# Kapitel 7

# Lösungen zu den Aufgaben

## 7.1 Aufgaben zu Kapitel 1

1. Ein Öltank hat die Gestalt eines liegenden Zylinders vom Radius 1m. Wie hoch steht das Öl, wenn der Tank zu $\frac{1}{4}$ seines Fassungsvermögens gefüllt ist?

   **Lösung:** Wir nehmen zunächst an, der Zylinder (bzw. der Öltank) habe den Radius $r$ und die Länge $l$. Die Höhe des Öls im Tank sei $h$. In der folgenden Abbildung 7.1 wird der Winkel $\alpha$ erklärt. Mit diesem Winkel $\alpha$ berechnet sich das Volumen des Öls

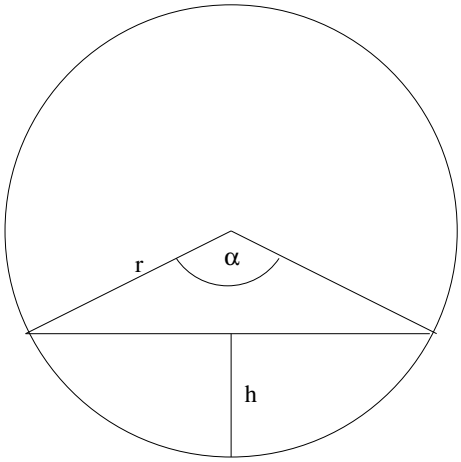


Abbildung 7.1: Ein Öltank

(Kreisausschnitt minus Dreieck) durch

$$V = r^2 l \frac{1}{2}(\alpha - \sin\alpha),$$

ferner ist die Höhe durch

$$h = r\left(1 - \cos\frac{\alpha}{2}\right)$$

gegeben. Ist der Tank mit einem $q \in (0,1)$ zu $100q\%$ gefüllt, so ist $V = r^2 l\pi q$ das Volumen des Öls. Damit erhalten wir für $\alpha$ die Bestimmungsgleichung

$$r^2 l \frac{1}{2}(\alpha - \sin\alpha) = r^2 l\pi q$$

bzw.

$$\alpha - \sin\alpha - 2\pi q = 0.$$

Für $q = \frac{1}{4}$ erhält man mit `alpha:=fsolve(alpha-sin(alpha)-Pi/2,alpha);` dass $\alpha^* = 2.309881460$, als zugehörige Höhe erhalten wir $h^* = 0.5960272467\,r$ m. In unserem Falle steht das Öl also etwa 59.6 cm hoch.

2. Man konstruiere eine möglichst billige Dose (mathematisch: Kreiszylinder) mit Radius $r$ und Höhe $h$, welche ein vorgegebenes Volumen $V > 0$ besitzt. Die Kosten des Bodens und des Deckels seien $c_1 > 0$ Geldeinheiten (etwa Euro) pro Quadrateinheit (etwa $\mathrm{cm}^2$), entsprechend die des Mantels $c_2 > 0$ Geldeinheiten.

   **Lösung:** Die Gesamtkosten sind gegeben durch $2\pi r^2 c_1 + 2\pi r h c_2$, diese gilt es unter der Nebenbedingung $\pi r^2 h = V$ (sowie $r > 0$, $h > 0$) zu minimieren. Eliminiert man hier $h$ durch die Nebenbedingung, so erhält man die Aufgabe

   $$f(r) := 2\pi r^2 c_1 + \frac{2Vc_2}{r}$$

   unter der Nebenbedingung $r > 0$ zu minimieren. Schon in der Schule weiß man, dass Extrema von $f$ notwendigerweise Nullstellen der Ableitung $f'$ sind. Als Nullstelle von

   $$f'(r) = 4\pi r c_1 - \frac{2Vc_2}{r^2}$$

   berechnet man den Radius der optimalen Dose als

   $$r^* = \left(\frac{Vc_2}{2\pi c_1}\right)^{1/3}.$$

   Wegen

   $$f''(r) = 4\pi c_1 + \frac{4Vc_2}{r^3} > 0$$

   ist $f$ strikt konvex auf $\mathbb{R}_+$ und daher $r^*$ die eindeutige Lösung. Die gesuchte Höhe der Dose erhält man aus $h^* = V/[\pi(r^*)^2]$.

3. Man finde in dem in Abbildung 7.2 angegebenen Graphen einen Euler-Zug.

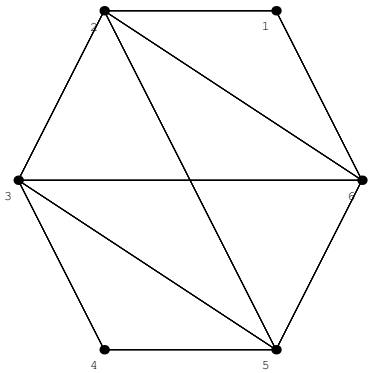


Abbildung 7.2: Das (erweiterte) Haus des Nikolaus

**Lösung:** Es ist

$$\{(4,5),(5,6),(6,3),(3,5),(5,2),(2,6),(6,1),(1,2),(2,3),(3,4)\}$$

ein Euler-Zug.

4. Gegeben sei ein Graph $G$ mit $4 \cdot 4 = 16$ Ecken. Diese denke man sich in einem Quadrat angeordnet. Zwei Ecken seien durch eine Kante verbunden, wenn man in einem Rösselsprung von der einen zur anderen Ecke gelangen kann.

   (a) Man rufe Maple auf und informiere sich durch `?networks` über das networks package.

   (b) Man generiere obigen Graphen und zeichne ihn mit Hilfe des `draw` Befehls.

   (c) Mit Hilfe von `degreeseq` bestimme man die Folge der Grade der Ecken und entscheide, ob es im Graphen einen Euler-Zug gibt.

   (d) Ist der Graph zusammenhängend, d. h. lassen sich je zwei Ecken durch einen Kantenzug verbinden?

   (e) Gibt es einen *Hamilton-Kreis* in dem Graphen, d. h. kann man mit dem Springer so über die 16 Felder springen, dass man jedes Feld (bis auf das erste) genau einmal trifft und am Schluss wieder im Ausgangsfeld ist?

   **Lösung:** Nach

```
with(networks):
new(G):
addvertex({1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16},G):
addedge([{1,7},{1,10},{2,8},{2,9},{2,11},{3,5},{3,10},{3,12},
{4,6},{4,11},{5,11},{5,14},{6,12},{6,13},{6,15},{7,9},{7,14},{7,16},
{8,10},{8,15},{9,15},{10,16},{11,13},{12,14}],G):
ross:=draw(Linear([1,2,3,4],[5,6,7,8],[9,10,11,12],[13,14,15,16]),G):
with(plots):
display(ross,thickness=4);
```

   erhalten wir die Abbildung 7.3. Nach `degreeseq(G);` erhalten wir

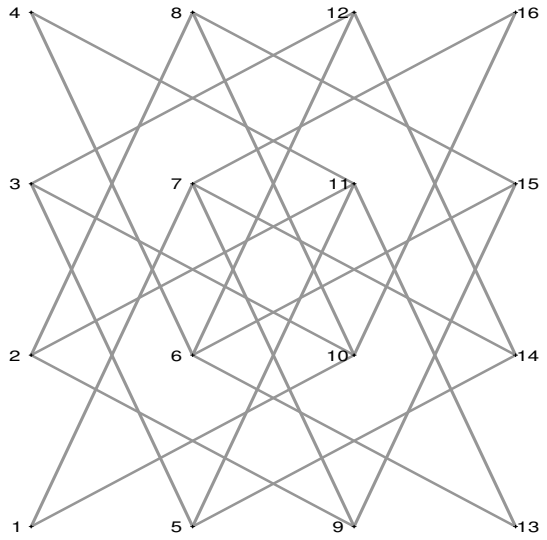


Abbildung 7.3: Rösselsprünge in einem Graphen der Ordnung 16

$$\{2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,4\},$$

es gibt weder einen geschlossenen noch einen offenen Euler-Zug. Nach `components(G)` erhält man, dass die einzige Zusammenhangskomponente von $G$ aus allen 16 Ecken besteht, der Graph ist also zusammenhängend.

Es gibt keinen Hamilton-Kreis in dem Graphen. Um dies einzusehen beachte man: Die Ecke "links oben" kann man nur von zwei Ecken erreichen. Die Ecke "rechts unten" kann nur von diesen zwei Ecken erreicht werden. Hierdurch hat man aber schon einen Kreis. Diese Ecken können nicht Teil eines Hamilton-Kreises sein. In Abbildung 7.3 hat man den Kreis $4, 11, 13, 6, 4$.

5. In[1] einem landwirtschaftlichen Betrieb sollen Roggen und Kartoffeln angebaut werden. Bezogen auf 1 Morgen Anbaufläche benötigt man hierzu bei Kartoffeln Anbaukosten von 5 Euro, einen Aufwand an Arbeitszeit von 2 Stunden und erhält dafür einen Reingewinn von 20 Euro. Die entsprechenden Daten für Roggen sind 10 Euro, 10 Stunden und 60 Euro. Die Anbauflächen für Roggen und Kartoffeln sind so zu wählen, dass der gesamte Reingewinn maximal wird. Hierbei stehen 1200 Morgen Land, 7000 Euro und 5200 Arbeitsstunden zur Verfügung.

   Man formuliere diese Aufgabe als lineare Optimierungsaufgabe, stelle die Menge der zulässigen Lösungen (hierzu kann man `inequal` aus dem `plots`-package von Maple benutzen) und die Zielfunktion in der Ebene graphisch dar und bestimme eine Lösung. Diese Lösung vergleiche man mit der durch Maple gefundenen Lösung.

   **Lösung:** Sei $x_1$ die Anbaufläche für Kartoffeln und $x_2$ die Anbaufläche von Roggen. Ein Anbauplan $x = (x_1, x_2)^T$ ist offenbar zulässig, wenn er den Nebenbedingungen

$$(*) \qquad \begin{cases} x_1 + x_2 \leq 1200 \\ 5x_1 + 10x_2 \leq 7000 \\ 2x_1 + 10x_2 \leq 5200 \\ x_1, x_2 \geq 0. \end{cases}$$

   Unter den Nebebedingungen $(*)$ ist die Gewinnfunktion

$$g(x_1, x_2) := 20x_1 + 60x_2$$

   zu maximieren. In Abbildung 7.4 stellen wir die zulässige Menge $M$ und die Zielfunktion dar. Als Lösung erhält man: 600 Morgen Land werden mit Kartoffeln, 400 Morgen mit Roggen angebaut, 200 Morgen liegen brach. Der maximale Reingewinn ist 36000 DM. Die obige Graphik haben wir erhalten durch:

```
with(plots): with(simplex):
bed:={x1+x2<=1200,5*x1+10*x2<=7000,2*x1+10*x2<=5200,x1>=0,x2>=0}:
pl1:=inequal(bed,x1=0..1300,x2=0..700,optionsfeasible=(color=grey),
optionsclosed=(color=blue,thickness=3),optionsexcluded=(color=yellow),
scaling=constrained):
pl2:=implicitplot(20*x1+60*x2=36000,x1=0..1300,x2=0..700,color=red):
pl3:=implicitplot(20*x1+60*x2=30000,x1=0..1300,x2=0..700,color=red):
pl4:=implicitplot(20*x1+60*x2=42000,x1=0..1300,x2=0..700,color=red):
display(pl1,pl2,pl3,pl4);
```

[1] Diese Aufgabe haben wir

L. Collatz, W. Wetterling (1971) *Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York entnommen.

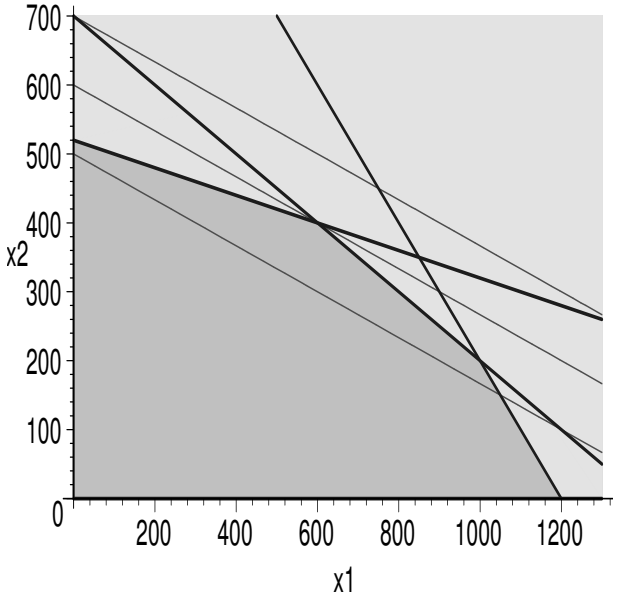


Abbildung 7.4: Menge der zulässigen Lösungen und Zielfunktion

Leider sieht man die angegebenen Farben nur im Maple-Fenster. Die Lösung selber (nach wie vor ist das `simplex`-Paket aktiviert) erhalten wir durch

```
maximize(20*x1+60*x2,bed);
```

6. Sie wollen Ihrer Tante (vielleicht eine reiche Erbtante?) zum Geburtstag eine Freude machen. Ihre Tante trinkt gerne einen süßen Wein und da Ihnen eine Beerenauslese zu teuer ist, kommen Sie auf die Idee, ihr einen Liter Wein zukommen zu lassen, den Sie selbst zusammengestellt haben.

   Hierzu können Sie einen Landwein für 1.00 Euro pro Liter, zur Anhebung der Süße Diäthylenglykol-haltiges Frostschutzmittel für 1.20 Euro pro Liter und für eine Verbesserung der Lagerungsfähigkeit eine Natriumacid-Lösung für 1.80 Euro pro Liter kaufen. Verständlicherweise wollen Sie eine möglichst billige Mischung herstellen, wobei aber folgende Nebenbedingungen zu beachten sind: Um eine hinreichende Süße zu garantieren, muss die Mischung mindestens 1/3 Frostschutzmittel enthalten. Andererseits muss (z. B. wegen gesetzlicher Bestimmungen) mindestens halb soviel Wein wie Frostschutzmittel enthalten sein. Der Natriumacid-Anteil muss mindestens halb so groß, darf aber andererseits höchstens so gross wie der Glykol-Anteil sein und darf die Hälfte des Weinanteils nicht unterschreiten.

   Man formuliere die Aufgabe, einen kostenminimalen Wein herzustellen, als eine lineare Optimierungsaufgabe und löse sie mit Hilfe von Maple.

   **Lösung:** Die gesuchte Mischung bestehe aus $x_1$ Liter Frostschutzmittel, $x_2$ Liter Natriumacid-Lösung und $x_3$ Liter Wein. Für eine "zulässige" Mischung erhalten wir die Nebenbedingungen

$$x_1 + x_2 + x_3 = 1, \quad x_1 \geq 0, \quad x_2 \geq 0, \quad x_3 \geq 0$$

   sowie

$$x_1 \geq \frac{1}{3}, \quad x_3 \geq \frac{1}{2}x_1, \quad \frac{1}{2}x_1 \leq x_2 \leq x_1, \quad \frac{1}{2}x_3 \leq x_2.$$

Die Kosten der Mischung ergeben sich zu $1.2x_1 + 1.8x_2 + x_3$ Euro, diese sind zu minimieren. Nach

```
with(simplex):
beding:={x_1+x_2+x_3=1,x_1>=1/3,x_3>=(1/2)*x_1,(1/2)*x_1<=x_2,
x_2<=x_1,(1/2)*x_3<=x_2}:
ziel:=1.2*x_1+1.8*x_2+x_3:
los:=minimize(ziel,beding,NONNEGATIVE);
opt:=subs(los,ziel);
```

erhält man die Lösung $x^* = (\frac{2}{5}, \frac{1}{5}, \frac{2}{5})^T$ mit den Kosten $\frac{31}{25} = 1.24$ Euro.

7. Man betrachte eine große Population von $N$ Individuen. Geburten, "natürliche Tode", Ein- und Auswanderungen mögen vernachlässigt werden. Es grassiere eine Krankheit, die sich durch Kontakt zwischen Individuen ausbreitet. Diese Krankheit sei so beschaffen, dass ein Individuum entweder durch sie stirbt oder nach einer Genesung immun gegen sie wurde. Die Population kann dann in drei Klassen eingeteilt werden.

   - In der Klasse $S$ sind die Anfälligen (**S**usceptibles) zusammengefasst, also diejenigen, die die Krankheit noch nicht bekommen haben und nicht gegen sie immun sind. Ihre Zahl zur Zeit $t$ sei $S(t)$.
   - In der Klasse $I$ sind die **I**nfizierten enthalten, also diejenigen, die die Krankheit haben und andere anstecken können. Zur Zeit $t$ sei ihre Zahl $I(t)$.
   - Zur Klasse $R$ gehört der **R**est, genauer also diejenigen, die tot, isoliert oder immun sind. $R(t)$ sei die Anzahl der Individuen der Klasse $R$ zur Zeit $t$.

   Die Krankheit genüge der folgenden Gesetzmäßigkeit.

   (a) Die Änderungsrate der anfälligen Population ist proportional zur Anzahl der Kontakte zwischen anfälliger und infizierter Population. Wir nehmen daher an, es sei

   $$S' = -\beta SI$$

   mit einer Konstanten (der sogenannten Infektionsrate) $\beta > 0$.

   (b) Individuen werden aus der Klasse $I$ der Infizierten mit einer Rate entfernt (sie sterben, werden isoliert oder immun), die proportional zu ihrer Anzahl ist. Daher ist

   $$I' = \beta SI - \gamma I, \qquad R' = \gamma I.$$

   Mit $S_0, I_0$ seien die positiven Populationen der Klassen $S$ und $I$ zur Anfangszeit $t = 0$ bezeichnet. Zu dieser Zeit sei noch niemand an der Krankheit gestorben bzw. ihretwegen isoliert oder immun. Man hat daher die Anfangswertaufgabe

   $$\text{(P)} \qquad \begin{array}{rclrcl} S' &=& -\beta SI, & S(0) &=& S_0, \\ I' &=& \beta SI - \gamma I, & I(0) &=& I_0, \\ R' &=& \gamma I, & R(0) &=& 0. \end{array}$$

   Dies ist das sogenannte Kermack-McKendrick-Modell für die Ausbreitung ansteckender Krankheiten. Wir gehen davon aus, dass obige Anfangswertaufgabe eine eindeutige Lösung $(S, I, R)$ auf $[0, \infty)$ besitzt. Man zeige (die ersten beiden Aussagen sind anschaulich völlig trivial, müssen aber trotzdem bewiesen werden):

(a) Es sind $I(\cdot)$ und $S(\cdot)$ auf $[0,\infty)$ positiv.

(b) Es ist $S(\cdot)$ auf $[0,\infty)$ monoton fallend. Daher existiert $S_\infty := \lim_{t\to\infty} S(t)$.

(c) Es ist $S(t) + I(t) - (\gamma/\beta)\ln S(t) = \text{const}$ für alle $t$.

(d) Ist $S_0 > \gamma/\beta$, so kommt es zu einer Epidemie in dem Sinne, dass es ein $t > 0$ mit $I(t) > I_0$ gibt. Weiter gibt es ein $t^* > 0$ derart, dass $I(\cdot)$ auf $[0,t^*]$ monoton wachsend und auf $[t^*,\infty)$ monoton fallend ist. Es ist $\lim_{t\to\infty} I(t) = 0$ und $S_\infty$ ist die eindeutige Lösung der transzendenten Gleichung

$$S_0 \exp\left(-\frac{(N-x)\beta}{\gamma}\right) - x = 0.$$

(e) Ist $S_0 < \gamma/\beta$, so ist $I(\cdot)$ auf $[0,\infty)$ monoton fallend und $\lim_{t\to\infty} I(t) = 0$. Es kommt also zu keiner Epidemie und die Krankheit verschwindet letztendlich.

Hinweis: Es kann zweckmäßig sein, zunächst die folgende Aussage zu beweisen:

- Sei $h\colon [0,\infty) \longrightarrow \mathbb{R}$ stetig. Dann besitzt die Anfangswertaufgabe $x' = h(t)x$, $x(0) = x_0$ die eindeutige Lösung

$$x(t) = x_0 \exp\left(\int_0^t h(\tau)\,d\tau\right).$$

**Lösung:** Wir beweisen zunächst die im Hinweis gemachte Aussage. Dass

$$x(t) := x_0 \exp(\int_0^t h(\tau)\,d\tau)$$

eine Lösung von $x' = h(t)x$, $x(0) = x_0$, ist, erkennt man einfach durch Einsetzen. Daher nehmen wir jetzt an, die auf $[0,\infty)$ stetig differenzierbare Funktion $x$ sei eine Lösung der angegebenen Anfangswertaufgabe. Dann ist

$$\frac{d}{dt}\left[\exp\left(-\int_0^t h(\tau)\,d\tau\right)x(t)\right] = \exp\left(-\int_0^t h(\tau)\,d\tau\right)[x'(t) - h(t)x(t)] = 0.$$

Daher ist

$$\exp\left(-\int_0^t h(\tau)\,d\tau\right)x(t) = x_0 \qquad \text{bzw.} \qquad x(t) = x_0 \exp\left(\int_0^t h(\tau)\,d\tau\right).$$

Nun kommen wir zu den eigentlichen Aussagen. $I(\cdot)$ ist Lösung von

$$I' = (\beta S(t) - \gamma)I, \qquad I(0) = I_0.$$

Aus der Aussage im Hinweis erhalten wir, dass

$$I(t) = I_0 \exp\left(\int_0^t (\beta S(\tau) - \gamma)\,d\tau\right) > 0.$$

Entsprechend zeigt man mit Hilfe der ersten Differentialgleichung, dass auch $S$ auf $[0,\infty)$ positiv ist. Wegen $S'(t) = -\beta S(t)I(t) < 0$, ist auch der zweite Teil bewiesen. Weiter ist

$$\begin{aligned}\frac{d}{dt}[S(t) + I(t) - (\gamma/\beta)\ln S(t)] &= S'(t) + I'(t) - (\gamma/\beta)\frac{S'(t)}{S(t)} \\ &= -\beta S(t)I(t) + \beta S(t)I(t) - \gamma I(t) + (\gamma/\beta)I(t) \\ &= 0,\end{aligned}$$

woraus auch die dritte Behauptung folgt.

In der vierten Aussage wird $S_0 > \gamma/\beta$ vorausgesetzt. Daher ist $I'(0) = I_0[\beta S_0 - \gamma] > 0$ und folglich $I(t) > I_0$ für alle hinreichend kleinen $t > 0$. Es ist $0 < S_\infty$, denn die Annahme $S_\infty = 0$ würde wegen der aus der dritten Behauptung folgenden Beziehung

$$(*) \qquad I(t) = \underbrace{I_0 + S_0}_{=N} - S(t) + (\gamma/\beta) \ln \frac{S(t)}{S_0}$$

einen Widerspruch ergeben. Wäre $S(t) > \gamma/\beta$ für alle $t \geq 0$, so wäre $I(t) \geq I_0$ für alle $t \geq 0$. Dann wäre aber $R'(t) \geq \gamma I_0$ und folglich $R(t) \geq \gamma I_0 t$ für alle $t \geq 0$, was wegen $R(t) \leq N$ natürlich nicht sein kann. Folglich existiert genau ein $t^* \in (0, \infty)$ mit $S(t^*) = \gamma/\beta$, ferner ist $S_\infty < \gamma/\beta$. Wegen

$$I'(t) = I(t)[\beta S(t) - \gamma]$$

ist $I(\cdot)$ auf $[0, t^*]$ monoton wachsend und auf $[t^*, \infty)$ monoton fallend. Weiter existiert ein $t^{**} > t^*$ mit $S(t) \leq \frac{1}{2}(S_\infty + \gamma/\beta)$ für alle $t \geq t^{**}$. Für diese $t$ ist

$$\begin{aligned}
0 &< I(t) \\
&= I_0 \exp\left[-\beta \int_0^t [\rho - S(\tau)]\, d\tau\right] \\
&= I_0 \exp\left[-\beta \int_0^{t^{**}} [\rho - S(\tau)]\, d\tau\right] \exp\left[-\beta \int_{t^{**}}^t [\rho - S(\tau)]\, d\tau\right. \\
&\leq I_0 \exp\left[-\beta \int_0^{t^{**}} [\rho - S(\tau)]\, d\tau\right] \exp\left[-\frac{\beta}{2}(\rho - S_\infty)(t - t^{**})\right]
\end{aligned}$$

und daher $\lim_{t\to\infty} I(t) = 0$. Aus $(*)$ erhält man

$$0 = N - S_\infty + \rho \ln \frac{S_\infty}{S_0}.$$

Hieraus erhält man sehr schnell einen Beweis der letzten Behauptung.

Nun zur fünften Aussage, in der wir annehmen, es sei $S_0 < \gamma/\beta$. Da $S(\cdot)$ monoton fallend ist, ist $S(t) < \gamma/\beta$ für alle $t$. Folglich ist

$$I'(t) = \underbrace{(\beta S(t) - \gamma)}_{<0} \underbrace{I(t)}_{>0} < 0,$$

also ist $I(\cdot)$ auf $[0, \infty)$ monoton fallend. Ferner ist

$$\begin{aligned}
0 &< I(t) \\
&= I_0 \exp\left(\int_0^t (\beta S(\tau) - \gamma)\, d\tau\right) \\
&\leq I_0 \exp\left(\int_0^t (\beta S_0 - \gamma)\, d\tau\right) \\
&= I_0 \exp(-\underbrace{(\gamma - \beta S_0)}_{>0} t) \\
&\to 0,
\end{aligned}$$

womit auch die fünfte Behauptung bewiesen ist.

8. Das Wachstumgesetz von B. Gompertz (1779-1865) soll das Wachsen von Tumoren gut beschreiben. Es basiert auf der Anfangswertaufgabe

$$V' = -rV \ln\left(\frac{V}{K}\right), \qquad V(0) = V_0.$$

Hierbei sind $r$ und $K > 0$ gegebene Konstanten, $V(t)$ die Größe des Tumors zur Zeit $t$ und $V_0 > 0$ der Anfangszustand. Mit Hilfe von Maple finde man einen Lösungskandidaten für diese Anfangswertaufgabe. Anschließend verifiziere man, dass es sich bei dem Kandidaten wirklich um eine Lösung handelt. Für $V_0 := 1$, $r := 2$ und $K := 3$ plotte man die Lösung schließlich auf dem Intervall $[0, 5]$.

**Lösung:** Nach

```
sol:=dsolve({D(V)(t)=-r*V(t)*ln(V(t)/K),V(0)=V_0},V(t));
simplify(sol);
```

erhalten wir

$$V(t) = \left(\frac{V_0}{K}\right)^{e^{-tr}} K.$$

Wir wollen uns davon überzeugen, dass dies wirklich eine Lösung ist. Es ist

$$V(t) = K \exp(\ln(V_0/K)e^{-rt})$$

und daher

$$\begin{aligned} V'(t) &= K \exp(\ln(V_0/K)e^{-rt})(-r \ln(V_0/K)e^{-rt}) \\ &= -rV(t) \ln(V(t)/K), \end{aligned}$$

daher genügt $V$ der angegebenen Differentialgleichung. Offensichtlich ist $V(0) = V_0$, daher $V$ Lösung der angegebenen Anfangswertaufgabe. In Abbildung 7.5 haben wir $V(t) = 3\exp(\log(1/3)\exp(-2t))$ auf $[0, 5]$ geplottet. Hierzu haben wir diesmal MAT-

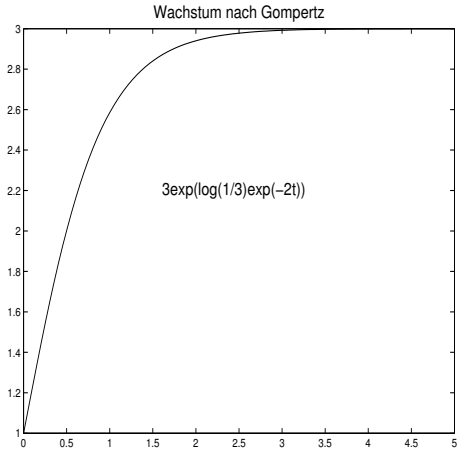


Abbildung 7.5: Wachstum nach Gompertz

LAB benutzt:

```
t=linspace(0,5);V=3*exp(log(1/3)*exp(-2*t));
plot(t,V);title('Wachstum nach Gompertz');
gtext('3exp(\log(1/3)exp(-2t))');
```

# 7.2 Aufgaben zu Kapitel 2

## 7.2.1 Aufgaben zu Abschnitt 2.1

1. Jeannot[2] spielt mit einer Waage, deren zwei Schalen sich im Gleichgewicht befinden, wenn man auf die eine ein Gewicht von 100 g und auf die andere zwei gleiche Schlüssel, zwei gleiche Münzen und drei gleiche Spielsoldaten, oder aber einen Apfel, einen Soldaten und eine Aprikose legt.

   Eine Münze, ein Schlüssel, ein Soldat und eine Pflaume wiegen zusammen 50 g.

   Die Aprikose, der Apfel und die Pflaume wiegen genausoviel wie eine Münze, ein Schlüssel, ein Soldat und die Feder des Weckers.

   Wieviel wiegt die Feder des Weckers?

   **Lösung:** Sei

$$\begin{aligned} x_1 &= \text{Gewicht des Schlüssels,} \\ x_2 &= \text{Gewicht der Münze,} \\ x_3 &= \text{Gewicht des Soldaten,} \\ x_4 &= \text{Gewicht des Apfels,} \\ x_5 &= \text{Gewicht der Aprikose,} \\ x_6 &= \text{Gewicht der Pflaume,} \\ x_7 &= \text{Gewicht der Feder des Weckers.} \end{aligned}$$

   Als Gleichungen hat man

$$\begin{aligned} 100 &= 2x_1 + 2x_2 + 3x_3, \\ 100 &= x_4 + x_3 + x_5, \\ x_2 + x_1 + x_3 + x_6 &= 50, \\ x_5 + x_4 + x_6 &= x_2 + x_1 + x_3 + x_7 \end{aligned}$$

   bzw.

$$\begin{pmatrix} 2 & 2 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} 100 \\ 100 \\ 50 \\ 0 \end{pmatrix}.$$

   Dies ist ein unterbestimmtes lineares Gleichungssystem, dessen allgemeine Lösung gegeben ist als Summe einer speziellen Lösung des inhomogenen Systems plus der allgemeinen Lösung der homogenen Aufgabe. Die linksstehende Matrix definiere man als $A$, den rechtsstehenden Vektor als $b$. Die letzte Komponente jedes Elementes aus dem Kern von $A$ verschwindet. Dies erkennt man, indem man im homogenen Gleichungssystem die

[2] Diese Aufgabe haben wir wörtlich

J. C. Baillif (1985, S. 11) *Denkpirouetten. Spiele aus Logik und Mathematik*. Hugendubel, München entnommen.

erste Gleichung mit $-1$ multipliziert und anschließend alle Gleichungen addiert. Daher ist $x_7$ bzw. das Gewicht der Feder des Weckers eindeutig bestimmt. Mit

```
with(LinearAlgebra):
A:=Matrix([[2,2,3,0,0,0,0],[0,0,1,1,1,0,0],[1,1,1,0,0,1,0],
     [1,1,1,-1,-1,-1,1]]):
NullSpace(A);
```

stellt man fest, dass eine Basis des Kerns von $A$ aus den Spalten von

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

besteht. Nach `b:=<100,100,50,0>: LinearSolve(A,b);` erhält man die allgemeine Lösung und insbesondere die eindeutig bestimmte letzte Komponente. Das Gewicht der Feder des Weckers ist 50 g.

2. Vor dem Pokerspiel entspricht die Summe, über die Paul verfügt, plus zweimal das, was André und Bernard besitzen, dem Zweifachen dessen, was Claude hat, plus dem Dreifachen dessen, über das Jean verfügt, und nochmals 300 Francs.

   Wenn André 1500 Francs mehr hätte, hätte er genausoviel wie Jean und Bernard, zuzüglich dem Zweifachen dessen, worüber Paul verfügt.

   Wenn Claude 1100 Francs mehr hätte, würde er soviel besitzen wie alle anderen vier Spieler zusammen.

   Das Dreifache dessen, was André hat, das Vierfache von dem, über das Jean verfügt, und weitere 1200 Francs machen das Dreifache von dem aus, was Bernard und Paul besitzen.

   Wieviel haben Jean und Paul zusammen?

   **Lösung:** Sei

$$\begin{aligned} x_1 &= \text{Besitz von Paul,} \\ x_2 &= \text{Besitz von André,} \\ x_3 &= \text{Besitz von Bernard,} \\ x_4 &= \text{Besitz von Claude,} \\ x_5 &= \text{Besitz von Jean.} \end{aligned}$$

   Man hat dann die Gleichungen

$$\begin{aligned} x_1 + 2(x_2 + x_3) &= 2x_4 + 3x_5 + 300, \\ x_2 + 1500 &= x_5 + x_3 + 2x_1, \\ x_4 + 1100 &= x_1 + x_2 + x_3 + x_5, \\ 3x_2 + 4x_5 + 1200 &= 3(x_3 + x_1) \end{aligned}$$

bzw.

$$\begin{pmatrix} 1 & 2 & 2 & -2 & -3 \\ 2 & -1 & 1 & 0 & 1 \\ 1 & 1 & 1 & -1 & 1 \\ 3 & -3 & 3 & 0 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 300 \\ 1500 \\ 1100 \\ 1200 \end{pmatrix}.$$

Die linksstehende Matrix wird mit $A$ bezeichnet, der rechtsstehende Vektor mit $b$. Der Kern von $A$ wird durch $(0,1,1,2,0)^T$ aufgespannt (dies haben wir in Maple durch `NullSpace(A)` erhalten), woraus man schon erkennt, dass die erste und die letzte Komponente einer Lösung von $Ax = b$ eindeutig ist, erst recht deren Summe. Mit Hilfe des Befehls `LinearSolve` aus dem `LinearAlgebra` Package erkennt man, dass Jean und Paul zusammen 700 Francs haben.

3. Seien reelle Zahlen $\alpha, \beta$ gegeben. Man zeige, dass sich jedes $p \in \mathcal{P}_3$, also jedes kubische Polynom, eindeutig darstellen lässt in der Form

$$p(x) := a + b(x-\alpha) + c(x-\alpha)^2 + d(x-\alpha)^2(x-\beta),$$

dass also durch $\{v_0, v_1, v_2, v_3\}$ mit

$$v_0(x) := 1, \quad v_1(x) := x - \alpha, \quad v_2(x) := (x-\alpha)^2, \quad v_3(x) := (x-\alpha)^2(x-\beta),$$

eine Basis von $\mathcal{P}_3$ gegeben ist.

**Lösung:** Es ist zu zeigen, dass $\{v_0, v_1, v_2, v_3\}$ linear unabhängig sind. Angenommen, es ist

$$p(x) := a + b(x-\alpha) + c(x-\alpha)^2 + d(x-\alpha)^2(x-\beta) \equiv 0.$$

Wir nehmen zunächst an, es sei $\alpha \neq \beta$. Wegen $p(\alpha) = a$ und $p'(\alpha) = b$ ist $a = b = 0$. Wegen $p(\beta) = c(\beta-\alpha)^2$ und $\alpha \neq \beta$ ist $c = 0$. Da außerdem $p'(\beta) = d(\beta-\alpha)^2$ ist auch $d = 0$. Damit ist die lineare Unabhängigkeit von $\{v_0, v_1, v_2, v_3\} \subset \mathcal{P}_3$ für $\alpha \neq \beta$ bewiesen. Für $\alpha = \beta$ ist das natürlich auch richtig, wie man aus $p''(\alpha) = 2c$ und $p'''(\alpha) = 6d$ erkennt. Damit ist die Behauptung bewiesen.

4. Sei mit $\Delta_n$ die Zerlegung $x_0 < \cdots < x_n$ von $[x_0, x_n]$ in $n$ Teilintervalle $[x_j, x_{j+1}]$, $j = 0, \ldots, n-1$, bezeichnet. Gegeben seien $f_0, \ldots, f_n$ und $f'_0, \ldots, f'_n$. Man zeige, dass es genau ein $s \in C^1[x_0, x_n]$ mit $s|_{[x_j, x_{j+1}]} \in \mathcal{P}_3$, $j = 0, \ldots, n-1$, und $s(x_j) = f_j$, $s'(x_j) = f'_j$, $j = 0, \ldots, n$, gibt und bestimme es.

Hinweis: Man berücksichtige Aufgabe 3 und mache für die Restriktion von $s$ auf das Intervall $[x_j, x_{j+1}]$ den Ansatz

$$s|_{[x_j, x_{j+1}]}(x) = a_j + b_j(x-x_j) + c_j(x-x_j)^2 + d_j(x-x_j)^2(x-x_{j+1})$$

mit noch zu bestimmenden $a_j, b_j, c_j, d_j$, $j = 0, \ldots, n-1$.

**Lösung:** Die Bedingungen $s(x_j + 0) = f_j$ und $s'(x_j + 0) = f'_j$, $j = 0, \ldots, n-1$, ergeben sofort, dass

$$a_j = f_j, \quad b_j = f'_j, \qquad j = 0, \ldots, n-1.$$

Die restlichen Parameter $c_j, d_j$, $j = 0, \ldots, n-1$, sind aus den Bedingungen $s(x_{j+1} - 0) = f_{j+1}$, $s'(x_{j+1} - 0) = f'_{j+1}$, $j = 0, \ldots, n-1$, zu berechnen. Zur Abkürzung definiere man $h_j := x_{j+1} - x_j$. Dann ist

$$s(x_{j+1} - 0) = f_j + f'_j h_j + c_j h_j^2,$$

so dass die Bedingung $s(x_{j+1}-0)=f_{j+1}$ auf

$$c_j=\frac{f_{j+1}-f_j-f'_jh_j}{h_j^2}$$

führt. Weiter ist

$$s'(x_{j+1}-0)=f'_j+2\frac{f_{j+1}-f_j-f'_jh_j}{h_j}+d_jh_j^2,$$

so dass die Bedingung $s'(x_{j+1}-0)=f'_{j+1}$ auf

$$d_j=\frac{f'_{j+1}-f'_j-2(f_{j+1}-f_j-f'_jh_j)/h_j}{h_j^2}=\frac{f'_{j+1}+f'_j-2(f_{j+1}-f_j)/h_j}{h_j^2}$$

führt. Insgesamt sind die gesuchten Parameter $a_j,b_j,c_j,d_j$, $j=0,\dots,n-1$, also zu bestimmen aus

$$a_j:=f_j,\quad b_j:=f'_j,\quad c_j:=\frac{f_{j+1}-f_j-f'_jh_j}{h_j^2},\quad d_j:=\frac{f'_{j+1}+f'_j-2(f_{j+1}-f_j)/h_j}{h_j^2}.$$

5. In Aufgabe 4 konnte gezeigt werden, dass es zu einer gegebenen Zerlegung $\Delta_n$ des Intervalls $[x_0,x_n]$ in $n$ Teilintervalle $[x_j,x_{j+1}]$, $j=0,\dots,n-1$, und gegebenen reellen Zahlen $f_0,\dots,f_n$ sowie $f'_0,\dots,f'_n$ genau ein $s\in C^1[x_0,x_n]$ mit $s|_{[x_j,x_{j+1}]}\in\mathcal{P}_3$, $j=0,\dots,n-1$, und $s(x_j)=f_j$, $s'(x_j)=f'_j$, $j=0,\dots,n$, gibt.

   (a) Welchen Bedingungen müssen $f'_0,\dots,f'_n$ genügen, damit $s$ sogar zweimal stetig differenzierbar ist, also $s\in S_3(\Delta_n)$ ein kubischer Spline zur Zerlegung $\Delta_n$ ist?

   (b) Man zeige, dass $f'_0,\dots,f'_n$ eindeutig dadurch bestimmt sind, dass $s\in S_3(\Delta_n)$ der

   - Hermiteschen Randbedingung (hier sind $f'_0$ und $f'_n$ gegeben),
   - natürlichen Randbedingung (hier sind $f''_0$ und $f''_n$ gegeben, die Forderung ist $s''(x_0)=f''_0$ und $s''(x_n)=f''_n$),
   - not-a-knot Bedingung (die Zusatzforderung ist, dass der kubische Spline $s$ in $x_1$ und $x_{n-1}$ dreimal stetig differenzierbar ist)

   genügt.

   Hinweis: Für den letzten Teil stelle man jeweils zur Bestimmung von $f'_1,\dots,f'_{n-1}$ ein lineares Gleichungssystem mit einer symmetrischen $(n-1)\times(n-1)$-Koeffizientenmatrix auf. Die Nichtsingularität der Koeffizientenmatrix folgt dann aus dem folgenden Resultat (siehe z. B. J. Werner (1992, S. 170)[3]:

   - *Sei $A=(a_{ij})\in\mathbb{R}^{n\times n}$ eine symmetrische Matrix mit positiven Diagonalelementen. Ist $A$ strikt diagonal dominant, d. h. gilt*

$$\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|<a_{ii},\qquad i=1,\dots,n,$$

[3] J. Werner (1992) *Numerische Mathematik 1*. Vieweg, Braunschweig-Wiesbaden.

*so ist $A$ positiv definit. Genauer gilt: Ist $\lambda$ ein Eigenwert von $A$, so ist*

$$\min_{i=1,\dots,n}\left(a_{ii}-\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\right)\le\lambda\le\max_{i=1,\dots,n}\left(a_{ii}+\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\right).$$

Man sollte dieses Ergebnis wenigstens anwenden können. Wer (mit Anstrengung) versucht, es zu beweisen, ist gut, wer dabei erfolgreich ist, sehr gut.

**Lösung:** Durch die (einfache) Lösung von Aufgabe 4 wissen wir, dass

$$s|_{[x_j,x_{j+1}]}(x)=a_j+b_j(x-x_j)+c_j(x-x_j)^2+d_j(x-x_j)^2(x-x_{j+1})$$

mit

$$a_j:=f_j,\quad b_j:=f'_j,\quad c_j:=\frac{f_{j+1}-f_j-f'_jh_j}{h_j^2},\quad d_j:=\frac{f'_{j+1}+f'_j-2(f_{j+1}-f_j)/h_j}{h_j^2}.$$

Hierbei ist $h_j:=x_{j+1}-x_j$, $j=0,\dots,n-1$. Für $j=1,\dots,n-1$ ist daher

$$s''(x_j+0)=2c_j-2d_jh_j,\qquad s''(x_j-0)=2c_{j-1}+4d_{j-1}h_{j-1}.$$

Es ist $s\in C^2[x_0,x_n]$, wenn $s''(x_j+0)=s''(x_j-0)$ bzw. $c_j-d_jh_j=c_{j-1}+2d_{j-1}h_{j-1}$, $j=1,\dots,n-1$. Einsetzen der oben angegebenen Formeln ergibt als gleichwertige Bedingung

$$3\frac{f_{j+1}-f_j}{h_j^2}-\frac{f'_{j+1}+2f'_j}{h_j}=-3\frac{f_j-f_{j-1}}{h_{j-1}^2}+\frac{2f'_j+f'_{j-1}}{h_{j-1}},\qquad j=1,\dots,n-1,$$

bzw.

$$\frac{1}{h_{j-1}}f'_{j-1}+2\left(\frac{1}{h_{j-1}}+\frac{1}{h_j}\right)f'_j+\frac{1}{h_j}f'_{j+1}=3\left(\frac{f_j-f_{j-1}}{h_{j-1}^2}+\frac{f_{j+1}-f_j}{h_j^2}\right),$$
$$j=1,\dots,n-1.$$

Setzt man noch $k_j:=1/h_j$, $j:=0,\dots,n-1$, so lauten diese Gleichungen

$$k_{j-1}f'_{j-1}+2(k_{j-1}+k_j)f'_j+k_jf'_{j+1}=3[k_{j-1}^2(f_j-f_{j-1})+k_j^2(f_{j+1}-f_j)],$$
$$j=1,\dots,n-1.$$

Man hat also $n-1$ Gleichungen für die $n+1$ Unbekannten $f'_0,\dots,f'_n$ als gleichwertige Bedingung dafür, dass $s$ ein kubischer Spline ist, der den Interpolationsforderungen $s(x_j)=f_j$, $j=0,\dots,n$, genügt.

