Algorithmische Mathematik

Skript zur Vorlesung
im
Wintersemester 2007/8
und
Sommersemster 2008

Helmut Harbrecht

Stand: 14. Oktober 2008

Vorwort

Diese Mitschrift kann und soll nicht ganz den Wortlaut der Vorlesung wiedergeben. Sie soll das Nacharbeiten des Inhalts der Vorlesung erleichtern. Der Dank des Vorlesenden richtet sich an die Studenten Kai Gödde, Chris Kerstan, Christoph Kunze und Benedikt Lemmen, die einen erheblichen Teil dieser Mitschrift gesetzt haben.

Literatur zur Vorlesung:

— P. Deuflhard und A. Hohmann: *Numerische Mathematik*, de Gruyter-Verlag

— M. Hanke-Bourgeois: *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, Teubner-Verlag

— J. Stoer: *Numerische Mathematik I*, Springer-Verlag

— N. Blum: *Algorithmen und Datenstrukturen*, Oldenbourg-Verlag

— B. Korte und J. Vygen: *Combinatorial Optimization: Theory and Algorithms*, Springer-Verlag

— O. Häggström: *Finite Markov Chains and Algorithmic Applications*, London Mathematical Society

— U. Krengel: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*, Vieweg-Verlag

Inhaltsverzeichnis

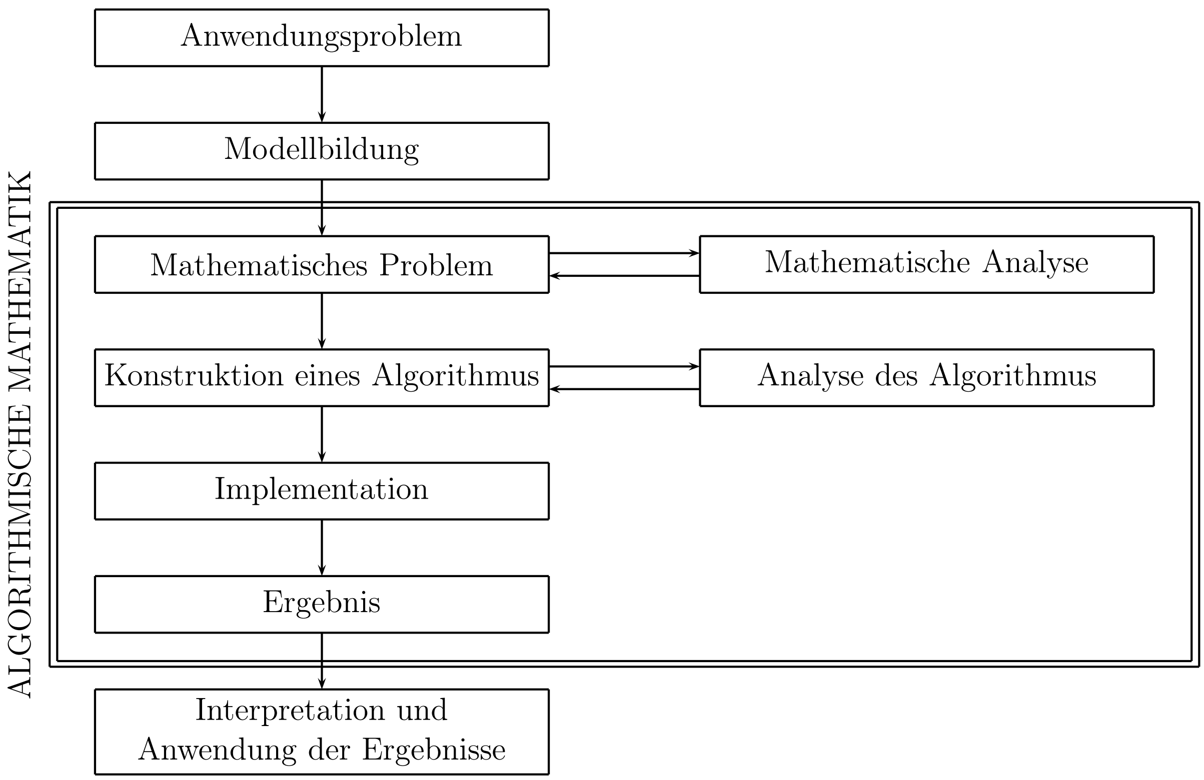
Einführung

Die Algorithmische Mathematik vereint verschiedene Gebiete der Angewandten Mathematik, nämlich

- Diskrete Mathematik,
- Numerik und
- Stochastik.

Ihre Aufgabe ist die Konstruktion und Analyse von Algorithmen zur Lösung mathematischer Aufgaben. Diese Aufgaben stammen zum Beispiel aus Technik, Naturwissenschaft, oder aus den Wirtschafts– und Sozialwissenschaften. Sobald Zahlenwerte erlaubt sind, treten überall ähnliche Probleme auf. Beispielsweise treten in 70% aller Anwendungen lineare Gleichungssysteme auf.

Übersicht zur Algorithmischen Mathematik:



1. Zahlendarstellung im Rechner

1.1 Zahlensysteme

Die Darstellung von Zahlen basiert auf sogenannten *Zahlensystemen* . Diese nutzen wiederum Zeichen eines Alphabets zu ihrer Darstellung. Es sei $\mathbb{N} := \{1, 2, \ldots\}$ die Menge der natürlichen Zahlen und $b \in \mathbb{N}$, $b > 1$ beliebig. Dann heißt die Menge $\Sigma_b := \{0, 1, \ldots, b-1\}$ das *Alphabet* des b-adischen Zahlensystems.

Beispiel 1.1

- Dem Dezimalsystem liegt das Alphabet $\Sigma_{10} := \{0, 1, \ldots, 9\}$ zugrunde. Worte aus diesem Alphabet sind zum Beispiel: 123, 734, 7806. Eine feste Wortlänge, etwa $n = 4$, erreicht man durch führende Nullen: 0123, 0734, 7806.
- $\Sigma_2 := \{0, 1\}$: Dual- oder Binäralphabet,
  $\Sigma_8 := \{0, 1, \ldots, 7\}$: Oktalalphabet,
  $\Sigma_{16} := \{0, 1, \ldots, 9, A, B, \ldots, F\}$: Hexadezimalalphabet.
  Die Basen 2, 8, 16 spielen in der Informatik eine entscheidende Rolle.
- Alte Basen sind $b = 12$ (Dutzend) und $b = 60$ (Zeitrechnung).

$\triangle$

Satz 1.2 Seien $b, n \in \mathbb{N}$ mit $b > 1$. Dann ist jede ganze, nicht-negative Zahl z mit $0 \leq z \leq b^n - 1$ eindeutig als Wort der Länge n über Σ_b darstellbar durch

$$z = \sum_{i=0}^{n-1} z_i \cdot b^i$$

mit $z_i \in \Sigma_b$ für alle $i = 0, 1, \ldots, n-1$. Vereinfachend wird die Ziffernschreibweise verwendet

$$z = (z_{n-1} z_{n-2} \ldots z_1 z_0)_b.$$

Beweis. Der Beweis erfolgt durch vollständige Induktion nach z.
Induktionsanfang: $z < b$ hat die eindeutige Darstellung $z_0 = z$ und $z_i = 0$ sonst.
Induktionsvorraussetzung: Behauptung sei wahr für alle $1, 2, \ldots, z-1$.
Induktionsschluss $z - 1 \mapsto z \geq b$: Es ist

$$z = \underbrace{\left\lfloor \frac{z}{b} \right\rfloor}_{=:\hat{z}} \cdot b + (z \bmod b).$$

Da $\widehat{z} < z$ ist, besitzt $\widehat{z}$ die eindeutige Darstellung

$$\widehat{z} = (\widehat{z}_{n-1}\widehat{z}_{n-2}\cdots\widehat{z}_0)_b.$$

Dabei ist $\widehat{z}_{n-1} = 0$, da

$$\widehat{z}_{n-1} \cdot b \leq z \leq b^n - 1.$$

Folglich ist $(z_{n-1}z_{n-2}z_{n-3}\cdots z_1 z_0)_b$ mit $z_{n-1} = \widehat{z}_{n-2}$, $z_{n-2} = \widehat{z}_{n-3}$, $\ldots$, $z_1 = \widehat{z}_0$ und $z_0 = z \bmod b$ eine n-stellige Darstellung von z.

Wir müssen noch die Eindeutigkeit der Darstellung zeigen. Angenommen, es existieren zwei verschieden Darstellungen von z, das heißt

$$z = \left(z_{n-1}^{(1)}z_{n-2}^{(1)}\cdots z_1^{(1)}z_0^{(1)}\right)_b = \left(z_{n-1}^{(2)}z_{n-2}^{(2)}\cdots z_1^{(2)}z_0^{(2)}\right)_b.$$

Sei $m \in \mathbb{N}$ der größte Index mit $z_m^{(1)} \neq z_m^{(2)}$, wobei ohne Einschränkung der Allgemeinheit $z_m^{(1)} > z_m^{(2)}$ gelte. Dann müssen die niedrigwertigen Stellen $z_{m-1}^{(2)}, z_{m-2}^{(2)}, \ldots, z_0^{(2)}$ eine höherwertige Stelle m wettmachen. Nun hat aber die größte durch niedrigwertigeren Stellen erreichbare Zahl die Form

$$\begin{aligned}
(b-1)\cdot b^0 + (b-1)\cdot b^1 + \cdots + (b-1)\cdot b^{m-1} &= (b-1)(\underbrace{b^0 + b^1 + \cdots + b^{m-1}}_{\text{geometrische Reihe}}) \\
&= (b-1)\frac{b^m - 1}{b-1} \\
&= b^m - 1.
\end{aligned}$$

Die größte durch niedrigwertige Stellen darstellbare Zahl ist also $b^m - 1$. Da aber $1 \cdot b^m$ gerade eine Einheit der fehlenden Stelle m ist, kann diese nicht wettgemacht werden. Damit folgt ein Widerspruch, das heißt, die Darstellung von z ist eindeutig. □

Aus dem Beweis erhält man direkt einen Algorithmus zur Umwandlung in die b-adische Zahlendarstellung.

Beispiel 1.3 Umwandlung von $z = (1364)_{10}$ in eine Oktalzahl:

$$\left.\begin{aligned}
1364 &= 170 \cdot 8 + 4 \\
170 &= 21 \cdot 8 + 2 \\
21 &= 8 \cdot 8 + 5 \\
2 &= 0 \cdot 8 + 2
\end{aligned}\right\} \Rightarrow z = (1364)_{10} = (2524)_8$$

△

Algorithmus 1.4 (b-adische Darstellung $(y_{n-1}y_{n-2}\cdots y_0)_b$ von x)

input: `unsigned int x, b, n;`
output: `unsigned int y[n];`

```
for (i=0; i<n; i++) y[i] = 0;
i=0;
while (x > 0)
```

```
{  y[i] = x % b;  /* entspricht x mod b   */
   x = x / b;     /* ganzzahlige Division */
   i++;
   }
```

Die Umwandlung b-adisch nach dezimal folgt mit dem *Horner-Schema*:

$$z = \sum_{i=0}^{n-1} z_i \cdot b^i = \Big(\dots \Big(\big((z_{n-1} \cdot b + z_{n-2}) \cdot b + z_{n-3} \big) \cdot b + z_{n-4} \Big) \cdot b + \dots + z_1 \Big) \cdot b + z_0.$$

Algorithmus 1.5 (Auswertung der b-adischen Darstellung)

```
input:    unsigned int y[n], b, n;
output:   unsigned int x;
x = 0;
for (i=n-1; i>=0; i--) x = x * b + y[i];
/* Beachte: i muss ein signed int sein! */
```

1.2 Darstellung ganzer Zahlen am Rechner

1.2.1 Vorzeichen-Betrag-Darstellung

Dies ist die im Alltag verwendete Darstellung. Wir wollen uns hier auf Dualzahlen beschränken. Bei einer Wortlänge von n (n Bit) wird das erste Bit als Vorzeichen($0 =$ '$+$', $1 =$ '$-$') verwendet, die restlichen Bits für den Betrag der Zahl. Da die Null zwei Darstellungen besitzt, nämlich ± 0, können $2^n - 1$ Zahlen dargestellt werden.

Beispiel 1.6 ($n = 3$)

Bitmuster	Dezimaldarstellung
000	+0
001	+1
010	+2
011	+3
100	-0
101	-1
110	-2
111	-3

△

Diese natürliche Darstellung ist auf Rechnern völlig unüblich. Der Grund dafür ist, dass hardwaremäßig nur Addierwerke (plus Zusatzlogik) verwendet werden. Daher benötigt man eine Darstellung, bei der die Subtraktion auf die Addition zurückgeführt werden kann.

1.2.2 Komplement-Darstellung

Definition 1.7 Sei $z = (z_{n-1}z_{n-2}\cdots z_1z_0)_b$ eine n-stellige b-adische Zahl. Das **$(b-1)$-Komplement** $K_{b-1}(z)$ ist definiert als

$$K_{b-1}(z) = (b-1-z_{n-1}, b-1-z_{n-2}, \ldots, b-1-z_1, b-1-z_0)_b.$$

Beispiel 1.8

- $K_9\big((325)_{10}\big) = (674)_{10}$
- $K_1\big((10110)_2\big) = (01001)_2$

$\triangle$

Definition 1.9 Das **b-Komplement** von z ist definiert als

$$K_b(z) = K_{b-1}(z) + 1.$$

Beispiel 1.10

- $K_{10}\big((325)_{10}\big) = (674)_{10} + (1)_{10} = (675)_{10}$
- $K_2\big((10110)_2\big) = (01001)_2 + (1)_2 = (01010)_2$

$\triangle$

Speziell im Dualsystem nennt man $K_{b-1} = K_1$ **Einerkomplement** und $K_b = K_2$ **Zweierkomplement**. Im Dezimalsystem spricht man bei $K_{b-1} = K_9$ vom **Neunerkomplement** und bei $K_b = K_{10}$ vom **Zehnerkomplement**.

Lemma 1.11 Für jede n-stellige b-adische Zahl z gilt

1. $z + K_b(z) = b^n$,
2. $z + K_{b-1}(z) = b^n - 1 = (b-1, \ldots, b-1)_b$,
3. $K_{b-1}\big(K_{b-1}(z)\big) = z$,
4. $K_b\big(K_b(z)\big) = z$.

Beweis. Nach Definition des $(b-1)$-Komplements ist

$$\begin{aligned} z + K_{b-1}(z) &= (b-1, \ldots, b-1)_b \\ &= \sum_{i=0}^{n-1} (b-1) \cdot b^i \\ &= (b-1) \cdot \sum_{i=0}^{n-1} \cdot b^i \\ &= (b-1) \cdot \frac{b^n - 1}{b-1} \\ &= b^n - 1. \end{aligned}$$

Daher gilt Aussage 1. Aussage 2 folgt aus der Definition $K_b(z) = K_{b-1}(z) + 1$. Die dritte Aussage ist offensichtlich, während aus der ersten folgt

$$K_b(z) + K_b\big(K_b(z)\big) = b^n = z + K_b(z).$$

Dies impliziert

$$K_b\big(K_b(z)\big) = z,$$

das heißt Aussage 4. □

Die erste Aussage von Lemma 1.11 bedeutet anschaulich, dass das b-Komplement von z bei n Stellen sich als Differenz von z zu b^n ergibt. Beispielsweise ergibt sich im Dezimalsystem mit $n = 3$ Stellen für $z = (374)_{10}$

$$K_{10}(374) = 1000 - 374 = 626.$$

Mithilfe des b-Komplements werden die b^n Zahlen z mit

$$-\left\lfloor \frac{b^n}{2} \right\rfloor \leq z \leq \left\lceil \frac{b^n}{2} \right\rceil - 1$$

dargestellt. Dieser Bereich heißt *darstellbarer Bereich*. Zahlen aus diesem Bereich werden wie folgt dargestellt:

- nicht-negative Zahlen durch ihre b-adische Darstellung,
- negative Zahlen z durch das b-Komplement ihres Betrags $|z|$.

Definition 1.12 Die ***b*-Komplement-Darstellung** $(z)_{K_b} = (z_{n-1}, z_{n-2}, \ldots, z_1, z_0)$ einer Zahl $z \in \mathbb{Z}$ mit darstellbarem Bereich $-\lfloor b^n/2 \rfloor \leq z \leq \lceil b^n/2 \rceil - 1$ ist definiert als

$$(z)_{K_b} = \begin{cases} (z)_b, & \text{falls } z \geq 0, \\ \big(K_b(|z|)\big)_b, & \text{sonst.} \end{cases}$$

Beispiel 1.13

- Im Fall $b = 10$, $n = 2$ ist der darstellbare Bereich $-50 \leq z \leq 49$. Konkrete Darstellungen sind:

Zahl	Darstellung
43	43
-13	87
-27	73
38	38

- Im Fall $b = 2$, $n = 3$ ist der darastellbare Bereich $-4 \leq z \leq 3$. Konkrete Darstellungen sind:

Bitmuster	Dezimalwert
000	0
001	1
010	2
011	3
100	-4
101	-3
110	-2
111	-1

$\triangle$

Wir betrachten nun die Addition von Zahlen in dieser Darstellung. Dazu sei $(x)_{K_b} \oplus (y)_{K_b}$ die ziffernweise Addition der Darstellungen von x und y, wobei ein eventueller Überlauf auf die $(n+1)$-te Stelle vernachlässigt wird, man rechnet also modulo b^n.

Satz 1.14 Seien x, y n-stellige Zahlen und x, y, $x+y$ im darstellbaren Bereich. Dann gilt

$$x + y = (x)_{K_b} \oplus (y)_{K_b}.$$

Beweis. Sind $x, y \geq 0$, so gilt

$$\begin{aligned}(x)_{K_b} \oplus (y)_{K_b} &= (x)_b + (y)_b \bmod b^n \\ &= x + y.\end{aligned}$$

Sind $x, y < 0$, so gilt nach Definition 1.12 und Lemma 1.11

$$\begin{aligned}(x)_{K_b} \oplus (y)_{K_b} &= \big(K_b(|x|)\big)_b + \big(K_b(|y|)\big)_b \bmod b^n \\ &= K_b(|x|) + K_b(|y|) \bmod b^n \\ &= b^n - |x| + b^n - |y| \bmod b^n \\ &= -(|x| + |y|) \bmod b^n \\ &= x + y.\end{aligned}$$

Ist $x \geq 0$ und $y < 0$, so folgt

$$\begin{aligned}(x)_{K_b} \oplus (y)_{K_b} &= (x)_b + \big(K_b(|y|)\big)_b \bmod b^n \\ &= x + K_b(|y|) \bmod b^n \\ &= x + b^n - |y| \bmod b^n \\ &= x - |y| \bmod b^n \\ &= x + y.\end{aligned}$$

$\square$

Beispiel 1.15 $b = 10$, $n = 2$, $-50 \leq z \leq 49$:

$$\begin{aligned}27 + 12 &= (27)_{K_{10}} \oplus (12)_{K_{10}} = (39)_{K_{10}} = 39 \\ 27 + (-15) &= (27)_{K_{10}} \oplus (85)_{K_{10}} = (12)_{K_{10}} = 12 \\ 27 + (-34) &= (27)_{K_{10}} \oplus (66)_{K_{10}} = (93)_{K_{10}} = -7 \\ (-27) + (-21) &= (73)_{K_{10}} \oplus (79)_{K_{10}} = (52)_{K_{10}} = -48\end{aligned}$$

△

Die Rückführung der Subtraktion auf die Addition beruht auf der Gleichung $x - y = x + (-y)$.

Satz 1.16 Seien x, y n-stellge Zahlen und $x - y$ im darstellbaren Bereich. Dann gilt

$$x - y = (x)_{K_b} \oplus \big(K_b(y)\big)_{K_b}.$$

Beweis. Gemäß der ersten Aussage aus Lemma 1.11 gilt $y + K_b(y) = b^n$. Daher ist $-y = K_b(y) - b^n$ woraus wegen der modulo-b^n-Rechnung folgt

$$(-y)_{K_b} = \big(K_b(y)\big)_{K_b}.$$

Damit ist

$$x - y = (x)_{K_b} \oplus (-y)_{K_b} = (x)_{K_b} \oplus \big(K_b(y)\big)_{K_b}.$$

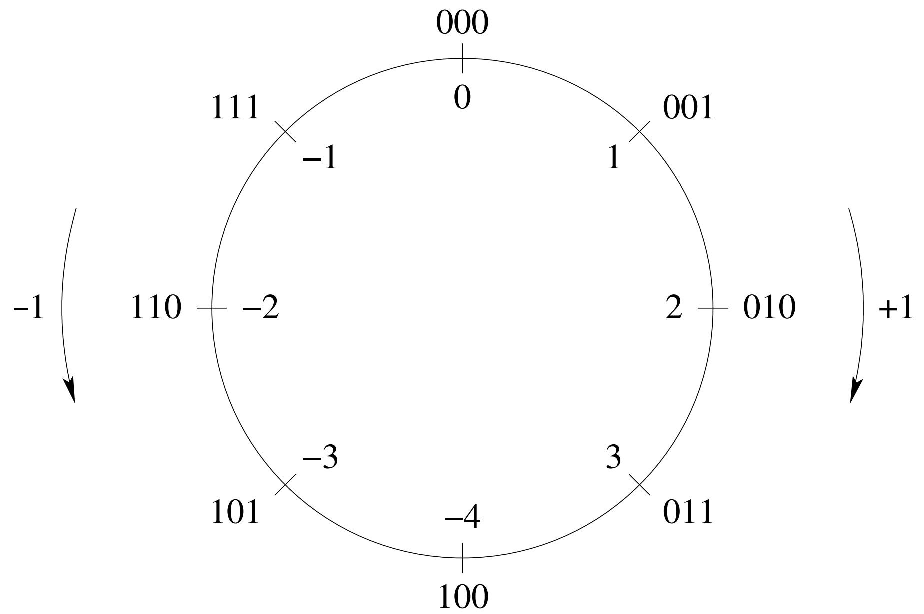
□

Beispiel 1.17 $b = 10$, $n = 2$, $-50 \leq z \leq 49$:

$$\begin{aligned} 37 - 28 =& (37)_{K_{10}} \oplus (K_{10}(28)_{K_{10}} = (37)_{K_{10}} \oplus (72)_{K_{10}} = (9)_{K_{10}} &&= 9 \\ 37 - 48 =& (37)_{K_{10}} \oplus (K_{10}(48)_{K_{10}} = (37)_{K_{10}} \oplus (52)_{K_{10}} = (89)_{K_{10}} &&= -11 \\ -12 - 24 =& (88)_{K_{10}} \oplus (K_{10}(24)_{K_{10}} = (88)_{K_{10}} \oplus (76)_{K_{10}} = (64)_{K_{10}} &&= -36 \end{aligned}$$

△

Die darstellbaren Zahlen kann man sich beim b-Komplement ringförmig vorstellen. Für $b = 2$, $n = 3$ ergibt sich:



Man beachte, dass ein eventueller Überlauf bzw. Unterlauf (Verlassen des darstellbaren Bereichs) im allgemeinen *nicht* durch den Rechner aufgefangen wird. So ist zum Beispiel bei n-stelliger Dualzahlarithmetik die größte darstellbare Zahl $x_{\max} = (011\cdots1)_2$ und $x_{\max} + 1 = (100\cdots0)_2$ wird als -2^{n-1} interpretiert.

1.3 Darstellung reeller Zahlen

1.3.1 Festkommadarstellung (Fixed Point Representation)

Definition 1.18 Bei der **Festkommadarstellung** wird bei vorgegebener Stellenzahl n eine feste Nachkommastelle vereinbart, das heißt

$$z = \pm(z_{k-1}z_{k-2}\cdots z_0.z_{-1}z_{-2}\cdots z_{k-n})_b = \pm\sum_{i=k-n}^{k-1} z_i \cdot b^i$$

mit k Vorkomma- und $(n-k)$ Nachkommastellen.

Beispiel 1.19

- $(271.314)_10 = 271.314$
- $(101.011)_2 = 2^2 + 2^0 + 2^{-2} + 2^{-3} = 5.375$

△

Im Gegensatz zur Darstellung von ganzen Zahlen treten bereits bei der Konvertierung von Dezimalzahlen in das b-adische Zahlensystem $b \neq 10$ Rundungsfehler auf. So ist die Dezimalzahl 0.8 im Binärsystem unendlich periodisch

$$(0.8)_{10} = (0.1100\overline{1100})_2.$$

Der große Nachteil der Festkommaarithmetik besteht jedoch darin, dass der darstellbare Bereich stark beschränkt ist, insbesondere ist die Genauigkeit zwischen zwei benachbarten Zahlen immer gleich. Liegt etwa ein Rechenergebnis x zwischen den beiden kleinsten positiven Zahlen

$$z_1 = (00\cdots 0.00\cdots 01)_b$$

und $z_2 = 2z_1$, so muss x zur internen Darstellung auf eine der beiden Zahlen gerundet werden. Für

$$x = \frac{z_1 + z_2}{2} = \frac{3}{2}\cdot z_1$$

ergibt sich bei der Rundung der *relative Fehler*

$$\left|\frac{x - z_1}{x}\right| = \left|\frac{x - z_2}{x}\right| = \left|\frac{\frac{1}{2}z_1}{\frac{3}{2}z_1}\right| = \frac{1}{3} \approx 33\%.$$

Dies ist für numerisch-wissenschaftliche Rechnungen völlig inakzeptabel.

1.3.2 Gleitkommadarstellung (Floating Point Representation)

Definition 1.20 Gleitkommazahlen haben die Form

$$z = \pm m \cdot b^{\pm e}$$

mit der **Mantisse** m, dem **Exponenten** e und der **Basis** b.

Die Basis der Mantisse kann von der Basis b des Exponenten verschieden sein. Wir werden hier jedoch die gleiche Basis vorraussetzen.

Beispiel 1.21

- $-213.78 = -2.1378 \cdot 10^2$
- $0.000031 = 0.31 \cdot 10^{-4} = 3.1 \cdot 10^{-5}$
- $(1101.011)_2 = 1.101011 = 3.1 \cdot 2^3$

$\triangle$

Die Beispiele zeigen, dass die Gleitkommadarstellung nicht eindeutig ist. Daher normalisiert man die Mantisse:

Definition 1.22 Die Mantisse m heißt **normalisiert** , falls

$$m = 0.m_1 m_2 \cdots m_t \quad \text{mit} \quad m_1 \neq 0,$$

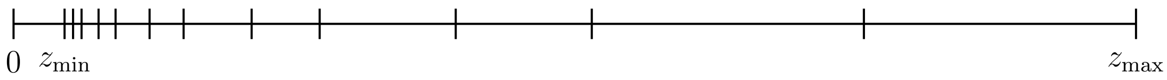
oder, äquivalent dazu,

$$1 \leq m_1 < b.$$

Bei der rechnerinternen Darstellung muss festgelegt werden, wie viele Stellen für die Mantisse bzw. den Exponenten zur Verfügung stehen. Dies resultiert in der Menge der Maschinenzahlen $F = F(b, t, e_{\text{min}}, e_{\text{max}})$ der Form

$$z = \pm 0.m_1 \ldots m_t \cdot b^e, \qquad e_{\text{min}} \leq e \leq e_{\text{max}}.$$

Maschinenzahlen sind wie folgt angeordnet:



Man sieht, dass der Abstand zwischen benachbarten Zahlen mit deren Betrag wächst. Ferner erkennt man eine Lücke zwischen Null und der kleinsten positiven Zahl z_{min}. Sie ergibt sich aus der Normalisierungsbedingung und ist deutlich größer als der Abstand zur zweitkleinsten positiven Zahl. Rundungsfehler werden wir im nächsten Abschnitt analysieren.

Da bei der Basis $b = 2$ das erste Bit wegen der Normalisierungsbedingung weggelassen werden kann, da es eindeutig festliegt ($m = 1$), wird es oft nicht dargestellt (*hidden bit*). In diesem Fall werden die t Mantissenbits zur Darstellung von $m_2, m_3, \ldots, m_{t+1}$ genutzt, allerdings muss dann die Null gesondert dargestellt werden, zum Beispiel als Mantisse 0 mit Exponenten $e_{\text{min}} - 1$.

Um Exponenten besser vergleichen zu können, verwendet man oft die *Exzess- oder Biasdarstellung*: Durch Addition des Exzesses $|e_{\text{min}}| + 1$ wird der Exponent auf den Bereich $1, 2, \ldots, |e_{\text{min}}| + e_{\text{max}} + 1$ transformiert.

Wir wollen den IEEE-Standard 754 näher betrachten. Für die gesamte Zahl stehen 64 Bit zur Verfügung, davon

- 52 Bit für die Mantisse in hidden-bit-Darstellung

$$1.\underbrace{m_2 m_3 \ldots m_{53}}_{\textit{Signifikant}},$$

- 11 Bit für den Exponenten mit

$$e_{\min} = -1022, \qquad e_{\max} = 1023,$$

gemäß der Exzessdarstellung gespeichert als

$$1 \le e \le |e_{\min}| + e_{\max} + 1 \le 2046,$$

- 1 Bit als Vorzeichen: 0 = '+', 1 = '−'.

Speziell gilt:

- Die Zahl 0 wird durch das Bitmuster $00\ldots0$ repräsentiert, das heißt

$$0 = 1.00\ldots0 \cdot 2^{e_{\min}-1}.$$

- Ein Overflow (± inf) erzeugt im Exponenten das Bitmuster $11\ldots1$, was 2^{1024} entspricht.
- Ein positiver Signifikant mit Exponent 0 entspricht einem Underflow (kleiner als die kleinste postive Zahl). Hier werden im IEEE-Standard weitere Unterscheidungen getroffen (*denormalized*), oftmals liefert der Computer nur ein ± NaN (Not a Number).
- Die größte Maschinenzahl ist

$$z_{\max} = 1.7976931348623157 \cdot 10^{308},$$

die kleinste positive Maschinenzahl ist

$$z_{\min} = 2.2250738585072014 \cdot 10^{-308}.$$

Beispiel 1.23 Die Dezimalzahl 10 wird im IEEE-Standard 754 dargestellt als

$$\underbrace{0}_{\substack{\text{Vorzeichen}\\ +}} \quad \underbrace{10000000010}_{\substack{\text{Exponent}\\ 1026-1023=3}} \quad \underbrace{0100\ldots00}_{\substack{\text{Signifikant}\\ 1.25}},$$

während −0.8 das (gerundete) Bitmuster

$$\underbrace{1}_{\substack{\text{Vorzeichen}\\ -}} \quad \underbrace{01111111110}_{\substack{\text{Exponent}\\ 1022-1023=-1}} \quad \underbrace{10011001\ldots10011010}_{\substack{\text{Signifikant}\\ 1.5999\ldots}}$$

ergibt. △

1.3.3 Genauigkeit der Gleitkommadarstellung

Der begrenzte Vorrat an Maschinenzahlen erzwingt bei der Konvertierung oder Darstellung von Zwischenergebnissen die Rundung auf Maschinenzahlen.

Definition 1.24 Die **Rundung** ist eine Abbildung $\mathrm{rd} : \mathbb{R} \to F(b, t, e_{\min}, e_{\max})$ mit den Eigenschaften:

1. $\mathrm{rd}(a) = a \quad$ für alle $a \in F$,
2. $|x - \mathrm{rd}(x)| = \min_{a \in F} |x - a|$ für alle $x \in \mathbb{R}$.

Definition 1.25 Den maximalen relativen Rundungsfehler **eps** für $z_{\min} \leq |x| \leq z_{\max}$ nennt man die **Rechnergenauigkeit**. Die Stellen der Mantisse heißen **signifikante** Stellen, eine t-stellige Mantisse bezeichnet man auch als ***t*-stellige Arithmetik**.

Satz 1.26 Die Rechnergenauigkeit eps für $F = F(b, t, e_{\min}, e_{\max})$ ist

$$\mathrm{eps} = \frac{1}{2} \cdot b^{1-t}.$$

Beweis. Wir betrachten zunächst den relativen Rundungsfehler μ, der durch das Abschneiden (*chopping*) nicht-signifikanter Stellen ansteht. Sei hierzu $z_{\min} < x < z_{\max}$ und $x_c \in F$ die durch Abschneiden entstehende Zahl. Dann gilt

$$\begin{aligned}
\mu &= \frac{x - x_c}{x} \\
&= \frac{0.x_1 x_2 \ldots x_t x_{t+1} x_{t+2} \ldots \cdot b^e - 0.x_1 x_2 \ldots x_t \cdot b^e}{x} \\
&= \frac{0.00 \ldots 0 x_{t+1} x_{t+2} \ldots \cdot b^e}{x} \\
&= \frac{0.x_{t+1} x_{t+2} \ldots \cdot b^{e-t}}{x} \\
&< \frac{b^{e-t}}{x} && (\text{da } 0.x_{t+1} x_{t+2} \ldots < 1) \\
&\leq \frac{b^{e-t}}{b^{e-1}} && (\text{da } x \geq b^{e-1}) \\
&= b^{1-t}.
\end{aligned}$$

Hieraus folgt die Behauptung, da die Rechnergenauigkeit offenbar halb so groß wie der Abschneidefehler μ ist. □

Die Rechnergenauigkeit eps ist die wichtigste Größe bei Genauigkeitsbetrachtungen für ein Gleitkommasystem. Sie bezieht sich auf die Anzahl t signifikanter Stellen zur Basis b. Die entsprechende Anzahl s signifikanter Stellen im Dezimalsystem erhält man durch Auflösen von

$$\mathrm{eps} = \frac{1}{2} \cdot b^{1-t} = \frac{1}{2} \cdot 10^{1-s}$$

nach s. Dies ergibt

$$s = \lfloor 1 + (t-1) \cdot \log_{10} b \rfloor.$$

Für den IEEE-Standard 754 mit $b = 2$ und $t = 53$ ergeben sich $s = 16$ signifikante Dezimalstellen.

2. Fehleranalyse

2.1 Rechnerarithmetik

Man hat auf dem Rechner nur eine *Pseudoarithmetik*. *Pseudoarithmetik* Die Menge $F = F(b, t, e_{\min}, e_{\max})$ ist *nicht* abgeschlossen bezüglich $+, -, *, /$.

Beispiel 2.1 Betrachte $F = F(10, 5, -4, 5)$. Dann gilt für

$$\begin{aligned} x &:= 0.25684 \cdot 10^1 &&= 2.5684, \\ y &:= 0.32791 \cdot 10^{-2} &&= 0.00032971 \end{aligned}$$

dass

$$\begin{aligned} x + y &= 2.5716791 \notin F, \qquad & \mathrm{rd}(x+y) &= 0.25717 \cdot 10^1, \\ x * y &= 0.008422044044 \notin F, \qquad & \mathrm{rd}(x*y) &= 0.84220 \cdot 10^{-2}, \\ x/y &= 783.2637004 \notin F, \qquad & \mathrm{rd}(x/y) &= 0.78326 \cdot 10^3. \end{aligned}$$

$\triangle$

Die Rechneroperationen $\boxplus, \boxminus, \circledast, \boxslash$ müssen so realisiert sein, dass das Ergebnis wieder in F ist. Im allgemeinen ist das Ergebnis der Rechneroperationen $\boxdot$, $\circ \in \{+, -, *, /\}$ festgelegt durch:

$$\mathrm{rd}(x \circ y) = x \boxdot y = \mathrm{rd}(x \boxdot y) \qquad \forall x, y \in \mathbb{R}. \tag{2.1}$$

Dies gilt beispielsweise gemäß dem IEEE-Standard. Hardwaremäßig wird üblicherweise mit einer längeren Mantisse gearbeitet und das Ergebnis dann normalisiert und gerundet. Unter der Annahme (2.1) gilt demnach für $|x|, |y, |x \circ y| \in [z_{\min}, z_{\max}]$ für den relativen Fehler:

$$\left| \frac{x \boxdot y - x \circ y}{x \circ y} \right| = \left| \frac{\mathrm{rd}(x \circ y) - x \circ y}{x \circ y} \right| \leq \mathrm{eps}\,.$$

In $\mathbb{R}$ gelten Assoziativ-, Kommutativ- und Distributivgesetze, in der Rechnerarithmetik in F im allgemeinen *nur* das *Kommutativgesetz* bei Addition und Multiplikation.

Beispiel 2.2 Betrachte $F = F(10, 5, -4, 5)$.

- Assoziativgesetz $(a + b) + c = a + (b + c)$: Exakt gilt

  $$0.98765 + 0.012424 - 0.6065432 = 0.9935308,$$

  während sich einerseits

  $$0.98765 \boxplus (0.012424 \boxminus 0.6065432) = 0.98765 \boxminus 0.0058808 = 0.9935308$$

ergibt, andererseits jedoch

$$(0.98765 \boxplus 0.012424) \boxminus 0.6065432 = 1.00001 \boxminus 0.6065432 = 0.99356.$$

- Distributivgesetz $(a-b)\cdot c = a\cdot c - b\cdot c$: Exakt gilt

$$(\underbrace{4.2832}_{=:a} - \underbrace{4.2821}_{=:b}) \cdot \underbrace{5.7632}_{=:c} = 0.00633952,$$

jedoch ist

$$(a \boxminus b) \circledast c = 0.0011000 \circledast 5.7632 = 0.00633952$$

und

$$(a \circledast c) \boxminus (b \circledast c) = 24.685 \boxminus 24.679 = 0.0060000.$$

△

Folgerung: Mathematisch äquivalente Algorithmen zur Auswertung eines rationalen Ausdrucks können bei Ausführung auf dem Rechner zu *wesentlich* verschiedenen Ergebnissen führen, selbst wenn die Eingangsdaten Maschinenzahlen sind. Nicht-rationale Operationen wie $\sqrt{x}$, $\sin(x)$, $\exp(x)$ sind (nicht immer gut) softwaremäßig realisiert. Auch hierfür gilt im allgemeinen, dass das Maschinenergebnis gleich dem gerundeten exakten Ergebnis ist.

