(1+h_jL)}_{\le e^{h_jL}} \left(\frac{e^{Lt_j}-1}{L}b + e^{Lt_j}a_0\right) + h_jb \\
&\le \left(\frac{e^{L(t_j+h_j)}-1-h_jL}{L} + h_j\right) b + e^{L(t_j+h_j)}a_0 \\
&= \frac{e^{Lt_{j+1}}-1}{L}b + e^{Lt_{j+1}}a_0
\end{aligned}
$$

□

6.1.5 Explizite Runge-Kutta-Verfahren

Verfahren hoher Konsistenzordnung kann man durch eine Verallgemeinerung des Ansatzes beim RK4-Verfahren gewinnen:

r-stufiges explizite Runge-Kutta-Verfahren:

Hier wählt man die Verfahrensfunktion

$$
\begin{aligned}
k_i &= f\left(t+\gamma_i h, u + h\sum_{j=1}^{i-1}\alpha_{ij}k_j\right), \quad i = 1,\dots,r, \\
\phi(t,h;u) &= \sum_{i=1}^{r}\beta_i k_i.
\end{aligned}
\tag{6.8}
$$

Hierbei heißt $k_i = k_i(t,u,h)$ die *i-te Stufe*. Zur kompakten Beschreibung von expliziten Runge-Kutta-Verfahren notiert man die Koeffizienten in einem Tableau, dem sogenannten

Butcher-Schema:

$$
\begin{array}{c|ccccc}
\gamma_1 & 0 & & & & \\
\gamma_2 & \alpha_{21} & 0 & & & \\
\gamma_3 & \alpha_{31} & \alpha_{32} & 0 & & \\
\vdots & \vdots & \vdots & \ddots & \ddots & \\
\gamma_r & \alpha_{r1} & \cdots & \cdots & \alpha_{r,r-1} & 0 \\
\hline
 & \beta_1 & \beta_2 & \cdots & \beta_{r-1} & \beta_r
\end{array}
$$

Beispiele für Butcher-Schemata:

Explizites Euler-Verfahren: Modifiziertes Euler-Verfahren: Verfahren von Heun:

$$
\begin{array}{c|c}
0 & 0 \\
\hline
 & 1
\end{array}
\qquad\qquad
\begin{array}{c|cc}
0 & 0 & \\
1/2 & 1/2 & 0 \\
\hline
 & 0 & 1
\end{array}
\qquad\qquad
\begin{array}{c|cc}
0 & 0 & \\
1 & 1 & 0 \\
\hline
 & 1/2 & 1/2
\end{array}
$$

Mit diesem Ansatz kann man Verfahren beliebiger Konsistenzordnung p erzeugen. Man muss hierzu die Stufenzahl r groß genug wählen. Taylorentwicklung des lokalen Abbruchfehlers liefert dann Gleichungen für die Koeffizienten.

Durch Taylorentwicklung läßt sich der folgende Satz beweisen.

Satz 6.1.6 *Betrachte ein Runge-Kutta Verfahren* (6.7) *mit Verfahrensfunktion* (6.8) *mit*

$$\gamma_i = \sum_{j=1}^{r} \alpha_{ij} \quad i = 1, \ldots, r.$$

Es besitzt genau dann für jede rechte Seite $f \in C^p([a,b] \times \mathbb{R})$ die Konsistenzordnung $p = 1$, falls die Koeffizienten der Gleichung

$$\sum_{i=1}^{r} \beta_i = 1$$

genügen; genau dann die Konsistenzordnung $p = 2$, falls die Koeffizienten zusätzlich der Gleichung

$$\sum_{i=1}^{r} \beta_i \gamma_i = 1/2$$

genügen; genau dann die Konsistenzordnung $p = 3$, falls die Koeffizienten zusätzlich den Gleichungen

$$\sum_{i=1}^{r} \beta_i \gamma_i^2 = 1/3$$

$$\sum_{i,j=1}^{r} \beta_i \alpha_{ij} \gamma_j = 1/6$$

genügen; genau dann die Konsistenzordnung $p = 4$, falls die Koeffizienten zusätzlich den Gleichungen

$$\sum_{i=1}^{r} \beta_i \gamma_i^3 = 1/4$$

$$\sum_{i,j=1}^{r} \beta_i \gamma_i \alpha_{ij} \gamma_j = 1/8$$

$$\sum_{i,j=1}^{r} \beta_i \alpha_{ij} \gamma_j^2 = 1/12$$

$$\sum_{i,j,k=1}^{r} \beta_i \alpha_{ij} \alpha_{jk} \gamma_k = 1/24$$

genügen.

Beweis: Siehe zum Beispiel Deuflhard und Bornemann [DB02]. □

6.2 Steife Differentialgleichungen

In zahlreichen Anwendungen (z.B. beim Ablauf chemischer Rekationen), aber auch bei Semidiskretisierung partieller Differentialgleichungen, treten *steife Systeme* auf. Obwohl es sich auch um Anfangswertprobleme handelt, erzwingen sie bei vielen – aber nicht bei allen – Verfahren inakzeptabel kleine Schrittweiten h, um eine genaue Lösung zu erhalten.

Ausgangspunkt ist ein Anfangswertproblem für ein System n gewöhnlicher Differentialgleichungen:

$$\text{(AWPn)} \qquad \begin{aligned} &y'(t) = f(t, y(t)), \quad t \in [a, b] \\ &y(a) = y_0 \end{aligned}$$

mit $f : [a, b] \times \mathbb{R}^n \to \mathbb{R}^n$, $y_0 \in \mathbb{R}^n$.

Der Begriff "steifes System" ist in der Literatur nicht ganz einheitlich definiert. Der wesentlich Punkt ist, dass die Lösung zusammengesetzt ist aus einem langsam veränderlichen Teil (der meist abklingt) und einem Anteil, die im Allgemeinen sehr schnell gedämpft wird.

Wir betrachten den Spezialfall, dass (AWPn) linear ist:

$$\text{(LAWPn)} \qquad \begin{aligned} &y'(t) = Ay(t) + b, \quad t \in [a, b] \\ &y(a) = y_0 \end{aligned}$$

mit einer Matrix $A \in \mathbb{R}^{n,n}$ und einem Vektor $b \in \mathbb{R}^n$.

Sei zudem $A \in \mathbb{R}^{n,n}$ diagonalisierbar mit zugehörigen Eigenwerten λ_i sowie Eigenvektoren v_i. Mit einer partikulären Lösung y_P ist dann die allgemeine Lösung von der Form

$$y(t) = y_H(t) + y_P(t), \quad y_H(t) = \sum_{i=1}^{n} C_i e^{\lambda_i t} v_i.$$

Ist nun $\mathrm{Re}(\lambda_i) < 0$ für $i = 1, \ldots, n$, so gilt

$$\lim_{t\to\infty} y_H(t) \to 0,$$

alle Lösungen nähern sich also y_P an. Hierbei klingen die Summanden in y_H mit $\mathrm{Re}(\lambda_i) \ll -1$ sehr schnell und Summanden mit $\mathrm{Re}(\lambda_i) \not\ll -1$ deutlich langsamer ab. Gibt es Eigenwerte mit $\mathrm{Re}(\lambda_i) \ll -1$ und Eigenwerte mit schwach negativem Realteil, so nennt man das System *steif*.

Beispiel: Betrachte zum Beispiel das Problem

$$y' = Ay, \quad y(0) = y_0 := \begin{pmatrix} C_1 + C_2 \\ C_1 - C_2 \end{pmatrix}$$

mit $C_1, C_2 \in \mathbb{R}$ und

$$A = \begin{pmatrix} \frac{\lambda_1+\lambda_2}{2} & \frac{\lambda_1-\lambda_2}{2} \\ \frac{\lambda_1-\lambda_2}{2} & \frac{\lambda_1+\lambda_2}{2} \end{pmatrix}.$$

A hat die Eigenwerte λ_1, λ_2 mit zugehörigen Eigenvektoren $\binom{1}{1}$ bzw. $\binom{1}{-1}$. Ist zum Beispiel $\lambda_1 = -1$ und $\lambda_2 = -1000$, dann lautet die Lösung

$$y(t) = C_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{-t} + C_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-1000t}.$$

Der zweite Term spielt nach kürzester Zeit so gut wie keine Rolle mehr. Der erste Term ist bestimmend und konvergiert für $t \to \infty$ ebenfalls gegen 0. Von einem geeigneten Integrationsverfahren wird man erwarten, dass es ohne große Einschränkungen an die Schrittweite Näherungen u_j liefert mit

$$\lim_{j\to\infty} u_j = 0.$$

Betrachten wir jedoch zum Beispiel die Anwendung des expliziten Euler-Verfahrens, so ergibt sich mit $u_0 = y_0 = C_1 \binom{1}{1} + C_2 \binom{1}{-1}$

$$u_1 = (I + hA)u_0 = C_1(1 + h\lambda_1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + C_2(1 + h\lambda_2) \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

und nun induktiv

$$u_j = C_1(1 + h\lambda_1)^j \begin{pmatrix} 1 \\ 1 \end{pmatrix} + C_2(1 + h\lambda_2)^j \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Ist $C_2 \neq 0$, so müssen wir $|1 + h\lambda_2| < 1$, also $-h\lambda_2 = 1000h < 2$ wählen, damit gilt $\lim_{j\to\infty} u_j = 0$. Ein geeignetes Verfahren sollte dies möglichst für alle $h > 0$ sicherstellen.
$\square$

Das Euler-Verfahren benötigt also sehr kleine Schrittweiten, obwohl sich die Lösung kaum ändert. Man nennt die Differentialgleichung dann *steif*. Die formale Definition ist uneinheitlich. Folgende Definition ist am weitesten verbreitet.

Definition 6.2.1 *Ein Anfangswertproblem (LAWPn) heißt* steif, *wenn A Eigenwerte mit* $\mathrm{Re}(\lambda_i) \ll -1$ *und Eigenwerte λ_i mit schwach negativem Realteil besitzt.*

Wir kommen nun zur numerischen Behandlung steifer Differentialgleichungen.

Um eine einfache Modellgleichung für steife Differentialgleichungen herzuleiten, betrachten wir zunächst (LAWPn) mit $b = 0$, also

(6.9) $$y' = Ay, y(0) = y_0.$$

A sei diagonalisierbar. Dann existiert $M \in \mathbb{R}^{n,n}$ mit

$$MAM^{-1} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$$

mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$ von A. Setzen wir $z = My$, dann gilt also

$$z' = MAy = MAM^{-1}z = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)z, \quad z(0) = My_0.$$

Die Komponenten z_i von $z = My$ erfüllen also

$$z_i' = \lambda_i z_i, \quad z_i(0) = (My_0)_i. \tag{6.10}$$

Für eine steife Differentialgleichung gilt zudem $\mathrm{Re}(\lambda_i) \le 0$, wobei einige $\mathrm{Re}(\lambda_i)$ stark, andere schwach negativ sind.

Beobachtung: Verhält sich ein numerisches Verfahren für alle Differentialgleichungen (6.10) gutartig, dann liefert es auch für das steife System (6.9) gute Ergebnisse.

Um Verfahren für steife Differentialgleichungen zu bewerten und zu analysieren, betrachtet man daher nach Dahlquist (1963) die

Modellgleichung

$$y' = \lambda y, \quad y(0) = 1, \quad \text{mit } \lambda \in \mathbb{C}, \ \mathrm{Re}(\lambda) < 0. \tag{6.11}$$

Die Lösung ist

$$y(t) = e^{\lambda t}$$

und wegen $\mathrm{Re}(\lambda) < 0$ gilt

$$\lim_{t\to\infty} y(t) = 0. \tag{6.12}$$

Die Lösung fällt also je nach größe von $|\mathrm{Re}(\lambda)|$ sehr unterschiedlich stark ab. Damit ein Verfahren gut für steife Differentialgleichungen geeignet ist, hat sich folgende Anforderung bewährt:

Forderung: Die numerische gewonnene Näherungslösung von (6.11) soll die Eigenschaften von der Lösung $y(t) = e^{\lambda t}$, also insbesondere (6.12), möglichst gut widerspiegeln.

Dies motiviert folgende

Definition 6.2.2 (A-stabil (absolut stabil), L-stabil)
Ein Verfahren heißt

a) absolut stabil (A-stabil)*, wenn seine Anwendung auf das Modellproblem* (6.11) *für jede Schrittweite $h > 0$ eine Folge $\{u_j\}_{j\in\mathbb{N}_0}$ produziert mit*

$$|u_{j+1}| \le |u_j| \quad \forall\, j \ge 0.$$

b) L-stabil*, wenn es A-stabil ist und zudem gilt*

$$\lim_{j\to\infty} u_j = 0.$$

Bei vielen Einschrittverfahren gilt bei Anwendung auf das Modellproblem (6.11) die Beziehung

$$u_{j+1} = R(q)u_j \quad \text{mit } q = \lambda h$$

und einer Funktion $R : D \to \mathbb{C}$, $0 \in D \subset \mathbb{C}$.

Beispiel: Wendet man das explizite Euler-Verfahren auf das Modellproblem (6.11) an, dann erhält man

$$u_{j+1} = u_j + h\lambda u_j = (1 + \lambda h)u_j = (1 + q)u_j,$$

also ist für das explizite Euler-Verfahren $R(q) = 1 + q$. □

Definition 6.2.3 *Man nennt R die* Stabilitätsfunktion *des Einschrittverfahrens. Die Menge*

$$S = \{q \in \mathbb{C} \,:\, |R(q)| < 1\}\,.$$

heißt Stabilitätsgebiet *des Einschrittverfahrens.*

Offensichtlich gilt

$$\text{A-stabil} \iff |R(q)| \leq 1 \quad \forall\, q \in \mathbb{C},\ \text{Re}(q) < 0.$$

$$\text{L-stabil} \iff |R(q)| < 1 \quad \forall\, q \in \mathbb{C},\ \text{Re}(q) < 0 \iff S \supset \{q \in \mathbb{C} \,:\, \text{Re}(q) < 0\}\,.$$

6.2.1 Stabilitätsgebiete einiger Verfahren

Explizites Euler-Verfahren

Anwendung des expliziten Euler-Verfahrens auf das Modellproblem (6.11) ergibt

$$u_{j+1} = u_j + h\lambda u_j = (1 + \lambda h)u_j,$$

die Stabilitätsfunktion ist daher $R(q) = 1 + q$. Das Stabilitätsgebiet ist also

$$S = \{q \in \mathbb{C} \,:\, |1 + q| < 1\}\,.$$

Bemerkung: Man kann leicht zeigen, dass alle expliziten Runge-Kutta-Verfahren nicht A-stabil sind!

Implizites Euler-Verfahren

Das implizite Euler-Verfahren liefert für das Modellproblem (6.11)

$$u_{j+1} = u_j + h\lambda u_{j+1}$$

und somit

$$u_{j+1} = \frac{1}{1-\lambda h} u_j.$$

Dies ergibt die Stabilitätsfunktion $R(q) = \frac{1}{1-q}$, $q \neq 1$, und das Stabilitätsgebiet

$$S = \{q \in \mathbb{C} \,:\, |1-q| > 1\} \supset \{q \in \mathbb{C} \,:\, \mathrm{Re}(q) < 0\}\,.$$

Das implizite Euler-Verfahren ist also A-stabil, sogar L-stabil!

Implizite Trapezregel

Die Verfahrensgleichung lautet

$$u_{j+1} = u_j + \frac{h}{2}(f(u_j) + f(u_{j+1})).$$

Wir erhalten für das Modellproblem (6.11)

$$u_{j+1} = u_j + \frac{h}{2}\lambda(u_j + u_{j+1})$$

und somit

$$u_{j+1} = \frac{1+\lambda h/2}{1-\lambda h/2} u_j.$$

Daher gilt $R(q) = \frac{1+q/2}{1-q/2}$, $q \neq 2$, und das Stabilitätsgebiet ist

$$S = \{q \in \mathbb{C} \,:\, |1+q/2| < |1-q/2|\} = \{q \in \mathbb{C} \,:\, \mathrm{Re}(q) < 0\}\,.$$

Implizite Runge-Kutta-Verfahren

Besonders gut geeignet für steife Differentialgleichungen sind implizite Runge-Kutta-Verfahren.

Implizite Runge-Kutta-Verfahren erhält man durch Butcher-Schemata, bei denen die Koeffizienten α_{ij} *keine* strikte untere Dreiecksmatrix bilden. Die Verfahrensgleichung ist gegeben durch

$$\begin{aligned} k_i &= f\left(t + \gamma_i h, u + h\sum_{l=1}^{r} \alpha_{il} k_l\right), \quad i = 1, \dots, r, \\ \phi(t, h; u) &= \sum_{i=1}^{r} \beta_i k_i. \end{aligned} \tag{6.13}$$

(beachte die Summation bis r anstelle $i-1$). Ein implizites Runge-Kutta-Verfahren ist ein explizites Einschrittverfahren, lediglich die Stufen k_i sind als Lösung eines nichtlinearen Gleichungssystems gegeben. Man kann nun die Koeffizienten α_{ij}, β_i, γ_i tatsächlich so wählen, dass ein L-stabiles Verfahren der Ordnung $p = 2r$ ensteht.

Kapitel 7

Verfahren zur Eigenwert- und Eigenvektorberechnung

7.1 Eigenwertprobleme

In vielen technischen und physikalischen Problemen, etwa bei der Untersuchung des Schwingungsverhaltens von mechanischen oder elektrischen Systemen, ist es von Bedeutung, die Eigenwerte und Eigenvektoren einer Matrix $A \in \mathbb{C}^{n,n}$ zu bestimmen.

Aber auch bei der PageRank-Bestimmung der Suchmaschine google werden große Eigenwertprobleme gelöst.

Beispiel: Grundschwingungen und Resonanzfrequenzen schwingender Strukturen

Wir betrachten eine mechansiche Struktur (z.B. Karosserie, Brücke, Gebäude) und interessieren uns dafür, auf welchen Frequenzen sie schwingen kann und wie die zugehörigen Schwingungen aussehen (für elektrische Schaltkreise ist die Situation ähnlich). Dies Fragestellung ist z.B. bei der Schwingungs- und Lärmbekämpfung sowie bei der Auslegung von Bauwerken, Flugzeugen, etc. von großer Relevanz.