Bei der Hermiteschen Randbedingung sind $f'_0$ und $f'_n$ vorgegeben. Die verbleibenden $f'_1,\dots,f'_{n-1}$ sind aus einem linearen Gleichungssystem mit der $(n-1)\times(n-1)$-Koeffizientenmatrix

$$A_H:=\begin{pmatrix}2(k_0+k_1) & k_1 & & & & \\ k_1 & 2(k_1+k_2) & k_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & k_{n-3} & 2(k_{n-3}+k_{n-2}) & k_{n-2} & \\ & & & k_{n-2} & 2(k_{n-2}+k_{n-1}) & \end{pmatrix}$$

und der rechten Seite

$$b_H := \begin{pmatrix} 3[k_0^2(f_1-f_0)+k_1^2(f_2-f_1)]-k_0f_0' \\ 3[k_1^2(f_2-f_1)+k_2^2(f_3-f_2)] \\ \vdots \\ 3[k_{n-3}^2(f_{n-2}-f_{n-3})+k_{n-2}^2(f_{n-1}-f_{n-2})] \\ 3[k_{n-2}^2(f_{n-1}-f_{n-2})+k_{n-1}^2(f_n-f_{n-1})]-k_{n-1}f_n' \end{pmatrix}$$

zu bestimmen.

Bei der natürlichen Randbedingung ist die Zusatzforderung, dass

$$s''(x_0) = 2c_0 - 2d_0h_0 = f_0'', \qquad s''(x_n) = 2c_{n-1} + 4d_{n-1}h_{n-1} = f_n''$$

mit vorgegebenen $f_0''$ und $f_n''$. Einsetzen ergibt, dass

$$3k_0^2(f_1-f_0) - 2k_0f_0' - k_0f_1' = \frac{f_0''}{2}$$

und

$$-3k_{n-1}^2(f_n-f_{n-1}) + 2k_{n-1}f_n' + k_{n-1}f_{n-1}' = \frac{f_n''}{2}.$$

Nun berücksichtige man noch die erste und die letzte der obigen $n-1$ Gleichungen, also

$$k_0f_0' + 2(k_0+k_1)f_1' + k_1f_2' = 3[k_0^2(f_1-f_0)+k_1^2(f_2-f_1)]$$

und

$$k_{n-2}f_{n-2}' + 2(k_{n-2}+k_{n-1})f_{n-1}' + k_{n-1}f_n' = 3[k_{n-2}^2(f_{n-1}-f_{n-2})+k_{n-1}^2(f_n-f_{n-1})].$$

Daher ist

$$(\tfrac{3}{2}k_0+2k_1)f_1' + k_1f_2' = \tfrac{3}{2}k_0^2(f_1-f_0) + 3k_1^2(f_2-f_1) + \tfrac{1}{4}f_0''$$

und entsprechend

$$k_{n-2}f_{n-2}' + (2k_{n-2}+\tfrac{3}{2}k_{n-1})f_{n-1}' = 3k_{n-2}^2(f_{n-1}-f_{n-2}) + \tfrac{3}{2}k_{n-1}^2(f_n-f_{n-1}) - \tfrac{1}{4}f_n''.$$

Insgesamt erhält man ein lineares Gleichungssystem für $f_1', \ldots, f_{n-1}'$ mit der $(n-1)\times(n-1)$-Koeffizientenmatrix

$$A_N := \begin{pmatrix} (\frac{3}{2}k_0+2k_1) & k_1 & & & & \\ k_1 & 2(k_1+k_2) & k_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & k_{n-3} & 2(k_{n-3}+k_{n-2}) & k_{n-2} & \\ & & & k_{n-2} & (2k_{n-2}+\frac{3}{2}k_{n-1}) \end{pmatrix}$$

und der rechten Seite

$$b_N := \begin{pmatrix} \frac{3}{2}k_0^2(f_1-f_0) + 3k_1^2(f_2-f_1) + \frac{1}{4}f_0'' \\ 3[k_0^2(f_1-f_0)+k_1^2(f_2-f_1)] \\ \vdots \\ 3[k_{n-2}^2(f_{n-1}-f_{n-2})+k_{n-1}^2(f_n-f_{n-1})] \\ 3k_{n-2}^2(f_{n-1}-f_{n-2}) + \frac{3}{2}k_{n-1}^2(f_n-f_{n-1}) - \frac{1}{4}f_n'' \end{pmatrix}.$$

Anschließend sind $f_0'$ und $f_n'$ aus

$$f_0' := -\tfrac{1}{2}f_1' + \tfrac{3}{2}k_0(f_1 - f_0) - \tfrac{1}{2}h_0 f_0''$$

und

$$f_n' := -\tfrac{1}{2}f_{n-1}' + \tfrac{3}{2}k_{n-1}(f_n - f_{n-1}) + \tfrac{1}{2}h_{n-1} f_n''$$

zu berechnen.

Die not-a-knot Bedingung lautet

$$s'''(x_1 - 0) = s'''(x_1 + 0), \qquad s'''(x_{n-1} - 0) = s'''(x_{n-1} + 0)$$

bzw. $d_0 = d_1$ und $d_{n-2} = d_{n-1}$. Einsetzen ergibt, dass

$$(*) \qquad f_0' = -f_1' + 2k_0(f_1 - f_0) + \left(\frac{k_1}{k_0}\right)^2 \left(f_1' + f_2' - 2k_1(f_2 - f_1)\right)$$

und

$$(**) \quad f_n' = -f_{n-1}' + 2k_{n-1}(f_n - f_{n-1}) + \left(\frac{k_{n-2}}{k_{n-1}}\right)^2 \left(f_{n-2}' + f_{n-1}' - 2k_{n-2}(f_{n-1} - f_{n-2})\right).$$

Nun berücksichtige man wieder die obigen $n-1$ Gleichungen für $f_0', \ldots, f_n'$. Die erste Gleichung lautet

$$k_0 f_0' + 2(k_0 + k_1) f_1' + k_1 f_2' = 3[k_0^2(f_1 - f_0) + k_1^2(f_2 - f_1)],$$

Einsetzen von $(*)$ liefert für $f_1', f_2'$ die Gleichung

$$\left(k_0 + 2k_1 + \frac{k_1^2}{k_0}\right) f_1' + \left(k_1 + \frac{k_1^2}{k_0}\right) f_2' = k_0^2(f_1 - f_0) + k_1^2\left(3 + 2\frac{k_1}{k_0}\right)(f_2 - f_1).$$

Um insgesamt auf ein lineares Gleichungssystem mit einer *symmetrischen* Koeffizientenmatrix zu kommen, multiplizieren wir diese Gleichung noch mit $1/(1+k_1/k_0)$. Dann lautet die erste Gleichung

$$\frac{k_0 + 2k_1 + k_1^2/k_0}{1 + k_1/k_0} f_1' + k_1 f_2' = \frac{k_0^2(f_1 - f_0) + k_1^2(3 + 2k_1/k_0)(f_2 - f_1)}{1 + k_1/k_0}.$$

Entsprechend lautet die $(n-1)$-te Gleichung

$$k_{n-2} f_{n-2}' + 2(k_{n-2} + k_{n-1}) f_{n-1}' + k_{n-1} f_n' = 3[k_{n-2}^2(f_{n-1} - f_{n-2}) + k_{n-1}^2(f_n - f_{n-1})],$$

Einsetzen von $(**)$ liefert nach Multiplikation mit $1/(1+k_{n-2}/k_{n-1})$ für $f_{n-2}'$ und $f_{n-1}'$ die Gleichung

$$\begin{aligned} & k_{n-2}' f_{n-2}' + \frac{2k_{n-2} + k_{n-1} + k_{n-2}^2/k_{n-1}}{1 + k_{n-2}/k_{n-1}} f_{n-1}' \\ = \; & \frac{k_{n-2}^2(3 + 2k_{n-2}/k_{n-1})(f_{n-1} - f_{n-2}) + k_{n-1}^2(f_n - f_{n-1})}{1 + k_{n-2}/k_{n-1}}. \end{aligned}$$

Setzt man daher

$$a_{11} := \frac{k_0 + 2k_1 + k_1^2/k_0}{1 + k_1/k_0}, \qquad a_{n-1,n-1} := \frac{2k_{n-2} + k_{n-1} + k_{n-2}^2/k_{n-1}}{1 + k_{n-2}/k_{n-1}}$$

sowie

$$\begin{aligned} b_1 &:= \frac{k_0^2(f_1-f_0)+k_1^2(3+2k_1/k_0)(f_2-f_1)}{1+k_1/k_0}, \\ b_{n-1} &:= \frac{k_{n-2}^2(3+2k_{n-2}/k_{n-1})(f_{n-1}-f_{n-2})+k_{n-1}^2(f_n-f_{n-1})}{1+k_{n-2}/k_{n-1}}, \end{aligned}$$

so erhält man für $f_1', \ldots, f_{n-1}'$ ein lineares Gleichungssystem mit der $(n-1)\times(n-1)$-Koeffizientenmatrix

$$A_K := \begin{pmatrix} a_{11} & k_1 & & & & \\ k_1 & 2(k_1+k_2) & k_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & k_{n-3} & 2(k_{n-3}+k_{n-2}) & k_{n-2} & \\ & & & k_{n-2} & a_{n-1,n-1} \end{pmatrix}$$

und der rechten Seite

$$b_K := \begin{pmatrix} b_1 \\ 3[k_1^2(f_2-f_1)+k_2^2(f_3-f_2)] \\ \vdots \\ 3[k_{n-3}^2(f_{n-2}-f_{n-3})+k_{n-2}^2(f_{n-1}-f_{n-2})] \\ b_{n-1} \end{pmatrix}.$$

Anschließend sind $f_0'$ und $f_n'$ aus $(*)$ und $(**)$ zu bestimmen.

Offenbar sind die Matrizen $A_H, A_N$ und $A_K$ symmetrisch, sie haben positive Diagonalelemente und sind strikt diagonal dominant. Aus dem angegebenen Resultat folgt, dass die Matrizen positiv definit und daher insbesondere nichtsingulär sind. Uns bleibt daher nur, den Hilfssatz zu beweisen. Sei hierzu $\lambda$ ein (wegen der Symmetrie von $A$ natürlich reeller) Eigenwert von $A$ und $x = (x_j) \neq 0$ ein zugehöriger Eigenvektor, der durch $\max_{i=1,\ldots,n} |x_i| = 1$ normiert sei. Man wähle ein $i \in \{1,\ldots,n\}$ mit $|x_i| = 1$. Aus

$$\sum_{j=1}^{n} a_{ij}x_j = \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j + a_{ii}x_i = \lambda x_i$$

erhält man

$$|\lambda - a_{ii}|\,|x_i| = \left|\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j\right| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\,\underbrace{|x_j|}_{\leq 1} \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

Insbesondere ist

$$0 < a_{ii} - \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \leq \lambda \leq a_{ii} + \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|.$$

Jeder Eigenwert von $A$ ist also positiv und daher $A$ positiv definit. Ferner folgt der Zusatz über die Lage der Eigenwerte von $A$.

6. Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ strikt zeilenweise diagonal dominant, d. h. es ist

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1, \ldots, n.$$

Man zeige:

(a) $A$ ist nichtsingulär.

(b) Alle Hauptabschnittsdeterminanten von $A$ sind von Null verschieden, so dass nach Satz 1.1 das Gaußsche Eliminationsverfahren ohne Spaltenpivotsuche durchführbar ist und eine Darstellung $A = LU$ mit einer unteren Dreiecksmatrix $L$ mit Einsen in der Diagonalen und einer oberen Dreiecksmatrix $U$ liefert.

**Lösung:** Angenommen, es ist $Ax = 0$. Sei $\|x\|_\infty$ die Maximumnorm von $x$, also $\|x\|_\infty = \max_{j=1,\ldots,n} |x_j|$. Man wähle $i \in \{1, \ldots, n\}$ mit $|x_i| = \|x\|_\infty$. Wegen

$$0 = a_{ii}x_i + \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j$$

folgt aus der diagonalen Dominanz von $A$, dass

$$|a_{ii}|\,\|x\|_\infty = |a_{ii}|\,|x_i| = \left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j\right| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_j| \leq \left(\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\right)\|x\|_\infty.$$

Daher ist

$$\underbrace{\left(|a_{ii}| - \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\right)}_{>0}\|x\|_\infty \leq 0$$

und damit ist $x = 0$. Folglich ist eine strik zeilenweise diagonal dominante Matrix nichtsingulär.

Auch die Matrizen $(a_{ij})_{1\leq i,j\leq k}$ sind strikt zeilenweise diagonal dominant und daher nichtsingulär. Folglich sind alle Hauptabschnittsdeterminanten von Null verschieden.

7. Sei $A \in \mathbb{R}^{n\times n}$ strikt spaltenweise diagonal dominant, d. h. es sei

$$\sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}| < |a_{jj}|, \qquad j = 1, \ldots, n.$$

Man zeige:

(a) Die Matrix $A$ ist nichtsingulär.

(b) Die Matrix $A$ besitzt nicht nur eine $LU$-Zerlegung der Form $A = LU$, sondern mehr noch: Das Gaußsche Eliminationsverfahren mit Spaltenpivotsuche benutzt keine Vertauschungen.

**Lösung:** Die Matrix $A^T$ ist wegen (a) in Aufgabe 6 nichtsingulär, da sie offenbar strikt zeilenweise diagonal dominant ist. Daher ist auch $A$ nichtsingulär, ferner sind alle Hauptabschnittsdeterminanten von Null verschieden.

Da $|a_{11}|$ in der ersten Spalte von $A$ das betragsgrößte Element ist, wird im ersten Schritt des Gaußschen Eliminationsverfahrens mit Spaltenpivotsuche keine Vertauschung vorgenommen. Ist

$$A = \begin{pmatrix} a_{11} & * \\ * & A_{22} \end{pmatrix}$$

mit $A_{22} = (a_{ij})_{2\le i,j\le n} \in \mathbb{R}^{(n-1)\times(n-1)}$, so liefert der nächste Schritt des Gaußschen Eliminationsverfahrens die Matrix

$$A^{(2)} = \begin{pmatrix} a_{11} & * \\ 0 & A_{22}^{(2)} \end{pmatrix}$$

mit

$$a_{ij}^{(2)} := a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}}, \qquad 2 \le i,j \le n.$$

Wir zeigen, dass auch $A_{22}^{(2)}$ spaltenweise diagonal dominant ist und erhalten durch Induktion die Behauptung. Denn für $j = 2,\dots,n$ ist

$$\begin{aligned}
\sum_{\substack{i=2\\ i\neq j}}^{n} |a_{ij}^{(2)}| &= \sum_{\substack{i=2\\ i\neq j}}^{n} \left| a_{ij} - \frac{a_{i1}a_{1j}}{a_{11}} \right| \\
&\le \sum_{\substack{i=2\\ i\neq j}}^{n} |a_{ij}| + \frac{|a_{1j}|}{|a_{11}|} \sum_{\substack{i=2\\ i\neq j}}^{n} |a_{i1}| \\
&< |a_{jj}| - |a_{1j}| + \frac{|a_{1j}|}{|a_{11}|}[|a_{11}| - |a_{j1}|] \\
&= |a_{jj}| - \left| \frac{a_{j1}a_{1j}}{a_{11}} \right| \\
&\le \left| a_{jj} - \frac{a_{j1}a_{1j}}{a_{11}} \right| \\
&= |a_{jj}^{(2)}|,
\end{aligned}$$

womit alles gezeigt ist.

8. Die nichtsinguläre Matrix $A \in \mathbb{R}^{n\times n}$ besitze eine Darstellung $A = LU$ mit einer unteren Dreiecksmatrix $L \in \mathbb{R}^{n\times n}$ mit Einsen in der Diagonalen und einer oberen Dreiecksmatrix $U \in \mathbb{R}^{n\times n}$. Man zeige, dass $L$ und $U$ hierdurch eindeutig bestimmt sind.

   **Lösung:** Sei $A = LU = L_1U_1$, wobei $L_1, U_1$ die entsprechenden Eigenschaften wie $L, U$ haben. Da $A$ nach Voraussetzung nichtsingulär ist, sind $U, U_1$ nichtsingulär und es gilt $L^{-1}L_1 = UU^{-1}$. Links steht eine untere Dreiecksmatrix mit Einsen in der Diagonalen, rechts eine obere Dreicksmatrix. Folglich ist $L^{-1}L_1 = UU^{-1} = I$ und daher $L_1 = L$ sowie $U_1 = U$.

9. Wieviele Multiplikationen/Divisionen werden bei der Anwendung des Gaußschen Eliminationsverfahrens auf eine Matrix $A \in \mathbb{R}^{n\times n}$ gemacht?

**Lösung:** Die Anzahl $C$ der Multiplikationen/Divisionen ist offenbar

$$C = \sum_{k=1}^{n-1}[(n-k)+(n-k)^2] = \sum_{j=1}^{n-1}[j+j^2] = \frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{1}{3}n^2 - \frac{1}{3}n.$$

10. Gegeben sei die Matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{pmatrix}.$$

(a) Welche Matrix würden Sie erhalten, wenn sie die $6 \times 6$-Matrix $A$ um eine Zeile und eine Spalte zu einer $7 \times 7$-Matrix erweitern müssten?

(b) Mit Hilfe des Gaußschen Eliminationsverfahrens mit Spaltenpivotsuche berechne man eine $LU$-Zerlegung von $A$. Hierzu sollte man Funktionen aus Maple, MATLAB und ein selbst (z. B. in MATLAB) geschriebenes Programm benutzen.

**Lösung:** Man erhält

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 3 & 6 & 10 & 15 & 21 & 28 \\ 1 & 4 & 10 & 20 & 35 & 56 & 84 \\ 1 & 5 & 15 & 35 & 70 & 126 & 210 \\ 1 & 6 & 21 & 56 & 126 & 252 & 462 \\ 1 & 7 & 28 & 84 & 210 & 462 & 924 \end{pmatrix},$$

wenn man das Pascalsche Dreieck dreht. Die angegebene Matrix erhält man in MATLAB durch die Anweisung `A=pascal(6)` bzw. `pascal(7)`.

Eine Anwendung der Funktion `LUDecomposition` aus dem Package `LinearAlgebra` von Maple liefert:

```
> with(LinearAlgebra):
> A:=Matrix([[1,1,1,1,1,1],[1,2,3,4,5,6],[1,3,6,10,15,21],[1,4,10,20,35
> ,56],[1,5,15,35,70,126],[1,6,21,56,126,252]]):
> (Q,L,U):=LUDecomposition(A);
```

$$Q,\, L,\, U := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 1 & 3 & 3 & 1 & 0 & 0 \\ 1 & 4 & 6 & 4 & 1 & 0 \\ 1 & 5 & 10 & 10 & 5 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 1 & 3 & 6 & 10 \\ 0 & 0 & 0 & 1 & 4 & 10 \\ 0 & 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

```
> P:=Transpose(Q);
```

$$P := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

```
> Equal(P.A,L.U);
```

*true*

hier wurden also keine Vertauschungen durchgeführt.

Wir geben nun die Ergebnisse an, die wir mit MATLAB erhalten haben. Der Befehl `[L,U,P]=lu(A)` liefert (`format short`, d. h. vier Stellen hinter dem Komma)

$$L = \begin{pmatrix} 1.0000 & & & & & \\ 1.0000 & 1.0000 & & & & \\ 1.0000 & 0.6000 & 1.0000 & & & \\ 1.0000 & 0.2000 & 0.6667 & 1.0000 & & \\ 1.0000 & 0.8000 & 0.6667 & -0.5000 & 1.0000 & \\ 1.0000 & 0.4000 & 1.0000 & 0.7500 & -0.5000 & 1.0000 \end{pmatrix}$$

sowie

$$U = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ & 5.0000 & 20.0000 & 55.0000 & 125.0000 & 251.0000 \\ & & -3.0000 & -14.0000 & -41.0000 & -95.6000 \\ & & & 1.3333 & 6.3333 & 18.5333 \\ & & & & -0.5000 & -2.8000 \\ & & & & & -0.1000 \end{pmatrix}$$

und

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

Eine Anwendung der oben angegebenen MATLAB-Funktion `GEpiv` liefert

$$L = \begin{pmatrix} 1.0000 & 0 & 0 & 0 & 0 & 0 \\ 1.0000 & 1.0000 & 0 & 0 & 0 & 0 \\ 1.0000 & 0.4000 & 1.0000 & 0 & 0 & 0 \\ 1.0000 & 0.8000 & 0.6667 & 1.0000 & 0 & 0 \\ 1.0000 & 0.2000 & 0.6667 & -0.5000 & 1.0000 & 0 \\ 1.0000 & 0.6000 & 1.0000 & 0.7500 & -0.5000 & 1.0000 \end{pmatrix}$$

sowie

$$U = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ 0 & 5.0000 & 20.0000 & 55.0000 & 125.0000 & 251.0000 \\ 0 & 0 & -3.0000 & -13.0000 & -36.0000 & -80.4000 \\ 0 & 0 & 0 & -1.3333 & -7.0000 & -22.2000 \\ 0 & 0 & 0 & 0 & -0.5000 & -2.7000 \\ 0 & 0 & 0 & 0 & 0 & 0.1000 \end{pmatrix}$$

und `piv=[1 6 3 5 2 4]`, was der Permutationsmatrix

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

entspricht. Letzten Endes haben wir also drei verschiedene (alles aber richtige) Ergebnisse erhalten.

11. Sogenannte Householder-Matrizen bilden die Grundlage des Verfahrens von Householder zur Berechnung einer $QR$-Zerlegung einer Matrix $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$. Hierbei heißt eine Matrix der Form

$$Q := I - \frac{2}{u^T u} uu^T$$

mit $u \in \mathbb{R}^m \setminus \{0\}$ eine *Householder-Matrix*. Man zeige:

(a) Eine Householder-Matrix ist symmetrisch und orthogonal.

(b) Ist $a \in \mathbb{R}^m \setminus \{0\}$ und definiert man $u := a + \text{sign}\,(a_1)\|a\|_2\, e_1$, so ist durch

$$Q := I - \frac{2}{u^T u} uu^T$$

eine Householder-Matrix gegeben, die sich in der Form

$$Q = I - \beta uu^T \qquad \text{mit} \quad \beta := \frac{1}{\|a\|_2(\|a\|_2 + |a_1|)}$$

darstellen lässt und die $a$ in ein Vielfaches des ersten Einheitsvektors überführt. Genauer ist $Qa = -\text{sign}\,(a_1)\|a\|_2\, e_1$. Hierbei sei $\text{sign}\,(0) := 1$.

**Lösung:** Eine Householder-Matrix $Q$ ist ganz offensichtlich symmetrisch. Wegen

$$Q^T Q = Q^2 = \left(I - \frac{2}{u^T u} uu^T\right) = I - \frac{4}{u^T u} uu^T + \frac{4}{u^T u} uu^T = I$$

sind sie auch orthogonal. Seien $u$ und $Q$ wie angegeben definiert. Dann ist

$$u^T u = \|a\|_2^2 + 2\underbrace{a_1 \text{sign}\,(a_1)}_{|a_1|}\|a\|_2 + \underbrace{\text{sign}\,(a_1)^2}_{=1}\|a\|_2^2 = 2\|a\|_2(\|a\|_2 + |a_1|),$$

woraus die angegebene Darstellung der Householder-Matrix folgt. Weiter ist

$$\begin{aligned} Qa &= a - \beta u^T a u \\ &= a - \frac{[a + \text{sign}\,(a_1)\|a\|_2\, e_1]^T a}{\|a\|_2(\|a\|_2 + |a_1|)}[a + \text{sign}\,(a_1)\|a\|_2\, e_1] \\ &= a - [a + \text{sign}\,(a_1)\|a\|_2\, e_1] \\ &= -\text{sign}\,(a_1)\|a\|_2\, e_1. \end{aligned}$$

Damit ist die Aufgabe gelöst.

12. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ und $\text{Rang}\,(A) = n$. Sei $A = \hat{Q}\hat{R} = Q^*R^*$ jeweils eine reduzierte $QR$-Zerlegung von $A$ (die Spalten der $m \times n$-Matrizen $\hat{Q}, Q^*$ bilden also ein Orthonormalsystem, die $n \times n$-Matrizen sind obere Dreiecksmatrizen). Man beachte, dass $\hat{R}, R^*$ wegen der Rangvoraussetzung an $A$ nichtsingulär sind, die Diagonalelemente also von Null verschieden sind. Man zeige: Ist $\text{sign}\,(\hat{r}_{jj}) = \text{sign}\,(r^*_{jj})$, $j = 1, \ldots, n$, so ist $\hat{Q} = Q^*$ und $\hat{R} = R^*$.

**Lösung:** Wegen $\hat{Q}^T\hat{Q} = I$ und der Nichtsingularität von $R^*$ folgt aus $\hat{Q}\hat{R} = Q^*R^*$, dass $\hat{R}(R^*)^{-1} = \hat{Q}^TQ^*$. Hier steht links eine obere Dreiecksmatrix, die nach Voraussetzung positive Diagonalelemente besitzt. Andererseits ist

$$[\hat{R}(R^*)^{-1}]^T = (Q^*)^T\hat{Q} = R^*\hat{R}^{-1}$$

ebenfalls eine obere Dreicksmatrix. Daher ist $D := \hat{R}(R^*)^{-1}$ eine Diagonalmatrix mit positiven Diagonalelementen, ferner ist $D^{-1} = D$. Daher ist $D = I$ und folglich $\hat{R} = R^*$ und $\hat{Q} = Q^*$ (aus der Darstellung $\hat{Q}\hat{R} = Q^*R^*$ unter Benutzung der Nichtsingularität von $\hat{R} = R^*$).

13. Sei $A = (\ a_1\ \cdots\ a_n\ ) \in \mathbb{R}^{n\times n}$. Man beweise die Hadamardsche Determinantenungleichung

$$|\det(A)| \leq \prod_{i=1}^{n} \|a_i\|_2.$$

Hinweis: Man benutze eine $QR$-Zerlegung von $A$ und zeige, dass $\|a_i\|_2^2 \geq r_{ii}^2$, $i = 1, \ldots, n$.

**Lösung:** Sei $A = QR$ eine $QR$-Zerlegung von $A$. Es ist

$$\|a_i\|_2^2 = (A^TA)_{ii} = (R^T\underbrace{Q^TQ}_{=I}R)_{ii} = (R^TR)_{ii} = \sum_{k=1}^{i} r_{ki}^2 \geq r_{ii}^2, \qquad i = 1, \ldots, n.$$

Dann ist

$$\det(A)^2 = \det(A^TA) = \det(R^TR) = \det(R)^2 = \prod_{i=1}^{n} r_{ii}^2 \leq \prod_{i=1}^{n} \|a_i\|_2^2,$$

woraus die Behauptung folgt.

14. Sei

$$A := \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Man benutze Maple, MATLAB und eine (z. B. in MATLAB) selbst geschriebene Funktion zur Berechnung einer (vollen) $QR$-Zerlegung von $A$.

Zunächst benutzen wir Maple und erhalten:

```
> with(LinearAlgebra):
> A:=Matrix([[1,1,1,1],[0,1,1,1],[0,0,1,1],[0,0,0,1],[1,2,3,4]]):
```

```
> (Q,R):=QRDecomposition(A,fullspan);
```

$$Q,\, R := \begin{bmatrix} \frac{1}{2}\sqrt{2} & -\frac{1}{6}\sqrt{6} & -\frac{1}{6}\sqrt{3} & -\frac{1}{10}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & \frac{1}{3}\sqrt{6} & -\frac{1}{6}\sqrt{3} & -\frac{1}{10}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & 0 & \frac{1}{2}\sqrt{3} & -\frac{1}{10}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ 0 & 0 & 0 & \frac{2}{5}\sqrt{5} & \frac{1}{5}\sqrt{5} \\ \frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6} & \frac{1}{6}\sqrt{3} & \frac{1}{10}\sqrt{5} & -\frac{1}{5}\sqrt{5} \end{bmatrix}, \begin{bmatrix} \sqrt{2} & \frac{3}{2}\sqrt{2} & 2\sqrt{2} & \frac{5}{2}\sqrt{2} \\ 0 & \frac{1}{2}\sqrt{6} & \frac{2}{3}\sqrt{6} & \frac{5}{6}\sqrt{6} \\ 0 & 0 & \frac{2}{3}\sqrt{3} & \frac{5}{6}\sqrt{3} \\ 0 & 0 & 0 & \frac{1}{2}\sqrt{5} \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

```
> Equal(Q.R,A);
```

*true*

```
> evalf(Q);
```

$$\begin{bmatrix} .7071067810 & -.4082482906 & -.2886751347 & -.2236067978 & .4472135956 \\ 0. & .8164965809 & -.2886751347 & -.2236067978 & .4472135956 \\ 0. & 0. & .8660254040 & -.2236067978 & .4472135956 \\ 0. & 0. & 0. & .8944271912 & .4472135956 \\ .7071067810 & .4082482906 & .2886751347 & .2236067978 & -.4472135956 \end{bmatrix}$$

```
> evalf(R);
```

$$\begin{bmatrix} 1.414213562 & 2.121320343 & 2.828427124 & 3.535533905 \\ 0. & 1.224744872 & 1.632993162 & 2.041241452 \\ 0. & 0. & 1.154700539 & 1.443375673 \\ 0. & 0. & 0. & 1.118033989 \\ 0. & 0. & 0. & 0. \end{bmatrix}$$

Mit der `qr`-Funktion in MATLAB erhalten wir (mit `format short`):

$$Q = \begin{pmatrix} -0.7071 & 0.4082 & 0.2887 & 0.2236 & -0.4472 \\ 0 & -0.8165 & 0.2887 & 0.2236 & -0.4472 \\ 0 & 0 & -0.8660 & 0.2236 & -0.4472 \\ 0 & 0 & 0 & -0.8944 & -0.4472 \\ -0.7071 & -0.4082 & -0.2887 & -0.2236 & 0.4472 \end{pmatrix}$$

und

$$R = \begin{pmatrix} -1.4142 & -2.1213 & -2.8284 & -3.5355 \\ 0 & -1.2247 & -1.6330 & -2.0412 \\ 0 & 0 & -1.1547 & -1.4434 \\ 0 & 0 & 0 & -1.1180 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Zum Schluss geben das mit der Funktion `QRGivens` erhaltene Ergebnis an. Hier ist

$$Q = \begin{pmatrix} 0.7071 & -0.4082 & -0.2887 & -0.2236 & -0.4472 \\ 0 & 0.8165 & -0.2887 & -0.2236 & -0.4472 \\ 0 & 0 & 0.8660 & -0.2236 & -0.4472 \\ 0 & 0 & 0 & 0.8944 & -0.4472 \\ 0.7071 & 0.4082 & 0.2887 & 0.2236 & 0.4472 \end{pmatrix}$$

und

$$R = \begin{pmatrix} 1.4142 & 2.1213 & 2.8284 & 3.5355 \\ 0 & 1.2247 & 1.6330 & 2.0412 \\ 0 & 0 & 1.1547 & 1.4434 \\ 0 & 0 & 0 & 1.1180 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Bis auf Vorzeichenverteilungen stimmen diese Ergebnisse also gut überein.

15. Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch, so ist

$$\lambda_{\min}(A)\,\|x\|_2^2 \le x^T A x \le \lambda_{\max}(A)\,\|x\|_2^2 \qquad \text{für alle } x \in \mathbb{R}^n,$$

wobei $\lambda_{\min}(A)$ den kleinsten und $\lambda_{\max}(A)$ den größten Eigenwert von $A$ bedeutet.

Hinweis: Man kann benutzen, dass eine symmetrische Matrix durch eine Ähnlichkeitstransformation mit einer orthogonalen Matrix auf Diagonalgestalt transformiert werden kann bzw., äquivalent dazu, ein vollständiges Orthonormalystem von Eigenvektoren existiert.

**Lösung:** Seien

$$\lambda_{\max}(A) = \lambda_1 \ge \cdots \ge \lambda_n = \lambda_{\min}(A)$$

die Eigenwerte von $A$ und $U \in \mathbb{R}^{n\times n}$ eine orthogonale Matrix mit

$$U^T A U = \Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Sei $x \in \mathbb{R}^n$ beliebig. Dann ist

$$x^T A x = (U^T x)^T \Lambda U^T x = \sum_{j=1}^{n} \lambda_j (U^T x)_j^2 \le \lambda_{\max}(A)\,\|U^T x\|_2^2 = \lambda_{\max}(A)\,\|x\|_2^2,$$

wobei wir benutzt haben, dass die euklidische Norm eines Vektors bei Transformation mit einer orthogonalen Matrix invariant bleibt. Die linke Ungleichung kann entsprechend bewiesen werden.

16. Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit. Man zeige, dass es positive Konstanten $c_0, C_0$ mit

$$c_0\,\|Ax\|_2^2 \le x^T A x \le C_0\,\|Ax\|_2^2 \qquad \text{für alle } x \in \mathbb{R}^n$$

gibt. Insbesondere gilt: Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semidefinit, so folgt aus $x^T A x = 0$, dass $Ax = 0$.

**Lösung:** Wir können o. B. d. A. annehmen, dass $A \neq 0$. Daher besitzt $A$ mindestens einen positiven Eigenwert. Seien $\lambda_1 \ge \cdots \ge \lambda_r$ die positiven, $\lambda_{r+1} = \cdot = \lambda_n = 0$ die restlichen Eigenwerte. Ferner sei $\{u_1, \ldots, u_n\}$ ein zugehöriges Orthonormalsystem von Eigenvektoren. Ein beliebiges $x \in \mathbb{R}^n$ hat die eindeutige Darstellung

$$x = \sum_{j=1}^{n} \alpha_j u_j.$$

Dann ist offenbar

$$x^T A x = \sum_{j=1}^{n} \lambda_j \alpha_j^2 = \sum_{j=1}^{r} \lambda_j \alpha_j^2$$

und

$$\|Ax\|_2^2 = \sum_{j=1}^{n} \lambda_j^2 \alpha_j^2 = \sum_{j=1}^{r} \lambda_j^2 \alpha_j^2.$$

Hieraus liest man ab, dass die Aussage mit $c_0 := 1/\lambda_1$ und $C_0 := 1/\lambda_r$ richtig ist.

## 7.2.2 Aufgaben zu Abschnitt 2.2

1. Gegeben[4] sei das nichtlineare Gleichungssystem

$$f(x_1, x_2) := \begin{pmatrix} x_2 \exp(x_1) - 2 \\ x_1^2 + x_2 - 4 \end{pmatrix} = 0.$$

Durch Elimination von $x_2$ führe man diese Aufgabe auf die Bestimmung der Schnittpunkte zweier Funktionen in einer unabhängigen Variablen zurück. Mit Hilfe von Maple oder MATLAB bestimme man Näherungen für Lösungen des gegebenen nichtlinearen Gleichungssystems.

**Lösung:** Offenbar hat man $2\exp(-x_1) = 4 - x_1^2$ zu lösen. Wie man an Abbildung 7.6 erkennt, gibt es zwei Lösungen, eine bei $x_1 \approx -0.6$, die andere bei $x_1 \approx 2$. Ohne

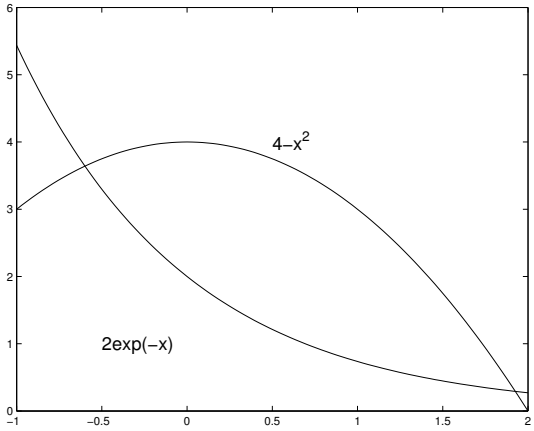


Abbildung 7.6: Die Gleichung $2\exp(-x) = 4 - x^2$

Elimination erhält man dank Maple

```
> fsolve({x_2*exp(x_1)-2=0,x_1^2+x_2-4=0},{x_1,x_2});
```

$$\{x\_1 = 1.925737122,\ x\_2 = .2915365365\}$$

```
> fsolve({x_2*exp(x_1)-2=0,x_1^2+x_2-4=0},{x_1,x_2},
> {x_1=-0.7..-0.5,x_2=3..4});
```

$$\{x\_1 = -.5991247824,\ x\_2 = 3.641049495\}$$

2. Als Variante zu Satz 2.2 zeige man: Sei $f \in C^2[a,b]$ mit $f'(x) > 0$ und $f''(x) \leq 0$ für alle $x \in [a,b]$, d. h. die auf $[a,b]$ zweimal stetig differenzierbare Funktion $f$ sei auf $[a,b]$ streng monoton wachsend und konkav. Ferner sei $f(a) < 0 < f(b)$ und $b - f(b)/f'(b) \geq a$. Dann liefert das Newton-Verfahren

$$x_{k+1} := x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, \ldots$$

für jedes $x_0 \in [a,b]$ eine Folge $\{x_k\}$ mit der Eigenschaft, dass $\{x_k\}_{k\in\mathbb{N}}$ monoton wachsend und von mindestens zweiter Ordnung gegen die einzige Nullstelle $x^*$ von $f$ in $[a,b]$ konvergiert, wobei natürlich $x_k \neq x^*$ für alle $k$ angenommen wird.

[4] Siehe

N. Köckler (1990, S. 138) *Numerische Algorithmen in Softwaresystemen*. B. G. Teubner, Stuttgart.

**Lösung:** Die Folge $\{x_k\}$ wird durch die Iterationsfunktion

$$F(x) := x - \frac{f(x)}{f'(x)}$$

erzeugt, d. h. es ist $x_{k+1} = F(x_k)$, $k = 0, \ldots,$. Dann ist

$$F'(x) = \frac{f(x)f''(x)}{f'(x)^2} \begin{cases} \geq 0, & x \in [a, x^*], \\ \leq 0, & x \in [x^*, b]. \end{cases}$$

Sei $x_0 \in [a, b]$ beliebig. Wir zeigen, dass $a \leq x_k \leq x_{k+1} \leq x^*$, $k = 1, \ldots$. Nach spätestens einem Schritt sind also alle Iterierten des Newton-Verfahrens links der einzigen Nullstelle $x^*$ von $f$ in $[a, b]$, ferner ist die Folge $\{x_k\}_{k\in\mathbb{N}}$ monoton wachsend. Denn ist $x^* \leq x_0 \leq b$, so ist

$$a \leq F(b) \leq F(x_0) = x_1 \leq F(x^*) = x^*,$$

d. h. nach einem Schritt ist man links von $x^*$, aber immer noch rechts von $a$. Ist aber $x_k \in [a, x^*]$, so ist

$$x^* = F(x^*) \geq F(x_k) = x_{k+1} \geq x_k \geq a.$$

Als monoton wachsende, nach oben durch $x^*$ beschränkte Folge ist $\{x_k\}_{k\in\mathbb{N}}$ konvergent. Der Limes ist notwendigerweise eine Nullstelle von $f$, stimmt also mit $x^*$ überein. Die quadratische Konvergenz folgt aus dem lokalen Konvergenzsatz.

3. Mit Hilfe von Aufgabe 2 überlege man sich, für welche Startwerte das Newton-Verfahren, angewandt auf die Nullstellenaufgabe $f(x) := x - e^{-x} = 0$, eine konvergente Folge bildet. Mit $x_0 := 1.0$ berechne man die ersten fünf Iterierten.

   **Lösung:** Es ist sehr leicht zu sehen, dass die Voraussetzungen von Aufgabe 2 mit $[a, b] := [0, 1]$ erfüllt sind. Wir erhalten die folgenden Werte:

| $k$ | $x_k$ |
|---|---|
| 0 | 1.00000000000000 |
| 1 | 0.53788284273999 |
| 2 | 0.56698699140541 |
| 3 | 0.56714328598912 |
| 4 | 0.56714329040978 |
| 5 | 0.56714329040978 |

   Nach einem Schritt ist man links der Nullstelle, danach hat man Monotonie.

4. Die Funktion $f(x) := xe^{-x}$ hat $x^* = 0$ als einzige Nullstelle. Man untersuche, für welche Startwerte das Newton-Verfahren eine gegen $x^*$ konvergente Folge liefert.

   **Lösung:** Das Newton-Verfahren liefert die Vorschrift

$$x_{k+1} := -\frac{x_k^2}{1 - x_k}, \qquad k = 0, 1, \ldots.$$

   Ist $x_0 \in (1, \infty)$, so ist $x_0 \leq x_1$, es kann also keine Konvergenz vorliegen. Ist $x_0 \in (0, 1)$, so ist $x_1 < 0$. Für $x_0 \in (-\infty, 0)$ ist $x_0 \leq x_1 < 0$. Daher liegt Konvergenz der Folge $\{x_k\}$ für jeden Startwert aus $(-\infty, 1)$ vor.

5. Durch $f_0 := 1$, $f_1 := 1$ und $f_{k+1} := f_k + f_{k-1}$, $k = 1, 2, \ldots$, sei die Fibonacci-Folge $\{f_k\}$ definiert. Ferner sei $\tau := (1 + \sqrt{5})/2$.

   (a) Man beweise die Binetsche Formel

   $$f_k = \frac{1}{\sqrt{5}}[\tau^{k+1} - (-\tau)^{-(k+1)}], \qquad k = 0, 1, \ldots.$$

   (b) Man zeige, dass

   $$\lim_{k\to\infty} \frac{f_{k+1}}{f_k} = \tau.$$

   **Lösung:** Den ersten Teil beweisen wir durch vollständige Induktion nach $k$. Es ist $\sigma := -1/\tau = (1 - \sqrt{5})/2$ die negative Lösung von $x^2 - x = 1$, während $\tau$ die positive ist. Hiermit lautet die Binetsche Formel

   $$f_k = \frac{1}{\sqrt{5}}[\tau^{k+1} - \sigma^{k+1}], \qquad k = 0, 1, \ldots.$$

   Dann ist

   $$\frac{1}{\sqrt{5}}[\tau - \sigma] = 1 = f_0$$

   und

   $$\frac{1}{\sqrt{5}}[\tau^2 - \sigma^2] = \frac{1}{\sqrt{5}}[\tau - \sigma] = 1 = f_1.$$

   Angenommen, die Formel sei für $k - 1$ und $k$ richtig. Dann ist

   $$\begin{aligned} f_{k+1} &= f_k + f_{k-1} \\ &= \frac{1}{\sqrt{5}}[\tau^{k+1} - \sigma^{k+1}] + \frac{1}{\sqrt{5}}[\tau^k - \sigma^k] \\ &= \frac{1}{\sqrt{5}}[\tau^k(\underbrace{\tau + 1}_{=\tau^2}) - \sigma^k(\underbrace{\sigma + 1}_{\sigma^2})] \\ &= \frac{1}{\sqrt{5}}[\tau^{k+1} - \sigma^{k+1}]. \end{aligned}$$

   Damit ist die Binetsche Formel bewiesen.