2.2 Fehlerfortpflanzung

Unglücklicherweise pflanzen sich einmal bereits gemachte Fehler, zum Beispiel durch Rundung, fort. Gegeben seien $|x|, |y|$ und $|\Delta x|, |\Delta y|$ mit

$$\left|\frac{\Delta x}{x}\right|, \left|\frac{\Delta y}{y}\right| \ll 1.$$

Was passiert bei *exakter* Durchführung einer elementaren Operation mit diesen Fehlern?

Lemma 2.3 Für den fortgepflanzten Fehler

$$\Delta(x \circ y) := (x + \Delta x) \circ (y + \Delta y) - x \circ y,$$

$\circ \in \{+, -, *, /\}$, gelten die Abschätzungen

$$\begin{aligned} \frac{\Delta(x \pm y)}{x \pm y} &= \frac{x}{x \pm y} \cdot \frac{\Delta x}{x} \pm \frac{y}{x \pm y} \cdot \frac{\Delta y}{y}, \\ \frac{\Delta(x * y)}{x * y} &\approx \frac{\Delta x}{x} + \frac{\Delta y}{y}, \\ \frac{\Delta(x/y)}{x/y} &\approx \frac{\Delta x}{x} - \frac{\Delta y}{y}. \end{aligned}$$

Dabei bedeutet "$\approx$" eine Vernachlässigung der Terme $(\Delta x)^2, (\Delta y)^2$ und $\Delta x \Delta y$.

Beweis. Der einfache Beweis verbleibt dem Leser zur Übung. □

Wir bemerken, dass sich die Fehler bei "$*$" beziehungsweise "$/$" im wesentlichen addieren bzw. subtrahieren, das heißt, es tritt keine wesentliche Verstärkung des Fehlers auf. Ist hingegen $|x \pm y|$ klein gegenüber $|x|$ oder $|y|$, so kann der relative Fehler außerordentlich verstärkt werden. Dieser Effekt heißt *Auslöschung*. Man muss bei der Aufstellung von Algorithmen darauf achten, dass dies soweit wie möglich vermieden wird.

Beispiel 2.4 Die quadratische Gleichung

$$x^2 - 2px + q = 0$$

besitzt die Lösungen

$$x_{1/2} = p \pm \sqrt{p^2 - q}.$$

Diese lassen sich berechnen gemäß

```
d = sqrt(p*p-q);
x1 = p+d;
x2 = p-d;
```

Ein numerisches Beispiel mit den Werten $p = 100$ und $q = 1$, ausgeführt mit einer dreistelligen dezimalen Rechnerarithmetik, ergibt:

$$\begin{aligned}
d &= \sqrt{p \cdot p - q} = \underbrace{10000 - 1}_{=0.1 \cdot 10^5} = \underbrace{9999}_{\widehat{=} 0.1 \cdot 10^5} = 0.100 \cdot 10^3, \\
x_1 &= p + d = 0.100 \cdot 10^3 + 0.100 \cdot 10^3 = 0.200 \cdot 10^3 = 200, \\
x_2 &= p - d = 0.100 \cdot 10^3 - 0.100 \cdot 10^3 = 0.
\end{aligned}$$

Die exakten Werte in dreistelliger Arithmetik lauten jedoch $x_1 = 200$ und $x_2 = 0.005$. Die Rechnergenauigkeit ist

$$\text{eps} = \frac{1}{2} \cdot 10^{1-3} = 0.005.$$

Die Abweichung des errechneten Ergebnis muss daher als inakzeptal betrachtet werden. Das Ergebnis $x_2 = 0$ ist schlicht falsch. Die *Auslöschung* bei Berechnung von x_2 kann mithilfe des *Wurzelsatzes von Vieta*

$$x_1 x_2 = q \tag{2.2}$$

vermieden werden. Es wird lediglich die betragsgrößere Nullstelle berechnet, die zweite wird dann via (2.2) berechnet:

```
d = sqrt(p*p-q);
if (p >= 0) x1 = p+d;
else        x1 = p-d;
x2 = q/x1;
```

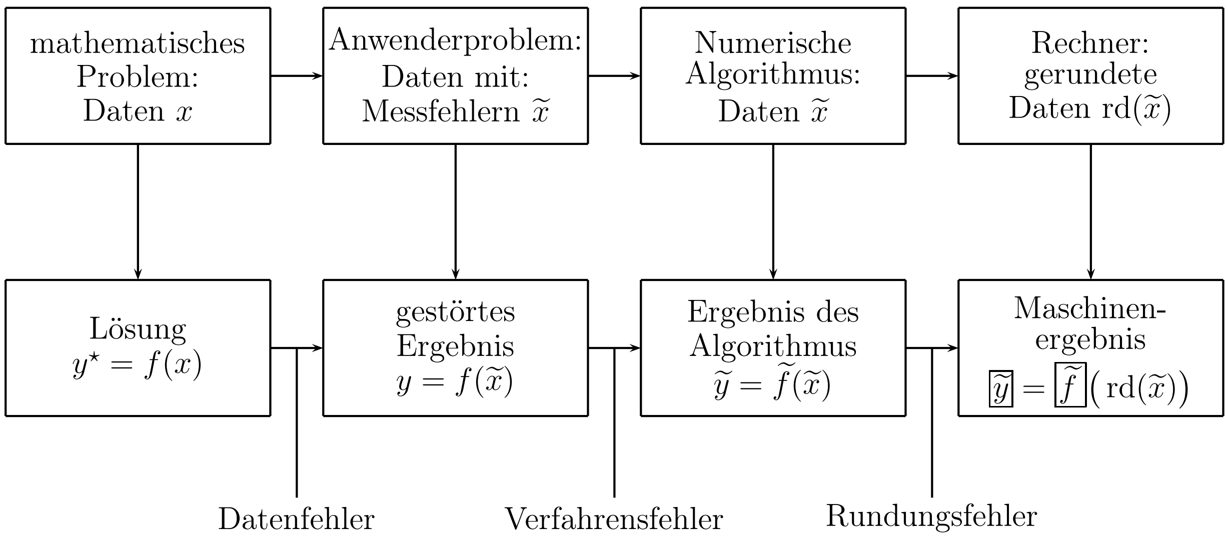
Konkret erhält man nun:

$$\begin{aligned}
d &= 0.100 \cdot 10^3, \\
x_1 &= p + d = 200, \\
x_2 &= 1/200 = 0.005.
\end{aligned}$$

△

Insgesammt gibt es drei verschiedene Fehlerarten:

1. **Rundungs-** oder **Reduktionsfehler**.
2. **Datenfehler**: Üblicherweise sind Eingangsdaten $x \in D$ ungenau, zum Beispiel Messdaten mit Messfehlern Δx. Das Lösen eines Problems entspricht dem Auswerten einer Funktion $f : D \to W$. Der Datenfehler ist dann $f(x + \Delta x) - f(x)$. Die Empfindlichkeit der Lösung $f(x)$ des Problems gegenüber kleiner Datenstörung Δx heißt *Kondition des Problems*. Sie ist eine Eigenschaft des Problems und *nicht* des Verfahrens.
3. **Verfahrensfehler**: Exakte Verfahren enden bei exakter Rechnung nach endlich vielen Operationen mit dem exakten Ergebnis. Näherungsverfahren, zum Beispiel Iterationsverfahren, enden in Abhängigkeit von bestimmten Kriterien mit einer Näherung $\widetilde{y}$ für die Lösung y. Der Verfahrensfehler ist $\widetilde{y} - y$.



Zur *Fehleranalyse* verfolgt man die Auswirkung aller Fehler, die in den einzelnen Schritten eines Algorithmus auftreten können:

Definition 2.5 Bei der **Vorwärtsanalyse** wird der Fehler von Schritt zu Schritt verfolgt und der akkumulierte Fehler für jedes Teilergebnis abgeschätzt.

Bei der **Rückwärtsanalyse** geschieht die Verfolgung des Fehlers hingegen so, dass jedes Zwischenergebnis als exakt berechnetes Ergebnis für gestörte Daten interpretiert wird, das heißt, der akkumulierte Fehler im Teilergebnis wird als Datenfehlereffekt interpretiert.

Beispiel 2.6 Wir untersuchen die Addition von zwei Zahlen, das heißt

$$f(x, y) = x + y.$$

Es ergibt sich:

Vorwärtsanalyse:	$\boxed{f}(x, y) = x \boxplus y = (x + y)(1 + \varepsilon)$
Rückwärtsanalyse:	$x \boxplus y = x(1 + \varepsilon) + y(1 + \varepsilon) = f\big(x(1 + \varepsilon), y(1 + \varepsilon)\big)$

mit $|\varepsilon| \le \mathrm{eps}$. △

Die Vorwärtsanalyse ist im allgemeinen kaum durchführbar, für die meisten Verfahren ist, wenn überhaupt, nur die Rückwärtsanalyse bekannt.

Beispiel 2.7 Wir analysieren den ersten Algorithmus aus Beispiel 2.4 zur Berechnung der betragskleineren Nullstelle x_2 rückwärts, dies bedeutet,

$$f(p,q) := p - \sqrt{p^2 - q}$$

mit $p > 0$. Es gilt

$$\begin{aligned}
&\left[p - \sqrt{\big(p^2(1+\varepsilon_1) - q\big)(1+\varepsilon_2)(1+\varepsilon_3)}\right](1+\varepsilon_4) \qquad \text{mit } |\varepsilon_i| \le \text{eps} \\
&\quad = p(1+\varepsilon_4) - \sqrt{p^2 \underbrace{(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)^2}_{=1+\varepsilon_5 \text{ mit } |\varepsilon_5| \le 4\,\text{eps}}(1+\varepsilon_4)^2 - q \underbrace{(1+\varepsilon_2)(1+\varepsilon_3)^2(1+\varepsilon_4)^2}_{=1+\varepsilon_6 \text{ mit } |\varepsilon_6| \le 5\,\text{eps}}} \\
&\quad = p(1+\varepsilon_4) - \sqrt{p^2(1+\varepsilon_4)^2 - q \underbrace{\left(1 + \varepsilon_6 - \frac{p^2(1+\varepsilon_4)^2\varepsilon_5}{q}\right)}_{=1+\varepsilon_7 \text{ mit } |\varepsilon_7| \le \text{eps}\left(5+4\frac{p^2}{q}\right)}} \\
&\quad = f\big(p(1+\varepsilon_4), q(1+\varepsilon_7)\big).
\end{aligned}$$

Der Rechner liefert demnach die exakte Lösung von

$$x^2 - 2p(1+\varepsilon_4) + q(1+\varepsilon_7) = 0$$

mit

$$|\varepsilon_4| \le \text{eps}, \qquad |\varepsilon_7| \le \text{eps}\left(5 + 4\frac{p^2}{q}\right).$$

Die Abschätzung für ε_7 explodiert, falls

$$0 < |q| \ll 1 < p.$$

△

2.3 Kondition und Stabilität

Gegeben sei eine Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto y = f(x).$$

Für fehlerbehaftete Daten $x + \Delta x$ gilt

$$\frac{f(x+\Delta x) - f(x)}{\Delta x} \approx f'(x).$$

Folglich gilt für den absoluten Datenfehler

$$\Delta y = f(x+\Delta x) - f(x) \approx f'(x)\Delta x$$

und für den relativen Datenfehler

$$\frac{\Delta y}{y} = \frac{f'(x)\Delta x}{f(x)} = \frac{f'(x)\cdot x}{f(x)}\frac{\Delta x}{x}.$$

Definition 2.8 Die Zahl

$$\kappa_{\text{abs}} = |f'(x)|$$

heißt **absolute Konditionszahl** des Problems $x \mapsto f(x)$. Für $x \cdot f(x) \neq 0$ ist

$$\kappa_{\text{rel}} = \left| \frac{f'(x) \cdot x}{f(x)} \right|$$

die entsprechende **relative Konditionszahl**. Ein Problem heißt **schlecht konditioniert**, falls eine der Konditionszahlen deutlich größer ist als 1, ansonsten heißt es **gut konditioniert**.

Beispiel 2.9

- Im Fall der Addition $f(x) = x + a$ haben wir

  $$\kappa_{\text{rel}} = \left| \frac{f'(x) \cdot x}{f(x)} \right| = \left| \frac{x}{x+a} \right|.$$

  Dies bedeutet, die relative Konditionszahl ist groß, wenn $|x + a| \ll |x|$.

- Im Fall der Multiplikation $f(x) = ax$ gilt

  $$\kappa_{\text{abs}} = |f'(x)| = |a|.$$

  Die absolute Kondition ist also schlecht, falls $1 \ll a$. Wegen

  $$\kappa_{\text{rel}} = \left| \frac{f'(x) \cdot x}{f(x)} \right| = \left| \frac{ax}{ax} \right| = 1$$

  ist die relative Kondition jedoch immer gut.

△

Definition 2.10 Erfüllt die Implementierung eines Algorithmus $\boxed{f}$ zur Lösung eines Problems $x \mapsto f(x)$

$$\left| \frac{\boxed{f}(x) - f(x)}{f(x)} \right| \leq C_V \kappa_{\text{rel}} \,\text{eps}$$

mit einem mäßig großen $C_V > 0$, so wird der Algorithmus $\boxed{f}$ **vorwärtsstabil** genannt. Ergibt die Rückwärtanalyse $\boxed{f}(x) = f(x + \Delta x)$ mit

$$\left| \frac{\Delta x}{x} \right| \leq C_R \,\text{eps}$$

und $C_R > 0$ ist nicht zu groß, so ist der Algorithmus $\boxed{f}$ **rückwärtsstabil**.

Bemerkung: Rückwärtsstabile Algorithmen sind auch vorwärtsstabil, denn es gilt

$$\left| \frac{\boxed{f}(x) - f(x)}{f(x)} \right| = \left| \frac{f(x + \Delta x) - f(x)}{f(x)} \right| \lesssim \kappa_{\text{rel}} \left| \frac{\Delta x}{x} \right| \leq \kappa_{\text{rel}} C_R \,\text{eps}\,.$$

Faustregel: Ein gut konditioniertes Problem in Verbindung mit einem stabilen Algorithmus liefert gute numerische Ergebnisse. Ein schlecht konditioniertes Problem oder ein instabiler Algorithmus liefern fragwürdige Ergebnisse.

Beispiel 2.11 Wir betrachten wieder Beispiel 2.4. Für

$$f(x) = p - \sqrt{p^2 - x}$$

mit $|x| \ll 1 < p$ gilt

$$\kappa_{\text{abs}} = |f'(x)| = \frac{1}{2\sqrt{p^2 - x}} < 1$$

und

$$\begin{aligned}
\kappa_{\text{rel}} = \left|\frac{f'(x)\cdot x}{f(x)}\right| &= \left|\frac{x}{2\sqrt{p^2 - x}(p - \sqrt{p^2 - x})}\right| \\
&= \frac{1}{2}\left|\frac{x(p + \sqrt{p^2 - x})}{\sqrt{p^2 - x}\,\underbrace{(p - \sqrt{p^2 - x})(p + \sqrt{p^2 - x})}_{=x}}\right| \\
&= \frac{1}{2}\left|\frac{p + \sqrt{p^2 - x}}{\sqrt{p^2 - x}}\right| \approx 1.
\end{aligned}$$

Die Nullstellenberechnung ist gut konditioniert, folglich ist der Algorithmus instabil. △

3. Dreitermrekursion

3.1 Theoretische Grundlagen

Definition 3.1 Eine Rekursion der Form

$$p_k = a_k p_{k-1} + b_k p_{k-2} + c_k, \qquad k = 2, 3, \ldots \tag{3.1}$$

mit $b_k \neq 0$ heißt **Dreitermrekursion**. Die zugehörige **Rückwärtsrekursion** ist

$$p_{k-2} = -\frac{a_k}{b_k} p_k + \frac{1}{b_k} p_{k-1} + \frac{c_k}{b_k}, \qquad k = n, n-1, \ldots, 2. \tag{3.2}$$

Geht (3.2) aus (3.1) durch Vertauschen von p_k und p_{k-2} hervor, das heißt gilt $b_k = -1$, so heißt die Rekursion **symmetrisch**. Verschwinden alle c_k, so heißt die Rekursion **homogen**, ansonsten **inhomogen**.

Beispiel 3.2

- Sei
$$p_k = a_k p_{k-1} + b_k p_{k-2}, \quad a_k = 2\cos(x), \quad b_k = -1, \qquad k = 2, 3, \ldots$$
mit $p_0 = 1$, $p_1 = \cos(x)$. Dann ist
$$p_k = 2\cos(kx), \qquad k = 2, 3, \ldots .$$
- Die Rekursionsformel für die Tschebbyscheff-Polynome lautet
$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \qquad k = 2, 3, \ldots$$
mit den Startwerten $T_0(x) = 1$ und $T_1(x) = x$.
- Die Fibonacci-Zahlen sind rekursiv derfiniert durch
$$f_k = f_{k-1} + f_{k-2}, \qquad k = 2, 3, \ldots$$
mit den Startwerten $f_0 = 0$ und $f_1 = 1$.

$\triangle$

Algorithmus 3.3 (Dreitermrekursion)

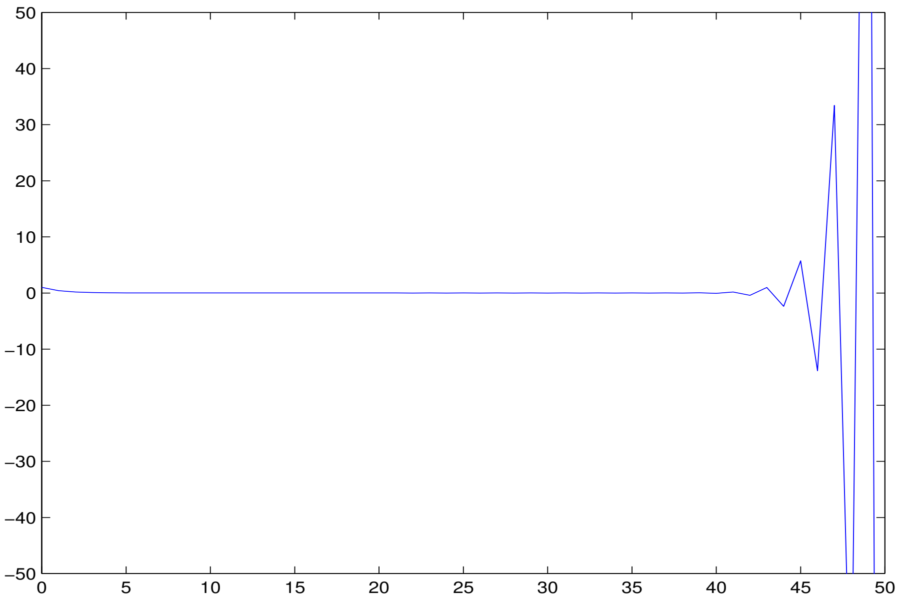
input: `unsigned int n;`
`double a[n], b[n], c[n], p0, p1;`
output: `double p[n]`

```
    p[0] = p0;
    p[1] = p1;
    for (k=2; k<n; k++) p[k] = a[k]*p[k-1]+b[k]*p[k-2]+c[k];
```

Beispiel 3.4 Im Falle der Dreitermrekursion

$$p_k = \begin{cases} 1, & k = 0, \\ \sqrt{2} - 1, & k = 1, \\ -2p_{k-1} + p_{k-2}, & \text{sonst}, \end{cases}$$

liefert der Algorithmus 3.3 das folgende Ergebnis:



△

Die homogene Dreitermrekursion

$$p_k = a_k p_{k-1} + b_k p_{k-2}, \qquad k = 2, 3, \ldots$$

kann umgeschrieben werden gemäß

$$\begin{bmatrix} p_k \\ p_{k-1} \end{bmatrix} = \begin{bmatrix} a_k & b_k \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_{k-1} \\ p_{k-2} \end{bmatrix} = \mathbf{A}_k \begin{bmatrix} p_{k-1} \\ p_{k-2} \end{bmatrix}, \qquad k = 2, 3, \ldots.$$

Rekursiv folgt somit

$$\begin{bmatrix} p_k \\ p_{k-1} \end{bmatrix} = \mathbf{A}_k \mathbf{A}_{k-1} \cdots \mathbf{A}_2 \begin{bmatrix} p_1 \\ p_0 \end{bmatrix} = \mathbf{B}_k \begin{bmatrix} p_1 \\ p_0 \end{bmatrix}$$

mit der Matrix $\mathbf{B}_k = \mathbf{A}_k \mathbf{A}_{k-1} \cdots \mathbf{A}_2 \in \mathbb{R}^{2\times 2}$. Offensichtlich gilt für alle $\alpha, \beta \in \mathbb{R}$

$$\mathbf{B}_k \begin{bmatrix} \alpha p_1 + \beta q_1 \\ \alpha p_0 + \beta q_0 \end{bmatrix} = \alpha \cdot \mathbf{B}_k \begin{bmatrix} p_1 \\ p_0 \end{bmatrix} + \beta \cdot \mathbf{B}_k \begin{bmatrix} q_1 \\ q_0 \end{bmatrix} = \alpha \cdot \begin{bmatrix} p_k \\ p_{k-1} \end{bmatrix} + \beta \cdot \begin{bmatrix} q_k \\ q_{k-1} \end{bmatrix}$$

für alle $\alpha, \beta \in \mathbb{R}$, das heißt die Lösungsfolge $\{p_k\}$ hängt *linear* von den Startwerten $[p_0, p_1]^T \in \mathbb{R}^2$ ab.

Im Fall konstanten Koeffizienten $a = a_k$ und $b = b_k$ folgt

$$\mathbf{B}_k = \mathbf{A}^{k-1}, \qquad \mathbf{A} = \begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix}.$$

Bestimmt man die Eigenwerte λ_1, λ_2 der Matrix $\mathbf{A}$, so kann die Lösung der homogenen Dreitermrekursion geschlossen angegeben werden.

Satz 3.5 Seien λ_1, λ_2 die Nullstellen des charakteristischen Polynoms

$$q(\lambda) = \lambda^2 - a\lambda - b.$$

Dann ist die Lösung der homogenen Dreitermrekursion

$$p_k = ap_{k-1} + bp_{k-2}, \qquad k = 2, 3, \ldots \tag{3.3}$$

gegeben durch

$$p_k = \alpha\lambda_1^k + \beta\lambda_2^k, \qquad k = 0, 1, \ldots$$

mit $\alpha, \beta \in \mathbb{R}$ aus

$$\alpha + \beta = p_0, \qquad \alpha\lambda_1 + \beta\lambda_2 = p_1.$$

Beweis. Wir führen den Beweis vermittels vollständiger Induktion nach k. Für $k = 0$ gilt $p_0 = \alpha + \beta$ während für $k = 1$ gilt $p_1 = \alpha\lambda_1 + \beta\lambda_2$. Wir wollen daher annehmen, dass das Behauptete gilt für ein $k \in \mathbb{N}$. Der Induktionsschritt $k \mapsto k + 1$ ergibt sich nun wie folgt:

$$\begin{aligned} p_{k+1} - ap_{k-1} - bp_{k-2} &= \alpha\lambda_1^{k+1} + \beta\lambda_2^{k+1} - a \cdot (\alpha\lambda_1^k + \beta\lambda_2^k) - b \cdot (\alpha\lambda_1^{k-1} + \beta\lambda_2^{k-1}) \\ &= \alpha\lambda_1^{k-1} \underbrace{(\lambda_1^2 - a\lambda_1 - b)}_{=0} + \beta\lambda_2^{k-1} \underbrace{(\lambda_2^2 - a\lambda_2 - b)}_{=0} = 0. \end{aligned}$$

$\square$

Wir wollen Satz 3.5 auf Beispiel 3.4 anwenden. Das charakteristische Polynom $q(\lambda) = \lambda^2 + 2\lambda - 1$ hat die Nullstellen

$$\lambda_1 = \sqrt{2} - 1, \qquad \lambda_2 = -\sqrt{2} - 1.$$

Aus

$$\alpha + \beta = 1, \qquad \alpha\lambda_1 + \beta\lambda_2 = \sqrt{2} - 1$$

erhält man die Koeffizienten $\alpha = 1$ und $\beta = 0$. Demnach ist die Lösung der Dreitermrekursion also

$$p_k = \lambda_1^k > 0, \qquad k = 0, 1, \ldots.$$

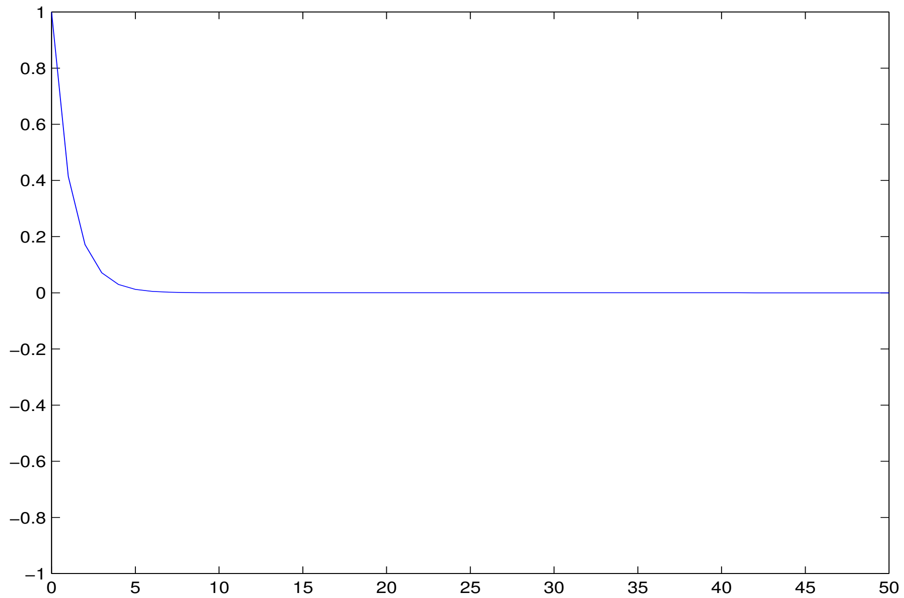
Allgemeiner liefert Satz 3.5 folgenden Algorithmus für die homogene Dreitermrekursion mit konstanten Koeffizienten:

Algorithmus 3.6 (geschlossene Lösung)

input: `unsigned int n;`
`double a, b, p0, p1;`
output: `double p[n];`

```
    lambda1 = -a/2+sqrt(a*a/4-b);
    lambda2 = -a/2-sqrt(a*a/4-b);
    beta = (p1-lambda1*p0)/(lambda2-lambda1);
    alpha = p0-beta;
    p[0] = p0;
    p[1] = p1;
    l1 = lambda1;
    l2 = lambda2;
    for(k=2; k<n; k++)
    {  l1 *= lambda1;
       l2 *= lambda2;
       p[k] = alpha*l1+beta*l2;
       }
```

Im Fall der Werte aus Beispiel 3.4 liefert dieser Algorithmus das folgende Ergebnis:



Dies ist das richtige Ergebnis, da gemäß Satz 3.5 gilt

$$p_k = \lambda_1^k = (\sqrt{2}-1)^k \ll 1.$$

Was läuft also bei Algorithmus 3.3 schief?

Um zu verstehen was passiert ist, untersuchen wir die Kondition unseres Problems. Dazu betrachten wir $f : \mathbb{R} \to \mathbb{R}$ mit $f(p_0, p_1) = p_k$. Gestörte Daten

$$\widehat{p}_0 = 1 + \varepsilon_0, \qquad \widehat{p}_1 = \lambda_1 \cdot (1 + \varepsilon_1), \qquad |\varepsilon_0|, |\varepsilon_1| \leq \text{eps}$$

ergeben gestörte Koeffizienten

$$\widehat{\alpha} = 1 + \varepsilon_0 \frac{\lambda_2}{\lambda_2 - \lambda_1} - \varepsilon_1 \frac{\lambda_1}{\lambda_2 - \lambda_1},$$
$$\widehat{\beta} = (\varepsilon_1 - \varepsilon_0) \frac{\lambda_1}{\lambda_2 - \lambda_1},$$

und schließlich die gestörte Lösung

$$\widehat{p}_k = \left(1 + \varepsilon_0 \frac{\lambda_2}{\lambda_2 - \lambda_1} - \varepsilon_1 \frac{\lambda_1}{\lambda_2 - \lambda_1}\right) \lambda_1^k + (\varepsilon_1 - \varepsilon_0) \frac{\lambda_1}{\lambda_2 - \lambda_1} \lambda_2^k.$$

Folglich ergibt sich der relative Fehler

$$\begin{aligned}\left|\frac{\widehat{p}_k - p_k}{p_k}\right| &= \left|\varepsilon_0 \underbrace{\frac{\lambda_2}{\lambda_2 - \lambda_1}}_{=1+\frac{\lambda_1}{\lambda_2-\lambda_1}} - \varepsilon_1 \frac{\lambda_1}{\lambda_2 - \lambda_1} + (\varepsilon_1 - \varepsilon_0) \frac{\lambda_1}{\lambda_2 - \lambda_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k\right| \\ &= \left|\varepsilon_0 + (\varepsilon_1 - \varepsilon_0) \frac{\lambda_1}{\lambda_2 - \lambda_1} \cdot \left(\left(\frac{\lambda_2}{\lambda_1}\right)^k - 1\right)\right|.\end{aligned}$$

Gilt $|\lambda_2| > |\lambda_1|$, so explodiert der relative Fehler. Genau dies ist in Beispiel 3.4 passiert: Der parasitäre Zweig $|\lambda_2|^k$ wächst exponentiell, während die Lösung selbst exponentiell fällt. Auch im Fall der allgemeinen Dreitermrekursion (3.1) beobachtet man das unterschiedliche Verhalten der Lösungen.

Definition 3.7 Die Lösung $\{p_k\}$ der Dreitermrekursion (3.1) zu den Startwerten p_0, p_1 heißt **Minimallösung**, falls für jede Lösung $\{q_k\}$ zu den von p_0, p_1 linear unabhängigen Startwerten q_0, q_1 gilt

$$\lim_{k\to\infty} \frac{p_k}{q_k} = 0.$$

Die Lösung $\{q_k\}$ wird auch **dominante Lösung** genannt.

Beispiel 3.8 In Beispiel 3.4 ist $\{p_k\}$ genau dann Minimallösung, wenn $\beta = 0$. △

Es ist klar, dass die Minimallösung nur bis auf einen skalaren Faktor eindeutig bestimmt ist. Daher führen wir die Normierung $p_0^2 + p_1^2 = 1$ ein.

Das berechnen der Minimallösung ist extrem schlecht konditioniert, da im Laufe der Rekursion die Rundungsfehler dominieren. Algorithmus 3.3 ist i.a. jedoch stabil, da nur Multiplikationen und Additionen ausgeführt werden.

3.2 Miller-Algorithmus

Betrachte die zur Dreitermrekursion (3.3) gehörende Rückwärtsrekursion

$$q_{k-1} = -\frac{a}{b} q_k + \frac{1}{b} q_{k+1}, \qquad k = n, n-1, \ldots, 1$$

mit den Startwerten $q_{n+1} = 0$ und $q_n = 1$. Das charakteristische Polynom

$$p(\mu) = \mu^2 + \frac{a}{b}\mu - \frac{1}{b}$$

besitzt die Nullstellen

$$\mu_{1,2} = \frac{-a \pm \sqrt{a^2 + 4b}}{2b} = \frac{1}{\lambda_{1/2}}.$$

Aus

$$\alpha + \beta = 0, \qquad \frac{\alpha}{\lambda_1} + \frac{\beta}{\lambda_2} = 1$$

ergeben sich die Koeffizienten

$$\alpha = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \neq 0, \qquad \beta = -\frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} \neq 0.$$

Aufgrund der Lösungsdarstellung

$$q_k = \frac{\alpha}{\lambda_1^{n-k}} + \frac{\beta}{\lambda_2^{n-k}}$$

folgt im Falle $|\lambda_1| < |\lambda_2|$ dass

$$p_k^{(n)} := \frac{q_k}{q_0} = \frac{\frac{\alpha}{\lambda_1^{n-k}} + \frac{\beta}{\lambda_2^{n-k}}}{\frac{\alpha}{\lambda_1^n} + \frac{\beta}{\lambda_2^n}} = \frac{\lambda_1^k + \frac{\beta}{\alpha}\lambda_2^k \left(\frac{\lambda_1}{\lambda_2}\right)^n}{1 + \frac{\beta}{\alpha}\left(\frac{\lambda_1}{\lambda_2}\right)^n} \xrightarrow{n\to\infty} \lambda_1^k.$$

Bei der Wahl von $n = 100$ wird p_{50} aus Beispiel 3.4 auf 13 Nachkommastellen genau berechnet! Dies motiviert folgendem Algorithmus zur Berechnung der Minimallösung:

Algorithmus 3.9 (Miller)

① Wähle n genügend groß.

② Rückwärtsrekursion: berechne

$$\widehat{p}_{k-2} = -\frac{a}{b}\widehat{p}_{k-1} + \frac{1}{b}\widehat{p}_k, \qquad k = n, n-1, \ldots, 2,$$

mit den Startwerten $\widehat{p}_n = 0$ und $\widehat{p}_{n-1} = 1$.

③ Normierung: setze

$$p_k^{(n)} = \frac{\widehat{p}_n}{\sqrt{\widehat{p}_0^2 + \widehat{p}_1^2}}, \qquad k = 0, 1, \ldots, n.$$

Satz 3.10 Sei $\{p_k\}$ Minimallösung der homogenen Dreitermrekursion mit $p_0^2 + p_1^2 = 1$. Dann gilt für die Lösung des Miller-Algorithmus

$$\lim_{n\to\infty} p_k^{(n)} = p_k, \qquad k = 0, 1, \ldots, n.$$

Beweis. Jede Lösung der Dreitermrekursion lässt sich schreiben als Linearkombination der Minimallösung $\{p_k\}$ und einer (normierten) dominanten Lösung $\{q_k\}$. Wegen $\widehat{p}_0 = 0$ und $\widehat{p}_1 = 1$ folgt daher

$$\widehat{p}_k = \frac{p_k q_n - q_k p_n}{p_{n-1} q_n - q_{n-1} p_n} = \frac{q_n}{p_{n-1} q_n - q_{n-1} p_n} \left(p_k - \frac{p_n}{q_n} q_k\right).$$

Aus

$$\begin{aligned}
\widehat{p}_0^2 + \widehat{p}_1^2 &= \frac{q_n^2}{(p_{n-1}q_n - q_{n-1}p_n)^2}\left(\underbrace{p_0^2 + p_1^2}_{=1} + 2\frac{p_n}{q_n}\underbrace{(p_0q_0 + p_1q_1)}_{=:c} + \underbrace{q_0^2 + q_1^2}_{=1}\right) \\
&= \frac{q_n^2}{(p_{n-1}q_n - q_{n-1}p_n)^2}\left(1 + 2c\frac{p_n}{q_n} + \frac{p_n^2}{q_n^2}\right)
\end{aligned}$$

folgt daher

$$p_k^{(n)} = \frac{\widehat{p}_k}{\sqrt{\widehat{p}_0^{\,2} + \widehat{p}_1^{\,2}}} = \frac{p_k - \overbrace{\frac{p_n}{q_n}}^{\to 0} q_k}{\sqrt{1 + 2c\underbrace{\frac{p_n}{q_n}}_{\to 0} + \underbrace{\frac{p_n^2}{q_n^2}}_{\to 0}}} \xrightarrow{n\to\infty} p_k.$$

$\square$

4. Sortieren

4.1 Bubblesort

Wir betrachten folgendes Sortierproblem: Gegeben seien n paarweise verschiedene Zahlen $z_1, z_2, \dots, z_n \in \mathbb{R}$. Gesucht ist eine Permutation $\pi_1, \pi_2, \dots, \pi_n$, so dass $z_{\pi_1} < z_{\pi_2} < \dots < z_{\pi_n}$.

Definition 4.1 Eine **Permutation** π von $\{1, 2, \dots, n\}$ ist eine bijektive Abbildung von $\{1, 2, \dots, n\}$ auf sich selbst. Wir schreiben $\pi(k) = \pi_k$ für alle $k = 1, \dots, n$.

Bemerkung: Die Zahlen $z_1, z_2, \dots z_n$ sollen also der Größe nach sortiert werden. Dieses Problem ist eindeutig lösbar.

Algorithmus 4.2 (Ausprobieren) Probiere so lange alle möglichen Permutationen durch, bis die gewünschte Eigenschaft vorliegt.

Dieser Algorithmus macht keinen sonderlich durchdachten Eindruck.

Satz 4.3 Es gibt $n!$ Permutationen von $\{1, 2, \dots, n\}$.