Sei jeweils $y_i(t) \in \mathbb{R}^3$ die Verschiebung der Struktur am Punkt $x_i \in \mathbb{R}^3$ zur Zeit t, $1 \leq i \leq n$. Im Falle einer ungedämpften Schwingung, die durch eine Kraft $f(t)$ angeregt wird erfüllt $y(t) = (y_i(t))_{1 \leq i \leq n}$ das Anfangswertproblem

$$My''(t) = -Ay(t) + f(t), \;\; y(0) = y^{(0)}, \;\; y'(0) = y^{(1)}.$$

mit der invertierbaren Massenmatrix $M \in \mathbb{R}^{3n,3n}$ und der Steifigkeitsmatrix $A \in \mathbb{R}^{3n,3n}$. Die Lösung ist die Summe aus einer Lösung der inhomogenen Gleichung und einer geeigneten Lösung der homogenen Gleichung

$$My''(t) = -Ay(t),$$

die äquivalent ist zu
$$y''(t) = -M^{-1}Ay(t).$$

Man kann zeigen, dass $B := M^{-1}A$ diagonalisierbar ist mit rellen Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{3n}$ und zugehörigen Eigenvektoren $v_1, \ldots, v_{3n}$. Nun ist jede der Funktionen

$$\phi_i(t) := (a_i \sin(\sqrt{\lambda_i}t) + b_i \cos(\sqrt{\lambda_i}t))v_i$$

eine Lösung der homogenen Gleichung, denn wegen $Bv_i = \lambda_i v_i$ gilt

$$\phi_i''(t) = -\sqrt{\lambda_i}^2 (a_i \sin(\sqrt{\lambda_i}t) + b_i \cos(\sqrt{\lambda_i}t))v_i = -\lambda_i \phi_i(t) = -B\phi_i(t).$$

Damit sind $\phi_i(t)$ die Grundschwingungen der Struktur, die i-te Grundschwingung hat also Frequenz $\sqrt{\lambda_i}/(2\pi)$ und die zugehörige Verformung der Struktur wird durch den Eigenvektor v_i gegeben. □

Beispiel: PageRank-Algorithmus von google

Betrachte N Webseiten. Webseite i enthalte k_i Links auf andere Seiten. Die Wahrscheinlichkeit, dass ein Nutzer von Seite i nach Seite j geht, wird modelliert als

$$p_{ij} = \begin{cases} \frac{\alpha}{k_i} + \frac{1-\alpha}{N}, & \text{falls Seite } i \text{ einen Link auf Seite } j \text{ enthält,} \\ \frac{1-\alpha}{N}, & \text{falls Seite } i \text{ keinen Link auf Seite } j \text{ enthält.} \end{cases}$$

Hierbei wird in der Regel $\alpha = 0,85$ gewählt. Sei nun $P = (p_{ij})_{1 \leq i,j \leq N}$. Als Gewichte der Seiten bestimmt man nun einen Vektor $\pi \in \mathbb{R}^N$, die sogenannte stationäre Verteilung, so dass gilt

$$\pi = P^T\pi, \quad \sum_{i=1}^{N} \pi_i = 1, \quad \pi_i \geq 0.$$

Anschauliche Erklärung: Ist π_i der Anteil der Internetnutzer, die sich im Mittel auf Seite i aufhalten, dann bleibt nach dem Übergangsverhalten gemäß den Wahrscheinlichkeiten p_{ij} dieser Anteil unverändert. Also gibt π_i den Anteil der Internetnutzer an, die sich nach Einstellen eines Gleichgewichtszustandes im Mittel auf Seite i befinden. □

7.1.1 Grundlagen

Definition 7.1.1 *Eine Zahl* $\lambda \in \mathbb{C}$ *heißt* Eigenwert *einer Matrix* $A \in \mathbb{C}^{n,n}$*, wenn es einen Vektor* $x \in \mathbb{C}^n$*,* $x \neq 0$ *gibt mit*
$$Ax = \lambda x.$$

Jeder solche Vektor $x \in \mathbb{C}^n$ *heißt* (Rechts-)Eigenvektor *zum Eigenwert* λ*. Die Menge* $\sigma(A)$ *aller Eigenwerte von* A *heißt* Spektrum *von* A*.* □

Der Unterraum

$$\mathrm{Eig}_A(\lambda) := \{x \in \mathbb{C}^n \,:\, (A - \lambda I)x = 0\}$$

ist der *Eigenraum* von A zum Eigenwert λ. Seine Dimension

$$\gamma(\lambda) := \dim \mathrm{Eig}_A(\lambda) = n - \mathrm{Rang}(A - \lambda I)$$

ist die *geometrische Vielfachheit* von λ und gibt die Maximalzahl linear unabhängiger Eigenvektoren zu λ an.

Offensichtlich ist λ genau dann Eigenwert von A, wenn gilt

$$\chi(\lambda) := \det(A - \lambda I) = 0,$$

also wenn λ Nullstelle des *charakteristischen Polynoms* $\chi(\mu)$ von A ist. χ ist ein Polynom n-ten Grades und hat die Form

$$\chi(\mu) = (-1)^n \mu^n + (-1)^{n-1}\mu^{n-1}\mathrm{spur}(A) + \cdots + \det(A).$$

Sind $\lambda_1, \ldots, \lambda_k$ die verschiedenen Nullstellen von χ (d.h. die verschiedenen Eigenwerte von A) mit Vielfachheiten ν_i, $i = 1, \ldots, k$, so gilt $\nu_1 + \ldots + \nu_k = n$ und χ hat die Linearfaktorzerlegung

$$\chi(\mu) = (-1)^n (\mu - \lambda_1)^{\nu_1} \cdots (\mu - \lambda_k)^{\nu_k}.$$

Man nennt $\nu(\lambda_i) = \nu_i$ die *algebraische Vielfachheit* von λ_i. Es ist nicht schwer zu zeigen, dass immer gilt

$$\gamma(\lambda_i) \leq \nu(\lambda_i).$$

Wir fassen einige grundlegende Eigenschaften von Eigenwerten und Eigenvektoren zusammen:

Proposition 7.1.2 *Sei $A \in \mathbb{C}^{n,n}$ ein beliebig. Dann gilt:*

a) Ist λ Eigenwert von A, so ist λ Eigenwert von A^T und $\bar{\lambda}$ Eigenwert von $A^H := \bar{A}^T$.

b) Für jede nichtsinguläre Matrix $T \in \mathbb{C}^{n,n}$ hat die zu A ähnliche Matrix *$B := T^{-1}AT$ dasselbe charakteristische Polynom und dieselben Eigenwerte wie A. Ist x Eigenvektor von A, so ist $y := T^{-1}x$ Eigenvektor von B.*

c) Ist A hermitesch*, also $A^H = A$ mit $A^H := \bar{A}^T$, dann hat A lauter reelle Eigenwerte. Ist A* unitär*, also $A^H = A^{-1}$, so gilt $|\lambda| = 1$ für jeden Eigenwert λ.*

Eine Matrix $A \in \mathbb{C}^{n,n}$ heißt *diagonalisierbar*, wenn sie n linear unabhängige Eigenvektoren $x_1, \ldots, x_n$ besitzt. Die zugehörige Matrix $T := (x_1, \ldots, x_n)$ ist dann invertierbar und mit den Eigenwerten λ_i zu x_i gilt

$$T^{-1}AT = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n) =: D.$$

Tatsächlich haben wir

$$AT = (\lambda_1 x_1, \dots, \lambda_n x_n) = TD.$$

Eine wichtige Rolle spielen *hermitesche* Matrizen $A \in \mathbb{C}^{n,n}$, d.h. $A^H = A$, und mithin *reelle symmetrische* Matrizen. Man kann recht einfach zeigen, dass eine hermitesche Matrix $A \in \mathbb{C}^{n,n}$ immer *diagonalisierbar* ist mit einer *unitären* Matrix $T = U$, also

$$U^{-1}AU = D, \quad U^H = U^{-1}.$$

Ist $A = A^T$ reell, dann kann $U \in \mathbb{R}^{n,n}$ orthogonal gewählt werden, also

$$U^{-1}AU = D, \quad U^T = U^{-1}.$$

Die wichtigsten Verfahren zur Berechnung von Eigenwerten und Eigenvektoren einer Matrix A nehmen zunächst eine Reihe von Ähnlichkeitstransformationen

$$A^{(0)} := A, \quad A^{(k+1)} := T_k^{-1} A^{(k)} T_k, \quad k = 0, 1, \dots$$

vor, um A in eine Matrix einfacherer Gestalt zu transformieren, deren Eigenwerte und Eigenvektoren leichter zu bestimmen sind.

7.1.2 Grundkonzepte numerischer Verfahren

Die im folgenden besprochenen numerischen Verfahren zur Berechnung von Eigenwerten und Eigenvektoren lassen sich in zwei Klassen einordnen. Die eine beruht auf der Vektoriteration, die andere auf der Anwendung von Ähnlichkeitstransformationen.

Vektoriteration

Bei der ersten Klasse von Verfahren handelt es sich um Vektoriterationen, die allgemein von der Form sind

$$x^{(k+1)} = \frac{Bx^{(k)}}{\|Bx^{(k)}\|}, \quad k = 0, 1, \dots$$

mit einem Startvektor $x^{(0)}$, einer Iterationsmatrix B und einer Vektornorm $\|\cdot\|$.

Ähnlichkeitstransformation auf einfachere Gestalt

Nach Proposition 7.1.2 bleiben die Eigenwerte einer Matrix A bei einer Ähnlichkeitstransformation $B = T^{-1}AT$ unverändert und aus einem Eigenvektor y von B erhält man durch $x = Ty$ einen Eigenvektor der Ausgangsmatrix A.

Es liegt daher nahe, A durch Ähnlichkeitstransformationen

$$A^{(0)} := A \to A^{(1)} \to \dots, \quad A^{(k+1)} = T_k^{-1} A^{(k)} T_k \tag{7.1}$$

in eine einfachere Form zu überführen, für die die Bestimmung von Eigenwerten und Eigenvektoren einfacher ist. Wir betrachten hier nur das QR-Verfahren, das eines der schnellsten Verfahren zur Lösung von Eigenwertproblemen darstellt.

QR-Verfahren: Beim QR-Verfahren wird durch Anwendung unitärer Matrizen T_i erreicht, dass die Elemente von $A^{(k)}$ im strikten unteren Dreieck gegen null konvergieren. Die Diagonaleinträge von $A^{(k)}$ konvergieren wiederum gegen die Eigenwerte von A.

7.1.3 Störungstheorie für Eigenwertprobleme

Bei oberen oder unteren Dreiecksmatrizen sind die Eigenwerte nichts anderes als die Diagonalelemente. Wir haben bereits angedeutet, dass das QR-Verfahren durch Ähnlichkeitstransformationen den Außerdiagonalteil bzw. das strikte untere Dreieck reduzieren. Störungsresultate für Eigenwerte liefern unter anderem Schranken, wie gut die Diagonalelemente mit den Eigenwerten übereinstimmen.

Wir haben das folgende fundamentale Resultat.

Satz 7.1.3 *Bezeichnet $\lambda_i(A)$, $i = 1, \dots, n$, die angeordneten Eigenwerte einer Matrix $A \in \mathbb{C}^{n,n}$ (zum Beispiel aufsteigend nach Realteil und bei gleichem Realteil aufsteigend nach Imaginärteil), dann sind die Abbildungen*

$$A \in \mathbb{C}^{n,n} \mapsto \lambda_i(A), \quad i = 1, \dots, n$$

stetig. Eigenwerte hängen also stetig von der Matrix ab.

Beweis: Siehe zum Beispiel Werner [We92]. □

Ein wichtiges Einschließungskriterium für Eigenwerte erhält man durch die *Gershgorin-Kreise*:

Satz 7.1.4 *Es sei $A = (a_{ij}) \in \mathbb{C}^{n,n}$ beliebig.*

a) Es gilt

$$\sigma(A) \subset \bigcup_{i=1}^{n} K_i$$

mit den Gershgorin-Kreisen

$$K_i := \left\{ \mu \in \mathbb{C} \;:\; |\mu - a_{ii}| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \right\}, \quad i = 1, \dots, n.$$

b) *Ist die Vereinigung G_1 von k Gershgorin-Kreisen disjunkt von der Vereinigung G_2 der restlichen $n-k$ Gershgorin-Kreise, dann enthält G_1 genau k Eigenwerte und G_2 genau $n-k$ Eigenwerte von A.*

Das folgende Resultat gilt für diagonalisierbare Matrizen:

Satz 7.1.5 (Bauer/Fike)
Es sei $A \in \mathbb{C}^{n,n}$ diagonalisierbar, also

$$T^{-1}AT = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: D.$$

Dann gilt für jede Matrix $\Delta A \in \mathbb{C}^{n,n}$

$$\forall\, \mu \in \sigma(A + \Delta A): \quad \min_{i=1,\ldots,n} |\mu - \lambda_i| \leq \operatorname{cond}_2(T)\|\Delta A\|_2.$$

Hierbei ist $\|\cdot\|_2$ die von der euklidischen Norm induzierte Matrix-Norm und $\operatorname{cond}_2(T) := \|T\|_2\|T^{-1}\|_2$ *die zugehörige Kondition von T.*

Bemerkung: Ist A hermitesch, so kann T unitär gewählt werden und es gilt $\operatorname{cond}_2(T) = 1$.

7.2 Die Vektoriteration

7.2.1 Definition und Eigenschaften der Vektoriteration

Definition 7.2.1 *Für eine Matrix $B \in \mathbb{C}^{n,n}$ ist die zugehörige Vektoriteration gegeben durch*

$$z^{(k+1)} = \frac{1}{\|Bz^{(k)}\|} Bz^{(k)}, \quad k = 0, 1, \ldots \tag{7.2}$$

mit einem Startvektor $z^{(0)} \in \mathbb{C}^n \setminus \{0\}$.

Bei geeigneter Wahl von B ergeben sich hieraus Näherungen $z^{(k)}$ für einen Eigenvektor zu einem Eigenwert λ. Eine Eigenwertnäherung für λ erhalten wir dann durch den *Rayleigh-quotienten*

$$R(z^{(k)}, B) = \frac{(z^{(k)})^H Bz^{(k)}}{(z^{(k)})^H z^{(k)}}.$$

Wir untersuchen die grundlegenden Eigenschaften für eine diagonalisierbare Matrix B mit Eigenwerten $\lambda_1, \ldots, \lambda_n$. Wir sagen, dass ein Vektor $x \in \mathbb{C}^n$ *einen Anteil in* $\operatorname{Eig}_B(\lambda_i)$ hat, falls in der eindeutigen Darstellung

$$x = u + v, \quad u \in \operatorname{Eig}_B(\lambda_i), \quad v \in \bigoplus_{\lambda_j \neq \lambda_i} \operatorname{Eig}_B(\lambda_j)$$

gilt $u \neq 0$. u ist der Anteil von x in $\operatorname{Eig}_B(\lambda_i)$.

Satz 7.2.2 *Es sei $B \in \mathbb{C}^{n,n}$ diagonalisierbar mit Eigenwerten $\lambda_1, \ldots, \lambda_n$,*

$$\lambda_1 = \ldots = \lambda_r, \quad |\lambda_r| > |\lambda_{r+1}| \geq \ldots \geq |\lambda_n|$$

mit $r < n$. Falls der Startvektor $z^{(0)}$ einen Anteil in $\mathrm{Eig}_B(\lambda_1)$ besitzt, gilt für die Vektoriteration (7.2)

$$R(z^{(k)}, B) = \frac{(z^{(k)})^H B z^{(k)}}{(z^{(k)})^H z^{(k)}} = \lambda_1 + O(q^k) \quad \textit{für } k \to \infty, \quad q = \frac{|\lambda_{r+1}|}{|\lambda_1|} < 1.$$

Zudem gilt

$$z^{(k)} = \frac{\lambda_1^k}{|\lambda_1|^k} \frac{x_1}{\|x_1\|} + O(q^k), \ \ k \geq 1.$$

mit einer beliebigen Vektornorm $\|\cdot\|$, wobei x_1 den Anteil von $z^{(0)}$ in $\mathrm{Eig}_B(\lambda_1)$ bezeichnet.

Beweis: Wir können genausogut die nicht normierte Folge $\tilde{z}^{(k+1)} = B\tilde{z}^{(k)}$, $\tilde{z}^{(0)} = z^{(0)}$ betrachten. Es gilt dann $z^{(k)} = \tilde{z}^{(k)}/\|\tilde{z}^{(k)}\|$, $k \geq 1$.

Es gibt eine Darstellung der Form $z^{(0)} = x_1 + \sum_{j=r+1}^n x_j$ mit $x_j \in \mathrm{Eig}_B(\lambda_j)$, $x_1 \neq 0$. Einsetzen in $\tilde{z}^{(k+1)} = B\tilde{z}^{(k)}$ ergibt

$$(7.3) \qquad \tilde{z}^{(k)} = B^k z^{(0)} = \lambda_1^k x_1 + \sum_{j=r+1}^n \lambda_j^k x_j = \lambda_1^k \left(x_1 + \sum_{j=r+1}^n \left(\frac{\lambda_j}{\lambda_1}\right)^k x_j \right), \quad k \geq 0.$$

Dies liefert

$$\tilde{z}^{(k)} = \lambda_1^k (x_1 + O(q^k))$$

und somit

$$\begin{aligned}
(\tilde{z}^{(k)})^H B \tilde{z}^{(k)} &= (\tilde{z}^{(k)})^H \tilde{z}^{(k+1)} = \bar{\lambda}_1^k \lambda_1^{k+1} (x_1 + O(q^k))^H (x_1 + O(q^k)) \\
&= \lambda_1 |\lambda_1|^{2k} (\|x_1\|_2^2 + O(q^k)) \\
(\tilde{z}^{(k)})^H \tilde{z}^{(k)} &= \bar{\lambda}_1^k \lambda_1^k (x_1 + O(q^k))^H (x_1 + O(q^k)) = |\lambda_1|^{2k} (\|x_1\|_2^2 + O(q^k)).
\end{aligned}$$

Wir erhalten

$$R(z^{(k)}, B) = R(\tilde{z}^{(k)}, B) = \lambda_1 \frac{\|x_1\|_2^2 + O(q^k)}{\|x_1\|_2^2 + O(q^k)} = \lambda_1 + O(q^k).$$

Analog haben wir

$$z^{(k)} = \frac{\tilde{z}^{(k)}}{\|\tilde{z}^{(k)}\|} = \frac{\lambda_1^k (x_1 + O(q^k))}{|\lambda_1|^k (\|x_1\| + O(q^k))} = \frac{\lambda_1^k}{|\lambda_1|^k} \frac{x_1}{\|x_1\|} + O(q^k)$$

□

Bemerkung: Selbst wenn $z^{(0)}$ keinen Anteil in $\mathrm{Eig}_B(\lambda_1)$ hat, was bei "genügend allgemeiner" Wahl vom $z^{(0)}$ unwahrscheinlich ist, so stellt sich in der Praxis diese Situation durch den Einfluß von Rundungsfehlern ein. $\square$

Im Falle hermitescher Matrizen erhält man lineare Konvergenzrate q^2 des Rayleigh-Quotienten gegen λ_1.