   Wir setzen weiter $\sigma := -1/\tau$. Wegen

   $$\frac{f_{k+1}}{f_k} = \frac{\tau^{k+2} - \sigma^{k+2}}{\tau^{k+1} - \sigma^{k+1}} = \frac{\tau - (\sigma/\tau)^{k+1}\sigma}{1 - (\sigma/\tau)^{k+1}} \overset{k\to\infty}{\to} \tau$$

   folgt der zweite Teil unmittelbar.

6. Gegeben sei die Nullstellenaufgabe $f(x) := x + \ln x = 0$. Durch einen Plot von $f$ über dem Intervall $[0.3, 1]$ verschaffe man sich einen Überblick über eventuelle Nullstellen. Mit Hilfe der Maple-Funktion `fsolve` oder die MATLAB-Funktion `fzero` bestimme man Näherungslösungen.

   **Lösung:** In Abildung 7.7 links ist ein in Maple durch `plot(x+ln(x),x=0.3..1);` erzeugtes Bild zu sehen. Die rechte Abbildung ist mit Hilfe von MATLAB erzeugt worden. Nach `x:=fsolve(x+ln(x),x);` erhält man $x = .5671432904$, während in MATLAB der Befehl `x=fzero(inline('x+log(x)'),0.6)` den Wert $x = 0.56714329040978$ ergibt (mit `format long`).

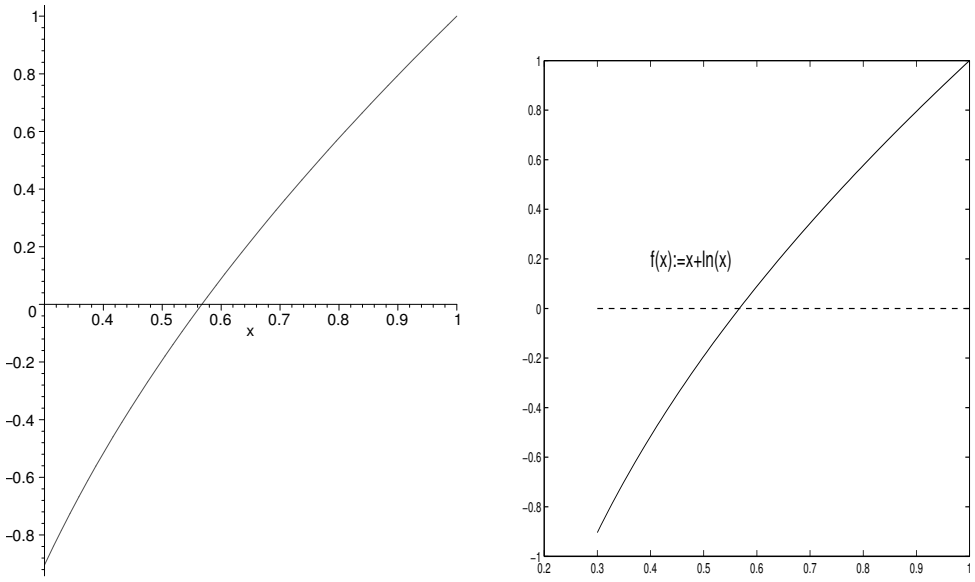


Abbildung 7.7: Die Funktion $f(x) := x + \ln(x)$

7. Die Frobeniusnorm $\|\cdot\|_F$ auf $\mathbb{R}^{n\times n}$ ist für $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ definiert durch $\|A\|_F := (\sum_{i,j=1}^n a_{ij}^2)^{1/2}$. Man zeige:

   (a) Es ist $\|A\|_F = \sqrt{\text{Spur}(A^T A)}$ für jedes $A \in \mathbb{R}^{n\times n}$. Hierbei ist die Spur einer Matrix die Summe ihrer Diagonalelemente.

   (b) Die Frobeniusnorm $\|\cdot\|_F : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$ hat die Eigenschaften einer Norm (Definitheit, Homogenität und Dreiecksungleichung) und ist darüberhinaus *submultiplikativ*, d. h. es ist $\|AB\| \le \|A\|\,\|B\|$ für alle $A, B \in \mathbb{R}^{n\times n}$.

   (c) Es ist $\|A\|_2 \le \|A\|_F$ für alle $A \in \mathbb{R}^{n\times n}$.

   **Lösung:** Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ beliebig. Dann ist

$$\text{Spur}(A^T A) = \sum_{j=1}^n (A^T A)_{jj} = \sum_{j=1}^n \sum_{i=1}^n (A^T)_{ji}(A)_{ij} = \sum_{i,j=1}^n a_{ij}^2 = \|A\|_F^2,$$

   woraus die erste Behauptung folgt.

   Die Definitheit und Homogenität der Frobeniusnorm sind klar. Die Gültigkeit der Dreiecksungleichg erkennt man sofort, wenn man $\|A\|_F$ für $A \in \mathbb{R}^{n\times n}$ als euklidische Norm eines Vektors im $\mathbb{R}^{n^2}$ auffasst, der die Einträge von $A$ als Komponenten besitzt. Für $A, B \in \mathbb{R}^{n\times n}$ ist unter Benutzung der Cauchy-Schwarzschen Ungleichung

$$\|AB\|_F^2 = \sum_{i,j=1}^n (AB)_{ij}^2 = \sum_{i,j=1}^n \left(\sum_{k=1}^n a_{ik} b_{kj}\right)^2 \le \sum_{i=1,j=1}^n \left(\sum_{k=1}^n a_{ik}^2 \sum_{k=1}^n b_{kj}^2\right) = \|A\|_F^2\,\|B\|_F^2.$$

   Die Frobeniusnorm ist also submultiplikativ.

   Für beliebige $A \in \mathbb{R}^{n\times n}$ und $x \in \mathbb{R}^n$ ist mit Hilfe der Cauchy-Schwarzschen Ungleichung

$$\|Ax\|_2^2 = \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij} x_j\right)^2 \le \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}^2 \sum_{j=1}^n x_j^2\right) = \|A\|_F^2\,\|x\|_2^2.$$

Daher ist

$$\|A\|_2 := \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \|A\|_F.$$

Damit ist die Aufgabe gelöst.

8. Man beweise, dass

$$\lim_{k\to\infty} \underbrace{\sqrt{2+\sqrt{2+\cdots+\sqrt{2}}}}_{k \text{ Wurzeln}} = 2.$$

**Lösung:** Man definiere $F : [0,\infty) \subset \mathbb{R} \longrightarrow \mathbb{R}$ durch $F(x) := \sqrt{2+x}$. Das Intervall $D := [0,\infty)$ ist abgeschlossen in $(\mathbb{R}, |\cdot|)$ und wird durch $F$ in sich abgebildet. Ferner kontrahiert $F$ auf $D$, da

$$q := \sup_{\xi\in D} |F'(\xi)| = \sup_{\xi\in[0,\infty)} \frac{1}{2\sqrt{2+\xi}} = \frac{1}{2\sqrt{2}} < 1.$$

Für jedes $x_0 \in [0,\infty)$ konvergiert die durch $x_{k+1} := F(x_k)$ gewonnene Folge $\{x_k\}$ wegen des Banachschen Fixpunktsatzes gegen den einzigen Fixpunkt $x^*$ von $F$ in $D$. Setzt man $x_0 := \sqrt{2}$, so wird $x_1 = \sqrt{2+\sqrt{2}}$ und allgemein

$$x_{k-1} = \underbrace{\sqrt{2+\sqrt{2+\cdots+\sqrt{2}}}}_{k \text{ Wurzeln}}.$$

Der einzige Fixpunkt von $F$ in $D$ ist aber $x^* = 2$, womit die Aufgabe gelöst ist.

9. Gegeben sei das nichtlineare Gleichungssystem

$$f(x) := \begin{pmatrix} x_1 - 0.1x_1^2 - \sin x_2 \\ x_2 - \cos x_1 - 0.1x_2^2 \end{pmatrix} = 0.$$

   (a) Aus irgendeinem Grund vermuten Sie, dass das nichtlineare Gleichungssystem $f(x) = 0$ eine Lösung in $[0,1] \times [0,1]$ besitzt. Verschaffen Sie sich durch den `implicitplot` Befehl im `plots`-package von Maple eine Näherung.

   (b) Benutzen Sie `fsolve` in Maple, um das System $f(x) = 0$ zu lösen. Bestimmen Sie mehr als eine Lösung.

   (c) Benutzen Sie eine selbst (z. B. in MATLAB) geschriebene Funktion, um das Gleichungssystem zu lösen.

**Lösung:** Mit Hilfe von

```
g1:=implicitplot(x_1-0.1*x_1^2-sin(x_2)=0,x_1=0..1,x_2=0..1):
g2:=implicitplot(x_2-cos(x_1)-0.1*x_2^2=0,x_1=0..1,x_2=0..1):
display([g1,g2]);
```

erhalten wir den in Abbildung 7.8 dargestellten plot: Hiernach erkennen wir, dass eine Lösung bei $x^* \approx (0.8, 0.8)$ existiert.

Mit `fsolve` erhalten wir:

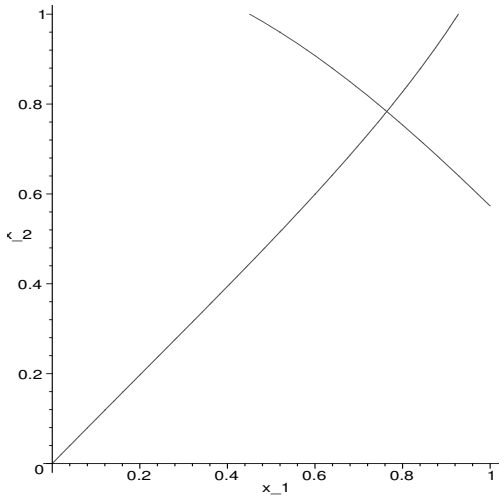


Abbildung 7.8: Schnitt von $x_1 - 0.1x_1^2 - \sin x_2 = 0,\ x_2 - \cos x_1 - 0.1x_2^2 = 0$

```
> fsolve({x_1-0.1*x_1^2-sin(x_2)=0,x_2-cos(x_1)-0.1*x_2^2=0},
> {x_1,x_2});
```

$$\{x\_2 = -.4925720046,\ x\_1 = 10.45242487\}$$

```
> fsolve({x_1-0.1*x_1^2-sin(x_2)=0,x_2-cos(x_1)-0.1*x_2^2=0},
> {x_1,x_2},{x_1=0..1,x_2=0..1});
```

$$\{x\_2 = .7833967743,\ x\_1 = .7640705508\}$$

Wir haben also zwei Lösungen erhalten.

Wir benutzen die von uns geschriebene MATLAB-Funktion `NewtonGLS`. Hierzu schreiben wir ein function file `fun.m`, in dem die Funktion und ihre Funktionalmatrix berechnet wird:

```
function [f,f_strich]=fun(x);
f=[x(1)-0.1*x(1)^2-sin(x(2));x(2)-cos(x(1))-0.1*x(2)^2];
f_strich=[1-0.2*x(1),-cos(x(2));sin(x(1)),1-0.2*x(2)];
```

Nach dem Aufruf von MATLAB geben wir ein:

```
[x,iter]=NewtonGLS(@fun,[0.8;0.8],1e-12,20);
```

Mit `format long` erhalten wir

$$x = \begin{pmatrix} 0.76407055081274 \\ 0.78339677430048 \end{pmatrix}, \qquad \text{iter} = 4.$$

Mit dem Startwert $x_0 = (0,0)^T$ benötigen wir 6 Iterationen, um denselben Wert zu erhalten. Nach

```
[x,iter]=NewtonGLS(@fun,[10;0],1e-12,20);
```

ist

$$x = \begin{pmatrix} 10.45242486716671 \\ -0.49257200461440 \end{pmatrix}, \qquad \text{iter} = 5,$$

nach

```
[x,iter]=NewtonGLS(@fun,[0;10],1e-12,20);
```

ist

$$x = \begin{pmatrix} 0.43776925371129 \\ 8.99286917618392 \end{pmatrix}, \qquad \text{iter} = 6$$

das Resultat.

10. Gegeben[5] sei die Abbildung $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ mit

$$f(x) := \begin{pmatrix} \exp(x_1^2 + x_2^2) - 3 \\ x_1 + x_2 - \sin(3(x_1 + x_2)) \end{pmatrix}.$$

Man bestimme die Funktionalmatrix $f'(x)$. Für welche $x$ ist $f'(x)$ singulär?

**Lösung:** Die Funktionalmatrix ist durch

$$f'(x) = \begin{pmatrix} 2\exp(x_1^2 + x_2^2)x_1 & 2\exp(x_1^2 + x_2^2)x_2 \\ 1 - 3\cos(3(x_1 + x_2)) & 1 - 3\cos(3(x_1 + x_2)) \end{pmatrix}.$$

Die Funktionalmatrix ist singulär für die $x$, für die

$$\det f'(x) = 2\exp(x_1^2 + x_2^2)[1 - 3\cos(3(x_1 + x_2))](x_1 - x_2) = 0.$$

Daher ist $f'(x)$ singulär, wenn $x_1 = x_2$ oder $\cos(3(x_1 + x_2)) = 1/3$.

## 7.3 Aufgaben zu Kapitel 3

### 7.3.1 Aufgaben zu Abschnitt 3.1

1. Man gebe einen Euler-Zug im vollständigen Graphen $K_5$ an. Ferner begründe man, weshalb $K_n$ für ungerades $n$ ein Euler-Graph ist.

   **Lösung:** In Abbildung 7.9 ist der $K_5$ dargestellt. Ein Euler-Zug ist z. B. gegeben durch (wir geben nur die Ecken an)

$$\{1, 4, 2, 5, 4, 3, 5, 1, 2, 3, 1\}.$$

   Für ungerades $n$ ist der Grad jeder Ecke in $K_n$ gerade, nämlich gleich $n - 1$. Wegen Satz 1.2 ist $K_n$ dann Eulersch.

2. Man zeige: Kann der Zusammenhang eines Graphen $G$ durch die Entnahme einer einzigen Ecke und sämtlicher mit dieser Ecke inzidierender Kanten zerstört werden, so ist $G$ kein Hamiltonscher Graph.

   **Lösung:** Sei $G = (V, E)$ ein zusammenhängender Graph und $v \in V$ eine Ecke mit der Eigenschaft, dass $(V \setminus \{v\}, E \setminus E(v))$ nicht zusammenhängend ist, wobei $E(v)$ die Menge der mit $v$ inzidierenden Kanten in $E$ ist. Wir nehmen an, es gäbe einen Kreis, der durch sämtliche Ecken von $G$ führt, insbesondere durch $v$. Dieser Kreis verlässt und erreicht $v$ jeweils über eine Kante aus $E(v)$, weitere Kanten aus $E(v)$ können im Kreis nicht vorkommen. Das wäre aber ein Widerspruch dazu, dass $(V \setminus \{v\}, E \setminus E(v))$ nicht zusammenhängend ist.

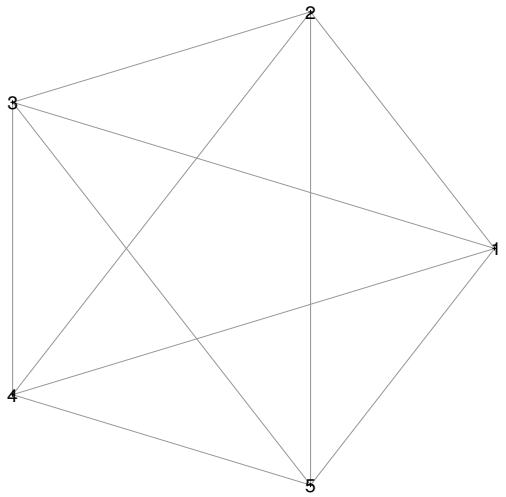


Abbildung 7.9: Der vollständige Graph $K_5$

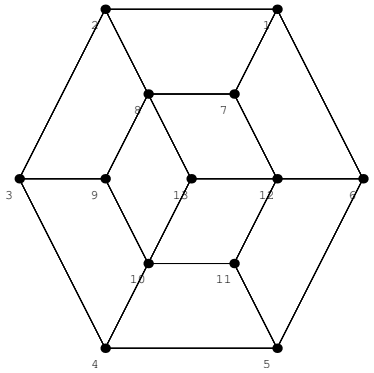


Abbildung 7.10: Bipartiter Graph mit ungerader Eckenzahl

3. Man zeige, dass ein bipartiter Graph mit einer ungeraden Zahl von Ecken nicht Hamiltonsch ist. Hiermit zeige man, dass der in Abbildung 7.10 angegebene Graph nicht Hamiltonsch ist.

   **Lösung:** Die Eckenmenge $V$ eines bipartiten Graphen $G = (V, E)$ kann so in zwei Mengen $U$ und $W$ zerlegt werden, dass jede Kante $e \in E$ in $U$ startet und in $W$ endet. Jeder Hamiltonsche Kreis muss zwischen diesen beiden Mengen alternieren und in der Menge enden, in der er startete. Daher müssen $U$ und $W$ dieselbe Anzahl von Elementen haben und daher $|V|$ gerade sein. Der in Abbildung 7.10 angegebene Graph ist bipartit (man nehme $U := \{1, 3, 5, 8, 10, 12\}$ und $W := \{2, 4, 6, 7, 9, 11, 13\}$) und hat eine ungerade Zahl von Ecken, ist also wegen der gerade eben bewiesenen Aussage nicht Hamiltonsch.

4. Der Graph $Q_n$, der sogenannte Hyperwürfel, ist folgendermaßen definiert: Die Eckenmenge $V_n$ besteht aus allen $0, 1$-Folgen der Länge $n$, offenbar ist $|V_n| = 2^n$. Zwei Ecken $x, y \in V$ werden durch eine Kante $(x, y)$ verbunden, wenn sich die beiden $0, 1$-Folgen

[5]Diese Aufgabe findet man bei

J. Stoer (1994, S. 356) *Numerische Mathematik 1*. Springer, Berlin-Heidelberg-New York.

an genau einer Stelle unterscheiden. Die Kantenmenge von $Q_n$ wird mit $E_n$ bezeichnet. Man zeige:

(a) Es ist $|E_n| = n2^{n-1}$.

(b) $Q_n$ ist bipartit.

(c) $Q_n$ ist für $n \geq 2$ ein Hamilton-Graph.

**Lösung:** Wir beweisen die erste Behauptung durch vollständige Induktion nach $n$. Für $n = 1$ ist die Aussage richtig. Denn $Q_1$ besitzt die beiden Ecken $\{0\}$ und $\{1\}$, die durch eine Kante verbunden sind, es ist also $|E(Q_1)| = 1$, wie behauptet. In der Induktionsannahme wird davon ausgegangen, dass $|E(Q_n)| = n2^{n-1}$ die Anzahl von Paaren von 0,1-Folgen der Länge $n$ ist, die sich an genau einer Stelle unterscheiden. Seien nun $\tilde{x} = (x, x_{n+1}), \tilde{y} = (y, y_{n+1}) \in \{0,1\}^n \times \{0,1\}$ zwei 0,1-Folgen der Länge $n+1$, die sich an genau einer Stelle unterscheiden. Dies kann daran liegen, dass $x, y$ zwei 0,1-Folgen der Länge $n$ sind, die sich an genau einer Stelle unterscheiden, und $x_{n+1} = y_{n+1} = 0$ oder $x_{n+1} = y_{n+1} = 1$ gilt (das sind unter Berücksichtigung der Induktionsannahme $n2^{n-1} + n2^{n-1} = n2^n$ Möglichkeiten) oder dass $x = y$ (hier gibt es $2^n$ Möglichkeiten) und $x_{n+1} = 0$, $y_{n+1} = 1$ bzw. $x_{n+1} = 1$, $y_{n+1} = 0$. Insgesamt hat daher $Q_{n+1}$ genau $n2^n + 2^n = (n+1)2^n$ Kanten), was zu zeigen war.

Dass der Hyperwürfel $Q_n = (V_n, E_n)$ bipartit ist, ist einfach zu sehen. Sei nämlich $U_n$ die Menge der Ecken $x \in V_n$, bei denen die Anzahl der Einsen gerade ist, entsprechend $W_n$ die Menge der Ecken, bei denen die Anzahl der Einsen ungerade ist. Kanten können nur zwischen Ecken aus $U_n$ und $W_n$ bestehen. Denn zwei Ecken aus $U_n$ bzw. $W_n$ sind entweder gleich oder unterscheiden sich an mehr als einer Stelle. Daher ist $Q_n$ bipartit.

Wir beweisen die Behauptung durch vollständige Induktion nach $n$. Für $n = 2$ ist die Behauptung richtig, denn $(0,0)$, $(1,0)$, $(1,1)$, $(0,1)$, $(0,0)$ ist ein Hamilton-Kreis in $Q_2$. Nun nehmen wir an, $Q_n$ sei ein Hamiltonscher Graph. Die Anzahl der Ecken in $Q_n$ ist $2^n$. Sei $0, z_2, \ldots, z_{2^n-1}, z_{2^n}, 0$ ein Hamilton-Kreis in $Q_n$. Dann ist

$$(0,0), (z_2,0), \ldots, (z_{2^n-1},0), (z_{2^n},0), (z_{2^n},1), (z_{2^n-1},1), \ldots, (z_2,1), (0,1), (0,0)$$

ein Hamilton-Kreis in $Q_{n+1}$. Die Behauptung ist bewiesen.

5. Gegeben sei ein $n \times n$-Schachbrett. Die Felder seien die Ecken eines Graphen. Zwei Ecken sind durch eine Kante verbunden, wenn ein Rösselsprung zwischen ihnen möglich ist. Man zeige, dass für ungerades $n$ der entsprechende Graph kein Hamilton-Graph ist.

   Hinweis: Man benutze die Aussage von Aufgabe 3, dass nämlich ein bipartiter Graph mit einer ungeraden Zahl von Ecken nicht Hamiltonsch ist.

   **Lösung:** Der entsprechende Graph ist bipartit. Denn die $n^2$ Ecken denke man sich in schwarze und weiße partitioniert. Bei einem Rösselsprung verändert sich die Farbe des Feldes. Mit anderen Worten ist der entsprechende Graph bipartit. Für ungerdes $n$ ist $n^2$ ebenfalls ungerade. Da aber ein bipartiter Graph mit einer ungeraden Zahl von Ecken nicht Hamiltonsch ist, folgt die Aussage.

6. Gegeben sei ein zusammenhängender Graph $G = (V, E)$, der genau $k$ Ecken mit ungeradem Grad besitzt. Man zeige:

   (a) Es ist $k$ gerade.

(b) Etwas lax formuliert: Die minimale Anzahl der Kantenzüge, mit denen die Kantenmenge $E$ gezeichnet werden kann, ist $k/2$.

(c) Gegeben sei der Graph in Abbildung 7.11. Man zeige, dass man die Kantenmenge

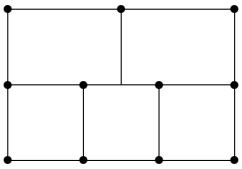

Abbildung 7.11: Wie oft muss man absetzen?

in vier, aber nicht in drei Zügen zeichnen kann.

Hinweis: Der erste Teil der Aufgabe ist eine einfache Folgerung aus dem sogenannten *Handshaking Lemma*, das ohne Beweis benutzt werden darf (obwohl dieser einfach ist). Dieses sagt aus:

- *Sei $G = (V, E)$ ein Graph. Dann ist $\sum_{x \in V} d(x) = 2|E|$, wobei natürlich $d(x)$ den Grad einer Ecke $x \in V$ bedeutet.*

Für den zweiten Teil der Aussage füge man $k/2$ Kanten zu $G$ so dazu, dass der entstehende Graph Eulersch ist.

**Lösung:** Wir bezeichnen mit $V_g$ die Menge der Ecken geraden Grades und mit $V_u$ die Menge der Ecken ungeraden Grades. Wegen des Handshaking Lemmas ist

$$\sum_{x \in V_u} d(x) = 2|E| - \sum_{x \in V_g} d(x)$$

eine gerade Zahl und folglich ist $|V_u|$ gerade.

Je zwei Ecken ungeraden Grades durch eine zusätzliche Kante verbunden. Der hierdurch entstehende Graph hat also $k/2$ zusätzliche Kanten. Hierdurch können zwar Mehrfachkanten entstehen, aber Satz 1.2 gilt, wie wir früher betonten, auch in einem Multigraphen. In ihm hat jede Ecke geraden Grad, es existiert in ihm also ein Euler-Zug. Lassen wir in diesem die $k/2$ hinzugefügten Kanten fort, so erhalten wir die gesuchten $k/2$ Kantenzüge. Mit weniger Kantenzügen kann man andererseits nicht auskommen.

Der dritte Teil folgt sofort aus dem zweiten.

### 7.3.2 Aufgaben zu Abschnitt 3.2

1. Man zeichne alle Bäume mit höchstens 5 Ecken.

   **Lösung:** In Abbildung 7.12 geben wir die acht verschiedenen (bis auf Isomorphie) Bäume mit bis zu 5 Ecken an.

2. Sei $G = (V, E)$ ein zusammenhängender Graph. Mit $d(x, y)$ bezeichnen wir den Abstand zweier Ecken $x, y \in V$, also die Länge eines kürzesten Weges von $x$ nach $y$. Für $x \in V$ definieren wir $r(x) := \max_{y \in V \setminus \{x\}} d(x, y)$ als die *Exzentrizität* in $x$, also die Länge eines längsten von $x$ ausgehenden kürzesten Weges. Dann heißt $r(G) := \min_{x \in V} r(x)$ der *Radius* von $G$ und $Z(G) := \{x \in V : r(x) = r(G)\}$ das *Zentrum* von $G$.

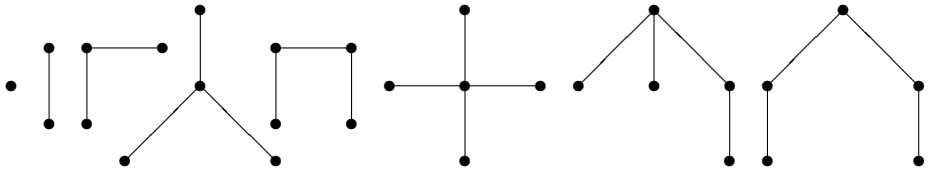

Abbildung 7.12: Bäume mit höchstens 5 Ecken

(a) Was ist der Radius und was das Zentrum zu dem in Abbildung 7.13 dargestellten Graphen $G$?

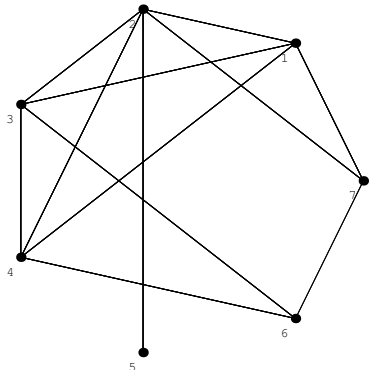


Abbildung 7.13: Was ist der Radius, was ist das Zentrum?

(b) Man zeige, dass das Zentrum eines Baumes entweder aus einer Ecke oder zwei benachbarten Ecken besteht.

**Lösung:** Offenbar ist $r(G) = 2$ und $Z(G) = \{1, 2, 3, 4\}$. Im *Mathematica*-Zusatzpaket `DiscreteMath‘Combinatorica‘` gibt es die Befehle `Radius` und `GraphCenter`. Wendet man `Eccentricity` auf einen Graphen an, so erhält man den Vektor (Länge ist die Ordnung des Graphen) der Exzentrizitäten. In unserem Falle ist dies $\{2, 2, 2, 2, 3, 3, 3\}$, woraus man obige Aussage noch einmal ablesen kann. Etwas vergleichbares scheint es im `networks` package von Maple nicht zu geben.

Wir geben ein Verfahren zur Berechnung des Zentrums eines Baumes an. Jeder Baum hat Ecken vom Grad 1. Diese kommen als Zentrum nicht in Frage. Man streiche sie und die Kante auf der sie liegen. Der entstehende Graph ist wieder ein Baum. Man wiederhole das Verfahren und endet entweder mit einer Ecke oder zwei benachbarten Ecken. In Abbildung 7.14 haben wir zwei Bäume mit 15 Knoten dargestellt. Der linke Baum hat $\{4\}$ als Zentrum, während der rechte Baum $\{1, 7\}$ als Zentrum besitzt.

3. In der folgenden Tabelle sind die Entfernungen (in Hunderten von Meilen) von sechs

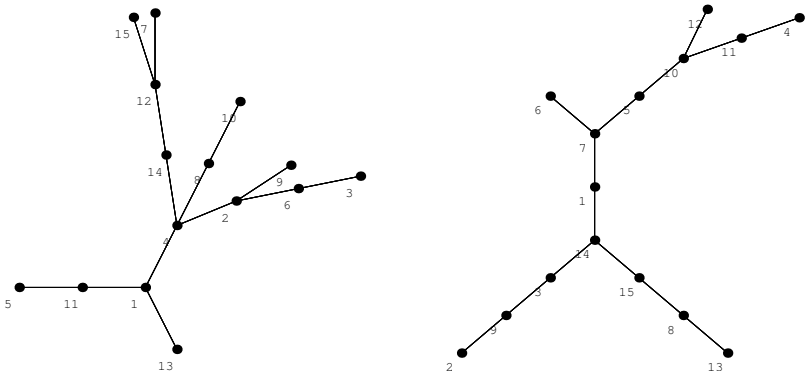

Abbildung 7.14: Zentrum eines Baumes

Städten angegeben.

| | Berlin | London | Moskau | Paris | Rom | Sevilla |
|---|---|---|---|---|---|---|
| Berlin | – | 7 | 11 | 7 | 10 | 15 |
| London | 7 | – | 18 | 3 | 12 | 11 |
| Moskau | 11 | 18 | – | 18 | 20 | 27 |
| Paris | 7 | 3 | 18 | – | 9 | 8 |
| Rom | 10 | 12 | 20 | 9 | – | 13 |
| Sevilla | 15 | 11 | 27 | 8 | 13 | – |

(a) In dem zugehörigen gewichteten vollständigen Graphen bestimme man einen minimalen aufspannenden Baum.

(b) Mit der Minimum Spanning Tree Heuristik bestimme man eine (suboptimale) Rundtour.

**Lösung:** Mit dem Kruskal-Verfahren erhält man offenbar das folgende "Movie" zur Bestimmung des minimalen aufspannenden Baumes. Die ersten drei Schritte sind in Abbildung 7.15 angegeben. Die Verbindung Berlin-Paris darf jetzt nicht gewählt werden,

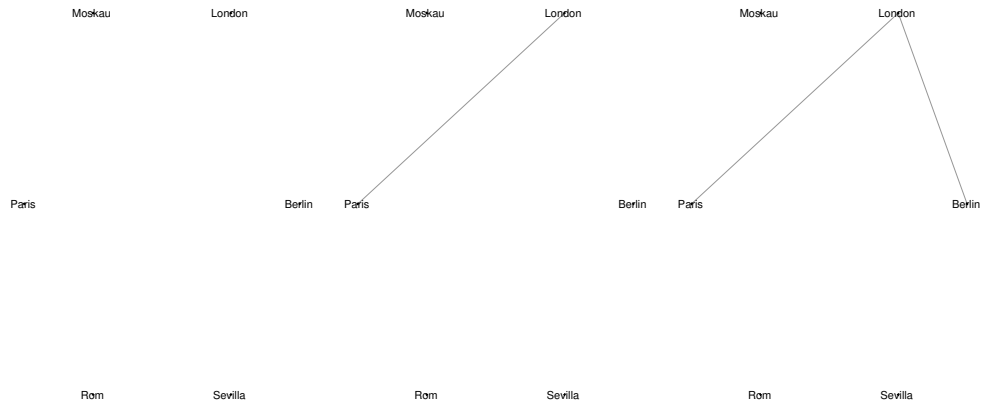


Abbildung 7.15: Die ersten drei Schritte im Kruskal-Verfahren

weil dadurch ein Kreis entstehen würde. Daher wird im nächsten Schritt Paris-Sevilla gewählt. Die restlichen Schritte sind in Abbildung 7.16 angegeben. Das Gesamtgewicht

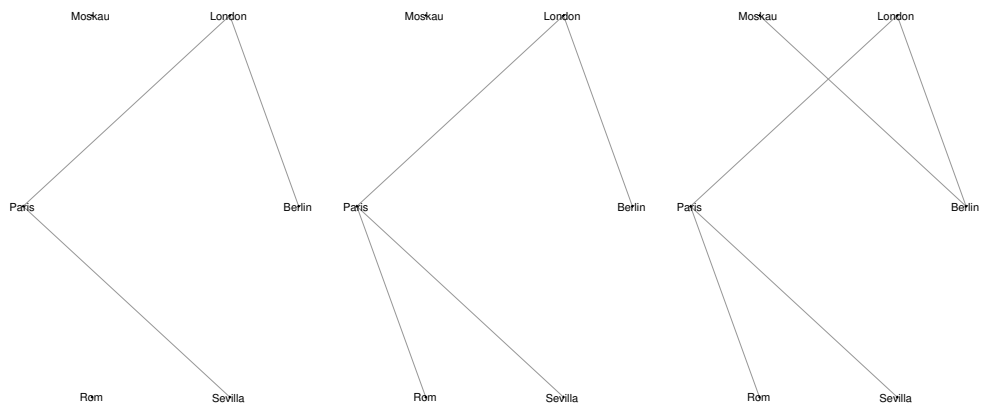


Abbildung 7.16: Die restlichen Schritte

des aufspannenden Baumes ist 38. Das selbe Ergebnis erhalten wir mit `spantree`.

Bei der Minimum Spanning Tree Heuristik müssen wir die Kanten im aufspannenden Baum verdoppeln. Da dann alle Ecken geraden Grad haben, gibt es einen Euler-Zug. Z. B.

$$(P,R),(R,P),(P,S),(S,P),(P,L),(L,B),(B,M),(M,B)(B,L),(L,P).$$

Durch Überspringen schon durchlaufener Ecken gewinnen wir einen Hamilton-Kreis:

$$\{P,R,S,L,B,M,P\}.$$

Die Gesamtdistanz ist 69.

4. Eine Firma hat sich zu überlegen, welches zusammenhängende Pipelinenetzwerk sie zwischen 7 Quellen $A,B,\ldots,G$ und einer Fabrik $H$ bauen sollte. Die möglichen Pipelines und ihre Konstruktionskosten (in gewissen Geldeinheiten) sind gegeben durch

| Pipeline | Kosten | Pipeline | Kosten |
|---|---|---|---|
| $AB$ | 23 | $CG$ | 10 |
| $AE$ | 17 | $DE$ | 14 |
| $AD$ | 19 | $DF$ | 20 |
| $BC$ | 15 | $EH$ | 28 |
| $BE$ | 30 | $FG$ | 11 |
| $BF$ | 27 | $FH$ | 35 |

Welches Pipelinenetzwerk sollte gebaut werden und was sind seine Konstruktionskosten? Was ist die kostengünstigste Verbindung von der Quelle $A$ zur Fabrik $H$?

**Lösung:** Die möglichen Pipelines zwischen den Quellen und der Fabrik mitsamt ihren Kosten ergeben einen gewichteten Graphen Die Schritte zum Aufbau eines minimalen aufspannenden Baumes durch das Verfahren von Kruskal sind die folgenden: Zunächst werden die Kanten $CG$, $FG$, $DE$, $BC$ und $AE$ aufgenommen, da keine Kreise gebildet werden. Die Kante $AD$ wird verworfen, da sonst ein Kreis $ADE$ entstehen würde. Statt dessen wird $DF$ aufgenommen. Auch die Kanten $AB$ bzw. $BF$ werden verworfen, weil dadurch die Kreise $ABCGFDE$ bzw. $BFGC$ entstehen würden. Schließlich wird $EH$ als siebte und letzte Kante aufgenommen. Der minimale aufspannende Baum bzw. das

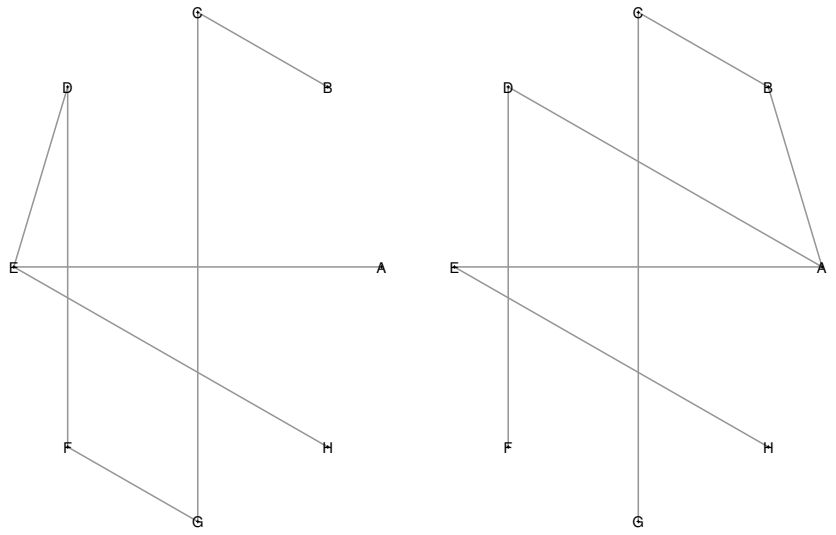


Abbildung 7.17: Billigstes Pipelinenetzwerk, Beste Verbindungen von $A$

kostengünstigste Pipelinenetzwerk ist in Abbildung 7.17 links angegeben. Die Kosten betragen 115 Geldeinheiten.

Nun wenden wir das Verfahren von Dijkstra an. Wir erhalten der Reihe nach die Kanten $AE$, $AD$, $AB$, $BC$, $DF$, $EH$ und $CG$ und damit den in 7.17 rechts dargestellten aufspannenden Baum. Hieraus liest man ab, dass $AEH$ die günstigste Verbindung von $A$ nach $H$ ist.

5. Ein zusammenhängender Graph $G = (V, E)$ habe paarweise verschiedene Gewichte auf den Kanten. Man zeige, dass $G$ einen *eindeutigen* minimalen aufspannenden Baum besitzt.

   **Lösung:** Seien $T^* = (V, E^*)$ und $T' = (V, E')$ zwei verschiedene minimale aufspannende Bäume und $e = uv \in E^* \setminus E'$ eine zu $T^*$, nicht aber zu $T'$ gehörende Kante. Auf dem eindeutigen Weg in $T'$, welcher $u$ mit $v$ verbindet, liegt eine Kante $e'$, welche die beiden Komponenten von $(V, E^* \setminus \{e\})$ miteinander verbindet. Dann sind $(V, (E^* \setminus \{e\}) \cup \{e'\})$ und $(V, (E' \setminus \{e'\}) \cup \{e\})$ jeweils aufspannende Bäume, von denen einer ein kleineres Gewicht als $T^*$ bzw. $T'$ hat (denn das Gewicht auf $e$ ist verschieden von dem auf $e'$). Dies ist ein Widerspruch.

## 7.3.3 Aufgaben zu Abschnitt 3.3

1. Sei $G = (V, E)$ ein zusammenhängender, planarer Graph, bei dem alle Länder Dreiecke sind, also von genau drei Kanten berandet werden. Man zeige, dass dann $|E| = 3|V| - 6$.

   **Lösung:** Wegen des Handshaking-Lemmas für planare Graphen (Lemma 3.6) ist ($F$ bezeichne wieder die Menge der Länder)

$$2|E| = \sum_{l \in F} \deg(l) = \sum_{l \in F} 3 = 3|F|$$

   Wegen der Eulerschen Polyederformel ist andererseits $|V| - |E| + |F| = 2$. Folglich ist

$$2 = |V| - |E| + |F| = |V| - |E| + \frac{2}{3}|E| = |V| - \frac{1}{3}|E|,$$

und hieraus folgt die Behauptung.

2. Ist der in Abbildung 7.18 links angegebene Graph planar?

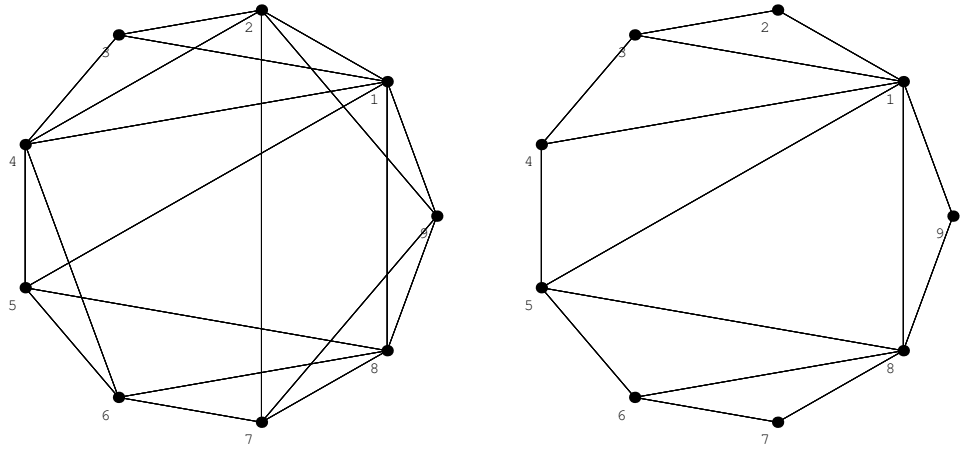


Abbildung 7.18: Ist der Graph links planar?

**Lösung:** Wir wollen zeigen, dass der angegebene Graph planar ist. Ein planarer Untergraph ist offensichtlich in Abbildung 7.18 rechts angegeben. Gegenüber dem gegebenen Graphen fehlen hier noch die Kanten 24, 27, 29, 46 und 79. Diese 5 Kanten können aber leicht "außen herum" überschneidungsfrei gelegt werden.

3. Sei $G = (V, E)$ ein zusammenhängender, planarer Graph mit $n := |V| \geq 5$ Ecken, $m := |E|$ Kanten, ferner habe jeder Kreis eine Länge $\geq 5$. Man zeige:

   (a) Es ist $m \leq \frac{5}{3}(n-2)$.

   (b) Der in Abbildung 7.19 dargestellte Petersen-Graph ist nicht planar.

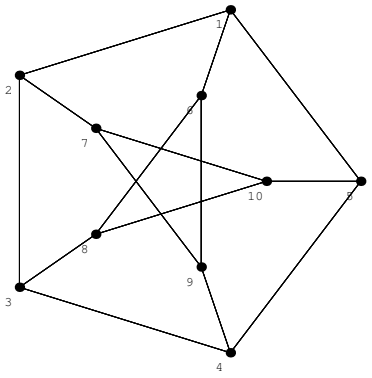


Abbildung 7.19: Der Petersen-Graph ist nicht planar

   (c) Man bestimme eine Eckenfärbung des Petersen-Graphen mit möglichst wenig Farben, wobei natürlich benachbarte Ecken unterschiedliche Farben haben sollen.