Beweis. Wir haben

$$\begin{aligned} n &\text{ Möglichkeiten, die erste Zahl auszuwählen,}\\ n-1 &\text{ Möglichkeiten, die zweite Zahl auszuwählen,}\\ &\vdots\\ 1 &\text{ Möglichkeit, die letzte Zahl auszuwählen,} \end{aligned}$$

woraus die Behauptung folgt. □

Bei Algorithmus 4.2 müssen wir im schlimmsten Fall (*worst case*)

$$\underbrace{(n-1)}_{\substack{\text{Vergleiche}\\ \text{pro Permutation}}} \cdot \underbrace{n!}_{\substack{\text{Anzal an}\\ \text{Permutationen}}}$$

Vergleiche durchführen.
Nachfolgender Algorithmus nutzt die transitive Struktur

$$x < y \text{ und } y < z \quad \Longrightarrow \quad x < z$$

der Ordnungsrelation "$<$" aus:

Algorithmus 4.4 (Bubblesort)

① setze $k = n$ und $S_n = \{z_1, z_2, \ldots, z_n\}$

② setze $z_{\pi_k} = \max S_k$

③ setze $S_{k-1} = S_k \setminus \{z_{\pi_k}\}$ und $k := k - 1$

④ falls $k > 0$, dann gehe nach ②.

Beispiel 4.5 (Bubblesort)

$$\begin{aligned} S_3 &= \{1,2,4\}, & k &= 3, & z_{\pi_3} &= \max S_3 = 4, \\ S_2 &= \{1,2\}, & k &= 2, & z_{\pi_2} &= \max S_2 = 2, \\ S_1 &= \{1\}, & k &= 1, & z_{\pi_1} &= \max S_1 = 1. \end{aligned}$$

△

Die zugehörige C-Funktion heißt `bubblesort`. Sie enthält gleich eine kleine Optimierung: Wenn bei einem der inneren Schleifendurchgänge keine Vertauschung mehr vorgenommen wurde, ist die Folge offenbar schon sortiert. Dann kann das Programm schon abgebrochen werden.

```
void bubblesort(z,n)
double          *z;
unsigned int    n;
{
double          x;
unsigned int    k, l, swapped;

for (k=n-1; k>0; k--)
{  swapped = 0;
   for (l=0; l<k; l++)
   {  if (z[l] > z[l+1])
      {  x    = z[l];
         z[l] = z[l+1];
         z[l] = x;
         }
      swapped++;
      }
   if (swapped == 0) break;
   }
return;
}
```

Wir wollen den Aufwand von Algorithmus 4.4 untersuchen. Dabei wollen wir uns auf das *asymptotische Verhalten* des Aufwandes für große n beschränken.

Definition 4.6 Wir schreiben $f(x) = \mathcal{O}\big(g(x)\big)$, falls Konstanten $c, C > 0$ existieren, so dass

$$f(x) \leq c \cdot g(x) \quad \text{für alle } x > C.$$

Bemerkung: Gilt $f(x) = \mathcal{O}\big(g(x)\big)$, so wächst f mit wachsendem x nicht schneller als g.

Beispiel 4.7

- $\sin(x) = O(1)$
- $x^2 + 2x + 1 = O(x^2)$

$\triangle$

Definition 4.8 Der **Aufwand** eines Algorithmus ist die kleinste obere Schranke für das **Aufwandsmaß**.

Bei der Wahl des Aufwandmaßes ignorieren wir hier den Speicherbedarf und berücksichtigen nur die Rechenzeit. Als einfaches Maß für die Rechenzeit wählen wir

$$\text{Aufwandsmaß} \mathrel{\widehat{=}} \text{Anzahl der Vergleiche.}$$

Zur Bestimmung des Aufwands von Algorithmus 4.4 haben wir also Vergleiche zu zählen. Da in der k-ten Schleife k Vergleiche ausgeführt werden, folgt

$$\sum_{k=1}^{n-1} k = \frac{n(n-1)}{2} = \mathcal{O}(n^2)$$

Dies ist eine dramatische Reduktion verglichen mit dem exponentiellen Aufwand von Algorithmus 4.2.

4.2 Mergesort

Lemma 4.9 Gegeben seien die beiden sortierten Mengen

$$S_x = \{x_1 < x_2 < \cdots < x_m\},$$
$$S_y = \{y_1 < y_2 < \cdots < y_n\}.$$

Dann lässt sich die Menge $S = S_x \cup S_y$ mit linearem Aufwand sortieren. Genauer, es werden höchstens $m + n + 1$ Vergleiche benötigt.

Wir führen einen konstruktiven Beweis, indem wir einen entsprechenden Algorithmus angeben:

Algorithmus 4.10 (Merge)

① Eingabe: $x_1 < x_2 < \cdots < x_m$
$y_1 < y_2 < \cdots < y_n$

② Initialisierung: setze $i := 1$, $j := 1$, $k := 1$

③ Direkte Vergleiche:
solange $(i \leq m)$ und $(j \leq n)$: falls $(x_i < y_j)$,
dann setze $z_k := x_i$, $k := k + 1$, $i := i + 1$
sonst setze $z_k := y_j$, $k := k + 1$, $j := j + 1$

④ Hinten anhängen: für alle $\ell = 0, 1, \ldots, m - i$: setze $z_{k+\ell} := x_{i+\ell}$
für alle $\ell = 0, 1, \ldots, n - j$: setze $z_{k+\ell} := y_{j+\ell}$

Wir nehmen nun der Einfachheit halber an, dass $n = 2^m$ für ein $m \in \mathbb{N}$ gilt. Die nachfolgenden Überlegungen lassen sich jedoch auf beliebige n übertragen. Wir zerlegen unser Sortierproblem in viele kleine Probleme (*divide and conquer*):

Algorithmus 4.11 (Mergesort)

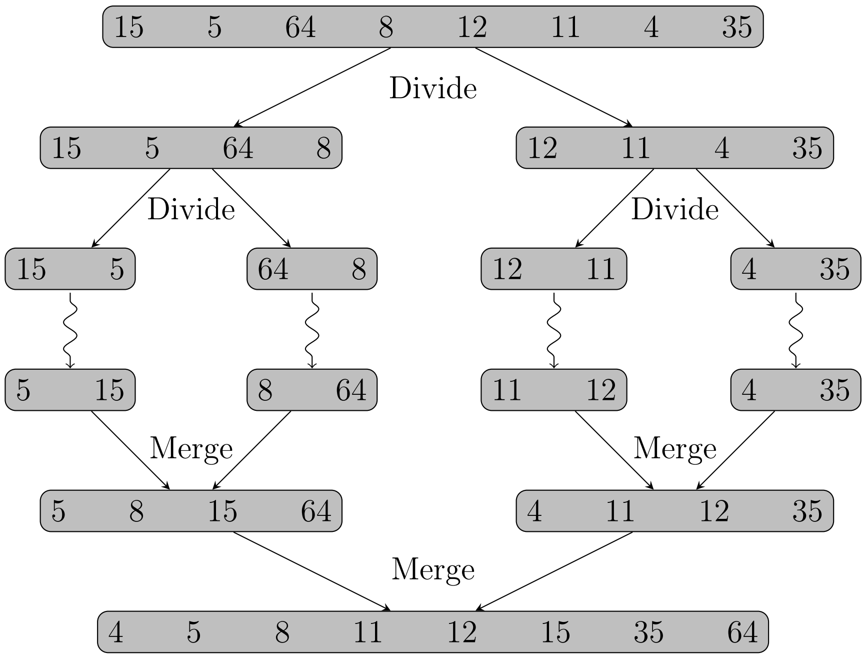
① Ist $n = 1$, so gibt es nichts zu sortieren. Andernfalls sortiere

$$S_x = \{z_1, z_2, \dots, z_{n/2}\}, \qquad S_y = \{z_{n/2+1}, z_{n/2+2}, \dots, z_n\}$$

mit Algorithmus 4.11.

② Verschmelze (merge) die sortierten Teilmengen mit Algorithmus 4.10.

Beispiel 4.12



△

Bemerkung: Beachte, dass Mergesort sich selbst aufruft. Man spricht von einem *rekursiven Algorithmus*. Im allgemeinen ist die Frage, ob ein rekursiver Algorithmus terminiert, höchst kompliziert. Im vorliegenden Fall ist die Sache jedoch klar: Nach m Rekursionsschritten ist man bei der Länge $n = 1$ angelangt und beginnt mit dem Verschmelzen.

Eine mögliche Implementierung von Mergesort für beliebig viele, nicht notwendig paarweise verschiedenen Zahlen ist:

```
void mergesort(z,n)
double          *z;             /* zu sortierendes Feld           */
unsigned int    n;              /* Laenge des Feldes              */
{
unsigned int    m;              /* hier wird das Feld unterteilt */
double          *x, *y;         /* Teilfelder von z               */
unsigned int    i, j, k;        /* Laufvariablen                  */
```

```
if (n == 1) return; /* es ist nichts zu tun */

/* Rekursiver Aufruf / Divide */
m = n/2;
x = (double*) calloc(  m,sizeof(double));
y = (double*) calloc(n-m,sizeof(double));
for (i=0; i<  m; i++) x[i] = z[  i];
for (i=0; i<n-m; i++) y[i] = z[m+i];
mergesort(x,m  );
mergesort(y,n-m);

/* Mergen */
i = j = k = 0;

/* direkte Vergleiche */
while ((i < m) && (j < n-m))
{  if (x[i] <= y[j]) z[k++] = x[i++];
   else              z[k++] = y[j++];
   }

/* hinten anhaengen */
while (i <   m) z[k++] = x[i++];
while (j < n-m) z[k++] = y[j++];

/* Speicherplatz wieder freigeben */
free(x);
free(y);
return;
```

Satz 4.13 Der Aufwand von Mergesort ist höchstens $\mathcal{O}(n \log n)$.

Beweis. Der Aufwand von Mergesort für $n = 2^m$ Elemente sei $A(n)$. Nun gilt

$$\begin{aligned}
A(n) &= \underbrace{2\left(\frac{n}{2} - 1\right)}_{\text{Merge}} + \underbrace{2 \cdot A\left(\frac{n}{2}\right)}_{\text{rekursiver Aufruf}} \\
&= (n-1) + 2\left(\frac{n}{2} - 1\right) + 4 \cdot A\left(\frac{n}{4}\right) \\
&= mn - \sum_{i=0}^{m-1} 2^i \\
&= (m-1)n + 1 \\
&= \mathcal{O}(n \log n),
\end{aligned}$$

da

$$m = \log_2 n = \frac{\log n}{\log 2}$$

□

Da $\log n \ll n$ für große n, haben wir demnach gegenüber Bubblesort eine signifikante Verbesserung erziehlt.

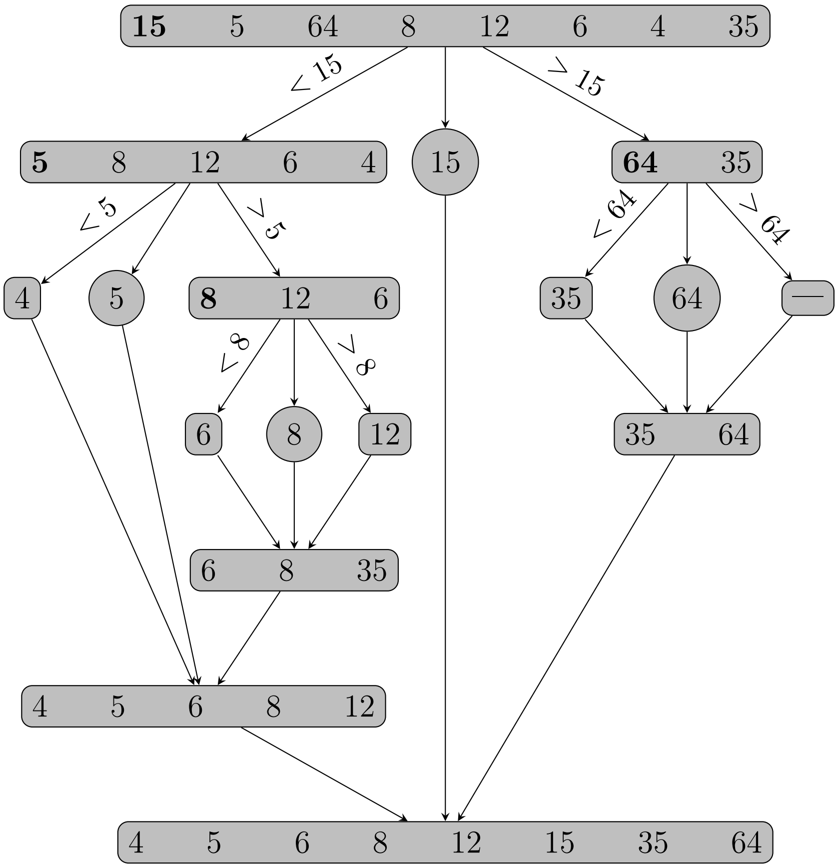
4.3 Quicksort

Quicksort (T. Hoare, 1962) ist eines der berühmtesten Sortierverfahren und in der Praxis weit verbreitet. Die Grundidee ist, die zu sortierenden $z_1, z_2, \dots, z_n$ bezüglich eines Pivotelement $x \in \{z_1, z_2, \dots, z_n\}$ in die zwei Folgen

$$\{z \in \{z_1, z_2, \dots, z_n\} : z < x\}, \qquad \{z \in \{z_1, z_2, \dots, z_n\} : z > x\}$$

zu unterteilen, diese (rekursiv) zu sortieren und dann die sortierten Teilfolgen zusammenzusetzen.

Beispiel 4.14



△

Algorithmus 4.15 (Quicksort)

① Initialisierung: $S = \{z_1, z_2, \dots, z_n\}$

② wähle ein Pivotelement $x \in S$

③ bestimme eine Permutation Π, so dass

$$x = z_{\pi_m}, \quad z_{\pi_1}, z_{\pi_2}, \dots, z_{\pi_{m-1}} < x, \quad z_{\pi_{m+1}}, z_{\pi_{m+2}}, \dots, z_{\pi_n} > x$$

④ Falls $\{z_{\pi_1}, \ldots, z_{\pi_{m-1}}\} \neq \emptyset$, dann $S := \{z_{\pi_1}, \ldots, z_{\pi_{m-1}}\}$ und gehe nach ②.
Falls $\{z_{\pi_{m+1}}, \ldots, z_{\pi_n}\} \neq \emptyset$, dann $S := \{z_{\pi_{m+1}}, \ldots, z_{\pi_n}\}$ und gehe nach ②.

Da in ③ $n-1$ Vergleiche benötigt werden, folgt für den Aufwand die Rekursion

$$A(n) = (n-1) + A(m-1) + A(n-m), \tag{4.1}$$

wobei $A(0) = A(1) = 0$ gilt.

Lemma 4.16 Im schlimmsten Fall ist der Aufwand von Quicksort $\mathcal{O}(n^2)$.

Beweis. Der Aufwand $A(n)$ aus (4.1) wird offensichtlich maximal, falls $m = 1$ oder $m = n$ gilt, das heißt, dass das Pivotelement jeweils an letzer oder erster Stelle steht:

$$A(n) = n - 1 + A(n-1).$$

Analog zu Bubblesort liefert dies den Aufwand

$$A(n) = \frac{n(n-1)}{2}.$$

□

Satz 4.17 Alle Permutationen der Zahlen $\{1, 2, \ldots, n\}$ seien gleichwahrscheinlich. Dann benötigt Quicksort im Mittel $\mathcal{O}(n \log n)$ Vergleiche zum Sortieren einer zufälligen Permutation.

Beweis. Sei Π die Menge aller Permutationen der Zahlen $\{1, 2, \ldots, n\}$, dann ist der mittlere Aufwand gegeben durch

$$\overline{A}(n) = \frac{1}{n!} \sum_{\pi \in \Pi} A(\pi)$$

Wir teilen Π in die Mengen $\Pi_1, \Pi_2, \ldots, \Pi_n$ mit

$$\Pi_k := \{\pi \in \Pi : \pi_1 = k\}.$$

Da das erste Element aus Π_k fest ist, gilt

$$|\Pi_k| = (n-1)!, \quad k = 1, 2, \ldots, n.$$

Für alle $\pi \in \Pi_k$ ergibt die erste Aufteilung in Quicksort zwei Mengen bestehend aus den Permutationen $\pi_<$ von $\{1, 2, \ldots, k-1\}$ und $\pi_>$ von $\{k+1, k+2, \ldots, n\}$. Damit folgt

$$A(\pi) = n - 1 + A(\pi_<) + A(\pi_>),$$

beziehungsweise

$$\sum_{\pi \in \Pi_k} A(\pi) = (n-1)!(n-1) + \sum_{\pi \in \Pi_k} A(\pi_<) + \sum_{\pi \in \Pi_k} A(\pi_>).$$

Wenn π alle Permutationen aus Π_k durchläuft, entstehen für $\pi_<$ alle Permutationen von $\{1, 2, \dots, k-1\}$ und zwar jede genau $\frac{(n-1)!}{(k-1)!}$-mal, da $|\Pi_k| = (n-1)!$. Folglich gilt

$$\sum_{\pi \in \Pi_k} A(\pi_<) = \frac{(n-1)!}{(k-1)!} \sum_{\substack{\pi_< \text{ ist Permutation} \\ \text{von } \{1,2,\dots,k-1\}}} A(\pi_<) = (n-1)! \cdot \overline{A}(k-1),$$

und analog

$$\sum_{\pi \in \Pi_k} A(\pi_>) = (n-1)! \cdot \overline{A}(n-k).$$

Zusammengesetzt folgt demnach die Rekursion

$$\begin{aligned}
\overline{A}(n) &= \frac{1}{n!} \sum_{\pi \in \Pi} A(\pi) \\
&= \frac{1}{n!} \sum_{k=1}^{n} \sum_{\pi \in \Pi_k} A(\pi) \\
&= \frac{1}{n!} \sum_{k=1}^{n} (n-1)! \Big\{ n - 1 + \overline{A}(k-1) + \overline{A}(n-k) \Big\} \\
&< n - 1 + \frac{1}{n} \sum_{k=1}^{n} \Big\{ \overline{A}(k-1) + \overline{A}(n-k) \Big\} \\
&= n - 1 + \frac{2}{n} \sum_{k=0}^{n-1} \overline{A}(k)
\end{aligned}$$

mit den Startwerten $\overline{A}(0) = \overline{A}(1) = 1$

Induktiv zeigt man

$$\overline{A}(n) = 2(n+1) \sum_{i=1}^{n} \frac{1}{i} - 4n.$$

Die Partialsumme $\sum_{i=1}^{n} \frac{1}{i}$ der harmonischen Reihe lässt sich elementar durch ein Integral abschätzen:

$$\sum_{i=1}^{n} \frac{1}{i} \leq 1 + \int_1^n \frac{1}{x} \, \mathrm{d}x = 1 + \log n - \log 1 = 1 + \log n.$$

Daher folgt

$$\overline{A}(n) \leq 2(n+1) \log n - 2(n-1).$$

□

4.4 Untere Schranken für das Sortierproblem

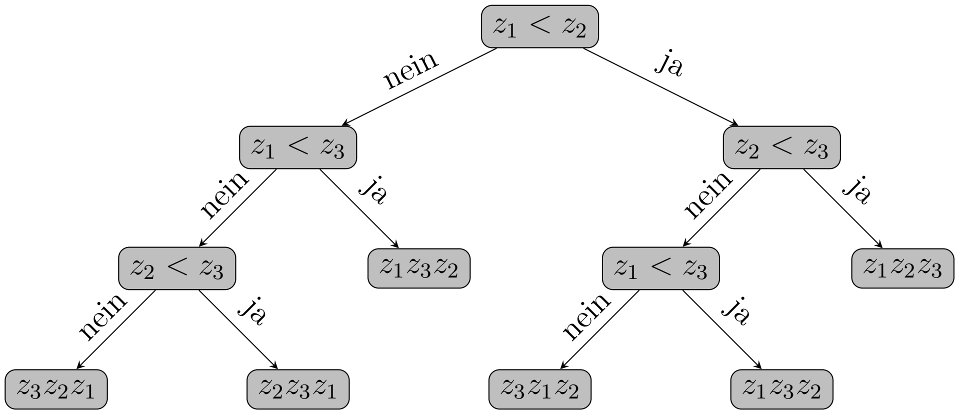
Mergesort ist optimal in dem Sinn, dass sich die asymptotische Laufzeit nicht unterbieten lässt. Diese Beobachtung ist in folgendem Satz formuliert.

Satz 4.18 Jedes deterministische Sortierverfahren, das auf paarweisen Vergleichen beruht und keine Vorkenntnisse über die zu sortierende Zahlenfolge hat, benötigt im schlimmsten Fall mindestens $\log_2(n!)$ Vergleiche zum Sortieren von n verschiedenen Zahlen.

Der Beweis beruht auf der Inspektion des *Entscheidungsbaums*, der jedem Sortierverfahren mit paarweisen Vergleichen für festes n zugeordnet werden kann. Allgemein können wir den Entscheidungsbaum wie folgt beschreiben:

- Innere Knoten des Entscheidungsbaums sind Vergleiche im Algorithmus.
- Ein Weg von der Wurzel zu einem Blatt entspricht der zeitlichen Abfolge der Vergleiche. Ein Weitergehen nach rechts entspricht einem richtigen Vergleich, nach links einem falschen.
- Die $n!$ Blätter des Baumes stellen die Permutationen der Eingabefolge dar, die jeweils zur Abfolge der Vergleiche gehört.

Beispiel 4.19 Der Entscheidungsbaum im Fall $\{z_1, z_2, z_3\}$ ist wie folgt gegeben:



△

Beweis. [Beweis von Satz 4.18] Wir müssen eine untere Schranke für die Höhe des Entscheidungsbaumes finden. Bestenfalls ist er ausbalanciert, das heißt alle Blätter liegen in der gleichen Ebene. Ein ausbalancierter Baum hat also pro Ebene $1, 2, 4, 8, \ldots, 2^m$ Knoten. Da wir $n!$ Blätter haben, folgt $2^m = n!$, beziehungsweise

$$\begin{aligned} m &= \log_2(n!) \\ &\geq \log_2\Bigg(\underbrace{\frac{n}{2}\cdot\frac{n}{2}\cdot\frac{n}{2}\cdot\cdots\cdot\frac{n}{2}}_{\lceil n/2\rceil \text{ Therme}}\Bigg)\underbrace{\left\lfloor\frac{n}{2}\right\rfloor}_{\geq 1} \\ &\geq \log_2\left(\left(\frac{n}{2}\right)^{n/2}\right) \\ &= \frac{n}{2}(\log_2 n - 1). \end{aligned}$$

□

Bemerkung: Die Funktion $f(n) = \log_2(n!)$ verhält sich im wesentlichen wie $n \log_2 n$, denn wir haben auch die Abschätzung

$$\log_2(n!) \leq \log_2(n^n) = n \log_2 n.$$

Satz 4.20 Alle Permutationen der Zahlen $z_1, z_2, \ldots, z_n$ seien gleichhäufig. Dann benötigt das Sortierverfahren aus Satz 4.18 im Mittel mindestens $\log_2 n!$ Vergleiche.

Beweis. Es bezeichne $h_\tau(\nu)$ die Höhe des Knotens ν bezüglich des Entscheidungsbaums τ. Sei $B(\tau)$ die Menge aller Blätter in τ und $\beta(\tau) = |B(\tau)|$. Die mittlere Höhe $\overline{H}(\tau)$ ist definiert als

$$\overline{H}(\tau) = \frac{1}{\beta(\tau)} \sum_{\nu \in B(\tau)} h_\tau(\nu).$$

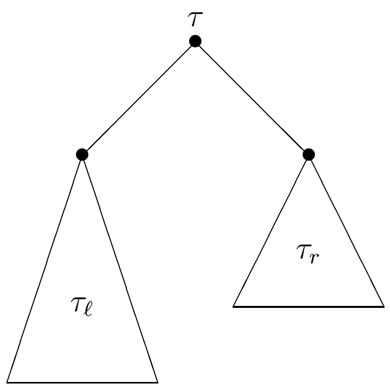
Der Satz ist bewiesen, wenn wir zeigen können, dass gilt

$$\overline{H}(\tau) \geq \log_2 \beta(\tau)$$

Dies zeigen wir durch vollständige Induktion über die Höhe von τ.
Ist $\overline{H}(\tau) = 0$, so besteht τ nur aus dem Wurzelknoten ν mit $h_\tau(\nu) = 0$ und es gilt

$$\overline{H}(\tau) = 0 = \log_2 1 = \log_2 \big(\beta(\tau)\big).$$

Wir nehmen nun an, die Behauptung sei wahr für $\overline{H}(\tau) = 0, 1, \ldots, h-1$. Seien nun $\overline{H}(\tau) = h$ und τ_ℓ, τ_r der linke beziehungsweise rechte Teilbaum des Wurzelknotens von τ:



Dann gilt

$$H(\tau_\ell), H(\tau_r) < h$$

und

$$B(\tau) = B(\tau_\ell) \cup B(\tau_r), \qquad B(\tau_\ell) \cap B(\tau_r) = \emptyset,$$

sowie

$$\beta(\tau) = \beta(\tau_\ell) + \beta(\tau_r), \qquad \beta(\tau_\ell), \beta(\tau_r) \geq 1.$$

Da für jeden Knoten $\nu \in \tau_\ell$ gilt

$$h_\tau(\nu) = h_{\tau_\ell}(\nu) + 1,$$

und analog für Knoten aus τ_r, folgt

$$\begin{aligned}
\overline{H}(\tau) &= \frac{1}{\beta(\tau)} \sum_{\nu \in B(\tau)} h_\tau(\nu) \\
&= \frac{1}{\beta(\tau)} \Bigg(\sum_{\nu \in B(\tau_\ell)} h_\tau(\nu) + \sum_{\nu \in B(\tau_r)} h_\tau(\nu) \Bigg) \\
&= \frac{1}{\beta(\tau)} \Bigg(\sum_{\nu \in B(\tau_\ell)} \big(h_{\tau_\ell}(\nu) + 1\big) + \sum_{\nu \in B(\tau_r)} \big(h_{\tau_r}(\nu) + 1\big) \Bigg) \\
&= \frac{1}{\beta(\tau)} \Bigg(\sum_{\nu \in B(\tau_\ell)} h_{\tau_\ell}(\nu) + \sum_{\nu \in B(\tau_r)} h_{\tau_r}(\nu) \Bigg) + \underbrace{\frac{\beta(\tau_\ell) + \beta(\tau_r)}{\beta(\tau)}}_{=1} \\
&= 1 + \frac{\beta(\tau_\ell)}{\beta(\tau)} \overline{H}(\tau_\ell) + \frac{\beta(\tau_r)}{\beta(\tau)} \overline{H}(\tau_r).
\end{aligned}$$

Einsetzten der Induktionsvorraussetzung

$$\overline{H}(\tau_\ell) \geq \log_2 \big(\beta(\tau_\ell)\big), \qquad \overline{H}(\tau_r) \geq \log_2 \big(\beta(\tau_r)\big)$$

in die obige Formel ergibt

$$\begin{aligned}
\overline{H}(\tau) &\geq 1 + \frac{\beta(\tau_\ell)}{\beta(\tau)} \log_2 \beta(\tau_\ell) + \frac{\beta(\tau_r)}{\beta(\tau)} \log_2 \beta(\tau_r) \\
&= 1 + \frac{1}{\beta(\tau)} \Big\{ \beta(\tau_\ell) \log_2 \big(\beta(\tau_\ell)\big) + \big(\beta(\tau) - \beta(\tau_\ell)\big) \log_2 \big(\beta(\tau) - \beta(\tau_\ell)\big) \Big\}.
\end{aligned}$$

Da wir nicht genau wissen, wie groß $\beta(\tau_\ell)$ ist, fassen wir die rechte Seite dieser Gleichung als Funktion von $x := \beta(\tau_\ell)$ auf und suchen ihr Minimum:

$$\overline{H}(\tau) \geq 1 + \frac{1}{b} \min_{x \in [0,b]} \Big\{ x \log_2 x + (b - x) \log_2 (b - x) \Big\}, \qquad b := \beta(\tau_\ell).$$

Kurvendiskusion von $f(x) := x \log_2 x + (b - x) \log_2 (b - x)$ liefert

$$\begin{aligned}
f'(x) &= \log_2 x + \frac{1}{\log 2} - \log_2 (b - x) - \frac{1}{\log 2} \\
&= \log_2 \frac{x}{b - x} \\
&\stackrel{!}{=} 0,
\end{aligned}$$

dies bedeutet $x = b/2$. Wegen

$$f(0) = f(b) = (b) \log_2 (b) \geq b \log_2 \Big(\frac{b}{2}\Big) = f\Big(\frac{b}{2}\Big),$$

muss $x = b/2$ das Minimum von f in $[0, b]$ sein. Daher folgt

$$\overline{H}(\tau) \geq 1 + \log_2 \big((\beta(\tau)\big) - 1 = \log_2 \big(\beta(\tau)\big).$$

□

Bemerkung: Wir haben gezeigt, dass der Entscheidungsbaum τ bezüglich n Zahlen den Abschätzungen

$$H(\tau) \geq \log_2\big(\beta(\tau)\big), \qquad \overline{H}(\tau) \geq \log_2\big(\beta(\tau)\big)$$

genügt. Diese werden auch als *informationstheoretische Schranken* bezeichnet: Jeder deterministische Sortieralgorithmus muss zwischen $n!$ Permutationen unterscheiden und hierzu $\log_2(n!)$ Bits an Informationen sammeln.

5. Graphen

5.1 Grundbegiffe

Definition 5.1 Ein **Graph** ist ein Paar $G = (V, E)$, bestehend aus endlichen Mengen V von **Knoten** (*vertices*) und E von **Kanten** (*edges*). Die Kanten stehen für Verbindungen zwischen verschiedenen Knoten und können gerichtet oder ungerichtet sein:

- Für **gerichtete Graphen**, kurz auch **Digraphen** genannt, ist

  $$E \subseteq \{(v, w) \in V \times V : v \neq w\}.$$

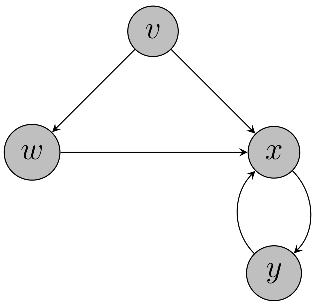
  Ist $e = (v, w) \in E$, dann heißt v der **Anfangsknoten** und w der **Endknoten** der Kante e.

- Für **ungerichtete Graphen** ist E eine Menge von ungeordneten Paaren (v, w) mit $v, w \in V$ mit $v \neq w$.

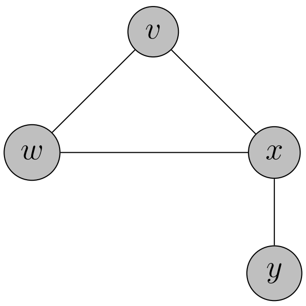
Beachte: Es gilt $(v, w) = (w, v)$, falls (v, w) eine Kante in einem ungerichteten Graphen ist. Dies gilt jedoch nicht im Fall gerichteter Graphen.

Beispiel 5.2

- Gerichteter Graph $G = \{(v, w), (w, x), (v, x), (x, y), (y, x)\}$:



- Ungerichteter Graph $G = \{(v, w), (w, x), (v, x), (x, y)\}$:



△

Bemerkung: In gerichteten Graphen werden manchmal auch Kanten der Form (v, v) zugelassen. Man spricht dann von *Schleifen*.

Definition 5.3 Sei $G = (V, E)$ ein Graph. Ein **Weg** oder **Pfad** in G ist eine Knotenfolge

$$\pi = v_0, v_1, \ldots, v_r$$

mit $r \geq 1$ und $(v_i, v_{i+1}) \in E$, $i = 0, 1, \ldots, r-1$. Wir sprechen von einem Weg, der **von v nach w** führt, wenn der Anfangsknoten $v_0 = v$ und der Endknoten $v_r = w$ ist. Der Weg ist **einfach**, falls gilt:

- entweder sind $v_0, v_1, \ldots, v_r$ paarweise verschieden
- oder $v_0 = v_r$ und $v_0, v_1, \ldots, v_{r-1}$ sind paarweise verschieden.

Die **Länge** $|\pi|$ von π ist r, das heißt, die Anzahl der Kanten, die in π durchlaufen werden. Ein Knoten w heißt von Knoten v **erreichbar**, falls ein Weg $\pi = v_0, v_1, \ldots, v_r$ in G existiert, so dass $v = v_0$ und $w = v_r$.

Beispiel 5.4 Beispiele für Wege in den beiden Graphen von Beispiel 5.2 sind

- $\pi_1 = v, w$ (Länge 1),
- $\pi_2 = v, w, x$ (Länge 2),
- $\pi_3 = v, w, x, y, x, y$ (Länge 5).

Im ungerichteten Fall ist auch $\pi_4 = v, w, x, v$ (Länge 3) ein Weg. △

Definition 5.5 Es sei $G = (V, E)$ ein Graph und $v \in V$. Wir definieren:

- die Menge der **(direkten) Nachfolger** von v
$$\mathrm{post}(v) = \{w \in V : (v, w) \in E\},$$
- die Menge der **(direkten) Vorgänger** von v
$$\mathrm{pre}(v) = \{w \in V : (w, v) \in E\},$$
- die Menge aller von v erreichbaren Knoten
$$\mathrm{post}^\star(v) = \{w \in V : \text{es existiert ein Weg von } v \text{ nach } w\},$$
- die Menge aller Knoten, die v erreichen können,
$$\mathrm{pre}^\star(v) = \{w \in V : v \in \mathrm{post}^\star(w)\}.$$

Ein Knoten $w \in \mathrm{post}(v) \cup \mathrm{pre}(v)$ ist ein **Nachbar** von v.
Die obigen Definition erweitern wir auf Knotenmengen, zum Beispiel ist für $W \subseteq V$

$$\mathrm{pre}(W) = \bigcup_{w \in W} \mathrm{pre}(w).$$

Beispiel 5.6 Für den gerichteten Graphen aus Beispiel 5.2 ist

$$\operatorname{post}(v) = \{w, x\}, \qquad \operatorname{post}^\star(v) = \{w, x, y\}, \qquad \operatorname{pre}(v) = \operatorname{pre}^\star(v) = \emptyset$$

und

$$\operatorname{post}(y) = \{x\}, \qquad \operatorname{post}^\star(y) = \{x, y\}.$$

Im ungerichteten Fall ist jedoch

$$\operatorname{post}(x) = \{v, w, y\}, \qquad \operatorname{post}^\star(x) = \{v, w, x, y\}.$$

△

Für ungerichtete Graphen gilt $\operatorname{post}(v) = \operatorname{pre}(v)$ und $\operatorname{post}^\star(v) = \operatorname{pre}^\star(v)$. Insbesondere gilt

$$\sum_{v \in V} |\operatorname{post}(v)| = \sum_{v \in V} |\operatorname{pre}(v)| = 2|E|.$$

Im gerichteten Fall können $\operatorname{post}(v)$ und $\operatorname{pre}(v)$ zwar völlig verschieden sein, jedoch gilt

$$\sum_{v \in V} |\operatorname{post}(v)| = \sum_{v \in V} |\operatorname{pre}(v)| = |E|.$$

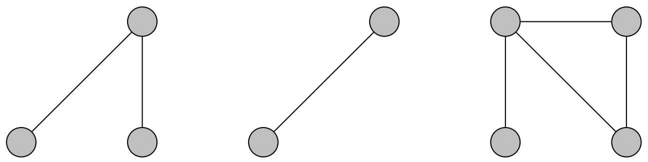
5.2 Zusammenhang

Definition 5.7 Sei $G = (V, E)$ ein ungerichteter Graph und $C \subseteq V$. C heißt **zusammenhängend**, falls je zwei Knoten $v, w \in C$, $v \neq w$, voneinander erreichbar sind, das heißt, falls gilt $w \in \operatorname{post}^\star(v)$ und $v \in \operatorname{post}^\star(w)$. Ist der Graph G gerichtet, so heißt C zusammenhängend, falls C im zugrundeliegenden ungerichteten Graphen zusammenhängend ist.

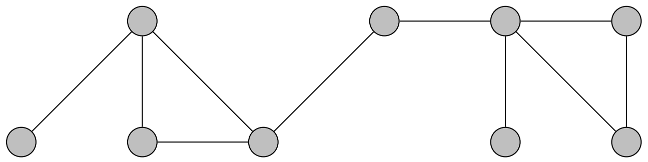
C heißt **Zusammenhangskomponente** von G, falls C eine nicht-leere maximale zusammenhängende Knotenmenge ist. Maximalität bedeutet hier, dass C in keiner anderen zusammenhängenden Menge $C' \subseteq V$ echt enthalten ist (also $C \neq C'$). Der Graph G heißt zusammenhängend, falls V zusammenhängend ist.

Beispiel 5.8

- Ein unzusammenhängender Graph mit drei Zusammenhangskomponenten ist:



- Ein zusammenhängender Graph ist:



△

Offenbar sind die Zusammenhangskomponenten eines ungerichteten Graphs die Äquivalenzklassen der Knotenmenge V unter der Erreichbarkeits-Äquivalenzrelation "$\equiv$", wobei

$$v \equiv w :\iff v \cup \mathrm{post}^\star(v) = w \cup \mathrm{post}^\star(w).$$

Insbesondere zerfällt G in paarweise disjunkte Zusammenhangskomponenten $C_1, C_2, \ldots, C_r$ mit

$$V = \bigcup_{i=1}^{r} C_i, \qquad E = \bigcup_{i=1}^{s} E_i,$$

wobei $E_i := E \cap (C_i \times C_i)$.