Satz 7.2.3 *Sei $B \in \mathbb{C}^{n,n}$ hermitesch. Dann gilt unter den Voraussetzungen von Satz 7.2.2 für den Rayleigh-Quotienten die Konvergenzaussage*

$$R(z^{(k)}, B) = \frac{(z^{(k)})^H B z^{(k)}}{(z^{(k)})^H z^{(k)}} = \lambda_1 + O(q^{2k}) \quad \textit{für } k \to \infty \textit{ mit } \quad q = \frac{|\lambda_{r+1}|}{|\lambda_1|} < 1.$$

7.2.2 Die Vektoriterationen nach v. Mises und Wielandt

Sei $A \in \mathbb{C}^{n,n}$ gegeben. Unterschiedliche Varianten der Vektoriteration entstehen durch die Wahl der Iterationsmatrix B.

Einfache Vektoriteration nach von Mises

Die einfache Vektoriteration erhält man durch die naheliegende Wahl $B = A$. Die Konvergenzeigenschaften können dann unmittelbar Satz 7.2.2 bzw. 7.2.3 entnommen werden.

Inverse Vektoriteration von Wielandt

Offensichtliche Nachteile der Vektoriteration sind die langsame Konvergenz bei schlechter Trennung der Eigenwerte und die Einschränkung auf die Bestimmung des betragsmäßig größten Eigenwerts. Dies kann durch die inverse Vektoriteration von Wielandt vermieden werden. Man braucht hierzu eine gute Näherung μ eines Eigenwerts λ_j, so dass

$$|\lambda_j - \mu| \ll |\lambda_i - \mu|, \quad \text{für } \lambda_i \neq \lambda_j.$$

Dann hat für $\mu \neq \lambda_j$ die Matrix $B = (A - \mu I)^{-1}$ die Eigenwerte

$$\mu_i = \frac{1}{\lambda_i - \mu},$$

wobei $|\mu_j| \gg |\mu_i|$ für alle $\mu_i \neq \mu_j$. Ferner ist x_j genau dann Eigenvektor von B zum Eigenwert μ_j, wenn x_j Eigenvektor von A zum Eigenwert λ_j ist.

Die zugehörige inverse Iteration von Wielandt lautet dann

$$z^{(k+1)} = \frac{\hat{z}^{(k+1)}}{\|\hat{z}^{(k+1)}\|} \quad \text{mit } \hat{z}^{(k+1)} = (A - \mu I)^{-1} z^{(k)}.$$

In der Praxis bestimmt man nicht $(A - \mu I)^{-1}$, sondern implementiert die Iteration in der Form

$$\text{Löse } (A - \mu I)\hat{z}^{(k+1)} = z^{(k)} \quad \text{und setze} \quad z^{(k+1)} = \frac{\hat{z}^{(k+1)}}{\|\hat{z}^{(k+1)}\|}.$$

Die inverse Iteration von Wielandt hat dann im Falle

$$q := \max_{1 \le i \le n, \lambda_i \ne \lambda_j} \frac{|\lambda_j - \mu|}{|\lambda_i - \mu|} < 1$$

nach Satz 7.2.2 die Konvergenzeigenschaften

$$R(z^{(k)}, (A - \mu I)^{-1}) = \frac{(z^{(k)})^H \hat{z}^{(k+1)}}{(z^{(k)})^H z^{(k)}} = \frac{1}{\lambda_j - \mu} + O(q^k),$$

$$z^{(k)} = \frac{|\lambda_j - \mu|^k}{(\lambda_j - \mu)^k} \frac{x_j}{\|x_j\|} + O(q^k),$$

wobei x_j den Anteil von $z^{(0)}$ in $\mathrm{Eig}_A(\lambda_j) = \mathrm{Eig}_{(A-\mu I)^{-1}}(1/(\lambda_j - \mu))$ bezeichnet. Ist A zudem hermitesch, so erfüllt der Rayleigh-Quotient nach Satz 7.2.3

$$R(z^{(k)}, (A - \mu I)^{-1}) = \frac{(z^{(k)})^H \hat{z}^{(k+1)}}{(z^{(k)})^H z^{(k)}} = \frac{1}{\lambda_j - \mu} + O(q^{2k}).$$

7.3 Das QR-Verfahren

Das im folgenden beschriebene QR-Verfahren von Francis bildet die Basis sehr leistungsfähiger Verfahren zur Eigenwert- und Eigenvektorberechnung. Ausgehend von einer Matrix $A^{(1)} = A \in \mathbb{C}^{n,n}$ führt man beim QR-Verfahren unitäre Ähnlichkeitstransformationen folgender Form durch:

Algorithmus 6 QR-Verfahren
Sei $A \in \mathbb{C}^{n,n}$ eine gegebene Matrix.

0. Setze $A^{(1)} := A$.

1. Für $l = 1, 2, \ldots$: Berechne

$$\begin{aligned} A^{(l)} &=: Q_l R_l, \quad Q_l \in \mathbb{C}^{n,n} \textit{ unitär}, \quad R_l \in \mathbb{C}^{n,n} \textit{ obere Dreiecksmatrix}, \\ A^{(l+1)} &:= R_l Q_l. \end{aligned} \tag{7.4}$$

In jedem Schritt ist also die Berechnung einer QR-Zerlegung

$$A^{(l)} = Q_l R_l, \quad R_l \in \mathbb{C}^{n,n} \text{ obere Dreiecksmatrix}, \quad Q_l \in \mathbb{C}^{n,n} \text{ unitär, also } Q_l^H = Q_l^{-1}$$

erforderlich. Eine solche Zerlegung kann mit Hilfe des Householder-Verfahrens berechnet werden, das wir für Interessierte am Ende dieses Kapitels kurz beschreiben.

7.3.1 Grundlegende Eigenschaften des QR-Verfahrens

Wir beginnen mit der offensichtlichen Feststellung, dass (7.4) tatsächlich eine Folge unitär ähnlicher Matrizen $A^{(l)}$ erzeugt.

Lemma 7.3.1 *Es seien Q_l und R_l von Algorithmus 6 erzeugt. Dann gilt mit den Bezeichnungen $Q_{1\dots l} := Q_1 Q_2 \cdots Q_l$, $R_{l\dots 1} := R_l R_{l-1} \cdots R_1$*

$$A^{(l+1)} = Q_l^{-1} A^{(l)} Q_l = Q_{1\dots l}^{-1} A Q_{1\dots l}, \quad l = 1, 2, \dots .$$

Beweis: Wegen (7.4) ist $R_l = Q_l^{-1} A^{(l)}$ und daher

$$A^{(l+1)} = R_l Q_l = Q_l^{-1} A^{(l)} Q_l.$$

Induktiv ergibt sich

$$A^{(l+1)} = Q_l^{-1} \cdots Q_1^{-1} A^{(1)} Q_1 \cdots Q_l = Q_{1\dots l}^{-1} A Q_{1\dots l}.$$

□

7.3.2 Konvergenz des QR-Verfahrens

Wir geben zunächst ein Resultat für Matrizen mit betragsmäßig getrennten Eigenwerten an. Unter gewissen Voraussetzungen konvergiert dann die vom QR-Verfahren generierte Folge $A^{(l)}$ nach unitärer Diagonalskalierung der Form $S_l^{-1} A^{(l)} S_l$ gegen eine obere Dreiecksmatrix U, wobei die Konvergenzgeschwindigkeit von der Trennung der Beträge der Eigenwerte abhängt.

Satz 7.3.2 *Die Matrix $A \in \mathbb{C}^{n,n}$ sei regulär mit betragsmäßig getrennten Eigenwerten $\lambda_1, \dots, \lambda_n$,*

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|.$$

Weiter seien $v_1, \dots, v_n$ zugehörige Eigenvektoren und die Inverse der Matrix $T = (v_1, \dots, v_n)$ besitze ohne Zeilenvertauschung eine LR-Faktorisierung. Dann gilt für das in Algorithmus 6 angegebene QR-Verfahren

$$A^{(l)} = S_l U S_l^{-1} + O(q^{l-1}) \quad \textit{für } l \to \infty, \quad q := \max_{j=1,\dots,n-1} \left| \frac{\lambda_{j+1}}{\lambda_j} \right|$$

mit einer oberen Dreiecksmatrix

$$U = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ & \ddots & \ddots & \vdots \\ & & \ddots & * \\ & & & \lambda_n \end{pmatrix}$$

und unitären Phasenmatrizen $S_l = \operatorname{diag}(\sigma_1^{(l)}, \dots, \sigma_n^{(l)})$, $|\sigma_i^{(l)}| = 1$. *Insbesondere gilt mit den Diagonaleinträgen* $a_{11}^{(l)}, \dots, a_{nn}^{(l)}$ *von* $A^{(l)}$

$$|a_{ii}^{(l)} - \lambda_i| = O(q^{l-1}).$$

Beweis: Siehe zum Beispiel Plato [Pl00]. □

Bemerkungen:

- Die zugehörigen Eigenvektoren kann man zum Beispiel durch Inverse Vektoriteration berechnen, wobei man jeweils die Diagonalelemente von $A^{(l)}$ als Shifts μ verwendet.
- Hat T^{-1} lediglich eine LR-Faktorisierung mit Zeilenvertauschungen, dann konvergiert das QR-Verfahren nach wie vor, die Eigenwerte erscheinen in der Diagonale der Grenzmatrix U jedoch unter Umständen in anderer Reihenfolge.
- Sind nicht alle Eigenwerte betragsmäßig getrennt, also etwa

  $$|\lambda_1| > \ldots > |\lambda_r| = |\lambda_{r+1}| > \ldots > |\lambda_n|,$$

  was zum Beispiel eintritt, wenn eine reelle Matrix A konjugiert komplexe Eigenwerte hat, dann konvergiert $S_l^{-1} A^{(l)} S_l$ mit Phasenmatrizen S_l außerhalb des mit $\times$ markierten Bereichs gegen eine Matrix der Form

  $$\begin{pmatrix} \lambda_1 & \cdots & * & \times & \times & * & \cdots & \\ & \ddots & \vdots & \vdots & \vdots & \vdots & & \\ & & \lambda_{r-1} & \times & \times & * & \cdots & \\ & & & \times & \times & * & \cdots & \\ & & & \times & \times & * & \cdots & \\ & & & & & \lambda_{r+2} & & \\ & & & & & & \ddots & \\ & & & & & & & \lambda_n. \end{pmatrix}$$

  Die Eigenwerte des Blocks $\begin{pmatrix} a_{r,r}^{(l)} & a_{r,r+1}^{(l)} \\ a_{r+1,r}^{(l)} & a_{r+1,r+1}^{(l)} \end{pmatrix}$ konvergieren gegen λ_r und λ_{r+1}.
- Die Konvergenz des QR-Verfahrens ist sehr langsam, wenn die Trennung der Eigenwerte schlecht ist. Die Konvergenz der letzten Zeile gegen $(0, \dots, 0, \lambda_n)$ kann durch *Shift*-Techniken entscheidend verbessert werden, auf die wir nun kurz eingehen.

7.3.3 Shift-Techniken

Eine genauere Analyse zeigt, dass die letzte Zeile von $A^{(l)}$ die Form hat $(O(|\lambda_n/\lambda_{n-1}|^{l-1}), a_{nn}^{(l)})$. Ist also $|\lambda_n| \ll |\lambda_{n-1}|$, dann konvergiert $a_{n,j}^{(l)}$, $1 \leq j < n$ sehr schnell gegen 0 und

$a_{nn}^{(l)}$ sehr schnell gegen λ_n. Nach genauer Bestimmung von λ_n kann man dann mit dem $(n-1)\times(n-1)$-Block von $A^{(l)}$ zur Bestimmung von λ_{n-1} fortfahren.

Um die Trennung von λ_n und λ_{n-1} zu verbessern, wendet man das QR-Verfahren in jedem Schritt auf $A^{(l)} - \mu_l I$ an mit $\mu_l \approx \lambda_n$ und korrigiert den Shift anschließend. Anstelle von (7.4) berechnet man also mit einem *Shift* $\mu_l \approx \lambda_n$

$$\begin{aligned} A^{(l)} - \mu_l I &=: Q_l R_l, \quad Q_l \in \mathbb{C}^{n,n} \text{ unitär}, \quad R_l \in \mathbb{C}^{n,n} \text{ obere Dreiecksmatrix}, \\ A^{(l+1)} &:= R_l Q_l + \mu_l I. \end{aligned}$$

Man prüft leicht nach, dass wieder gilt $A^{(l+1)} = Q_l^{-1} A^{(l)} Q_l$.

Verbreitete Shift-Strategie: Eine effizente Shift-Strategie erhält man, wenn man μ_l als denjenigen Eigenwert von $\left(\begin{smallmatrix} a_{n-1,n-1}^{(l)} & a_{n-1,n}^{(l)} \\ a_{n,n-1}^{(l)} & a_{n,n}^{(l)} \end{smallmatrix}\right)$ wählt, der am nächsten bei $a_{n,n}^{(l)}$ liegt. Im Zweifelsfall wähle den mit positivem Imaginärteil.

Das QR-Verfahren mit Shift liefert recht schnell eine Matrix $A^{(l)}$, deren letzte Zeile auf hohe Genauigkeit mit $(0,\ldots,0,\lambda_n)$ übereinstimmt. Man wendet nun das QR-Verfahren mit Shift auf den oberen linken $(n-1)\times(n-1)$-Block von $A^{(l)}$ zur Bestimmung von λ_{n-1} an und so fort.

Bemerkung: Das QR-Verfahren mit Shift gilt zur Zeit als eines der besten Iterationsverfahren zur Lösung des vollständigen Eigenwertproblems.

Berechnung der Eigenvektoren: Die Eigenvektoren kann man nun zum Beispiel wieder durch Inverse Vektoriteration bestimmen, wobei man als Shifts μ die vom QR-Verfahren berechneten Eigenwerte verwendet.

7.3.4 Berechnung einer QR-Zerlegung (Ergänzung für Interessierte)

Wir geben zum Abschluss ein numerisches Verfahren an zur Berechnung einer

QR-Zerlegung:

Für $B \in \mathbb{C}^{n,n}$ bestimme eine unitäre Matrix $Q \in \mathbb{C}^{n,n}$ und eine obere Dreiecksmatrix $R \in \mathbb{C}^{n,n}$ mit

$$B = QR. \tag{7.5}$$

Householder-Verfahren zur Berechnung einer QR-Zerlegung:

Beim Householder-Verfahren berechnet man (7.5) in $n-1$ Schritten:

Initialisierung:

$$B^{(0)} := B = \left(\begin{array}{c|cc} & * & \cdots \\ b^{(0)} & \vdots & \\ & * & \cdots \end{array}\right)$$

Schritt 0: Bestimme eine unitäre Matrix T_0 (siehe (7.7), (7.8)) mit

$$B^{(1)} := T_0 B^{(0)} = \left(\begin{array}{c|ccc} * & * & * & \cdots \\ \hline 0 & * & * & \cdots \\ \vdots & \vdots & \vdots & \\ 0 & * & * & \cdots \end{array}\right) =: \left(\begin{array}{c|c|c} B_1^{(1)} & \multicolumn{2}{c}{B_2^{(1)}} \\ \hline 0 & & \\ \vdots & b^{(1)} & B_3^{(1)} \\ 0 & & \end{array}\right)$$

Schritt 1: Bestimme eine unitäre Matrix T_1 (siehe (7.7), (7.8)) mit

$$B^{(2)} := T_1 B^{(1)} = \left(\begin{array}{cc|ccc} * & * & * & * & \cdots \\ 0 & * & * & * & \cdots \\ \hline 0 & 0 & * & * & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & * & * & \cdots \end{array}\right) =: \left(\begin{array}{cc|c|c} \multicolumn{2}{c|}{B_1^{(2)}} & \multicolumn{2}{c}{B_2^{(2)}} \\ \hline 0 & 0 & & \\ \vdots & \vdots & b^{(2)} & B_3^{(2)} \\ 0 & 0 & & \end{array}\right)$$

Schritt k, $k = 2, \ldots, n-2$: Bestimme eine unitäre Matrix T_k (siehe (7.7), (7.8)) mit

(7.6)
$$B^{(k+1)} := T_k B^{(k)} = \left(\begin{array}{ccc|cc} * & \cdots & * & * & \cdots \\ & \ddots & \vdots & & \\ 0 & & * & * & \cdots \\ \hline 0 & \cdots & 0 & * & \cdots \\ \vdots & & \vdots & \vdots & \\ 0 & \cdots & 0 & * & \cdots \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c} * \\ \ddots \\ 0 \end{array}}\right\} k+1 \\ \left.\vphantom{\begin{array}{c} 0 \\ \vdots \\ 0 \end{array}}\right\} n-(k+1) \end{array}$$

$$= \left(\begin{array}{ccc|c|c} & B_1^{(k+1)} & & \multicolumn{2}{c}{B_2^{(k+1)}} \\ \hline 0 & \cdots & 0 & & \\ \vdots & & \vdots & b^{(k+1)} & B_3^{(k+1)} \end{array}\right).$$

Ergebnis: $R := B^{(n-1)}$, $Q := (T_{n-2} \cdot \ldots \cdot T_0)^H = T_0^H \cdot \ldots \cdot T_{n-2}^H$.

Rechtfertigung des Verfahrens:

Dann gilt tatsächlich

$R = B^{(n-1)} =$ obere Dreiecksmatrix, $\quad Q = T_0^H \cdot \ldots \cdot T_{n-2}^H$ unitär als Produkt unitärer Matrizen und

$$R = B^{(n-1)} = \underbrace{T_{n-2} \cdot \ldots \cdot T_0}_{=Q^H} B = Q^H B, \quad \text{also} \quad QR = B.$$

Berechnung der Transformationen T_k:

Es bleibt, die Berechnung von T_k anzugeben. Beim Householder-Verfahren wählt man jeweils T_k von der Form

(7.7)
$$T_k = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & H_k \end{array} \right)$$

mit

$$I_k = \text{Einheitsmatrix in } \mathbb{R}^{k,k},$$

und $H_k \in \mathbb{R}^{n-k,n-k}$ als **Householder Transformation** der Form

(7.8)
$$H_k = I - \frac{2}{w_k^H w_k} w_k w_k^H, \quad w_k = b^{(k)} + \sigma_k \|b^{(k)}\|_2 \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}, \quad \sigma_k = \begin{cases} 1 & \text{falls } b_1^{(k)} = 0, \\ \frac{b_1^{(k)}}{|b_1^{(k)}|} & \text{sonst.} \end{cases}$$

Man kann zeigen, dass mit dieser Wahl gilt

$$H_k \ \text{ unitär und hermitesch}, \quad H_k b^{(k)} = \begin{pmatrix} \omega_k \|b^{(k)}\|_2 \\ 0 \\ \vdots \end{pmatrix}, \quad \omega_k \in \mathbb{C}, \ \ |\omega_k| = 1.$$

Man sieht leicht, dass dann tatsächlich jeweils $B^{(k+1)}$ die Form (7.6) hat.