**Lösung:** Man betrachte einen zu $G$ isomorphen ebenen Graphen. Jedes Land wird von mindestens fünf Kanten begrenzt, jede Kante gehört zum Rand von höchstens zwei

Ländern. Bezeichnet $f$ die Anzahl der Länder, so ist daher (siehe auch Korollar 3.6) $5f \leq 2m$. Wegen der Eulerschen Polyederformel ist $f = m - n + 2$, Einsetzen liefert die Behauptung.

Der Petersen-Graph besitzt offenbar keinen Kreis mit kleinerer Länge als 5. Er besitzt $n = 10$ Ecken und $m = 15$ Kanten. Es ist $15 > \frac{5}{3}(10 - 2)$, der erste Teil der Aufgabe liefert daher, dass der Petersen-Graph nicht planar ist.

Drei Farben genügen zum Färben der 10 Knoten des Petersen-Graphen. Mögliche Gruppen gleich gefärbter Ecken sind z. B. $\{1, 3, 9, 10\}$, $\{2, 4, 6\}$ und $\{5, 7, 8\}$. Andererseits ist das Färben der Ecken mit nur zwei Farben nicht möglich. Daher ist 3 die chromatische Zahl des Petersen-Graphen.

4. Sei $G = (V, E)$ ein Graph mit $d := \max_{x \in V} d(x)$, wobei $d(x)$ den Grad der Ecke $x \in V$ bedeutet. Dann ist die chromatische Zahl von $G$ kleiner oder gleich $d + 1$. D. h. es ist eine Färbung der Ecken von $G$ durch höchstens $d + 1$ Farben derart möglich, dass benachbarte Ecken verschieden gefärbt sind.

   Hinweis: Man mache einen Induktionsbeweis nach der Anzahl $n$ der Ecken von $G$ und schließe wie beim Beweis des Sechsfarbensatzes.

   **Lösung:** Genau wie beim Beweis des Sechs- und des Fünffarbensatzes ist der Induktionsanfang trivial. Wir nehmen an, die Aussage sei für Graphen mit weniger als $n$ Ecken richtig. Für den Induktionsschluss sei $G$ ein Graph mit $n$ Ecken und $v$ eine Ecke mit dem Grad $d$, also mit $d$ Nachbarn. Man gewinne den Graphen $H$ aus $G$, indem man die Ecke $v$ und die Kanten zu den $d$ Nachbarn weglässt. Da $H$ weniger als $n$ Kanten besitzt und der Grad jeder Ecke in $H$ kleiner oder gleich $d$ ist, ist $H$ (bzw. die Ecken von $H$) nach Voraussetzung durch $d + 1$ Farben färbbar. Die Ecke $v$ färbe man nun mit einer Farbe, die noch nicht von den $d$ Nachbarn benutzt wurde. Dadurch ist die gewünschte Eckenfärbung von $G$ mit höchstens $d + 1$ Farben erreicht.

5. Zehn Personen $A, B, \ldots, I, J$ sind in acht Kommisionen $1, \ldots, 8$ vertreten. In der folgenden Tabelle 7.1 wird die Zusammensetzung der Kommissionen angegeben:Z.B. hat

| | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* | *I* | *J* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | * | * | * | * | | | | | | |
| 2 | * | | * | * | * | | | | | |
| 3 | | * | | * | | * | * | | | |
| 4 | | | * | | | * | * | * | | |
| 5 | * | | | | | | | * | | * |
| 6 | | | | | | | | * | * | * |
| 7 | | | | | | | * | * | | * |
| 8 | | | | | * | | | | * | |

Tabelle 7.1: Zusammensetzung der Kommissionen

   die Kommision 4 die Mitglieder $C, F, G, H$. Zwei Kommissionen können nicht am selben Tag tagen, wenn sie ein gemeinsames Mitglied haben. Man bestimme die minimale Zahl von Tagen, in denen sich alle Kommissionen zu einer Sitzung treffen können.

   **Lösung:** Die acht Kommisionen seien die Ecken in einem Graphen. Zwei Ecken werden durch eine Kante verbunden, wenn die entsprechenden Kommissionen ein gemeinsames

Mitglied haben. Die chromatische Zahl des so definierten Graphen ist die minimale Zahl erforderlicher Sitzungstage. In Abbildung 7.20 geben wir den so erhaltenen Gra-

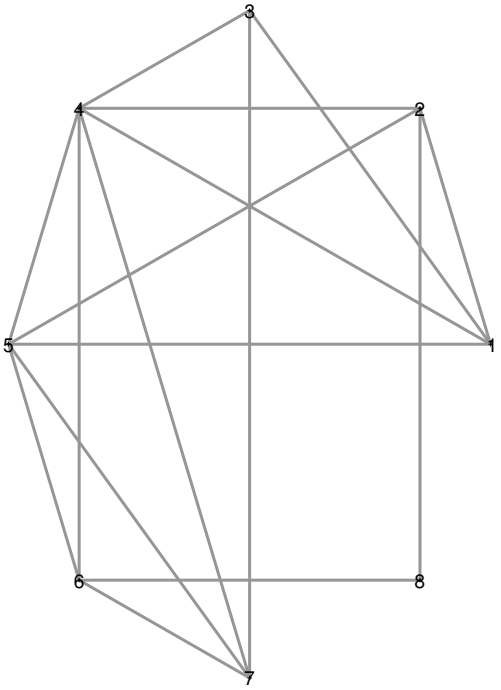


Abbildung 7.20: Tagungen von 8 Kommissionen

phen wieder. Die vier Ecken $1, 2, 3, 4$ bilden einen $K_4$, daher sind für eine Eckenfärbung mindestens vier Farben nötig. Andererseits genügen vier Sitzungstage, wenn die Kommissionen $\{1, 7, 8\}$, $\{3, 5\}$, $\{2, 6\}$ und $\{4\}$ jeweils an einem Tag tagen.

6. Bisher hatten wir uns nur mit dem *Eckenfärbungsproblem* für einen Graphen $G = (V, E)$ beschäftigt. Die *chromatische Zahl* $\chi(G)$ ist die minimale Anzahl der Farben, um die Ecken von $G$ so zu färben, dass benachbarte Ecken unterschiedlich gefärbt sind. Entsprechend kann man auch das *Kantenfärbungsproblem* für einen Graphen $G = (V, E)$ betrachten. Der *chromatische Index* $\chi'(G)$ ist die minimale Anzahl von Farben, um die Kanten von $G$ so zu färben, dass Kanten, die eine gemeinsame Ecke haben, unterschiedlich gefärbt sind.

   Man zeige, dass $\chi'(K_4) = 3$ und $\chi'(K_5) = 5$. Hierbei bezeichne $K_4$ bzw. $K_5$ den vollständigen Graphen mit 4 bzw. 5 Ecken.

   **Lösung:** Es ist $\chi'(K_4) \geq 3$, da jede Ecke von $K_4$ den Grad 3 hat (es würde natürlich genügen, wenn es nur eine Ecke vom Grad 3 gibt). Andererseits genügen drei Farben, siehe Abbildung 7.21 links. Es ist trivialerweise $\chi'(K_5) \geq 4$. Eine Kantenfärbung des $K_5$ mit 4 Farben ist aber nicht möglich, d. h. es ist $\chi'(K_5) \geq 5$. Denn der $K_5$ hat 10 Kanten. Könnte man diese Kanten durch vier Farben färben, so müsste es wenigstens drei Kanten mit der selben Farbe geben. Da es im $K_5$ nur fünf Ecken gibt, müssen zwei dieser drei Kanten eine Ecke gemein haben, was ein Widerspruch ist. Andererseits ist eine Färbung mit 5 Farben möglich, siehe Abbildung 7.21 rechts. Folglich ist $\chi'(K_5) = 5$.

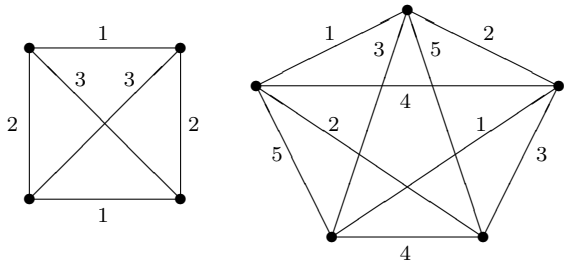


Abbildung 7.21: Der chromatische Index vollständiger Graphen

# 7.4 Aufgaben zu Kapitel 4

## 7.4.1 Aufgaben zu Abschnitt 4.1

1. Gegeben seien $t, b \in \mathbb{R}^m$ ($m$ Beobachtungen $b_i$ zu Zeiten $t_i$, $i = 1, \ldots, m$). Zur Bestimmung der *Regressionsgeraden* hat man die Aufgabe

$$\text{(P)} \qquad \text{Minimiere} \quad f(x_1, x_2) := \frac{1}{2}\sum_{i=1}^{m}(x_1 + x_2 t_i - b_i)^2, \quad (x_1, x_2) \in \mathbb{R} \times \mathbb{R},$$

zu lösen. Man gebe eine explizite Darstellung für die Lösung.

**Lösung:** Die Lösung ist dadurch charakterisiert, dass der Gradient von $f$ in ihr verschwindet. Nun ist

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \sum_{i=1}^{m}(x_1 + x_2 t_i - b_i) = x_1 m + x_2 \sum_{i=1}^{m} t_i - \sum_{i=1}^{m} b_i$$

und

$$\frac{\partial f}{\partial x_2}(x_1, x_2) = \sum_{i=1}^{m}(x_1 + x_2 t_i - b_i)t_i = x_1 \sum_{i=1}^{m} t_i + x_2 \sum_{i=1}^{m} t_i^2 - \sum_{i=1}^{m} t_i b_i.$$

Daher sind die optimalen Parameter aus dem linearen Gleichungssystem

$$\begin{pmatrix} m & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & \sum_{i=1}^m t_i^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m b_i \\ \sum_{i=1}^m t_i b_i \end{pmatrix}$$

zu bestimmen. Als Lösung erhält man

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{m\sum_{i=1}^m t_i^2 - (\sum_{i=1}^m t_i)^2} \begin{pmatrix} \sum_{i=1}^m t_i^2 & -\sum_{i=1}^m t_i \\ -\sum_{i=1}^m t_i & m \end{pmatrix} \begin{pmatrix} \sum_{i=1}^m b_i \\ \sum_{i=1}^m t_i b_i \end{pmatrix}.$$

Mit

$$\bar{t} := \frac{1}{m}\sum_{i=1}^{m} t_i, \qquad \bar{b} := \frac{1}{m}\sum_{i=1}^{m} b_i$$

ist daher (nach leichter Rechnung)

$$x_1 = \frac{\bar{b}\sum_{i=1}^m t_i^2 - \bar{t}\sum_{i=1}^m t_i b_i}{\sum_{i=1}^m (t_i - \bar{t})^2}, \qquad x_2 = \frac{\sum_{i=1}^m (t_i - \bar{t})(b_i - \bar{b})}{\sum_{i=1}^m (t_i - \bar{t})^2}.$$

Damit ist die gesuchte explizite Darstellung gefunden.

2. In der folgenden Tabelle gibt $t$ die Länge eines Säuglings bei der Geburt und $b$ die Schwangerschaftsdauer an:

| $t$ [cm] | 48 | 49 | 50 | 51 | 52 |
|---|---|---|---|---|---|
| $b$ [Tage] | 277.1 | 279.3 | 281.4 | 283.2 | 284.8 |

Hierbei kann man sich vorstellen, dass die Daten in den fünf Gruppen schon Mittelwerte aus zahlreichen weiteren Messungen sind. Es wird ein linearer Zusammenhang zwischen der Länge bei der Geburt und der Schwangerschaftsdauer vermutet. Man bestimme mit der Methode der kleinsten Quadrate die beiden Parameter bzw. löse das entsprechende lineare Ausgleichsproblem. Hierbei kann Maple (oder ein anderes mathematisches Anwendersystem) oder Aufgabe 1 benutzt werden.

**Lösung:** Zu lösen ist ein lineares Ausgleichsproblem mit den Daten

$$A := \begin{pmatrix} 1 & 48 \\ 1 & 49 \\ 1 & 50 \\ 1 & 51 \\ 1 & 52 \end{pmatrix}, \qquad b := \begin{pmatrix} 277.1 \\ 279.3 \\ 281.4 \\ 283.2 \\ 284.8 \end{pmatrix}.$$

Zur Lösung benutzen wir Maple:

```
> with(LinearAlgebra):
> A:=Matrix([[1,48],[1,49],[1,50],[1,51],[1,52]]):
> b:=<277.1,279.3,281.4,283.2,284.8>:
> x:=LeastSquares(A,b);
```

$$x := \begin{bmatrix} 184.659999999998746 \\ 1.93000000000002414 \end{bmatrix}$$

Offensichtlich sind wieder Rundungsfehler aufgetreten. Wendet man MATLAB an und schreibt nach dem MATLAB-prompt:

```
A=[1,48;1,49;1,50;1,51;1,52];
b=[277.1;279.3;281.4;283.2;284.8];
format long g
x=A\b
```

so erhält man:

$$x = \begin{pmatrix} 184.660000000001 \\ 1.92999999999999 \end{pmatrix}.$$

In der folgenden Abbildung 7.22 wird das Ergebnis visualisiert.

3. Gegeben sei ein Approximationsproblem mit den Daten $(X, \|\cdot\|)$, $M \subset X$ und $z \in X$. Man zeige:

   (a) Ist $M \subset X$ konvex, so ist die Menge der besten Approximierenden an $z$ bezüglich $M$ ebenfalls eine konvexe Menge.

   (b) Ist $M \subset X$ konvex und die Norm $\|\cdot\|$ *strikt*, d. h. gilt die Implikation

   $$\|x + y\| = \|x\| + \|y\| \Longrightarrow \; x \text{ und } y \text{ sind linear abhängig},$$

   so existiert höchstens eine beste Approximierende an $z$ bezüglich $M$.

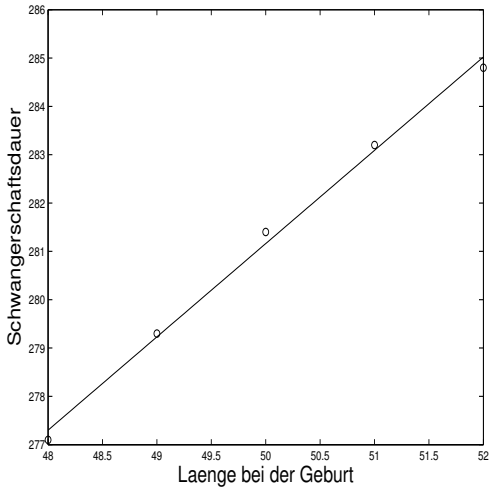


Abbildung 7.22: Länge bei der Geburt und Schwangerschaftsdauer

(c) Ist $\|\cdot\|$ durch ein inneres Produkt $(\cdot,\cdot)$ erzeugt, so ist die Norm $\|\cdot\|$ strikt. Ferner gilt die sogenannte *Parallelogrammgleichung*, d. h. für alle $x, y \in X$ ist

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

**Lösung:** Seien $x_1^*, x_2^* \in M$ beste Approximierende an $z$ in $M$, also

$$\|x_1^* - z\| = \|x_2^* - z\| = \inf_{x\in M} \|x - z\|.$$

Ferner sei $\lambda \in [0,1]$. Wegen der Konvexität von $M$ ist $x^* := (1-\lambda)x_1^* + \lambda x_2^* \in M$. Dann ist

$$\begin{aligned} \|x^* - z\| &= \|(1-\lambda)(x_1^* - z) + \lambda(x_2^* - z)\| \\ &\leq (1-\lambda)\|x_1^* - z\| + \lambda\|x_2^* - z\| \\ &= \inf_{x\in M} \|x - z\|. \end{aligned}$$

Daher ist auch $x^*$ eine beste Approximierende an $z$ bezüglich $M$, womit der erste Teil der Aufgabe gelöst ist.

Nun seien $x_1^*, x_2^* \in M$ zwei beste Approximierende an $z$ bezüglich $M$. Wegen des ersten Teiles dieser Aufgabe ist auch $\frac{1}{2}(x_1^* + x_2^*)$ eine beste Approximierende. Daher ist

$$\|\tfrac{1}{2}(x_1^* + x_2^*) - z\| = \tfrac{1}{2}\|x_1^* - z\| + \tfrac{1}{2}\|x_2^* - z\|$$

bzw.

$$(*) \qquad \|(x_1^* - z) + (x_2^* - z)\| = \|x_1^* - z\| + \|x_2^* - z\|.$$

O. B. d. A. können wir annehmen, dass $z \notin M$ (andernfalls ist $z$ selber die eindeutige beste Approximierende). Da die Norm $\|\cdot\|$ nach Voraussetzung strikt ist existiert $\alpha \neq 0$ mit $x_1^* - z = \alpha(x_2^* - z)$. Einsetzen in $(*)$ liefert $|1+\alpha| = 1 + |\alpha|$, woraus $\alpha > 0$ folgt. Daher ist

$$\|x_1^* - z\| = \alpha\,\|x_2^* - z\| = \alpha\,\|x_1^* - z\|$$

und folglich $\alpha = 1$, womit wir schließlich mit $x_1^* = x_2^*$ die Eindeutigkeit einer Lösung erhalten.

Sei nun die Norm $\|\cdot\|$ durch ein inneres Produkt $(\cdot,\cdot)$ erzeugt und $\|x+y\| = \|x\| + \|y\|$. Wir haben zu zeigen, dass $x$ und $y$ linear abhängig sind, wobei wir natürlich $y \neq 0$ annehmen können. Durch Quadrieren der letzten Gleichung erhält man $(x,y) = \|x\|\,\|y\|$, d. h. Gleichheit in der Dreiecksungleichung zieht die Gleichheit bei der Cauchy-Schwarzschen Ungleichung nach sich. Dann ist aber

$$\begin{aligned}\left\| x - \frac{\|x\|}{\|y\|} y \right\|^2 &= \|x\|^2 - 2\frac{\|x\|}{\|y\|}(x,y) + \left(\frac{\|x\|}{\|y\|}\right)\|y\|^2 \\ &= \|x\|^2 - 2\frac{\|x\|}{\|y\|}\|x\|\,\|y\| + \|x\|^2 \\ &= 0,\end{aligned}$$

d. h. es ist $x = (\|x\|/\|y\|)y$, womit gezeigt ist, dass $x$ und $y$ linear abhängig sind. Der Nachweis der Parallegrammgleichung ist trivial, da sich auf der linken Seite die gemischten Terme weghenen und nur je zweimal die quadratischen Terme übrig bleiben.

4. Man bestimme alle Lösungen des linearen Ausgleichsproblems zu den Daten

$$A := \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad b := \begin{pmatrix} 2 \\ 0 \\ 0 \\ -1 \end{pmatrix}.$$

Hinweis: Man wende den ersten Teil von Satz 1.5 an.

**Lösung:** Es ist

$$A^T A = \begin{pmatrix} 4 & 4 \\ 4 & 4 \end{pmatrix}$$

mit den Eigenwerten $\lambda_1 := 8$, $\lambda_2 := 0$. Daher sind $\sigma_1 := 2\sqrt{2}$, $\sigma_2 := 0$ die singulären Werte von $A$. Die Spalten von

$$V := \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} =: \begin{pmatrix} v_1 & v_2 \end{pmatrix}$$

bilden ein Orthonormalsystem von Eigenvektoren zu $A^T A$. Wir definieren nun (siehe den Beweis zur Existenz der Singuläwertzerlegung)

$$u_1 := \frac{1}{\sigma_1} A v_1 = \frac{1}{2\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Die allgemeine Lösung des zu $(A,b)$ gehörenden linearen Ausgleichsproblems ist gegeben durch

$$x = \frac{u_1^T b}{\sigma_1} v_1 + \alpha v_2 = \frac{1}{8} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

mit beliebigem $\alpha \in \mathbb{R}$.

5. Sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n$ gegeben, es sei $r := \text{Rang}\,(A)$. Sei $A = U\Sigma V^T$ eine Singulärwertzerlegung von $A$ (also $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ orthogonal, $\Sigma \in \mathbb{R}^{m\times n}$ eine Diagonalmatrix mit den singulären Werten $\sigma_1 \geq \cdots \geq \sigma_n$ in der Diagonalen). Man definiere $A^+ := V\Sigma^+U^T$, wobei $\Sigma_+ \in \mathbb{R}^{n\times m}$ gegeben ist durch $\Sigma^+ := (\ \hat{\Sigma}^+ \quad 0\ )$ mit

$$\hat{\Sigma}^+ := \text{diag}\,(1/\sigma_1, \ldots, 1/\sigma_r, \underbrace{0, \ldots, 0}_{n-r}).$$

Man zeige:

(a) Für jedes $b \in \mathbb{R}^m$ ist $A^+b$ die eindeutige Lösung minimaler euklidischer Norm des zu den Daten $(A, b)$ gehörenden linearen Ausgleichsproblems. Insbesondere zeigt dies die Wohldefiniertheit der Pseudoinversen.

(b) Ist $m \geq n$ und Rang $(A) = n$, so ist $A^+ = (A^TA)^{-1}A^T$.

(c) Ist $A \in \mathbb{R}^{n\times n}$ nichtsingulär, so ist $A^+ = A$.

**Lösung:** Mit $u_i$, $v_i$ seien die $i$-te Spalte von $U$ bzw. $V$ bezeichnet. Für ein beliebiges $b \in \mathbb{R}^m$ ist

$$A^+b = V\Sigma^+U^Tb = V\begin{pmatrix} u_1^T/\sigma_1 \\ \vdots \\ u_r^Tb/\sigma_r \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{i=1}^{r} \frac{u_i^Tb}{\sigma_i} v_i.$$

Nach dem zweiten Teil von Satz 1.5 ist der rechtstehende Vektor die eindeutige Lösung minimaler euklidischer Norm des zu den Daten $(A, b)$ gehörenden linearen Ausgleichsproblems.

Den zweiten Teil der Aufgabe könnte man durch Einsetzen der Singulärwertzerlegung

$$A = U\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$$

in $(A^TA)^{-1}A^T$ unter Benutzung von $\hat{\Sigma}^+ = \Sigma^{-1}$ (wegen Rang $(A) = n$) lösen:

$$\begin{aligned} (A^TA)^{-1}A^T &= \left[V(\ \hat{\Sigma} \quad 0\ )U^TU\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T\right]^{-1} V(\ \hat{\Sigma} \quad 0\ )U^T \\ &= (V\hat{\Sigma}^2V^T)^{-1}V(\ \hat{\Sigma} \quad 0\ )U^T \\ &= V(\ \hat{\Sigma}^{-1} \quad 0\ )U^T \\ &= A^+. \end{aligned}$$

Etwas einfacher ist der Beweis, wenn man beachtet, dass $A^+b$ die (wegen Rang $(A) = n$) eindeutige Lösung der Bormalgleichungen $A^TAx = A^Tb$ ist.

Der dritte Teil der Aufgabe folgt z. B. sofort aus dem zweiten.

6. Sei $z \in C^2[a,b]$ eine Funktion, deren zweite Ableitung auf dem Intervall $[a,b]$ nichtnegativ oder nichtpositiv ist, die also konvex oder konkav ist. Wegen des Mittelwertsatzes existiert ein $t_1 \in (a,b)$ mit $z(b) - z(a) = z'(t_1)(b-a)$. Man zeige, dass

$$x^*(t) := \frac{z(b)-z(a)}{b-a}\left(t - \frac{a+t_1}{2}\right) + \frac{1}{2}[z(a)+z(t_1)]$$

die beste Tschebyscheff-Approximierende an $z$ bezüglich $\mathcal{P}_1$ ist.

**Lösung:** Wir definieren $d(t) := x^*(t) - z(t)$. Nach Konstruktion ist $d'(t_1) = 0$, weiter ist $d$ auf $[a,b]$ konvex oder konkav. Daher nimmt $d$ oder $-d$ in $t_1$ ein Minimum auf $[a,b]$ an. Weiter ist $d(a) = -d(t_1) = d(b)$. Aus dem Alternantensatz (hier brauchen wir nur die einfache Richtung bzw. die Bemerkung im Anschluss an den Satz von de la Vallée-Poussin) folgt die Behauptung.

7. Bei gegebener nichtnegativer ganzer Zahl $n$ heißt die durch $T_n(t) := \cos(n \arccos t)$ definierte Funktion $T_n\colon [-1,1] \longrightarrow \mathbb{R}$ *Tschebyscheff-Polynom (erster Art) vom Grad* $n$. Man beweise die folgenden Aussagen, von denen durch die erste überhaupt erst nachgewiesen wird, dass es sich um Polynome handelt.

   (a) Es gilt die Rekursionsformel

$$T_0(t) = 1, \quad T_1(t) = 1, \quad T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t) \quad (n = 2,3,\ldots).$$

   (b) Es ist $T_n \in \mathcal{P}_n$ und $T_n(t) = 2^{n-1}t^n + p(t)$ mit $p \in \mathcal{P}_{n-2}$.

   (c) Die Nullstellen von $T_n$ sind

$$s_j := \cos\left(\frac{2(n-j)+1}{2n}\,\pi\right), \qquad j = 1,\ldots,n.$$

   (d) Es ist $\|T_n\|_\infty = 1$ und $T_n(t_i) = (-1)^{n-i}$ mit

$$t_i := \cos\left(\frac{n-i}{n}\,\pi\right), \qquad i = 0,\ldots,n.$$

   (e) Es ist $T_n(-t) = (-1)^n T_n(t)$.

   (f) Es gilt

$$\frac{2}{\pi}\int_{-1}^{1} \frac{T_m(t)T_n(t)}{\sqrt{1-t^2}}\,dt = \begin{cases} 2 & \text{für} \quad m = n = 0, \\ 1 & \text{für} \quad m = n \neq 0, \\ 0 & \text{für} \quad m \neq n. \end{cases}$$

   Hinweis: Mache die Variablentransformation $t = \cos\phi$.

   (g) Es ist

$$T_n(t) = \frac{1}{2}\left[(t + \sqrt{t^2-1})^n + (t - \sqrt{t^2-1})^n\right].$$

   (h) Sei $n \in \mathbb{N}$ und $p^* \in \mathcal{P}_{n-1}$ die beste Tschebyscheff-Approximierende an $z(t) := t^n$ auf dem Intervall $[a,b] := [-1,1]$. Dann ist $t^n - p^*(t) = 2^{-n+1}T_n(t)$ und daher $d(z, \mathcal{P}_{n-1}) = 2^{-n+1}$ der Abstand von $z$ zu $\mathcal{P}_{n-1}$.

**Lösung:** Offenbar ist $T_0(t) = 1$ und $T_1(t) = 1$. Wegen des Additionstheorems für den Cosinus:

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

ist

$$\cos(\alpha+\beta) + \cos(\alpha-\beta) = 2\cos\alpha\cos\beta.$$

Setzt man hier $\alpha := n\phi$, $\beta := \phi$, so folgt

$$\cos(n+1)\phi + \cos(n-1)\phi = 2\cos n\phi\cos\phi.$$

Mit $\phi := \arccos t$ erhält man die in (a) behauptete Rekursionsformel. Hieraus folgt auch sofort (b). Die Aussagen (c)–(e) prüft man durch einfaches Nachrechnen nach. Macht man in (f) die Variablentransformation $t = \cos\phi$, so erhält man wegen

$$dt = -\sin\phi\, d\phi = -\sqrt{1-t^2}\, d\phi,$$

dass

$$\begin{aligned}\frac{2}{\pi}\int_{-1}^{1}\frac{T_m(t)T_n(t)}{\sqrt{1-t^2}}\,dt &= -\frac{2}{\pi}\int_{\pi}^{0}\cos m\phi\cos n\phi\, d\phi\\ &= \frac{1}{\pi}\int_{-\pi}^{\pi}\cos m\phi\cos n\phi\, d\phi\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}[\cos(m+n)\phi + \cos(m-n)\phi]\, d\phi.\end{aligned}$$

Nun ist

$$\int_{-\pi}^{\pi}\cos k\phi\, d\phi = 0 \qquad \text{für alle } k \in \mathbb{Z}\setminus\{0\}.$$

Insgesamt folgt dann auch (f). Zum Beweis von (g) beachte man, dass

$$\cos n\phi \pm i\,\sin\phi = \exp(\pm in\phi) = \exp(\pm i\phi)^n = (\cos\phi \pm i\,\sin\phi)^n,$$

daher

$$\cos n\phi = \frac{1}{2}\left[(\cos\phi + i\sin\phi)^n + (\cos\phi - i\sin\phi)^n\right].$$

Mit $t = \cos\phi$ und $i\sin\phi = \sqrt{t^2-1}$ erhält man (g).

Wir wenden die hinreichende Optimalitätsbedingung im Alternantensatz 1.7 (also die einfache Richtung, die wir durch den Satz von dela Vallée-Poussin bzw. die anschließende Bemerkung bewiesen haben) an und zeigen hiermit, dass $p^* \in \mathcal{P}_{n-1}$, definiert durch $p^*(t) := t^n - 2^{-n+1}T_n(t)$ beste Tschebyscheff-Approximierende an $z(t) := t^n$ in $\mathcal{P}_{n-1}$ ist. Hierzu definiere man

$$t_i := \cos\left(\frac{n-i}{n}\,\pi\right), \qquad i = 0,\ldots,n.$$

Dann ist $-1 = t_0 < \cdots < t_n = 1$. Weiter ist

$$|p^*(t_i) - z(t_i)| = 2^{-n+1}|T_n(t_i)| = 2^{-n+1} = \|p^* - z\|_\infty,$$

d. h. in den Punkten $t_i$ nimmt $p^* - z$ sein betragsmäßiges Maximum an. Dies geschieht sogar alternierend, wie man aus

$$p^*(t_i) - z(t_i) = -2^{n-i}(-1)^{n-i}, \qquad i = 0,\ldots,n,$$

erkennt. Aus der einfachen Richtung des Alternantensatzes folgt die Behauptung.

8. Sei $(X, (\cdot,\cdot))$ ein Prä-Hilbertraum, also $(\cdot,\cdot)$ ein inneres Produkt und $\|cdot\|$ die durch $\|x\| := (x,x)^{1/2}$ induzierte Norm. Sei ferner $M \subset X$ nichtleer, konvex und $z \in X$. Dann gilt:

   (a) Es ist $x^* \in M$ genau dann beste Approximierende an $z$ bezüglich $M$, wenn $(x - x^*, z - x^*) \leq 0$ für alle $x \in M$.

   (b) Ist $(X, (\cdot,\cdot))$ sogar ein *Hilbertraum*, ist also jede Cauchy-Folge in $(X, \|\cdot\|)$ konvergent, und ist $M$ zusätlich abgeschlossen, so existiert (genau eine: das folgt schon aus den Aussagen von Aufgabe 3) eine eindeutige beste Approximierende an $z$ bezüglich $M$.

**Lösung:** Wir definieren die Abbildung $f : X \longrightarrow \mathbb{R}$ durch

$$f(x) := \frac{1}{2}\|x - z\|^2.$$

Zum Beweis des ersten Teils der Aufgabe nehmen wir zunächst an, $x^* \in M$ sei beste Approximende an $z$ bezüglich $M$. Sei weiter $x \in M$ beliebig vorgegeben. Für alle $t \in (0,1]$ ist dann $x^* + t(x - x^*) \in M$ und daher

$$\begin{aligned} 0 &\leq \frac{f(x^* + t(x - x^*)) - f(x^*)}{t} \\ &= \frac{1}{2t}[\|x^* - z + t(x - x^*)\|^2 - \|x^* - z\|^2] \\ &= \frac{1}{2t}[2t(x - x^*, x^* - z) + t^2\|x - x^*\|^2] \\ &= -(x - x^*, z - x^*) + \frac{t}{2}\|x - x^*\|^2. \end{aligned}$$

Mit $t \to 0$ folgt $(x - x^*, z - x^*) \leq 0$. Ist umgekehrt $(x - x^*, z - x^*) \leq 0$ für alle $x \in M$, so erhält man für beliebiges $x \in M$ die folgende Gleichungs-Ungleichungskette

$$\begin{aligned} f(x) - f(x^*) &= \frac{1}{2}\|x - z\|^2 - \frac{1}{2}\|x^* - z\|^2 \\ &= \frac{1}{2}\|x - x^* + x^* - z\|^2 - \frac{1}{2}\|x^* - z\|^2 \\ &= \frac{1}{2}\|x - x^*\|^2 + \underbrace{(x - x^*, x^* - z)}_{\geq 0} \\ &\geq 0, \end{aligned}$$

dass also $x^*$ beste Approximierende an $z$ bezüglich $M$ ist.

Im zweiten Teil der Aufgabe wird zusätzlich angenommen, $(X, (\cdot,\cdot))$ sei ein Hilbertraum und $M \subset X$ sei abgeschlossen (was ja wegen des zweiten Teiles von Satz 1.1 eine notwendige Bedingung an die Existenz einer besten Approximierenen für jedes $z \in X$ ist). Zu zeigen bleibt die *Existenz* eine besten Approximierenden, wobei wir $z \notin M$ annehmen können. Sei $\{x_k\} \subset M$ eine *Minimalfolge*, d. h. es gelte

$$\lim_{k\to\infty} \|x_k - z\| = d(z, M) := \inf_{x\in M} \|x - z\|.$$

Wir zeigen, dass $\{x_k\}$ eine Cauchyfolge und daher wegen der vorausgesetzten Vollständigkeit von $(X, (\cdot,\cdot))$ konvergent ist. Es ist (Parallelogrammgleichung)

$$\begin{aligned}\|x_k - x_l\|^2 + 4\underbrace{\left\|\frac{x_k + x_l}{2} - z\right\|^2}_{\geq d(z,M)^2} &= \|(x_k - z) + (z - x_l)\|^2 + \|(x_k - z) - (z - x_l)\|^2 \\ &= 2(\|x_k - z\|^2 + \|x_l - z\|^2).\end{aligned}$$

Daher ist

$$\|x_k - x_l\|^2 \leq 2(\|x_k - z\|^2 + \|x_l - z\|^2) - 4d(z, M)^2 \to 4d(z, M)^2 - 4d(z, M)^2 = 0.$$

Also existiert $x^* \in X$ mit $\lim_{k\to\infty} x_k = x^*$. Da $\{x_k\} \subset M$ und $M$ abgeschlossen, ist $x^* \in M$. Wegen der Stetigkeit der Norm ist $\|x^* - z\| = d(z, M)$ und das bedeutet, dass $x^*$ beste Approximierende an $z$ bezüglich $M$ ist.

### 7.4.2 Aufgaben zu Abschnitt 4.2

1. Man beweise den dritten Teil des starken Dualitätssatzes 2.4, also: Gegeben seien die lineare Optimierungsaufgabe

   (P) $\qquad$ Minimiere $\quad c^T x \quad$ auf $\quad M := \{x \in \mathbb{R}^n : x \geq 0,\ Ax = b\}$

   und die hierzu duale lineare Optimierungsaufgabe

   (D) $\qquad$ Maximiere $\quad b^T y \quad$ auf $\quad N := \{y \in \mathbb{R}^m : A^T y \leq c\}.$

   Man zeige: Ist $M = \emptyset$ und $N \neq \emptyset$, so ist $\sup_{y\in N} b^T y = +\infty$, die Zielfunktion von (D) ist also auf der Menge $N$ der zulässigen Lösungen von (D) nicht nach oben beschränkt.

   **Lösung:** Ist $M = \emptyset$, so liefert das Farkas-Lemma die Existenz eines $q$ mit $A^T q \leq 0$, $b^T q > 0$. Mit einem beliebigen $y \in N$ ist daher $y+tq \in N$ für alle $t \geq 0$ und $b^T(y+tq) \to +\infty$ mit $t \to \infty$. Damit ist die Behauptung bewiesen.

2. Die[6] Spieler P und D haben je 3 Karten auf der Hand, und zwar P die Karten Pik As, Karo As und Pik Zwei, D die Karten Pik As, Karo As und Karo Zwei. Beide Spieler legen jeweils zugleich eine ihrer Karten auf den Tisch. D gewinnt, wenn die hingelegten Karten die gleiche Farbe haben, andernfalls P. Ein As hat den Wert 1, eine Zwei den Wert 2. Die Höhe des Gewinnes ist gleich dem Wert derjenigen Karte, die der Gewinner hingelegt hat. Das Spiel hat also die Auszahlungsmatrix

| D \ P | ♢ | ♠ | ♠♠ |
|---|---|---|---|
| ♢ | 1 | −1 | −2 |
| ♠ | −1 | 1 | 1 |
| ♢♢ | 2 | −1 | −2 |

   Man hat den Eindruck, das Spiel sei unfair, weil die Auszahlungsmatrix 5 negative Elemente gegenüber 4 positiven enthält. Das gibt Anlass zur Formulierung der

[6] Siehe

L. Collatz, W. Wetterling (1971) *Optimierungsaufgaben*. Springer-Verlag, Berlin-Heidelberg-New York.

*Zusatzregel:* Wenn beide Spieler ihre Zweierkarte hinlegen, so soll keiner an den anderen etwas zahlen, d. h. das Element $-2$ in der rechten unteren Ecke der Auszahlungsmatrix wird durch 0 ersetzt.

Man berechne für das Spiel ohne und mit Zusatzregel mit Hilfe von Maple jeweils optimale gemischte Strategien für P und D und entscheide damit, welches der beiden Spiele fair ist.

**Lösung:** Für das Spiel ohne Zusatzregel erhalten wir durch

```
> with(simplex):
> pziel:=alpha:
> prestr:={x1>=0,x2>=0,x3>=0,-x1+x2+2*x3+alpha>=0,
> x1-x2-x3+alpha>=0,-2*x1+x2+2*x3+alpha>=0,x1+x2+x3=1}:
> minimize(pziel,prestr);
```

$$\{\alpha = 0,\ x2 = 0,\ x3 = \frac{1}{2},\ x1 = \frac{1}{2}\}$$

```
> dziel:=beta:
> drestr:={y1>=0,y2>=0,y3>=0,-y1+y2-2*y3+beta<=0,y1-y2+y3+beta<=0,
> 2*y1-y2+2*y3+beta<=0,y1+y2+y3=1}:


> maximize(dziel,drestr);
```

$$\{\beta = 0,\ y1 = 0,\ y3 = \frac{1}{3},\ y2 = \frac{2}{3}\}$$

die optimale Strategie $x^* = (\frac{1}{2}, 0, \frac{1}{2})^T$ mit Wert $\alpha^* = 0$ für P und für die die optimale Strategie $y^* = (0, \frac{2}{3}, \frac{1}{3})^T$ mit Wert $\beta^* = 0$. Das Spiel ohne Zusatzregel ist also fair. Dagegen ist für das Spiel mit Zusatzregel die optimale Strategie für P durch $x^* = (\frac{2}{5}, \frac{3}{5}, 0)^T$ mit dem Wert $\alpha^* = \frac{1}{5}$ gegeben, die für D ist $y^* = (0, \frac{3}{5}, \frac{2}{5})^T$. Es stellt sich also heraus, dass nur das Spiel ohne Zusatzregel fair ist.

3. Seien $\alpha$, $\beta$ und $\gamma$ Winkel in einem (spitzwinkligen) Dreieck mit $90^\circ \geq \alpha \geq \beta \geq \gamma \geq 0$ und natürlich $\alpha + \beta + \gamma = 180^\circ$. Unter allen solchen Winkeln bestimme man diejenigen, für die

$$g(\alpha, \beta, \gamma) := \min\{\gamma, \beta - \gamma, \alpha - \beta, 90^\circ - \alpha\}$$

maximal ist. Hierzu formuliere man diese Aufgabe als ein lineares Programm und löse es mit Hilfe eines mathematischen Anwendersystems.

Hinweis: Die obige Aufgabenstellung steht im engen Zusammenhang mit einem (sehr witzigen) Aufsatz von B. Tergan (1980)[7], in dem gezeigt wird, dass es (bis auf Ähnlichkeit) genau ein allgemeines, spitzwinkliges Dreieck gibt, dessen Winkel durch $\alpha^* := 75^\circ$, $\beta^* := 60^\circ$ und $\gamma^* := 45^\circ$ gegeben sind.

**Lösung:** Als äquivalente Aufgabe betrachten wir das Problem, die Zahl $\delta$ unter den Nebenbedingungen $\delta \leq g(\alpha, \beta, \gamma)$ sowie $90^\circ \geq \alpha \geq \beta \geq \gamma \geq 0$ und $\alpha + \beta + \gamma = 180^\circ$ zu

[7] Siehe den Anhang 2 bei

F. Wille (1982) *Humor in der Mathematik*. Vandenhoek & Ruprecht, Göttingen.

maximieren. Dies ergibt die lineare Optimierungsaufgabe

$$\begin{gathered}\text{Maximiere} \quad \delta \quad \text{unter den Nebenbedingungen} \\ \gamma \geq \delta, \quad \beta - \gamma \geq \delta, \quad \alpha - \beta \geq \delta, \quad 90 - \alpha \geq \delta, \\ 90 \geq \alpha, \quad \alpha \geq \beta, \quad \beta \geq \gamma, \quad \gamma \geq 0, \\ \alpha + \beta + \gamma = 180.\end{gathered}$$

Wir benutzen Maple, um diese Aufgabe zu lösen:

```
> with(simplex):
> ziel:=delta:
> restr:=
> {gam>=delta,beta-gam>=delta,alpha-beta>=delta,90-alpha>=delta,
> 90>=alpha,alpha>=beta,beta>=gam,gam>=0,alpha+beta+gam=180}:
> maximize(ziel,restr);
```

$$\{gam = 45,\ \delta = 15,\ \beta = 60,\ \alpha = 75\}$$

Das gesuchte Dreieck hat also in der Tat die Winkel $(\alpha^*, \beta^*, \gamma^*) = (75, 60, 45)$.

4. In einer Molkerei[8] werden zwei Sorten Käse hergestellt, etwa Gouda und Edamer. Die Fabrik hat Verträge, bis zu bestimmten Daten eine gewisse Menge (gemessen in einer bestimmten Einheit) von Käse mindestens herzustellen, nämlich

| Zeitpunkt | Gouda | Edamer |
|---|---|---|
| 30. Juni | 5 000 | 3 000 |
| 31. Juli | 6 000 | 3 000 |
| 31. August | 4 000 | 5 000 |

Zur Produktion stehen zwei Typen von Maschinen zur Verfügung. Die Anzahl der zur Verfügung stehenden Produktionsstunden für die beiden Maschinen während der Sommermonate sind:

| Monat | Maschine A | Maschine B |
|---|---|---|
| Juni | 700 | 1 500 |
| Juli | 300 | 400 |
| August | 1 000 | 300 |

Die Produktionsraten (Stunden pro Mengeneinheit Käse) auf den beiden Typen von Maschinen sind

| Typ | Maschine A | Maschine B |
|---|---|---|
| Gouda | 0.15 | 0.16 |
| Edamer | 0.12 | 0.14 |

[8]Die Aufgabe ist im wesentlichen

M. Asghar Bhatti (2000) *Practical Optimization Methods. With Mathematica Applications*. Springer-Verlag, New York-Berlin-Heidelberg

entnommen. Hier handelt es sich allerdings um eine Reifenfabrik, in der Sommer- und Winterreifen produziert werden. Da die Produktion eines Bruchteils eines Reifens keinen Sinn macht, handelt es sich bei dem dort geschilderten Problem aber um eine *ganzzahlige* lineare Optimierungsaufgabe. Um dies zu vermeiden (es kommen nämlich nicht ganzzahlige Werte heraus) haben wir die Aufgabenstellung ein wenig verändert. Inwiefern diese Aufgabenstellung sinnvoll ist, sei dahin gestellt. Es kommt letzten Endes darauf an, das mathematische Modell aufzustellen.