Satz 5.9 Ist $G = (V, E)$ ein ungerichteter Graph mit $n = |V| \geq 1$ Knoten und $m = |E|$ Kanten, so gilt: Ist G zusammenhängend, dann ist $m \geq n - 1$.

Beweis. Wir beweisen die Aussage durch Induktion nach n. Für $n = 1$ folgt $m = 0 = n-1$; im Fall $n = 2$ ist G zusammenhängend genau dann, wenn $m = 1 = n - 1$. Wir nehmen nun an, dass $n \geq 3$ und G zusammenhängend ist. Wähle $v \in V$, so dass

$$|\,\mathrm{post}(v)| = \min_{w \in V} |\,\mathrm{post}(w)| =: k.$$

Es gilt $k > 0$, denn sonst wäre v ein isolierter Knoten, was im Widerpruch zu G zusammenhängend steht. Im Fall $k \geq 2$ folgt

$$2m = 2|E| = \sum_{w \in V} \underbrace{|\,\mathrm{post}(w)|}_{\geq k} \geq n \cdot k \geq 2 \cdot n$$

und folglich $m \geq n \geq n - 1$.

Für $k = 1$ ergibt sich die Aussage wie folgt: Es sei $G' = (E', V')$ derjenige Graph, der durch Streichen des Knotens v sowie der ausgehenden Kante entsteht. Mit G ist auch G' zusammenhängend und nach Induktionsvoraussetzung folgt wegen $|V'| = n - 1$ und $|E'| = m - 1$

$$m - 1 = |E'| \geq (n - 1) - 1 = n - 2,$$

das heißt $n \geq n - 1$. $\square$

Definition 5.10 Ein **einfacher Zyklus** oder **einfacher Kreis** in einem Graph $G = (V, E)$ ist ein einfacher Weg $\pi = v_0, v_1, \ldots, v_r$ mit $v_0 = v_r$ und $r \geq 2$ (gerichteter Graph) bzw. $r \geq 3$ (ungerichteter Graph). Ein **Zyklus** oder **Kreis** ist ein Weg $v_0, v_1, \ldots, v_r$, der sich aus einfachen Zyklen zusammensetzt.

Bemerkung: Formal bedeutet diese Definition:

(i) jeder einfache Zyklus ist ein Zyklus,

(ii) sind $\pi_1 = v_0, v_1, \ldots, v_r$ und $\pi_2 = w_0, w_1, \ldots, w_\ell$ Zyklen mit $v_i = w_0 = w_\ell$, so ist auch

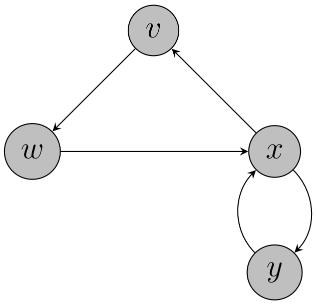
$$\pi = v_0, v_1, \ldots, v_{i-1}, w_0, w_1, \ldots, w_\ell, v_{i+1}, v_{i+2}, \ldots, v_r$$

ein Zyklus,

(iii) nur die durch (i) und (ii) generierbaren Wege sind Zyklen.

Beispiel 5.11

- Der gerichtete Graph



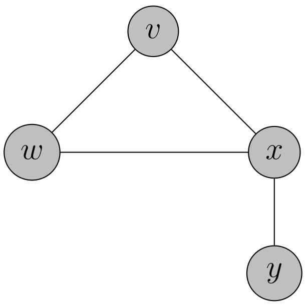
besitzt die einfachen Zyklen

$$\pi_1 = x, y, x, \qquad \pi_2 = v, w, x, v,$$

und die nicht einfachen Zyklen

$$\pi_3 = x, y, x, y, x, \qquad \pi_4 = v, w, x, y, x, v.$$

- Der ungerichtete Graph



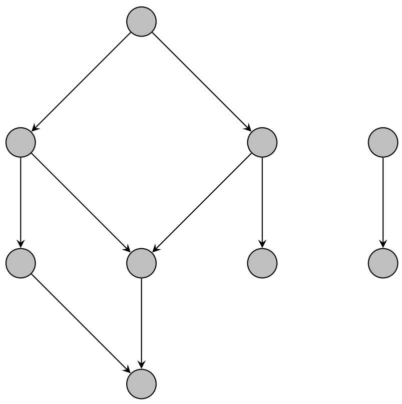
besitzt den (einfachen) Zyklus $\pi_1 = v, w, x, v$. Die Wege $\pi_2 = x, y, x$ und $\pi_3 = v, w, x, y, x, v$ sind hingegen keine Zyklen!

△

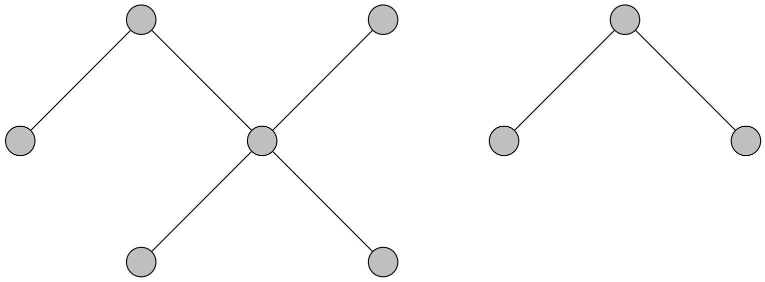
Definition 5.12 Ein Graph heißt **azyklisch** oder **zyklenfrei**, falls es keine Zyklen in G gibt. Ein ungerichteter, azyklischer und zusammenhängender Graph ist ein **Baum**.

Beispiel 5.13

- Azyklischer, gerichteter Graph:



- Azyklischer, ungerichteter Graph:



△

Satz 5.14 Sei $G = (V, E)$ ein ungerichteter Graph mit n Knoten. Dann sind folgende Aussagen äquivalent:

1. G ist ein Baum.
2. G hat $n-1$ Kanten und ist zusammenhängend.
3. G hat $n-1$ Kanten und ist azyklisch.
4. G ist azyklisch und das Hinzufügen einer beliebigen Kante erzeugt einen Zyklus.
5. G ist zusammenhängend und das Entfernen einer Kante erzeugt einen unzusammenhängenden Graphen.
6. Jedes Paar von verschiedenen Knoten in G ist durch genau einen einfachen Weg miteinander verbunden.

Beweis.

$1 \Rightarrow 6$: Dies folgt aus der Tatsache, dass die Vereinigung zweier disjunkter einfacher Wege mit gleichen Anfangs- und Endpunkten ein Zyklus ist.

$6 \Rightarrow 5$: G ist zusammenhängend gemäß Definition. Das Entfernen der Kante (v, w) macht w unerreichbar von v.

$5 \Rightarrow 4$: G ist azyklisch, denn sonst kann eine Kante entfernt werden, so dass G weiterhin zusammenhängend ist. Da es in G stets einen Weg von v nach w gibt, liefert das Hinzufügen einer Kante (v, w) einen Zyklus.

$4 \Rightarrow 3 \Rightarrow 2$: Die Behauptung folgt, falls für einen azyklischen, ungerichteten Graphen gilt

$$n = m + p, \tag{5.1}$$

wobei $m = |E|$ und p die Anzahl der Zusammenhangskomponenten ist. Da (5.1) klar ist für $m = 0$, nehmen wir an, (5.1) gilt für ein $|E| = m$. Fügen wir eine zusätzliche Kante hinzu, dies bedeutet $|E| = m + 1$, so muss sich p um eins reduzieren, denn sonst würde ein Zyklus entstehen.

$2 \Rightarrow 1$: Wir zerstören Zyklen aus G durch Entfernen von Kanten. Haben wir etwa k Kanten entfernt, so folgt aus (5.1)

$$\underbrace{n - 1 - k}_{\text{Kanten}} + \underbrace{p}_{=1} = n,$$

das heißt $k = 0$.

□

5.3 Implementierung von Graphen

Die einfachste Art Graphen am Rechner zu speichern ist die Verwendung von *Adjazenzmatrizen*.

Definition 5.15 Ein Graph $G = (V, E)$ mit $V = \{1, 2, \ldots, n\}$ kann durch eine **Adjazenzmatrix** oder **Nachbarschaftsmatrix**

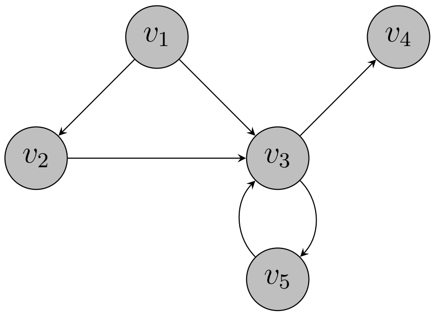
$$\mathbf{A} = [a_{i,j}]_{i,j=1}^{n} \in \mathbb{R}^{n \times n}$$

mit

$$a_{i,j} = \begin{cases} 1, & \text{falls } (i,j) \in E, \\ 0, & \text{sonst,} \end{cases}$$

dargestellt werden.

Beispiel 5.16 Der gerichtete Graph



besitzt die Adjazenzmatrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Beim zugrundeliegenden ungerichteten Graphen ist

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

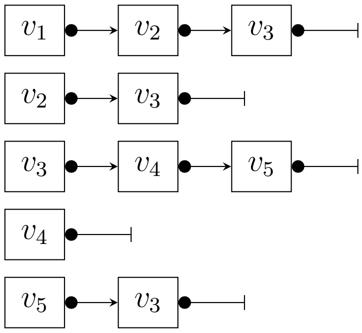
$\triangle$

Bemerkung: Bei ungerichteten Graphen ist $\mathbf{A}$ stets symmetrisch, das heißt es gilt $a_{i,j} = a_{j,i}$ für alle $1 \leq i, j \leq n$.

Die Adjazenzmatrix besitzt immer den Speicherplatzbedarf $|V|^2$, unabhängig von der Anzahl $|E|$ der Kanten. Platzeffizienter ist die Darstellung von Graphen $G = (V, E)$ durch *Adjazenzlisten*:

Definition 5.17 Die **Adjazenzliste** oder **Nachbarschaftsliste** zu einem Knoten $v \in V$ enthält einen Knoten $w \in V$ genau dann, wenn $(v, w) \in E$.

Beispiel 5.18 Die Adjazenzlisten zum gerichteten Graphen aus Beispiel 5.16 werden wie folgt dargestellt



△

Adjazenzlisten können gespeichert werden als *einfach verkettete Listen* mit Elementen der Form

```
struct node
{  unsigned int  number;
   struct node  *next;
   };
```

Sind `A` und `B` vom Typ `struct node`, so enthält `A.number` bzw. `B.number` die Nummer des Knotens, während die Zuweisung

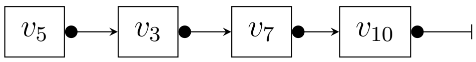
```
A.next = &B;
```

`B` als Nachfolger von `A` definiert. Anstelle von `B` kann man dann auch `A.next` schreiben. Das Ende der Liste wird durch

```
B.next = NULL;
```

markiert. Merken muss man sich jeweils den Anfang der Liste, auch *Listenkopf* genannt. Dies geschieht meist durch ein Feld der Länge $|V|$.

Bemerkung: Eine weitere Möglichkeit, Listen zu speichern, bietet die Darstellung durch zwei Felder `number[n]`, `next[n]` und einer Variablen `Kopf`. Beispielsweise kann die Liste



auch realisiert werden durch

	number	next
1	7	4
2	3	1
3	5	2
4	10	0

`Kopf`: 3

Für gerichtete Graphen habe Adjazenzlisten den Seicherplatzbedarf $|V| + |E|$. Für ungerichtete Graphen ist der Platzbedarf $|V| + 2|E|$, da jede Kante in zwei Adjazenzlisten vorkommt.

5.4 Graphendurchmusterung

Häufig muss ein Graph durchmustert werden. Populäre Graphendurchmusterungsmethoden sind die *Tiefensuche* und die *Breitensuche*. Beide lassen sich auf folgenden Algorithmus zurückführen, der alle von einem Startknoten s erreichbaren Knoten durchsucht.

Algorithmus 5.19 (algorithmische Suche)

input: Graph $G = (V, E)$ und Startknoten $s \in V$

output: azyklischer Graph $G' = (R, T)$ mit $R = \{s\} \cup \mathrm{post}^\star(s)$ und $T \subseteq E$

① Initialisierung: $R := \{s\}$, $Q := \{s\}$, $T = \emptyset$.

② Falls $Q = \emptyset$ dann `stop`, sonst wähle $v \in Q$.

③ Wähle $w \in V \setminus R$ mit $e = (v, w) \in E$. Falls kein solches w existiert, dann setze $Q := Q \setminus \{v\}$ gehe nach ②.

④ Setze $R := R \cup \{w\},\ Q := Q \cup \{w\},\ T = T \cup \{e\}$ und gehe nach ②.

Satz 5.20 Algorithmus 5.19 liefert einen azyklischen Graphen $G' = (R, T)$ mit $R = \{s\} \cup \mathrm{post}^\star(s)$ und $T \subseteq E$.

Beweis. Zu jedem Zeitpunkt des Algotithmus ist (R, T) zusammenhängend. Insbesondere ist (R, T) azyklisch, denn eine neue Kante $e = (v, w)$ verbindet stets Knoten $v \in Q \subseteq R$ und $w \in V \setminus R$.

Angenommen, am Ende existiert ein von s erreichbarer Knoten $w \in V \setminus R$. Sei $\pi = s, v_1, \ldots, v_r, w$ ein einfacher s-w-Weg in G und $(x, y) \in E$ eine Kante mit $x \in R$ und $y \notin R$. Da $x \in R$ ist, muss irgendwann bei Ausführung des Algorithmus $x \in Q$ gelten. Der Algorithmus terminiert jedoch nicht bevor x aus Q entfernt ist. Dies geschieht jedoch nur, falls $(x, y) \notin E$. □

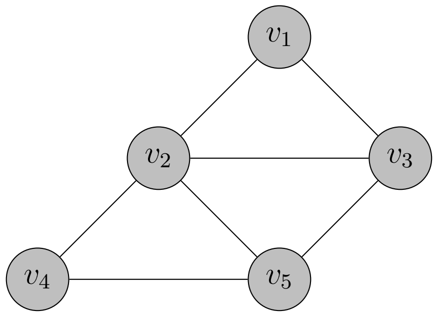
Satz 5.21 Die Ausführung von "wähle $v \in Q$" und "wähle $w \in V \setminus R$ mit $e = (v, w) \in E$" sei in $\mathcal{O}(1)$ durchführbar. Dann besitzt Algorithmus 5.19 die Komplexität $\mathcal{O}(|V| + |E|)$.

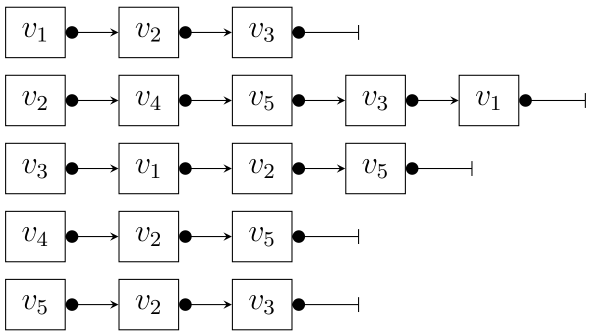
Beweis. Jeder Knoten $v \in V$ wird höchstens $(|\mathrm{post}(v)| + 1)$-mal und jede Kante $e \in E$ höchstens einmal betrachtet. □

Je nachdem, wie die Menge Q verwaltet wird, ergeben sich verschiedene Resultate. Wir unterscheiden:

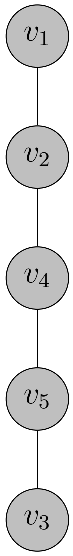
Definition 5.22 Bei der **Tiefensuche** oder **Depth-First-Search (DFS)** wird derjenige Knoten $v \in Q$ ausgewählt, der zuletzt zu Q hinzugefügt wurde. Bei der **Breitensuche** oder **Breadth-First-Search (BFS)** wird derjenige Knoten $v \in Q$ ausgewählt, der zuerst zu Q hinzugefügt wurde.

Beispiel 5.23 Der Graph $G = (V, E)$ sei

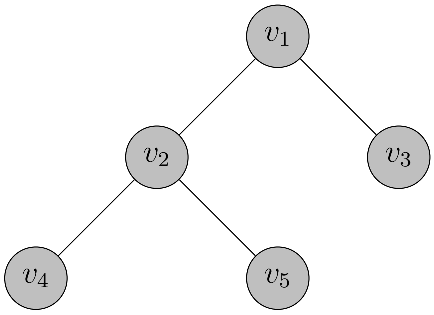




und $s = v_1$. Bei der Tiefensuche ergibt sich die Besuchsreihenfolge v_1, v_2, v_4, v_5, v_3 und somit der Baum



Die Breitensuche liefert die Besuchsreihenfolge v_1, v_2, v_3, v_4, v_5, das heißt den Baum



$\triangle$

Satz 5.24 Seien $G = (V, E)$ ein ungerichteter Graph, $s, v \in V$ und

$$\mathrm{dist}_G(s, v) := \min\{|\pi| : \pi = s, u_1, \ldots, u_r, v \text{ Weg in } G\}.$$

Dann enthält der BFS-Graph $G' = (R, T)$ zum Startknoten $s \in V$ einen kürzesten Weg zu jedem $v \in \mathrm{post}^\star(s)$. Dies bedeutet, der einfache Weg $\pi = s, u_1, \ldots, u_r, v$ erfüllt $|\pi| = \mathrm{dist}_G(s, v)$.

Beweis. Zuerst bemerken wir, dass π eindeutig bestimmt ist, da $G' = (R, T)$ azyklisch ist. Wir modifizieren nun Algorithmus 5.19 wie folgt: In ① setzen wir $\ell(s) := 0$ und in ④ $\ell(w) := \ell(v) + 1$. Dann gilt offenbar zu jedem Zeitpunkt

$$\ell(v) = \mathrm{dist}_{(R,T)}(s, v) \quad \text{für alle } v \in R.$$

Außerdem, falls in ② ein $v \in Q$ ausgewählt wird, so gibt es kein $w \in R$ mit

$$\ell(w) > \ell(v) + 1. \tag{5.2}$$

Angenommen, der Algorithmus bricht ab und es existiert ein Knoten $w \in V$ mit

$$\mathrm{dist}_G(s, w) < \mathrm{dist}_{G'}(s, w).$$

Falls es mehr als einen solchen Knoten gibt, so wählen wir denjenigen mit dem kleinsten Abstand $\mathrm{dist}_G(s, w)$. Sei $\pi = s, u_1, \ldots, u_r, v, w$ ein kürzester Weg in G. Dann gilt $\mathrm{dist}_G(s, v) = \mathrm{dist}_{G'}(s, v)$ und $(v, w) \in E$. Weiter ist

$$\begin{aligned}\ell(w) = \mathrm{dist}_{G'}(s, w) > \mathrm{dist}_G(s, w) = \mathrm{dist}_G(s, v) + 1 = \mathrm{dist}_{G'}(s, v) + 1 \\ = \ell(v) + 1.\end{aligned}$$

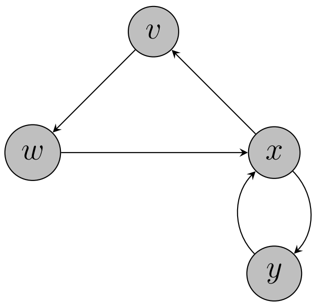
Gemäß (5.2) gilt $w \notin R$ zu dem Zeitpunkt, zu dem $v \in Q$ entfernt wird. Dies widerspricht jedoch ③, denn $(v, w) \in E$. □

Bemerkung: Gemäß Satz 5.21 berechnet obiger Algorithmus die Werte $\mathrm{dist}_{G'}(s, v) = \mathrm{dist}_G(s, v)$ für alle $v \in R$ in Komplexität $\mathcal{O}(|V| + |E|)$.

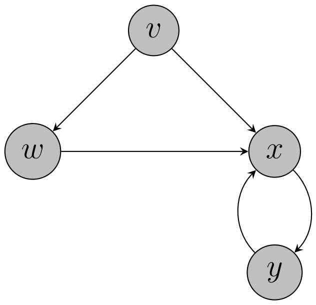
5.5 Starker Zusammenhang

Definition 5.25 Ein gerichteter Graph $G = (V, E)$ heißt **stark zusammenhängend**, falls für jedes Paar von Knoten $v, w \in V$ mit $v \neq w$ gilt $v \in \mathrm{post}^\star(w)$ und $w \in \mathrm{post}^\star(v)$, das heißt, es gibt einen Weg von v nach w und einen Weg von w nach v. Die **starken Zusammenhangskomponenten** sind die maximalen stark zusammenhängenden Teilgraphen.

Beispiel 5.26 Der Graph



ist stark zusammenhängend. Hingegen ist der Graph



nicht stark zusammenhängend, besitzt jedoch die starke Zusammenhangskomponente $\{x, y\}$. △

Nachfolgender, auf der Tiefensuche basierender Algorithmus bestimmt die starken Zusammenhangskomponenten eines gerichteten Graphen.

Algorithmus 5.27 (Bestimmung starker Zusammenhangskomponenten)

input: gerichteter Graph $G = (V, E)$

output: eine Funktion comp : $V \to \mathbb{N}$, die die Zugehörigkeit zu einer starken Zusammenhangskomponente kennzeichnet

① setze $R := \emptyset$, $N := 0$

② für alle $v \in V$:
falls $v \notin R$, dann visit1(v)

③ setze $R := \emptyset$, $K := 0$

④ für alle $i = |V|, |V| - 1, \ldots, 1$:
falls $\Psi^{-1}(i) \notin R$, dann setze $K := K + 1$ und visit2($\Psi^{-1}(i)$)

Hilfsprogramm visit1(v):

① setze $R := R \cup \{v\}$

② für alle $w \in V \setminus R$ mit $(v, w) \in E$: visit1(w)

③ setze $N := N + 1$, $\Psi(v) := N$ und $\Psi^{-1}(N) := v$

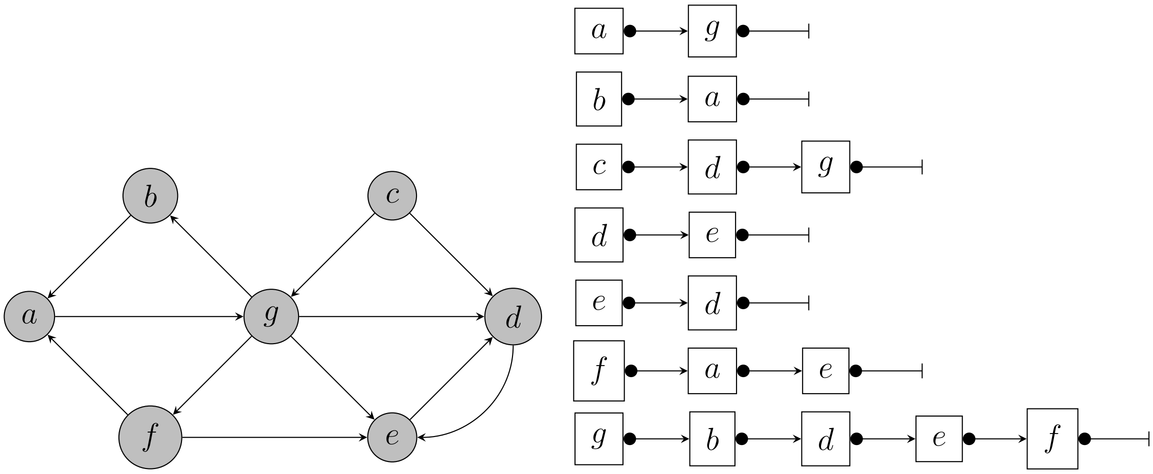
Hilfsprogramm visit2(v):

① setze $R := R \cap \{v\}$

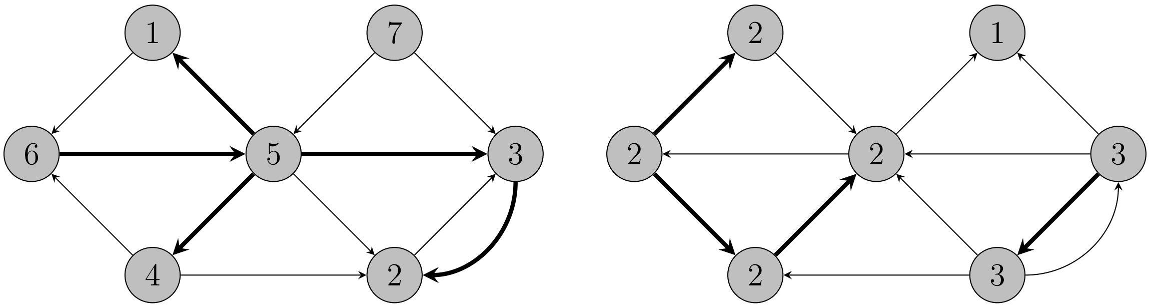
② für alle $w \in V \setminus R$ mit $(w, v) \in E$: visit2(w)

③ setze comp(v) $:= K$

Beispiel 5.28 Gegeben sei der gerichtete Graph



Dann ergibt sich in ② für die erste Tiefensuche `visit1` die Besuchsreihenfolge $\{a, g, b, d, e, f\}$. Der Knoten c ist der einzige Knoten, der nicht von a erreichbar ist. Er bekommt die Nummer 7.



Die Tiefensuche `visit2` ist eine Tiefensuche im inversen Graphen $G^{-1} := (V, E^{-1})$, wobei $E^{-1} := \{(w, v) : (v, w) \in E\}$. In ④ startet die erste Tiefensuche `visit2` mit c (da $\Psi(c) = 7$), kann aber keine weiteren Knoten erreichen. Folglich wird nun die Tiefensuche für a gestartet, da $\Psi(a) = 6$. Es können b, f, g erreicht werden. Schließlich wird e von d aus erreicht. Damit ergeben sich die starken Zusammenhangskomponenten $\{c\}$, $\{a, b, f, g\}$, $\{d, e\}$. △

Satz 5.29 Algorithmus 5.27 identifiziert die starken Zusammenhangskomponenten in linearem Aufwand $\mathcal{O}(|V| + |E|)$.

Beweis. Die Laufzeit ist offensichtlich $\mathcal{O}(|V| + |E|)$.

Seien nun $v, w \in V$ zwei Knoten derselben starken Zusammenhangskomponente, das heißt, in G gibt es einen Weg von v nach w und umgekehrt. Somit gibt es in G^{-1} ebenfalls einen Weg von v nach w und umgekehrt. Die Tiefensuche `visit2` markiert beide Knoten als zur selben Zusammenhangskomponente gehörig, also $\text{comp}(v) = \text{comp}(w)$.

Es verbleibt zu zeigen, dass zwei Knoten $v, w \in V$ mit $\text{comp}(v) = \text{comp}(w)$ auch zur selben starken Zusammenhangskomponente gehören. Dazu sei $r(v)$ bzw. $r(w)$ derjenige von v bzw. w erreichbare Knoten mit dem höchsten Ψ-Wert. Wegen $\text{comp}(v) = \text{comp}(w)$ liegen beide Knoten im selben durch `visit2` erzeugten DFS-Baum. Dessen Startknoten r erfüllt $r = r(v) = r(w)$. Da r von v erreichbar ist und r einen höheren Ψ-Wert erhalten hat, muss r vor v zu R hinzugefügt worden sein bei der Tiefensuche `visit1`. Daher erhält der entsprechende von `visit1` erzeugte DFS-Baum einen r-v-Weg, das heißt, v ist auch von

r erreichbar. Analog ist auch w von r erreichbar. Zusammengefasst ist v von w erreichbar und umgekehrt, was zu zeigen war. □

Definition 5.30 Es sei $G = (V, E)$ ein gerichteter Graph. Eine Numerierung

$$V = \{v_1, v_2, \dots, v_n\}$$

der Knoten heißt **topologische Ordnung**, falls für alle Kanten $(v_i, v_j) \in E$ gilt $i < j$.

Bemerkung: Der gerichtete Graph $G = (V, E)$ besitzt eine topologische Ordnung genau dann, wenn er azyklisch ist.

Satz 5.31 Algorithmus 5.27 bestimmt eine topologische Ordnung des gerichteten Graphen $G = (V, E)$ in linearem Aufwand $\mathcal{O}(|V| + |E|)$, falls diese existiert. Gibt es eine solche Ordnung nicht, so erfährt man dies ebenfalls in linearem Aufwand.

Beweis. Seien $X, Y \subseteq V$ zwei starke Zusammenhangskomponenten mit $\operatorname{comp}(x) = k$, $\operatorname{comp}(y) = \ell$ für alle $x \in X$, $y \in Y$ und $k < \ell$. Wir zeigen, dass keine Kanten $e = (y, x) \in E$ existieren mit $x \in X$, $y \in Y$.

Angenommen eine solche Kante existiert. Alle Knoten aus X werden in der Tiefensuche `visit2` vor den Knoten aus Y zu R hinzugefügt. Insbesondere gilt $x \in R$ und $y \notin R$ beim überprüfen der Kante $e = (y, x)$. Dies bedeutet jedoch, dass y zu R hinzugefügt wird, bevor K erhöht wird, was $\operatorname{comp}(x) \neq \operatorname{comp}(y)$ widerspricht.

Das Hintereinanderreihen der starken Zusammenhangskomponenten liefert folglich eine topologische Vorordnung. Da eine topologische Ordnung nur dann existiert, wenn der Graph azyklisch ist, ergibt sich eine topologische Ordnung genau dann, wenn alle starken Zusammenhangskomponenten einknotig sind. □

6. Algorithmen auf Graphen

6.1 Kürzeste-Wege-Probleme

Definition 6.1 Sei $G = (V, E)$ ein Graph. Eine **Gewichtsfunktion**, manchmal auch **Kostenfunktion** genannt, für die Kanten von G ist eine Abbildung $w : E \to \mathbb{R}$. Ist $\pi = v_0, v_1, \ldots, v_r$ ein Weg in G, dann wird der Wert

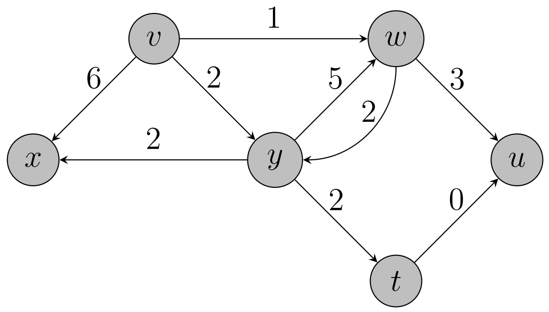
$$w(\pi) = \sum_{i=0}^{r-1} w(v_i, v_{i+1})$$

die **Weglänge** von π bezüglich w genannt. Das Tripel $G = (V, E, w)$ heißt **gewichteter Graph**.

Definition 6.2 Sei $G = (V, E, w)$ ein gewichteter Graph und $v, w \in V$. Ein **kürzester Weg** von v nach w in G bezüglich w ist ein s-w-Weg π mit $w(\pi) \leq w(\pi')$ für jeden s-w-Weg π'. Die **kürzeste Weglänge** $\delta(v, w)$ von v nach w ist definiert durch

$$\delta(v,w) := \begin{cases} \min\{w(\pi) : \pi \text{ ist Weg von } v \text{ nach } w\}, & \text{falls ein solcher Weg existiert,} \\ \infty, & \text{sonst.} \end{cases}$$

Beispiel 6.3 Gegeben sei der gewichtete Graph $G = (V, E, w)$:



Ein kürzester Weg von v nach x ist $\pi = v, y, x$. Seine Weglänge ist $w(\pi) = 2 + 2 = 4$ und ist kürzer als $w(x, y) = 6$. Zum Knoten u gibt es von v zwei kürzeste Wege, nämlich $\pi_1 = v, w, u$ und $\pi_2 = v, y, t, u$ mit $w(\pi_1) = w(\pi_2) = \delta(v, u) = 4$. △

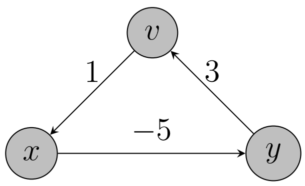
Man unterscheidet verschiedene Varianten des *kürzeste-Wege-Problems*. Gegeben sei stets ein gewichteter, gerichteter Graph $G = (V, E, w)$.

1. **Einzelpaar-kürzeste-Wege-Problem (*single pair shortest path problem*):** gesucht ist für $v, w \in V$ ein kürzester Weg von v nach w.
2. **Einzelquelle-kürzeste-Wege-Problem (*single source shortest path problem*):** für $v \in V$ berechne einen kürzesten Weg zu allen $w \in \mathrm{post}^\star(v)$.
3. **Alle-Paare-kürzeste-Wege-Problem (*all pair shortest path problem*):** finde für jedes Paar $v, w \in V$ einen kürzesten Weg von v nach w.

Natürlich kann das erste Problem durch das zweite gelöst werden. Es ist auch kein asymptotisch besseres Verfahren bekannt. Aus diesem Grund werden wir gleich dieses allgemeine Problem betrachten.

Negative Gewichte sind gemäß Definition 6.1 zugelassen. Problematisch sind jedoch Zyklen mit negativer Weglänge wie folgendes Beispiel zeigt:

Beispiel 6.4 Gegeben sei folgender gewichtete, gerichtete Graph $G = (V, E, w)$:



In diesem Graphen gibt es keine kürzesten Wege, denn

$$\begin{aligned} w(v, x, y) &= -4, \\ w(v, x, y, v, x, y) &= -5, \\ w(v, x, y, v, x, y, v, x, y) &= -6, \\ &\vdots \end{aligned}$$

△

Lemma 6.5 Sei $G = (V, E, w)$ ein gerichteter, gewichteter Graph. Falls es in G keine Zyklen mit negativer Weglänge gibt, dann gibt es für je zwei Knoten $v, w \in V$ mit $w \in \mathrm{post}^\star(v)$ einen kürzesten Weg π mit

$$\delta(v, w) = w(\pi) > -\infty.$$

Beweis. Da es keine negativen Zyklen gibt, genügt es alle einfachen Wege von v nach w zu betrachten. Weil $|V|$ und $|E|$ endlich sind, sind dies nur endlich viele, woraus die Behauptung folgt. □

Lemma 6.6 Sei $G = (V, E, w)$ ein gerichteter, gewichteter Graph ohne negative Zyklen. Ist $\pi = v_0, v_1, \ldots, v_{r-1}, v_r$ ein kürzester Weg von v_0 nach v_r, dann ist für alle $0 \leq i < j \leq r$ der Teilweg $\pi_{i,j} = v_i, v_{i+1}, \ldots, v_j$ von π ein kürzester Weg von v_i nach v_j.

Beweis. Angenommen, es existiert ein kürzester Weg $\pi_{i,j}$ von v_i nach v_j mit $w(\pi'_{i,j}) <$

$w(\pi_{i,j})$. Dann erfüllt der zusammengesetzte Weg $\widehat{\pi} = \pi_{1,i}, \pi'_{i,j}, \pi_{j,r}$ die Abschätzung

$$\begin{aligned} w(\widehat{\pi}) &= \pi_{1,i} + \pi'_{i,j} + \pi_{j,r} \\ &< \pi_{1,i} + \pi_{i,j} + \pi_{j,r} \\ &= w(\pi). \end{aligned}$$

Dies ist ein Widerspruch zur Voraussetzung, dass π ein kürzester Weg von v_0 nach v_r ist. □

Korollar 6.7 Seien $G = (V, E, w)$ ein gerichteter, gewichteter Graph ohne negative Zyklen und $\pi = v_0, v_1, \ldots, v_r$ ein kürzester Weg von v_0 nach v_r. Dann gilt

$$\delta(v_0, v_r) = \delta(v_0, v_{r-1}) + w(v_{r-1}, v_r).$$

Beweis. Gemäß Lemma 6.6 ist $\pi' = v_0, v_1, \ldots, v_{r-1}$ ein kürzester Weg von v_0 nach v_{r-1}, das heißt $w(\pi') = \delta(v_0, v_{r-1})$. Dies bedeutet

$$\delta(v_0, v_r) = w(\pi) = w(\pi') + w(v_{r-1}, v_r).$$

□

Basierend auf diesem Korollar löst der nachfolgende Algorithmus das Einzelquelle-kürzeste-Wege-Problem im Fall *nicht-negativer* Gewichte:

Algorithmus 6.8 (Dijkstra)

input: Ein gerichteter, gewichteter Graph $G = (V, E, w)$ mit nicht-negativen Gewichten und ein Startknoten $s \in V$.

output: Kürzeste Wege von s zu allen $v \in V$ samt Weglänge $\ell(v)$. Genauer ist $\ell(v)$ die Länge eines kürzesten s-v-Wegs, der aus einem kürzesten s-$p(v)$-Weg und der Kante $\big(p(v), v\big)$ besteht. Für $v \notin \text{post}^\star(s)$ ist $\ell(v) = \infty$ und $p(v)$ undefiniert.