Statistik

Statistische Methoden werden in allen empirischen Wissenschaften verwendet. In der ”Beschreibenden Statistik” geht es zunächst darum, Beobachtungsdaten übersichtlich darzustellen und durch Berechnung von Kenngrößen (Mittelwerte, Streuungen) zu charakterisieren.

Da jedoch Beobachtungsdaten in der Regel zufallsbehaftet sind (zufällige Meßfehler bei Experimenten, Möglichkeit unterschiedlicher Ergebnisse), besteht das Risiko von Fehlschlüssen. Die sogenannte ”Schließende Statistik” stellt daher Methoden bereit, bei denen diese Fehlerrisiken abgeschätzt werden können. Die Abschätzung dieser Risiken beruht auf mathematischen Modellen für zufallsabhängige Vorgänge, die in der ”Wahrscheinlichkeitstheorie” behandelt werden.

Dieser Teil des Skripts basiert in Teilen auf dem Buch v. Finckenstein, Lehn, Schellhaas, Wegmann; *Arbeitsbuch für Ingenieure II*, Teubner Verlag, 2006, das als vertiefende Literatur empfohlen wird.

Kapitel 8

Grundbegriffe der Statistik und Wahrscheinlichkeitstheorie

8.1 Messreihen

Im Folgenden werden zwei Typen von *Merkmalen* betrachtet:

- *quantitativ-diskrete*, z.B. Alter in Jahren, Geschosszahl eines Gebäudes,...
  Die *Merkmalausprägungen* sind dann ganze Zahlen.
- *quantitativ-stetige*, z.B. Gebäudehöhe, Temperatur,...
  Die *Merkmalausprägungen* sind dann reelle Zahlen.

Am Beginn einer statistischen Untersuchung steht immer die mehrfache Beobachtung eines Merkmals. Das Beobachtungsergebnis ist dann eine *Messreihe* von n Zahlen

$$x_1, x_2, \ldots, x_n.$$

Definition 8.1.1 *Sei $x_1, x_2, \ldots, x_n$ eine Messreihe. Ordnet man die Werte der Messreihe der Größe nach, so ensteht die zugehörige* geordnete Messreihe

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

Sie besteht aus den gleichen Zahlen, aber so umgeordnet, dass gilt $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Die empirische Verteilungsfunktion *einer Messreihe $x_1, x_2, \ldots, x_n$ ist die Funktion*

$$F(z; x_1, x_2, \ldots, x_n) = \frac{\textit{Zahl der } x_i \textit{ mit } x_i \leq z}{n} = \frac{\max\left\{i \, : \, x_{(i)} \leq z\right\}}{n}.$$

Wählt man $r-1$ Zahlen $a_1 < a_2 < \ldots < a_{r-1}$, so entsteht die Unterteilung von $\mathbb{R}$ in r Klassen

$$\mathbb{R} =]-\infty, a_1] \cup]a_1, a_2] \cup \cdots \cup]a_{r-2}, a_{r-1}] \cup]a_{r-1}, \infty[.$$

Mit der Abkürzung $F(z) = F(z; x_1, x_2, \dots, x_n)$ ergeben sich dann die *relativen Klassenhäufigkeiten* für diese r Klassen zu

$$F(a_1), F(a_2) - F(a_1), \dots, F(a_{r-1}) - F(a_{r-2}), 1 - F(a_{r-1}).$$

Wählt man noch zwei zusätzliche Zahlen

$$a_0 < \min\{a_1, x_{(1)}\}, \quad a_r > \max\{a_{r-1}, x_{(n)}\},$$

so können die relativen Klassenhäufigkeiten in einem *Histogramm* dargestellt werden: über jedem der Intervalle $]a_{j-1}, a_j]$, $j = 1, \dots, r$, wird ein Rechteck erreichtet, das die jeweilige Klassenhäufigkeit als Fläche hat.

Beispiel:

Die zur Messreihe

$$2.2 \quad 4.5 \quad 0.8 \quad 1.7 \quad 5.8 \quad 1.2 \quad 5.6 \quad 2.5 \quad 3.9 \quad 1.7$$

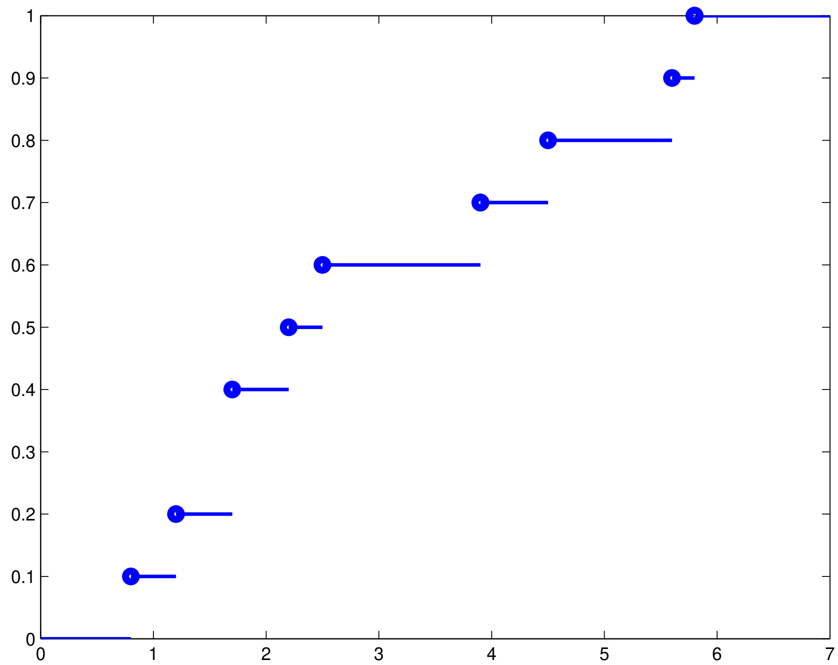
gehörige geordnete Messreihe ist

$$0.8 \quad 1.2 \quad 1.7 \quad 1.7 \quad 2.2 \quad 2.5 \quad 3.9 \quad 4.5 \quad 5.6 \quad 5.8.$$

Hierbei ist

$$x_{(1)} = x_3,\ x_{(2)} = x_6,\ x_{(3)} = x_{(4)} = x_4 = x_{10},\ x_{(5)} = x_1,\ x_{(6)} = x_8,$$
$$x_{(7)} = x_9,\ x_{(8)} = x_2,\ x_{(9)} = x_7,\ x_{(10)} = x_5.$$

Die empirische Verteilungsfunktion hat folgenden Graphen:



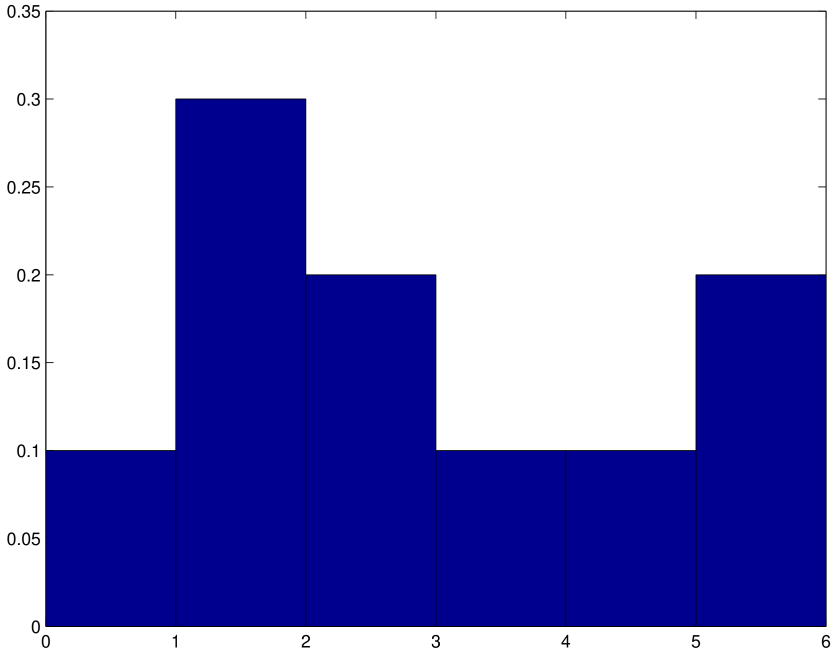
Zu der Unterteilung

$$0 < 1 < 2 < 3 < 4 < 5 < 6$$

ergeben sich die Klassenhäufigkeiten

$$\frac{1}{10}, \quad \frac{4-1}{10} = \frac{3}{10}, \quad \frac{6-4}{10} = \frac{2}{10}, \quad \frac{7-6}{10} = \frac{1}{10}, \quad \frac{8-7}{10} = \frac{1}{10}, 1 - \frac{8}{10} = \frac{2}{10},$$

also das Histogramm



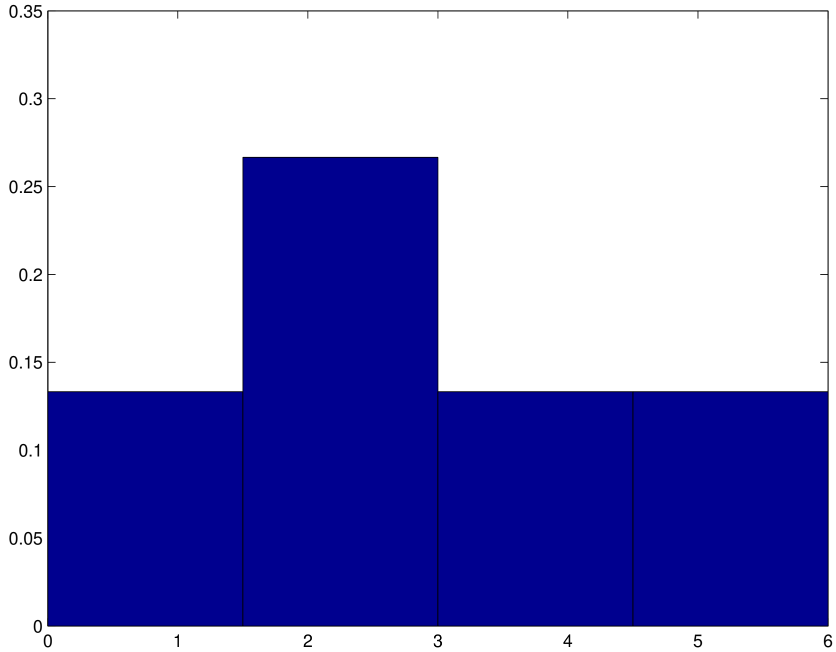
Zur Unterteilung

$$0 < 1.5 < 3 < 4.5 < 6$$

ergeben sich die Klassenhäufigkeiten

$$\frac{2}{10}, \quad \frac{6-2}{10} = \frac{4}{10}, \quad \frac{8-6}{10} = \frac{2}{10} \quad 1 - \frac{8}{10} = \frac{2}{10},$$

mit zugehörigem Histogramm



.

8.2 Lage- und Streumaßzahlen

Zur Beschreibung und Charakterisierung von Messreihen dienen Lage- und Streumaßzahlen.

Sei $x_1, \ldots, x_n$ eine Messreihe.

8.2.1 Lagemaßzahlen

Beispiele für Lagemaßzahlen sind

Arithmetisches Mittel:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Median:

$$\tilde{x} = \begin{cases} x_{(\frac{n}{2})}, & \text{falls } n \text{ gerade,} \\ x_{(\frac{n+1}{2})}, & \text{falls } n \text{ ungerade.} \end{cases}$$

p**-Quantil** $(0 < p < 1)$**:**

$$x_p = \begin{cases} x_{(np)}, & \text{falls } np \text{ ganzzahlig,} \\ x_{([np]+1)}, & \text{falls } np \text{ nicht ganzzahlig.} \end{cases}$$

Hierbei ist

$$[x] = \max\{z \in \mathbb{Z} : z \leq x\} \quad (Gaussscheklammer)$$

die größte ganze Zahl $\leq x$.

α-gestutztes Mittel $(0 < \alpha < 0.5)$**:**

$$\bar{x}_\alpha = \frac{1}{n-2k}(x_{(k+1)} + \ldots + x_{(n-k)}), \quad k = [n\alpha].$$

Das 0.25-Quantil $x_{0.25}$ wird als *unteres Quartil*, das 0.75-Quantil $x_{0.75}$ wird als *oberes Quartil* bezeichnet.

8.2.2 Streuungsmaße

Beispiele für Streuungsmaße sind:

Empirische Varianz oder **empirische Stichprobenvarianz:**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Empirische Streuung:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Spannweite:

$$v = x_{(n)} - x_{(1)}$$

Quartilabstand:

$$q = x_{0.75} - x_{0.25}.$$

8.2.3 Zweidimensionale Messreihen

Werden bei einer statistischen Erhebung zwei verschiedene Merkmale gleichzeitig ermittelt, so entstehen *zweidimensionale Messreihen*, d.h. endliche Folgen

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Analog wie oben definieren wir folgende Maßzahlen:

Arithmetische Mittel:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n), \quad \bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n)$$

Empirische Varianzen:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2$$

Empirische Streuungen:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2}, \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2}$$

Weiterhin sind auch folgende Maßzahlen zwischen den x_i und y_i von Bedeutung:

Empirische Kovarianz:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})\,(y_i-\bar{y})$$

Empirischer Korrelationskoeffizient:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}.$$

Bemerkung: Es gilt immer

$$-1 \le r_{xy} \le 1.$$

Beweis: Sei $u = (x_i-\bar{x})_{1\le i\le n}$, $v = (y_i-\bar{y})_{1\le i\le n}$. Dann gilt nach der Cauchy-Schwartz-Ungleichung

$$r_{xy} = \frac{u^T v}{\|u\|_2\|v\|_2} \in [-1,1].$$

□

Bemerkung: Die empirische Varianz s^2 lässt sich auch nach der Formel berechnen

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \tag{8.1}$$

Ebenso gilt für die empirische Kovarianz

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right) \tag{8.2}$$

Beweis: Es gilt

$$(n-1)\,s^2 = \sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.$$

Die Formel für s_{xy} folgt ganz analog. □

Zur Veranschaulichung einer zweidimensionalen Messreihe dient das *Punktediagramm*, bei dem die Punkte (x_i, y_i), $i = 1, \ldots, n$, als Punkte in einem $x-y$-Koordinatensystem eingetragen werden.

8.2.4 Regressionsgerade

Der Korrelationskoeffizient r_{xy} gibt Hinweise, ob die y-Werte tendenziell monoton wachsend oder monoton fallend von den x-Werten abhängen. Für diesen Zusammenhang soll nun angenommen werden, dass er sich im wesentlichen durch eine lineare Gleichung der Form

$$y = ax + b$$

beschreiben lässt. Wir nehmen also an, dass sich die Datenpunkte um eine Gerade mit der Steigung a und Achsenabschnitt b gruppieren. Wir wollen nun a und b bestimmen, damit die Gerade möglichst gut zu den Datenpunkten passt. Das Quadrat des Abstands zwischen Datenpunkt (x_i, y_i) und einem Punkt $(x_i, ax_i + b)$ auf der Geraden mit demselben x-Wert ist $(y_i - ax_i - b)^2$. Steigung a und Achsenabschnitt b der Geraden sollen nun so bestimmt werden, dass die Summe all dieser Quadrate

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

minimal wird. Wir erhalten dann die sogenannte *Regressionsgerade*. Wir suchen also eine Lösung des Problems

$$\min_{(a,b)\in\mathbb{R}^2} S(a,b). \tag{8.3}$$

Bestimmung der Minimalstelle:

$$\frac{\partial S(a,b)}{\partial a} = \sum_{i=1}^{n} 2(y_i - ax_i - b)\,(-x_i) = -2\sum_{i=1}^{n}(x_i y_i - ax_i^2 - bx_i) = 0.$$

$$\frac{\partial S(a,b)}{\partial b} = \sum_{i=1}^{n} 2(y_i - ax_i - b)\,(-1) = -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0.$$

Die zweite Gleichung ergibt

$$n\bar{y} - an\bar{x} - nb = 0,$$

also

$$b = \bar{y} - a\bar{x}.$$

Einsetzen in die erste Gleichung ergibt

$$\sum_{i=1}^{n}(x_i y_i - ax_i^2 - \bar{y}x_i + a\bar{x}x_i) = 0$$

und somit

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = a\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

Insgesamt ergibt sich für die Lösung $(\hat{a}, \hat{b})$ von (8.3) unter Verwendung von (8.1), (8.2):

Berechnung der Regressionsgerade:

$$y = \hat{a}\, x + \hat{b},$$

mit

$$\hat{a} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{s_{xy}}{s_x^2}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

Wie bereits erwähnt, heisst die so gefundene Gerade *Regressionsgerade*. Die Abweichungen der Punkte (x_i, y_i) von der Regressionsgerade in vertikaler Richtung

$$r_i = y_i - \hat{a}x_i - \hat{b}, \quad i = 1, \dots, n$$

heißen *Residuen*. Nach kurzer Rechnung erhält man folgende

Formel für das Residuenquadrat:

$$\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - \bar{y})^2\, (1 - r_{xy}^2).$$

Die vertikale Abweichung von der Regressionsgerade hängt also eng mit dem Korrelationskoeffizienten r_{xy} zusammen. Für die extremen Werte $r_{xy} = 1$ bzw. $r_{xy} = -1$ verschwinden die Residuen, alle Punkte (x_i, y_i) liegen also auf der Regressionsgeraden.

Da die Werte

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad \text{und} \quad \hat{a} = \frac{s_{xy}}{s_x^2}$$

gleiches Vorzeichen haben, ergibt sich also für $r_{xy} > 0$ eine streng monoton steigende, für $r_{xy} < 0$ eine streng monoton fallende und für $r_{xy} = 0$ eine horizontale Regressionsgerade.

Das Vorzeichen von r_{xy} gibt also den Trend der Abhängigkeit der y-Werten von den x-Werten an.