Unabhängig von den benutzten Typen und dem produzierten Käse kostet eine Arbeitsstunde 100 Euro. Das Material für eine Mengeneinheit Gouda kostet 52.50 Euro, das für Edamer 41.50 Euro. Pro Mengeneinheit Käse kommen noch 4 Euro hinzu. Überschüssiger Käse kann in den nächsten Monat (also von Juni in den Juli und von Juli in den August) übernommen werden, die Lagerkosten sind 1.50 Euro pro Mengeneinheit Käse. Eine Mengeneinheit des produzierten Käses wird für 200 Euro (Gouda) bzw. 150 Euro (Edamer) verkauft. Wie sollte die Produktion organisiert werden, um einerseits den Lieferbedingungen nachzukommen und andererseits den Gewinn der Molkerei zu maximieren?

Hinweis: Als Variable führen man ein:

| | |
|---|---|
| $x_1$ | Menge des im Juni auf Machine A produzierten Gouda |
| $x_2$ | Menge des im Juli auf Maschine A produzierten Gouda |
| $x_3$ | Menge des im August auf Maschine A produzierten Gouda |
| $x_4$ | Menge des im Juni auf Machine A produzierten Edamer |
| $x_5$ | Menge des im Juli auf Machine A produzierten Edamer |
| $x_6$ | Menge des im August auf Machine A produzierten Edamer |
| $x_7$ | Menge des im Juni auf Machine B produzierten Gouda |
| $x_8$ | Menge des im Juli auf Maschine B produzierten Gouda |
| $x_9$ | Menge des im August auf Maschine B produzierten Gouda |
| $x_{10}$ | Menge des im Juni auf Machine B produzierten Edamer |
| $x_{11}$ | Menge des im Juli auf Machine B produzierten Edamer |
| $x_{12}$ | Menge des im August auf Machine B produzierten Edamer |

**Lösung:** Zunächst stellen wir die Zielfunktion auf. Durch den Verkauf erhält man

$$v = 200(x_1 + x_2 + x_3 + x_7 + x_8 + x_9) + 150(x_4 + x_5 + x_6 + x_{10} + x_{11} + x_{12})$$

Euro. Die Materialkosten betragen

$$m = 52.50(x_1 + x_2 + x_3 + x_7 + x_8 + x_9) + 41.50(x_4 + x_5 + x_6 + x_{10} + x_{11} + x_{12})$$

Euro. Arbeitskosten sind

$$a = 100[0.15(x_1+x_2+x_3)+0.16(x_7+x_8+x_9)+0.12(x_4+x_5+x_6)+0.14(x_{10}+x_{11}+x_{12})]$$

Euro. Zusätzliche Kosten sind

$$z = 4(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12})$$

Euro, Lagerkosten

$$\begin{aligned} l \;=\; & 1.5[(x_1 + x_7 - 5000) + (x_4 + x_{10} - 3000) \\ & + (x_1 + x_2 + x_7 + x_8 - 11000) + (x_4 + x_5 + x_{10} + x_{11} - 6000)] \end{aligned}$$

Euro. Der Gesamtgewinn ist $v - (m + a + z + l)$, dieser ist zu maximieren. Die Produktionsbeschränkungen führen auf die Restriktionen

$$0.15x_1 + 0.12x_4 \leq 700, \quad 0.15x_2 + 0.12x_5 \leq 300, \quad 0.15x_3 + 0.12x_6 \leq 1000$$

und

$$0.16x_7 + 0.14x_{10} \leq 1500, \quad 0.16x_8 + 0.14x_{11} \leq 400, \quad 0.16x_9 + 0.14x_{12} \leq 300.$$

Wegen der eingegangenen Verträge kommen die weiteren Restriktionen

$$\begin{aligned} x_1 + x_7 &\geq 5000, \\ x_4 + x_{10} &\geq 3000, \\ x_1 + x_2 + x_7 + x_8 &\geq 11000, \\ x_4 + x_5 + x_{10} + x_{11} &\geq 6000, \\ x_1 + x_2 + x_3 + x_7 + x_8 + x_9 &\geq 15000, \\ x_4 + x_5 + x_6 + x_{10} + x_{11} + x_{12} &\geq 11000 \end{aligned}$$

hinzu, weiter sind natürlich alle Variablen nichtnegativ. Mit Maple kann dieses lineare Programm folgendermaßen gelöst werden:

```
> with(simplex):
> v:=200*(x1+x2+x3+x7+x8+x9)+150*(x4+x5+x6+x10+x11+x12):
> m:=52.5*(x1+x2+x3+x7+x8+x9)+41.5*(x4+x5+x6+x10+x11+x12):
> a:=100*(0.15*(x1+x2+x3)+0.16*(x7+x8+x9)+0.12*(x4+x5+x6)+0.14*(x10+x11
> +x12)):
> z:=4*(x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12):
> l:=1.5*((x1+x7-5000)+(x4+x10-3000)+(x1+x2+x7+x8-5000)+(x4+x5+x10+x11-
> 6000)):
> gewinn:=v-(m+a+z+l):
> nebenbed:=
> {0.15*x1+0.12*x4<=700,0.15*x2+0.12*x5<=300,0.15*x3+0.12*x6<=1000,
> 0.16*x7+0.14*x10<=1500,0.16*x8+0.14*x11<=400,0.16*x9+0.14*x12<=300,
> x1+x7>=5000,x4+x10>=300,x1+x2+x7+x8>=11000,x4+x5+x10+x11>=6000,
> x1+x2+x3+x7+x8+x9>=15000,x4+x5+x6+x10+x11+x12>=11000}:
> loesung:=maximize(gewinn,nebenbed,NONNEGATIVE);
```

$loesung := \{x10 = 0.,\ x11 = 0.,\ x1 = 1866.666667,\ x4 = 3500.000000,\ x8 = 2500.000000,$
$x3 = 2666.666667,\ x2 = 0.,\ x7 = 9375.000000,\ x12 = 0.,\ x9 = 1875.000000,$
$x5 = 2500.000000,\ x6 = 5000.000000\}$

```
> opt:=subs(loesung,gewinn);
```

$$opt := .3329933333\,10^7$$

Wir erhalten also

$$x^* = (1866.67, 0, 2666.67, 3500, 2500, 5000, 9375, 2500, 1875, 0, 0, 0)^T$$

als Lösung. Nun wollen wir auch noch das Ergebnis für den Fall bestimmen, dass zum Schluss der Produktionszeit (also Ende August) kein Überschuss vorhanden ist. Die letzten beiden Restriktionen sind dann zu ersetzen durch

$$x_1 + x_2 + x_3 + x_7 + x_8 + x_9 = 15000, \qquad x_4 + x_5 + x_6 + x_{10} + x_{11} + x_{12} = 11000.$$

Als Lösung mit Maple erhalten wir

$$x^* = (1866.67, 0, 2666.67, 3500, 2500, 5000, 6333.33, 2500, 1333.33, 0, 0, 0)^T.$$

Dies stimmt genau mit dem Ergebnis bei M. Asghar Bhatti (2000, S. 417) überein.

5. Man beweise den ersten Teil des Max-Flow Min-Cut Theorems von Ford-Fulkerson, also: Gegeben sei ein Digraph $(\mathcal{N}, \mathcal{A})$, in dem zwei Knoten $s$ (Quelle) und $t$ (Senke) ausgezeichnet sind. Auf den Pfeilen seien nichtnegative Kapazitäten gegeben. Ist dann $x = (x_{ij})_{(i,j)\in\mathcal{A}}$ ein zulässiger Fluss mit dem Wert $v = \sum_{j:(s,j)\in\mathcal{A}} x_{sj}$ und ist $(\mathcal{N}_1, \mathcal{N}_2)$ ein Schnitt mit Kapazität $C(\mathcal{N}_1, \mathcal{N}_2)$, so ist $v \leq C(\mathcal{N}_1, \mathcal{N}_2)$.

   **Lösung:** Es ist

$$
\begin{aligned}
v &= \sum_{j:(s,j)\in\mathcal{A}} x_{sj} - \underbrace{\sum_{i:(i,s)\in\mathcal{A}} x_{is}}_{=0} + \sum_{k\in\mathcal{N}_1\setminus\{s\}} \underbrace{\left( \sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} \right)}_{=0} \\
&= \sum_{k\in\mathcal{N}_1} \left( \sum_{j:(k,j)\in\mathcal{A}} x_{kj} - \sum_{i:(i,k)\in\mathcal{A}} x_{ik} \right) \\
&= \sum_{k\in\mathcal{N}_1} \left( \sum_{\substack{j\in\mathcal{N}_2 \\ (k,j)\in\mathcal{A}}} x_{kj} - \sum_{\substack{i\in\mathcal{N}_2 \\ (i,k)\in\mathcal{A}}} x_{ik} \right) + \underbrace{\sum_{k\in\mathcal{N}_1} \left( \sum_{\substack{j\in\mathcal{N}_1 \\ (k,j)\in\mathcal{A}}} x_{kj} - \sum_{\substack{i\in\mathcal{N}_1 \\ (i,k)\in\mathcal{A}}} x_{ik} \right)}_{=0} \\
&\leq \sum_{k\in\mathcal{N}_1} \sum_{\substack{j\in\mathcal{N}_2 \\ (k,j)\in\mathcal{A}}} x_{kj} \\
&= \sum_{i\in\mathcal{N}_1} \sum_{\substack{j\in\mathcal{N}_2 \\ (i,j)\in\mathcal{A}}} x_{ij} \\
&\leq \sum_{\substack{(i,j)\in\mathcal{A} \\ i\in\mathcal{N}_1,\, j\in\mathcal{N}_2}} u_{ij} \\
&= C(\mathcal{N}_1, \mathcal{N}_2).
\end{aligned}
$$

   Damit ist die Aufgabe gelöst.

6. Eine Gruppe von 11 Personen trifft sich in San Francisco. Möglichst viele von ihnen sollen nach New York geschickt werden. Es gibt keine Direktflüge, sondern es muss umgestiegen werden, wobei der Anschluss jeweils gesichert ist. In der folgenden Tabelle sind diese Flüge und die jeweils noch vorhandenen freien Sitze aufgelistet.

| Von | Nach | Zahl freier Sitze |
|---|---|---|
| San Francisco | Denver | 5 |
| San Francisco | Houston | 6 |
| Denver | Atlanta | 4 |
| Denver | Chicago | 2 |
| Houston | Atlanta | 5 |
| Atlanta | New York | 7 |
| Chicago | New York | 4 |

   (a) Man formuliere die Aufgabe als Maximaler-Fluss-Problem in einem geeigneten Digraphen.

   (b) Man rate einen maximalen Fluss und beweise seine Optimalität mit dem Max-Flow Min-Cut Theorem.

**Lösung:** Durch die 6 Städte und die 7 Verbindungen ist ein Digraph gegeben, den wir in Abbildung 7.23 darstellen. Die Kapazitäten bzw. die Anzahl freier Sitze haben wir

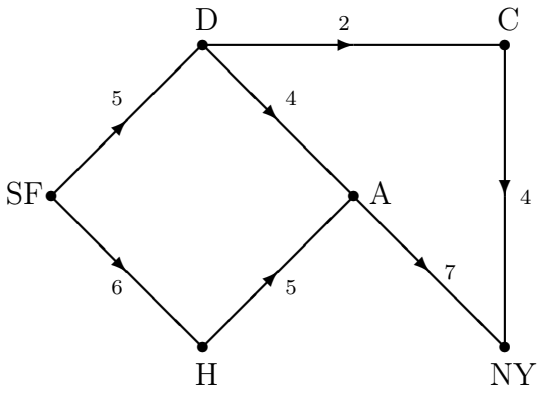


Abbildung 7.23: Flug von San Francisco nach New York

an die jeweiligen Pfeile geschrieben. Man hat also in natürlicher Weise ein Maximaler-Fluss-Problem mit San Francisco als Quelle und New York als Senke zu lösen.

In Abbildung 7.24 geben wir einen zulässigen Fluss mit dem Wert $v = 9$ an, weiter einen

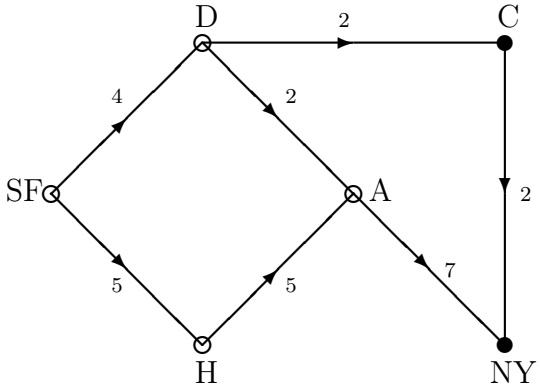


Abbildung 7.24: Optimaler Fluss von San Francisco nach New York

durch $\circ$ bzw. $\bullet$ gekennzeichneten Schnitt der Knoten, der die Kapazität 9 besitzt. Wegen des Max Flow-Min Cut Theorems hat man auch eine Lösung des Maximaler-Fluss- bzw. Minimaler-Schnitt-Problems gefunden.

7. Für die Aufgabe

$$\text{(P)} \qquad \begin{cases} \text{Minimiere} \quad f(x) := x_1^2 + 4x_2^2 + 16x_3^2 \quad \text{unter der Nebenbedingung} \\ \qquad\qquad\qquad\qquad h(x) := x_1 x_2 x_3 - 1 = 0 \end{cases}$$

bestimme man mit Hilfe von Maple alle zulässigen Punkte, in denen die notwendigen Optimalitätsbedingungen erster Ordnung erfüllt sind.

**Lösung:** Ein Punkt $x^* = (x_1^*, x_2^*, x_3^*)$ ist zulässig für (P) und genügt den notwendigen Optimalitätsbedingungen erster Ordnung, wenn $h(x^*) = 0$ und ein $v^* \in \mathbb{R}$ mit $\nabla f(x^*) +$

$v^*\nabla h(x^*) = 0$ existiert. Die notwendigen Optimalitätsbedingungen sind in $x^*$ erfüllt, wenn ein $v^* \in \mathbb{R}$ mit

$$\begin{pmatrix} 2x_1^* \\ 8x_2^* \\ 32x_3^* \end{pmatrix} + v^* \begin{pmatrix} x_2^* x_3^* \\ x_1^* x_3^* \\ x_1^* x_2^* \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

$$x_1^* x_2^* x_3^* - 1 = 0$$

existiert. Wir benutzen einmal wieder Maple zur Lösung dieses nichtlinearen Gleichungssystems und erhalten als reelle Lösungen

$$x^{(1)} := \begin{pmatrix} 2 \\ 1 \\ \frac{1}{2} \end{pmatrix}, \qquad x^{(2)} := \begin{pmatrix} -2 \\ -1 \\ \frac{1}{2} \end{pmatrix}, \qquad x^{(3)} := \begin{pmatrix} 2 \\ -1 \\ -\frac{1}{2} \end{pmatrix}, \qquad x^{(4)} := \begin{pmatrix} -2 \\ 1 \\ -\frac{1}{2} \end{pmatrix},$$

der zugehörige Multiplikator ist jeweils $v^* = -8$.

8. Bei einer *ganzzahligen linearen Optimierungsaufgabe* handelt es sich um eine lineare Optimierungsaufgabe, bei der die Variablen ganzzahlig sind.

   In der $x_1$-$x_2$-Ebene veranschauliche man sich die folgende ganzzahlige lineare Optimierungsaufgabe

$$\begin{cases} \text{Minimiere} \quad \begin{pmatrix} -2 \\ 1 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{unter den Nebenbedingungen} \\ \begin{pmatrix} 5 & 7 \\ -2 & 1 \\ 1 & -5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \le \begin{pmatrix} 45 \\ 1 \\ 5 \end{pmatrix}, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \ge \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad x_1, x_2 \in \mathbb{Z} \end{cases}$$

   und gebe die Lösung an. Zum Vergleich bestimme man die Lösung des *relaxierten* Problems, also des Problems, bei dem die Ganzzahligkeitsforderung gestrichen wird.

   **Lösung:** In Abbildung 7.25 stellen wir die Menge der zulässigen Lösungen (markiert durch •) und die Zielfunktion dar. Die gesuchte Lösung ist also $x^* = (7, 1)$. Mit Hilfe von Maple erhalten wir die Lösung des relaxierten Problems:

```
> with(simplex):
> ziel:=-2*x_1+x_2:
> restr:={5*x_1+7*x_2<=45,-2*x_1+x_2<=1,x_1-5*x_2<=5}:
> minimize(ziel,restr,NONNEGATIVE);
```

$$\{x\_1 = \frac{65}{8},\ x\_2 = \frac{5}{8}\}$$

# 7.5 Aufgaben zu Kapitel 5

## 7.5.1 Aufgaben zu Abschnitt 5.1

1. Sei[9] $p$ die Lösung der Anfangswertaufgabe für die logistische Differentialgleichung

$$p' = ap - bp^2, \qquad p(t_0) = p_0,$$

[9]Diese Aufgabe findet man auf S. 88 von

M. Braun (1983) "Single species population models". In: *Differential Equation Models* (eds. M. Braun et al.), Springer-Verlag, New York-Heidelberg-Berlin.

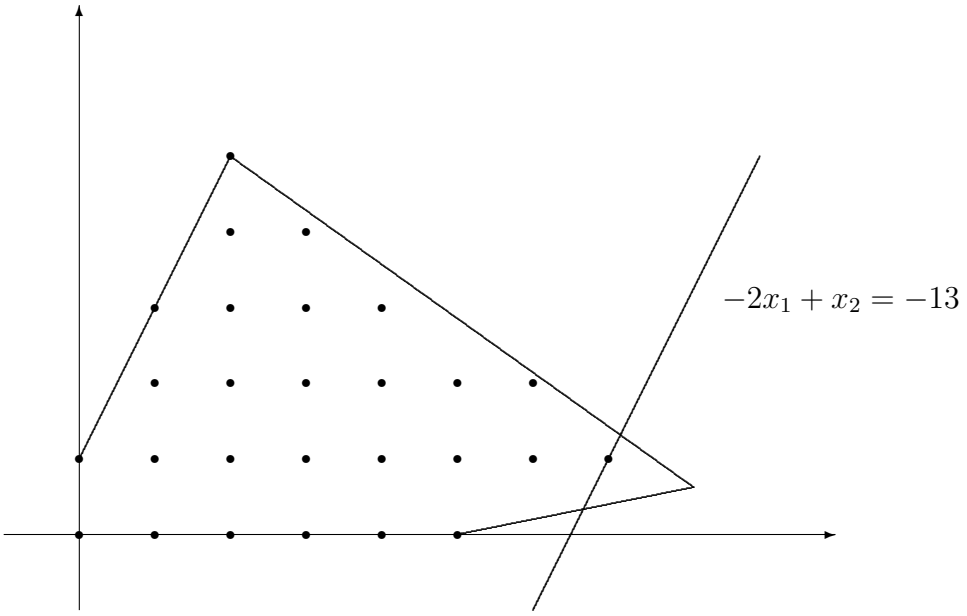


Abbildung 7.25: Ein ganzzahliges Optimierungsproblem

wobei $a, b, p_0$ positive Konstanten mit $p_0 < \frac{1}{2}(a/b)$ sind.

(a) Seien $t_1 < t_2$ mit $t_1 > t_0$ und $t_1 - t_0 = t_2 - t_1$ gegeben. Man zeige, dass $a$ und $b$ eindeutig durch $p_0 = p(t_0), p(t_1), p(t_2)$ bestimmt sind. Dies bedeutet: Legt man das logistische Wachstumsmodell zugrunde und sind die Populationen $p_0, p_1, p_2$ zu äquidistanten Zeiten $t_0, t_1, t_2$ bekannt, so sind hierdurch die Parameter $a, b$ im Modell eindeutig festgelegt.

(b) Man zeige, dass genau ein $t^* > t_0$ mit $p(t^*) = \frac{1}{2}(a/b)$ existiert und die Darstellung

$$p(t) = \frac{a/b}{1 + e^{-a(t-t^*)}}$$

gilt.

(c) Aus

| $k$ | $t_k$ | $p(t_k)$ |
|---|---|---|
| 0 | 1790 | 3 929 000 |
| 1 | 1850 | 23 192 000 |
| 2 | 1910 | 91 972 000 |

bestimme man $a$ und $b$. Anschließend berechne man $t^*$ mit $p(t^*) = \frac{1}{2}(a/b)$.

Hinweis: Es darf Maple eingesetzt werden.

**Lösung:** Zur Abkürzung setzen wir $p_1 := p(t_1)$ und $p_2 := p(t_2)$. Es ist

$$p(t) = \frac{ap_i}{bp_i + (a - bp_i)\exp[-a(t - t_i)]}, \qquad i = 0, 1, 2,$$

da dies die Lösung der logistischen Differentialgleichung mit der Anfangsbedingung $p(t_i) = p_i$, $i = 0, 1, 2$, ist. Aus den Gleichungen

$$p_1[bp_0 + (a - bp_0)\exp(-a(t_1 - t_0))] = ap_0$$

und

$$p_2[bp_1 + (a - bp_1)\exp(-a(t_2 - t_1))] = ap_1$$

erhält man unter Berücksichtigung von $t_1 - t_0 = t_2 - t_1$, dass

$$\frac{p_0(a - bp_1)}{p_1(a - bp_0)} = \frac{p_1(a - bp_2)}{p_2(a - bp_1)}.$$

Hieraus kann man wenigstens in eindeutiger Weise das Verhältnis $\xi = a/b$ berechnen, denn es ist

$$\xi = \frac{p_1[p_1(p_0 + p_2) - 2p_0p_2]}{p_1^2 - p_0p_2}.$$

Da wir insbesondere $p_0 < a/b$ vorausgesetzt haben, ist $p(\cdot)$ auf $(t_0, \infty)$ monoton wachsend mit $\lim_{t\to\infty} p(t) = \xi$. Folglich ist

$$\eta := \left(\frac{p_0}{p_1}\right)\frac{\xi - p_1}{\xi - p_0} \in (0, 1),$$

so dass genau ein $a > 0$ mit

$$\exp(-a(t_1 - t_0)) = \eta$$

existiert, nämlich

$$a = -\frac{\log \eta}{t_1 - t_0}.$$

Die eindeutige Existenz von $t^* > t_0$ mit $p(t^*) = \frac{1}{2}(a/b)$ ist völlig trivial. Auch die behauptete Darstellung von $p(\cdot)$ ist einfach einzusehen, denn es ist

$$\begin{aligned} p(t) &= \frac{ap(t^*)}{bp(t^*) + (a - bp(t^*))e^{-a(t-t^*)}} \\ &= \frac{a\frac{1}{2}(a/b)}{b\frac{1}{2}(a/b) + (a - b\frac{1}{2}(a/b))e^{-a(t-t^*)}} \\ &= \frac{a/b}{1 + e^{-a(t-t^*)}}. \end{aligned}$$

Im letzten Teil der Aufgabe berechnen wir zunächst $\xi = a/b$ aus

$$\xi = \frac{p_1[p_1(p_0 + p_2) - 2p_0p_2]}{p_1^2 - p_0p_2}.$$

Wir erhalten durch

```
p_0:=3929000;p_1:=23192000;p_2:=91972000;
xi:=p_1*(p_1*(p_0+p_2)-2*p_0*p_2)/(p_1^2-p_0*p_2);
```

das Resultat

$$\xi = \frac{8705233252768000}{44127719} = 1.97273583363 \cdot 10^9,$$

letzteres nach Eingabe von `evalf(xi)`. Durch

```
a:=-ln((p_0/p_1)*(xi-p_1)/(xi-p_0))/60;
```

und `a:=evalf(a);` erhalten wir $a = 0.03133953992$, dann entsprechend $b = a/\xi = 0.158863378 \cdot 10^{-9}$. Bei der Berechnung von $t^*$ mit $p(t^*) = \frac{1}{2}(a/b)$ machen wir es uns einfach. Aus

```
p:=t->a*p_0/(b*p_0+(a-b*p_0)*exp(-a*(t-1790)));
fsolve({p(t)=0.5*xi},{t});
```

erhalten wir als Lösung $t^* = 1914.318643$. Natürlich hätte man die Berechnung von $a$ und $b$ mit Maple noch einfacher bekommen können, sozusagen ohne überhaupt ein bisschen nachzudenken:

```
> eqn:={p_1*(b*p_0+(a-b*p_0)*exp(-a*T))=a*p_0,
p_2*(b*p_1+(a-b*p_1)*exp(-a*T))=a*p_1};
```

$$eqn := \{p\_1\,(b\,p\_0 + (a - b\,p\_0)\,e^{(-a\,T)}) = a\,p\_0,\ p\_2\,(b\,p\_1 + (a - b\,p\_1)\,e^{(-a\,T)}) = a\,p\_1\}$$

```
> sol:=solve(eqn,{a,b});
```

$$sol := \{b = b,\ a = 0\},$$

$$\left\{a = -\frac{\ln(-\frac{p\_0\,(p\_1 - p\_2)}{p\_2\,(p\_1 - p\_0)})}{T},\ b = -\frac{\ln(-\frac{p\_0\,(p\_1 - p\_2)}{p\_2\,(p\_1 - p\_0)})\,(-p\_0\,p\_2 + p\_1^{\,2})}{T\,(p\_2\,p\_1 - 2\,p\_0\,p\_2 + p\_1\,p\_0)\,p\_1}\right\}$$

```
> p_0:=3929000: p_1:=23192000: p_2:=91972000: T:=60:
> evalf(sol);
```

$$\{b = b,\ a = 0.\},\ \{a = .03133953992,\ b = .1588633378\,10^{-9}\}$$

2. Das vollständige elliptische Integral erster Ordnung ist mit dem Gaußschen arithmetisch-geometrischen Mittel (AGM) verwandt[10]. Insbesondere zeige man:

   (a) Gegeben seien Zahlen $a, b$ mit $0 < b \le a$. Auf die folgende Weise erzeuge man Folgen $\{a_k\}$, $\{b_k\}$.

   - Setze $a_0 := a$, $b_0 := b$.
   - Für $k = 0, 1, \ldots$:
     - Berechne $a_{k+1} := \frac{1}{2}(a_k + b_k)$.
     - Berechne $b_{k+1} := \sqrt{a_k b_k}$.

   Man zeige: Die Folgen $\{a_k\}$ und $\{b_k\}$ konvergieren monoton nicht wachsend bzw. monoton nicht fallend gegen einen gemeinsamen Grenzwert $M(a, b)$, das sogenannte arithmetisch-geometrische Mittel von $a$ und $b$.

   (b) Für $0 < b \le a$ und $\lambda > 0$ ist $M(\lambda a, \lambda b) = \lambda M(a, b)$.

   (c) Für $0 < b \le a$ ist

$$M(a, b) = M\left(\frac{a + b}{2}, \sqrt{ab}\right).$$

[10] Hierüber kann man sich sehr gut auf den ersten Seiten von

J. M. Borwein, P. B. Borwein (1987) *Pi and the AGM*. J. Wiley, New York

informieren. Zu recht bezeichnen sie das arithmetisch-geometrische Mittel als eine der Juwelen der klassischen Analysis.

(d) Für $0 < b \leq 1$ ist

$$M(1,b) = \frac{1+b}{2} M\left(1, \frac{2\sqrt{b}}{1+b}\right).$$

(e) Für $0 < x \leq 1$ ist

$$\frac{1}{M(1,x)} = \frac{2}{\pi} K(\sqrt{1-x^2}),$$

wobei mit

$$K(k) := \int_0^{\pi/2} \frac{d\theta}{\sqrt{1-k^2\sin^2\theta}}$$

das vollständige elliptische Integral erster Art bezeichnet wird[11].

(f) Man zeige, dass

$$\frac{1}{M(\sqrt{2},1)} = \frac{2}{\pi} \int_0^1 \frac{dt}{\sqrt{1-t^4}}.$$

**Lösung:** Wir nehmen an, es sei $0 < b_k \leq a_k$, was für $k = 0$ richtig ist. Wegen der Ungleichung vom geometrisch-arithmetischen Mittel ist

$$b_k \leq b_{k+1} = \sqrt{a_k b_k} \leq \frac{a_k + b_k}{2} = a_{k+1} \leq a_k.$$

Als monotone, nach unten bzw. oben beschränkte Folgen sind $\{a_k\}$ und $\{b_k\}$ konvergent gegen $a_\infty$ bzw. $b_\infty$. Z. B. wegen $a_\infty = \frac{1}{2}(a_\infty + b_\infty)$ ist $a_\infty = b_\infty$. Die Folgen $\{a_k\}$ und $\{b_k\}$ besitzen also denselben Grenzwert.

Mit $\lambda > 0$ und $0 < b \leq a$ setze man $a_0 := a$, $b_0 := b$ sowie $a_0' := \lambda a$, $b_0' := \lambda b$. Anschließend definiere man die Folgen $\{a_k\}$, $\{b_k\}$ sowie $\{a_k'\}$, $\{b_k'\}$ durch

$$a_{k+1} := \frac{a_k + b_k}{2}, \qquad b_{k+1} := \sqrt{a_k b_k}$$

bzw.

$$a_{k+1}' := \frac{a_k' + b_k'}{2}, \qquad b_{k+1}' := \sqrt{a_k' b_k'}.$$

Durch vollständige Induktion nach $k$ zeigt man, dass $a_k' = \lambda a_k$, $b_k' = \lambda b_k$. , woraus natürlich $M(\lambda a, \lambda b) = \lambda M(a,b)$ folgt.

Die Gleichung

$$M(a,b) = M\left(\frac{a+b}{2}, \sqrt{ab}\right)$$

für $0 < b \leq a$ ist völlig selbstverständlich, denn ob man im obigen Algorithmus bei $a_0, b_0$ oder $a_1, b_1$ beginnt, ist für den gemeinsamen Grenzwert irrelevant.

Für $0 < b \leq 1$ ist unter Berücksichtigung der beiden letzten Ergebnisse

$$M(1,b) = M\left(\frac{1+b}{2}, \sqrt{b}\right) = \frac{1+b}{2} M\left(1, \frac{2\sqrt{b}}{1+b}\right).$$

[11] Für einen Beweis kann man J. M. Borwein, P. B. Borwein (1987, S. 5) oder auch J. Todd (1979, S. 18) *Basic Numerical Mathematics*, vol. 1. Birkhäuser Verlag, Basel-Stuttgart konsultieren. Aber selbst dann wird der Beweis nicht ganz einfach sein

Sei $0 < b \le a$. Man definiere

$$T(a,b) := \frac{2}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{a^2\cos^2\theta + b^2\sin^2\theta}}.$$

Mit der Substitution $t := b\tan\theta$ wird

$$\cos^2\theta = \frac{b^2}{b^2+t^2}, \qquad \sin^2\theta = \frac{t^2}{b^2+t^2}, \qquad d\theta = \frac{b}{b^2+t^2}\,dt$$

und folglich

$$T(a,b) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}}.$$

Nun ist

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}} &= \int_{-\infty}^{0} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}} + \int_0^{\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}} \\
&= \int_{-\infty}^{\infty} \Bigg\{ \frac{1 - u/\sqrt{ab+u^2}}{\sqrt{[a^2+(u-\sqrt{ab+u^2})^2][b^2+(u-\sqrt{ab+u^2})^2]}} \\
&\qquad + \frac{1 + u/\sqrt{ab+u^2}}{\sqrt{[a^2+(u+\sqrt{ab+u^2})^2][b^2+(u+\sqrt{ab+u^2})^2]}} \Bigg\}\, du \\
&\qquad (t = u - \sqrt{ab+u^2} \text{ bzw. } t = u + \sqrt{ab+u^2}) \\
&= \int_{-\infty}^{\infty} \frac{1}{C(u)} \Bigg\{ \frac{C(u)-u}{\sqrt{[a^2+(C(u)-u)^2][b^2+(C(u)-u)^2]}} \\
&\qquad + \frac{C(u)+u}{\sqrt{[a^2+(C(u)+u)^2][b^2+(C(u)+u)^2]}} \Bigg\}\, du,
\end{aligned}$$

wobei wir zur Abkürzung

$$C(u) := \sqrt{ab+u^2}$$

gesetzt haben. Nun ist

$$\begin{aligned}
&\frac{C(u)-u}{\sqrt{[a^2+(C(u)-u)^2][b^2+(C(u)-u)^2]}} + \frac{C(u)+u}{\sqrt{[a^2+(C(u)+u)^2][b^2+(C(u)+u)^2]}} \\
= \;& \frac{1}{2\sqrt{[(a+b)/2]^2+u^2}} + \frac{1}{2\sqrt{[(a+b)/2]^2+u^2}} \\
= \;& \frac{1}{2\sqrt{[(a+b)/2]^2+u^2}},
\end{aligned}$$

wenn man den ersten Summanden mit $C(u)+u$ und den zweiten mit $C(u)-u$ erweitert. Insgesamt ist

$$\int_{-\infty}^{\infty} \frac{dt}{\sqrt{(a^2+t^2)(b^2+t^2)}} = \int_{-\infty}^{\infty} \frac{du}{\sqrt{\{[(a+b)/2]^2+u^2\}\{ab+u^2\}}}$$

und daher

$$T(a,b) = T\left(\frac{a+b}{2}, \sqrt{ab}\right).$$

Eine Fortsetzung liefert

$$T(a,b) = T(M(a,b), M(a,b)) = \frac{1}{M(a,b)}.$$

Für $0 < x \leq 1$ ist insbesondere

$$\begin{aligned}
\frac{1}{M(1,x)} &= T(1,x) \\
&= \frac{2}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{\cos^2\theta + x^2\sin^2\theta}} \\
&= \frac{2}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{1-(1-x^2)\sin^2\theta}} \\
&= \frac{2}{\pi}K(\sqrt{1-x^2}),
\end{aligned}$$

was zu zeigen war.

Es ist

$$\begin{aligned}
\frac{1}{M(\sqrt{2},1)} &= \frac{1}{\sqrt{2}M(1,1/\sqrt{2})} \\
&= \frac{\sqrt{2}}{\pi}K(1/\sqrt{2}) \\
&\qquad \text{(wegen des gerade eben bewiesenen Teils)} \\
&= \frac{\sqrt{2}}{\pi}\int_0^{\pi/2} \frac{d\theta}{\sqrt{1-\frac{1}{2}\sin^2\theta}} \\
&= \frac{\sqrt{2}}{\pi}\int_0^{1} \frac{dt}{\sqrt{(1-t^2)(1-\frac{1}{2}t^2)}} \\
&\qquad \text{(Substitution } t = \sin\theta\text{)} \\
&= \frac{2}{\pi}\int_0^{1} \frac{dt}{\sqrt{(1-t^2)(2-t^2)}} \\
&= \frac{2}{\pi}\int_0^{1} \frac{dx}{\sqrt{1-x^4}} \\
&\qquad \text{(Substitution } x^2 = t^2/(2-t^2)\text{)},
\end{aligned}$$

womit die Behauptung bewiesen ist.

3. Wir betrachten den Fall eines Körpers der Masse $m$, der unter dem Einfluss der Schwerkraft sich senkrecht nach unten bewegt, wobei der zur Geschwindigkeit proportionale Luftwiderstand berücksichtigt werde. Mit einer Konstanten $\rho > 0$ und (konstantem) $g = 9.81\,\mathrm{m/sec^2}$ hat man die Anfangswertaufgabe

$$m\ddot{x} = mg - \rho\dot{x}, \qquad x(0) = 0, \quad \dot{x}(0) = v_0$$

zu lösen. Man zeige, dass $\lim_{t\to\infty}\dot{x}(t)$ existiert, der Körper also eine endliche Endgeschwindigkeit erreicht[12]

[12] Bei H. Heuser (1989, S. 30) findet man hierzu die Bemerkung: Von dieser Tatsache profitiert der Fallschirmspringer immer dann, wenn sein Schirm überhaupt aufgeht.

**Lösung:** Die Geschwindigkeit $v := \dot{x}$ ist Lösung der Anfangswertaufgabe

$$m\dot{v} = mg - \rho v, \qquad v(0) = v_0.$$

Nach

```
> dsolve({m*diff(v(t),t)=m*g-rho*v(t),v(0)=v_0},v(t));
```

$$\mathrm{v}(t) = \frac{g\,m}{\rho} - \frac{e^{(-\frac{\rho\,t}{m})}\,(m\,g - v\_0\,\rho)}{\rho}$$

erkennt man, dass

$$\lim_{t\to\infty} v(t) = \frac{gm}{\rho}.$$

4. Bei W. Walter (1996, S. 5)[13] findet man ein System von zwei Differentialgleichungen zweiter Ordnung, durch das die Bewegung eines Satelliten im Gravitationsfeld zweier Körper (z. B. Erde und Mond) modelliert wird, nämlich

$$\begin{aligned}
\ddot{x} &= x + 2\dot{y} - \mu' \frac{x+\mu}{[(x+\mu)^2+y^2]^{3/2}} - \mu \frac{x-\mu'}{[(x-\mu')^2+y^2]^{3/2}},\\
\ddot{y} &= y - 2\dot{x} - \mu' \frac{y}{[(x+\mu)^2+y^2]^{3/2}} - \mu \frac{y}{[(x-\mu')^2+y^2]^{3/2}}.
\end{aligned}$$

Hierbei ist $\mu$ eine gegebene Konstante und $\mu' := 1 - \mu$. Für $\mu := 0.01213$ und die Anfangsbedingungen

$$x(0) = 1.2, \quad \dot{x}(0) = 0, \quad y(0) = 0, \quad \dot{y}(0) = -1.04936$$

plotte man die Bahn $\{(x(t), y(t)) : 0 \le t \le 10\}$.

**Lösung:** Wir benutzen die folgenden Maple-Befehle:

```
mu:=0.01213: must:=1-mu:
eqn1:=diff(x(t),t,t)=x(t)+2*diff(y(t),t)
    -must*(x(t)+mu)/((x(t)+mu)^2+y(t)^2)^(3/2)
    -mu*(x(t)-must)/((x(t)-must)^2+y(t)^2)^(3/2):
eqn2:=diff(y(t),t,t)=y(t)-2*diff(x(t),t)-must*y(t)/((x(t)+mu)^2+y(t)^2)^(3/2)
    -mu*y(t)/((x(t)-must)^2+y(t)^2)^(3/2):
initial1:=x(0)=1.2,D(x)(0)=0:
initial2:=y(0)=0,D(y)(0)=-1.04936;
sol:=dsolve({eqn1,eqn2,initial1,initial2},{x(t),y(t)},type=numeric):
with(plots):
odeplot(sol,[x(t),y(t)],0..10,numpoints=300,labels=["x","y"],
      title="Bahn eines Satelliten");
```

Das Resultat ist in Abbildung 7.26 angegeben. Erstaunlicherweise erhält man eine periodische Bahn.

[13] W. Walter (1996) *Gewöhnliche Differentialgleichungen. 6. Auflage*. Springer-Verlag, Berlin-Heidelberg-New York.

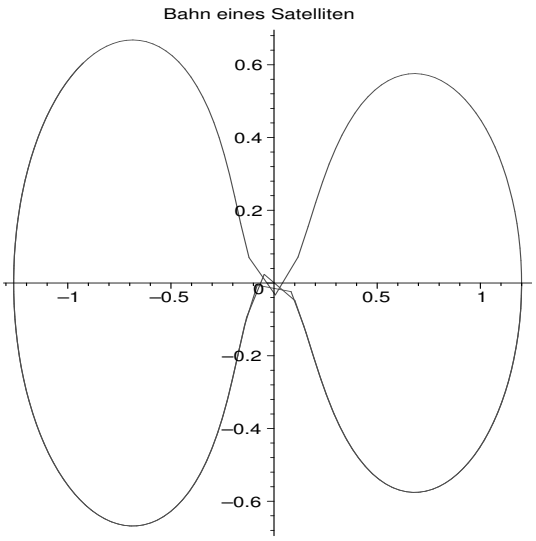


Abbildung 7.26: Bewegung im Gravitationsfeld zweier Körper

### 7.5.2 Aufgaben zu Abschnitt 5.2

1. Der lineare Raum $C_n[a,b]$ aller stetigen Abbildungen $x\colon [a,b] \longrightarrow \mathbb{R}^n$, versehen mit der Norm

$$\|x\| := \max_{t\in[a,b]} \|x(t)\|,$$

wobei auf der rechten Seite $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ ist, ist ein Banachraum.

**Lösung:** Sei $\{x_k\}$ eine Cauchyfolge in $C_n[a,b]$. Zu vorgegebenem $\epsilon > 0$ gibt es dann ein $K(\epsilon) \in \mathbb{N}$ mit

$$(*) \qquad \|x_k(t) - x_l(t)\| \le \|x_k - x_l\| \le \epsilon \qquad \text{für alle } k,l \ge K(\epsilon) \text{ und alle } t \in [a,b].$$

Für jedes $t \in [a,b]$ ist daher $\{x_k(t)\} \subset \mathbb{R}^n$ eine Cauchyfolge. Da der $\mathbb{R}^n$ versehen mit einer beliebigen Norm ein Banachraum ist, konvergiert die Folge $\{x_k(t)\}$ für jedes $t \in [a,b]$, es existiert also eine Abbildung $x\colon [a,b] \longrightarrow \mathbb{R}^n$ mit $\lim_{k\to\infty} x_k(t) = x(t)$. Aus $(*)$ folgt mit $l \to \infty$, dass

$$\|x_k(t) - x(t)\| \le \epsilon \qquad \text{für alle } k \ge K(\epsilon) \text{ und alle } t \in [a,b].$$

Folglich ist $\|x_k - x\| \le \epsilon$ für alle $k \ge K(\epsilon)$, d. h. die Folge $\{x_k\} \subset C_n[a,b]$ konvergiert gleichmäßig gegen $x$. Der gleichmäßige Limes stetiger Funktionen ist bekanntlich eine stetige Funktion, wie man leicht mit einem $\epsilon/3$-Argument nachweist. Also ist $x \in C_n[a,b]$, die Behauptung ist bewiesen.

2. Der lineare Raum $C_n^1[a,b]$ aller stetig differenzierbaren Abbildungen $x\colon [a,b] \longrightarrow \mathbb{R}^n$, versehen mit der Norm

$$\|x\| := \max\Big(\max_{t\in[a,b]} \|x(t)\|, \max_{t\in[a,b]} \|x'(t)\|\Big),$$

wobei auf der rechten Seite $\|\cdot\|$ eine beliebige Norm auf dem $\mathbb{R}^n$ ist, ist ein Banachraum.