① setze $\ell(s) := 0$ und $\ell(v) := \infty$ für alle $v \in V \setminus \{s\}$, setze $R := \emptyset$

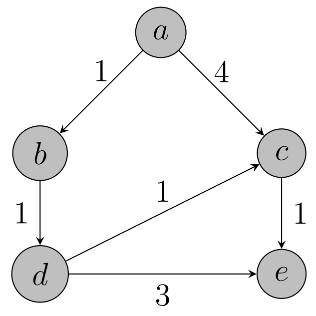
② finde $u \in V \setminus R$ mit $\ell(u) = \min_{v \in V \setminus R} \ell(v)$

③ setze $R := R \cup \{u\}$

④ für alle $v \in V \setminus R$ mit $(u, v) \in E$:
  falls $\ell(v) > \ell(u) + w(u, v)$, dann
    setze $\ell(v) := \ell(u) + w(u, v)$ und $p(v) := u$

⑤ falls $R \neq V$ gehe nach ②

Beispiel 6.9 Gegeben sei folgender gewichtete Graph $G = (V, E, w)$:



Ausgehend vom Startknoten ist die Arbeitsweise dieses Algorithmus wie folgt:

Iteration	a	b	c	d	e	u	$\ell(u)$	$p(u)$
0	0	∞	∞	∞	∞	a	0	—
1	—	1	4	∞	∞	b	1	a
2	—	—	4	2	∞	d	2	b
3	—	—	3	—	5	c	3	d
4	—	—	—	—	4	e	4	c

△

Satz 6.10 (Dijkstra) Der Algorithmus von Dijkstra arbeitet korrekt, wobei seine Laufzeit $\mathcal{O}(n^2)$ mit $n = |V|$ ist.

Beweis. Der Übersichtlichkeit halber schreiben wir den Iterationsindex als Suffix an alle Variablen des Dijkstra-Algorithmus. Wir beweisen, dass folgende Aussagen bei jeder Ausführung von ② gelten:

(a) Für alle $v \in R^{(k)}$ und alle $y \in V \setminus R^{(k)}$ gilt $\ell^{(k)}(v) \leq \ell^{(k)}(y)$.

(b) Für alle $v \in R^{(k)}$ ist $\ell^{(k)}(v)$ die kürzeste Weglänge von s nach v. Ist $\ell^{(k)}(v) < \infty$ und $v \neq s$, dann existiert ein kürzester Weg von s nach v mit Knoten aus $R^{(k)}$ und letzter Kante $\big(p^{(k)}(v), v\big)$.

(c) Für alle $v \in V \setminus R^{(k)}$ ist $\ell^{(k)}(v)$ die kürzeste Weglänge von s nach v im aus den Knoten $R^{(k)} \cup \{v\}$ bestehenden Teilgraphen von G. Ist $\ell^{(k)}(v) \neq \infty$, dann ist $p^{(k)}(v) \in R^{(k)}$ und

$$\ell^{(k)}(v) = \ell^{(k)}\big(p^{(k)}(v)\big) + w\big(p^{(k)}(v), v\big).$$

Da diese Aussagen für $k = 0$ gelten, das heißt, nach Ausführung von ①, zeigen wir, dass ③ und ④ die Aussagen erhalten, das heißt, wir zeigen den Induktionsschritt $k \mapsto k+1$. Dazu sei $k \geq 0$ beliebig und u der in ② ausgewählte Knoten.

Für beliebige $v \in R^{(k)}$ und $y \in V \setminus R^{(k)}$ gilt wegen (a)

$$\ell^{(k)}(v) \leq \ell^{(k)}(u) = \min_{x \in V \setminus R^{(k)}} \ell^{(k)}(x) \leq \ell^{(k+1)}(y).$$

Folglich gilt (a) auch nach ③ und ④, da $R^{(k+1)} = R^{(k)} \cup \{u\}$.

Um zu zeigen, dass (b) nach ④ gilt, müssen wir nur den Knoten u betrachten. Da (c) für k gilt, genügt es zu zeigen, dass in G kein Weg π von s nach u existiert mit einem Knoten $y \in V \setminus R^{(k+1)}$ und $w(\pi) < \ell^{(k)}(u) = \ell^{(k+1)}(u)$.

Angenommen, es gibt einen solchen Weg π, etwa

$$\pi = \underbrace{s, v_1, \ldots, v_r,}_{\in R^{(k)}} \underbrace{y}_{\notin R^{(k)}}, v_{r+1}, \ldots, v_{r+m}, u.$$

Da (c) für k gilt, ist $\ell^{(k)}(y) = w(s, v_1, \ldots, v_r, y)$, und es folgt wegen der Nicht-Negativität der Gewichte

$$w(s, v_1, \ldots, v_r, y) \leq w(\pi) < \ell^{(k)}(u).$$

Dies bedeutet $\ell^{(k)}(y) < \ell^{(k)}(u)$, was $\ell^{(k)}(u) = \min_{x \in V \setminus R^{(k)}} \ell^{(k)}(x)$ widerspricht.

Nun zeigen wir, dass ③ und ④ auch Aussage (c) erhalten. Falls für ein $v \in V \setminus R^{(k+1)}$ in ④

$$p^{(k+1)}(v) := u, \qquad \ell^{(k+1)}(v) := \ell^{(k+1)}(u) + w(u, v)$$

gesetzt wird, muss ein Weg von s nach v existieren im von den Knoten $R^{(k+1)} \cup \{v\}$ aufgespannten Teilgraphen $G'(v)$ von G mit Länge $\ell^{(k+1)}(u) + w(u, v)$ und letzter Kante (u, v).

Angenommen, es existiert ein $v \in V \setminus R^{(k+1)}$ und ein Weg π von s nach v in $G'(v)$ mit $w(\pi) < \ell^{(k+1)}(v)$. Der Knoten u muss in π enthalten sein, da nur u zu $R^{(k+1)}$ hinzugefügt worden ist und sich sonst ein Widerspruch zu (c) vor Ausführung von ③ und ④ ergäbe (denn $\ell^{(k+1)}(v)$ ist höchstens kleiner als $\ell^{(k)}(v)$ geworden).

Sei x der Vorgänger von v in π. Wegen $x \in R^{(k+1)}$ folgt aus (a)

$$\ell^{(k+1)}(x) \leq \ell^{(k+1)}(u)$$

und aus ④

$$\ell^{(k+1)}(v) \leq \ell^{(k+1)}(x) + w(x, v) \leq \ell^{(k+1)}(u) + w(x, v) \leq w(\pi).$$

Hierin gilt die letzte Ungleichung, da π den Knoten u enthält und die Kante (x, v). Die Ungleichung

$$\ell^{(k+1)}(v) \leq w(\pi)$$

widerspricht jedoch unserer Annahme.

Folglich gelten zu jedem Zeitpunkt k die Aussagen (a)–(c), insbesondere gilt (b) bei Abbruch des Algorithmus, das heißt, der Algorithmus arbeitet korrekt.

Die Aufwandsabschätzung ist offensichtlich: Es werden $n = |V|$ Iterationen ausgeführt, die jeweils einen Aufwand $\mathcal{O}(n)$ haben. □

Im Falle von Graphen mit *negativen* Gewichten, aber ohne negativen Zyklen, muss folgender teurerer Algorithmus verwendet werden. Er ist der schnellste bisher bekannte Algorithmus für dieses Problem.

Algorithmus 6.11 (Moore-Bellman-Ford)

input: Ein gerichteter, gewichteter Graph $G = (V, E, w)$ ohne negative Zyklen und ein Startknoten $s \in V$.

output: Kürzeste Wege von s zu allen $v \in V$ samt Weglänge $\ell(v)$. Genauer ist $\ell(v)$ die Länge eines kürzesten s-v-Wegs, der aus einem kürzesten s-$p(v)$-Weg und der Kante $\big(p(v), v\big)$ besteht. Für $v \notin \text{post}^\star(s)$ ist $\ell(v) = \infty$ und $p(v)$ undefiniert.

① setze $\ell(s) := 0$ und $\ell(v) := \infty$ für alle $v \in V \setminus \{s\}$

② für alle $k = 1, 2, \ldots, n-1$:
für jede Kante $(u, v) \in E$:
falls $\ell(v) > \ell(u) + w(u, v)$, dann
setze $\ell(v) := \ell(u) + w(u, v)$ und $p(v) := u$

Satz 6.12 (Moore-Bellman-Ford) Der Moore-Bellman-Ford-Algorithmus arbeitet korrekt, wobei seine Laufzeit $\mathcal{O}(m \cdot n)$ ist mit $m = |E|$ und $n = |V| \leq m$.

Beweis. Die Aufwandsabschätzung ist offensichtlich.
Zu jedem Zeitpunkt des Algorithmus bezeichne

$$R := \{v \in V : \ell(v) < \infty\},$$
$$F := \{(u, v) \in E : u = p(v)\},$$

dann zeigen wir:

(a) $\ell(v) \geq \ell(u) + w(u, v)$ für alle $(u, v) \in F$,

(b) der Graph (R, F) ist azyklisch,

(c) der Graph (R, F) ist ein gerichteter Baum mit Startknoten s, das heißt, jeder Knoten $v \in R$ ist von s aus über genau einen Weg erreichbar.

Wenn in ② $p(v) := u$ gesetzt wird, dann gilt gerade

$$\ell(v) = \ell(u) + w(u, v).$$

Da $\ell(u)$ danach höchstens verkleinert wird, folgt Aussage (a).
Um (b) zu zeigen, nehmen wir an, dass zu einem Zeitpunkt ein Zyklus

$$\pi = v_0, v_1, \ldots, v_{r-1}, v_r, \quad v_0 = v_r$$

entsteht durch Setzen von $p(v_r) := v_{r-1}$. Dann galt aber zuvor

$$\ell(v_0) = \ell(v_r) > \ell(v_{r-1}) + w(v_{r-1}, v_r)$$

und gemäß (a)

$$\ell(v_i) \geq \ell(v_{i-1}) + w(v_{i-1}, v_i), \quad i = 1, 2, \ldots, r-2.$$

Aufsummieren ergibt

$$\sum_{i=1}^{r} w(v_{i-1}, v_i) = \left(\sum_{i=1}^{r-1} \underbrace{w(v_{i-1}, v_i)}_{\leq \ell(v_i) - \ell(v_{i-1})} \right) + \underbrace{w(v_{r-1}, v_r)}_{< v_r - v_{r-1}} < \sum_{i=1}^{r} \big(\ell(v_i) - \ell(v_{i-1}) \big) = 0,$$

das heißt, der Zyklus ist negativ, was der Voraussetzung widerspricht.
Schließlich folgt aus $x \in R \setminus \{s\}$ auch $p(x) \in R$, dies ist die Aussage (c).
Gemäß (a)–(c) ist also zu jedem Zeitpunkt $\ell(y)$ mindestens die Länge des (eindeutigen) s-y-Wegs im Graphen (R, F). Wir zeigen nun, dass nach k Iterationen $\ell(y)$ auch höchstens die Länge eines kürzesten s-y-Wegs in G mit höchstens k Kanten ist. Da für $k = 1$ die Aussage klar ist, nehmen wir an, sie gilt auch nach Iteration $k - 1$. Nun sei

$$\pi = s, v_1, \ldots, v_r, x, y, \quad r \leq k - 2$$

ein kürzester s-x-Weg in G mit höchstens k Kanten. Dann ist $\pi' = s, v_1, \ldots, v_r, x$ ein kürzester s-x-Weg mit höchstens $k-1$ Kanten. Nach Induktionsannahme folgt $\ell(x) \leq w(\pi')$ und somit

$$\ell(y) \leq \ell(x) + w(x,y) \leq w(\pi') + w(x,y) \leq w(\pi).$$

Da kein Weg mehr als $n-1$ Kanten besitzt, impliziert dies die Korrektheit des Algorithmus. □

Wir betrachten nun einen Algorithmus zum Lösen des Alle-Paare-kürzeste-Wege-Problems. Ohne Einschränkung der Allgemeinheit seien die Knoten mit $1, 2, \ldots, n$ beschriftet.

Algorithmus 6.13 (Floyd-Warshall)

input: Ein gerichteter, gewichteter Graph $G = (V, E, w)$ mit $V = \{1, 2, \ldots, n\}$ ohne negative Zyklen.

output: Matrizen $[\ell_{i,j}]_{1 \leq i,j \leq n}$ und $[p_{i,j}]_{1 \leq i,j \leq n}$ mit der kürzesten Weglänge $\ell_{i,j}$ von i nach j und der letzten Kante $(p_{i,j}, j)$ eines i-j-Wegs, falls ein solcher existiert.

① setze $\ell_{i,j} := w(i,j)$ für alle $(i,j) \in E$
setze $\ell_{i,j} := \infty$ für alle $(i,j) \in (V \times V) \setminus E$ mit $i \neq j$
setze $\ell_{i,i} := 0$ für alle i
setze $p_{i,j} := i$ für alle $i, j \in V$

② für alle $j = 1, 2, \ldots, n$:
    für alle $i = 1, 2, \ldots, n$ mit $i \neq j$:
        für alle $k = 1, 2, \ldots, n$ mit $k \neq j$:
        falls $\ell_{i,k} > \ell_{i,j} + \ell_{j,k}$, dann
            setze $\ell_{i,k} := \ell_{i,j} + \ell_{j,k}$ und $p_{i,k} := p_{j,k}$

Satz 6.14 (Floyd-Warshall) Der Algorithmus von Floyd und Warshall arbeitet korrekt, wobei seine Laufzeit $\mathcal{O}(n^3)$ mit $n = |V|$ ist.

Beweis. Die Aussage zur Laufzeit ist klar. Wir schreiben wieder den Iterationsindex j als Suffix an alle Variablen und zeigen: Nach j_0 äußeren Iterationen ist $\ell_{i,k}^{(j_0)}$ die Länge eines kürzesten i-k-Wegs bestehend nur aus den Zwischenknoten $v \in \{1, \ldots, j_0\}$ und mit Endkante $(p_{i,k}^{(j_0)}, k)$. Diese Aussage beweisen wir mit vollständiger Induktion über j_0.

Für $j_0 = 0$ gilt sie gemäß ① und für $j_0 = n$ impliziert sie die Korrektheit des Algorithmus. Wir nehmen an, dass obige Aussage gilt für ein $j_0 \in \{0, 1, \ldots, n-1\}$, das heißt, für alle $i, k \in V$ enthält $\ell_{i,k}^{(j_0)}$ einen kürzesten i-k-Weg bestehend nur aus Zwischenknoten $v \in \{1, \ldots, j_0\}$.

In der (j_0+1)-ten Iteration wird $\ell_{i,k}^{(j_0)}$ durch $\ell_{i,j_0+1}^{(j_0)} + \ell_{j_0+1,k}^{(j_0)}$ ersetzt, falls dieser Wert kleiner ist. Es verbleibt daher zu zeigen, dass dann die entsprechenden Wege

$$\pi_1 = i, u_1, \ldots, u_r, j_0 + 1,$$
$$\pi_2 = j_0 + 1, v_1, \ldots, v_s, k$$

keine gemeinsamen inneren Knoten haben.

Angenommen, beide Wege haben den gemeinsamen Knoten $u_p = v_q$, dann ist

$$\pi = i, u_1, \dots, u_p, v_{q+1}, \dots, v_s, k$$

ein Weg von i nach k, bestehend nur aus Knoten $v \in \{1, \dots, j_0\}$. Wegen

$$w(u_p, u_{p+1}, \dots, u_r, j_0 + 1, v_1, \dots, v_q) \geq 0$$

ist

$$\ell_{i,k}^{(j_0)} \leq w(\pi) \leq w(\pi_1 \cup \pi_2) \leq \ell_{i,j_0+1}^{(j_0)} + \ell_{j_0+1,k}^{(j_0)},$$

was ein Widerspruch zu $\ell_{i,k}^{(j_0)} > \ell_{i,j_0+1}^{(j_0)} + \ell_{j_0+1,k}^{(j_0)}$ ist. □

6.2 Netzwerkflussprobleme

Motivation: In den Seehäfen $A_1, A_2, \dots, A_p$ liegen $r_1, r_2, \dots, r_p$ Tonnen Bananen zum Verschiffen bereit. In den Zielhäfen $B_1, B_2, \dots, B_q$ besteht die Nachfrage nach $d_1, d_2, \dots, d_q$ Tonnen. Die Kapazität der Schifffahrtslinie von Hafen A_i nach Hafen B_j ist maximal $c(A_i, B_j)$.

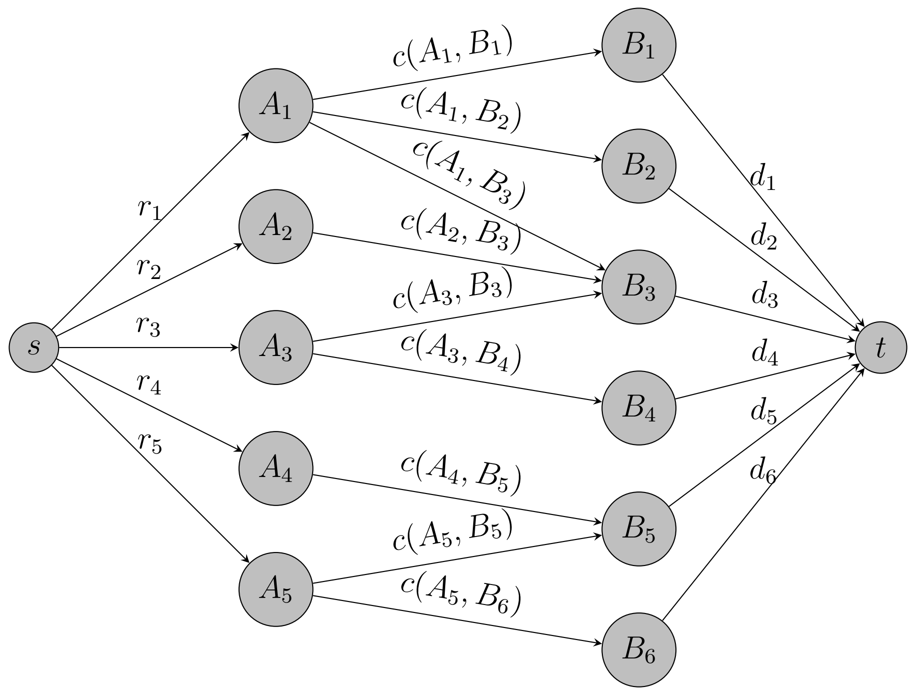
Es stellen sich die folgenden Fragen:

1. Ist es möglich, alle Anforderungen zu befriedigen?
2. Falls nein, wie viel Bananen können maximal zu den Zielhäfen gebracht werden?
3. Wie sollen die Bananen verschifft werden?

Zur Lösung konstruieren wir einen gerichteten Graphen $G = (V, E)$ mit

$$V = \{A_1, A_2, \dots, A_p, B_1, B_2, \dots, B_q\}, \qquad E = \{(A_i, B_j) : 1 \leq i \leq p,\ 1 \leq j \leq r\}.$$

Der Kante (A_i, B_j) ordnen wir die Kapazität $c(A_i, B_j)$ zu. Um Angebots- und Nachfragemengen zu modellieren, führen wir zwei weitere Knoten s, t und Kanten (s, A_i) beziehungsweise (B_j, t) mit Kapazität r_i beziehungsweise d_j ein.



Zur Beantwortung der drei Fragen lösen wir folgendes Problem: Was ist der maximale Fluss von s nach t in G und wie sieht dieser aus? Dabei ist der Fluss auf einer Kante durch ihre Kapazität beschränkt. Des Weiteren muss der gesamte Fluss, der einen Knoten A_i oder B_j betritt, diesen auch wieder verlassen.

Definition 6.15 Ein **Netzwerk** ist ein Tupel $N = (V, E, c, s, t)$ bestehend aus

- einem gerichteten Graphen $G = (V, E)$,
- einer **Kapazitätsfunktion** $c : E \to \mathbb{R}_{\geq 0}$,
- einer **Quelle** $s \in V$ mit $\mathrm{pre}(s) = \emptyset$,
- einer **Senke** $t \in V$ mit $\mathrm{post}(t) = \emptyset$.

Ein **Fluss** $f : E \to \mathbb{R}_{\geq 0}$ ist eine Funktion, die folgende Bedingungen erfüllt:

1. **Kapazitätsbedingung**:
$$f(v, w) \leq c(v, w) \quad \text{für alle } (v, w) \in E,$$
2. **Kirchhoffsches Gesetz**:
$$\sum_{u \in \mathrm{pre}(v)} f(u, v) = \sum_{w \in \mathrm{post}(v)} f(v, w) \quad \text{für alle } v \in V \setminus \{s, t\}.$$

Der **Wert** des Flusses ist
$$\mathrm{flow}(f) = \sum_{w \in \mathrm{post}(s)} f(s, w).$$

Definition 6.16 Der **maximale Fluss** von N ist gegeben durch
$$\mathrm{MaxFlow}(N) = \max\{\mathrm{flow}(f) : f \text{ ist Fluss für } N\}.$$
Eine Flussfunktion f für N wird **optimal** genannt, falls
$$\mathrm{flow}(f) = \mathrm{MaxFlow}(N).$$

Definition 6.17 Ein **Schnitt** für $N = (V, E, c, s, t)$ ist eine Knotenmenge $S \subseteq V$ mit $s \in S$ und $t \notin S$. Die **Kapazität** eines Schnitts ist gegeben durch
$$\mathrm{cap}(S) = \sum_{\substack{v \in S \\ w \in \mathrm{post}(v) \setminus S}} c(v, w).$$
Die **minimale Schnittkapazität** von N ist
$$\mathrm{MinCut}(N) := \min\{\mathrm{cap}(S) : S \text{ ist Schnitt für } N\}.$$

Lemma 6.18 Sei S ein Schnitt für $N = (V, E, c, s, t)$, dann gilt für jeden Fluss f

(i) $\text{flow}(f) = \sum\limits_{\substack{v \in S \\ w \in \text{post}(v) \setminus S}} f(v, w) - \sum\limits_{\substack{v \in S \\ u \in \text{pre}(v) \setminus S}} f(u, v)$,

(ii) $\text{flow}(f) \leq \text{cap}(S)$.

Beweis. Aussage (i) folgt aus dem Kirchhoffschen Gesetz

$$\begin{aligned}
\text{flow}(f) &= \sum_{w \in \text{post}(s)} f(s, w) \\
&= \sum_{v \in S} \Bigg(\sum_{w \in \text{post}(v)} f(v, w) - \sum_{w \in \text{pre}(v)} f(u, v) \Bigg) \\
&= \sum_{\substack{v \in S \\ w \in \text{post}(v) \setminus S}} f(v, w) - \sum_{\substack{v \in S \\ u \in \text{pre}(v) \setminus S}} f(u, v) \\
&+ \underbrace{\sum_{\substack{v \in S \\ w \in \text{post}(v) \cap S}} f(v, w) - \sum_{\substack{v \in S \\ u \in \text{pre}(v) \cap S}} f(u, v)}_{=0}
\end{aligned}$$

Da $0 \leq f(e) \leq c(e)$ für alle $e \in E$, folgt weiter

$$\text{flow}(f) \overset{\text{(i)}}{\leq} \sum_{\substack{v \in S \\ w \in \text{post}(v) \setminus S}} f(v, w) \leq \sum_{\substack{v \in S \\ w \in \text{post}(v) \setminus S}} c(v, w) = \text{cap}(S),$$

dies ist Aussage (ii). ☐

Satz 6.19 (Max-Flow-Min-Cut-Theorem) Sei $N = (V, E, c, s, t)$ ein Netzwerk, dann gilt

$$\text{MinCut}(N) = \text{MaxFlow}(N).$$

Beweis. Aus Lemma 6.18 folgt $\text{MaxFlow}(N) \leq \text{MinCut}(N)$. Daher genügt es zu zeigen, dass ein Schnitt S existiert mit $\text{MaxFlow}(N) = \text{cap}(S)$. Hierzu geben wir eine Prozedur an, die für einen gegebenen Fluss f mit $\text{flow}(f) = \text{MaxFlow}(N)$ einen Schnitt S mit $\text{cap}(S) = \text{MaxFlow}(N)$ konstruiert.

Wir starten mit $S = \{s\}$. In jedem Schritt erweitern wir S um einen Knoten $y \in V \setminus S$, der benötigt wird, damit die Behauptung überhaupt erfüllt sein kann:

① setze $S := \{s\}$

② solange $x \in S$, $y \in V \setminus S$ existiert mit

$$\begin{aligned}
&c(x, y) > f(x, y), && \text{falls } (x, y) \in E, \\
&f(x, y) > 0, && \text{falls } (y, x) \in E,
\end{aligned}$$

setze $S := S \cup \{y\}$.

Wir zeigen zunächst, dass S stets ein Schnitt für N ist, das heißt, es gilt stets $t \notin S$.
Angenommen, es gilt $t \in S$, dann gibt es einen Knoten $v_{r-1} \in S$, der dafür verantwortlich ist, dass $t = v_r$ zu S hinzugenommen wurde, das heißt, $c(v_{r-1}, v_r) > f(v_{r-1}, v_r)$ oder $f(v_r, v_{r-1}) > 0$. Genauso gibt es einen Knoten $v_{r-2} \in S$, der dafür verantwortlich ist, dass v_{r-1} hinzugenommen worden ist, usw. Folglich existiert ein ungerichteter Weg

$$\pi = v_0, v_1, \ldots, v_r, \quad v_i \in S \text{ für alle } 0 \leq i \leq r,$$

wobei $v_0 = s$. Setzen wir für alle $i = 0, 1, \ldots, r-1$

$$\varepsilon_i := \begin{cases} c(e) - f(e), & \text{falls } e = (v_i, v_{i+1}) \in E \text{ und } e^{-1} = (v_{i+1}, v_i) \notin E, \\ f(e^{-1}), & \text{falls } e = (v_i, v_{i+1}) \notin E \text{ und } e^{-1} = (v_{i+1}, v_i) \in E, \\ \max\{c(e) - f(e), f(e^{-1})\}, & \text{falls } e = (v_i, v_{i+1}) \in E \text{ und } e^{-1} = (v_{i+1}, v_i) \in E, \end{cases} \tag{6.1}$$

so folgt nach Konstruktion stets $\varepsilon_i > 0$. Wir setzen

$$\varepsilon := \min_{0 \leq i \leq r} \varepsilon_i > 0. \tag{6.2}$$

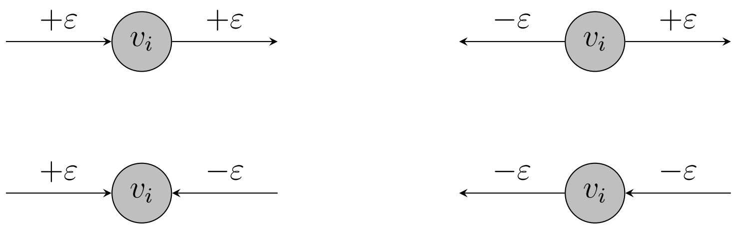
und führen einen Widerspruch herbei, indem wir nun einen Fluss $f^\star$ konstruieren mit

$$\text{flow}(f^\star) = \text{MaxFlow}(N) + \varepsilon.$$

Hierzu definieren wir $f^\star$ für alle $0 \leq i < r$ wie folgt:

$$\begin{aligned} f^\star(e) &:= f(e) + \varepsilon, \quad \text{falls } e = (v_i, v_{i+1}) \in E \text{ und } e^{-1} = (v_{i+1}, v_i) \notin E, \\ f^\star(e^{-1}) &:= f(e^{-1}) - \varepsilon, \text{ falls } e = (v_i, v_{i+1}) \notin E \text{ und } e^{-1} = (v_{i+1}, v_i) \in E. \end{aligned}$$

Gilt $e = (v_i, v_{i+1}) \in E$ und $e^{-1} = (v_{i+1}, v_i) \in E$, so erhöhen wir $f(e)$ um ε falls $c(e) - f(e) > f(e^{-1})$, ansonsten verringern wir $f(e^{-1})$ um ε.
(6.1), (6.2) garantieren, dass $f^\star$ die Kapazitätsbedingung nicht verletzt. Das Kirchhoffsche Gesetz bleibt beim Übergang $f \mapsto f^\star$ erhalten, da es nur folgende vier Möglichkeiten der Flussänderung pro Knoten v_i gibt:



Also ist $f^\star$ ein Fluss. Weiter gilt

$$\begin{aligned} \text{flow}(f^\star) &= \sum_{v \in \text{post}(s)} f^\star(s, v) \\ &= \sum_{u \in \text{pre}(t)} f^\star(u, t) \\ &= \sum_{u \in \text{pre}(t) \setminus \{v_{r-1}\}} f(u, t) + \underbrace{f^\star(v_{r-1}, t)}_{= f(v_{r-1}, t) + \varepsilon} \\ &= \text{flow}(f) + \varepsilon, \end{aligned}$$

was ein Widerspruch zu $\text{flow}(f) = \text{MaxFlow}(N)$ ist. Damit ist S ein Schnitt in N und gemäß Konstruktion gilt für alle $x \in S$, $y \in V \setminus S$ dass $f(x, y) = c(x, y)$ beziehungsweise $f(y, x) = 0$. Damit folgt $\text{flow}(f) = \text{cap}(S)$. $\square$

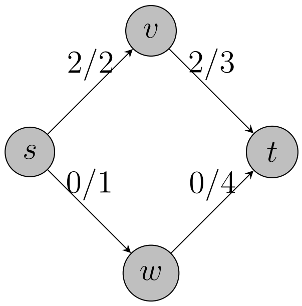
Der im Beweis des Max-Flow-Min-Cut-Theorems konstruierte Weg π heißt *augmentierter Weg*.

Definition 6.20 Sei f ein Fluss im Netzwerk $N = (V, E, c, s, t)$. Eine Kante $e = (x, y) \in E$ heißt **Vorwärtskante**, falls $f(e) < c(e)$. Eine Kante $e = (x, y)$ mit $e^{-1} = (y, x) \in E$ heißt **Rückwärtskante**, falls $f(e^{-1}) > 0$. Der **Restgraph** für f ist der gerichtete Graph $G' = (V, E')$ mit

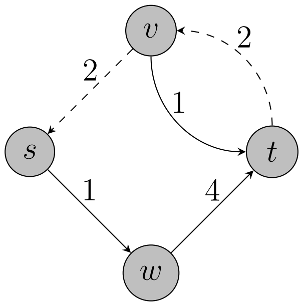
$$E' = \{(x, y) \in V \times V : (x, y) \text{ ist Vorwärts- oder Rückwärtskante}\}.$$

Die Größen $c(e) - f(e)$ beziehungsweise $f(e^{-1})$ heißen **Restkapazitäten**. Ein **augmentierter Weg** $\pi = v_0, v_1, \ldots, v_r$ ist ein Weg im Restgraph mit $v_0 = s$ und $v_r = t$.

Beispiel 6.21 Gegeben sei folgendes Netzwerk mit Fluss/Kapazitäten:



Der Restgraph ist



wobei Vorwärtskanten durch durchgezogene und Rückwärtskanten durch gestrichelte Pfeile markiert sind. △

Im Fall, dass der Fluss f nicht maximal ist, kann mit Hilfe eines augmentierten Weges der Fluss vergrößert werden. Somit erhalten wir folgenden Algorithmus:

Algorithmus 6.22 (Ford-Fulkerson)

input: Netzwerk $N = (V, E, c, s, t)$
output: Fluss f mit $\text{flow}(f) = \text{MaxFlow}(N)$

① Setze $f(e) = 0$ für alle $e \in E$.

② Suche einen augmentierten Weg π von s nach t. Falls keiner existiert, dann `stop`.

③ Berechne ε gemäß (6.1), (6.2). Augmentiere f um ε und gehe nach ②.

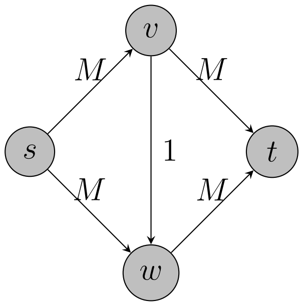
Ford und Fulkerson haben anhand eines Beispiels gezeigt, dass bei irrationalen Kapazitäten der Algorithmus möglicherweise nicht terminiert. Im Fall ganzzahliger Kapazitäten ist dies jedoch nicht der Fall.

Satz 6.23 (Integral-Flow-Theorem) Sei $N = (V, E, c, s, t)$ ein Netzwerk mit ganzzahligen Kapazitäten. Dann terminiert der Ford-Fulkerson-Algorithmus nach maximal $\sum_{(x,y)\in E} c(x,y)$ Augmentierungsschritten mit einem ganzzahligen maximalen Fluss.

Beweis. Da alle Kapazitäten ganzzahlig sind und wir mit dem Nullfluss starten, ist während der Durchführung des Algorithmus $\mathrm{flow}(f)$ stets ganzzahlig. Da ein Augmentierungsschritt die Größe des Flusses mindesten um 1 erhöht, ergibt sich die Behauptung. □

Auch bei ganzzahligen Kapazitäten kann der Ford-Fulkerson-Algorithmus viele Augmentierungsschritte benötigen:

Beispiel 6.24 Wie betrachten das Netzwerk



Offensichtlich gilt $\mathrm{MaxFlow}(N) = 2M$. Starten wir mit dem Nullfluss und augmentieren stets entlang eines s-t-Wegs der Länge 3, erhöht sich $\mathrm{flow}(f)$ jeweils nur um $\varepsilon = 1$. Folglich werden $2M$ Schritte benötigt. △

Das Beispiel zeigt, dass bei willkürlicher Wahl des augmentierenden Wegs (also Wege von s nach t im Restgraphen) die Anzahl der Augmentierungsschritte sehr groß sein kann. Polynomielle Laufzeitbeschränkung im Ford-Fulkerson-Algorithmus kann durch die Wahl eines *kürzesten* augmentierten Weg erreicht werden. Hierbei bezieht dich die Länge auf die Anzahl der Kanten.

Algorithmus 6.25 (Edmonds-Karp)

input: Netzwerk $N = (V, E, c, s, t)$

output: Fluss f mit $\mathrm{flow}(f) = \mathrm{MaxFlow}(N)$

① Setze $f(e) = 0$ für alle $e \in E$.

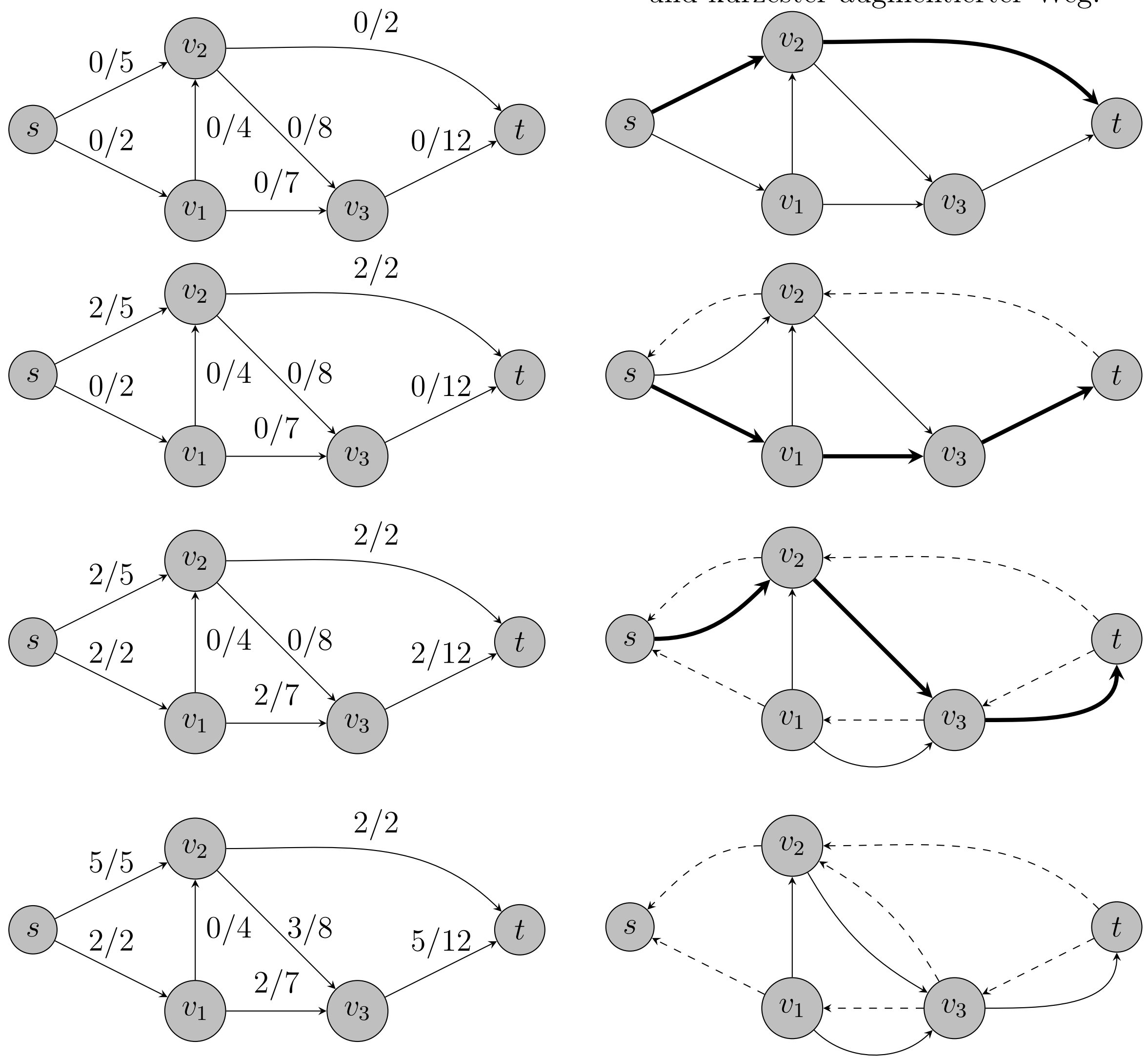
② Suche einen kürzesten augmentierten Weg π von s nach t. Falls keiner existiert, dann `stop`.