8.3 Zufallsexperimente und Wahrscheinlichkeit

8.3.1 Zufallsexperimente

Ein Vorgang, der so genau beschrieben ist, dass er als beliebig oft wiederholbar betrachtet werden kann, und dessen Ergebnisse vom Zufall abhängen, nennen wir *Zufallsexperiment*. Es wird angenommen, dass die Menge der möglichen Ergebnisse soweit bekannt ist, dass jedem Ergebnis ein Element ω einer Menge Ω zugeordnet werden kann.

Definition 8.3.1 *Ω heißt* Ergebnismenge*, seine Elemente ω* Ergebnisse*. Teilmengen $A \subset \Omega$ heißen* Ereignisse*. Ein Ereignis $A \subset \Omega$ tritt ein, falls ein Ergebnis $\omega \in A$ beobachtet wird.*

Beispiele:

1. Wurf eines Würfels: $\Omega = \{1, 2, 3, 4, 5, 6\}$. Das Ereignis $A = \{1, 3, 5\}$ tritt ein, falls eine ungerade Zahl gewürfelt wird.

2. Wurf zweier unterscheidbarer Würfel: $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$. Ω hat $6 \cdot 6 = 36$ Elemente.

3. Wurf zweier nicht unterscheidbarer Würfel: $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}, \ i \leq j\}$. Ω hat 21 Elemente.

4. Lebensdauer eines Gerätes: $\Omega =]0, \infty[$. Das Ereignis $A = \{\omega \in \mathbb{R} : \omega > 100\}$ tritt ein, wenn das Gerät mehr als 100 Stunden fehlerfrei funktioniert.

Definition 8.3.2 *Das aus zwei Ereignissen A und B* zusammengesetzte Ereignis *$A \cup B$ tritt ein, falls ein Ergebnis ω mit $\omega \in A$ oder $\omega \in B$ (d.h. $\omega \in A \cup B$) beobachtet wird.*

Entsprechend tritt das Ereignis $A \cap B$ ein, falls ein Ergebnis ω mit $\omega \in A$ und $\omega \in B$ (d.h. $\omega \in A \cap B$) beobachtet wird.

$A^c = \Omega \setminus A$ heißt zu A komplementäres Ereignis.

Zwei Ereignisse A und B heißen unvereinbar*, falls $A \cap B = \emptyset$.*

Die leere Menge $\emptyset$ heißt unmögliches Ereignis *und Ω das* sicheres Ereignis.

Die einelementigen Mengen $\{\omega\}$ von Ω heißen Elementarereignisse.

Auch für Folgen $A_1, A_2, \ldots$ von Ereignissen definieren wir das zusammengesetzte Ereignis $\bigcup_{i=1}^{\infty} A_i$, das eintritt, wenn mindestens ein A_i eintritt, und das Ereignis $\bigcap_{i=1}^{\infty} A_i$, das eintritt, wenn alle A_i zugleich eintreten.

8.3.2 Wahrscheinlichkeit

Fragt man im Falle der Betriebsdauer eines Gerätes danach, wie wahrscheinlich es ist, dass das Gerät exakt nach 100 Stunden (keinen Augenblick früher oder später!) seinen ersten Defekt hat, dann ist dies praktisch ausgeschlossen. Fragt man jedoch danach, dass der erste Defekt zwischen 90 und 100 Stunden auftritt, also nach der Wahrscheinlichkeit des Ereignisses $A = [90, 100]$, dann ist dies eine sachgerechte Fragestellung.

Dies zeigt, dass es sinnvoll ist, die Wahrscheinlichkeit des Eintretens von Ereignissen zu betrachten. Wir haben dabei die Vorstellung, dass die Wahrscheinlichkeit $P(A)$ für das

Eintreten des Ereignisses A immer genauer der relativen Häufigkeit des Eintretens von A in Versuchsserien entspricht, je länger die Versuchsreihe wird.

Dazu betrachten wir ein System $\mathcal{A}$ von Ereignissen (es muss nicht die Potenzmenge $\mathcal{P}(\Omega)$, also die Menge aller Teilmengen von Ω sein!), das folgende Eigenschaften hat:

Definition 8.3.3 *ein System $\mathcal{A} \subset \mathcal{P}(\Omega)$ von Ereignissen heißt σ-*Algebra*, wenn gilt:*

a) $\Omega \in \mathcal{A}$.

b) Falls $A \in \mathcal{A}$, dann gilt auch $A^c \in \mathcal{A}$.

c) Mit jeder Folge $A_1, A_2, \ldots \in \mathcal{A}$ gilt auch $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Bemerkung: Sind $A, B \in \mathcal{A}$, dann ist wegen b) und c) auch

$$A \cap B = (A^c \cup B^c)^c \in \mathcal{A}.$$

□

Eine σ-Algebra erlaubt gerade die Verknüpfungen von Ereignissen, die in der Praxis nützlich sind. Um jedem Ereignis eine Wahrscheinlichkeit zuzuordnen, betrachtet man eine Abbildung $P : \mathcal{A} \to \mathbb{R}$ mit folgenden Eigenschaften:

Definition 8.3.4 *Eine Abbildung $P : \mathcal{A} \to \mathbb{R}$ heißt* Wahrscheinlichkeitsmaß*, wenn sie den folgenden* Axiomen von Kolmogorov *genügt:*

a) $P(A) \geq 0$ *für* $A \in \mathcal{A}$,

b) $P(\Omega) = 1$,

c) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ *für paarweise unvereinbare* $A_1, A_2, \ldots \in \mathcal{A}$.

Bemerkung: c) umfasst auch endliche disjunkte Vereinigungen $\bigcup_{i=1}^{n} A_i$ durch die Wahl $A_i = \emptyset$, $i \geq n+1$.

Aus diesen Axiomen folgen nützliche Regeln für das Rechnen mit Wahrscheinlichkeiten von Ereignissen $A, B, A_1, \ldots, A_n$:

$$\begin{aligned}
&P(\emptyset) = 0,\\
&0 \leq P(A) \leq 1,\\
&P(A^c) = 1 - P(A),\\
&A \subset B \implies P(A) \leq P(B),\\
&P(A \cup B) = P(A) + P(B) - P(A \cap B),\\
&P(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i), \quad \text{falls } A_1, \ldots, A_n \text{ paarweise unvereinbar (Additivität).}
\end{aligned}$$

Falls die Ergebnismenge endlich ist, also $\Omega = \{\omega_1, \dots, \omega_n\}$, und man wie beim Würfelwurf annehmen kann, dass jedes Elementarereignis $\{\omega_i\}$ gleich wahrscheinlich ist, dann folgt aus der Additivität

$$P(\{\omega_i\}) = \frac{1}{n}, \quad i = 1, \dots, n$$

und für beilebige Ereignisse mit Elementzahl $\#A$ gilt

$$P(A) = \frac{\text{Elementzahl von } A}{n} = \frac{\#A}{\#\Omega}.$$

Die Annahme gleicher Wahrscheinlichkeit für die Elementarereignisse heißt *Laplace-Annahme*.

Beispiel: Drei Würfel werden geworfen. Wie groß ist die Wahrscheinlichkeit, dass die Würfelsumme 11 ist?

Wir wählen $\Omega = \{(i, j, k) \,:\, i, j, k = 1, \dots, 6\}$. Dann ist $\#\Omega = 6^3 = 216$.

A sei das Ereignis "Würfelsumme ist 11".

Möglichkeiten für die drei Summanden (der Größe nach geordnet):

$$11 = 1 + 4 + 6 = 1 + 5 + 5 = 2 + 3 + 6 = 2 + 4 + 5 = 3 + 3 + 5 = 3 + 4 + 4.$$

Wir summieren die Anzahl der Tripel auf, die auf die angegebenen Summanden führen:

$$\#A = 6 + 3 + 6 + 6 + 3 + 3 = 27.$$

Dies ergibt

$$P(A) = \frac{\#A}{\#\Omega} = \frac{27}{216} = \frac{1}{8} = 0,125.$$

8.3.3 Elementare Formeln der Kombinatorik

Zur Berechnung der Elementezahlen von Ereignissen werden häufig kombinatorische Formeln verwendet. Wir geben einige wichtige Formeln an:

Sei Ω eine Menge mit n Elementen und $k \in \mathbb{N}$.

Geordnete Probe mit Wiederholungen:
Ein k-Tupel $(x_1, \dots, x_k)$ mit $x_i \in \Omega$, $i = 1, \dots, k$, heißt *geordnete Probe* von Ω vom Umfang k *mit Wiederholungen*. Es gibt

$$n^k \quad \text{(Anzahl geordneter Proben mit Wiederholungen)}$$

solcher Proben (für jede Stelle gibt es n Möglichkeiten).

Geordnete Probe ohne Wiederholungen:
Ein k-Tupel $(x_1, \dots, x_k)$, $k \leq n$, mit $x_i \in \Omega$, $i = 1, \dots, k$, und $x_i \neq x_j$ für $i \neq j$ heißt *geordnete Probe* von Ω vom Umfang k *ohne Wiederholungen*. Es gibt

$$n(n-1)(n-2) \cdot \ldots \cdot (n-k+1) \quad \text{(Anzahl geordneter Proben ohne Wiederholungen)}$$

solcher Proben (für die erste Stelle gibt es n Möglichkeiten, für die zweite $n-1$, usw.).

Im Fall $k=n$ spricht man von einer *Permutation* der Menge Ω. Davon gibt es

$$n! = n(n-1)(n-2)\cdot\ldots\cdot 2\cdot 1 \quad \text{(Anzahl von Permutationen)}$$

Ungeordnete Probe ohne Wiederholungen:
Eine Teilmenge $\{x_1,\ldots,x_k\}$, $k\le n$, von Ω heißt *ungeordnete Probe* von Ω vom Umfang k *ohne Wiederholungen*. Es gibt

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} \quad \text{(Anzahl } k\text{-elem. Teilmengen)}$$

solcher Proben (es gibt $n(n-1)(n-2)\cdot\ldots\cdot(n-k+1)$ geordnete Proben, aber jeweils $k!$ bestehen aus den gleichen k Elementen).

Beispiel:

1. Wie viele Möglichkeiten gibt es, k Einsen und $n-k$ Nullen anzuordnen?

   Lösung: Jede Anordnung der Einsen entspricht einer k-elementigen Teilmenge von $\{1,\ldots,n\}$, welche die Positionen der Einsen angibt. Also: $\binom{n}{k}$ Möglichkeiten.

2. Beim Austeilen gemischter Karten (32 Karten, davon 4 Asse) sei A das Ereignis "die ersten drei Karten sind Asse". Dann gilt unter der Laplace-Annahme

$$P(A) = \frac{4\cdot 3\cdot 2\cdot 29!}{32!} = \frac{24}{32\cdot 31\cdot 30} = \frac{1}{1240}.$$

8.4 Bedingte Wahrscheinlichkeit, Unabhängigkeit

8.4.1 Bedingte Wahrscheinlichkeit

Seien A, B zwei Ereignisse mit $P(B) > 0$. Oft interessiert die Wahrscheinlichkeit von A unter der Bedingung, dass B eintritt. Man definiert diese *bedingte Wahrscheinlichkeit* $P(A|B)$ von A unter der Bedingung B durch

Bedingte Wahrscheinlichkeit von A unter der Bedingung B:

$$P(A|B) = \frac{P(A\cap B)}{P(B)}.$$

Beispiel: Beim Ziehen von einem gemischten Kartenstapel (32 Karten, 4 Asse) betrachte die Ereignisse A "die zweite Karte ist ein Ass" und B "die erste Karte ist ein Ass". Dann gilt

$$P(B) = \frac{4\cdot 31!}{32!} = \frac{1}{8}, \quad P(A\cap B) = \frac{4\cdot 3\cdot 30!}{32!} = \frac{12}{32\cdot 31}.$$

Dies ergibt

$$P(A|B) = \frac{12 \cdot 8}{32 \cdot 31} = \frac{3}{31}.$$

Direkte Rechnung: Wenn schon ein Ass gezogen ist, dann ist die Wahrscheinlichkeit, dass die zweite Karte wieder ein Ass ist

$$P(A|B) = \frac{3 \cdot 30!}{31!} = \frac{3}{31}.$$

Im Folgenden seien $A_1, \dots, A_n$ paarweise unvereinbare Ereignisse, d.h. $A_i \cap A_j = \emptyset$ für $i \neq j$, und es sei $\bigcup_{i=1}^n A_i = \Omega$. Man spricht von einer *vollständigen Ereignisdisjunktion*.

Es gelten die folgenden Rechenregeln:

Regel von der vollständigen Wahrscheinlichkeit:

$A_1, \dots, A_n$ sei eine vollständige Ereignisdisjunktion mit $P(A_i) > 0$, $i = 1, \dots, n$. Dann gilt:

$$P(B) = \sum_{i=1}^n P(A_i) \cdot P(B|A_i).$$

Beweis: Es gilt $P(A_i) \cdot P(B|A_i) = P(B \cap A_i)$. Die Mengen $C_i = B \cap A_i$ sind paarweise disjunkt mit $\bigcup_{i=1}^n C_i = B$. Wegen der Additivität gilt also

$$\sum_{i=1}^n P(A_i) \cdot P(B|A_i) = \sum_{i=1}^n P(C_i) = P\left(\bigcup_{i=1}^n C_i\right) = P(B).$$

□

Formel von Bayes:

$A_1, \dots, A_n$ sei eine vollständige Ereignisdisjunktion mit $P(A_i) > 0$, $i = 1, \dots, n$, und B sei ein Ereignis mit $P(B) > 0$. Dann gilt für $i = 1, \dots, n$:

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{k=1}^n P(A_k) \cdot P(B|A_k)}.$$

Beweis: Der Nenner ist $P(B)$ nach der Regel von der vollständigen Wahrscheinlichkeit. Also ist die rechte Seite gegebn durch

$$\frac{P(A_i) \cdot P(B|A_i)}{P(B)} = \frac{P(A_i) \cdot P(B \cap A_i)/P(A_i)}{P(B)} = \frac{P(B \cap A_i)}{P(B)} = P(A_i|B).$$

□

Beispiel: Bei einer Reihenuntersuchung sind die Ereignisse A: "untersuchter Patient ist erkrankt" und B: "Befund positiv" von Interesse. Es sei $P(A) = 0,001$ die Wahrscheinlichkeit, dass ein Patient erkrankt ist. Weiter seien $P(B|A) = 0,92$ und $P(B|A^c) = 0,01$

die Wahrscheinlichkeiten für einen positiven Befund bei einem erkrankten bzw. nicht erkrankten Patienten.

Gesucht ist die bedingte Wahrscheinlichkeit, dass ein Patient bei einem positiven Befund tatsächlich erkrankt ist, also $P(A|B)$.

Mit $A_1 = A$, $A_2 = A^c$ ergibt die Bayessche Formel

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)} = \frac{0,001 \cdot 0,92}{0,001 \cdot 0,92 + 0,999 \cdot 0,01} = 0,0844.$$

Multiplikationsformel:

$A_1, \ldots, A_n$ seien Ereignisse mit $P(A_1 \cap A_2 \cap \cdots \cap A_n) > 0$. Dann gilt

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}).$$

Beweis: Vollständige Induktion nach n: Für $n = 2$ gilt

$$P(A_1) \cdot P(A_2|A_1) = P(A_1 \cap A_2).$$

Induktionsschritt:

$$\begin{aligned} &P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1}) \\ \overset{\text{Ind.Ann.}}{=}\; &P(A_1 \cap \cdots \cap A_{n-1}) \cdot P(A_n|A_1 \cap \cdots \cap A_{n-1}) = P(A_n \cap A_1 \cap \cdots \cap A_{n-1}). \end{aligned}$$

□

8.4.2 Unabhängigkeit

Beim zweifachen Werfen eines Würfels erkennt man, dass die Ereignisse

$$A = \text{"1 beim zweiten Wurf"}, \quad B = \text{"1 beim ersten Wurf"}$$

von völlig unabhängig ablaufenden Teilexperimenten bestimmt wird und für die bedingte Wahrscheinlichkeit gilt

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6 \cdot 1/6}{1/6} = 1/6 = P(A).$$

Wir haben also

$$P(A \cap B) = P(A) \cdot P(B).$$

Dies motiviert die

Definition 8.4.1 *Zwei Ereignisse A und B heißen* unabhängig, *falls gilt*

$$P(A \cap B) = P(A) \cdot P(B).$$

Ereignisse $A_1, \ldots, A_n$ heißen vollständig unabhängig, *falls für alle $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ gilt*

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k}).$$

Bemerkung: Aus der paarweisen Unabhängigkeit von mehr als zwei Ereignissen folgt nicht immer die vollständige Unabhängigkeit. □

8.5 Zufallsvariablen und Verteilungsfunktion

Es sei Ω die Ergebnismenge und $\mathcal{A}$ das Ereignissystem, auf dem die Wahrscheinlichkeit P erklärt ist. Oft ist man in der Statistik an einem dem Ergebnis $\omega \in \Omega$ zugeordneten Zahlenwert $X(\omega)$ interessiert.

Definition 8.5.1 *Eine* Zufallsvariable *ist eine Abbildung*

$$X : \Omega \to \mathbb{R}$$

mit der Eigenschaft, dass für jedes Intervall $I \subset \mathbb{R}$ die Urbildmenge

$$A = \{\omega \in \Omega \,:\, X(\omega) \in I\}$$

zum Ereignissystem $\mathcal{A}$ gehört. Die Wahrscheinlichkeit dieses Ereignisses "X nimmt Werte im Intervall I an" bezeichnet man abkürzend mit $P(X \in I)$ und schreibt entsprechend

$$P(a \leq X \leq b), \;\; P(X \leq x), \;\; P(X < x), \;\; P(|X - a| < b), \;\; P(X = b) \;\; \textit{usw.}$$

Beispiel: Zwei Würfel werden geworfen. Wir wählen $\Omega = \{(i, j) \,:\, i, j = 1, \ldots, 6\}$. Wir betrachten die Zufallsvariable $X : \Omega \to \mathbb{R}$,

$$X((i, j)) = i + j.$$

X beschreibt also die Summe der beiden gewürfelten Zahlen. Nun gilt zum Beipiel

$$P(X = 1) = 0, \;\; P(X = 2) = P(\{(1, 1\}) = \frac{1}{36}, \;\; P(X = 3) = P(\{(1, 2), (2, 1)\}) = \frac{2}{36}, \ldots$$

Definition 8.5.2 *Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Die Abbildung $F : \mathbb{R} \to \mathbb{R}$*

$$F(x) = P(X \leq x), \quad x \in \mathbb{R},$$

heißt Verteilungsfunktion *der Zufallsvariable X.*

Man kann zeigen, dass mit den Abkürzungen

$$F(x+) = \lim_{h \searrow 0} F(x+h), \quad F(x-) = \lim_{h \searrow 0} F(x-h),$$
$$F(-\infty) = \lim_{x \to -\infty} F(x), \quad F(\infty) = \lim_{x \to \infty} F(x)$$

gilt: Verteilungsfunktionen sind monoton wachsende Funktionen mit

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad F(x+) = F(x) \quad \forall\, x \in \mathbb{R}.$$

Zudem lassen sich alle interessierenden Wahrscheinlichkeiten im Zusammenhang mit der Zufallsvariable X berechnen:

$$\begin{aligned} P(X = a) &= F(a) - F(a-), \\ P(a < X \le b) &= F(b) - F(a), \\ P(a \le X < b) &= F(b-) - F(a-), \\ P(a \le X \le b) &= F(b) - F(a-), \\ P(X > a) &= 1 - F(a). \end{aligned}$$

Eine Zufallsvariable X heißt *diskret verteilt*, wenn sie nur endlich viele oder abzählbar unendlich viele Werte $x_1, x_2, \ldots$ annimmt. Ihre Verteilungsfunktion ist eine monoton wachsende Treppenfunktion, die jeweils an den Stellen x_i Sprünge der Höhe $P(X = x_i)$ hat.