**Lösung:** Sei $\{x_k\}$ eine Cauchyfolge in $C_n^1[a,b]$ bezüglich der angegebenen Norm. Dann sind $\{x_k\}, \{x_k'\}$ Cauchyfolgen in $C_n[a,b]$, versehen mit der Norm

$$\|x\| := \max_{t\in[a,b]} \|x(t)\|.$$

Folglich konvergieren $\{x_k\}$ und $\{x_k'\}$ gleichmäßig gegen $x \in C_n[a,b]$ bzw. $y \in C_n[a,b]$. Zunächst zeigen wir, dass

$$x(t) = x(a) + \int_a^t y(s)\,ds \qquad \text{für alle } t \in [a,b],$$

woraus $x \in C_n^1[a,b]$ und $x' = y$ folgt. Denn für beliebiges $t \in [a,b]$ ist

$$x(t) - x(a) - \int_a^t y(s)\,ds = [x(t) - x_k(t)] - [x(a) - x_k(a)] - \int_a^t [y(s) - x_k'(s)]\,ds$$

und folglich

$$\begin{aligned}
\left\| x(t) - x(a) - \int_a^t y(s)\,ds \right\| &\le \|x(t) - x_k(t)\| + \|x(a) - x_k(a)\| \\
&\quad + \left\| \int_a^t [y(s) - x_k'(s)]\,ds \right\| \\
&\le \|x(t) - x_k(t)\| + \|x(a) - x_k(a)\| \\
&\quad + \int_a^t \|y(s) - x_k'(s)\|\,ds \\
&\le \underbrace{\|x(t) - x_k(t)\|}_{\to 0} + \underbrace{\|x(a) - x_k(a)\|}_{\to 0} \\
&\quad + (t-a) \underbrace{\max_{s\in[a,b]} \|y(s) - x_k'(s)\|}_{\to 0},
\end{aligned}$$

woraus $x(t) = x(a) + \int_a^t y(s)\,ds$ für alle $t \in [a,b]$ folgt. Dann konvergiert $\{x_k\}$ bezüglich der auf $C_n^1[a,b]$ gegebenen Norm gegen $x$, so dass $C_n^1[a,b]$ bezüglich dieser Norm ein Banachraum ist.

3. Die lineare Anfangswertaufgabe erster Ordnung

$$x' = 2tx + t, \qquad x(0) = x_0$$

besitzt die Lösung

$$x(t) = x_0 e^{t^2} + \frac{1}{2}(e^{t^2} - 1),$$

wie man durch `dsolve({diff(x(t),t)=2*t*x(t)+t,x(0)=x_0},x(t));` oder eigene Rechnung feststellt. Mit $x_0(t) := x_0$ sei

$$x_{k+1}(t) := x_0 + \int_0^t (2s x_k(s) + s)\,ds.$$

Man stelle die Iterierten $x_k$ geschlossen dar und begründe, weshalb die Folge $\{x_k\}$ auf jedem kompakten Intervall in $\mathbb{R}$ gleichmäßig gegen die Lösung der gegebenen Anfangswertaufgabe konvergiert.

**Lösung:** Wir zeigen durch vollständige Induktion nach $k$, dass

$$x_k(t) = x_0\left(1 + t^2 + \frac{t^4}{2!} + \cdots + \frac{t^{2k}}{k!}\right) + \frac{1}{2}\left(t^2 + \frac{t^4}{2!} + \cdots + \frac{t^{2k}}{k!}\right).$$

Hierzu genügt es, den Induktionsschluss durchzuführen. Nun ist

$$\begin{aligned}
x_{k+1}(t) &= x_0 + \int_0^t [2sx_k(s) + s]\, ds \\
&= x_0 + \frac{t^2}{2} + 2\int_0^t \left[x_0 s\left(1 + s^2 + \frac{s^4}{2!} + \cdots + \frac{s^{2k}}{k!}\right)\right. \\
&\qquad \left. + \frac{s}{2}\left(s^2 + \frac{s^4}{2!} + \cdots + \frac{s^{2k}}{k!}\right)\right] ds \\
&= x_0 + \frac{t^2}{2} + x_0\left(t^2 + \frac{t^4}{2!} + \frac{t^6}{3!} + \cdots + \frac{t^{2(k+1)}}{(k+1)!}\right) \\
&\qquad + \frac{1}{2}\left(\frac{t^4}{2!} + \frac{t^6}{3!} + \cdots + \frac{t^{2(k+1)}}{(k+1)!}\right) \\
&= x_0\left(1 + t^2 + \frac{t^4}{2!} + \cdots + \frac{t^{2(k+1)}}{(k+1)!}\right) + \frac{1}{2}\left(t^2 + \frac{t^4}{2!} + \cdots + \frac{t^{2(k+1)}}{(k+1)!}\right),
\end{aligned}$$

womit die Behauptung bewiesen ist. Unter Beachtung der Potenzreihenentwicklung

$$e^{t^2} = \sum_{j=0}^{\infty} \frac{t^{2j}}{j!}$$

und der Tatsache, dass Partialsummen auf kompakten Teilmengen gleichmäßig gegen den Limes konvergieren, folgt der Rest der Behauptung. Natürlich hätte aber auch mit der Bemerkung im Anschluss an den Satz von Picard-Lindelöf argumentiert werden können.

4. Gegeben sei die Anfangswertaufgabe

$$x' = tx^2, \qquad x(0) = 1.$$

Mit $x_0 := 1$ und

$$x_{k+1}(t) := 1 + \int_0^t sx_k(s)^2\, ds$$

berechne man $x_1, x_2, x_3$. Man bestimme ein Intervall $[0, \alpha]$ mit $\alpha > 0$, auf dem eine Lösung eindeutig existiert und mache eine Fehlerabschätzung.

**Lösung:** Mit Hilfe von Maple erhalten wir

$$\begin{aligned}
x_1(t) &= 1 + \frac{1}{2}t^2, \\
x_2(t) &= 1 + \frac{1}{2}t^2 + \frac{1}{4}t^4 + \frac{1}{24}t^6, \\
x_3(t) &= 1 + \frac{1}{2}t^2 + \frac{1}{4}t^4 + \frac{1}{8}t^6 + \frac{1}{24}t^8 + \frac{1}{96}t^{10} + \frac{1}{576}t^{12} + \frac{1}{8064}t^{14}.
\end{aligned}$$

Für eine Fehlerabschätzung untersuchen wir, unter welchen Voraussetzungen an noch unbestimmte Parameter $\alpha >$ und $\beta \in (0,1]$ die Abbildung

$$F(x)(t) := 1 + \int_0^t sx(s)^2\,ds$$

die abgeschlossene Menge

$$K := \{x \in C[0,\alpha] : \|x - x_0\| \le \beta\}$$

kontrahierend in sich abbildet, wobei $x_0(t) := 1$ und als Norm in $C[0,\alpha]$ die (ungewichtete) Maximumnorm zugrunde gelegt sei. Seien $t \in [0,\alpha]$, $x \in K$ beliebig. Es ist

$$\begin{aligned} |F(x)(t) - x_0(t)| &= \left| \int_0^t sx(s)^2\,ds \right| \\ &\le (1+\beta)^2 \int_0^t s\,ds \\ &\le \frac{1}{2}\alpha^2(1+\beta)^2. \end{aligned}$$

Daher bildet die Abbildung $F$ die Menge $K$ in sich ab, wenn

$$\frac{1}{2}\alpha^2(1+\beta)^2 \le \beta.$$

Für beliebige $x, y \in K$ und $t \in [0,\alpha]$ ist

$$\begin{aligned} |F(x)(t) - F(y)(t)| &= \left| \int_0^t s[x(s)^2 - y(s)^2]\,ds \right| \\ &\le \int_0^t s|x(s)+y(s)|\,|x(s)-y(s)|\,ds \\ &\le 2(1+\beta)\int_0^t s\,ds\,\|x-y\| \\ &\le \alpha^2(1+\beta)\,\|x-y\|. \end{aligned}$$

Insgesamt bildet $F$ die (abgeschlossene) Menge $K$ kontrahierend in sich ab, wenn

$$\frac{1}{2}\alpha^2(1+\beta)^2 \le \beta, \qquad \alpha^2(1+\beta) < 1.$$

Setzt man z. B. $\beta := 1$, so sind beide Ungleichungen erfüllt, wenn $0 < \alpha < \sqrt{2}/2$. Die exakte Lösung der gegebenen Anfangswertaufgabe ist übrigens $x(t) = 2/(2-t^2)$. Eine Lösung existiert also nur auf $(-\sqrt{2}, \sqrt{2})$.

5. Man zeige, dass die Anfangswertaufgabe für das mathematische Pendel, also

$$x'' + \omega_0^2 \sin x = 0, \quad x(0) = x_0, \quad x'(0) = 0,$$

für beliebige $\omega_0$ und $x_0$ genau eine Lösung besitzt. Diese existiert auf ganz $\mathbb{R}$ und ist gerade, also $x(t) = x(-t)$ für alle $t$. Für $\omega_0 := 2$ und $x_0 := 1$ berechne man mit Hilfe des Gaußschen Verfahrens vom arithmetisch-geometrischen Mittel die Periodenlänge $T = (4/\omega_0)K(\sin\frac{1}{2}x_0)$. Schließlich plotte man die Lösung auf $[0, 2T]$.

**Lösung:** Schreibt man die Anfangswertaufgabe für die Differentialgleichung zweiter Ordnung als eine Anfangswertaufgabe für ein System von zwei Differentialgleichungen erster Ordnung, so erhält man

$$\begin{array}{rclcrcl} x_1' &=& x_2, & & x_1(0) &=& x_0, \\ x_2' &=& -\omega_0^2 \sin x_1, & & x_2(0) &=& 0. \end{array}$$

Die rechte Seite

$$f(t,x) := \begin{pmatrix} x_2 \\ -\omega_0^2 \sin x_1 \end{pmatrix}$$

ist global lipschitzstetig. Denn ist $\|\cdot\|$ die Maximumnorm auf dem $\mathbb{R}^2$, so ist für beliebige $x, y \in \mathbb{R}^2$ offenbar

$$\|f(t,x) - f(t,y)\| = \max(|x_2 - y_2|, \omega_0^2 |\sin x_1 - \sin y_1|) \le \max(1, \omega_0^2)\|x - y\|.$$

Aus dem Korollar 4.8 zum Satz von Picard-Lindelöf folgt die Existenz genauer einer Lösung $x$ auf $\mathbb{R}$. Um zu zeigen, dass diese gerade ist, definieren wir $y(t) := x(-t)$. Dann genügt $y$ denselben Anfangsbedingungen wie $x$, da $y(0) = x(0) = x_0$ und $y'(0) = -x'(0) = 0$, aber auch derselben Differentialgleichung, da

$$y''(t) + \omega_0^2 \sin y(t) = x''(-t) + \omega_0^2 \sin x(-t) = 0.$$

Da gerade gezeigt wurde, dass diese Anfangswertaufgabe genau eine Lösung besitzt, ist $x(t) = y(t)$ für alle $t$ bzw. $x$ gerade. Für $\omega_0 = 2$ und $x_0 = 1$ ist die Periodenlänge

$$T = \frac{4}{\omega_0} K(\sin \tfrac{1}{2} x_0) = 2K(\sin \tfrac{1}{2}) = \frac{\pi}{M(1, \cos \frac{1}{2})}.$$

Bei der Berechnung von $M(1, \cos \frac{1}{2})$ erhalten wir die folgenden Werte

| $k$ | $b_k$ | $a_k$ |
|---|---|---|
| 0 | 0.877582561890373 | 1.000000000000000 |
| 1 | 0.936793767000172 | 0.938791280945186 |
| 2 | 0.937791992130215 | 0.937792523972679 |
| 3 | 0.937792258051410 | 0.937792258051447 |
| 4 | 0.937792258051429 | 0.937792258051429 |

Daher erhalten wir als Periodenlänge

$$T = 3.34998783218523.$$

In Abbildung 7.27 findet man einen Plot über dem Intervall $[0, 2T]$.

6. Gegeben sei die Anfangswertaufgabe

(P) $$x' = t + \sin x, \qquad x(0) = 0.$$

   (a) Man zeige, dass (P) auf jedem kompakten Teilintervall $I$ von $\mathbb{R}$ mit $0 \in I$ genau eine Lösung besitzt.

   (b) Mit Hilfe von Maple-Befehlen plotte man die Lösung von (P) auf dem Intervall $[-1, 1]$.

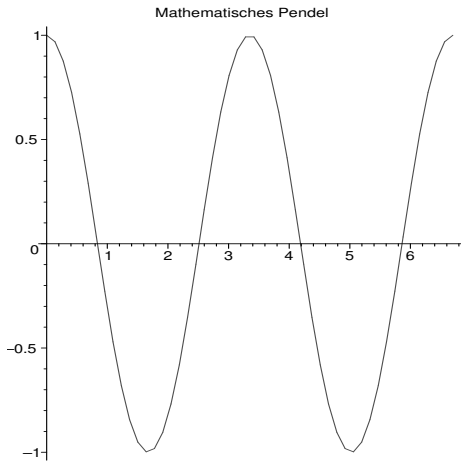


Abbildung 7.27: Die Lösung von $x'' + 4\sin x = 0,\ x(0) = 1,\ x'(0) = 0$

**Lösung:** Die rechte Seite von (P) ist in $x$ global lipschitzstetig. Denn mit $f(t,x) := t + \sin x$ ist

$$|f(t,x) - f(t,y)| \leq |x - y| \qquad \text{für alle } (t,x)\,(t,y) \in \mathbb{R}^2.$$

Aus dem Korollar 4.8 zum Satz von Picard-Lindelöf folgt die Existenz genau einer Lösung von (P) auf jedem kompakten Intervall, das 0 enthält.

Mit Hilfe von

```
kor:=dsolve({diff(x(t),t)=t+sin(x(t)),x(0)=0},x(t),type=numeric);
plots[odeplot](kor,[t,x(t)],-1..1,labels=["t","x"]);
```

erhalten wir den in Abbildung 7.28 dargestellten Plot.

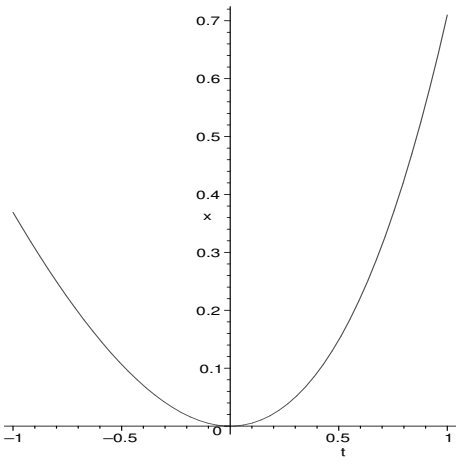


Abbildung 7.28: Lösung von $x' = t + \sin x,\ x(0) = 0$

### 7.5.3 Aufgaben zu Abschnitt 5.3

1. Für $A \in \mathbb{C}^{n\times n}$ bezeichne $e^{At}$ das durch $X(0) = I$ normierte Fundamentalsystem $X(t)$ zu $x' = Ax$. Man zeige:

   (a) Es ist

   $$\frac{d}{dt}e^{At} = Ae^{At} = e^{At}A, \qquad e^{A(t+s)} = e^{At}e^{As}, \qquad (e^{At})^{-1} = e^{(-A)t} = e^{A(-t)}$$

   für alle $t, s \in \mathbb{R}$.

   (b) Ist $C \in \mathbb{C}^{n\times n}$ nichtsingulär und $J := C^{-1}AC$, so ist $e^{At} = Ce^{Jt}C^{-1}$.

   (c) Mit $\lambda \in \mathbb{C}$ sei $J \in \mathbb{C}^{n\times n}$ definiert durch

   $$J := \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & \ddots & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & \ddots & 1 \\ 0 & 0 & \cdots & \cdots & \lambda \end{pmatrix}.$$

   Dann ist

   $$e^{Jt} = e^{\lambda t}\begin{pmatrix} 1 & \frac{t}{1!} & \frac{t^2}{2!} & \cdots & \frac{t^{n-2}}{(n-2)!} & \frac{t^{n-1}}{(n-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{n-3}}{(n-3)!} & \frac{t^{n-2}}{(n-2)!} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

   (d) Die inhomogene Anfangswertaufgabe

   $$x' = Ax + b(t), \qquad x(t_0) = x_0$$

   besitzt die eindeutige Lösung

   $$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}b(s)\,ds.$$

   (e) Sind $A, B \in \mathbb{C}^{n\times n}$ mit $AB = BA$, so ist $e^{(A+B)t} = e^{At}e^{Bt}$ für alle $t$.

   (f) Es ist

   $$e^{At} = \sum_{j=0}^{\infty} \frac{A^j t^j}{j!}.$$

   Genauer ist

   $$\lim_{k\to\infty} \sum_{j=0}^{k} \frac{A^j t^j}{j!} = e^{At} \qquad \text{für jedes } t \in \mathbb{R},$$

   wobei diese Konvergenz auf kompakten Teilmengen von $\mathbb{R}$ gleichmäßig ist.

(g) Sei

$$A := \begin{pmatrix} 0 & 1 & 0 \\ 4 & 3 & -4 \\ 1 & 2 & -1 \end{pmatrix}.$$

Man zeige, dass

$$e^{At} = \begin{pmatrix} 1 & 4 & 0 \\ 0 & 4 & 4 \\ 1 & 6 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^t & te^t \\ 0 & 0 & e^t \end{pmatrix} \begin{pmatrix} 1 & 4 & 0 \\ 0 & 4 & 4 \\ 1 & 6 & 1 \end{pmatrix}^{-1}.$$

Man vergleiche die hierdurch gewonnene Matrix $e^A$ mit $\sum_{j=0}^{10} A^j/j!$.

**Lösung:** Die erste Gleichung folgt daraus, dass $e^{At}$ nach Definition ein Fundamentalsystem von $x' = Ax$ ist. Um $Ae^{At} = e^{At}A$ zu erhalten, definieren wir $Y(t) := Ae^{At}$ und $Z(t) := e^{At}A$. Dann ist $Y(0) = Z(0) = A$ und

$$\begin{aligned} \frac{d}{dt}Y(t) &= A\frac{d}{dt}e^{At} = AAe^{At} = AY(t) \\ \frac{d}{dt}Z(t) &= \frac{d}{dt}e^{At}A = Ae^{At}A = AZ(t) \end{aligned}$$

und folglich $Y(t) = Z(t)$ wegen der Eindeutigkeit bei linearen Anfangswertaufgaben.

Die Aussage $e^{A(t+s)} = e^{At}e^{As}$ ist für $t = 0$ richtig. Da beide Seiten außerdem der Matrix-Differentialgleichung $Z' = AZ$ genügen, ist die Aussage richtig.

Es ist

$$0 = \frac{d}{dt}[(e^{At})^{-1}e^{At}] = \frac{d}{dt}[(e^{At})^{-1}]e^{At} + A$$

und damit

$$\frac{d}{dt}[(e^{At})^{-1}] = -A(e^{At})^{-1}.$$

Folglich ist $(e^{At})^{-1} = e^{(-A)t}$. Ferner ist $e^{(-A)t} = e^{A(-t)}$, da dies für $t = 0$ richtig ist und beide Seiten der Matrix-Differentialgleichung $Z' = -AZ$ genügen.

Für $t = 0$ stimmt die behauptete Gleichung $e^{At} = Ce^{Jt}C^{-1}$. Ferner ist

$$\frac{d}{dt}[Ce^{Jt}C^{-1}] = CJe^{Jt}C^{-1} = \underbrace{CJC^{-1}}_{=A}Ce^{Jt}C^{-1} = ACe^{Jt}C^{-1}$$

und daraus folgt die Behauptung.

Man definiere

$$x_j(t) := \begin{pmatrix} \frac{t^{j-1}}{(j-1)!} \\ \vdots \\ \frac{t}{1!} \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad j = 1, \ldots, n.$$

Dann ist

$$x_j'(t) = \lambda x_j(t) + x_{j-1}(t) = Jx_j(t), \qquad j = 1, \ldots, n.$$

Da außerdem $x_j(0) = e_j$ der $j$-te Einheitsvektor ist, ist die Behauptung bewiesen.

Die Lösung der inhomogenen Anfangswertaufgabe ist durch

$$x(t) = X(t)[X^{-1}(t_0)x_0 + \int_{t_0}^{t} X^{-1}(s)b(s)\,ds]$$

gegeben, wobei $X(t)$ ein beliebiges Fundamentalsystem zu $x' = Ax$ ist. Einsetzen von $X(t) = e^{At}$ und Berücksichtigung der oben nachgewiesenen Rechenregeln liefert die Behauptung.

Zunächst folgt aus der Vertauschbarkeit von $A$ und $B$, dass $e^{At}B = Be^{At}$. Dies ist nämlich für $t = 0$ richtig, ferner ist

$$\frac{d}{dt}[e^{At}B - Be^{At}] = Ae^{At}B - BAeAt = A(e^{At}B - Be^{At}),$$

woraus die Zwischenbehauptung folgt. Die Behauptung selbst sieht man ebenfalls nach vertrautem Muster ein. Sie ist für $t = 0$ richtig und es ist

$$\frac{d}{dt}[e^{At}e^{Bt}] = Ae^{At}e^{Bt} + e^{At}Be^{Bt} = (A + B)e^{At}e^{Bt},$$

woraus die Behauptung folgt.

Man gebe sich $\alpha > 0$ vor und definiere $I_\alpha := [-\alpha, \alpha]$. Ferner sei die Abbildung $F\colon C(I_\alpha; \mathbb{C}^{n\times n}) \longrightarrow C(I_\alpha; \mathbb{C}^{n\times n})$ definiert durch

$$F(X)(t) := I + \int_0^t AX(s)\,ds.$$

Offenbar ist $X(t) = e^{At}$ Fixpunkt von $F$. Auf $C(I_\alpha; \mathbb{C}^{n\times n})$ definiere man die Norm

$$\|X\| := \max_{t\in I_\alpha} e^{-2\|A\|\,|t|}\,\|X(t)\|,$$

wodurch $C(I_\alpha; \mathbb{C}^{n\times n})$ zu einem Banachraum wird. Hierbei ist rechts $\|\cdot\|$ eine Matrixnorm auf $\mathbb{C}^{n\times n}$. Wie wir schon früher in einer ganz ähnlichen Situation beim Beweis des Satzes von Picard-Lindelöf ausgerechnet haben, kontrahiert die Abbildung $F$ auf $C(I_\alpha; \mathbb{C}^{n\times n})$ mit einer Lipschitzkonstanten $q := \frac{1}{2}$. Mit

$$X_0(t) := I, \qquad X_{k+1}(t) := F(X_k)(t), \qquad k = 0, 1, \ldots$$

ist

$$X_k(t) = \sum_{j=0}^{k} \frac{A^j t^j}{j!}.$$

Der Fixpunktsatz für kontrahierende Abbildungen liefert die Abschätzung

$$\|X - X_k\| \le \frac{q^k}{1-q}\,\|X_1 - X_0\|$$

bzw.

$$e^{-2\|A\|\alpha}\left\|e^{At}-\sum_{j=0}^{k}\frac{A^jt^j}{j!}\right\|\le\left(\frac{1}{2}\right)^{k-1}\|A\|\alpha \qquad \text{für alle } t\in I_\alpha.$$

Folglich ist

$$\left\|e^{At}-\sum_{j=0}^{k}\frac{A^jt^j}{j!}\right\|\le\|A\|\alpha e^{2\|A\|\alpha}\left(\frac{1}{2}\right)^{k-1} \qquad \text{für alle } t\in I_\alpha$$

und hieraus folgt die Behauptung.

Es ist

$$J:=C^{-1}AC=\begin{pmatrix}0&0&0\\0&1&1\\0&0&1\end{pmatrix} \qquad \text{mit} \quad C:=\begin{pmatrix}1&4&0\\0&4&4\\1&6&1\end{pmatrix}$$

die Jordansche Normalform von $A$ (in Maple kann die Dunktion `JordanForm` benutzt werden, die Matrix $C$ ist dann aber etwas anders). Wegen

$$e^{Jt}=\begin{pmatrix}1&0&0\\0&e^t&te^t\\0&0&e^t\end{pmatrix}$$

folgt die Behauptung. Direktes Ausrechnen liefert

$$\begin{aligned}e^A &= Ce^JC^{-1}\\ &= \begin{pmatrix}5&1+e&-4\\4e&3e&-4e\\5+e&1+2e&-4-e\end{pmatrix}\\ &= \begin{pmatrix}5.00000000000&3.71828182846&-4.00000000000\\10.87312731380&8.15484548538&-10.87312731380\\7.71828182846&6.43656365692&-6.71828182846\end{pmatrix}.\end{aligned}$$

Die Partialsumme $\sum_{j=0}^{10}A^j/j!$ kann man in Maple folgendermaßen erhalten:

```
> with(LinearAlgebra):
> A:=Matrix([[0,1,0],[4,3,-4],[1,2,-1]]):
> s:=IdentityMatrix(3): summand:=IdentityMatrix(3):
> for j from 1 to 10 do
> summand:=summand.A/j:
> s:=s+summand:
> end do:
> evalf(s);
```

$$\begin{bmatrix}4.99999889771&3.71828125000&-3.99999889771\\10.8731261023&8.15484485229&-10.8731261023\\7.71828014771&6.43656277557&-6.71828014771\end{bmatrix}$$

Noch einfacher kann man $e^A$ in *Mathematica* durch die Funktion `MatrixExp` berechnen, mit MATLAB durch `expm`. In Maple gibt es (im eigentlich veralteten `linalg`-package die Funktion `exponential`. Hier kann man $e^A$ berechnen durch:

```
> with(linalg):

Warning, the protected names norm and trace have been redefined and
unprotected
```

```
> A:=Matrix([[0,1,0],[4,3,-4],[1,2,-1]]):
> evalf(exponential(A,1));
```

$$\begin{bmatrix} 5. & 3.718281828 & -4. \\ 10.87312731 & 8.154845484 & -10.87312731 \\ 7.718281828 & 6.436563656 & -6.718281828 \end{bmatrix}$$

2. Für zwei durch eine Feder gekoppelte Pendel gleicher Masse $m = 1$ und gleicher Länge $l$ lauten die Bewegungsgleichungen

$$\begin{aligned} \ddot{x} &= -\alpha x - k(x-y) \\ \ddot{y} &= -\alpha y - k(y-x), \end{aligned}$$

wobei $g$ die Erdbeschleunigung und $k$ die (positive) Federkonstante bedeuten und $\alpha := g/l$ gesetzt ist. Schreibt man die beiden Differentialgleichungen zweiter Ordnung als ein System von vier Differentialgleichungen erster Ordnung, so erhält man ein homogenes System mit der Koeffizientenmatrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -(\alpha+k) & 0 & k & 0 \\ 0 & 0 & 0 & 1 \\ k & 0 & -(\alpha+k) & 0 \end{pmatrix}.$$

Man bestimme ein zu $x' = Ax$ gehörendes (nicht notwendig normiertes) Fundamentalsystem. Ferner löse man die Anfangswertaufgabe für den Fall, dass zur Zeit $t = 0$ ein Pendel angestoßen wird bzw. die Anfangswerte $x(0) = y(0) = y'(0) = 0$, $x'(0) = 1$ vorgegeben werden.

**Lösung:** Durch die Eingabe

```
A:=Matrix([[0,1,0,0],[-(alpha+k),0,k,0],[0,0,0,1],[k,0,-(alpha+k),0]]);
(lambda,C):=LinearAlgebra[Eigenvectors](A);
```

(bzw. `(lambda,C):=Eigenvectors(A);` wenn vorher durch `with(LinearAlgebra):` das `LinearAlgebra`-Paket geladen ist) erhält man die Eigenwerte

$$\lambda_{1,2} = \pm i\sqrt{\alpha}, \qquad \lambda_{3,4} = \pm i\sqrt{\alpha + 2k}$$

mit zugehörigen Eigenvektoren

$$c_{1,2} = \begin{pmatrix} 1 \\ \pm i\sqrt{\alpha} \\ 1 \\ \pm i\sqrt{\alpha} \end{pmatrix}, \qquad c_{3,4} = \begin{pmatrix} 1 \\ \pm i\sqrt{\alpha+2k} \\ -1 \\ \mp i\sqrt{\alpha+2k} \end{pmatrix}.$$

Daher erhält man durch

$$x_1(t) = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \cos\sqrt{\alpha}t - \begin{pmatrix} 0 \\ \sqrt{\alpha} \\ 0 \\ \sqrt{\alpha} \end{pmatrix} \sin\sqrt{\alpha}t,$$

$$
\begin{aligned}
x_2(t) &= \begin{pmatrix} 0 \\ \sqrt{\alpha} \\ 0 \\ \sqrt{\alpha} \end{pmatrix} \cos\sqrt{\alpha}t + \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \sin\sqrt{\alpha}t, \\
x_3(t) &= \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \cos\sqrt{\alpha+2k}t - \begin{pmatrix} 0 \\ \sqrt{\alpha+2k} \\ 0 \\ -\sqrt{\alpha+2k} \end{pmatrix} \sin\sqrt{\alpha+2k}t, \\
x_4(t) &= \begin{pmatrix} 0 \\ \sqrt{\alpha+2k} \\ 0 \\ -\sqrt{\alpha+2k} \end{pmatrix} \cos\sqrt{\alpha+2k}t + \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} \sin\sqrt{\alpha+2k}t,
\end{aligned}
$$

Spalten eines Fundamentalsystems $X(t)$ zu $x' = Ax$. Die allgemeine Lösung ist eine Linearkombination dieser Spalten. Wir lösen das lineare Gleichungssystem mit der Koeffizientenmatrix $X(0)$ und der rechten Seite $(0,1,0,0)^T$. Mit

```
X_0:=Matrix([[1,0,1,0],[0,sqrt(alpha),0,sqrt(alpha+2*k)],[1,0,-1,0],
[0,sqrt(alpha),0,-sqrt(alpha+2*k)]]);
b:=Vector([0,1,0,0]);
LinearAlgebra[LinearSolve](X_0,b);
```

erhält man, dass die Lösung der Anfangswertaufgabe

$$
\begin{aligned}
\ddot{x} &= -\alpha x - k(x-y) \\
\ddot{y} &= -\alpha y - k(y-x),
\end{aligned}
\qquad
\begin{pmatrix} x(0) \\ x'(0) \\ y(0) \\ y'(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}
$$

durch

$$
\begin{pmatrix} x(t) \\ x'(t) \\ y(t) \\ y'(t) \end{pmatrix} = \frac{1}{2\sqrt{\alpha}} x_2(t) + \frac{1}{2\sqrt{\alpha+2k}} x_4(t)
$$

gegeben ist. Daher ist

$$
\begin{aligned}
x(t) &= \frac{1}{2\sqrt{\alpha}} \sin\sqrt{\alpha}t + \frac{1}{2\sqrt{\alpha+2k}} \sin\sqrt{\alpha+2k}t, \\
y(t) &= \frac{1}{2\sqrt{\alpha}} \sin\sqrt{\alpha}t - \frac{1}{2\sqrt{\alpha+2k}} \sin\sqrt{\alpha+2k}t.
\end{aligned}
$$

3. Man bestimme (wie auch immer)[14] ein reelles Fundamentalsystem von Lösungen der Differentialgleichungssysteme $x' = Ax$ mit

$$
A := \begin{pmatrix} 3 & 6 \\ -2 & -3 \end{pmatrix}, \qquad A := \begin{pmatrix} 8 & 1 \\ -4 & 4 \end{pmatrix}.
$$

[14] Die Aufgabe ist W. Walter (1996, S. 159) entnommen.

**Lösung:** Im ersten Fall erhält man als Eigenwerte und zugehörige Eigenvektoren

$$\lambda_{1,2} = \pm i\sqrt{3}, \qquad c_{1,2} = \begin{pmatrix} \frac{1}{2}(-3 \mp \sqrt{3}) \\ 1 \end{pmatrix}.$$

Als Spalten eines Fundamentalsystems erhält man

$$\begin{aligned} x_1(t) &= \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix} \cos\sqrt{3}t + \begin{pmatrix} \frac{\sqrt{3}}{2} \\ 0 \end{pmatrix} \sin\sqrt{3}t, \\ x_2(t) &= \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ 0 \end{pmatrix} \cos\sqrt{3}t + \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix} \sin\sqrt{3}t. \end{aligned}$$

Im zweiten Fall hat $A$ den doppelten Eigenwert $\lambda_{1,2} = 6$ und ist nicht diagonalisierbar. Bei Maple wird dies nach dem Aufruf von `Eigenvectors(A)` dadurch deutlich gemacht, dass neben dem Eigenvektor $(1, -2)^T$ noch ein Nullvektor ausgegeben wird. Durch eine Ähnlichkeitstransformation mit der Matrix

$$C := \begin{pmatrix} 2 & 1 \\ -4 & 0 \end{pmatrix}$$

erhält man die Jordansche Normalform von $A$, d. h. es ist

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

Dies kann man z. B. durch den Maple-Befehl `JordanForm` erhalten. Daher ist

$$e^{At} = \begin{pmatrix} 2 & 1 \\ -4 & 0 \end{pmatrix} e^{6t} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ -4 & 0 \end{pmatrix}^{-1} = e^{6t} \begin{pmatrix} 1+2t & t \\ -4t & 1-2t \end{pmatrix}.$$

Etwas schneller hätten wir dies erhalten können, indem wir

```
dsolve({diff(x(t),t)=8*x(t)+y(t),diff(y(t),t)=-4*x(t)+4*y(t),
    x(0)=1,y(0)=0},{x(t),y(t)});
```

und

```
dsolve({diff(x(t),t)=8*x(t)+y(t),diff(y(t),t)=-4*x(t)+4*y(t),
    x(0)=0,y(0)=1},{x(t),y(t)});
```

eingegeben hätten. Wesentlich schneller erhält man das Ergebnis mit der `exponential`-Funktion aus dem `linalg`-package von Maple:

```
> with(linalg):
> A:=Matrix([[3,6],[-2,-3]]):
> exponential(A,t);
```

$$\begin{bmatrix} \cos(\sqrt{3}\,t) + \%1 & 2\,\%1 \\ -\frac{2}{3}\,\%1 & \cos(\sqrt{3}\,t) - \%1 \end{bmatrix}$$

$$\%1 := \sqrt{3}\sin(\sqrt{3}\,t)$$

```
> A:=Matrix([[8,1],[-4,4]]):
> exponential(A,t);
```

$$\begin{bmatrix} e^{(6\,t)} + 2\,t\,e^{(6\,t)} & t\,e^{(6\,t)} \\ -4\,t\,e^{(6\,t)} & e^{(6\,t)} - 2\,t\,e^{(6\,t)} \end{bmatrix}$$

4. Man bestimme die allgemeine Lösung von

$$x'' - 6x' + 25x = e^{2t}.$$

Anschließend bestimme man die Lösung zu den Anfangswerten $x(0) = 1$, $x'(0) = 0$.

**Lösung:** Eine spezielle Lösung ist $\frac{1}{17}e^{2t}$. Die Nullstellen von $\lambda^2 - 6\lambda + 25 = 0$ sind $\lambda_{1,2} = 3 \pm 4i$. Die allgemeine Lösung der homogenen Aufgabe ist daher $e^{3t}(\alpha \cos 4t + \beta \sin 4t)$, folglich die allgemeine Lösung der inhomogenen Aufgabe

$$x(t) = e^{3t}(\alpha \cos 4t + \beta \sin 4t) + \frac{1}{17}e^{2t}$$

mit beliebigen $\alpha, \beta$. Die gewünschte Lösung der Anfangswertaufgabe ist

$$x(t) = \frac{1}{17}e^{2t} + e^{3t}\left(\frac{16}{17}\cos 4t - \frac{25}{34}\sin 4t\right).$$

Mit Maple hätte man natürlich dieselben Ergebnisse erzielt.

## 7.5.4 Aufgaben zu Abschnitt 5.4

1. Man zeige: Ist $p \in \mathcal{P}_3$ (Menge der kubischen Polynome), so ist

$$\int_a^b p(t)\,dt = \frac{b-a}{6}\,[p(a) + 4\,p(\tfrac{1}{2}\,(a+b)) + p(b)].$$

**Lösung:** Es bilden $1$, $t - \frac{1}{2}(a+b)$, $(t - \frac{1}{2}(a+b))^2$ und $(t - \frac{1}{2}(a+b))^3$ eine Basis von $\mathcal{P}_3$. Es genügt, die behauptete Gleichung für die Basiselemente nachzuweisen. Wegen

$$b - a = \int_a^b 1\,dt = \frac{b-a}{6}[1 + 4 + 1] = b - a$$

und

$$\tfrac{1}{12}(b-a)^3 = \int_a^b [t - \tfrac{1}{2}(a+b)]^2\,dt = \frac{b-a}{6}[\tfrac{1}{4}(b-a)^2 + \tfrac{1}{4}(b-a)^2] = \tfrac{1}{12}(b-a)^3$$

ist dies für das erste und dritte Basiselement richtig, was trivialerweise auch für das zweite und vierte gilt.

2. Man betrachte ein Einschrittverfahren mit der Verfahrensfunktion

$$\Phi(h, f)(t, u) := a_1 f(t, u) + a_2 f(t + b_1 h, u + b_2 h f(t, u))$$

und zeige, dass dieses die Ordnung 2 besitzt, falls

$$a_1 + a_2 = 1, \qquad a_2 b_1 = \frac{1}{2}, \qquad a_2 b_2 = \frac{1}{2}.$$

Spezialfälle erhält man übrigens für $a_1 = 0$, $a_2 = 1$, $b_1 = b_2 = \frac{1}{2}$ (modifiziertes Euler-Verfahren) und für $a_1 = a_2 = \frac{1}{2}$, $b_1 = b_2 = 1$ (Heun-Verfahren).

**Lösung:** Bei gegebenem $(t,u) \in [t_0,T] \times \mathbb{R}^n$ sei $z$ die Lösung von $z' = f(s,z)$, $z(t) = u$. Der lokale Diskretisierungsfehler ist gegeben durch

$$\begin{aligned}
\Delta(h,f)(t,u)) &= \frac{z(t+h)-z(t)}{h} - \Phi(h,f)(t,z(t)) \\
&= z'(t) + \frac{h}{2}z''(t) + O(h)^2 \\
&\quad - [a_1 f(t,z(t)) + a_2 f(t+b_1 h, z(t) + b_2 h f(t,z(t))] \\
&= f(t,z(t)) + \frac{h}{2}[f_t(t,z(t)) + f_x(t,z(t))f(t,z(t))] + O(h^2) \\
&\quad - \{(a_1+a_2)f(t,z(t)) \\
&\quad + a_2[b_1 h f_t(t,z(t)) + b_2 h f_x(t,z(t))f(t,z(t))]\} \\
&= [1-(a_1+a_2)]f(t,z(t)) + h f_t(t,z(t))\left[\frac{1}{2} - a_2 b_1\right] \\
&\quad + h f_x(t,z(t))f(t,x(t))\left[\frac{1}{2} - a_2 b_2\right] + O(h^2),
\end{aligned}$$

woraus man die Behauptung abliest.

3. Man betrachte ein Einschrittverfahren mit der Verfahrensfunktion

$$\Phi(h,f)(t,u) := \frac{1}{4}k_1 + \frac{3}{4}k_3,$$

wobei

$$k_1 := f(t,u), \qquad k_2 := f(t+\tfrac{1}{3}h, u+\tfrac{1}{3}hk_1), \qquad k_3 := f(t+\tfrac{2}{3}h, u+\tfrac{2}{3}hk_2).$$

Man zeige, dass dies ein Verfahren der Ordnung 3 ist. Hierbei darf man sich auf den Fall einer Differentialgleichung erster Ordnung, also $n = 1$, beschränken. Anschließend löse man diese Aufgabe mit Maple.

**Lösung:** Bei gegebenem $(t,u) \in [t_0,T] \times \mathbb{R}^n$ sei $z$ die Lösung von $z' = f(s,z)$, $z(t) = u$. Statt des Argumentes $(t,z(t))$ schreiben wir im folgenden nur $t$. Es ist also $z'(t) := f(t,z(t)) = f(t)$, daher

$$z''(t) = f_t(t) + f_x(t)f(t)$$

und folglich

$$\begin{aligned}
z'''(t) &= f_{tt}(t) + f_{tx}(t)f(t) + [f_{xt}(t) + f_{xx}(t)f(t)]f(t) + f_x(t)[f_t(t) + f_x(t)f(t)] \\
&= f_{tt}(t) + 2f_{xt}(t)f(t) + f_{xx}(t)f^2(t) + f_x(t)[f_t(t) + f_x(t)f(t)].
\end{aligned}$$

Daher ist der lokale Diskretisationsfehler gegeben durch

$$\begin{aligned}
\Delta(h,f)(t,u) &= \frac{z(t+h)-z(t)}{h} - \Phi(h,f)(t,z(t)) \\
&= z'(t) + \frac{h}{2}z''(t) + \frac{h^2}{6}z'''(t) + O(h^3) - \Phi(h,f)(t,z(t)) \\
&= f(t) + \frac{h}{2}[f_t(t) + f_x(t)f(t)] \\
&\quad + \frac{h^2}{6}\{f_{tt}(t) + 2f_{xt}(t)f(t) + f_{xx}(t)f^2(t) \\
&\quad + f_x(t)[f_t(t) + f_x(t)f(t)]\} + O(h^3) - \Phi(h,f)(t,z(t)).
\end{aligned}$$

Andererseits ist

$$\begin{aligned} k_2(t) &= f(t+\tfrac{1}{3}h, z(t)+\tfrac{1}{3}hk_1(t)) \\ &= f(t) + \frac{h}{3}\begin{pmatrix} f_t(t) \\ f_x(t) \end{pmatrix}^T \begin{pmatrix} 1 \\ f(t) \end{pmatrix} \\ &\qquad + \frac{h^2}{18}\begin{pmatrix} 1 \\ f(t) \end{pmatrix}^T \begin{pmatrix} f_{tt}(t) & f_{tx}(t) \\ f_{xt}(t) & f_{xx}(t) \end{pmatrix} \begin{pmatrix} 1 \\ f(t) \end{pmatrix} + O(h^3) \\ &= f(t) + \frac{h}{3}[f_t(t) + f_x(t)f(t,u)] \\ &\qquad + \frac{h^2}{18}[f_{tt}(t) + 2f_{xt}(t)f(t) + f_{xx}(t)f^2(t)] + O(h^3). \end{aligned}$$

Entsprechend ist

$$\begin{aligned} k_3(t) &= f(t+\tfrac{2}{3}h, z(t)+\tfrac{2}{3}hk_2(t)) \\ &= f(t) + \frac{2h}{3}\begin{pmatrix} f_t(t) \\ f_x(t) \end{pmatrix}^T \begin{pmatrix} 1 \\ k_2(t) \end{pmatrix} \\ &\qquad + \frac{2h^2}{9}\begin{pmatrix} 1 \\ k_2(t) \end{pmatrix}^T \begin{pmatrix} f_{tt}(t) & f_{tx}(t) \\ f_{xt}(t) & f_{xx}(t) \end{pmatrix} \begin{pmatrix} 1 \\ k_2(t) \end{pmatrix} + O(h^3) \\ &= f(t) + \frac{2h}{3}[f_t(t) + f_x(t)k_2(t)] \\ &\qquad + \frac{2h^2}{9}[f_{tt}(t) + 2k_2 f_{xt}(t) + k_2^2 f_{xx}(t)] + O(h^3) \\ &= f(t) + \frac{2h}{3}[f_t(t) + f_x(t)f(t)] + \frac{2h^2}{9}f_x(t)[f_t(t) + f_x(t)f(t)] \\ &\qquad + \frac{2h^2}{9}[f_{tt}(t,u) + 2f_{xt}(t,u)f(t,u) + f_{xx}(t,u)f^2(t,u)] + O(h^3). \end{aligned}$$

Folglich ist

$$\begin{aligned} \Phi(h,f)(t,z(t)) &= \frac{1}{4}k_1(t) + \frac{3}{4}k_3(t) \\ &= f(t) + \frac{h}{2}[f_t(t) + f_x(t)f(t)] + \frac{h^2}{6}\{f_x(t)[f_t(t) + f_x(t)f(t)] \\ &\qquad + f_{tt}(t) + 2f_{xt}(t)f(t) + f_{xx}(t)f^2(t)\} + O(h^3). \end{aligned}$$

Wir erkennen also, dass in der Tat

$$\Delta(h,f)(t,u) = O(h^3),$$

die Konsistenzordnung des Verfahrens also 3 ist. Genau das war zu zeigen. Mit Maple geht es wesentlich einfacher:

```
> restart;
> g:=t->f(t,z(t)):
> k_1:=h->f(t,z(t)):
> k_2:=h->f(t+(h/3),z(t)+(h/3)*k_1(h)):
> k_3:=h->f(t+(2*h/3),z(t)+(2*h/3)*k_2(h)):
> Delta:=h->(z(t+h)-z(t))/h-(1/4)*k_1(h)-(3/4)*k_3(h):
```

```
> s:=series(Delta(h),h,4):
> s1:=subs((D@@3)(z)(t)=(D@@2)(g)(t),s):
> s2:=subs((D@@2)(z)(t)=D(g)(t),s1):
> s3:=subs(D(z)(t)=g(t),s2):
> simplify(%);
```

$$\mathrm{O}(h^3)$$

4. Man betrachte ein Einschrittverfahren mit der Verfahrensfunktion

$$\Phi(h,f)(t,u) := \frac{1}{6}(k_1 + 4k_2 + k_3),$$

wobei

$$k_1 := f(t,u), \qquad k_2 := f(t+\tfrac{1}{2}h, u+\tfrac{1}{2}hk_1), \qquad k_3 := f(t+h, u-hk_1+2hk_2).$$

Man zeige, dass dies ein Verfahren der Ordnung 3 ist. Hierbei darf man sich auf den Fall einer Differentialgleichung erster Ordnung, also $n = 1$, beschränken und die Aufgabe mit Maple lösen.