③ Berechne ε gemäß (6.1), (6.2). Augmentiere f um ε und gehe nach ②.

Bemerkung: Schritt ② kann durch eine Breitensuche im Restgraphen realisiert werden, vergleiche Satz 5.24.

Beispiel 6.26 Wir illustrieren den Edmonds-Karp-Algorithmus anhand des folgende konkreten Beispiels.

Netzwerk mit aktuellem Fluss: | Restgraph und kürzester augmentierter Weg:



△

Wir benötigen folgendes Lemma:

Lemma 6.27 Sei $(f_0, \pi_0), (f_1, \pi_1), (f_2, \pi_2), \ldots$ die von Algorithmus 6.25 erzeugte Folge von Flussfunktionen f_i und zugehörigen kürzesten augmentierten Wegen π_i im Restgraphen von f_i. Dann gelten die folgenden Aussagen:

1. Für alle i gilt $|\pi_i| \leq |\pi_{i+1}|$.
2. Kommt $e = (v, w)$ in π_i und $e^{-1} = (w, v)$ in π_j vor mit $i < j$, so gilt
$$|\pi_i| + 2 \leq |\pi_j|.$$

Beweis. Es sei $\ell_i(x, y)$ die Länge eines kürzesten Wegs von x nach y im Restgraphen von f_i. Insbesondere gilt also $|\pi_i| = \ell_i(s, t)$. Wir zeigen zuerst, dass für alle $v \in V$ gilt

$$\ell_{i+1}(s, v) \geq \ell_i(s, v). \tag{6.3}$$

Falls $\ell_{i+1}(s,v) = \infty$, dann ist die Ungleichung trivialerweise erfüllt. Wir können also annehmen, dass v von s im Restgraph f_{i+1} erreichbar ist, das heißt $\ell_{i+1}(s,v) = r < \infty$. Sei $\pi = s, v_1, v_2, \ldots, v_r$ ein kürzester Weg von s nach $v = v_r$ im Restgraph von f_{i+1}. Wir zeigen, dass dann gilt

$$\ell_i(s, v_{j+1}) \leq \ell_i(s, v_j) + 1, \quad 1 \leq j < r. \tag{6.4}$$

Falls (v_j, v_{j+1}) eine Kante im Restgraph von f_i ist, gilt (6.4) trivialerweise. Ist (v_j, v_{j+1}) keine Kante im Restgraph von f_i, dann muss sich der Flusswert der inversen Kante (v_{j+1}, v_j) im Augmentierungsschritt $f_i \mapsto f_{i+1}$ verändert haben. Andernfalls könnte (v_j, v_{j+1}) im Restgraph von f_{i+1} nicht vertreten sein. Folglich liegt die Kante (v_{j+1}, v_j) auf dem Weg π_i. Da π_i ein kürzester Weg von s nach t ist und v_{j+1} unmittelbar vor v_j in π_i vorkommt, ergibt sich

$$\ell_i(s, v_{j+1}) = \ell_i(s, v_j) - 1,$$

das heißt, es gilt ebenfalls (6.4).

Aus (6.4) folgt dann (6.3), denn

$$\begin{aligned}
\ell_i(s,v) &= \ell_i(s, v_r) \\
&\overset{(6.4)}{\leq} \ell_i(s, v_{r-1}) + 1 \\
&\overset{(6.4)}{\leq} \ell_i(s, v_{r-2}) + 2 \\
&\;\vdots \\
&\overset{(6.4)}{\leq} \ell_i(s, v_1) + r - 1 \\
&\overset{(6.4)}{\leq} \ell_i(s, s) + r \\
&= \ell_{i+1}(s, v).
\end{aligned}$$

Insbesondere liefert die Wahl $v = t$ Aussage 1.

Analog zu (6.3) zeigt man, dass auch

$$\ell_{i+1}(v,t) \geq \ell_i(v,t) \tag{6.5}$$

gilt für alle $v \in V$.

Sei nun $e = (v, w)$ bzw. $e^{-1} = (w, v)$ eine Kante im Weg π_i bzw. π_j mit $i < j$, das heißt

$$\pi_i = s, \ldots, v, w, \ldots, t, \qquad \pi_j = s, \ldots, w, v, \ldots, t.$$

Da beide jeweils kürzeste Wege im entsprechenden Restgraph sind, gilt

(i) $|\pi_i| = \ell_i(s,v) + \ell_i(v,t)$,

(ii) $|\pi_j| = \ell_j(s,w) + 1 + \ell_j(v,t)$,

(iii) $\ell_i(s,w) = \ell_i(s,v) + 1$,

während (6.3) und (6.5) wegen $i < j$ implizieren

(iv) $\ell_j(s,w) \geq \ell_i(s,w)$, $\quad \ell_j(v,t) \geq \ell_i(v,t)$.

Kombination der Beziehungen (i)–(iv) liefert Aussage 2:

$$\begin{aligned}
|\pi_j| &\overset{\text{(ii)}}{=} \ell_j(s,w) + 1 + \ell_j(v,t) \\
&\overset{\text{(iv)}}{\geq} \ell_i(s,w) + 1 + \ell_i(v,t) \\
&\overset{\text{(iii)}}{=} \ell_i(s,v) + 2 + \ell_i(v,t) \\
&\overset{\text{(i)}}{=} |\pi_i| + 2.
\end{aligned}$$

□

Satz 6.28 (Edmonds-Karp) Unabhängig von den Kapazitäten terminiert Algorithmus 6.25 nach höchstens $(n \cdot m)/2$ Augmentierungsschritten, wobei $n = |V|$ und $m = |E|$ ist.

Beweis. Sei $(f_0, \pi_0), (f_1, \pi_1), (f_2, \pi_2), \ldots$ die von Algorithmus 6.25 erzeugte Folge von Flussfunktionen f_i und zugehörigen kürzesten Wegen π_i im Restgraph von f_i. In jedem Augmentierungsschritt wird mindestens eine Kante $e = (v, w)$ des Wegs π_i voll ausgeschöpft, das heißt, dass eine Flussveränderung um die Restkapazität stattfindet:

- Ist e eine Vorwärtskante im Restgraph von f_i, so ist $f_{i+1}(e) = c(e)$.
- Ist e eine Rückwärtskante im Restgraph von f_i, so ist $f_{i+1}(e^{-1}) = 0$.

In keinem der beiden Fälle ist e eine Kante im Restgraph von f_{i+1}. Bevor dieselbe Kante in einem späteren Augmentierungsschritt $f_k \mapsto f_{k+1}$ in π_k vorkommt und wieder voll ausgeschöpft wird, muss die inverse Kante $e^{-1} = (w, v)$ im Weg π_j mit $i < j < k$ vorgekommen sein. Aus Lemma 6.27 folgt

$$|\pi_i| \leq |\pi_j| - 2 \leq |\pi_k| - 4.$$

Wird also e in den Wegen $\pi_{i_0}, \pi_{i_1}, \pi_{i_2}, \ldots, \pi_{i_\ell}$ voll ausgeschöpft, dann existiert eine Indexfolge $j_0, j_1, \ldots, j_{\ell-1}$ derart, dass

- $i_0 < j_0 < i_1 < j_1 < \cdots < i_{\ell-1} < j_{\ell-1} < i_\ell$,
- $e^{-1} = (w, v)$ kommt in $\pi_{j_0}, \pi_{j_1}, \ldots, \pi_{j_{\ell-1}}$ vor,
- $1 \leq |\pi_{i_0}| \leq |\pi_{j_0}| - 2 \leq |\pi_{i_1}| - 4 \leq |\pi_{j_1}| - 6 \leq \cdots \leq |\pi_{i_\ell}| - 4\ell$.

Da kürzeste Wege stets einfach sind, ist π_{i_k} stets ein einfacher Weg im Restgraph von f_{i_k} und es folgt

$$|\pi_{i_\ell}| < n.$$

Hieraus folgt jedoch, dass jede Kante $e \in E \cup E^{-1}$ höchstens $n/4$-mal voll ausgeschöpft werden kann, das heißt, $\ell < n/4$. Da nur $|E \cup E^{-1}| \leq 2|E|$ Kanten vorhanden sind, werden maximal

$$2m \cdot \frac{n}{4} = \frac{m \cdot n}{2}$$

Augmentierungsschritte durchgeführt. □

Bemerkung: Wir haben soeben die Existenz einer Lösung des Netzwerkflussproblems gezeigt!

Korollar 6.29 Gilt $m \leq n$, so ist der Aufwand des Edmonds-Karp-Algorithmus $\mathcal{O}(m^2 n)$, wobei $n = |V|$ und $m = |E|$.

Beweis. Gemäß Satz 6.28 benötigen wir höchstens $(m \cdot n)/2$ Augmentierungsschritte. Da hierzu jeweils eine Breitensuche benötigt wird, die den Aufwand $\mathcal{O}(m)$ besitzt, ergibt sich das Behauptete. □

6.3 Bipartites Matching

Definition 6.30 Sei $G = (V, E)$ ein ungerichteter Graph. Ein **Matching** von G ist eine Kantenmenge $M \subseteq E$, so dass jeder Knoten von G höchstens auf einer Kante von M liegt, das heißt, wenn für alle Kanten $(v, w), (x, y) \in M$ gilt

$$(v, w) \neq (x, y) \Rightarrow \{v, w\} \cap \{x, y\} = \emptyset.$$

Ein Matching M heißt **maximal**, wenn $|M| \geq |M'|$ für alle Matchings M' von G.

Wir wollen uns im folgenden darauf beschränken, ein maximales Matching in einem *bipartiten* Graph zu suchen.

Definition 6.31 Ein ungerichteter Graph $G = (V, E)$ heißt **bipartit** oder **zweigeteilt**, falls nichtleere Knotenmengen V_L und V_R existieren, so dass

(i) $V = V_L \cup V_R$, $V_L \cap V_R = \emptyset$,

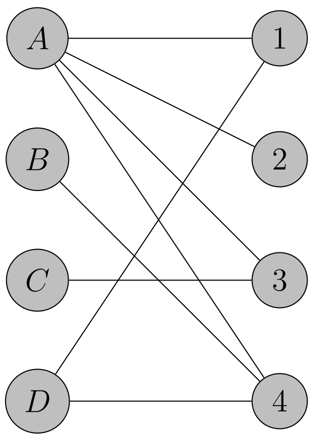
(ii) für jede Kante $(v, w) \in E$ ist

$$\{v, w\} \cap V_L \neq \emptyset, \ \{v, w\} \cap V_R \neq \emptyset.$$

Die Mengen V_L, V_R heißen **(Bi-) Partition**.

Viele Anwendungen führen auf die Bestimmung eines maximalen Matchings in einem bipartiten Graph. Wir betrachten exemplarisch das sogenannte "Heiratsproblem".

Beispiel 6.32 Vier Frauen $V_L = \{A, B, C, D\}$ haben unter vier Männern $V_R = \{1, 2, 3, 4\}$ diejenigen ausgewählt, die sie sich als Ehepartner wünschen, und umgekehrt. Eine Heiratsagentur soll anhand dieser Information potentielle Paare bilden. Gesucht ist folglich eine Paarbildung, bei der nur Wunschpaare zulässig sind und die Zahl der Heiratsvermittlungen maximal ist. Wir erhalten beispielsweise den Graph



wobei die Kanten für Wunschpaare stehen. Ein maximales Matching ist

$$M = \{(A, 2), (B, 4), (C, 3), (D, 1)\}.$$

△

Wir führen das Matchingprobem auf ein äquivalentes Flussproblem zurück.

Definition 6.33 Sei $G = (V, E)$ ein bipartiter Graph mit Partition $V = V_L \cup V_R$. Wir definieren das zugehörige Netzwerk $N_G = (V \cup \{s, t\}, E', c, s, t)$ gemäß:

- $s, t \notin V$ und $s \neq t$,
- $E' = E \cup \{(s, v) : v \in V_L\} \cup \{(w, t) : w \in V_R\}$,
- $c(e) = 1$ für alle $e \in E'$.

Eine **0-1-Flussfunktion** für N_G ist eine Flussfunktion f für N_G mit $f(e) = \{0, 1\}$ für alle $e \in E'$.

Satz 6.34 Sei $G = (V, E)$ ein bipartiter Graph mit Partition $V = V_L \cup V_R$, dann gilt:

(i) Zu jedem Matching M gibt es eine 0-1-Flussfunktion f_M für N_G mit $\text{flow}(f_M) = |M|$.

(ii) Zu jeder 0-1-Flussfunktion f für N_G gibt es ein Matching M_f für G mit $\text{flow}(f) = |M_f|$.

Beweis. (i) Sei f_M definiert gemäß

$$f_M(v, w) := \begin{cases} f_M(s, v) = f_M(w, t) = 1, & \text{falls } v \in V_L, w \in V_R, (v, w) \in M, \\ 0, & \text{sonst.} \end{cases}$$

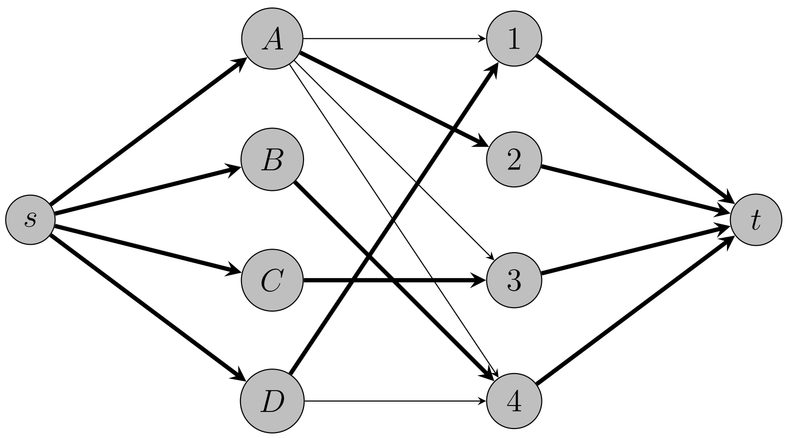
Da aufgrund der Matchingbedingungen jeder Knoten auf höchstens einer Kante von E liegt, erfüllt f_M das Kirchhoffsche Gesetz. Folglich ist f_M eine 0-1-Flussfunktion, insbesondere gilt $\text{flow}(f_M) = |M|$.

(ii) Definiere

$$M_f := \{(v, w) \in V_L \times V_R : f(v, w) = 1\}.$$

Das Kirchhoffsche Gesetz für f entspricht der Matchingeigenschaft von M_f und offensichtlich gilt $\text{flow}(f) = |M_f|$. □

Beispiel 6.35 Für das Heiratsproblem aus Beispiel 6.32 sind N_G und f_M gegeben durch:



Hierin entsprechen dicke Pfeile $f_M(v, w) = 1$, dünne Pfeile $f_M(v, w) = 0$. △

Korollar 6.36 Für G wie zuvor gilt

$$\text{MaxFlow}(N_G) = \max\{|M| : M \text{ ist Matching für } G\}.$$

Eine maximale 0-1-Flussfunktion für N_G kann mit dem Ford-Fulkerson-Algorithmus in Laufzeit $\mathcal{O}(n \cdot m)$ bestimmt werden, wobei $m = |E|$ und $n = |V| \leq m$.

Beweis. Die ersten Aussagen folgen sofort aus Satz 6.34. Die Aussage hinsichtlich der Laufzeit folgt aus der Tatsache, dass die Anzahl der Flusserhöhungsschritte beschränkt ist durch $\text{MaxFlow}(N_G)$ und

$$\text{MaxFlow}(N_G) \leq \sum_{v \in \text{post}(s)} c(s,v) = \sum_{v \in V_L} \underbrace{c(s,v)}_{=1} = |V_L| \leq n.$$

□

Bemerkung: Die Suche nach augmentierten Wegen kann mit einer Tiefen- oder Breitensuche in Laufzeit $\mathcal{O}(m)$ durchgeführt werden.

Definition 6.37 Dei $G = (V, E)$ ein bipartiter Graph mit Partition $V = V_L \cup V_R$ und $|V_L| \leq |V_R|$. Ein Matching M heißt **perfekt**, falls $|M| = |V_L|$ gilt.

Satz 6.38 (Heiratssatz von Hall) Sei $G = (V, E)$ ein bipartiter Graph mit Partition $V = V_L \cup V_R$ und $|V_L| \leq |V_R|$. Dann existiert ein perfektes Matching genau dann, wenn

$$|\,\text{post}(W)| \geq |W| \quad \text{für alle } W \subseteq V_L.$$

Beweis. "$\Rightarrow$" Ist M perfekt und $W \subseteq V_L$, so enthält M für alle $w \in W$ genau eine Kante $(w, w_M) \in E$. Die Knoten w_M liegen also in $\text{post}(W)$ und sind paarweise verschieden, dies bedeutet

$$|\,\text{post}(W)| \geq |W|.$$

"$\Leftarrow$" Sei $|\,\text{post}(W)| \geq |W|$ für alle $W \subseteq V_L$. Seien N_G das zum Matchingproblem gehörige Netzwerk und f eine maximale 0-1-Flussfunktion. Gilt

$$\text{flow}(f) = |V_L|,$$

so ist das entsprechende Matching gemäß Korollar 6.36 perfekt.
Sei S_f die Menge aller von s erreichbaren Knoten im Restgraph von f. Wegen $t \notin S$ und

$$\begin{aligned}
\text{cap}(S_f) &= \sum_{\substack{v \in S_f \\ w \in \text{post}(v) \setminus S_f}} c(v,w) \\
&= \sum_{\substack{v \in S_f \\ w \in \text{post}(v) \setminus S_f}} f(v,w) - \sum_{\substack{w \in S_f \\ v \in \text{pre}(w) \setminus S_f}} \underbrace{f(v,w)}_{=0} \\
&= \text{flow}(f)
\end{aligned}$$

ist S_f ein Schnitt mit minimaler Schnittkapazität, vergleiche den Beweis des Max-Flow-Min-Cut-Theorems. Wir zeigen zunächst, dass

$$\operatorname{post}(V_L \cap S_f) \subseteq S_f \tag{6.6}$$

gilt. Sei hierzu $v \in S_f \cap V_L$ beliebig und $w \in \operatorname{post}(v)$. Angenommen, (6.6) gilt nicht, dann ist $w \notin S_f$. Folglich ist (v, w) keine Kante im Restgraph von f. Die Kante (v, w) ist somit gesättigt, das heißt

$$f(v, w) = 1.$$

Da $\operatorname{pre}(v) = \{s\}$, folgt aus dem Kirchhoffschen Gesetz, dass $f(s, v) = 1$. Damit ist auch (s, v) keine Kante im Restgraph. Der Knoten v kann also im Restgraph nur über eine Rückwärtskante von der Quelle s erreicht werden. Daher gibt es ein $u \in \operatorname{post}(v) \cap S_f$ mit

$$f(v, u) = 1.$$

Es gibt folglich zwei direkte Nachfolger von v mit

$$f(v, w) = f(v, u) = 1.$$

Dies widerspricht aber dem Kirchhoffschen Gesetz, da (s, v) die einzige zu v führende Kante ist.

Damit gilt (6.6) und es folgt für alle Kanten (u, v) in N_G mit $u \in S_f$ und $v \in V \setminus S_f$, dass $u = s$ oder $v = t$ ist. Dies impliziert

$$\begin{aligned}
\operatorname{cap}(S_f) &= \sum_{\substack{u \in S_f \\ v \in \operatorname{post}(u) \setminus S_f}} \underbrace{c(u, v)}_{=1} \\
&= |\{(s, v) : (s, v) \text{ Kante in } N_G \text{ und } v \notin S_f\}| \\
&\qquad + |\{(v, t) : (v, t) \text{ Kante in } N_G \text{ und } v \in S_f\}| \\
&= |V_L \setminus S_f| + |\underbrace{S_f \cap V_R}_{\overset{(6.6)}{\supseteq} \operatorname{post}(V_L \cap S_f)}| \\
&\geq |V_L \setminus S_f| + |\operatorname{post}(V_L \cap S_f)|.
\end{aligned}$$

Nach Voraussetzung ist $|\operatorname{post}(W)| \geq |W|$, insbesondere

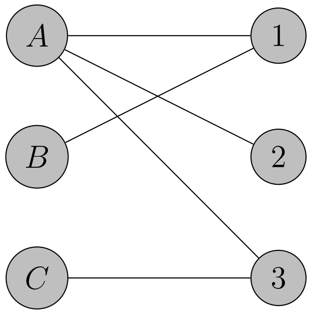
$$|\operatorname{post}(V_L \cap S_f)| \geq |V_L \cap S_f|,$$

und wir erhalten

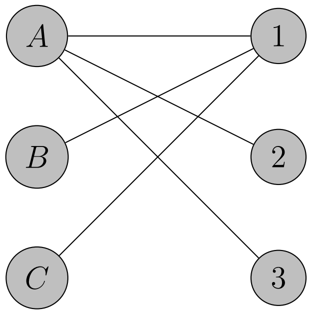
$$\begin{aligned}
\operatorname{flow}(f) &= \operatorname{cap}(S_f) \\
&\geq |V_L \setminus S_f| + |\operatorname{post}(V_L \cap S_f)| \\
&\geq |V_L \setminus S_f| + |V_L \cap S_f| \\
&= |V_L|.
\end{aligned}$$

Andererseits gilt offensichtlich auch $\operatorname{flow}(f) \leq |V_L|$ und somit $\operatorname{flow}(f) = |V_L|$. □

Beispiel 6.39 Beim Graph



ist das Hallsche Kriterium erfüllt, denn A, B, C haben jeweils mindestens einen Nachfolger, $\{A, B\}$, $\{B, C\}$, $\{A, C\}$ mindestens zwei Nachfolger und $\{A, B, C\}$ drei Nachfolger.
Zum Graph



existiert kein perfektes Matching, denn $\{B, C\}$ hat nur einen Nachfolger. △

Der Heiratssatz von Hall liefert zwar kein zufriedenstellendes algorithmisches Kriterium für die Existenz eines perfekten Matchings, da alle Teilmengen $W \subseteq V_L$ betrachtet werden müssen. Jedoch ermöglicht er diesen Nachweis in Graphen, in denen alle Knoten denselben Verzweigungsgrad haben. Solche Graphen heißen auch *regulär*.

Satz 6.40 Sei $G = (V, E)$ ein bipartiter Graph mit Partition $V = V_L \cup V_R$ und $|V_L| \leq |V_R|$. Gilt $|\operatorname{post}(v)| = k > 0$ für alle $v \in V$, so existiert ein perfektes Matching.

Beweis. Seien $W \subseteq V_L$ und

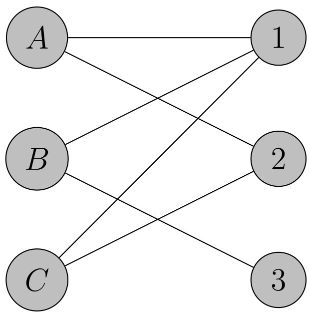
$$E_1 := \{(w, v) \in E : w \in W\},$$
$$E_2 := \{(u, v) \in E : v \in \operatorname{post}(W)\}.$$

Aus $E_1 \subseteq E_2$ folgt

$$k|W| = |E_1| \leq |E_2| = k|\operatorname{post}(W)|,$$

das heißt $|W| \leq |\operatorname{post}(W)|$. Satz 6.38 liefert dann die Behauptung. □

Beispiel 6.41 Der Graph



erfüllt $|\operatorname{post}(v)| = 2$ für alle Knoten $v \in V$. Hier ist

$$M = \{(A, 3), (B, 1), (C, 2)\}$$

ein perfektes Matching. $\triangle$

7. Lineare Gleichungssysteme

7.1 Vektor- und Matrixnormen

Im folgenden bezeichnet $\mathbb{R}^n$ ($\mathbb{C}^n$) den Raum der reellwertigen (komplexwertigen) Vektoren

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad x_i \in \mathbb{R} \quad (\mathbb{C})$$

und $\mathbb{R}^{m\times n}$ ($\mathbb{C}^{m\times n}$) den Raum der reellwertigen (komplexen) Matrizen

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & & a_{2,n} \\ \vdots & & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix}, \qquad a_{i,j} \in \mathbb{R} \quad (\mathbb{C}).$$

Meist ist es unerheblich, ob der zugrundeliegende Körper reell oder komplex ist. In diesem Fall schreiben wir $\mathbb{K}$ statt $\mathbb{R}$ oder $\mathbb{C}$.

Definition 7.1 Sei $X = \mathbb{K}^n$ oder $X = \mathbb{K}^{m\times n}$. Eine Abbildung

$$\|\cdot\| : X \to \mathbb{R}_{\geq 0}$$

heißt **Norm** auf X, wenn gilt

1. $\|\mathbf{x}\| > 0 \quad \forall \mathbf{x} \in X \setminus \{\mathbf{0}\}$
2. $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\| \quad \forall \mathbf{x} \in X,\ \alpha \in \mathbb{K}$
3. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X$

Bemerkung: Wegen $\mathbf{x} = \mathbf{x} - \mathbf{0}$ kann $\|\mathbf{x}\|$ als Abstand von $\mathbf{x}$ zum Nullpunkt in X interpretiert werden. In der Tat hat $\operatorname{dist}(\mathbf{x}, \mathbf{y}) := \|\mathbf{x}-\mathbf{y}\|$ die Eigenschaften einer *Distanz* von zwei Elementen. Der Begriff Distanz ist allerdings allgemeiner, und nicht nur auf (normierte) Vektorräume beschränkt. Insofern liefern Normen spezielle Distanzbegriffe.

Häufig verwendete Normen:

1. $X = \mathbb{K}^n$:

Betragssummennorm: $$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|$$

Euklidnorm: $$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{\mathbf{x}^\star \cdot \mathbf{x}}, \quad \mathbf{x}^\star = [\overline{x}_1, \overline{x}_2, ..., \overline{x}_n]$$

($\mathbf{x}^\star$ ist der zu $\mathbf{x}$ konjugiert komplexe Vektor)

Maximumnorm: $$\|\mathbf{x}\|_\infty := \max_{1 \le i \le n} |x_i|$$

2. $X = \mathbb{K}^{m \times n}$:

Spaltensummennorm: $$\|\mathbf{A}\|_1 := \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{i,j}|$$

Zeilensummennorm: $$\|\mathbf{A}\|_\infty := \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{i,j}|$$

Frobeniusnorm: $$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j} |a_{i,j}|^2}$$

Beispiel 7.2 Für

$$\mathbf{A} = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix}$$

gilt

$$\|\mathbf{A}\|_1 = 4, \qquad \|\mathbf{A}\|_\infty = 5, \qquad \|\mathbf{A}\|_F = \sqrt{15}.$$

△

Satz 7.3 Alle Normen auf $\mathbb{K}^n$ sind äquivalent, das heißt für zwei Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ auf $\mathbb{K}^n$ gibt es positive Konstanten $c, C > 0$ mit

$$c\|\mathbf{x}\|_a \le \|\mathbf{x}\|_b \le C\|\mathbf{x}\|_a \quad \forall \mathbf{x} \in \mathbb{K}^n.$$

Beweis. Es genügt die Behauptung für $\|\cdot\|_a = \|\cdot\|_\infty$ zu zeigen. Dazu seien $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ beliebig und $\|\cdot\|$ eine Norm im $\mathbb{K}^n$. Wegen

$$\mathbf{x} - \mathbf{y} = \sum_{i=1}^{m} (x_i - y_i)\mathbf{e}_i$$

folgt

$$\big|\|\mathbf{x}\| - \|\mathbf{y}\|\big| \le \|\mathbf{x} - \mathbf{y}\| \le \sum_{i=1}^{m} |x_i - y_i| \|\mathbf{e}_i\| \le \|\mathbf{x} - \mathbf{y}\|_\infty \sum_{i=1}^{m} \|\mathbf{e}_i\|.$$

Folglich ist $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$ eine Lipschitz-stetige Funktion mit Lipschitz-Konstante $L := \sum_{i=1}^m \|\mathbf{e}_i\|$. Als solche nimmt $\|\cdot\|$ auf der kompakten Einheitssphäre $\{\mathbf{x} \in \mathbb{K}^n : \|\mathbf{x}\|_\infty = 1\}$ sowohl ihr Maximum C als auch ihr Minimum c an. Wegen der ersten Normeigenschaft aus Definition 7.1 ist insbesondere $c > 0$. Daher folgt für beliebiges $\mathbf{z} \in \mathbb{K}^n$, dass

$$c \leq \left\| \frac{\mathbf{z}}{\|\mathbf{z}\|_\infty} \right\| \leq C,$$

beziehungsweise

$$c\|\mathbf{z}\|_\infty \leq \|\mathbf{z}\| \leq C\|\mathbf{z}\|_\infty.$$

$\square$

Beispiel 7.4 Für $\mathbf{x} \in \mathbb{K}^n$ folgt aus

$$\max_{1\leq i\leq n} |x_i|^2 \leq \sum_{k=1}^n \underbrace{|x_k|^2}_{\leq \max_{i=1}^n |x_i|^2} \leq n \cdot \max_{1\leq i\leq n} |x_i|^2$$

sofort die Ungleichung

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n} \cdot \|\mathbf{x}\|_\infty.$$

$\triangle$

Bemerkung: Satz 7.3 gilt auch für Matrizen, das heißt, im Falle des $\mathbb{K}^{m\times n}$.

Der Vektorraum $\mathbb{K}^{n\times n}$ unterscheidet sich von allen anderen genannten Räumen dadurch, dass eine weitere Operation definiert ist, nämlich die Multiplikation $\mathbf{A} \cdot \mathbf{B}$ mit $\mathbf{A}, \mathbf{B} \in \mathbb{K}^{n\times n}$.

Definition 7.5 Eine Matrixnorm $\|\cdot\|_M$ auf $\mathbb{K}^{n\times n}$ heißt **submultiplikativ**, falls gilt

$$\|\mathbf{A} \cdot \mathbf{B}\|_M \leq \|\mathbf{A}\|_M \cdot \|\mathbf{B}\|_M \quad \forall \mathbf{A}, \mathbf{B} \in \mathbb{K}^{n\times n}.$$

Eine Matrixnorm $\|\cdot\|_M$ auf $\mathbb{K}^{n\times n}$ heißt **verträglich** mit einer Vektornorm $\|\cdot\|_V$ auf $\mathbb{K}^n$, wenn gilt

$$\|\mathbf{A} \cdot \mathbf{x}\|_V \leq \|\mathbf{A}\|_M \cdot \|\mathbf{x}\|_V \quad \forall \mathbf{A} \in \mathbb{K}^{n\times n},\ \mathbf{x} \in \mathbb{K}^n.$$

Beispiel 7.6

1. $\|\mathbf{A}\| := \max_{1\leq i,j\leq n} |a_{i,j}|$ ist eine Matrixnorm auf $\mathbb{K}^{n\times n}$, aber nicht submultiplikativ:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \|\mathbf{A}\| = 1,$$
$$\mathbf{A}^2 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \quad \|\mathbf{A}^2\| = 2 \neq 1 = \|\mathbf{A}\|^2.$$

2. Die Frobeniusnorm ist mit der Euklidnorm verträglich:

$$\begin{aligned}\underbrace{[\mathbf{Ax}]_i^2}_{i\text{-te Komponente von } \mathbf{Ax}} &= \left(\sum_{j=1}^n a_{i,j} \cdot x_j\right)^2 \overset{\text{CSU}}{\leq} \left(\sum_{j=1}^n |a_{i,j}|^2\right) \cdot \left(\sum_{j=1}^n |x_j|^2\right) \\ &= \left(\sum_{j=1}^n |a_{i,j}|^2\right) \cdot \|\mathbf{x}\|_2^2\end{aligned}$$

$$\Longrightarrow \quad \|\mathbf{Ax}\|_2^2 = \sum_{i=1}^n [\mathbf{Ax}]_i^2 \leq \underbrace{\sum_{i=1}^n \left(\sum_{j=1}^n |a_{i,j}|^2\right)}_{=\|\mathbf{A}\|_F^2} \cdot \|\mathbf{x}\|_2^2 = \|\mathbf{A}\|_F^2 \cdot \|\mathbf{x}\|_2^2.$$

△

Definition 7.7 Sei $\|\cdot\|_V$ eine Vektornorm auf $\mathbb{K}^n$. Dann ist

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

eine Norm auf $\mathbb{K}^{n\times n}$, die sogenannte **induzierte Norm** von $\|\cdot\|_V$. (Die Normeigenschaften sind trivial nachgerechnet.)

Lemma 7.8 Die von $\|\cdot\|_V$ induzierte Norm $\|\cdot\|$ ist submultiplikativ und ist mit der Ausgangsnorm verträglich. Ist $\|\cdot\|_M$ eine mit $\|\cdot\|_V$ verträgliche Norm, dann gilt

$$\|\mathbf{A}\| \leq \|\mathbf{A}\|_M \quad \forall \mathbf{A} \in \mathbb{K}^{n\times n}.$$

Beweis. 1. Sei $m = n$ und $B \neq 0$, dann gilt

$$\begin{aligned}\|\mathbf{AB}\| &= \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_V}{\|\mathbf{x}\|_V} = \sup_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_V}{\|\mathbf{x}\|_V} = \sup_{\mathbf{Bx}\neq\mathbf{0}} \left(\frac{\|\mathbf{ABx}\|_V}{\|\mathbf{Bx}\|_V} \cdot \frac{\|\mathbf{Bx}\|_V}{\|\mathbf{x}\|_V}\right) \\ &\leq \sup_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_V}{\|\mathbf{Bx}\|_V} \cdot \sup_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|_V}{\|\mathbf{x}\|_V} \leq \sup_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|_V}{\|\mathbf{y}\|_V} \cdot \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|_V}{\|\mathbf{x}\|_V} \\ &= \|\mathbf{A}\| \cdot \|\mathbf{B}\|.\end{aligned}$$

2. Aus

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_V}{\|\mathbf{x}\|_V} \geq \frac{\|\mathbf{Ax}\|_V}{\|\mathbf{x}\|_V} \quad \forall \mathbf{x} \in \mathbb{K}^n \backslash \{\mathbf{0}\}$$

folgt

$$\|\mathbf{Ax}\|_V \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|_V \quad \forall \mathbf{x} \in \mathbb{K}^n \backslash \{\mathbf{0}\}.$$

Im Falle $\mathbf{x} = \mathbf{0}$ folgt sofort

$$0 = \|\mathbf{Ax}\|_V \leq \|\mathbf{A}\| \cdot \underbrace{\|\mathbf{x}\|_V}_{=0} = 0.$$

3. Die Behauptung ergibt sich aus
$$\|\mathbf{A}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_V}{\|\mathbf{x}\|_V} \leq \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\|_M \cdot \|\mathbf{x}\|_V}{\|\mathbf{x}\|_V} = \|\mathbf{A}\|_M.$$

□

Bemerkung: Die Spaltensummennorm ist von der Betragssummennorm induziert, die Zeilensummennorm ist von der Maximumsnorm induziert.

Beachte: Verallgemeinerungen von Definitionen 7.5 und 7.7 auf Matrixnormen im $\mathbb{K}^{m\times n}$ gelten entsprechend; in diesem Fall müssen dann Vektornormen sowohl für den $\mathbb{K}^m$ als auch den $\mathbb{K}^n$ spezifiziert werden.