Eine Zufallsvariable X heißt *stetig verteilt mit der Dichte* f, wenn ihre Verteilungsfunktion F durch

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \quad x \in \mathbb{R},$$

gegeben ist. Die Dichte ist hierbei eine nichtnegative Funktion, die Verteilungsfunktion F ist stetig und es gilt $F'(x) = f(x)$ für alle Stetigkeitsstellen x von f.

8.5.1 Beispiele für diskrete Verteilungen

Geometrische Verteilung

Es sei $0 < p < 1$. Eine Zufallsvariable X mit dem Wertebereich $\mathbb{N} = \{1, 2, \ldots\}$ heißt *geometrisch verteilt* mit dem Parameter p, falls

$$P(X = i) = (1-p)^{i-1}\, p, \quad i = 1, 2, \ldots$$

Anwendung: Wird ein Zufallsexperiment, bei dem ein bestimmtes Ereignis mit Wahrscheinlichkeit p eintritt, so lange unabhängig wiederholt, bis zum ersten Mal dieses Ereignis eintritt, dann kann die Anzahl der dazu benötigten Versuche durch eine geometrisch verteilte Zufallsvariable modelliert werden ("Warten auf den ersten Erfolg").

Binomialverteilung

Seine $n \in \mathbb{N}$ und $0 < p < 1$. Eine Zufallsvariable X mit dem Wertebereich $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ heißt *binomialverteilt* mit Parametern n und p, kurz $B(n, p)$-verteilt, falls

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \dots, n.$$

Anwendung: Wird ein Zufallsexperiment, bei dem ein bestimmtes Ereignis mit Wahrscheinlichkeit p eintritt, n-mal unabhängig wiederholt, und dabei gezählt, wie oft dieses Ereignis eintritt, so kann diese zufällige Anzahl als $B(n, p)$-verteilte Zufallsvariable X beschrieben werden ("Anzahl der Erfolge bei n Versuchen").

Poissonverteilung

Sei $\lambda > 0$. Eine Zufallsvariable X mit dem Wertebereich $\mathbb{N} \cup \{0\}$ heißt *Poisson-verteilt*, falls gilt

$$P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 0, 1, 2, \dots$$

Sie eignet sich zur Modellierung von Zählergebnissen folgenden Typs: In einer Telefonzentrale wird die Anzahl der innerhalb von 10 Minuten eingehenden Anrufe gezählt. λ gibt die "mittlere Anzahl" der eingehenden Anrufe an.

8.5.2 Beispiele für stetige Verteilungen

Rechteckverteilung

Es sei $a < b$. Eine stetig verteilte Zufallsvariable mit der Dichte

$$f(t) = \begin{cases} \frac{1}{b-a}, & a \le t \le b \\ 0 & \text{sonst} \end{cases}$$

heißt *rechteckverteilt* im Intervall $[a, b]$, kurz $R(a, b)$-verteilt. Für die Verteilungsfunktion ergibt sich

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \begin{cases} 0, & x \le a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1 & x \ge b. \end{cases}$$

Exponentialverteilung

Sei $\lambda > 0$. Eine stetig verteilte Zufallsvariable X mit der Dichte

$$f(t) = \begin{cases} 0, & t < 0, \\ \lambda e^{-\lambda t}, & t \geq 0, \end{cases}$$

heißt *exponentialverteilt* mit Parameter λ, kurz $Ex(\lambda)$-verteilt. Für die Verteilungsfunktion ergibt sich

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Normalverteilung

E seien $\mu \in \mathbb{R}$ und $\sigma > 0$. Eine stetig verteilte Zufallsvariable X mit der Dichte

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}, \quad t \in \mathbb{R},$$

heißt *normalverteilt* mit Parameter μ und σ^2, kurz: $N(\mu, \sigma^2)$-verteilt.

Im Fall $\mu = 0, \sigma^2 = 1$ spricht man von einer *Standard-Normalverteilung* und bezeichnet ihre Verteilungsfunktion mit

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt.$$

Φ ist nicht geschlossen angebbar und muss tabelliert oder numerisch ausgewertet werden. Offensichtlich gilt

$$\Phi(0) = \frac{1}{2}, \quad \Phi(-x) = 1 - \Phi(x), \quad x \geq 0.$$

Ist X eine $N(\mu, \sigma^2)$-verteilte Zufallsvariable, dann rechnet man leicht nach, dass die Verteilungsfunktion durch

$$F_{\mu,\sigma^2}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

gegeben ist. Tatsächlich ergibt sich durch die Substitution $s = \frac{t-\mu}{\sigma}$

$$F_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}\,dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{1}{2}s^2}\,ds = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

8.6 Erwartungswert und Varianz

Ist X eine diskret verteilte Zufallsvariable mit den Werten $x_1, x_2, \ldots$, so heißt

$$E(X) = \sum_i x_i P(X = x_i)$$

Erwartungswert von X, falls $\sum_i |x_i| P(X = x_i) < \infty$.

Ist X eine stetig verteilte Zufallsvariable mit Dichte f, so heißt

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

Erwartungswert von X, falls $\int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty$.

Beispiele:

1. Sei X Poisson-verteilt mit Parameter $\lambda > 0$.

$$E(X) = \sum_{i=0}^{\infty} i \frac{\lambda^i}{i!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

2. Sei X exponentialverteilt. Dann gilt

$$E(X) = \int_0^{\infty} x\, \lambda e^{-\lambda x}\, dx = -x e^{-\lambda x}|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}\, dx = 0 - \frac{1}{\lambda} e^{-\lambda x}|_0^{\infty} = \frac{1}{\lambda}.$$

Ist $h : \mathbb{R} \to \mathbb{R}$ eine stückweise stetige Funktion. Dann hat die Zufallsvariable $h(X)$ für eine diskret verteilt Zufallsvariable X den Erwartungswert (im Falle seiner Existenz)

$$E(h(X)) = \sum_i h(x_i) P(X = x_i).$$

Ist X stetig verteilt mit Dichte f, dann hat $h(X)$ den Erwartungswert (im Falle seiner Existenz)

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)\, f(x)\, dx.$$

Der Erwartungswert der quadratischen Abweichung der Zufallsvariablen X von ihrem Erwartungswert $E(X)$ heißt *Varianz* von X:

$$\mathrm{Var}(X) = E([X - E(X)]^2).$$

Die *Standardabweichung von* X ist definiert durch $\sqrt{\mathrm{Var}(X)}$.

8.6.1 Rechenregeln

Es gelten folgende Rechenregeln:

$$\begin{aligned} &E(aX+b) = aE(X)+b \\ &E(h_1(X)+h_2(X)) = E(h_1(X))+E(h_2(X)). \end{aligned}$$

Daraus erhält man

$$\text{Var}(X) = E(X^2) - (E(X))^2.$$

Begründung: Es gilt

$$\begin{aligned} \text{Var}(X) &= E([X-E(X)]^2) = E(X^2 - 2E(X)X + E(X)^2) \\ &= E(X^2) - E(2E(X)X) + E(E(X)^2) = E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned}$$

Außerdem gilt

$$\text{Var}(aX+b) = a^2\text{Var}(X),$$

da

$$\begin{aligned} \text{Var}(aX+b) &= E((aX+b)^2) - (E(aX+b))^2 = E(a^2X^2+2abX+b^2) - (aE(X)+b)^2 \\ &= a^2E(X^2) + 2abE(X) + b^2 - a^2E(X)^2 - 2abE(X) - b^2 \\ &= a^2(E(X^2) - E(X)^2) = a^2\text{Var}(X). \end{aligned}$$

Einige Beispiele:

Verteilung	$E(X)$	Var(X)
$N(\mu,\sigma^2)$	μ	σ^2
$Ex(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$B(n,p)$	np	$np(1-p)$

Die Tschebyschevsche Ungleichung stellt einen Zusammenhang zwischen Erwartungswert und Varianz her:

Tschebyschevsche Ungleichung:
Es gilt

$$P(|X-E(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2}, \quad c > 0.$$

Aus der Definition des Erwartungswerts ist klar, dass für Zufallsvariablen $X_1, X_2, \ldots, X_n$ gilt

$$E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n).$$

Die Frage, ob eine entsprechende Formel auch für die Varianz gilt, führt auf den Begriff der Unabhängigkeit von Zufallsvariablen $X_1, X_2, \ldots, X_n$.

Definition 8.6.1 *Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen mit Verteilungsfunktionen $F_1, \ldots, F_n$. Die* gemeinsame Verteilungsfunktion *von $X_1, X_2, \ldots, X_n$ ist definiert durch*

$$F(x_1, \ldots, x_n) = P(X_1 \le x_1, \cdots, X_n \le x_n), \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Die Zufallsvariablen heißen unabhängig, *wenn für alle $(x_1, \ldots, x_n) \in \mathbb{R}^n$ die Ereignisse*

$$\{X_1 \le x_1\}, \cdots, \{X_n \le x_n\}$$

vollständig unabhängig sind, also

$$P(X_1 \le x_1, \cdots, X_n \le x_n) = P(X_1 \le x_1) \cdot \ldots \cdot P(X_n \le x_n)$$

oder kurz

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n).$$

Satz 8.6.2 *Die Zufallsvariablen $X_1, X_2, \ldots, X_n$ seien unabhängig. Dann gilt*

$$\mathit{Var}(X_1 + \ldots + X_n) = \mathit{Var}(X_1) + \ldots + \mathit{Var}(X_n).$$

8.7 Gesetz der großen Zahlen, zentraler Grenzwertsatz

8.7.1 Das schwache Gesetz der großen Zahlen

Durch die Mittelung vieler unabhängiger identisch verteilter Zufallsvariablen erhält man eine Zufallsvariable, die mit großer Wahrscheinlichkeit Werte nahe beim Erwartungswert liefert.

Satz 8.7.1 (Das schwache Gesetz der großen Zahlen)
Ist $X_1, X_2, \ldots$ eine Folge unabhängiger identisch verteilter Zufallsvariablen (d.h. je endlich viele von ihnen sind unabhängig) mit $E(X_i) = \mu$, $\mathit{Var}(X_i) = \sigma^2$, dann gilt

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \varepsilon\right) = 0 \quad \forall \varepsilon > 0.$$

Beweis: Setze $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Dann gilt $E(Y_n) = \mu$ und wegen der Unabhängigkeit $\mathrm{Var}(Y_n) = \frac{1}{n^2}\mathrm{Var}(X_1 + \ldots + X_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$. Die Tschebyschevsche Ungleichung ergibt nun

$$P(|Y_n - \mu| \ge \varepsilon) \le \frac{\mathrm{Var}(Y_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{für } n \to \infty.$$

□

8.7.2 Zentraler Grenzwertsatz

Wir betrachten eine Zufallsvariable

$$Y = X_1 + \ldots + X_n$$

mit unabhängigen Summanden $X_1, \ldots, X_n$. Extrem große Werte von Y treten nur dann auf, wenn sehr viele X_i gleichzeitig große Werte annehmen. Wegen der Unabhängigkeit ist es sehr wahrscheinlich, dass große Werte eines Summanden durch kleine Werte eines anderen Summanden kompensiert werden. Es zeigt sich, dass die Verteilung von Y für großes n mehr und mehr einer Normalverteilung entspricht:

Satz 8.7.2 *(*Zentraler Grenzwertsatz*)*
Ist $X_1, X_2, \ldots$ eine Folge unabhängiger Zufallsvariablen mit

$$E(X_i) = \mu_i, \quad \mathit{Var}(X_i) = \sigma_i^2, \quad i = 1, 2, \ldots,$$

so gilt unter schwachen zusätzlichen Voraussetzungen, z.B. dass X_i identisch verteilt sind:

$$\lim_{n\to\infty} P\left(\frac{X_1 + \ldots + X_n - (\mu_1 + \ldots, \mu_n)}{\sqrt{\sigma_1^2 + \ldots + \sigma_n^2}} \le y\right) = \Phi(y) \quad \forall\, y \in \mathbb{R}.$$

Ein arithmetisches Mittel

$$\bar{X}_{(n)} = \frac{1}{n}(X_1 + \ldots + X_n)$$

ist für großes n also näherungsweise $N(\mu, \sigma^2)$-verteilt, wobei

$$\mu = \frac{1}{n}E(X_1 + \ldots + X_n) = \frac{1}{n}(\mu_1 + \ldots + \mu_n), \quad \sigma^2 = \frac{1}{n^2}\mathit{Var}(X_1 + \ldots + X_n) = \frac{1}{n^2}(\sigma_1^2 + \ldots + \sigma_n^2).$$

Bemerkung: Hat X Erwartungswert μ und Varianz σ^2, dann hat $\frac{X-\mu}{\sigma}$ den Erwartungswert 0 und Varianz 1. □

Wahrscheinlichkeitstheoretisches Modell für Messreihen

Als mathematisches Modell für das Entstehen von Messreihen werden im folgenden unabhängige, identisch verteilte Zufallsvariablen $X_1, \ldots, X_n$ verwendet.
Eine Messreihe $x_1, \ldots, x_n$ wird als Realisierung der Zufallsvariablen $X_1, \ldots, X_n$ angesehen, wir nehmen also an, dass ein Ergebnis $\omega \in \Omega$ existiert mit

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

Haben die X_i Erwartungswert μ und Varianz σ^2, dann sagt Satz 8.7.2 insbesondere aus, dass dann das arithmetische Mittel $\bar{X}_{(n)} = \frac{1}{n}(X_1 + \ldots + X_n)$ für grosse n näherungsweise $N(\mu, \sigma^2/n)$-verteilt ist.

Die Verteilungsfunktion der Zufallsvariablen $X_1, \ldots, X_n$ sei mit F bezeichnet. Was sagt die Messreihe über F aus?

Es ist intuitiv einleuchtend, dass die empirische Verteilungsfunktion

$$F_n(z; x_1, x_2, \ldots, x_n) = \frac{\text{Zahl der } x_i \text{ mit } x_i \le z}{n}$$

in engem Zusammenhang zur Wahrscheinlichkeit

$$F(z) = P(X_1 \le z)$$

stehen muss. Es gilt:

Satz 8.7.3 (Zentralsatz der Statistik)
Sei $X_1, X_2, \ldots$ eine Folge unabhängiger identisch verteilter Zufallsvariablen mit der Verteilungsfunktion F und bezeichne

$$D_n(X_1, \ldots, X_n) = \sup_{z \in \mathbb{R}} |F_n(z; X_1, \ldots, X_n) - F(z)|$$

die zufällige Maximalabweichung zwischen empirischer und "wahrer" Verteilungsfunktion. Dann gilt

$$P(\lim_{n\to\infty} D_n(X_1, \ldots, X_n) = 0) = 1,$$

$D_n(X_1, \ldots, X_n)$ konvergiert also mit Wahrscheinlichkeit 1 gegen 0.

8.8 Testverteilungen und Quantilapproximationen

In der Statistik, insbesondere in der Testtheorie, werden die folgenden Verteilungen benötigt, die von der Normalverteilung abgeleitet werden:

Seien $Z_1, \ldots, Z_n$ unabhängige, identisch $N(0, 1)$-verteilte Zufallsgrößen.

χ_r^2-Verteilung:

Es sei $1 \le r \le n$. Eine Zufallsvariable X heißt χ_r^2-verteilt, falls sie die Verteilungsfunktion besitzt

$$F(x) = P(Z_1^2 + \ldots + Z_r^2 \le x), \quad x \in \mathbb{R}.$$

t_r-Verteilung:

Es sei $1 \le r \le n-1$. Eine Zufallsvariable X heißt t_r-verteilt, falls sie die Verteilungsfunktion besitzt

$$F(x) = P\left(\frac{Z_{r+1}}{\sqrt{(Z_1^2 + \ldots + Z_r^2)/r}} \le x\right), \quad x \in \mathbb{R}.$$

$F_{r,s}$-Verteilung:

Es sei $1 \leq r, s \leq n-1$ mit $r+s \leq n$. Eine Zufallsvariable X heißt F-verteilt mit r und s Freiheitsgraden, falls sie die Verteilungsfunktion besitzt

$$F(x) = P\left(\frac{(Z_1^2 + \ldots + Z_r^2)/r}{(Z_{r+1}^2 + \ldots + Z_{r+s}^2)/s} \leq x\right), \quad x \in \mathbb{R}.$$

Die Dichten dieser Verteilungen können unter Verwendung der Gamma-Funktion

$$\Gamma(x) = \int_0^\infty e^{-t}\, t^{x-1}, \quad x > 0$$

und der Beta-Funktion

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt, \quad \alpha, \beta > 0$$

angegeben werden.

Bezeichnungen für Quantile: Allgemein ist das p-Quantil x_p für eine stetig verteilte Zufallsgröße mit Verteilungsfunktion F gegeben durch

$$F(x_p) = p.$$

Bezeichnungen:

u_p	p-Quantil der $N(0,1)$-Verteilung
$t_{r;p}$	p-Quantil der t_r-Verteilung
$\chi_{r;p}$	p-Quantil der χ_r^2-Verteilung
$F_{r,s;p}$	p-Quantil der $F_{r,s}$-Verteilung

Für gängige Werte von p existieren Tabellen für diese Quantile.