**Lösung:** Wir geben fast genau wie zur Bearbeitung der vorigen Aufgabe ein:

```
g:=t->f(t,z(t)):
k_1:=h->f(t,z(t)):
k_2:=h->f(t+(h/2),z(t)+(h/2)*k_1(h)):
k_3:=h->f(t+h,z(t)-h*k_1(h)+2*h*k_2(h)):
Delta:=h->(z(t+h)-z(t))/h-(1/6)*(k_1(h)+4*k_2(h)+k_3(h)):
s:=series(Delta(h),h,4):
s_1:=subs((D@@3)(z)(t)=(D@@2)(g)(t),s):
s_2:=subs((D@@2)(z)(t)=D(g)(t),s_1):
s_3:=subs(D(z)(t)=g(t),s_2);
```

Anschließendes `simplify(%);` liefert $O(h^3)$, die Konsistenzordnung ist also 3.

5. Man bestimme[15] die exakte Lösung der Anfangswertaufgabe $x' = (2/t)x$, $x(1) = 1$. Anschließend bestimme man einen analytischen Ausdruck für die durch das Eulersche Polygonzugverfahren erhaltene Näherung und gebe den globalen und den lokalen Diskretisierungsfehler an.

**Lösung:** Durch

```
dsolve({D(x)(t)=(2/t)*x(t),x(1)=1},x(t));
```

erhält man die Lösung $x(t) = t^2$. Beim Euler-Verfahren ist $u(t+h) = [1 + 2h/t]u(t)$ und natürlich $u(1) = 1$. Sei $t > 1$ fest und $h = (t-1)/m$ mit $m \in \mathbb{N}$. Die durch das Euler-Verfahren gewonnene Näherung $u(t;h)$ ist offenbar gegeben durch

$$u(t;h) = \prod_{i=0}^{m-1} \left(1 + \frac{2h}{1+ih}\right).$$

[15] Diese Aufgabe ist dem Lehrbuch

R. Kress (1998) *Numerical Analysis*. Springer-Verlag, New York-Berlin-Heidelberg.

entnommen.

Nun ist (für beliebiges $h$)

$$\prod_{i=0}^{m-1}\left(1+\frac{2h}{1+ih}\right)=\frac{(1+mh)(1+(m+1)h)}{1+h}.$$

Dies zeigt man natürlich durch vollständige Induktion nach $m$. Für $m=1$ (oder noch einfacher, für $m=0$) ist die Behauptung richtig. Im Induktionsschritt ist

$$\begin{aligned}\prod_{i=0}^{m}\left(1+\frac{2h}{1+ih}\right) &= \frac{(1+mh)(1+(m+1)h)}{1+h}\left(1+\frac{2h}{1+mh}\right)\\ &= \frac{(1+(m+1)h)(1+(m+2)h)}{1+h},\end{aligned}$$

das war im Induktionsschritt zu zeigen. Folglich ist

$$u(t;h)-x(t)=\frac{t(t+h)}{1+h}-t^2=t(1-t)\frac{h}{1+h}$$

der globale Diskretisierungsfehler.

Nun sei $z$ die Lösung von $z'=(1/s)z$, $z(t)=u$. Wir erhalten $z(s)=us^2/t^2$. Daher ist der lokale Diskretisierungsfehler

$$\begin{aligned}\Delta(h,f)(t,u) &= \frac{z(t+h)-z(t)}{h}-\Phi(h,f)(t,u)\\ &= \frac{u[(t+h)^2/t^2-1]}{h}-\frac{2u}{t}\\ &= \frac{u}{t^2}h.\end{aligned}$$

6. Man schreibe ein MATLAB-Programm für das klassische Runge-Kutta-Verfahren mit automatischer Schrittweitensteuerung. Anschließend teste man das Programm an der Anfangswertaufgabe (siehe Stoer-Bulirsch)

$$x'=-200tx^2,\qquad x(-3)=\frac{1}{901},$$

welche auf $[-3,0]$ zu lösen sei.

**Lösung:** Die exakte Lösung ist

$$x(t)=\frac{1}{1+100t^2}.$$

Ein Plot (siehe Abbildung 7.29) zeigt, wo die Schwierigkeiten sind. Etwa auf $[-3,\frac{1}{2}]$ kann mit einer verhältnismäßig großen Schrittweite gerechnet werden, danach muss zunehmend verfeinert werden. Wir haben die MATLAB-Funktion `RKauto` geschrieben, welche die dort enthaltene Funktion `RKstep` benutzt.

```
function  [tvals,xvals,steps]=RKauto(fname,x_0,t_0,t_max,H,epsilon);
%*********************************************************************
%Input:   fname     string fuer die rechte Seite, eine Funktion der Form
%                   f(t,x), wobet t ein Skalar und x ein Spaltenvektor
%         x_0       Vektor des Anfangzustandes
```

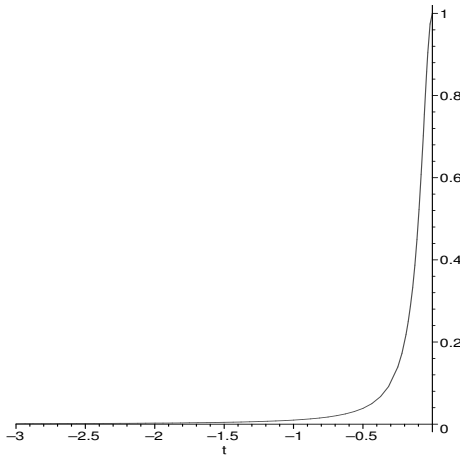


Abbildung 7.29: Die Lösung von $x' = -200tx^2$, $x(-3) = 1/901$ auf $[-3, 0]$

```
%         t_0       Anfangszeit
%         t_max     Endzeit
%         H         Schrittweite beim Start
%         epsilon   kleine Zahl, der Fehler ist ungefaehr gleich epsilon
%Output:  tvals     tvals(k) ist (k-1)-te Zeit
%         xvals     xvals(:,k) ist Naeherungsloesung zur Zeit t=tvals(k)
%         steps     Anzahl der Schritte. Die Funktion RKstep wird
%                   3*steps aufgerufen
%*********************************************************************
t_c=t_0; x_c=x_0; tvals=t_c; xvals=x_c; f_c=feval(fname,t_c,x_c);
steps=0;
while (t_c<t_max)
     step_small=0;
     while (step_small==0)
          [xH,fH]=RKstep(fname,t_c,x_c,f_c,H);
          [xH2,fH2]=RKstep(fname,t_c,x_c,f_c,H/2);
          [xH2,fH2]=RKstep(fname,t_c+H/2,xH2,fH2,H/2);
          h=H*(15*epsilon/(16*norm(xH-xH2,inf)))^(1/5);
          steps=steps+1;
          if (H>=3*h)
              H=2*h;
          else
              step_small=1;
              t_c=t_c+H;
              x_c=xH2;
              f_c=fH2;
              H=2*h;
          end;
     end;
     xvals=[xvals x_c];
     tvals=[tvals t_c];
end;
%*********************************************************************
function [x_new,f_new]=RKstep(fname,t_c,x_c,f_c,h);
%Input:    fname   string fuer die rechte Seite, eine Funktion der Form
%                  f(t,x), wobet t ein Skalar und x ein Spaltenvektor
```

```
%          x_c     eine Naeherung fuer x'=f(t,x) in t=t_c
%          f_c     f(t_c,x_c)
%          h       Schrittweite
%Output:   x_new   x_new ist eine Naeherung in t_new=t_c+h, and
%          f_new   f(t_new,x_nrw)

   k_1=f_c;
   k_2=feval(fname,t_c+(h/2),x_c+(h/2)*k_1);
   k_3=feval(fname,t_c+(h/2),x_c+(h/2)*k_2);
   k_4=feval(fname,t_c+h,x_c+h*k_3);
   x_new=x_c+(h/6)*(k_1+2*k_2+2*k_3+k_4);
   f_new=feval(fname,t_c+h,x_new);
```

Die rechte Seite wird in ein File `f.m` geschrieben:

```
function out=f(t,x);
   out=-200*t*x^2;
```

Mit dem Aufruf

```
[t,x,steps]=RKauto('f',1/901,-3,0,0.1,1.e-14);
```

erhalten wir `steps=687`, es wurden also $3 \cdot 687 = 2061$ Runge-Kutta-Schritte durchgeführt. Nach dem Aufruf sind `t` und `x` jeweils Zeilenvektoren mit 687 Komponenten, so viele Zeitschritte wurden also durchgeführt. Am Schluss ist

```
t(687)=2.353705998324984e-04
x(687)-1/(1+100*t(687)^2)=-4.300241132071392e-07
```

Die kleinste auftretende Schrittweite ist $h = 6.246407336988336 \cdot 10^{-4}$, die größte $h = 0.03265113115694$.

# 7.6 Aufgaben zu Kapitel 6

## 7.6.1 Aufgaben zu Abschnitt 6.1

1. Sei $\phi\colon [0,\pi] \longrightarrow \mathbb{R}$ definiert durch

$$\phi(x) := \begin{cases} x, & x \in [0, \frac{1}{2}\pi], \\ \pi - x, & x \in [\frac{1}{2}\pi, \pi]. \end{cases}$$

Man setze $\phi$ ungerade auf $[-\pi, \pi]$ und anschließend $2\pi$-periodisch auf ganz $\mathbb{R}$ fort. Man zeige, dass

$$a_k := \frac{1}{\pi}\int_0^{2\pi} \phi(x)\cos kx\,dx = 0, \qquad k = 0, 1, \ldots$$

und

$$b_k := \frac{1}{\pi}\int_0^{2\pi} \phi(x)\sin kx\,dx = \begin{cases} 0, & \text{falls } k \text{ gerade}, \\ \dfrac{4}{\pi k^2}(-1)^{(k-1)/2}, & \text{falls } k \text{ ungerade}, \end{cases}$$

also

$$\frac{4}{\pi}\sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)^2}\sin(2j+1)x$$

die Fourierreihe von $\phi$ ist.

**Lösung:** Es ist

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_0^{2\pi}\phi(x)\cos kx\,dx \\ &= \frac{1}{\pi}\left(\int_0^{\pi}\phi(x)\cos kx\,dx - \int_{\pi}^{2\pi}\phi(2\pi-x)\cos kx\,dx\right) \\ &= 0. \end{aligned}$$

Weiter ist

$$\begin{aligned} b_k &= \frac{1}{\pi}\int_0^{2\pi}\phi(x)\sin kx\,dx \\ &= \frac{1}{\pi}\left(\int_0^{\pi}\phi(x)\sin kx - \int_{\pi}^{2\pi}\phi(2\pi-x)\sin kx\,dx\right) \\ &= \frac{2}{\pi}\int_0^{\pi}\phi(x)\sin kx\,dx \\ &= \frac{2}{\pi}\left(\int_0^{\pi/2}x\sin kx\,dx + \int_{\pi/2}^{\pi}(\pi-x)\sin kx\,dx\right) \\ &= \frac{2}{\pi}[1+(-1)^{k+1}]\int_0^{\pi/2}x\sin kx\,dx \\ &= \begin{cases} 0, & \text{falls } k \text{ gerade,} \\ \dfrac{4}{\pi k^2}(-1)^{(k-1)/2}, & \text{falls } k \text{ ungerade,} \end{cases} \end{aligned}$$

und das war zu zeigen.

2. Gegeben sei mit einem $\alpha \geq 0$ die Tridiagonalmatrix

$$A := \begin{pmatrix} 1+2\alpha & -\alpha & \cdots & 0 \\ -\alpha & 1+2\alpha & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\alpha \\ 0 & \cdots & -\alpha & 1+2\alpha \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Man zeige:

   (a) $A$ ist symmetrisch und positiv definit. Genauer sind sogar alle Eigenwerte von $A$ größer oder gleich Eins.

   (b) Alle Einträge der (nichtsingulären) Matrix $A^{-1}$ sind nichtnegativ.

**Lösung:** Sei $\alpha > 0$ (andernfalls ist $A$ die Einheitsmatrix, die Aussagen sind dann trivialerweise richtig). Offenbar ist $A$ symmetrisch. Wir zeigen, dass alle Eigenwerte von $A$ positiv sind und folglich $A$ positiv definit ist. Sei $\lambda$ ein Eigenwert von $A$ mit

zugehörigem Eigenvektor $u = (u_j)$. Man wähle ein $i \in \{1, \ldots, n\}$ mit $|u_i| = \|u\|_\infty$. Die $i$-te Gleichung in $Au = \lambda u$ lautet dann

$$-\alpha u_{i-1} + (1 + 2\alpha)u_i - \alpha u_{i+1} = \lambda u_i$$

bzw.

$$(1 + 2\alpha - \lambda)u_i = \alpha(u_{i-1} + u_{i+1}),$$

wobei hier $u_0 := 0$ bzw. $u_{n+1} := 0$ zu setzen ist, wenn $i = 1$ bzw. $i = n$. Aus

$$|1 + 2\alpha - \lambda| \, \|u\|_\infty \leq \alpha(|u_{i-1}| + |u_{i+1}|) \leq 2\alpha \, \|u\|_\infty.$$

Also ist $|1 + 2\alpha - \lambda| \leq 2\alpha$ und folglich $1 \leq \lambda \leq 1 + 4\alpha$. Damit ist der erste Teil der Aufgabe gezeigt.

Wir zeigen für den zweiten Teil der Aufgabe: Ist $b \in \mathbb{R}^n$ ein beliebiger nichtnegativer Vektor (d. h. alle Komponenten von $b$ sind nichtnegativ) und $u$ die (wegen der Nichtsingularität von $A$) eindeutige Lösung von $Au = b$, so ist auch $u$ ein nichtnegativer Vektor. Ist dies gelungen, so ist die Behauptung bewiesen, denn als rechte Seite $b$ brauchen nur die $n$ Einheitsvektoren gewählt zu werden. Man bestimme einen Index $i \in \{1, \ldots, n\}$ mit $u_i \leq u_j$, $j = 1, \ldots, n$, also eine minimale Komponente von $u$. Indem man die $i$-te Gleichung in $Au = b$ betrachtet, erhält man

$$0 \leq b_i = -\alpha u_{i-1} + (1 + 2\alpha)u_i - \alpha u_{i+1} \leq u_i.$$

Folglich ist auch $u$ ein nichtnegativer Vektor, auch der zweite Teil der Aufgabe ist gelöst.

3. Gegeben sei mit einem $\alpha > 0$ die Tridiagonalmatrix

$$A := \begin{pmatrix} 1 + 2\alpha & -\alpha & \cdots & 0 \\ -\alpha & 1 + 2\alpha & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\alpha \\ 0 & \cdots & -\alpha & 1 + 2\alpha \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Man zeige, dass durch den folgenden Algorithmus die Lösung $u$ des linearen Gleichungsssystems $Au = b$ bestimmt wird:

- Berechne $\gamma_1 := 1 + 2\alpha$, setze $c_1 := b_1$.
- Für $i = 2, \ldots, n$:
  - Berechne $\beta_i := -\alpha/\gamma_{i-1}$.
  - Berechne $\gamma_i := 1 + 2\alpha + \beta_i\alpha$.
  - Berechne $c_i := b_i - \beta_i c_{i-1}$.
- Berechne $u_n := c_n/\gamma_n$.
- Für $i = n - 1, \ldots, 1$:
  - Berechne $u_i := (c_i + \alpha u_{i+1})/\gamma_i$.
- Output: Durch $u = (u_i)$ ist die Lösung von $Au = b$ gegeben.

Hinweis: Man zeige, dass $A = LR$ mit

$$L := \begin{pmatrix} 1 & & & \\ \beta_2 & 1 & & \\ & \ddots & \ddots & \\ & & \beta_n & 1 \end{pmatrix}, \qquad R := \begin{pmatrix} \gamma_1 & -\alpha & & \\ & \gamma_2 & \ddots & \\ & & \ddots & -\alpha \\ & & & \gamma_n \end{pmatrix}.$$

**Lösung:** Die behauptete Zerlegung von $A$ folgt durch einfaches Nachrechnen. Die Lösung $u$ von $Au = b$ erhält man daher in zwei Schritten. Zunächst bestimmt man den Vektor $c$ als Lösung von $Lc = b$ durch Vorwärtseinsetzen, anschließend die gesuchte Lösung $u$ aus $Ru = c$ durch Rückwärtseinsetzen.

4. Gegeben sei die Anfangsrandwertaufgabe für die Wärmeleitungsgleichung

$$\begin{aligned} u_t &= \sigma u_{xx} \quad \text{auf} \quad (0,\pi) \times (0,\infty), \\ u(x,0) &= \phi(x) \quad \text{auf} \quad (0,\pi) \end{aligned}$$

und $u(0,t) = u(\pi,t) = 0$ auf $(0,\infty)$. Hierbei sei $\phi \in C[0,\pi]$ und $\phi(0) = \phi(\pi) = 0$. Man schreibe ein Programm zur Lösung dieser Aufgabe mit Hilfe des Crank-Nicholson-Verfahrens. Außer $\sigma$ und $\phi$ seien *Eingabe*parameter: $J \in \mathbb{N}$ (bestimmt die Schrittweite $\Delta x$ in Ortsrichtung), $\Delta t$ (die Schrittweite in Zeitrichtung), $\theta \in [0,1]$ (Parameter bei Crank-Nicholson), $k_{\max} \in \mathbb{N}$ (Anzahl der Zeitschritte, die durchgeführt werden sollen). *Ausgabe*parameter sei $u$, ein Vektor bzw. Liste mit den Näherungen zur Zeit $T = k_{\max}\Delta t$. Man teste das Programm für $\sigma := 1$,

$$\phi(x) := \begin{cases} x, & x \in [0, \frac{1}{2}\pi], \\ \pi - x, & x \in [\frac{1}{2}\pi, \pi], \end{cases}$$

für $k_{\max} := 100$ Zeitschritte, $\Delta t = 0.001$ (es wird also eine Näherung zur Zeit $T = 0.1$ ausgegeben) und plotte die erhaltenenen Werte für $\theta = 0,\ 0.5,\ 1.0$ und $J = 10,\ 100$.

**Lösung:** Wir benutzen weitgehend dieselben Bezeichnungen wie in dem entsprechenden Abschnitt über die Wärmeleitungsgleichung[16].

```
ClearAll[phi,heat];
phi[x_]:=If[x<=0.5*Pi,x,Pi-x];
heat[sigma_,J_,deltat_,kmax_,theta_]:=Block[
{u,i,j,k,deltax,kappa,alpha,Jminus,b,beta,gamma},
deltax=N[Pi/J]; kappa=sigma*deltat/deltax^2;
alpha=theta*kappa;Jminus=J-1;
u=Table[N[phi[j*deltax]],{j,0,J}];
```

[16] Bei

V. G. Ganzha, E. V. Vorozhtsov (1996) *Numerical Solutions for Partial Differential Equations. Problem Solving Using Mathematica*. CRC Press, Boca Raton-New York-London-Tokyio

findet man auf S. 314 ff. ein *Mathematica*-Programm zur Lösung der Wärmeleitungsgleichung, welches auf dem expliziten Differenzenverfahren beruht. Ohne ein Experte in *Mathematica* zu sein stellt man fest, dass dieses nicht besonders gut ist. Schon besser ist da die Vorgehensweise auf S. 441 ff. von

R. Glass (1998) *Mathematica for Scientists and Engineers. Using Mathematica to do Science*. Prentice Hall, Upper Saddle River.

```
For[k=1,k<=kmax,k++,
b=Take[u+(1-theta)*kappa*(RotateRight[u]-2*u+RotateLeft[u]),{2,-2}];
beta=gamma=Table[1.0,{j,1,Jminus}];
gamma[[1]]=1+2*alpha;
For[i=2,i<=Jminus,i++,
{beta[[i]]=-alpha/gamma[[i-1]];gamma[[i]]=1+2*alpha+beta[[i]]*alpha;
b[[i]]=b[[i]]-beta[[i]]*b[[i-1]]};
b[[Jminus]]=b[[Jminus]]/gamma[[Jminus]];
For[i=Jminus-1,i>=1,i--,b[[i]]=(b[[i]]+alpha*b[[i+1]])/gamma[[i]]];
u=Prepend[b,0];u=Append[u,0]];Return[u]];
```

In der ersten Abbildung 7.30 geben wir die Ergebnisse für $J = 10$, also $\Delta x = \pi/10$, und $\Delta t = 0.001$ an. Mit $\sigma := 1$ ist $\sigma\Delta t/(\Delta x)^2 \approx 0.01$. Es wird $\theta = 0,\ 0.5,\ 1.0$ benutzt. Man erkennt praktisch keine Unterschiede. In Abbildung 7.31 geben wir die entspre-

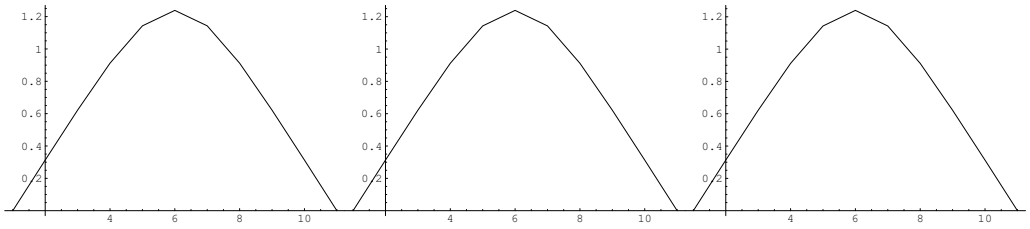

Abbildung 7.30: $\kappa = \sigma\Delta t/(\Delta x)^2 \approx 0.01$, $\theta = 0.0,\ 0.5,\ 1.0$

chenden Ergebnisse für $J = 100$ wieder. Hier ist $\sigma\Delta t/(\Delta x)^2 \approx 1.01$. Man erkennt, dass

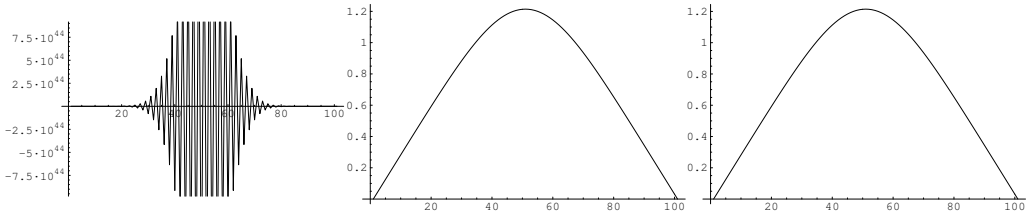

Abbildung 7.31: $\kappa = \sigma\Delta t/(\Delta x)^2 \approx 1.01$, $\theta = 0.0,\ 0.5,\ 1.0$

das explizite Verfahren katastrophale Ergebnisse liefert, während die beiden impliziten Verfahren gute Werte liefern. Den letzten Plot haben wir z. B. durch

```
u=heat[1.0,100,0.001,100,1.0];
ListPlot[u,PlotJoined->True]
```

erzeugt. Obwohl die Ortsschrittweite $\Delta x$ im zweiten Fall kleiner ist als im ersten, ist das explizite Verfahren in diesem Falle unbrauchbar.

5. Zur Lösung der Wärmeleitungsgleichung $u_t = \sigma u_{xx}$ in $(0,\pi)\times(0,\infty)$ mit der Anfangsbedingung $u(x,0) = \phi(x)$ in $[0,\pi]$ und den Randbedingungen $u(0,t) = u_x(\pi,t) = 0$ für $t > 0$ bestimme man mit der Methode der Trennung der Variablen und dem Superpositionsprinzip einen Lösungskandidaten.

   **Lösung:** Man mache den Lösungsansatz $u(x,t) = X(x)T(t)$. Einsetzen in die Wärmeleitungsgleichung liefert wieder, dass

   $$\frac{1}{\sigma}\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} = -\lambda$$

   mit einer Konstanten $\lambda$. Daher sind $X$ und $T$ aus

   $$X'' + \lambda X = 0, \qquad X(0) = X'(\pi) = 0$$

   und

   $$T' = -\sigma\lambda T$$

   zu berechnen. Die erste Gleichung besitzt für $\lambda \le 0$ keine nichttriviale Lösung, der Schluss hierfür ist genau wie für die Randbedingungen $X(0) = X(\pi) = 0$. Für $\lambda > 0$ ist die allgemeine Lösung von $X'' + \lambda X = 0$ durch

   $$X(x) = c_1 \sin\sqrt{\lambda}x + c_2 \cos\sqrt{\lambda}x$$

   gegeben. Aus der Randbedingung $X(0) = 0$ erhält man $c_2 = 0$. Für eine nichttriviale Lösung ist $c_1 \neq 0$. Die zweite Randbedingung $X'(\pi) = 0$ ergibt die Bedingung $\cos\sqrt{\lambda}\pi = 0$ bzw. $\sqrt{\lambda} = (2k+1)/2$ mit $k = 0, 1, \ldots$. Definiert man also

   $$u_k(x,t) := \sin\left(\frac{2k+1}{2}x\right)\exp\left(-\sigma\frac{(2k+1)^2}{4}t\right), \qquad k = 0, 1, \ldots,$$

   so ist $u_k$ eine nichttriviale Lösung der Wärmeleitungsgleichung, welche den homogenen Randbedingungen genügt. Wegen

   $$\int_0^\pi \sin\left(\frac{2k+1}{2}x\right)\sin\left(\frac{2l+1}{2}x\right)dx = \frac{\pi}{2}\delta_{kl}$$

   ist

   $$u(x,t) = \sum_{k=0}^{\infty} c_k \exp\left(-\sigma\frac{(2k+1)^2}{4}t\right)$$

   mit

   $$c_k := \frac{2}{\pi}\int_0^\pi \phi(x)\sin\left(\frac{2k+1}{2}x\right)dx$$

   ein Lösungskandidat für die gestellte Aufgabe.

6. Mit Hilfe von Aufgabe 5 bestimme man einen Lösungskandidaten für $u_t = u_{xx}$ in $(0,\pi)\times(0,\infty)$ mit $u(x,0) = x^2$ in $(0,\pi)$ und $u(0,t) = u_x(\pi,t) = 0$ für $t > 0$. Anschließend mache man einen dreidimensionalen Plot des Lösungskandidaten über $[0,\pi]\times[0,1]$.

   **Lösung:** Gibt man in *Mathematica* den Befehl

```
Integrate[x^2*Sin[(2k+1)*x/2],{x,0,Pi}];
Simplify[%]
```

so erhält man

$$\frac{2(-8+4(1+2k)\pi\cos k\pi+(-8+\pi^2+4k\pi^2+4k^2\pi^2)\sin k\pi)}{(1+2k)^3}.$$

Daher ist

$$c_k=\frac{2}{\pi}\int_0^\pi x^2\sin\left(\frac{2k+1}{2}x\right)dx=\frac{16}{\pi}\left[\frac{-2+(1+2k)\pi(-1)^k}{(1+2k)^3}\right],\qquad k=0,1,\ldots.$$

Mit Hilfe von

```
u10[x_,t_]:=(16/Pi)*
    Sum[(-2+(1+2j)*Pi*(-1)^j)/(1+2j)^3*Sin[(2j+1)*x/2]
    *Exp[-(2j+1)^2*t/4],{j,0,10}];
Plot3D[u10[x,t],{x,0,Pi},{t,0,1}]
    AxesLabel->{"Ort","Zeit","Temperatur"}]
```

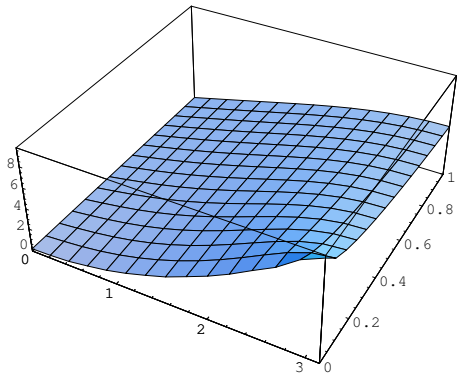


Abbildung 7.32: Temperaturverteilung auf $[0,\pi]\times[0,1]$

erhält man die in Abbildung 7.32 dagestellte Temperaturverteilung.

## 7.6.2 Aufgaben zu Abschnitt 6.2

1. Seien $\phi\in C^2(\mathbb{R})$, $\psi\in C^1(\mathbb{R})$ und $c>0$ gegeben. Man zeige, dass durch

$$u(x,t):=\frac{1}{2}[\phi(x-ct)+\phi(x+ct)]+\frac{1}{2c}\int_{x-ct}^{x+ct}\psi(\xi)\,d\xi$$

eine Lösung der Wellengleichung $u_{tt}=c^2u_{xx}$ auf $\mathbb{R}\times[0,\infty)$ gegeben ist, welche den Anfangsbedingungen

$$u(x,0)=\phi(x),\qquad u_t(x,0)=\psi(x)$$

auf $\mathbb{R}$ genügt.

**Lösung:** Die Behauptung folgt natürlich durch einfaches Nachrechnen. Wir führen die Rechnung trotzdem durch, wobei die Kettenregel wichtigstes Hilfsmittel ist. Die erste Anfangsbedingung $u(x,0) = \phi(x)$ gilt offensichtlich. Weiter ist

$$u_t(x,t) = \frac{c}{2}[-\phi'(x-ct) + \phi'(x+ct)] + \frac{1}{2}[\psi(x+ct) + \psi(x-ct)]$$

und insbesondere $u_t(x,0) = \psi(x)$. Erneutes Differenzieren liefert

$$u_{tt}\frac{c^2}{2}[\phi''(x-ct) + \phi''(x+ct)] + \frac{c}{2}[\psi'(x+ct) - \psi'(x-ct)].$$

Andererseits ist

$$u_x(x,t) = \frac{1}{2}[\phi'(x-ct) + \phi'(x+ct)] + \frac{1}{2c}[\psi(x+ct) - \psi(x-ct)]$$

und anschließend

$$u_{xx}(x,t) = \frac{1}{2}[\phi''(x-ct) + \phi''(x+ct)] + \frac{1}{2c}[\psi'(x+ct) - \psi'(x-ct)].$$

Damit genügt $u$ der Wellengleichung und die Behauptung ist bewiesen.

2. Seien $\phi \in C^2[0,\pi]$, $\psi \in C^1[0,\pi]$ vorgegebene Funktionen mit

$$\phi(0) = \phi(\pi) = \psi(0) = \psi(\pi) = \phi''(0) = \phi''(\pi) = 0,$$

ferner sei $c > 0$ gegeben. Man gewinne $\Phi, \Psi\colon \mathbb{R} \longrightarrow \mathbb{R}$ dadurch, dass man $\phi$ bzw. $\psi$ ungerade auf $[-\pi,\pi]$ und anschließend $2\pi$-periodisch auf ganz $\mathbb{R}$ fortsetzt. Man zeige, dass durch

$$u(x,t) := \frac{1}{2}[\Phi(x-ct) + \Phi(x+ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} \Psi(\xi)\,d\xi$$

eine Lösung der Wellengleichung $u_{tt} = c^2 u_{xx}$ auf $(0,\pi)\times(0,\infty)$ gegeben ist, die den homogenen Randbedingungen $u(0,t) = u(\pi,t) = 0$ auf $(0,\infty)$ und auf $(0,\pi)$ den Anfangsbedingungen $u(x,0) = \phi(x)$ und $u_t(x,0) = \psi(x)$ genügt.

**Lösung:** Zunächst beachten wir, dass wegen der Voraussetzungen an $\phi, \psi$ offenbar $\Phi \in C^2(\mathbb{R})$ und $\Psi \in \mathbb{C}^1(\mathbb{R})$. Wegen Aufgabe 1 ist die Angegebene Funktion $u$ eine Lösung der Wellengleichung, welche den Anfangsbedingung $u(x,0) = \Phi(x)$ und $u_t(x,0) = \Psi(x)$ auf $\mathbb{R}$ genügt. Da $\Phi$ bzw. $\Psi$ Fortsetzungen von $\phi$ bzw. $\psi$ von $[0,\pi]$ auf $\mathbb{R}$ sind, genügt $u$ trivialerweise den Anfangsbedingungen. Weiter ist

$$u(0,t) = \frac{1}{2}[\Phi(-ct) + \Phi(ct)] + \frac{1}{2}\int_{-ct}^{ct} \Psi(\xi)\,d\xi = 0,$$

da $\Phi$ und $\Psi$ ungerade Funktionen sind. Schließlich ist

$$u(\pi,t) = \frac{1}{2}[\Phi(\pi-ct) + \Phi(\pi+ct)] + \frac{1}{2}\int_{\pi-ct}^{\pi+ct} \Psi(\xi)\,d\xi = 0,$$

da $\Phi$ und $\Psi$ ungerade und $2\pi$-periodisch sind. Damit ist die Aufgabe gelöst.

3. Mit Hilfe der Methode der Fourierreihen bestimme man einen Lösungskandidaten für das Anfangsrandwertproblem

$$\begin{cases} u_{tt} = u_{xx} \quad \text{auf } (0,\pi)\times(0,\infty), \qquad u(0,t) = u(\pi,t) = 0 \quad \text{auf } (0,\infty), \\ \qquad\qquad u(x,0) = 1, \quad u_t(x,0) = 0 \quad \text{auf } (0,\pi). \end{cases}$$

**Lösung:** Als Fourier-Entwicklung einer Lösung erhält man offenbar

$$u(x,t) = \sum_{k=1}^{\infty} a_k \sin kx \cos kt$$

mit

$$a_k := \frac{2}{\pi}\int_0^\pi \sin kx\, dx = \begin{cases} 0, & k \text{ gerade}, \\ \dfrac{4}{k\pi}, & k \text{ ungerade}, \end{cases},$$

also

$$u(x,t) = \frac{4}{\pi}\sum_{j=0}^{\infty} \frac{1}{2j+1}\sin(2j+1)x\cos(2j+1)t.$$

In Abbildung 7.33 plotten wir die ersten 21 bzw. 51 Summanden der Darstellung von $u(x,0)$. Außerdem haben wir einen 3D-Plot von $u(x,t)$ (mit 51 Summanden) auf $[0,\pi]\times$

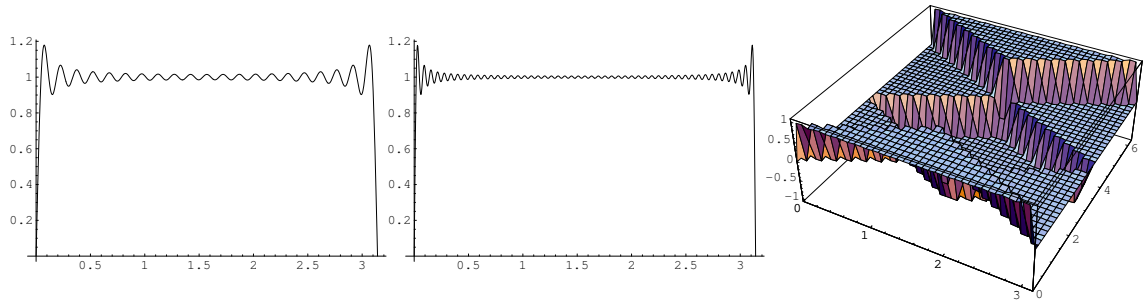

Abbildung 7.33: Lösung der Wellengleichung?

$[0,2\pi]$ gemacht. Ist $u$ eine Lösung?

4. Man bestimme eine Lösung der Wellengleichung $u_{tt} = u_{xx}$ in $(0,\pi)\times(0,\infty)$, welche den homogenen Anfangsbedingungen $u(x,0) = u_t(x,0)$ in $[0,\pi]$ und den Randbedingungen $u(0,t) = 0$ und $u_x(\pi,t) = 1$ auf $(0,\infty)$ genügt.

Hinweis: Man bestimme $v(x,t) := u(x,t) - x$ durch die Methode der Trennung von Variablen und das Superpositionsprinzip.

**Lösung:** Sei $u$ die gesuchte Lösung und $v(x,t) := u(x,t) - x$. Dann genügt $v$ ebenfalls der Wellengleichung, ferner den Anfangsbedingungen $v(x,0) = -x$, $v_t(x,0) = 0$ und den homogenen Randbedingungen $v(0,t) = v_x(\pi,t) = 0$. Der Ansatz $v(x,t) = X(x)T(t)$ führt auf

$$X'' + \lambda X = 0, \qquad X(0) = X'(\pi) = 0$$

und

$$T'' + \lambda T = 0.$$

Die erste Gleichung hat, wie man sich leicht überlegt, nichttriviale Lösungen nur für $\lambda = (k + \frac{1}{2})^2$, nämlich Vielfache von $X_k(x) := \sin(k + \frac{1}{2})x$, $k = 0, 1, \ldots$. Damit erhält man als Lösung der zweiten Gleichung

$$T_k(t) = a_k \cos(k + \tfrac{1}{2})t + b_k \sin(k + \tfrac{1}{2})t.$$

Als Lösungskandidaten hat man also

$$v(x,t) = \sum_{k=0}^{\infty} \sin(k + \tfrac{1}{2})x[a_k \cos(k + \tfrac{1}{2})t + b_k \sin(k + \tfrac{1}{2})t].$$

Wegen der Anfangsbedingung $v_t(x, 0) = 0$ ist $b_k = 0$, $k = 0, 1, \ldots$. Da ferner

$$\begin{aligned} \int_0^\pi \sin(k + \tfrac{1}{2})x \sin(l + \tfrac{1}{2})x \, dx &= \frac{1}{2} \int_0^\pi [\cos(k - l)x - \cos(k + l + 1)x] \, dx \\ &= \frac{\pi}{2} \delta_{kl}, \end{aligned}$$

ist

$$a_k = -\frac{2}{\pi} \int_0^\pi x \sin(k + \tfrac{1}{2})x \, dx$$

wegen der Anfangsbedingung $v(x, 0) = -x$. Nach

```
Integrate[x*Sin[(k+1/2)*x],{x,0,Pi}]
Simplify[%]
```

erhält man als Output

$$\frac{4\cos(k\pi) + 2(1 + 2k)\pi \sin(k\pi))}{(1 + 2k)^2}.$$

Berücksichtigt man noch, dass $k$ eine ganze Zahl ist, so erhält man

$$a_k = -\frac{2}{\pi} \frac{(-1)^k}{(k + \frac{1}{2})^2}.$$

Daher ist

$$u(x,t) = x - \frac{2}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{(k + \frac{1}{2})^2} \sin(k + \tfrac{1}{2})x \cos(k + \tfrac{1}{2})t.$$

In Abbildung 7.34 geben wir $u(x, 0)$ auf $[0, \pi]$ und $u(x, t)$ auf $[0.\pi] \times [0, 4\pi]$ an, wobei die ersten 51 Summanden in der Fourierentwicklung berücksichtigt worden.

5. Sei $m$ eine nichtnegative ganze Zahl. Für eine Lösung der Besselschen Differentialgleichung

$$z^2 Q'' + zQ' + (z^2 - m^2)Q = 0,$$

welche für $z = 0$ endlich ist, mache man einen Potenzreihenansatz

$$Q(z) = \sum_{n=0}^{\infty} a_n z^{k+n}$$

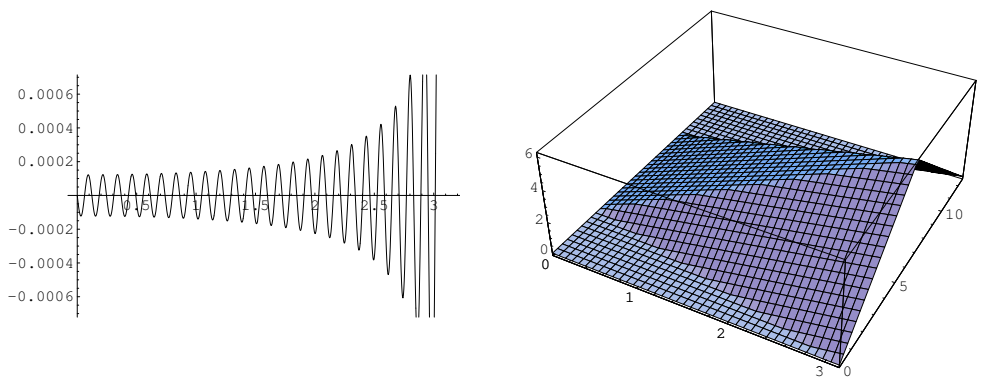


Abbildung 7.34: Lösung der Wellengleichung mit inhomogenen Randdaten

mit einer noch unbekannten nichtnegativen ganzen Zahl $k$ und unbekannten Koeffizienten $a_0 \neq 0, a_1, \ldots$. Man bestimme $k$ und $a_0, a_1, \ldots$, letztere bis auf ein gemeinsames Vielfaches.