Frage: Welche Matrixnorm wird durch die Euklidnorm induziert?
Um diese Frage zu beantworten, betrachten wir
$$\|\mathbf{A}\|_2 := \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \sqrt{(\mathbf{A}\mathbf{x})^\star(\mathbf{A}\mathbf{x})} = \max_{\|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x}}.$$

Satz 7.9 Es gilt
$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^\star\mathbf{A})} = \sqrt{\max\{\lambda : \lambda \text{ ist Eigenwert von } \mathbf{A}^\star\mathbf{A}\}}.$$

Beweis. $\mathbf{A}^\star\mathbf{A}$ ist hermitesch (d.h. $\mathbf{A}^\star = \mathbf{A}$) und positiv semidefinit (d.h. alle Eigenwerte ≥ 0). Also hat $\mathbf{A}^\star\mathbf{A}$ n nichtnegative Eigenwerte $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ und zugehörige, paarweise orthonormale Eigenvektoren $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Jeder Vektor $\mathbf{x} \in \mathbb{K}^n$, $\|\mathbf{x}\|_2 = 1$, lässt sich entwickeln
$$\mathbf{x} = \sum_{i=1}^n \xi_i \cdot \mathbf{v}_i,$$
woraus folgt
$$1 = \|\mathbf{x}\|_2^2 = \mathbf{x}^\star\mathbf{x} = \left(\sum_{i=1}^n \overline{\xi}_i \mathbf{v}_i^\star\right) \cdot \left(\sum_{j=1}^n \xi_j \mathbf{v}_j\right) = \sum_{i,j=1}^n \overline{\xi}_i \xi_j \cdot \underbrace{\mathbf{v}_i^\star\mathbf{v}_j}_{=\delta_{i,j}} = \sum_{i=1}^n |\xi_i|^2.$$
Einerseits ergibt sich nun
$$\begin{aligned}\mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x} &= \left(\sum_{i=1}^n \overline{\xi}_i \mathbf{v}_i^\star\right) \cdot \underbrace{\mathbf{A}^\star\mathbf{A} \cdot \left(\sum_{j=1}^n \xi_j \mathbf{v}_j\right)}_{=\sum_{j=1}^n \xi_j \mathbf{A}^\star\mathbf{A}\mathbf{v}_j} = \sum_{i,j=1}^n \overline{\xi}_i \xi_j \lambda_j \cdot \underbrace{\mathbf{v}_i^\star\mathbf{v}_j}_{=\delta_{i,j}} \\ &= \sum_{i=1}^n |\xi_i|^2 \cdot \lambda_i \leq \lambda_1 \cdot \underbrace{\sum_{i=1}^n |\xi_i|^2}_{=1} = \lambda_1.\end{aligned}$$
Andererseits gilt aber auch
$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\star\mathbf{A}^\star\mathbf{A}\mathbf{x} \geq \mathbf{v}_1^\star\mathbf{A}^\star\mathbf{A}\mathbf{v}_1 = \lambda_1 \cdot \mathbf{v}_1^\star\mathbf{v}_1 = \lambda_1.$$

Dies bedeutet aber

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\star \mathbf{A}^\star \mathbf{A}\mathbf{x} = \lambda_1,$$

woraus die Behauptung folgt. □

Bemerkung: Wegen Satz 7.9 nennt man die $\|\cdot\|_2$-Matrixnorm auch *Spektralnorm*.

Beispiel 7.10 (Fortsetzung von Beipiel (7.2)) Wir wollen für

$$\mathbf{A} = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix}$$

die Spektralnorm berechnen. Die Eigenwerte der Matrix

$$\mathbf{A}^\star \mathbf{A} = \begin{bmatrix} 5 & -5 \\ 5 & 10 \end{bmatrix}$$

kann man mit Hilfe der Regel von Sarrus über

$$\det(\mathbf{A}^\star\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} 5-\lambda & -5 \\ 5 & 10-\lambda \end{vmatrix} = \underbrace{(5-\lambda)(10-\lambda)}_{\lambda^2 - 15\lambda + 50} + 25 \stackrel{!}{=} 0$$

bestimmen. Es folgt

$$\lambda_{1/2} = \frac{15 \pm 5\sqrt{5}}{2}$$

und damit

$$\|\mathbf{A}\|_2 = \sqrt{\frac{1}{2} \cdot (15 + 5\sqrt{5})}.$$

△

7.2 Fehlerbetrachtungen

Für eine nichtsinguläre Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ wollen wir das lineare Gleichungssystem $\mathbf{A}\mathbf{x} = \mathbf{b}$ lösen. Offensichtlich ist bei Eingangsfehler $\Delta\mathbf{b}$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \qquad \mathbf{x} + \Delta\mathbf{x} = \mathbf{A}^{-1}(\mathbf{b} + \Delta\mathbf{b}) = \mathbf{A}^{-1}\mathbf{b} + \mathbf{A}^{-1}\Delta\mathbf{b},$$

das heißt

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}.$$

Für ein verträgliches Matrix/Vektornormpaar ergibt sich somit

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}^{-1}\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}^{-1}\|\|\mathbf{A}\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{7.1}$$

Definition 7.11 Der Faktor

$$\operatorname{cond}_M(\mathbf{A}) := \|\mathbf{A}^{-1}\|_M \|\mathbf{A}\|_M$$

wird als **Kondition** der Matrix $\mathbf{A}$ bzgl. der Matrixnorm $\|\cdot\|_M$ bezeichnet.

Wie in Kapitel 2 beschreibt die Kondition die relative Fehlerverstärkung in diesem Problem, diesmal allerdings normweise für den schlimmstmöglichen Fall. Ist die Matrixnorm $\|\cdot\|_M$ durch eine Vektornorm induziert, so kann man einfache Beispiele für $\mathbf{b}$ und $\Delta\mathbf{b}$ konstruieren, für die diese Fehlerverstärkung exakt ist ("=").

Beispiel 7.12 Sei

$$\mathbf{A} = \begin{bmatrix} 10^{-3} & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Die Matrix $\mathbf{A}$ ist gut konditioniert, denn mit

$$\mathbf{A}^{-1} \approx \begin{bmatrix} -2.004 & 1.002 \\ 1.002 & -0.001 \end{bmatrix}$$

folgt $3 = \|\mathbf{A}\|_\infty \approx \|\mathbf{A}^{-1}\|_\infty$ und daher ist

$$\mathrm{cond}_\infty(\mathbf{A}) \approx 9.$$

Die Lösung von $\mathbf{Ax} = \mathbf{b}$ ist also gut konditioniert. Sie lautet

$$\mathbf{x} \approx \begin{bmatrix} 1.002\ldots \\ 0.9989\ldots \end{bmatrix}.$$

Mit dem Gauß-Algorithmus und dreistelliger Gleitkommaarithmetik ergibt sich bei kleiner Datenstörung

$$\left[\begin{array}{cc|c} 0.001 & 1 & 1.01 \\ 1 & 2 & 3.01 \end{array}\right] \begin{array}{c} -1000 \\ \hookleftarrow \end{array} \longrightarrow \left[\begin{array}{cc|c} 0.001 & 1 & 1.01 \\ 0 & -998 & -1010 \end{array}\right]$$

$$\Longrightarrow \quad \begin{aligned} x_2 &= 1010 \boxslash 998 = 1.01 \\ x_1 &= 1000 \boxdot (1.01 \boxminus 1.01) = 0 \end{aligned}$$

△

Der Grund: Das kleine (1,1)-Element bewirkt einen großen Faktor (nämlich -1000) und damit starke Fehlerverstärkung, das heißt Instabilität.

Die Diagonalelemente, die bei der Gauß-Elimination im i-ten Schritt an Position (i,i) auftreten, werden *Pivotelemente* genannt.

Zur Stabilisierung der Gauß-Elimination vertauscht man daher vor jedem Eliminierungsschritt die i-te und die k-te Zeile derart, dass das Pivotelement am betragsgrößten ist ("Spaltenpivotsuche" oder "partial pivoting"). Am exakten Resultat ändert das nichts!

Beispiel 7.13 (Fortsetzung von Beipiel 7.12) In unserem Fall würde man also die beiden Zeilen vertauschen, da $1 > 0.001$:

$$\left[\begin{array}{cc|c} 1 & 2 & 3.01 \\ 0.001 & 1 & 1.01 \end{array}\right] \begin{array}{c} -\frac{1}{1000} \\ \hookleftarrow \end{array} \longrightarrow \left[\begin{array}{cc|c} 1 & 2 & 3.01 \\ 0 & 0.998 & 1.01 \end{array}\right]$$

$$\Longrightarrow \quad \begin{aligned} x_2 &= 1.01 \boxslash 0.998 = 1.01 \\ x_1 &= 3.01 \boxminus (2 \boxdot 1.01) = 3.01 \boxminus 2.02 = 0.99 \end{aligned}$$

△

Die Auswahl des Pivotelements hängt stark von der Skalierung des linearen Gleichungssystems ab. Beispielsweise könnte man auch einfach die erste Gleichung mit 1000 multiplizieren und das (1,1)-Element als Pivot behalten. Das Ergebnis wäre dann wieder so verheerend wie vorher. Man wählt daher im i-ten Teilschritt das Element, welches am betragsgrößten ist ("totale Pivotisierung" oder "total pivoting"):

Im i-ten Teilschritt sei

$$\mathbf{A}_i = \begin{bmatrix} * & \cdots & * & * & \cdots & * \\ & \ddots & \vdots & \vdots & & \vdots \\ & & * & * & \cdots & * \\ & & & a_{i,i}^{(i)} & \cdots & a_{i,n}^{(i)} \\ & \mathbf{0} & & a_{i+1,i}^{(i)} & \cdots & a_{i+1,n}^{(i)} \\ & & & \vdots & & \\ & & & a_{n,i}^{(i)} & \cdots & a_{n,n}^{(i)} \end{bmatrix}.$$

Das entsprechende Element $a_{\ell,m}^{(i)}$ mit

$$|a_{\ell,m}^{(i)}| = \max_{i \le j,k \le n} |a_{j,k}^{(i)}|$$

kann dadurch in Position (i,i) gebracht werden, indem man wie zuvor die Zeilen i und ℓ und zusätzlich noch die Spalten i und m vertauscht. Dies ist die *stabilste* Variante des Gauß-Algorithmus.

7.3 LR-Zerlegung

Erinnerung: Im i-ten Teilschritt $(1 \le i \le n-1)$ des Gauß-Algorithmus geht man wie folgt vor:

$$\left[\begin{array}{c|c} \underbrace{\begin{matrix} \star & \cdots & \star & \star & \cdots & \star \\ & \ddots & \vdots & \vdots & & \vdots \\ & & \star & \star & \cdots & \star \\ & & & a_{i,i}^{(i)} & \cdots & a_{i,n}^{(i)} \\ & \mathbf{0} & & a_{i+1,i}^{(i)} & \cdots & a_{i+1,n}^{(i)} \\ & & & \vdots & & \vdots \\ & & & a_{n,i}^{(i)} & \cdots & a_{n,n}^{(i)} \end{matrix}}_{= \mathbf{A}_i} & \underbrace{\begin{matrix} \star \\ \vdots \\ \star \\ b_i^{(i)} \\ b_{i+1}^{(i)} \\ \vdots \\ b_n^{(i)} \end{matrix}}_{= \mathbf{b}_i} \end{array}\right] \quad -\tau_{i+1}^{(i)} \quad \cdots \quad -\tau_n^{(i)}$$

mit

$$\tau_j^{(i)} = \frac{a_{j,i}^{(i)}}{a_{i,i}^{(i)}}, \qquad i < j \le n.$$

Der gemeinsame Nenner $a_{i,i}^{(i)}$ der Faktoren $\tau_j^{(i)}$ wird *Pivotelement* genannt. Der obige Elimininationsschritt kann in Matrixnotation wie folgt geschrieben werden:

$$\underbrace{\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & \mathbf{0} & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & -\tau_{i+1}^{(i)} & 1 & & \\ & \mathbf{0} & & \vdots & & \ddots & \\ & & & -\tau_n^{(i)} & & & 1 \end{bmatrix}}_{=\mathbf{L}_i} \left[\, \mathbf{A}_i \,\middle|\, \mathbf{b}_i \,\right] = \left[\, \mathbf{A}_{i+1} \,\middle|\, \mathbf{b}_{i+1} \,\right],$$

wobei

$$\mathbf{L}_i = \mathbf{I} - \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tau_{i+1}^{(i)} \\ \vdots \\ \tau_n^{(i)} \end{bmatrix}}_{=:\,\boldsymbol{l}_i} \underbrace{\begin{bmatrix} 0 & \cdots & 0 & \overbrace{1}^{\text{Stelle } i} & 0 & \cdots & 0 \end{bmatrix}}_{=:\,\mathbf{e}_i^\star}.$$

Mit $\mathbf{A}_1 = \mathbf{A}$ und $\mathbf{b}_1 = \mathbf{b}$ ergibt sich durch Auflösen der Rekursion

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1 \left[\, \mathbf{A} \,\middle|\, \mathbf{b} \,\right] = \left[\, \mathbf{A}_n \,\middle|\, \mathbf{b}_n \,\right] = \left[\, \mathbf{R} \,\middle|\, \mathbf{c} \,\right] = \left[\begin{array}{ccc|c} \star & \cdots & \star & \star \\ & \ddots & \vdots & \vdots \\ \mathbf{0} & & \star & \star \end{array}\right].$$

Insbesondere gilt

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2}\cdots\mathbf{L}_1\mathbf{A} = \mathbf{R},$$

das heißt die Faktorisierung

$$\mathbf{A} = \underbrace{\mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_{n-1}^{-1}}_{=:\mathbf{L}}\mathbf{R} = \mathbf{L}\mathbf{R}.$$

Die inversen Matrizen $\mathbf{L}_i^{-1}$ sowie $\mathbf{L}$ lassen sich explizit angeben:

Lemma 7.14

1. Die Inverse von $\mathbf{L}_i = \mathbf{I} - \boldsymbol{l}_i\mathbf{e}_i^\star$ berechnet sich gemäß

$$\mathbf{L}_i^{-1} = \mathbf{I} + \boldsymbol{l}_i\mathbf{e}_i^\star = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & \mathbf{0} & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & \tau_{i+1}^{(i)} & 1 & & \\ & \mathbf{0} & & \vdots & & \ddots & \\ & & & \tau_n^{(i)} & & & 1 \end{bmatrix}.$$

2. Die Matrix $\mathbf{L}$ erfüllt

$$\mathbf{L} = \mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_{n-1}\mathbf{e}_{n-1}^\star = \begin{bmatrix} 1 & & & & \\ \tau_2^{(1)} & 1 & & \mathbf{0} & \\ \tau_3^{(1)} & \tau_3^{(2)} & \ddots & & \\ \vdots & \vdots & & 1 & \\ \tau_n^{(1)} & \tau_n^{(2)} & \cdots & \tau_n^{(n-1)} & 1 \end{bmatrix}. \tag{7.2}$$

Beweis. Aufgrund der Nulleinträge in $\boldsymbol{l}_i$ und $\mathbf{e}_i$ ist $\mathbf{e}_i^\star\boldsymbol{l}_j = 0$ für $i \leq j$. Daraus folgt

$$\underbrace{(\mathbf{I} - \boldsymbol{l}_i\mathbf{e}_i^\star)}_{=\mathbf{L}_i}(\mathbf{I} + \boldsymbol{l}_i\mathbf{e}_i^\star) = \mathbf{I} - \boldsymbol{l}_i\mathbf{e}_i^\star + \boldsymbol{l}_i\mathbf{e}_i^\star - \boldsymbol{l}_i \underbrace{\mathbf{e}_i^\star\boldsymbol{l}_i}_{=0} \mathbf{e}_i^\star = \mathbf{I}.$$

Weiter ergibt sich induktiv aus

$$\mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_i^{-1} = \mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star$$

dass

$$\begin{aligned} \mathbf{L}_1^{-1}\mathbf{L}_2^{-1}\cdots\mathbf{L}_{i+1}^{-1} &= (\mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star)\mathbf{L}_{i+1}^{-1} \\ &= (\mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star)(\mathbf{I} + \boldsymbol{l}_{i+1}\mathbf{e}_{i+1}^\star) \\ &= \mathbf{I} + \boldsymbol{l}_1\mathbf{e}_1^\star + \boldsymbol{l}_2\mathbf{e}_2^\star + \cdots + \boldsymbol{l}_i\mathbf{e}_i^\star + \boldsymbol{l}_{i+1}\mathbf{e}_{i+1}^\star + \sum_{j=1}^{i} \boldsymbol{l}_j \underbrace{\mathbf{e}_j^\star\boldsymbol{l}_{i+1}}_{=0} \mathbf{e}_{i+1}^\star. \end{aligned}$$

□

Wird im Verlauf des Gauß-Algorithmus ein Pivotelement $a_{i,i}^{(i)}$ Null, dann bricht das Verfahren in dieser Form zusammen. Sind hingegen alle Pivotelemente für $i = 1, 2, \ldots, n$ von Null verschieden, so haben wir das folgende Resultat bewiesen.

Satz 7.15 Falls kein Pivotelement Null wird, bestimmt der Gauß-Algorithmus neben der Lösung $\mathbf{x}$ von $\mathbf{A}\mathbf{x} = \mathbf{b}$ eine LR-Zerlegung $\mathbf{A} = \mathbf{L} \cdot \mathbf{R}$ in eine linke untere und eine rechte obere Dreiecksmatrix. Die Matrix $\mathbf{L}$ ist dabei durch (7.2) gegeben.

Beispiel 7.16

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{bmatrix} \begin{matrix} -2 & -3 \\ \hookleftarrow & \\ & \hookleftarrow \end{matrix} \longrightarrow \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{bmatrix} \begin{matrix} \\ -2 \\ \hookleftarrow \end{matrix} \longrightarrow \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{bmatrix},$$

das heißt

$$\mathbf{R} = \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}.$$

△

Bemerkung: Bei der Realisierung der LR-Zerlegung am Computer überschreibt man die ursprünglichen Einträge $a_{i,j}^{(1)} = a_{i,j}$ der Matrix $\mathbf{A}$ mit den jeweils aktuellen Einträgen $a_{i,j}^{(i)}$. Die Matrix $\mathbf{L}$ lässt sich sukzessive in die nicht mehr benötigte untere Hälfte von $\mathbf{A}$ schreiben. Damit wird kein zusätzlicher Speicherplatz für die LR-Zerlegung gebraucht.

Die Lösung des linearen Gleichungssystems $\mathbf{Ax} = \mathbf{b}$ wird mit Hilfe der LR-Zerlegung wie folgt berechnet:

① zerlege $\mathbf{A} = \mathbf{LR}$ mit dem Gauß-Algorithmus

② löse $\mathbf{Ax} = \mathbf{LRx} = \mathbf{b}$ in zwei Schritten:

- löse $\mathbf{Ly} = \mathbf{b}$ durch Vorwärtssubstitution
- löse $\mathbf{Rx} = \mathbf{y}$ durch Rückwärtssubstitution

Aufwand:

① Im i-ten Teilschritt werden

$$(n-i+1)(n-i) = (n-i)^2 + n - i$$

Multiplikationen benötigt, das sind insgesamt

$$\sum_{i=1}^{n-1} \left\{(n-i)^2 + n - i\right\} \overset{j:=n-i}{=} \sum_{j=1}^{n-1} (j^2 + j) = \frac{1}{3}n^3 + \mathcal{O}(n^2)$$

Multiplikationen.

② Hier werden

$$2\sum_{i=1}^{n} i = n(n+1) = \mathcal{O}(n^2)$$

Multiplikationen benötigt.

Demnach werden also insgesamt zum Lösen eines linearen Gleichungssystems mit Hilfe der LR-Zerlegung $n^3/3 + \mathcal{O}(n^2)$ Multiplikationen (und, wie man leicht nachrechnet, nochmals ebensoviele Additionen) benötigt. Der Speicherplatzbedarf ist dabei allerdings nur von der Ordnung $\mathcal{O}(n^2)$.

Beispiel 7.17 (Fortsetzung von Beipiel 7.16) Für $\mathbf{b} = [1, 1, 1]^T$ wollen wir das lineare Gleichungssystem $\mathbf{Ax} = \mathbf{b}$ lösen:

1. bestimme $\mathbf{y}$ mit $\mathbf{Ly} = \mathbf{b}$ durch Vorwärtssubstitution:

$$\left[\ \mathbf{L} \,\middle|\, \mathbf{b}\ \right] = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 1 \end{array}\right] \quad \Longrightarrow \quad \begin{array}{l} y_1 = 1 \\ y_2 = 1 - 2 = -1 \\ y_3 = 1 - 3 + 2 = 0 \end{array}$$

2. bestimme $\mathbf{x}$ mit $\mathbf{Rx} = \mathbf{y}$ durch Rückwärtssubstitution:

$$\left[\ \mathbf{R} \,\middle|\, \mathbf{y}\ \right] = \left[\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & -1 \\ 0 & 0 & 1 & 0 \end{array}\right] \quad \Longrightarrow \quad \begin{array}{l} x_3 = 0 \\ x_2 = (-1 + 0)/(-3) = 1/3 \\ x_1 = 1 - 4/3 - 0 = -1/3 \end{array}$$

$\triangle$

Bemerkung: Eine weitere Möglichkeit, das lineare Gleichungssystem $\mathbf{A}\mathbf{x} = \mathbf{b}$ zu lösen, bietet bekanntlich die Cramersche Regel. Danach lautet die i-te Komponente x_i der Lösung

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}},$$

wobei hier $\mathbf{A}_i \in \mathbb{R}^{n\times n}$ diejenige Matrix ist, die aus $\mathbf{A}$ entsteht, wenn man die i-te Spalte durch $\mathbf{b}$ ersetzt. Berechnet man die Determinante nach dem Laplaceschen Entwicklungssatz, so benötigt man i.a. $n!$ Operationen. Bei Verwendung eines Rechners mit 10^8 Gleitkommaoperationen pro Sekunde (100 Megaflops) ergäben sich dann die folgenden Rechenzeiten:

n	10	12	14	16	18	20
Rechenzeit	0.4 s	1 min	3.6 h	41 Tage	38 Jahre	16 000 Jahre

Mit Spaltenpivotsuche wird die Matrizenformulierung des Gauß-Algorithmus schwieriger. Werden im i-ten Teilschritt die i-te und die k-te Zeile der Matrix $\mathbf{A}_i$ $(i < k \leq n)$ vertauscht, dann kann dies durch die zugehörige Permutationsmatrix

$$\mathbf{P}_i = \begin{array}{c} \begin{bmatrix} 1 &&&&&&&&&& \\ & \ddots &&&&&&&&& \\ && 1 &&&&&&&& \\ &&& 0 &&&& 1 &&& \\ &&&& 1 &&&&&& \\ &&&&& \ddots &&&&& \\ &&&&&& 1 &&&& \\ &&& 1 &&&& 0 &&& \\ &&&&&&&& 1 && \\ &&&&&&&&& \ddots & \\ &&&&&&&&&& 1 \end{bmatrix} \begin{array}{l} \\ \\ \\ \leftarrow i\text{-te Zeile} \\ \\ \\ \\ \leftarrow k\text{-te Zeile} \\ \\ \\ \\ \end{array} \\ \begin{array}{cc} \uparrow & \uparrow \\ i\text{-te Spalte} & k\text{-te Spalte} \end{array} \end{array} \tag{7.3}$$

beschrieben werden. Es gelten nämlich die folgenden Rechenregeln:

Rechenregeln:

1. Multiplikation mit $\mathbf{P}_i$ von *links* $\longrightarrow$ vertausche *Zeilen* i und k

2. Multiplikation mit $\mathbf{P}_i$ von *rechts* $\longrightarrow$ vertausche *Spalten* i und k

3. Insbesondere gilt $\mathbf{P}_i^2 = \mathbf{I}$ (dies gilt für jede Permutationsmatrix)

Folglich wird aus der Matrix $\mathbf{A}_i$ durch Vertauschung die Matrix $\mathbf{P}_i\mathbf{A}_i$. Diese Matrix wird nun im Eliminationsschritt weiter reduziert. Der vollständige i-te Teilschritt des Gauß-Algorithmus mit Spaltenpivotsuche transformiert $\mathbf{A}_i$ also wie folgt:

$$\mathbf{A}_i \mapsto \mathbf{A}_{i+1} = \mathbf{L}_i\mathbf{P}_i\mathbf{A}_i. \tag{7.4}$$

Wir benötigen folgendes Lemma:

Lemma 7.18 Sei $j < i$ und $\mathbf{P}_i$ durch (7.3) gegeben. Dann ist $\mathbf{P}_i\mathbf{L}_j = \widetilde{\mathbf{L}}_j\mathbf{P}_i$, wobei $\widetilde{\mathbf{L}}_j$ wieder die gleiche Form hat wie $\mathbf{L}_j$, außer dass $\tau_k^{(j)}$ und $\tau_i^{(j)}$ vertauscht sind.

Beweis. Wegen $\mathbf{P}_i^2 = \mathbf{I}$ gilt

$$\mathbf{P}_i\mathbf{L}_j = \mathbf{P}_i\mathbf{L}_j\mathbf{P}_i^2 = (\mathbf{P}_i\mathbf{L}_j\mathbf{P}_i)\mathbf{P}_i,$$

dies bedeutet

$$\widetilde{\mathbf{L}}_j = \mathbf{P}_i\mathbf{L}_j\mathbf{P}_i.$$

Mit Hilfe der obigen Rechenregeln folgt

$$\mathbf{P}_i\mathbf{L}_j\mathbf{P}_i = \begin{array}{r c}
\begin{array}{r} \\ \\ \\ \\ i\text{-te Zeile} \rightarrow \\ \\ \\ \\ k\text{-te Zeile} \rightarrow \\ \\ \\ \\ \end{array} &
\left[\begin{array}{cccccccccccc}
\ddots &&&&&&&&&&& \\
& 1 &&&&&&&&&& \\
& -\tau_{j+1}^{(j)} & 1 &&&&&&&&& \\
& \vdots && \ddots &&&&&&&& \\
& -\tau_k^{(j)} &&& 0 &&&& 1 &&& \\
& &&&& 1 &&&&&& \\
& \vdots &&&&& \ddots &&&&& \\
& &&&&&& 1 &&&& \\
& -\tau_i^{(j)} &&& 1 &&&& 0 &&& \\
& &&&&&&&& 1 && \\
& \vdots &&&&&&&&& \ddots & \\
& -\tau_n^{(j)} &&&&&&&&&& 1
\end{array}\right] \cdot \mathbf{P}_i \\
& \uparrow \qquad\qquad \uparrow \\
& i\text{-te Spalte} \quad k\text{-te Spalte}
\end{array}$$

$$= \begin{array}{r c}
\begin{array}{r} \\ \\ \\ \\ i\text{-te Zeile} \rightarrow \\ \\ \\ \\ k\text{-te Zeile} \rightarrow \\ \\ \\ \\ \end{array} &
\left[\begin{array}{cccccccccccc}
\ddots &&&&&&&&&&& \\
& 1 &&&&&&&&&& \\
& -\tau_{j+1}^{(j)} & 1 &&&&&&&&& \\
& \vdots && \ddots &&&&&&&& \\
& -\tau_k^{(j)} &&& 1 &&&&&&& \\
& &&&& 1 &&&&&& \\
& \vdots &&&&& \ddots &&&&& \\
& &&&&&& 1 &&&& \\
& -\tau_i^{(j)} &&&&&&& 1 &&& \\
& &&&&&&&& 1 && \\
& \vdots &&&&&&&&& \ddots & \\
& -\tau_n^{(j)} &&&&&&&&&& 1
\end{array}\right] . \\
& \uparrow \qquad\qquad \uparrow \\
& i\text{-te Spalte} \quad k\text{-te Spalte}
\end{array}$$

□

Damit können wir den folgenden Satz für den Gauß-Algorithmus mit Spaltenpivotsuche beweisen:

Satz 7.19 Ist $\mathbf{A}$ nichtsingulär, dann bestimmt der Gauß-Algorithmus mit Spaltenpivotsuche eine Zerlegung der Matrix $\mathbf{P} \cdot \mathbf{A} = \widetilde{\mathbf{L}}\mathbf{R}$, wobei $\mathbf{R}$ wie zuvor die rechte obere Dreiecksmatrix $\mathbf{A}_n$, $\mathbf{P} = \mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdots \mathbf{P}_1$ eine Permutationsmatrix und

$$\widetilde{\mathbf{L}} = \widetilde{\mathbf{L}}_1^{-1}\widetilde{\mathbf{L}}_2^{-1} \cdots \widetilde{\mathbf{L}}_{n-1}^{-1}$$

eine linke untere Dreiecksmatrix ist mit

$$\begin{aligned}
\widetilde{\mathbf{L}}_{n-1} &= \mathbf{L}_{n-1}, \\
\widetilde{\mathbf{L}}_{n-2} &= \mathbf{P}_{n-1}\mathbf{L}_{n-2}\mathbf{P}_{n-1}, \\
\widetilde{\mathbf{L}}_{n-3} &= \mathbf{P}_{n-1}\mathbf{P}_{n-2}\mathbf{L}_{n-3}\mathbf{P}_{n-2}\mathbf{P}_{n-1}, \\
&\ \vdots \\
\widetilde{\mathbf{L}}_1 &= \mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdots \mathbf{P}_2\mathbf{L}_1\mathbf{P}_2 \cdots \mathbf{P}_{n-2}\mathbf{P}_{n-1}.
\end{aligned}$$

Beweis. Nehmen wir zunächst an, dass der Gauß-Algorithmus mit Spaltenpivotsuche nicht zusammenbricht. Dann ergibt sich aus (7.4) durch sukzessive Anwendung von Lemma 7.18 dass

$$\begin{aligned}
\mathbf{R} = \mathbf{A}_n &= \mathbf{L}_{n-1}\mathbf{P}_{n-1}\mathbf{A}_{n-1} \\
&= \widetilde{\mathbf{L}}_{n-1}\mathbf{P}_{n-1}\mathbf{L}_{n-2}\mathbf{P}_{n-2}\mathbf{A}_{n-2} \\
&= \widetilde{\mathbf{L}}_{n-1}\widetilde{\mathbf{L}}_{n-2}\mathbf{P}_{n-1}\mathbf{P}_{n-2}\mathbf{L}_{n-3}\mathbf{P}_{n-3}\mathbf{A}_{n-3} \\
&\ \vdots \\
&= \widetilde{\mathbf{L}}_{n-1}\widetilde{\mathbf{L}}_{n-2} \cdots \widetilde{\mathbf{L}}_1\mathbf{P}_{n-1}\mathbf{P}_{n-2} \cdots \mathbf{P}_1\mathbf{A}.
\end{aligned}$$

Zu klären bleibt schließlich der Punkt, dass der Gauß-Algorithmus mit Spaltenpivotsuche nicht abbricht, also dass alle Pivotelemente nach der Spaltenpivotsuche von Null verschieden sind. Wäre das Pivotelement nach dem i-ten Teilschritt tatsächlich Null, dann gälte zwangsläufig

$$\mathbf{A}_i = \begin{array}{c} \mathbf{B}\left[\vphantom{\begin{array}{c}\star\\ \ddots\\ \star\end{array}}\right. \\ \vphantom{\begin{array}{c}0\\ \vdots\\ 0\end{array}} \end{array} \left[\begin{array}{ccc|cccc}
\star & \cdots & \star & & & & \\
 & \ddots & \vdots & & \star & & \\
\mathbf{0} & & \star & & & & \\ \hline
 & & & 0 & \star & \cdots & \star \\
 & \mathbf{0} & & \vdots & \vdots & & \vdots \\
 & & & 0 & \star & \cdots & \star
\end{array}\right].$$

Daraus folgt jedoch

$$\det \mathbf{A}_i = \det \mathbf{B} \cdot \det \begin{bmatrix} 0 & \star & \cdots & \star \\ \vdots & \vdots & & \vdots \\ 0 & \star & \cdots & \star \end{bmatrix} = 0$$

und weiter

$$0 = \det \mathbf{A}_i = \det(\mathbf{L}_{i-1}\mathbf{P}_{i-1} \cdots \mathbf{L}_1\mathbf{P}_1\mathbf{A}) = \prod_{j=1}^{i-1} \underbrace{\det \mathbf{L}_j}_{=1} \cdot \prod_{j=1}^{i-1} \underbrace{\det \mathbf{P}_j}_{=\pm 1} \cdot \det \mathbf{A}.$$

Dies impliziert $\det \mathbf{A} = 0$ im Widerspruch zur Voraussetzung. $\square$

Beispiel 7.20

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \\ 1 & -7 & 9 & 10 \end{bmatrix} \begin{matrix} -1/2 & +1 & -1 \\ & & \\ & & \\ & & \end{matrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 0 & 2 & -2 \\ 0 & 1 & -1/8 & -3 \\ 0 & -8 & 9 & 8 \end{bmatrix}$$

$$\xrightarrow{\mathbf{P}_2} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 1 & -1/8 & -3 \\ 0 & 0 & 2 & -2 \end{bmatrix} \begin{matrix} \\ +1/8 \\ \\ \\ \end{matrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & -2 \end{bmatrix}$$

$$\xrightarrow{\mathbf{P}_3} \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 1 & -2 \end{bmatrix} \begin{matrix} \\ \\ -1/2 \\ \\ \end{matrix} \longrightarrow \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

Damit ergibt sich (beachte: $\mathbf{P}_1 = \mathbf{I}$)

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & -8 & 9 & 8 \\ 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \qquad \widetilde{\mathbf{L}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ -1 & -1/8 & 1/2 & 1 \end{bmatrix},$$

$$\mathbf{PA} = \begin{bmatrix} 1 & 1 & 0 & 2 \\ 1 & -7 & 9 & 10 \\ 1/2 & 1/2 & 2 & -1 \\ -1 & 0 & -1/8 & -5 \end{bmatrix}.$$

△

Faustregel: Um auf $\widetilde{\mathbf{L}}$ zu kommen, erstellt man zunächst eine Matrix $\mathbf{L}$ wie gewohnt, und führt dann in jeder Spalte *alle* Vertauschungen (in der Reihenfolge $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{n-1}$) durch, bei denen nur Elemente unterhalb der Diagonalen betroffen sind.

Totale Pivotisierung: Im i-ten Teilschritt wird das Element $a^{(i)}_{k,\ell}$ $(i \le k, \ell \le n)$ als Pivotelement gewählt, das in der gesamten verbliebenen Restmatrix betragsmäßig am größten ist. Da man hierzu Zeilen und Spalten tauschen muss, benötigt man formal zwei Permutationsmatrizen $\mathbf{P}_i$ bzw. $\mathbf{\Pi}_i$:

$$\mathbf{A}_i \mapsto \mathbf{A}_{i+1} = \mathbf{L}_i \mathbf{P}_i \mathbf{A}_i \mathbf{\Pi}_i.$$

Man erhält so schließlich eine *LR*-Zerlegung der Matrix $\mathbf{PA\Pi}$ mit $\mathbf{\Pi} = \mathbf{\Pi}_1 \mathbf{\Pi}_2 \cdots \mathbf{\Pi}_{n-1}$.

Wird die Totalpivotsuche bei der Lösung eines linearen Gleichungssystems eingesetzt, dann entsprechen Spaltenvertauschungen Permutationen der Lösung $\mathbf{x}$. Der Ergebnisvektor ist also nicht mehr in der richtigen Reihenfolge.

Die totale Pivotisierung ist stabil, wird aber in der Praxis nur selten eingesetzt, da die Suche nach dem betragsgrößten Element im i-ten Schritt einem Aufwand $(n-i)^2$ entspricht. Der Gesamtaufwand

$$\sum_{i=1}^{n-1} (n-i)^2 \overset{j:=n-i}{=} \sum_{j=1}^{n-1} j^2 = \frac{1}{3} n^3$$

ist nicht mehr vernachlässigbar gegenüber der eigentlichen Rechnung.

7.4 Cholesky-Zerlegung

Wir betrachten zu gegebener Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$ die Blockpartitionierung

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix}$$

mit $\mathbf{A}_{1,1} \in \mathbb{K}^{p\times p}$ und $\mathbf{A}_{2,2} \in \mathbb{K}^{(n-p)\times(n-p)}$. Ist $\mathbf{A}_{1,1}$ nichtsingulär, so kann man das lineare Gleichungssystem

$$\mathbf{A}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \tag{7.5}$$

vermittels Block-Gauß-Elimination lösen:

$$\mathbf{A} = \left[\begin{array}{cc|c} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{b} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{c} \end{array}\right] \begin{matrix} -\mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1} \\ \hookleftarrow \end{matrix} \quad \longrightarrow \quad \left[\begin{array}{cc|c} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{b} \\ 0 & \mathbf{A}_{2,2} - \mathbf{A}_{2,1}\mathbf{A}^{-1}\mathbf{A}_{1,2} & \mathbf{c} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{b} \end{array}\right]$$

Definition 7.21 Die Matrix

$$\mathbf{S} := \mathbf{A}_{2,2} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2} \in \mathbb{K}^{(n-p)\times(n-p)} \tag{7.6}$$

heißt **Schurkomplement** von $\mathbf{A}$ bezüglich $\mathbf{A}_{1,1}$.

Für die Lösung von (7.5) folgt nun

$$\begin{aligned} \mathbf{y} &= \mathbf{S}^{-1}\big(\mathbf{c} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{b}\big), \\ \mathbf{x} &= \mathbf{A}_{1,1}^{-1}(\mathbf{b} - \mathbf{A}_{1,2}\mathbf{y}). \end{aligned}$$

Lemma 7.22 Sei $\mathbf{A} \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit. Dann ist das Schurkomplement $\mathbf{S}$ wohldefiniert und sowohl $\mathbf{A}_{1,1}$ als auch $\mathbf{S}$ sind hermitesch und positiv definit.