8.8.1 Wichtige Anwendungsbeispiele

Seien $X_1, \ldots, X_n$ unabhängige, identisch $N(\mu, \sigma^2)$-verteilte Zufallsvariable. Bilden wir ihr arithmetisches Mittel

$$\bar{X}_{(n)} := \frac{1}{n}\sum_{i=1}^n X_i$$

und die Stichprobenvarianz

$$S_{(n)}^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_{(n)})^2$$

dann gilt:

Satz 8.8.1 *Es seien $X_1, \ldots, X_n$ unabhängige, identisch $N(\mu, \sigma^2)$-verteilte Zufallsvariable. Dann gilt:*

- $\bar{X}_{(n)}$ *ist $N(\mu, \sigma^2/n)$-verteilt,*
- $\frac{n-1}{\sigma^2} S_{(n)}^2$ *ist χ_{n-1}^2-verteilt,*
- $\bar{X}_{(n)}$ *und* $S_{(n)}^2$ *sind unabhängig,*
- $\sqrt{n}\frac{\bar{X}_{(n)}-\mu}{\sqrt{S_{(n)}^2}}$ *ist t_{n-1}-verteilt.*

Kapitel 9

Schätzverfahren und Konfidenzintervalle

9.1 Grundlagen zu Schätzverfahren

Für eine Messreihe $x_1, \ldots, x_n$ wird im Folgenden angenommen, dass sie durch n gleiche Zufallsexperimente unabhängig voneinander ermittelt werden. Jeden Messwert sehen wir als unabhängige Realisierung einer Zufallsvariable X an. Als mathematisches Modell für das Entstehen von Messreihen werden im folgenden unabhängige, identisch wie X verteilte Zufallsvariablen $X_1, \ldots, X_n$ verwendet. Eine Messreihe $x_1, \ldots, x_n$ wird als Realisierung der Zufallsvariablen $X_1, \ldots, X_n$ angesehen, wir nehmen also an, dass ein Ergebnis $\omega \in \Omega$ existiert mit

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

Es wird nun angenommen, dass die Verteilungsfunktion F von X, die auch die Verteilungsfunktion der unabhängigen, identisch verteilten Zufallsvariablen X_i, $1 \leq i \leq n$, ist, einer durch einen Parameter $\theta \in \Theta \subset \mathbb{R}^k$ parametrisierten Familie

$$F_\theta, \quad \theta \in \Theta,$$

von Verteilungsfunktionen angehört. Dieser Parameter oder ein durch ihn bestimmter Zahlenwert $\tau(\theta)$ mit einer Abbildung $\tau : \Theta \to \mathbb{R}$ sei unbekannt und soll aufgrund der Messreihe näherungsweise geschätzt werden.

Beispiel: X und alle $X_1, \ldots, X_n$ seien normalverteilt. F_θ mit

$$\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times]0, \infty[$$

ist dann die Verteilungsfunktion einer $N(\mu, \sigma^2)$-Verteilung. Soll der Erwartungswert geschätzt werden, so ist $\tau(\theta) = \mu$. Will man die Varianz schätzen, dann ist $\tau(\theta) = \sigma^2$.

Definition 9.1.1 *Ein* Schätzverfahren *oder eine* Schätzfunktion *oder kurz ein* Schätzer *ist eine Abbildung*

$$T_n : \quad \mathbb{R}^n \to \mathbb{R}.$$

Sie ordnet einer Messreihe $x_1, \ldots, x_n$ einen Schätzwert $T_n(x_1, \ldots, x_n)$ *für den unbekannten Wert $\tau(\theta)$ zu.*

Die Zufallsvariable $T_n(X_1, \ldots, X_n)$ heißt Schätzvariable.

Erwartungswert und Varianz der Schätzvariablen $T_n(X_1, \ldots, X_n)$ sowie aller X_i hängen von der Verteilungsfunktion F_θ ab, die seiner Berechnung zugrundegelegt wird. Um dies zu verdeutlichen, schreiben wir

$$E_\theta(T_n(X_1, \ldots, X_n)), \quad E_\theta(X_1), \ldots$$

sowie

$$\mathrm{Var}_\theta(T_n(X_1, \ldots, X_n)), \quad \mathrm{Var}_\theta(X_1), \ldots$$

Außerdem schreiben wir für durch F_θ berechnete Wahrscheinlichkeiten

$$P_\theta(a \le T_n(X_1, \ldots, X_n) \le b), \quad P_\theta(a \le X_1 \le b), \ldots$$

Definition 9.1.2 *Ein Schätzer $T_n : \ \mathbb{R}^n \to \mathbb{R}$ heißt* erwartungstreu *für $\tau : \Theta \to \mathbb{R}$, falls gilt*

$$E_\theta(T_n(X_1, \ldots, X_n)) = \tau(\theta) \quad \textit{für alle } \theta \in \Theta.$$

Beispiele:

1. τ sei gegeben durch $\tau(\theta) = E_\theta(X) = \mu$. Das arithmetische Mittel $\bar{X}_{(n)} = \frac{1}{n}(X_1 + \ldots + X_n)$ ist ein erwartungstreuer Schätzer für $\tau(\theta)$. Tatsächlich gilt

$$E_\theta(\bar{X}_{(n)}) = E_\theta(\frac{1}{n}(X_1 + \ldots + X_n)) = \frac{1}{n}(E_\theta(X_1) + \ldots + E_\theta(X_n)) = \frac{1}{n} n\mu = \mu.$$

2. τ sei gegeben durch $\tau(\theta) = \mathrm{Var}_\theta(X)$. Die Stichprobenvarianz $S^2_{(n)} := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_{(n)})^2$ ist ein erwartungstreuer Schätzer für $\tau(\theta)$. Denn es gilt

$$\begin{aligned}
\sum_{i=1}^n (X_i - \bar{X}_{(n)})^2 &= \sum_{i=1}^n ((X_i - \mu) - (\bar{X}_{(n)} - \mu))^2 \\
&= \sum_{i=1}^n ((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_{(n)} - \mu) + (\bar{X}_{(n)} - \mu)^2) \\
&= \sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X}_{(n)} - \mu)^2 + n(\bar{X}_{(n)} - \mu)^2 \\
&= \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X}_{(n)} - \mu)^2.
\end{aligned}$$

Nun gilt $E_\theta((\bar{X}_{(n)} - \mu)^2) = \text{Var}_\theta(\bar{X}_{(n)}) = \frac{1}{n^2} n \text{Var}_\theta(X)$, also

$$E_\theta \left(\sum_{i=1}^n (X_i - \bar{X}_{(n)})^2 \right) = \sum_{i=1}^n E_\theta((X_i - \mu)^2) - n E_\theta((\bar{X}_{(n)} - \mu)^2)$$
$$= n \text{Var}_\theta(X) - n \frac{1}{n} \text{Var}_\theta(X) = (n-1) \text{Var}_\theta(X).$$

Als Abschwächung der Erwartungstreue betrachtet man asymptotische Erwartungstreue bei wachsender Stichprobenlänge.

Definition 9.1.3 *Ein Folge von Schätzer $T_n : \mathbb{R}^n \to \mathbb{R}$, $n = 1, 2, \ldots$ heißt* asymptotisch erwartungstreu *für $\tau : \Theta \to \mathbb{R}$, falls gilt*

$$\lim_{n\to\infty} E_\theta(T_n(X_1, \ldots, X_n)) = \tau(\theta) \quad \textit{für alle } \theta \in \Theta.$$

Zur Beurteilung der Güte eines Schätzers dient der

Mittlere quadratische Fehler (mean squared error):

$$MSE_\theta(T) := E_\theta((T - \tau(\theta))^2).$$

Offensichtlich gilt

$$T \text{ erwartungstreu} \quad \Longrightarrow \quad MSE_\theta(T) = \text{Var}_\theta(T).$$

Sind T_1 und T_2 zwei Schätzer für τ, dann heißt T_1 *effizienter* als T_2, wenn gilt

$$MSE_\theta(T_1) \leq MSE_\theta(T_2) \quad \forall\, \theta \in \Theta.$$

Sind T_1, T_2 erwartungstreu, dann bedeutet dies

$$\text{Var}_\theta(T_1) \leq \text{Var}_\theta(T_2) \quad \forall\, \theta \in \Theta.$$

Definition 9.1.4 *Ein Folge von Schätzern $T_1, T_2, \ldots$ heißt* konsistent *für τ, wenn für alle $\varepsilon > 0$ und alle $\theta \in \Theta$ gilt*

$$\lim_{n\to\infty} P_\theta \left(|T_n(X_1, \ldots, X_n) - \tau(\theta)| > \varepsilon \right) = 0.$$

Sie heißt konsistent im quadratischen Mittel *für τ, wenn für alle $\theta \in \Theta$ gilt*

$$\lim_{n\to\infty} MSE_\theta(T_n) = 0.$$

Es gilt folgender

Satz 9.1.5 *Ist $T_1, T_2, \ldots$ eine Folge von Schätzern, die erwartungstreu für τ sind und gilt*

$$\lim_{n\to\infty} \mathit{Var}_\theta(T_n(X_1, \ldots, X_n)) = 0 \quad \textit{für alle } \theta \in \Theta,$$

dann ist die Folge von Schätzern konsistent für τ.

Beweis: Wegen $E_\theta(T_n(X_1, \ldots, X_n)) = \tau(\theta)$ gilt nach der Ungleichung von Tschebyschev

$$P_\theta(|T_n(X_1, \ldots, X_n) - \tau(\theta)| > \varepsilon) \le \frac{\mathrm{Var}_\theta(T_n(X_1, \ldots, X_n))}{\varepsilon^2} \to 0.$$

□

Allgemeiner haben wir mit ganz ähnlichem Beweis

Satz 9.1.6 *Ist $T_1, T_2, \ldots$ eine Folge von Schätzern, die konsistent im quadratischen Mittel für τ ist, dann ist die Folge von Schätzern konsistent für τ.*

Beispiel: Es sei X $N(\mu, \sigma^2)$-verteilt, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}\times]0, \infty[$ und $\tau(\theta) = \mu$. Der Schätzer

$$T_n(X_1, \ldots, X_n) = \bar{X}_{(n)} = \frac{1}{n}(X_1 + \ldots + X_n)$$

ist nach Satz 8.8.1 $N(\mu, \sigma^2/n)$-verteilt, also erwartungstreu mit Varianz

$$\mathrm{Var}_\theta(T_n(X_1, \ldots, X_n)) = \sigma^2/n \to 0 \quad \text{für } n \to \infty.$$

Daher ist die Schätzerfolge nach Satz 9.1.5 auch konsistent.

9.2 Maximum-Likelihood-Schätzer

Bei gegebener Verteilungsklasse F_θ, $\theta \in \Theta$, lassen sich Schätzer für den Parameter θ oft mit der Maximum-Likelihood-Methode gewinnen.

Sind die zugrundeliegenden Zufallsvariablen $X_1, \ldots, X_n$ stetig verteilt, so ist die Verteilungsfunktion F_θ durch eine Dichte

$$f_\theta(x), \quad x \in \mathbb{R},$$

bestimmt. Im Fall diskreter Zufallsvariablen X, bzw. $X_1, \ldots, X_n$ definieren wir

$$f_\theta(x) = P_\theta(X = x) \quad \text{für alle } x \text{ aus dem Wertevorrat } \mathbb{X} \text{ von } X.$$

Definition 9.2.1 *Für eine Messreihe $x_1, \ldots, x_n$ heißt die Funktion $L(\cdot; x_1, \ldots, x_n)$ mit*

$$L(\theta; x_1, \ldots, x_n) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \ldots \cdot f_\theta(x_n)$$

die zu $x_1, \ldots, x_n$ gehörige Likelihood-Funktion.

Ein Parameterwert

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$$

mit

$$L(\hat{\theta}; x_1, \ldots, x_n) \geq L(\theta; x_1, \ldots, x_n) \quad \text{für alle } \theta \in \Theta$$

heißt Maximum-Likelihood-Schätzwert *für θ. Existiert zu jeder möglichen Messreihe $x_1, \ldots, x_n$ ein Maximum-Likelihood-Schätzwert $\hat{\theta}(x_1, \ldots, x_n)$, dann heißt*

$$T_n : \mathbb{X}^n \to \mathbb{R}, \quad T_n(x_1, \ldots, x_n) = \hat{\theta}(x_1, \ldots, x_n)$$

Maximum-Likelihood-Schätzer.

Beispiel: Die Zufallsvariablen seien Poisson-verteilt mit Parameter $\theta > 0$, also

$$f_\theta(x) = \frac{\theta^x}{x!} e^{-\theta}, \quad x \in \mathbb{N} \cup \{0\}.$$

Dies ergibt

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{x_1! \cdots x_n!} \cdot \theta^{x_1 + \ldots + x_n} \cdot e^{-n\theta}, \quad x_i \in \mathbb{N} \cup \{0\}.$$

L wird genau dann maximal, wenn die Log-Likelihood-Funktion $\ln(L)$, also

$$\ln L(\theta; x_1, \ldots, x_n) = -n\theta - \ln(x_1! \cdots x_n!) + (x_1 + \ldots + x_n) \ln \theta,$$

maximal wird. Die erste Ableitung dieser Funktion nach θ ist

$$\frac{d \ln L}{d\theta} = -n + \frac{x_1 + \ldots + x_n}{\theta}$$

mit der eindeutigen Nullstelle

$$\hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \ldots + x_n).$$

Da die zweite Ableitung negativ ist, ist $\hat{\theta}$ der Maximum-Likelihood-Schätzer für θ und ist nichts anderes als das arithmetische Mittel.

9.3 Konfidenzintervalle

Die Situation sei wie beim Schätzen. Es wird eine Messreihe $x_1, \dots, x_n$ beobachtet und es sollen diesmal Ober- und Unterschranken für den Wert $\tau(\theta)$ aus der Messreihe ermittelt werden. Durch ein Paar

$$U : \mathbb{R}^n \to \mathbb{R}, \quad O : \mathbb{R}^n \to \mathbb{R}$$

von Schätzern mit

$$U(x_1, \dots, x_n) \leq O(x_1, \dots, x_n)$$

wird ein "zufälliges Intervall"

$$I(X_1, \dots, X_n) = [U(X_1, \dots, X_n), O(X_1, \dots, X_n)]$$

definiert.

Definition 9.3.1 *Sei* $0 < \alpha < 1$. *Das zufällige Intervall* $I(X_1, \dots, X_n)$ *heißt* Konfidenzintervall *für* $\tau(\theta)$ *zum* Konfidenzniveau $1 - \alpha$*, falls gilt*

$$P_\theta(U(X_1, \dots, X_n) \leq \tau(\theta) \leq O(X_1, \dots, X_n)) \geq 1 - \alpha \quad \textit{für alle } \theta \in \Theta.$$

Das zu einer bestimmten Messreihe $x_1, \dots, x_n$ *gehörige Intervall*

$$I(x_1, \dots, x_n) = [U(x_1, \dots, x_n), O(x_1, \dots, x_n)]$$

heißt konkretes Schätzintervall *für* $\tau(\theta)$.

Die Forderung stellt sicher, dass mit Wahrscheinlichkeit $1-\alpha$ ein konkretes Schätzintervall den Wert $\tau(\theta)$ enthält.

9.3.1 Konstruktion von Konfidenzintervallen

Wir nehmen an, dass $X_1, \dots, X_n$ unabhängig, identisch normalverteilt sind. Die Verteilungsfunktion F_θ ist dann durch den zweidimensionalen Parameter $\theta = (\mu, \sigma^2)$ bestimmt durch

$$F_\theta(x) = F_{(\mu,\sigma^2)}(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Mit den bereits eingeführten Bezeichnungen

$$\bar{X}_{(n)} := \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_{(n)}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_{(n)})^2$$

erhält man folgende Konfidenzintervalle zum Niveau $1 - \alpha$:

Konfidenzintervall für μ bei bekannter Varianz $\sigma^2 = \sigma_0^2$:

Hier ist $\Theta = \{(\mu, \sigma_0^2) \, : \, \mu \in \mathbb{R}\}$ und $\tau(\theta) = \mu$. Das Konfidenzintervall für μ lautet

$$I(X_1, \dots, X_n) = \left[\bar{X}_{(n)} - u_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \bar{X}_{(n)} + u_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right].$$

mit dem $(1 - \alpha/2)$-Quantil $u_{1-\alpha/2}$ der $N(0,1)$-Verteilung, also

$$\Phi(u_{1-\alpha/2}) = 1 - \alpha/2.$$

Begründung: $\bar{X}_{(n)}$ ist nach Satz 8.8.1 $N(\mu, \sigma_0^2/n)$-verteilt. Also gilt:

$$Y_n := \frac{\bar{X}_{(n)} - \mu}{\sqrt{\sigma_0^2/n}} \quad \text{ist } N(0,1)\text{-verteilt.}$$

Wegen $\Phi(-u_{1-\alpha/2}) = \alpha/2$ gilt

$$P_\theta(-u_{1-\alpha/2} \leq Y_n \leq u_{1-\alpha/2}) = 1 - \alpha.$$

Einsetzen und Umformen ergibt

$$P_\theta(-u_{1-\alpha/2} \leq \frac{\bar{X}_{(n)} - \mu}{\sigma_0/\sqrt{n}} \leq u_{1-\alpha/2}) = P_\theta \left(\bar{X}_{(n)} - u_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X}_{(n)} + u_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right) = 1 - \alpha.$$

Konfidenzintervall für μ bei unbekannter Varianz σ^2:

Hier ist $\Theta = \{(\mu, \sigma^2) \, : \, \mu \in \mathbb{R}, \sigma^2 > 0\}$ und $\tau(\theta) = \mu$. Das Konfidenzintervall für μ lautet

$$I(X_1, \dots, X_n) = \left[\bar{X}_{(n)} - t_{n-1;1-\alpha/2} \sqrt{\frac{S_{(n)}^2}{n}}, \bar{X}_{(n)} + t_{n-1;1-\alpha/2} \sqrt{\frac{S_{(n)}^2}{n}} \right]$$

mit dem $(1 - \alpha/2)$-Quantil $t_{n-1;1-\alpha/2}$ der t_{n-1}-Verteilung.

Begründung: Nach Satz 8.8.1 ist

$$Y_n := \frac{\bar{X}_{(n)} - \mu}{\sqrt{S_{(n)}^2/n}} \quad \text{ist } t_{n-1}\text{-verteilt.}$$

Eine Rechnung völlig analog wie eben liefert das Konfidenzintervall.

Konfidenzintervall für σ^2 bei bekanntem Erwartungswert $\mu = \mu_0$:

Hier ist $\Theta = \{(\mu_0, \sigma^2) \,:\, \sigma^2 > 0\}$ und $\tau(\theta) = \sigma^2$. Das Konfidenzintervall für σ^2 lautet

$$I(X_1, \ldots, X_n) = \left[\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n;1-\alpha/2}}, \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n;\alpha/2}} \right].$$

Begründung: Jedes $\frac{X_i - \mu_0}{\sigma}$ ist $N(0,1)$-verteilt. Wegen der Unabhängigkeit ist also nach 8.8 $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2 \quad \chi^2_n$-verteilt. Dies ergibt

$$P_\theta \left(\chi^2_{n;\alpha/2} \leq \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2 \leq \chi^2_{n;1-\alpha/2} \right) = 1 - \alpha$$

und Auflösen nach σ^2 liefert das Konfidenzintervall.