**Lösung:** Einsetzen des Ansatzes in die Differentialgleichung liefert

$$\begin{aligned} z^2Q''(z) + zQ'(z) + (z^2 - m^2)Q(z) &= z^2 \sum_{n=0}^{\infty} (k+n)(k+n-1)a_n z^{k+n-2} \\ &\quad + z \sum_{n=0}^{\infty} (k+n)a_n z^{k+n-1} \\ &\quad + z^2 \sum_{n=0}^{\infty} a_n z^{k+n} - m^2 \sum_{n=0}^{\infty} a_n z^{k+n} \\ &= \sum_{n=0}^{\infty} (k+n)(k+n-1)a_n z^{k+n} \\ &\quad + \sum_{n=0}^{\infty} (k+n)a_n z^{k+n} \\ &\quad + \sum_{n=2}^{\infty} a_{n-2} z^{k+n} - m^2 \sum_{n=0}^{\infty} a_n z^{k+n}. \end{aligned}$$

Damit $Q$ eine Lösung der Besselschen Differentialgleichung ist, muss der Koeffizient von $z^k$ verschwinden, was auf

$$[k(k-1) + k - m^2]a_0 = 0$$

bzw. $k = m$ führt, da $a_0 \neq 0$. Dann ist aber

$$\begin{aligned} & z^2Q''(z) + zQ'(z) + (z^2 - m^2)Q(z) \\ = & \sum_{n=0}^{\infty} [(m+n)(m+n-1) + (m+n) - m^2]a_n z^{m+n} + \sum_{n=2}^{\infty} a_{n-2} z^{m+n} \\ = & \sum_{n=0}^{\infty} n(2m+n)a_n z^{m+n} + \sum_{n=2}^{\infty} a_{n-2} z^{m+n} \end{aligned}$$

$$= (2m+1)a_1 z^{m+1} + \sum_{n=2}^{\infty} [n(2m+n)a_n + a_{n-2}] z^{m+n}.$$

Also ist $a_1 = 0$ und

$$a_n = -\frac{a_{n-2}}{n(2m+n)}, \qquad n = 2, \ldots.$$

Hieraus folgt, dass die Koeffizienten mit einem ungeraden Index verschwinden. Setzt man $n = 2l$ und $a'_l = a_{2l}$, so hat man die Rekursionsformel

$$a'_l = -\frac{1}{2^2 l(m+l)} a'_{l-1}, \qquad l = 1, \ldots,$$

was durch vollständige Induktion sehr leicht auf

$$a'_l = \frac{(-1)^l m!}{2^{2l} l!\,(m+l)!} a_0, \qquad l = 0, \ldots,$$

führt. Also ist

$$Q(z) = a_0 \sum_{l=0}^{\infty} \frac{(-1)^l m!}{2^{2l} l!\,(m+l)!} z^{m+2l}$$

eine Lösung der Besselschen Differentialgleichung, die für $z = 0$ endlich ist (die Reihe ist offenbar "beliebig gut" konvergent).

Mit

$$a_0 := \frac{1}{2^m m!}$$

erhält für $m \in \mathbb{N}_0$ die Besselfunktion $J_m$, es ist also

$$J_m(z) = \sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2l} l!(m+l)!} z^{m+2l}.$$

6. Für $m \in \mathbb{N}_0$ kann die Besselfunktion $J_m$ durch

$$J_m(z) = \sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2l} l!(m+l)!} z^{m+2l}$$

definiert werden. Man zeige, dass

$$\frac{d}{dz}[z^m J_m(z)] = z^m J_{m-1}(z), \qquad m \in \mathbb{N}$$

und

$$J'_0(z) = -J_1(z).$$

**Lösung:** Es ist

$$\begin{aligned}
\frac{d}{dz}[z^m J_m(z)] &= \frac{d}{dz}\left[\sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2l} l!(m+l)!} z^{2m+2l}\right] \\
&= \sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2l} l!(m+l)!} 2(m+l) z^{2m+2l-1} \\
&= z^m \sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m-1+2l} l!(m-1+l)!} z^{m-1+2l} \\
&= z^m J_{m-1}(z),
\end{aligned}$$

womit die erste Behauptung schon bewiesen ist. Benutzt man die angegebene Reihenentwicklung, so ist

$$\begin{aligned}
J_0'(z) &= \sum_{l=0}^{\infty} \frac{(-1)^l}{2^{2l} l! l!} 2l z^{2l-1} \\
&= \sum_{l=1}^{\infty} \frac{(-1)^l}{2^{2l} l! l!} 2l z^{2l-1} \\
&= \sum_{l=1}^{\infty} \frac{(-1)^l}{2^{2l-1} l! (l-1)!} z^{2l-1} \\
&= \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{2^{1+2k} (k+1)! k!} z^{1+2k} \\
&= -\sum_{k=0}^{\infty} \frac{(-1)^k}{2^{1+2k} (k+1)! k!} z^{1+2k} \\
&= -J_1(z).
\end{aligned}$$

Damit ist auch die zweite Behauptung bewiesen.

7. Für $m \in \mathbb{N}$ kann die Ableitung $J_m'$ der $m$-ten Besselfunktion $J_m$ erster Art nach einer der folgenden Formeln berechnet werden:

$$J_m'(z) = \begin{cases} -J_{m+1}(z) + \dfrac{m}{z} J_m(z), \\ J_{m-1}(z) - \dfrac{m}{z} J_m(z), \\ \dfrac{1}{2}[J_{m-1}(z) - J_{m+1}(z)]. \end{cases}$$

**Lösung:** Mit *Mathematica* könnte man es sich einfach machen. Z. B. erhält man nach der Eingabe

```
D[BesselJ[m,z],z]==(1/2)*(BesselJ[m-1,z]-BesselJ[m+1,z])
```

die Auskunft `True`. Seltsamerweise (wer weiß eine Erklärung?) klappt dies für die ersten beiden Aussagen nicht. Auf die Eingabe

```
D[BesselJ[m,z],z]==-BesselJ[m+1,z]+(m/z)*BesselJ[m,z]
```

erhält man z. B. als Reaktion die Gleichung (in Formeln geschrieben)

$$\frac{1}{2}(J_{m-1}(z) - J_{m+1}(z)) = \frac{m J_m(z)}{z} - J_{m+1}(z),$$

was wohl so zu verstehen ist, dass die behauptete Gleichung genau dann richtig ist, wenn es die ausgegebene ist. Auch nach der Aufforderung `Simplify[%]` erhält man nur die Gleichung

$$\frac{1}{2} J_{m-1}(z) - \frac{m J_m(z)}{z} + \frac{1}{2} J_{m+1}(z) = 0,$$

die richtig ist, aber nicht als wahr erkannt wird. Daher beweisen wir jetzt zunächst die ersten beiden Aussagen. Es ist

$$\begin{aligned}
J'_m(z)+J_{m+1}(z)-\frac{m}{z}J_m(z) &= \sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m+l)!}(m+2l)z^{m+2l-1} \\
&\quad +\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+1+2l}l!(m+1+l)!}z^{m+1+2l} \\
&\quad -\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m+l)!}mz^{m+2l-1} \\
&= \sum_{l=1}^{\infty}\frac{(-1)^l}{2^{m+2l-1}(l-1)!(m+l)!}z^{m+2l-1} \\
&\quad +\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+1+2l}l!(m+1+l)!}z^{m+1+2l} \\
&= 0,
\end{aligned}$$

wie man sofort erkennt, wenn man in der ersten Summe den Summationsindex $l$ durch $l+1$ ersetzt. Damit ist die erste Gleichung bewiesen. Die zweite erhält man entsprechend aus

$$\begin{aligned}
J'_m(z)-J_{m-1}(z)+\frac{m}{z}J_m(z) &= \sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m+l)!}(m+2l)z^{m+2l-1} \\
&\quad -\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m-1+2l}l!(m-1+l)!}z^{m-1+2l} \\
&\quad +\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m+l)!}mz^{m+2l-1} \\
&= \sum_{l=0}^{\infty}\frac{(-1)^l z^{m-1+2l}}{2^{m+2l}l!(m-1+l)!}\Bigg[\underbrace{\frac{m+2l}{m+l}-2+\frac{m}{m+l}}_{=0}\Bigg] \\
&= 0.
\end{aligned}$$

Damit ist auch die zweite Gleichung bewiesen. Auch die dritte Gleichung kann man leicht beweisen (oder *Mathematica* vertrauen). Es ist nämlich

$$\begin{aligned}
J'_m(z)-\frac{1}{2}[J_{m-1}(z)-J_{m+1}(z)] &= \sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m+l)!}(m+2l)z^{m+2l-1} \\
&\quad -\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m-1+l)!}z^{m-1+2l} \\
&\quad +\sum_{l=0}^{\infty}\frac{(-1)^l}{2^{m+2+2l}l!(m+1+l)!}z^{m+1+2l} \\
&= \sum_{l=1}^{\infty}\frac{(-1)^l}{2^{m+2l}l!(m+l)!}(m+2l)z^{m+2l-1}
\end{aligned}$$

$$
\begin{aligned}
& -\sum_{l=1}^{\infty} \frac{(-1)^l}{2^{m+2l} l!(m-1+l)!} z^{m-1+2l} \\
& +\sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2+2l} l!(m+1+l)!} z^{m+1+2l} \\
= & -\sum_{l=0}^{\infty} \frac{(-1)^l (m+2l+2)}{2^{m+2l+2}(l+1)!(m+l+1)!} z^{m+2l+1} \\
& +\sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2l+2}(l+1)!(m+l)!} z^{m+2l+1} \\
& +\sum_{l=0}^{\infty} \frac{(-1)^l}{2^{m+2+2l} l!(m+1+l)!} z^{m+1+2l} \\
= & \ 0,
\end{aligned}
$$

womit auch die dritte Gleichung bewiesen ist.

8. Man zeige, dass

$$\frac{d}{dz}\left[\frac{1}{2}z^2(J_0^2(z)+J_1^2(z))\right] = zJ_0^2(z)$$

und

$$\frac{d}{dz}\left[\frac{1}{2}z^2(J_m^2(z)-J_{m-1}(z)J_{m+1}(z))\right] = zJ_m^2(z), \qquad m \in \mathbb{N}.$$

Mit Hilfe dieser Beziehungen berechne man

$$c_{mn} := \int_0^1 zJ_m^2(z_{mn}z)\,dz,$$

wobei $z_{mn}$ die $n$-te positive Nullstelle von $J_m$ ist.

**Lösung:** Mit Hilfe von Aufgabe 6 ist

$$
\begin{aligned}
\frac{d}{dz}\left[\frac{1}{2}z^2(J_0^2(z)+J_1^2(z))\right] &= z[J_0^2(z)+J_1^2(z)] + \frac{1}{2}z^2[2J_0(z)J_0'(z)+2J_1(z)J_1')z)] \\
&= zJ_0^2(z)+zJ_1^2(z)+z^2[J_0(z)J_0'(z)+J_1(z)J_1'(z)] \\
&= zJ_0^2(z)+zJ_1^2(z)+z^2[-J_0(z)J_1(z)+J_1(z)J_1'(z)] \\
&= zJ_0^2(z)+zJ_1(z)[J_1(z)-zJ_0(z)+zJ_1'(z)] \\
&= zJ_0^2(z)+zJ_1(z)[\underbrace{J_1(z)-(zJ_1(z))'+zJ_1'(z)}_{=0}] \\
&= zJ_0^2(z).
\end{aligned}
$$

Weiter ist mit Hilfe von Aufgabe 7

$$
\begin{aligned}
& \frac{d}{dz}\left[\frac{1}{2}z^2(J_m^2(z)-J_{m-1}(z)J_{m+1}(z))\right] \\
= & \ z[J_m^2(z)-J_{m-1}(z)J_{m+1}(z)] \\
& \quad +\frac{1}{2}z^2[2J_m(z)J_m'(z)-J_{m-1}'(z)J_{m+1}(z)-J_{m-1}(z)J_{m+1}'(z)] \\
= & \ zJ_m^2(z)-zJ_{m-1}(z)J_{m+1}(z)+z^2J_m(z)J_m'(z)
\end{aligned}
$$

$$
\begin{aligned}
& -\frac{1}{2}z^2[J'_{m-1}(z)J_{m+1}(z)+J_{m-1}(z)J'_{m+1}(z)] \\
= \; & zJ_m^2(z)-zJ_{m-1}(z)J_{m+1}(z)+z^2J_m(z)\frac{1}{2}[J_{m-1}(z)-J_{m+1}(z)] \\
& -\frac{1}{2}z^2\left[-J_m(z)+\frac{m-1}{z}J_{m-1}(z)\right]J_{m+1}(z) \\
& -\frac{1}{2}z^2J_{m-1}(z)\left[J_m(z)-\frac{m+1}{z}J_{m+1}(z)\right] \\
= \; & zJ_m^2(z).
\end{aligned}
$$

Schließlich ist

$$
c_{0n}=\int_0^1 zJ_0^2(z_{0n}z)\,dz=\frac{1}{z_{0n}^2}\int_0^{z_{0n}} zJ_0^2(z)\,dz=\frac{1}{2z_{0n}^2}z^2[J_0^2(z)+J_1^2(z)]\Big|_0^{z_{0n}}=\frac{J_1^2(z_{0n})}{2}
$$

und für $m\in\mathbb{N}$ entsprechend

$$
\begin{aligned}
c_{mn} &= \int_0^1 zJ_m^2(z_{mn}z)\,dz \\
&= \frac{1}{z_{mn}^2}\int_0^{z_{mn}} zJ_m^2(z)\,dz \\
&= \frac{1}{2z_{mn}^2}z^2[J_m^2(z)-J_{m-1}(z)J_{m+1}(z)]\Big|_0^{z_{mn}} \\
&= -\frac{J_{m-1}(z_{mn})J_{m+1}(z_{mn})}{2}.
\end{aligned}
$$

Damit ist die Aufgabe gelöst.

9. Die folgende Aufgabe beschäftigt sich zwar nicht mit der Wellengleichung, sondern mit der Wärmeleitungsgleichung, zu ihrer Bearbeitung werden aber Besselfunktionen benötigt, so dass die Aufgabe erst jetzt gestellt wird.

   Sei $\Omega:=\{(x,y)\in\mathbb{R}^2: x^2+y^2<1\}$ die Einheitskreisscheibe mit dem Rand $\partial\Omega$. Gesucht sei eine Lösung der Wärmeleitungsgleichung $u_t=u_{xx}+u_{yy}$ auf $\Omega\times(0,\infty)$, welche der Randbedingung $u(x,y,t)=1$ für $(x,y)\in\partial\Omega$ und $t>0$ und der Anfangsbedingung $u(x,y,0)=0$ für $(x,y)\in\Omega$ genügt. Man transformiere das Problem durch $v:=1-u$ zunächst auf homogene Randdaten, führe eine Trennung der Veränderlichen durch, führe Polarkoordinaten ein und benutze das Superpositionsprinzip.

   **Lösung:** Mit einer Lösung $u$ der gestellten Aufgabe setzen wir $v:=1-u$. Dann genügt $v$ ebenfalls der Wärmeleitungsgleichung, hat homogene Randbedingungen und die Anfangsverteilung $v(x,y,0)=1$ für $(x,y)\in\Omega$. Wir machen den Ansatz $v(x,y)=Z(x,y)T(t)$, was auf

$$
Z_{xx}+Z_{yy}+\lambda Z=0 \qquad \text{auf } \Omega, \qquad Z=0 \quad \text{auf } \partial\Omega
$$

   und

$$
T'+\lambda T=0
$$

   führt. Man mache den rotationssymmetrischen Ansatz $Z(r\cos\theta,r\sin\theta)=R(r)$. Wie bei der Wellengleichung erhält man nach der Variablentransformation $z=\sqrt{\lambda}r$, dass

$R(r) = J_0(\sqrt{\lambda}r)$ und $\lambda = z_{0n}^2$ das Quadrat eine Nullstelle der Besselfunktion $J_0$ ist. Als Lösungskandidaten haben wir daher bisher

$$u(x,y,t) = 1 - v(x,y,t) = 1 - \sum_{n=1}^{\infty} A_n J_0(z_{0n}r)e^{-z_{0n}^2 t}.$$

Nun sind die Koeffizienten $A_1, A_2, \ldots$ noch so zu bestimmen, dass die Anfangsbedingung erfüllt ist, also

$$\sum_{n=1}^{\infty} A_n J_0(z_{0n}r) = 1, \qquad r \in (0,1),$$

gilt. Wegen der Orthogonalität der Besselfunktionen ist

$$A_n = \int_0^1 rJ_0(z_{0n}r)\,dr \;\Big/\; \int_0^1 rJ_0^2(z_{0n}r)\,dr.$$

Unter Benutzung von Aufgabe 6 erhält man

$$\int_0^1 rJ_0(z_{0n}r)\,dr = \frac{1}{z_{0n}^2}\int_0^{z_{0n}} rJ_0(r)\,dr = \frac{1}{z_{0n}^2}\int_0^{z_{0n}} \frac{d}{dr}[rJ_1(r)]\,dr = \frac{J_1(z_{0n})}{z_{0n}}.$$

Entsprechend ist (siehe auch Aufgabe 7)

$$\int_0^1 rJ_0^2(z_{0n}r)\,dr = \frac{1}{z_{0n}^2}\int_0^{z_{0n}} rJ_0^2(r)\,dr = \frac{1}{2z_{0n}^2}r^2[J_0^2(r)+J_1^2(r)]\Big|_0^{z_{0n}} = \frac{J_1^2(z_{0n})}{2}.$$

Insgesamt ist daher

$$A_n = \frac{2}{z_{0n}J_1(z_{0n})}.$$

In Abbildung 7.35 haben wir $u(r,t)$ auf $[0,1]\times[0,5]$ aufgetragen. Nach

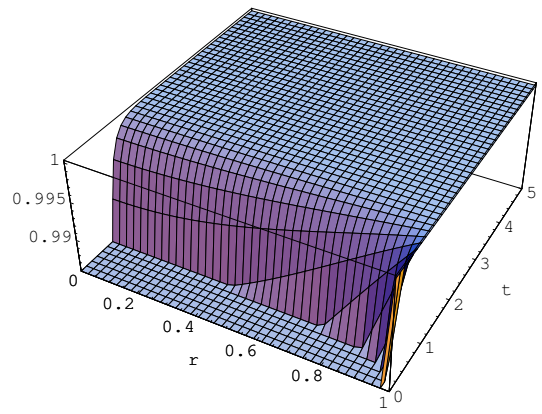


Abbildung 7.35: Eine rotationssymmetrische Lösung der Wellengleichung

```
Needs["NumericalMath`BesselZeros`"]
```

haben wir dies durch

```
z0n=BesselJZeros[0,10];
u[r_,t_]:=
  1-2*Sum[BesselJ[0,z0n[[n]]*r]/(z0n[[n]]*BesselJ[1,z0n[[n]]])*
          Exp[-z0n[[n]]^2*t],{n,10}];
Plot3D[u[r,t],{r,0,1},{t,0,5},PlotPoints->40,AxesLabel->{"r","t",""}]
```

erreicht.

10. Gegeben sei die Anfangswertaufgabe für die Wellengleichung. Bei gegebenen $c > 0$, $\phi \in C^2(\mathbb{R})$ und $\psi \in C^1(\mathbb{R})$ sei also eine Lösung von $u_{tt} = c^2 u_{xx}$ in $\mathbb{R} \times (0, \infty)$ gesucht, welche den Anfangsbedingungen $u(x,0) = \phi(x)$, $u_t(x,0) = \psi(x)$ genügt.Man überziehe die obere $(x,t)$-Halbebene mit einem Gitter mit der Maschenweite $\Delta x > 0$ in $x$-Richtung und $\Delta t > 0$ in $t$-Richtung. Als Gitterpunkte hat man dann $(x_j, t_k) := (j\Delta x, k\Delta t)$, $(j,k) \in \mathbb{Z} \times \mathbb{N}_0$. Mit $u_j^k$ bezeichne man wieder eine Näherung für die Lösung in $(x_j, t_k)$. Zur Abkürzung setze man ferner

$$\phi_j := \phi(x_j), \qquad \psi_j := \psi(x_j), \qquad j \in \mathbb{Z}.$$

Man setze $r := c\Delta t/\Delta x$.

(a) Man begründe, wie man darauf kommt, die ersten beiden Zeit-Level $u_j^0$, $u_j^1$, $j \in \mathbb{Z}$, aus

$$u_j^0 := \phi_j, \quad u_j^1 := \phi_j + \Delta t\psi_j + \frac{r^2}{2}(\phi_{j+1} - 2\phi_j + \phi_{j-1}), \quad j \in \mathbb{Z},$$

zu berechnen. Insbesondere erläutere man, weshalb es besser ist, den Zeit-Level $u_j^1$ in der angegebenen Weise zu berechnen als durch

$$u_j^1 := \phi_j + \Delta t\psi_j, \qquad j \in \mathbb{Z}.$$

(b) Man erläutere, wodurch die Vorschrift, den $(k+1)$-ten Zeit-Level aus dem $k$-ten und $(k-1)$-ten Zeit-Level durch

$$u_j^{k+1} := 2(1-r^2)u_j^k + r^2(u_{j+1}^k + u_{j-1}^k) - u_j^{k-1}, \qquad j \in \mathbb{Z},$$

zu berechnen, motiviert werden kann.

(c) Die exakte Lösung in $(x_j, t_k)$ hängt von den Anfangswerten in $[x_j - ct_k, x_j + ct_k]$ (analytischer Abhängigkeitsbereich) ab. Man stelle fest, von welchen Anfangswerten die berechnete Lösung $u_j^k$ abhängt, bestimme also den sogenannten numerischen Abhängigkeitsbereich. Weshalb sollte der analytische im numerischen Abhängigkeitsbereich enthalten sein? Auf welche Forderung an die Schrittweiten führt dies? Wenn Sie diese Fragen beantwortet haben, haben Sie im wesentlichen die Courant-Friedrichs-Lewy Bedingung hergeleitet.

**Lösung:** Dass der erste Zeit-Level durch $u_j^0 = u(x_j, 0) = \phi(x_j) = \phi_j$ bestimmt wird, ist völlig klar. Bezeichnet man mit $u$ die exakte Lösung, so ist

$$u(x, \Delta t) \quad = \quad u(x,0) + \Delta t u_t(x,0) + \frac{(\Delta t)^2}{2} u_{tt}(x,0) + O((\Delta t)^3)$$

$$
\begin{aligned}
&= u(x,0) + \Delta t u_t(x,0) + c^2\frac{(\Delta t)^2}{2}u_{xx}(x,0) + O((\Delta t)^3)\\
&= u(x,0) + \Delta t u_t(x,0) + c^2\frac{(\Delta t)^2}{2}\phi''(x) + O((\Delta t)^3)\\
&= u(x,0) + \Delta t u_t(x,0)\\
&\qquad + c^2\frac{(\Delta t)^2}{2}\left[\frac{\phi(x-\Delta x) - 2\phi(x) + \phi(x+\Delta x)}{(\Delta x)^2} + O((\Delta x)^2)\right]\\
&\qquad + O((\Delta t)^3)\\
&= \phi(x) + \Delta t\psi(x) + \frac{r^2}{2}[\phi(x-\Delta x) - 2\phi(x) + \phi(x+\Delta x)]\\
&\qquad + O((\Delta t\Delta x)^2) + O((\Delta t)^3).
\end{aligned}
$$

Setzt man hier $x = x_j$, so kommt man auf das angegebene Verfahren zur Berechnung des Zeit-Levels zur Zeit $\Delta t$. Wegen $u(x,\Delta t) = u(x,0) + \Delta t u_t(x,0) + O((\Delta t)^2)$ ist das zweite, naive Verfahren ungünstiger.

Diskretisiert man in der Wellengleichung die zweiten Ableitungen jeweils durch den zentralen Differenzenquotienten, so erhält man

$$
\begin{aligned}
\frac{u(x,t-\Delta t) - 2u(x,t) + u(x,t+\Delta t)}{(\Delta t)^2} &= u_{tt}(x,t) + O((\Delta t)^2)\\
&= c^2 u_{xx}(x,t) + O((\Delta t)^2)\\
&= c^2\frac{u(x-\Delta x,t) - 2u(x,t) + u(x+\Delta x,t)}{(\Delta x)^2}\\
&\qquad + O((\Delta x)^2) + O((\Delta t)^2).
\end{aligned}
$$

Dies motiviert das Verfahren

$$
\frac{u_j^{k-1} - 2u_j^k + u_j^{k+1}}{(\Delta t)^2} = c^2\frac{u_{j-1}^k - 2u_j^k + u_{j+1}^k}{(\Delta x)^2}
$$

bzw.

$$
u_j^{k+1} := 2(1-r^2)u_j^k + r^2(u_{j-1}^k + u_{j+1}^k) - u_j^{k-1}.
$$

Die berechnete Näherung $u_j^1$ hängt von den Werten in

$$
[x_{j-1}, x_{j+1}] = [x_j - \Delta x, x_j + \Delta x] = \left[x_j - \frac{\Delta x}{\Delta t}t_1, x_j + \frac{\Delta x}{\Delta t}t_1\right]
$$

ab. Durch vollständige Induktion nach $k$ zeigt man, dass $u_j^k$, $j \in \mathbb{Z}$, von den Anfangswerten in $[x_j - (\Delta x/\Delta t)t_k, x_j + (\Delta x/\Delta t)t_k]$ abhängt. Für $k = 1$ ist dies richtig, wir nehmen an, es sei auch für $k$ richtig. Da $u_j^{k+1}$ eine Linearkombination von $u_{j-1}^k, u_j^k, u_{j+1}^k$ ist, hängt dieser Wert von den Anfangswerten in

$$
\left[x_{j-1} - \frac{\Delta x}{\Delta t}t_k, x_{j+1} + \frac{\Delta x}{\Delta t}t_k\right] = \left[x_j - \frac{\Delta x}{\Delta t}t_{k+1}, x_j + \frac{\Delta x}{\Delta t}t_{k+1}\right]
$$

ab, womit auch die Induktionsbehauptung bewiesen ist. Der numerische Abhängigkeitsbereich sollte den analytischen enthalten, da andernfalls in die numerische Berechnung

einer Näherung auch Anfangswerte eingehen, von der die exakte Lösung gar nicht abhängt oder es würden nicht alle die Lösung bestimmenden Anfangswerte in die Rechnung eingehen. Die Forderung, dass der analytische im numerischen Abhängigkeitsbereich enthalten ist, führt auf $c \leq \Delta x/\Delta t$ bzw.

$$r := c\frac{\Delta t}{\Delta x} \leq 1.$$

### 7.6.3 Aufgaben zu Abschnitt 6.3

1. Sei $\Omega := \{(x,y) \in \mathbb{R}^2 : 0 < x, y < 1\}$ das Einheitsquadrat, mit $\partial\Omega$ sei der Rand von $\Omega$ bezeichnet. Gegeben sei das Dirichlet-Problem für die Laplace-Gleichung, gesucht ist also $u \in C^2(\Omega) \cap C(\overline{\Omega})$ mit $\Delta u = 0$ in $\Omega$ und $u = f$ auf $\partial\Omega$ bei gegebenem $f \in C(\Omega)$. In Abbildung 7.36 ist ein Gitter zur Maschenweite $h = 1/4$ eingetragen, wobei die inneren

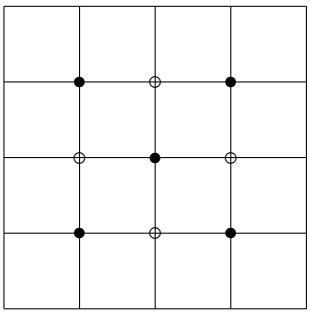

Abbildung 7.36: Schachbrettartiges Gitter

Gitterpunkte schachbrettartig in zwei Klassen zerlegt und deren Elemente durch $\bullet$ bzw. $\circ$ bezeichnet seien. Nach Diskretisierung mit dem 5-Punkt Differenzenstern stelle man das zugehörige lineare Gleichungssustem auf. Welche Aussagen können Sie über die Koeffizientenmatrix beweisen?

**Lösung:** Offenbar erhält man das lineare Gleichungssystem

$$\left(\begin{array}{rrrrr|rrrr}
4 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\
0 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\
0 & 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 4 & 0 & 0 & -1 & 0 & -1 \\
0 & 0 & 0 & 0 & 4 & 0 & 0 & -1 & -1 \\
\hline
-1 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & 0 \\
-1 & 0 & -1 & -1 & 0 & 0 & 4 & 0 & 0 \\
0 & -1 & -1 & 0 & -1 & 0 & 0 & 4 & 0 \\
0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 4
\end{array}\right)
\left(\begin{array}{c} u_{1,1} \\ u_{3,1} \\ u_{2,2} \\ u_{1,3} \\ u_{3,3} \\ \hline u_{2,1} \\ u_{1,2} \\ u_{3,2} \\ u_{2,3} \end{array}\right)
=
\left(\begin{array}{c} f_{1,0} + f_{0,1} \\ f_{3,0} + f_{4,1} \\ 0 \\ f_{0,3} + f_{1,4} \\ f_{4,3} + f_{3,4} \\ \hline f_{2,0} \\ f_{0,2} \\ f_{4,2} \\ f_{2,4} \end{array}\right).$$

Auch hier genügt die symmetrische Koeffizientenmatrix dem schwachen Zeilensummenkriterium und ist nicht zerlegbar, ferner sind die Diagonalelemente positiv und die Nichtdiagonalelemente nichtpositiv. Wegen Satz 3.2 ist $A$ nichtsingulär und positiv definit, ferner $A^{-1}$ nichtnegativ.

2. Eine Tridiagonalmatrix ist genau dann nicht zerfallend, wenn jedes ihrer Nebendiagonalelemente von Null verschieden ist.

   **Lösung:** Sei $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ eine Tridiagonalmatrix. Ist $a_{k,k+1} = 0$ für ein $k \in \{1,\dots,n-1\}$, so ist durch $N_1 := \{1,\dots,k\}$, $N_2 := \{k+1,\dots,n\}$ eine nichttriviale Zerlegung von $N := \{1,\dots,n\}$ gegeben mit $a_{ij} = 0$ für alle $(i,j) \in N_1 \times N_2$, also $A$ zerfallend. Ähnlich schließt man, wenn $a_{k+1,k} = 0$ für ein $k \in \{1,\dots,n-1\}$. Sei umgekehrt $A$ eine Tridiagonalmatrix mit von Null verschiedenen Nebendiagonalelementen. Ist $N_1, N_2$ eine nichttriviale Partition von $N$ mit $1 \in N_1$, so zeigt man durch vollständige Induktion nach $i$ sehr leicht, dass $\{1,\dots,i\} \subset N_1$. Dann ist aber $N_1 = N$ bzw. $N_2 = \emptyset$, ein Widerspruch.

3. Die Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ genüge dem *starken Zeilensummenkriterium*, d. h. es sei

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1,\dots,n.$$

   Man zeige:

   (a) Die Matrix $A$ ist nichtsingulär.

   (b) Mit der Iterationsmatrix $G := -A_D^{-1}(A_L + A_R)$ des Gesamtschrittverfahrens ist die affin lineare Abbildung $F(x) := Gx + c$ auf dem $\mathbb{R}^n$ kontrahierend bezüglich $\|\cdot\|_\infty$, d. h. es existiert eine Konstante $L < 1$ mit

$$\|F(x) - F(y)\|_\infty \le L\,\|x - y\|_\infty \qquad \text{für alle } x, y \in \mathbb{R}^n.$$

   Für $x = (x_j) \in \mathbb{R}^n$ ist hierbei $\|x\|_\infty$ definiert durch

$$\|x\|_\infty := \max_{j=1,\dots,n} |x_j|.$$

   **Lösung:** Der erste Teil kann fast genau wie der entsprechende Teil von Satz 3.2 bewiesen werden. Etwas einfacher sind die Verhältnisse hier, weil aus dem starken Zeilensummenkriterium sofort folgt, dass die Diagonalelemente von $A$ von Null verschieden sind. Wir schreiben den entsprechenden Teil aus dem Beweis von Satz 3.2 praktisch ab: Wir nehmen an, es gäbe einen vom Nullvektor verschiedenen Vektor $x$ aus dem Kern von $A$. O. B. d. A. ist $\|x\|_\infty = 1$. Für $i \in \{1,\dots,n\}$ mit $|x_i| = \|x\|_\infty$ ist

$$|a_{ii}| = |a_{ii}|\,|x_i| = \left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j\right| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\,|x_j| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|,$$

   ein Widerspruch. Damit ist nachgewiesen, dass eine Matrix, die dem starken Zeilensummenkriterium genügt, nichtsingulär ist.

   Nun sei $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ in der angegebenen Weise durch $F(x) := Gx + c$ definiert. Für beliebiges $x, y \in \mathbb{R}^n$ ist

$$\begin{aligned} \|F(x) - F(y)\|_\infty &= \|G(x-y)\|_\infty \\ &= \max_{i=1,\dots,n} |[-G(x-y)]_i| \end{aligned}$$

$$
\begin{aligned}
&= \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} (-g_{ij})(x_j - y_j) \right| \\
&= \max_{i=1,\dots,n} \left| \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} (x_j - y_j) \right| \\
&= \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \left| \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}(x_j - y_j) \right| \\
&\leq \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \, |x_j - y_j| \\
&\leq \left( \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right) \|x - y\|_\infty \\
&= L \, \|x - y\|_\infty,
\end{aligned}
$$

wobei

$$
L := \max_{i=1,\dots,n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < 1
$$

wegen des starken Zeilensummenkriteriums.

4. Gegeben sei das Dirichlet-Problem für die Laplace-Gleichung auf dem Einheitsquadrat $\Omega$, gesucht ist also $u \in C^2(\Omega) \cap C(\overline{\Omega})$ mit $\Delta u = 0$ und $u = f$ auf $\partial\Omega$ mit vorgegebenen Randdaten $f \in C(\partial\Omega)$. Mit der Maschenweite $h := 1/(n+1)$ diskretisiere man den Laplace-Operator mit Hilfe des 5 Punkt-Differenzensterns und gelange, wie in der Vorlesung beschrieben, zum linearen Gleichungssystem $Au = b$. Man zeige, dass

$$
\min_{(x,y)\in\partial\Omega} f(x,y) \leq u_{i,j} \leq \max_{(x,y)\in\partial\Omega} f(x,y), \qquad 1 \leq i,j \leq n.
$$

**Lösung:** Wir nehmen zunächst an, es sei $f(x,y) := 1$ für alle $(x,y) \in \partial\Omega$. Die Aussage der Aufgabe kann nur dann richtig sein, wenn $Ae = \hat{b}$, wobei $e$ der Vektor im $\mathbb{R}^{n^2}$ ist, dessen Komponenten sämtlich gleich 1 sind, und $\hat{b} \in \mathbb{R}^{n^2}$ durch

$$
\hat{b}^T := \begin{array}{ccccccccc}
(2 & , & 1 & , & \cdots & , & 1 & , & 2 & , \\
1 & , & 0 & , & \cdots & , & 0 & , & 1 & , \\
\vdots & , & \vdots & , & \cdots & , & \vdots & , & \vdots & , \\
1 & , & 0 & , & \cdots & , & 0 & , & 1 & , \\
2 & , & 1 & , & \cdots & , & 1 & , & 2) &
\end{array}
$$

gegeben ist. Beachtet man, dass $(Ae)_i$ die Summe der Elemente in der $i$-ten Zeile von $A$ ist, so stellt man unter Berücksichtigung der Struktur von $A$ leicht fest, dass dies richtig ist. Nun ist aber offensichtlich

$$
\left( \min_{(x,y)\in\partial\Omega} f(x,y) \right) \hat{b} \leq b = Au \leq \left( \max_{(x,y)\in\partial\Omega} f(x,y) \right) \hat{b}.
$$

Multiplikation dieser Ungleichungskette mit der nichtnegativen Matrix $A^{-1}$ liefert unter Berücksichtigung von $A^{-1}\hat{b} = e$ die Behauptung.

5. Für $n \in \mathbb{N}$ definiere man $A \in \mathbb{R}^{n^2 \times n^2}$ durch

$$A := \begin{pmatrix} A_n & -I_n & \cdots & 0 \\ -I_n & A_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & -I_n \\ 0 & \cdots & -I_n & A_n \end{pmatrix}$$

mit der Tridiagonalmatrix

$$A_n := \begin{pmatrix} 4 & -1 & \cdots & 0 \\ -1 & 4 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & -1 & 4 \end{pmatrix}$$

und der $n \times n$-Einheitsmatrix $I_n$. Weiter sei

$$H := \begin{pmatrix} H_n & 0 & \cdots & 0 \\ 0 & H_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & H_n \end{pmatrix}, \qquad V := \begin{pmatrix} V_n & -I_n & \cdots & 0 \\ -I_n & V_n & \ddots & \vdots \\ \vdots & \ddots & \ddots & -I_n \\ 0 & \cdots & -I_n & V_n \end{pmatrix}$$

mit

$$H_n := \begin{pmatrix} 2 & -1 & \cdots & 0 \\ -1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & -1 & 2 \end{pmatrix}, \qquad V_n := \begin{pmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2 \end{pmatrix}.$$

Man zeige:

(a) Die symmetrischen Matrizen $H$ und $V$ besitzen ein gemeinsames System von Eigenvektoren, nämlich die Vektoren $u^{p,q}$, $p, q = 1, \ldots, n$, mit

$$u_{i,j}^{p,q} := \sin \frac{ip\pi}{n+1} \sin \frac{jq\pi}{n+1}, \qquad i, j = 1, \ldots, n,$$

mit den zugehörigen Eigenwerten

$$\lambda_{pq} := 4 \sin^2 \frac{p\pi}{2(n+1)}, \qquad \mu_{pq} = 4 \sin^2 \frac{q\pi}{2(n+1)}, \qquad p, q = 1, \ldots, n.$$

(b) Ist $b \in \mathbb{R}^{n^2}$ vorgegeben, so konvergiert für einen beliebigen Startwert $u^0 \in \mathbb{R}^{n^2}$ und jedes $r > 0$ die durch die Vorschriften

$$(H + rI)u^{k+1/2} = (rI - V)u^k + b, \qquad (V + rI)u^{k+1} = (rI - H)u^{k+1/2} + b$$

gewonnene Folge $\{u^k\}$ gegen die eindeutige Lösung $u^*$ von $Au = b$.

**Lösung:** Der erste Teil kann durch einfaches Nachrechnen nachgewiesen werden, hierauf wollen wir nicht eingehen. Die angegebenen Vektoren $u^{p,q}$ sind linear unabhängig, weil sie vom Nullvektor verschieden sind und paarweise aufeinander senkrecht stehen. Denn für $(p,q) \neq (k,l)$ ist

$$\begin{aligned}(u^{p,q})^T u^{k,l} &= \sum_{i,j=1}^{n} \sin\frac{ip\pi}{n+1}\sin\frac{jq\pi}{n+1}\sin\frac{ik\pi}{n+1}\sin\frac{jl\pi}{n+1}\\ &= \left(\sum_{i=1}^{n}\sin\frac{ip\pi}{n+1}\sin\frac{ik\pi}{n+1}\right)\left(\sum_{j=1}^{n}\sin\frac{jq\pi}{n+1}\sin\frac{jl\pi}{n+1}\right)\\ &= 0,\end{aligned}$$

da einer der beiden Faktoren verschwindet. Denn für $q \neq l$ ist z. B.

$$\begin{aligned}\sum_{j=1}^{n}\sin\frac{jq\pi}{n+1}\sin\frac{jl\pi}{n+1} &= \frac{1}{2}\sum_{j=1}^{n}\left[\cos\frac{j(q-l)\pi}{n+1}-\cos\frac{j(q+l)\pi}{n+1}\right]\\ &= \frac{1}{2}\Re\left(\sum_{j=1}^{n}[(e^{i(q-l)\pi/(n+1)})^j-(e^{i(q+l)\pi/(n+1)})^j]\right)\\ &= \frac{1}{2}\Re\left(\sum_{j=0}^{n}[(e^{i(q-l)\pi/(n+1)})^j-(e^{i(q+l)\pi/(n+1)})^j]\right)\\ &= \frac{1}{2}\Re\left(\frac{1-e^{i(q-l)\pi}}{1-e^{i(q-l)\pi/(n+1)}}-\frac{1-e^{i(q+l)\pi}}{1-e^{i(q+l)\pi/(n+1)}}\right)\\ &= \frac{1}{2}\Re\left(\frac{1-e^{i(q-l)\pi}}{1-e^{i(q-l)\pi/(n+1)}}\right)-\frac{1}{2}\Re\left(\frac{1-e^{i(q+l)\pi}}{1-e^{i(q+l)\pi/(n+1)}}\right).\end{aligned}$$

Ist $k-l$ (und damit auch $k+l$) gerade, so verschweindet jeder der beiden Summanden. Ist dagegen $k-l$ (und damit auch $k+l$) ungerade, so ist

$$\begin{aligned}\frac{1}{2}\Re\left(\frac{1-e^{i(q-l)\pi}}{1-e^{i(q-l)\pi/(n+1)}}\right)-\frac{1}{2}\Re\left(\frac{1-e^{i(q+l)\pi}}{1-e^{i(q+l)\pi/(n+1)}}\right) &= \underbrace{\Re\left(\frac{1}{1-e^{i(q-l)\pi/(n+1)}}\right)}_{=1/2}\\ &\quad-\underbrace{\Re\left(\frac{1}{1-e^{i(q+l)\pi/(n+1)}}\right)}_{=1/2}\\ &= 0.\end{aligned}$$

Hierbei haben wir benutzt, dass

$$\Re\left(\frac{1}{1-e^{i\phi}}\right)=\frac{1}{2}$$

für beliebiges $\phi \notin 2\pi\mathbb{Z}$. Damit ist gezeigt, dass die Vektoren $u^{p,q}$ paarweise aufeinander senkrecht stehen und folglich linear unabhängig sind.

Es ist

$$\begin{aligned}u^{k+1} &= (V+rI)^{-1}[(rI-H)u^{k+1/2}+b]\\ &= (V+rI)^{-1}(rI-H)(H+rI)^{-1}(rI-V)u^k+2r(V+rI)^{-1}(H+rI)^{-1}b.\end{aligned}$$

Dies ist also ein Iterationsverfahren mit der Iterationsmatrix

$$M(r) := (V + rI)^{-1}(rI - H)(H + rI)^{-1}(rI - V).$$

Nach Satz 3.3 ist die Folge $\{u^k\}$ für jedes Startelement $u^0$ konvergent gegen die eindeutige Lösung von $u = M(r)u + c(r)$ mit $c(r) := 2r(V + rI)^{-1}(H + rI)^{-1}b$, falls $\rho(M(r)) < 1$. Bezeichnet man mit $\lambda_{pq}$ den Eigenwert von $H$ zum Eigenvektor $u^{p,q}$ und entsprechend mit $\mu_{pq}$ den Eigenwert von $V$ zum Eigenvektor $u^{p,q}$, so erkennt man dass die Eigenwerte von $M(r)$ durch

$$\nu_{pq} := \frac{(r - \lambda_{pq})(r - \mu_{pq})}{(\lambda_{pq} + r)(\mu_{pq} + r)}, \qquad p, q = 1, \ldots, n,$$

gegeben sind. Da die Eigenwerte $\lambda_{pq}$ und $\mu_{pq}$ positiv sind, ist $-1 < \nu_{pq} < 1$ bzw. $|\nu_{pq}| < 1$, $p, q = 1, \ldots, n$, bzw. $\rho(M(r)) < 1$. Da $(rI - H)$ und $H + rI)^{-1}$ miteinander vertauschbar sind, folgt aus $u = M(r)u + c(r)$, dass

$$(H + rI)(V + rI)u = (rI - H)(rI - V)u + 2rb$$

bzw.

$$[HV + r(H + V) + r^2 I]u = [r^2 I - r(H + V) + HV]u + 2rb.$$

berücksichtigt man nun, dass $A = H + V$, so erhält man $Au = b$. Genau das war zu zeigen.