Konfidenzintervall für σ^2 bei unbekanntem Erwartungswert:

Hier ist $\Theta = \{(\mu, \sigma^2) \,:\, \mu \in \mathbb{R}, \sigma^2 > 0\}$ und $\tau(\theta) = \sigma^2$. Das Konfidenzintervall für σ^2 lautet

$$I(X_1, \ldots, X_n) = \left[\frac{(n-1) S^2_{(n)}}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1) S^2_{(n)}}{\chi^2_{n-1;\alpha/2}} \right].$$

Begründung: Nach Satz 8.8.1 ist $\frac{n-1}{\sigma^2} S^2_{(n)} \quad \chi^2_{n-1}$-verteilt. Dies ergibt

$$P_\theta \left(\chi^2_{n-1;\alpha/2} \leq \frac{n-1}{\sigma^2} S^2_{(n)} \leq \chi^2_{n-1;1-\alpha/2} \right) = 1 - \alpha$$

und Auflösen nach σ^2 liefert das Konfidenzintervall.

Kapitel 10

Tests bei Normalverteilungsannahmen

10.1 Grundlagen

Beim Testen geht es um die Frage, ob eine Messreihe $x_1, \ldots, x_n$, die als Realisierung von unabhängigen, identisch verteilten Zufallsvariablen $X_1, \ldots, X_n$ angesehen wird, zu einer bestimmten Annahme über die Verteilung der X_i passt oder ihr widerspricht. Die zu prüfende Annahme heißt *Nullhypothese* H_0 und das Verfahren, mit dem entschieden wird, ob ein Widerspruch vorliegt, d.h. ob die Nullhypothese H_0 verworfen werden soll, heißt *Test*.

Seien also $X_1, \ldots, X_n$ unabhängige, identisch verteilte Zufallsvariablen, so dass eine Messreihe $x_1, \ldots, x_n$ als Realisierung von $X_1, \ldots, X_n$ aufgefasst werden kann. Dann ist ein Test durch die Angabe seines *kritischen Bereichs* $K \subset \mathbb{R}^n$ vollständig beschrieben: Es werde eine Messreihe $x_1, \ldots, x_n$ beobachtet.

Test:

Falls $(x_1, \ldots, x_n) \in K$: Lehne H_0 ab.
Sonst: Lehne H_0 nicht ab.

Es gibt zwei wichtige Fehlermöglichkeiten:

Fehler 1. Art: H_0 wird abgelehnt, obwohl H_0 zutrifft.

Fehler 2. Art: H_0 wird nicht abgelehnt, obwohl H_0 nicht zutrifft.

Natürlich soll K so gewählt werden, dass die Wahrscheinlichkeit für einen Fehler 1. Art klein ist.

Hierzu wird ein *Testniveau* α vorgegeben und gefordert, dass gilt:

Unter der Nullhypothese gilt $P((X_1, \ldots, X_n) \in K) \leq \alpha$.

Im folgenden wird der kritische Bereich mit Hilfe einer zur Nullhypothese passenden Funktion

$$T : \mathbb{R}^n \to \mathbb{R},$$

der sogenannten *Testgröße*, und geeignete kritische Schranken c bzw. c_1 und c_2 beschrieben. Wir betrachten folgende vier Möglichkeiten:

Falls sowohl große als auch kleine Werte von T gegen H_0 sprechen:

- $K = \{(x_1, \dots, x_n) \in \mathbb{R}^n \ : \ |T(x_1, \dots, x_n)| > c\}$,
- $K = \{(x_1, \dots, x_n) \in \mathbb{R}^n \ : \ T(x_1, \dots, x_n) < c_1 \ \text{ oder } \ T(x_1, \dots, x_n) > c_2\}$.

Falls nur große bzw. kleine Werte von T gegen H_0 sprechen:

- $K = \{(x_1, \dots, x_n) \in \mathbb{R}^n \ : \ T(x_1, \dots, x_n) > c\}$,
- $K = \{(x_1, \dots, x_n) \in \mathbb{R}^n \ : \ T(x_1, \dots, x_n) < c\}$.

Tests lassen sich nach dem folgenden allgemeinen Prinzip konstruieren.

Konstruktionsprinzip für Test zum Niveau α:

1. Verteilungsannahme formulieren.
2. Nullhypothese H_0 formulieren.
3. Testgröße T wählen und ihre Verteilung unter H_0 bestimmen.
4. $I \subset \mathbb{R}$ so wählen, dass unter H_0 gilt $P(T(X_1, \dots, X_n) \in I) \leq \alpha$.

I wird durch die kritischen Schranken festgelegt und ist von der Form

$$I = \mathbb{R} \setminus [-c, c], \quad I = \mathbb{R} \setminus [c_1, c_2], \quad I =]c, \infty[, \ \text{oder} \ \ I =]-\infty, c[.$$

Als Werte für das Niveau α werden oft 0.1, 0.05 und 0.01 gewählt.

10.2 Wichtige Test bei Normalverteilungsannahme

Wir nehmen nun an, dass $X_1, \dots, X_n$ unabhängig, identisch $N(\mu, \sigma^2)$-verteilt sind. Die wichtigsten Tests verwenden Nullhypothesen über Erwartungswert und Varianz.

Wir geben die Konstruktion verschiedener Tests nach obigem Prinzip an.

Gauß-Test

1. $X_1, \ldots, X_n$ unabhängig, identisch $N(\mu, \sigma_0^2)$-verteilt, σ_0^2 bekannt.
2. a) $H_0: \quad \mu = \mu_0,$ b) $H_0: \quad \mu \leq \mu_0,$ c) $H_0: \quad \mu \geq \mu_0$
3. Die Testgröße
$$T(X_1, \ldots, X_n) = \frac{\sqrt{n}}{\sigma_0}(\bar{X}_{(n)} - \mu_0)$$
ist nach Satz 8.8.1 $N(0,1)$-verteilt, falls $\mu = \mu_0$ gilt.
4. Ablehnung, falls
a) $|T| > u_{1-\alpha/2},$ b) $T > u_{1-\alpha},$ c) $T < u_\alpha.$

t-Test

1. $X_1, \ldots, X_n$ unabhängig, identisch $N(\mu, \sigma^2)$-verteilt, σ^2 unbekannt.
2. a) $H_0: \quad \mu = \mu_0,$ b) $H_0: \quad \mu \leq \mu_0,$ c) $H_0: \quad \mu \geq \mu_0$
3. Die Testgröße
$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\bar{X}_{(n)} - \mu_0}{\sqrt{S_{(n)}^2}}$$
ist nach Satz 8.8.1 t_{n-1}-verteilt, falls $\mu = \mu_0$ gilt.
4. Ablehnung, falls
a) $|T| > t_{n-1;1-\alpha/2},$ b) $T > t_{n-1;1-\alpha},$ c) $T < t_{n-1;\alpha}.$

χ^2-Streuungstest

1. $X_1, \ldots, X_n$ unabhängig, identisch $N(\mu, \sigma^2)$-verteilt, μ unbekannt.
2. a) $H_0: \quad \sigma^2 = \sigma_0^2,$ b) $H_0: \quad \sigma^2 \leq \sigma_0^2,$ c) $H_0: \quad \sigma^2 \geq \sigma_0^2$
3. Die Testgröße
$$T(X_1, \ldots, X_n) = \frac{n-1}{\sigma_0^2} \cdot S_{(n)}^2$$
ist nach Satz 8.8.1 χ_{n-1}^2-verteilt, falls $\sigma^2 = \sigma_0^2$ gilt.
4. Ablehnung, falls
a) $T < \chi_{n-1;\alpha/2}^2$ oder $T > \chi_{n-1;1-\alpha/2}^2,$ b) $T > \chi_{n-1;1-\alpha}^2,$ c) $T < \chi_{n-1;\alpha}^2.$

Kapitel 11

Multivariate Verteilungen und Summen von Zufallsvariablen

Bisher hatten wir hauptsächlich unabhängige, identische verteilte Zufallsvariablen $X_1, \ldots, X_n$ betrachtet, um Messreihen als Realisierungen von Zufallsvariablen zu modellieren.

Manchmal möchte man aber auch das Zusammenwirken mehrerer Zufallsvariablen $X_1, \ldots, X_n$ untersuchen, die nicht unabhängig voneinander sind (z.B. Kursverläufe von Aktien). Wir betrachten nun die gemeinsame Verteilung des Zufallsvektors $X = (X_1, \ldots, X_n)^T$ und beschäftigen uns auch mit der Frage, wie die Summe von Zufallsvariablen verteilt ist.

11.1 Grundlegende Definitionen

Wir haben bereits die gemeinsame Verteilungsfunktion von Zufallsvariablen $X_1, \ldots, X_n$ kennengelernt. Wir bauen nun auf dieser Definition auf.

Definition 11.1.1 *Seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen mit Verteilungsfunktionen $F_1, \ldots, F_n$. Die* gemeinsame Verteilungsfunktion *von $X_1, X_2, \ldots, X_n$ ist definiert durch*

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \cdots, X_n \leq x_n), \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Man nennt F auch die Verteilung des Zufallsvektors $X = (X_1, \ldots, X_n)^T$.

Eine Funktion $f : \mathbb{R}^n \to [0, \infty)$ heißt gemeinsame Dichte *von $X_1, \ldots, X_n$, wenn gilt*

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(s_1, \ldots, s_n)\, ds_1 \ldots ds_n \quad \forall\, (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Der Vektor

$$\mu = (E(X_1), \ldots, E(X_n))^T$$

heißt (im Falle seiner Existenz) Erwartungswert(vektor) von $X = (X_1, \dots, X_n)^T$. Die Matrix

$$\Sigma = \begin{pmatrix} \mathit{Var}(X_1) & \mathit{Cov}(X_1, X_2) & \cdots & \mathit{Cov}(X_1, X_n) \\ \mathit{Cov}(X_2, X_1) & \mathit{Var}(X_2) & \cdots & \mathit{Cov}(X_2, X_n) \\ \vdots & & \ddots & \\ \mathit{Cov}(X_n, X_1) & & & \mathit{Var}(X_n) \end{pmatrix}$$

heißt (im Falle ihrer Existenz) Kovarianzmatrix *von $X = (X_1, \dots, X_n)^T$.*

Hierbei ist die Kovarianz *zweier Zufallsvariablen X_i, X_j definiert durch*

$$\mathit{Cov}(X_i, X_j) := E((X_i - E(X_i))(X_j - E(X_j))),$$

sofern $\mathit{Var}(X_i) < \infty$, $i = 1, \dots, n$.

Bemerkungen: $X = (X_1, \dots, X_n)^T$ habe die gemeinsame Dichte $f(x_1, \dots, x_n)$.

- Es gilt

  $$F_i(x_i) = P(X_i \le x_i) = \int_{-\infty}^{x_i} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(s_1, \dots, s_n)\, ds_1 \dots ds_{i-1} ds_{i+1} \dots ds_n}_{=f_i(s_i)}\, ds_i.$$

  und somit

  $$E(X_i) = \int_{-\infty}^{\infty} x_i f(x_i)\, dx_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i\, f(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

  sowie

  $$\mathrm{Cov}(X_i, X_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - E(X_i))(x_j - E(X_j))\, f(x_1, \dots, x_n)\, dx_1 \dots dx_n.$$

  Offensichtlich ist $\mathrm{Var}(X_i) = \mathrm{Cov}(X_i, X_i)$.

- Man zeigt leicht, dass

  $$P(a_1 \le X_1 \le b_1, \dots, a_n \le X_n \le b_n) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(s_1, \dots, s_n)\, ds_1 \dots ds_n.$$

Beispiel: Die multivariate Normalverteilung

Die wichtigste multivariate Verteilung ist die multivariate Normalverteilung $N_n(\mu, \Sigma)$: Sei $X = 1 + (X_1, \dots, X_n)^T$ ein Vektor von normalverteilten Zufallsvariablen mit Erwartungswert $\mu = (E(X_1), \dots, E(X_n))^T$ und Kovarianzmatrix Σ, dann besitzt die multivariate Normalverteilung die Dichte

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^n.$$

□

Im Fall von Unabhängigkeit ist die gemeinsame Dichte das Produkt der Einzeldichten:

Satz 11.1.2 *Seien $X_1, ..., X_n$ unabhängige Zufallsvariablen mit Dichten $f_1(x_1), \dots, f_n(x_n)$. Dann hat $X = (X_1, \dots, X_n)^T$ die gemeinsame Dichte*

$$f(x_1, \dots, x_n) = f_1(x_1) \cdot \dots \cdot f_n(x_n). \tag{11.1}$$

Hat umgekehrt $X = (X_1, \dots, X_n)^T$ eine gemeinsame Dichte mit der Produktdarstellung (11.1)*, dann sind $X_1, ..., X_n$ unabhängig.*

Definition 11.1.3 *Zufallsvariablen $X_1, ..., X_n$ mit Var$(X_i) < \infty$, $i = 1, \dots, n$, heißen* paarweise unkorreliert*, wenn gilt*

$$\mathit{Cov}(X_i, X_j) = 0 \quad \textit{für alle } i \neq j.$$

Unabhängigkeit hat paarweise Unkorreliertheit zur Folge:

Satz 11.1.4 *Seien $X_1, \dots, X_n$ unabhängige Zufallsvariablen mit Var$(X_i) < \infty$, $i = 1, \dots, n$. Dann gilt*

$$\mathit{Cov}(X_i, X_j) = 0 \quad \textit{für alle } i \neq j.$$

Für gemeinsam normalverteilte Zufallsvariablen sind Unabhängigkeit und paarweise Unkorreliertheit sogar äquivalent.

Lemma 11.1.5 *$X = (X_1, \dots, X_n)^T$ sei $N_n(\mu, \Sigma)$-verteilt. Dann sind $X_1, \dots, X_n$ genau dann unabhängig, wenn sie paarweise unkorreliert sind.*

Beweis: Wegen Satz 11.1.4 ist nur zu zeigen, dass aus paarweiser Unkorreliertheit die Unabhängigkeit folgt. Sind $X_1, \dots, X_n$ paarweise unkorreliert, dann gilt $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ und daher lautet die gemeinsame Dichte

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} = \frac{1}{(2\pi)^{n/2}\sigma_1 \cdot \dots \cdot \sigma_n} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} = f_1(x_1) \cdot \dots \cdot f_n(x_n)$$

mit

$$f_i(x_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i-\mu_i}{\sigma_i})^2}.$$

Also sind $X_1, \dots, X_n$ unabhängig nach Satz 11.1.2. □

11.2 Verteilung der Summe von Zufallsvariablen

Seien X_1, X_2 unabhängige stetige Zufallsvariablen mit Dichten $f_1(x_1)$ und $f_2(x_2)$. Wie sieht die Dichte von $X_1 + X_2$ aus? Wir benötigen die folgende Definition.

Definition 11.2.1 *Falls für die Funktionen $f, g : \mathbb{R} \to \mathbb{R}$ das Integral*

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x-y)g(y)\,dy$$

für alle x existiert, dann heißt $f * g$ die Faltung *von f und g.*

Dann gilt

Satz 11.2.2 *Seien X_1, X_2 unabhängige stetige Zufallsvariablen mit Dichten $f_1(x_1)$ und $f_2(x_2)$. Dann hat $X_1 + X_2$ die Dichte $f * g$.*

Damit läßt sich zeigen:

Satz 11.2.3 *Seien $X_1, X_2, \ldots, X_n$ unabhängige Zufallsvariablen, die $N(\mu_i, \sigma_i^2)$-verteilt sind. Dann ist $X = X_1 + X_2 + \ldots + X_n$ $N(\mu, \sigma^2)$-verteilt mit*

$$\mu = \mu_1 + \mu_2 + \ldots + \mu_n, \quad \sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2.$$

Im Fall diskreter Zufallsvariablen gibt es analoge Aussagen.

Definition 11.2.4 *Für $f = (f_i)_{i \in \mathbb{Z}}$, $g = (g_i)_{i \in Z}$ ist die diskrete Faltung von f und g definiert durch*

$$(f * g)_i := \sum_{j \in \mathbb{Z}} f_{i-j} g_j.$$

Analog zu Satz 11.2.3 gilt nun

Satz 11.2.5 *Seien X_1, X_2 unabhängige diskrete, $\mathbb{Z}$-wertige und setze $f_{X_1} := (P(X_1 = i))_{i \in \mathbb{Z}}$, $f_{X_2} := (P(X_2 = i))_{i \in \mathbb{Z}}$. Dann ist $f_{X_1+X_2} := (P(X_1 + X_2 = i))_{i \in \mathbb{Z}}$ gegeben durch*

$$f_{X_1+X_2} = f_{X_1} * f_{X_2}.$$

Als Anwendung erhält man z.B.

Satz 11.2.6 *Seien X_1, X_2 unabhängige Poisson-verteilte Zufallsvariablen mit Parameter λ_1 bzw. λ_2. Dann ist $X_1 + X_2$ Poisson-verteilt mit Parameter $\lambda_1 + \lambda_2$*

Beispiel: Beim radioaktiven Zerfall einer Substanz werden ionisierende Teilchen frei. Mit einem Geiger-Müller-Zählrohr zählt man die innerhalb einer Minute eintreffenden Teilchen. Deren Anzahl ist Poisson-verteilt. Hat man zwei radioaktive Substanzen mit Poisson-Verteilungen zu Parametern λ_1 und λ_2, so genügt die Gesamtheit der pro Zeitintervall produzierten Teilchen einer Poisson-Verteilungen zum Parameter $\lambda_1 + \lambda_2$.

Literaturverzeichnis

[DB02] P. Deuflhard, F. Hohmann. *Numerische Mathematik I.* de Gruyter, Berlin, 2008.

[DB02] P. Deuflhard, F. Bornemann. *Numerische Mathematik II.* de Gruyter, Berlin, 2002.

[Pl00] R. Plato. *Numerische Mathematik kompakt.* Vieweg Verlag, Braunschweig, 2000.

[St94] J. Stoer. *Numerische Mathematik 1.* Springer Verlag, Berlin, 1994.

[TS90] W. Törnig, P. Spellucci. *Numerische Mathematik für Ingenieure und Physiker 2.* Springer Verlag, Berlin, 1990.

[We92] J. Werner. *Numerische Mathematik 2.* Vieweg Verlag, Braunschweig, 1992.