Beweis. Sei $\left[\begin{smallmatrix} \mathbf{x} \\ \mathbf{y} \end{smallmatrix}\right]$ entsprechend zu $\mathbf{A}$ partitioniert. Wegen

$$\begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{bmatrix} = \mathbf{A} = \mathbf{A}^\star = \begin{bmatrix} \mathbf{A}_{1,1}^\star & \mathbf{A}_{2,1}^\star \\ \mathbf{A}_{1,2}^\star & \mathbf{A}_{2,2}^\star \end{bmatrix}$$

ergibt sich

$$\mathbf{A}_{1,1} = \mathbf{A}_{1,1}^\star, \qquad \mathbf{A}_{2,2} = \mathbf{A}_{2,2}^\star, \qquad \mathbf{A}_{1,2} = \mathbf{A}_{2,1}^\star.$$

Folglich ist $\mathbf{A}_{1,1}$ hermitesch und es gilt

$$0 \le \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix}^\star \mathbf{A} \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix}^\star \begin{bmatrix} \mathbf{A}_{1,1}\mathbf{x} \\ \mathbf{A}_{2,1}\mathbf{x} \end{bmatrix} = \mathbf{x}^\star\mathbf{A}_{1,1}\mathbf{x},$$

wobei sich Gleichheit nur für $\mathbf{x} = \mathbf{0}$ ergibt. Also ist $\mathbf{A}_{1,1}$ positiv definit und $\mathbf{A}_{1,1}^{-1}$ existiert. $\mathbf{S}$ ist somit wohldefiniert und

$$\mathbf{S}^\star = \mathbf{A}_{2,2}^\star - \mathbf{A}_{1,2}^\star\mathbf{A}_{1,1}^{-\star}\mathbf{A}_{2,1}^\star = \mathbf{A}_{2,2} - \mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2} = \mathbf{S}.$$

Schließlich betrachten wir $\left[\begin{smallmatrix} \mathbf{x} \\ \mathbf{y} \end{smallmatrix}\right]$ mit $\mathbf{x} = -\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\mathbf{y}$:

$$\begin{aligned} 0 &\le \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\star \mathbf{A} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\star \begin{bmatrix} \mathbf{A}_{1,1}\mathbf{x} + \mathbf{A}_{1,2}\mathbf{y} \\ \mathbf{A}_{2,1}\mathbf{x} + \mathbf{A}_{2,2}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\star \begin{bmatrix} -\mathbf{A}_{1,2}\mathbf{y} + \mathbf{A}_{1,2}\mathbf{y} \\ -\mathbf{A}_{2,1}\mathbf{A}_{1,1}^{-1}\mathbf{A}_{1,2}\mathbf{y} + \mathbf{A}_{2,2}\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}^\star \begin{bmatrix} \mathbf{0} \\ \mathbf{S}\mathbf{y} \end{bmatrix} = \mathbf{y}^\star\mathbf{S}\mathbf{y}. \end{aligned}$$

Da Gleichheit nur im Falle $\mathbf{x} = \mathbf{0}$ und $\mathbf{y} = \mathbf{0}$ gilt, ist $\mathbf{S}$ ebenfalls positiv definit. $\square$

Im weiteren betrachten wir nur hermitesche Matrizen $\mathbf{A} = \mathbf{A}^\star \in \mathbb{K}^{n\times n}$.

Definition 7.23 Eine Zerlegung $\mathbf{A} = \mathbf{L}\mathbf{L}^\star$ mit unterer Dreiecksmatrix $\mathbf{L}$ mit positiven Diagonaleinträgen heißt **Cholesky-Zerlegung** von $\mathbf{A}$.

Proposition 7.24 Hat $\mathbf{A}$ eine Cholesky-Zerlegung, dann ist $\mathbf{A}$ hermitesch und positiv definit.

Beweis. Aus $\mathbf{A} = \mathbf{L}\mathbf{L}^\star$ folgt

$$\mathbf{A}^\star = \big(\mathbf{L}^\star\big)^\star\mathbf{L}^\star = \mathbf{L}\mathbf{L}^\star = \mathbf{A},$$

das heißt $\mathbf{A}$ ist hermitsch. Wegen

$$\mathbf{x}^\star\mathbf{A}\mathbf{x} = \mathbf{x}^\star\mathbf{L}\mathbf{L}^\star\mathbf{x} = \big(\mathbf{L}^\star\mathbf{x}\big)^\star\mathbf{L}^\star\mathbf{x} = \|\mathbf{L}^\star\mathbf{x}\|_2^2 \geq 0$$

ist $\mathbf{A}$ auch positiv semidefinit. Da $\mathbf{L}$ nach Voraussetzung nichtsingulär ist, impliziert $\mathbf{L}^\star\mathbf{x} = \mathbf{0}$ auch $\mathbf{x} = \mathbf{0}$. Damit ist $\mathbf{A}$ sogar definit. □

Satz 7.25 Ist $\mathbf{A}$ hermitesch und positiv definit, dann existiert eine Cholesky-Zerlegung von $\mathbf{A}$.

Beweis. Induktion über n:

$\underline{n=1}$: Da $\mathbf{A} = [a_{1,1}]$ positiv definit ist gilt $a_{1,1} > 0$. Wegen

$$\mathbf{A} = [a_{1,1}] \stackrel{!}{=} [\ell_{1,1}] \cdot [\overline{\ell_{1,1}}] = \mathbf{L}\cdot\mathbf{L}^\star$$

folgt damit $\ell_{1,1} = \sqrt{a_{1,1}} > 0$.

$\underline{n-1 \mapsto n}$: Betrachte

$$\mathbf{A} = \left[\begin{array}{c|c} a_{1,1} & \mathbf{A}_{1,2} \\ \hline \mathbf{A}_{2,1} & \mathbf{A}_{2,2}\end{array}\right]$$

mit $\mathbf{A}_{2,1} = \mathbf{A}_{1,2}^\star$ und das Schurkomplement

$$\mathbf{S} = \mathbf{A}_{2,2} - \frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} \tag{7.7}$$

von $\mathbf{A}$ bezüglich $a_{1,1}$. Nach Lemma 7.22 ist $a_{1,1} > 0$ und $\mathbf{S}$ hermitesch und positiv definit. Also ist $\ell_{1,1} = \sqrt{a_{1,1}} > 0$ und aufgrund der Induktionsannahme hat $\mathbf{S}$ eine Cholesky-Zerlegung

$$\mathbf{S} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star.$$

Definiere damit

$$\mathbf{L} = \left[\begin{array}{c|c} \ell_{1,1} & \mathbf{0} \\ \hline \frac{1}{\ell_{1,1}}\mathbf{A}_{2,1} & \widetilde{\mathbf{L}}\end{array}\right], \qquad \mathbf{L}^\star = \left[\begin{array}{c|c} \ell_{1,1} & \frac{1}{\ell_{1,1}}\mathbf{A}_{1,2} \\ \hline \mathbf{0} & \widetilde{\mathbf{L}}^\star\end{array}\right].$$

Es ergibt sich

$$\mathbf{L}\mathbf{L}^\star = \left[\begin{array}{c|c} \ell_{1,1} & \mathbf{0} \\ \hline \frac{1}{\ell_{1,1}}\mathbf{A}_{2,1} & \widetilde{\mathbf{L}}\end{array}\right] \cdot \left[\begin{array}{c|c} \ell_{1,1} & \frac{1}{\ell_{1,1}}\mathbf{A}_{1,2} \\ \hline \mathbf{0} & \widetilde{\mathbf{L}}^\star\end{array}\right] = \left[\begin{array}{c|c} a_{1,1} & \mathbf{A}_{1,2} \\ \hline \mathbf{A}_{2,1} & \frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} + \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star\end{array}\right].$$

Wegen

$$\frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} + \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\star = \frac{1}{a_{1,1}}\mathbf{A}_{2,1}\mathbf{A}_{1,2} + \mathbf{S} \overset{(7.7)}{=} \mathbf{A}_{2,2}$$

folgt hieraus

$$\mathbf{L}\mathbf{L}^\star = \left[\begin{array}{c|c} a_{1,1} & \mathbf{A}_{1,2} \\ \hline \mathbf{A}_{2,1} & \mathbf{A}_{2,2} \end{array}\right] = \mathbf{A}.$$

□

Bemerkung: Durch Kombination von Proposition 7.24 und Satz 7.25 erhalten wir die Aussage, dass eine Cholesky-Zerlegung genau dann existiert, falls $\mathbf{A}$ hermitesch und positiv definit ist.

Die Berechnung von $\mathbf{L}$ ergibt sich durch Koeffizientenvergleich: Aus

$$\underbrace{\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}}_{=\mathbf{A}} = \underbrace{\begin{bmatrix} \ell_{1,1} & & & \mathbf{0} \\ \ell_{2,1} & \ell_{2,2} & & \\ \vdots & \vdots & \ddots & \\ \ell_{n,1} & \ell_{n,2} & \cdots & \ell_{n,n} \end{bmatrix}}_{=\mathbf{L}} \cdot \underbrace{\begin{bmatrix} \overline{\ell_{1,1}} & \overline{\ell_{2,1}} & \cdots & \overline{\ell_{n,1}} \\ & \overline{\ell_{2,2}} & & \overline{\ell_{n,2}} \\ & & \ddots & \vdots \\ \mathbf{0} & & & \overline{\ell_{n,n}} \end{bmatrix}}_{=\mathbf{L}^\star}$$

folgt

$$\begin{aligned}
a_{1,1} &= |\ell_{1,1}|^2 & &\rightsquigarrow \ell_{1,1} = \sqrt{a_{1,1}} > 0 \\
a_{2,1} &= \ell_{2,1}\overline{\ell_{1,1}} & &\rightsquigarrow \ell_{2,1} = a_{2,1}/\overline{\ell_{1,1}} \\
a_{3,1} &= \ell_{3,1}\overline{\ell_{1,1}} & &\rightsquigarrow \ell_{3,1} = a_{3,1}/\overline{\ell_{1,1}} \\
&\vdots & &\vdots \\
a_{n1} &= \ell_{n1}\overline{\ell_{1,1}} & &\rightsquigarrow \ell_{n1} = a_{n1}/\overline{\ell_{1,1}} \\
a_{2,2} &= |\ell_{2,1}|^2 + |\ell_{2,2}|^2 & &\rightsquigarrow \ell_{2,2} = \sqrt{a_{2,2} - |\ell_{2,1}|^2} > 0 \\
a_{3,2} &= \ell_{3,1}\overline{\ell_{2,1}} + \ell_{3,2}\overline{\ell_{2,2}} & &\rightsquigarrow \ell_{3,2} = (a_{3,2} - \ell_{3,1}\overline{\ell_{2,1}})/\overline{\ell_{2,2}} \\
&\vdots & &\vdots \\
a_{n,2} &= \ell_{n,1}\overline{\ell_{2,1}} + \ell_{n,2}\overline{\ell_{2,2}} & &\rightsquigarrow \ell_{n,2} = (a_{n,2} - \ell_{n,1}\overline{\ell_{2,1}})/\overline{\ell_{2,2}}
\end{aligned}$$

und allgemein

$$\begin{aligned}
\ell_{j,j} &= \sqrt{a_{j,j} - \sum_{k=1}^{j-1} |\ell_{j,k}|^2} > 0, & & 1 \le j \le n, \\
\ell_{i,j} &= \frac{1}{\overline{\ell_{j,j}}}\left(a_{i,j} - \sum_{k=1}^{j-1} \ell_{i,k}\overline{\ell_{j,k}}\right), & & j < i \le n.
\end{aligned} \tag{7.8}$$

Die Berechenbarkeit ist durch den Existenzbeweis (Satz 7.25) gewährleistet (alle $\ell_{i,i} \neq 0$). Aus (7.8) ergibt sich sofort die folgende Aussage:

Korollar 7.26 Die Cholesky-Zerlegung von $\mathbf{A} = \mathbf{A}^\star$ positiv definit ist eindeutig.

Aufwand: Zur Berechnung von $\ell_{i,j}$ $(j \leq i \leq n)$ sind j Multiplikationen bzw. Wurzeln auszuführen. Damit werden insgesamt

$$\sum_{j=1}^{n} (n-j+1)j = \frac{n^2(n+1)}{2} - \frac{n(n+1)(2n+1)}{6} + \mathcal{O}(n^2) = \frac{1}{6}n^3 + \mathcal{O}(n^2)$$

Multiplikationen bzw. Wurzeln benötigt. Der Aufwand ist demnach nur halb so groß wie für die LR-Zerlegung.

Beispiel 7.27 Für

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 10 \end{bmatrix}$$

ergibt sich wegen

$$\begin{aligned} \ell_{1,1} &= \sqrt{1} = 1 & \qquad \ell_{3,1} &= 1/1 = 1 \\ \ell_{2,1} &= 2/1 = 2 & \ell_{3,2} &= (2-2)/1 = 0 \\ \ell_{2,2} &= \sqrt{5-4} = 1 & \ell_{3,3} &= \sqrt{10-1} = 3 \end{aligned}$$

die Cholesky-Zerlegung $\mathbf{L}\mathbf{L}^\star$ mit

$$\mathbf{L} = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 1 & 0 & 3 \end{bmatrix}, \qquad \mathbf{L}^\star = \begin{bmatrix} 1 & 2 & 1 \\ & 1 & 0 \\ & & 3 \end{bmatrix}.$$

△

Bemerkung: Im Gegensatz zur LR-Zerlegung ist die Cholesky-Zerlegung immer stabil.

8. Wahrscheinlichkeitsräume

8.1 Zufällige Ereignisse

Wir werden zunächst eine umgangssprachliche Definition des Wahrscheinlichkeitsraumes angeben. Am Ende des Kapitels wird dann eine mathematische Definition folgen.

Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$**:** Zusammenfassung aller Teile eines mathematischen Modells zur Beschreibung einer Zufallssituation. Verschiedene Zufallssituationen führen auf verschiedene Wahrscheinlichkeitsräume.

Zufallssituation: Eine Zufallssituation ist gekennzeichnet durch zwei Eigenschaften:

- sie ist beliebig oft wiederholbar (zumindest gedanklich),
- ihr Ergebnis ist absolut nicht vorhersagbar.

Versuch: Ein Versuch ist eine Realisierung einer Zufallssituation, wobei wir mit

- ω das Ergebnis des Versuchs und
- Ω die Menge aller möglichen Ergebnisse bezeichnen.

Annahme: Jedes Ergebnis eines Versuchs ist eindeutig einem Element ω der Ergebnismenge Ω zuzuordnen.

Beispiele 8.1 (Zufallssituationen und Ergebnismengen)

1. Beim Werfen eines Würfels ist $\Omega = \{1, 2, 3, 4, 5, 6\}$ eine endliche Ergebnismenge mit 6 möglichen verschiedenen Ergebnissen.
2. Wir betrachten die Lebensdauer einer Glühlampe. Die Ergebnismenge $\Omega = \{\omega \in \mathbb{R} : \omega \geq 0\}$ ist überabzählbar unendlich, wobei das Ergebnis $\omega \in \Omega$ der Lebensdauer der Glühlampe entspricht.
3. Bei der Überprüfung von n verschiedenen Geräten sei das Ergebnis ω_i definiert gemäß
$$\omega_i = \begin{cases} 0, & \text{i-tes Gerät defekt,} \\ 1, & \text{i-tes Gerät in Ordnung.} \end{cases}$$
Dann ist $\Omega = \big\{(\omega_1, \ldots, \omega_n) \in \{0, 1\}^n\big\}$ eine endliche Ergebnismenge mit 2^n Elementen.

$\triangle$

Definition 8.2 Ein **zufälliges Ereignis** ist eine Teilmenge $A \subseteq \Omega$. Wenn $\omega \in A$ gilt, so sagt man, dass das Ereignis A **eingetreten ist**. Nicht jede Teilmenge $A \subseteq \Omega$ muss sich als zufälliges Ereignis betrachten lassen, aber alle zufälligen Ereignisse sind Teilmengen von Ω.

Beispiele 8.3 (Zufällige Ereignisse (Fortsetzung der Beispiele 8.1))

1. Beim Werfen eines idealen Würfels entspricht das Ereignis

$$A \mathrel{\widehat{=}} \text{“es wird eine gerade Zahl gewürfelt”}$$

der Menge $A = \{2, 4, 6\}$.

2. Für eine Glühlampe wird das Ereignis

$$A \mathrel{\widehat{=}} \text{“Brenndauer liegt zwischen 500 und 5000 Stunden”}$$

beschrieben durch

$$A = \{\omega \in \mathbb{R} : 500 \leq \omega \leq 5000\} \subseteq \Omega = \mathbb{R}_{\geq 0}.$$

3. Bei der Überprüfung von n Geräten gilt für das Ereignis

$$A \mathrel{\widehat{=}} \text{“es funktionieren mindestens 2 Geräte”}$$

die Beziehung

$$A = \left\{(\omega_1, \dots, \omega_n) \in \{0,1\}^n : \sum_{i=1}^{n} \omega_i \geq 2\right\}.$$

△

8.2 Rechnen mit zufälligen Ereignissen

Bezeichnungen:

- “*A oder B*”: Das Ereignis tritt ein, wenn entweder A oder B oder beide Ereignisse A und B eintreten, kurz
$$\omega \in A \cup B.$$
- “*A und B*”: Das Ereignis tritt ein, wenn A und B gleichzeitig eintreten, kurz
$$\omega \in A \cap B.$$
- “*nicht A bzw. $\overline{A}$*”: Das Ereignis tritt ein, wenn A nicht eintritt, kurz
$$\omega \notin A \Leftrightarrow \omega \in \Omega \setminus A =: \overline{A}.$$
Hierbei nennen wir $\overline{A}$ das *Komplementärereignis* von A.
- “*sicheres Ereignis*”: A heißt das sichere Ereignis, falls gilt
$$A = \Omega.$$

- "*unmögliches Ereignis* ": A heißt das unmögliche Ereignis, falls gilt
$$A = \overline{\Omega} = \emptyset.$$
- "*Elementarereignis* ": Ein Elementarereignis ist eine einelementige Menge mit dem Ergebnis $\omega \in \Omega$, kurz
$$A = \{\omega\}.$$
- "*unvereinbares Ereignis* ": A und B heißen unvereinbar, wenn gilt
$$A \cap B = \emptyset.$$

Bemerkung: Man kann die "und"- bzw. "oder"-Operationen auch auf endlich viele oder abzählbar unendlich viele Ereignisse A_i im Sinne von $\bigcup_{i=1}^n A_i$, $\bigcap_{i=1}^n A_i$ bzw. $\bigcup_{i=1}^\infty A_i$, $\bigcap_{i=1}^\infty A_i$ anwenden.

Definition 8.4 Die Menge $\mathcal{A}$ von Teilmengen der Ergebnismenge Ω, welche die zufälligen Ereignisse beschreiben, heißt **σ-Algebra** bzw. **Ereignisalgebra** bezogen auf eine feste Zufallssituation, wenn gilt:

1. $\Omega \in \mathcal{A}$,
2. $A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A}$,
3. $A_i \in \mathcal{A}\, \forall i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{A}$.

Satz 8.5 (Rechenregeln für zufällige Ereignisse) Es gelten:

1. Kommutativgesetze:
$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$
2. Assoziativgesetze:
$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$
3. Distributivgesetze:
$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$
4. De Morgansche Regeln:
$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$
$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$
5. Verallgemeinerte De Morgansche Regeln für beliebige Indexmengen $\mathcal{I}$:
$$\overline{\bigcup_{i \in \mathcal{I}} A_i} = \bigcap_{i \in \mathcal{I}} \overline{A_i}$$
$$\overline{\bigcap_{i \in \mathcal{I}} A_i} = \bigcup_{i \in \mathcal{I}} \overline{A_i}$$

6.

$$\begin{aligned} A \cup \emptyset &= A, & A \cup \Omega &= \Omega \\ A \cap \emptyset &= \emptyset, & A \cap \Omega &= A \end{aligned}$$

Bemerkung: $A \subseteq B$ ist die Schreibweise für “A *zieht* B *nach sich*”, das heißt

$$\omega \in A \Rightarrow \omega \in B.$$

Äquivalent hierzu sind auch folgenden Varianten:

$$\begin{aligned} A \subseteq B &\Leftrightarrow A \cap B = A \\ &\Leftrightarrow A \cup B = B \\ &\Leftrightarrow \overline{B} \subseteq \overline{A} \\ &\Leftrightarrow \overline{A} \cap \overline{B} = \overline{B} \\ &\Leftrightarrow \overline{A} \cup \overline{B} = \overline{A} \\ &\Leftrightarrow A \cap \overline{B} = \emptyset. \end{aligned}$$

Bemerkung: Wie man leicht zeigt, gelten für eine σ-Algebra $\mathcal{A}$ zusätzlich die Aussagen

1. $\emptyset \in \mathcal{A}$,
2. $A_i \in \mathcal{A} \, \forall i \in \mathbb{N} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$,
3. $A, B \in \mathcal{A} \Rightarrow A \setminus B \in \mathcal{A}$.

Zusammen mit den Eigenschaften 1–3 aus Definition 8.4 ist damit sichergestellt, dass jedes mögliche Ergebnis von endlich vielen oder abzählbar unendlich vielen Mengenoperationen wieder in der σ-Algebra liegt.

8.3 Rechnen mit Wahrscheinlichkeiten

$A \in \mathcal{A}$ sei ein festes Ereignis innerhalb einer Zufallssituation. Wir betrachten n unabhängige, das heißt sich gegenseitig nicht beeinflussende, Versuche. Mit $h_n(A)$ bezeichnen wir die *absolute Häufigkeit* des Ereignisses A, dies ist die Anzahl des Eintretens von A bei n Versuchen.

Definition 8.6 Wir bezeichnen mit

$$H_n(A) = \frac{h_n(A)}{n}$$

die **relative Häufigkeit** für das Eintreten von A bei n Versuchen.

Die Erfahrung lehrt, dass für $n \to \infty$ die relative Häufigkeit gegen eine feste Zahl $P(A)$ strebt. Dies nennt man den Stabilisierungseffekt der Folge $H_n(A)$ der relativen Häufigkeiten.

Eigenschaften der relativen Häufigkeit:

1. *Positivität*: Es ist $0 \leq H_n(A)$ für alle A.
2. *Normierung*: $H_n(\Omega) = 1$.
3. *Additivität*: A, B seien unvereinbar, das heißt $A \cap B = \emptyset$, dann gilt

$$H_n(A \cup B) = \frac{h_n(A) + h_n(B)}{n} = \frac{h_n(A)}{n} + \frac{h_n(B)}{n} = H_n(A) + H_n(B).$$

Dies bedeutet, dass sich für unvereinbare Ereignisse die relativen Häufigkeiten addieren.

Diese Rechenregeln bilden die Grundlage des Kolmogorovschen Axiomsystems zum Rechnen mit Wahrscheinlichkeiten.

Definition 8.7 (Kolmogorovsches Axiomsystem, 1933) Gegeben sei eine Ergebnismenge Ω und ein Ereignisfeld $\mathcal{A}$, welche eine Zufallssituation beschreiben. Dann ist jedem Ereignis $A \in \mathcal{A}$ eine reelle Zahl $P(A)$ zugeordnet, die **Wahrscheinlichkeit** des Ereignisses A. Dabei gelten die folgenden Axiome:

A1 *Positivität*: $0 \leq P(A)$ für alle $A \in \mathcal{A}$.

A2 *Normierung*: $P(\Omega) = 1$.

A3 *σ-Additivität*: Für eine Folge von Ereignissen $A_i \in \mathcal{A}$, welche paarweise unvereinbar sind, das heißt $A_i \cap A_j = \emptyset$ für $i \neq j$, ist

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Folgerungen aus den Axiomen (Rechenregeln):

1. $P(\emptyset) = 0$

   Beweis. Die Ereignisse $A_1 := \Omega$ und $A_i := \emptyset$, $i > 1$, sind paarweise unvereinbar wegen $A_i \cap A_j = \emptyset$, $i \neq j$. Aus den Axiomen A1 und A3 folgt, dass

$$1 = P(\Omega) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = 1 + \sum_{i=1}^{\infty} P(\emptyset).$$

   Hieraus ergibt sich die Behauptung $P(\emptyset) = 0$. □

2. Für jede endliche Folge von paarweise unvereinbaren Ereignissen $A_i \in \mathcal{A}$ gilt

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

   Beweis. Setzen wir $A_i := \emptyset$ für alle $i > n$, so folgt die Behauptung sofort aus Axiom A3. □

3. $P(\overline{A}) = 1 - P(A)$

Beweis. Wegen $A \cup \overline{A} = \Omega$ und $A \cap \overline{A} = \emptyset$ folgt nach den Axiomen A2 und A3

$$1 = P(\Omega) = P(A) + P(\overline{A}).$$

□

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   Beweis. Wir setzen $C := B \cap \overline{A} \subseteq B$, dann folgt $A \cup B = A \cup C$ und $A \cap C = \emptyset$, und daher mit Axiom A3

   $$P(A \cup B) = P(A) + P(C). \tag{8.1}$$

   Weiter gilt für $D := A \cap B$, dass $B = C \cup D$ und $C \cap D = \emptyset$, und folglich

   $$P(B) = P(C) + P(D) = P(C) + P(A \cap B) \tag{8.2}$$

   gemäß Axiom A3. Aus (8.1) und (8.2) ergibt sich dann die Behauptung. □

5. Bilden die Ereignisse $A_i \in \mathcal{A}$ eine *Zerlegung von* Ω, dies bedeutet die A_i sind jeweils paarweise unvereinbare Ereignisse und $\Omega = \bigcup_{i=1}^{\infty} A_i$, dann gilt

   $$\sum_{i=1}^{\infty} P(A_i) = 1.$$

   Beweis. Aus den Axiomen A2 und A3 folgt

   $$1 = P(\Omega) = P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

   □

Den *klassischen* Wahrscheinlichkeitsbegriff werden wir dann zur Berechnung von Wahrscheinlichkeiten heranziehen, wenn ein Versuch nur endlich viele gleichmögliche Elementarereignisse besitzt. Wir sprechen hier von *Laplace-Modellen*:

Satz 8.8 (Laplace-Experiment) Die Ergebnismenge Ω erfülle die folgenden beiden Voraussetzungen:

a) $\Omega = \{\omega_1, \ldots, \omega_n\}$ besteht aus N Elementarereignissen,

b) alle Elementarereignisse sind gleich wahrscheinlich, dies bedeutet

$$P(\{\omega_1\}) = \ldots = P(\{\omega_n\}) = \frac{1}{N}.$$

Dann gilt für ein Ereignis $A = \{\omega_{i_1}, \ldots, \omega_{i_m}\} \in \mathcal{P}(\Omega)$

$$P(A) = \frac{M}{N}.$$

Beweis. Die Behauptung folgt sofort aus den Kolmogorovschen Axiomen. □

Bemerkung: Die Aussage von Satz 8.8 kann auch kurz dargestellt werden durch

$$P(A) = \frac{\text{Anzahl der für A günstigen Ereignisse}}{\text{Anzahl aller möglichen Ereignisse}}.$$

Beispiel 8.9 (Werfen eines idealem Würfel) Für die Ergebnismenge beim einmaligen Werfen eines idealen Würfels gilt $\Omega = \{1, 2, 3, 4, 5, 6\}$. Das Ereignis

$$A \mathrel{\widehat{=}} \text{“Primzahl wird gewürfelt”}$$

ist durch die Menge $A = \{2, 3, 5\}$ gegeben. Damit gilt also

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}.$$

△

Zusammenfassung: Eine Zufallssituation wird durch folgende drei Komponenten des mathematischen Modells vollständig charakterisiert:

- *Ergebnismenge* Ω: nichtleere Menge, deren Elemente die möglichen Versuchsausgänge darstellen.
- *σ-Algebra* $\mathcal{A}$: Menge der Teilmengen von Ω, die zufällige Ereignisse bilden und für die die Eigenschaften 1–3 aus Definition 8.4 gelten.
- *Wahscheinlichkeitsmaß* P: jedem Ereignis $A \in \mathcal{A}$ ist in eindeutiger Weise eine Zahl $P(A)$ zugeordnet, die Wahrschscheinlichkeit genannt wird.

Definition 8.10 Das Tripel $(\Omega, \mathcal{A}, P)$ heißt **Wahrscheinlichkeitsraum** der gegebenen Zufallssituation.

8.4 Grundformeln der Kombinatorik

Wir betrachten eine Urne mit n unterscheidbaren (zum Beipiel durch Nummerierung) Kugeln. Wir wollen die Anzahl der Möglichkeiten beim Ziehen von m Kugeln bestimmen. Dabei müssen wir berücksichtigen, ob eine entnommene Kugel vor der Entnahme der nächsten Kugel wieder zurückgelegt wird und ob die Reihenfolge der Kugeln eine Rolle spielt.

1. Reihenfolge der Entnahme wichtig $\longrightarrow$ *Variationen*
   - Anzahl der Möglichkeiten mit Zurücklegen: $\underbrace{n \cdot n \cdots n}_{m\text{-mal}} = n^m$
   - Anzahl der Möglichkeiten ohne Zurücklegen: $n\cdot(n-1)\cdots(n-m+1) = \dfrac{n!}{(n-m)!}$
2. Reihenfolge der Entnahme spielt keine Rolle $\longrightarrow$ *Kombinationen*
   - Anzahl der Möglichkeiten ohne Zurücklegen: $\dbinom{n}{m} = \dfrac{n!}{(n-m)!m!}$

Beweis. Mit C_n^m bezeichnen wir die Anzahl von Möglichkeiten beim Ziehen von m aus n Kugeln ohne Beachtung der Reihenfolge und ohne Zurücklegen. Es gilt $C_1^0 = 1$ und $C_n^n = 1$, ferner setzen wir $C_n^m := 0$ im Falle $m < 0$. Wir erhalten

$$\begin{aligned} C_{n+1}^m &= |\{ \underbrace{(a_1, \ldots, a_m)}_{a_i \widehat{=} \text{ Nummer der } i\text{-ten Kugel}} : 1 \leq a_1 < a_2 < \ldots < a_m \leq n+1\}| \\ &= |\underbrace{\{(a_1, \ldots, a_m) : 1 \leq a_1 < a_2 < \ldots < a_m \leq n\}}_{=C_n^m}| \\ &+ |\underbrace{\{(a_1, \ldots, a_{m-1}, n+1) : 1 \leq a_1 < \ldots < a_{m-1} \leq n\}}_{=C_n^{m-1}}| \\ &= C_n^{m-1} + C_n^m. \end{aligned}$$

Wie man leicht mit vollständiger Induktion zeigt entspricht dies der Rekursion

$$\binom{n}{m-1} + \binom{n}{m} = \binom{n+1}{m},$$

dies bedeutet $C_m^n = \binom{n}{m}$. □

- Anzahl der Möglichkeiten mit Zurücklegen: $\displaystyle\binom{n+m-1}{m}$

  Beweis. Die Menge aller $\{(a_1, \ldots, a_m) : 1 \leq a_1 \leq \ldots \leq a_m \leq n\}$ wird durch $b_i := a_i + i - 1$ bijektiv auf die Menge aller Tupel

  $$\{(b_1, \ldots, b_m) : 1 \leq b_1 < b_2 < \ldots < b_m \leq n+m-1\}$$

  abgebildet. Letztere Menge hat aber die Mächtigkeit $\binom{n+m-1}{m}$. □

Beispiel 8.11 (Geburtstagsparadoxon) Wir wollen die Wahrscheinlichkeit dafür bestimmen, dass zwei Personen in einem Raum am gleichen Tag Geburtstag haben. Dabei setzen wir die folgenden Annahmen voraus:

- n Personen sind im Raum,
- keiner hat am 29.2. Geburtstag,
- die übrigen 365 Tage seien als Geburtstag gleichwahrscheinlich.

Wir interessieren uns für das Ereignis

$$A \widehat{=} \text{“mindestens 2 Personen haben am gleichen Tag Geburtstag”}$$

bzw. für die Wahrscheinlichkeit $P(A)$. A tritt also ein, wenn $2, 3, 4, \ldots$ Personen am selben Tag Geburtstag haben.

Um die Anzahl dieser Möglichkeiten einzugrenzen, bietet es sich an, mit dem Komplementärereignis zu arbeiten, dies ist

$$\begin{aligned} \overline{A} &\widehat{=} \text{“keiner hat am gleichen Tag Geburtstag”} \\ &\widehat{=} \text{“alle Geburtstage sind verschieden”}. \end{aligned}$$

Die dritte Annahme sichert uns zu, dass ein Laplace-Experiment vorliegt, womit sich

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{\text{Anzahl der günstigen Ereignisse}}{\text{Anzahl aller Ereignisse}}$$

ergibt. Als nächstes stellt sich die Frage nach dem zu wählenden Urnenmodell. Das Modell "Ziehen mit Zurücklegen unter Beachtung der Reihenfolge" ist das passende, um den Ergebnisraum Ω zu charakterisieren, während das Modell "Ziehen ohne Zurücklegen unter Beachtung der Reihenfolge" auf die für $\overline{A}$ günstigen Ereignisse zutrifft. Es folgt daher

$$P(A) = 1 - \frac{\overbrace{365 \cdot 364 \cdot 363 \cdots (365 - n + 1)}^{n \text{ Faktoren}}}{365^n}.$$

Für $n = 23$ Personen folgt beispielsweise $P(A) = 0.507$, während sich für $n = 70$ Personen schon $P(A) = 0.999$ ergibt. △

Beispiel 8.12 (Lotto (6 aus 49)) Beim Lotto werden 6 Kugeln aus einer Urne mit 49 Kugeln ohne Zurücklegen und ohne Beachtung der Reihenfolge gezogen. Für das Ereignis

$$A \mathrel{\widehat{=}} \text{“}\textit{genau } k \text{ Richtige getippt”}$$

gilt

$$P(A) = \frac{\text{Anzahl der günstigen Ereignisse}}{\text{Anzahl aller Ereignisse}} = \frac{\overbrace{\binom{6}{k}}^{\text{Richtige}} \overbrace{\binom{43}{6-k}}^{\text{Nieten}}}{\binom{49}{6}}.$$

△

9. Bedingte Wahrscheinlichkeiten und Unabhängigkeit

9.1 Definition der bedingten Wahrscheinlichkeit

Gegeben sei ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ zu einer Zufallssituation und ein festes Ereignis $A \in \mathcal{A}$ mit der Wahrscheinlichkeit $P(A)$. Es kann sein, dass die Wahrscheinlichkeit sich verändert, wenn man beachtet, dass ein anderes Ereignis $B \in \mathcal{A}$ bereits eingetreten ist.

Definition 9.1 Es seien A und B mit $P(B) > 0$ zufällige Ereignisse, die zu einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ gehören. Dann heißt die Größe

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{9.1}$$

bedingte Wahrscheinlichkeit des Ereignisses A unter der Bedingung, dass B bereits eingetreten ist.

Motivation über relative Häufigkeiten: Es werden n Versuche durchgeführt, wobei $h_n(A)$, $h_n(B)$ und $h_n(A \cap B)$ jeweils die Anzahl der Versuche angeben, bei denen A, B bzw. beide Ereignisse eintreten. Die relative Häufigkeit für A unter der Bedingung B, kurz $A|B$, ergibt sich dann zu

$$H_n(A|B) = \frac{h_n(A \cap B)}{h_n(B)} = \frac{\frac{h_n(A \cap B)}{n}}{\frac{h_n(B)}{n}} = \frac{H_n(A \cap B)}{H_n(B)}.$$

Beispiel 9.2 (Würfelproblem) Mit einem idealen Würfel werden zwei Würfe ausgeführt. Dabei bezeichnen A und B die folgenden Ereignisse

$$A \mathrel{\widehat{=}} \text{“Ereignis, dass beim ersten Wurf eine 6 gewürfelt wird”},$$
$$B \mathrel{\widehat{=}} \text{“Ereignis, dass die Augensumme beider Würfel 8 ist”}.$$

Die Augensumme zweier Würfel kann durch folgendes Tableau veranschaulicht werden:

	1	2	3	4	5	6	(2. Wurf)
1	2	3	4	5	6	7	
2	3	4	5	6	7	8	
3	4	5	6	7	8	9	
4	5	6	7	8	9	10	
5	6	7	8	9	10	11	
6	7	8	9	10	11	12	
(1. Wurf)							

Wir sehen, dass fünf der insgesamt 36 Elementarereignisse günstig für B sind, das heißt, es gilt $P(B) = \frac{5}{36}$. Weiter ist $P(A) = \frac{1}{6}$. Nur im Fall, dass erst eine 6 und dann eine 2 gewürfelt wird, treten A und B ein, dies bedeutet $P(A \cap B) = \frac{1}{36}$ und $P(A|B) = \frac{1}{5}$. Die Gleichung

$$P(A|B) = \frac{\frac{1}{36}}{\frac{5}{36}} = \frac{1}{5}$$

bestätigt Definition 9.1. △

Rechenregeln für bedingte Wahrscheinlichkeiten:

1. $P(C|C) = 1$.

   Beweis. Es gilt
   $$P(C|C) = \frac{P(C \cap C)}{P(C)} = \frac{P(C)}{P(C)} = 1.$$
   □

2. $P(A|C) = 1 - P(\overline{A}|C)$

   Beweis. Wegen $A \cup \overline{A} = \Omega$ folgt $(A \cap C) \cup (\overline{A} \cap C) = C$ und weiter
   $$\begin{aligned} 1 &= \frac{P(C)}{P(C)} = \frac{P((A \cap C) \cup (\overline{A} \cap C))}{P(C)} = \frac{P(A \cap C) + P(\overline{A} \cap C)}{P(C)} \\ &= P(A|C) + P(\overline{A}|C) \end{aligned}$$
   □

3. $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$

   Beweis.
   $$\begin{aligned} P(A \cup B|C) &= \frac{P(\overbrace{(A \cup B) \cap C}^{=(A \cap C) \cup (B \cap C)})}{P(C)} \\ &= \frac{P(A \cap C)}{P(C)} + \frac{P(B \cap C)}{P(C)} - \frac{P(\overbrace{(A \cap C) \cap (B \cap C)}^{=(A \cap B) \cap C})}{P(C)} \\ &= P(A|C) + P(B|C) - P(A \cap B|C). \end{aligned}$$
   □

Bemerkung: Die Rechenregeln für die bedingte Wahrscheinlichkeit $P(A|C)$ sind wie für $P(A)$, das heißt, im ersten Argument darf wie bisher umgeformt werden. Nichts an der Bedingung umformen, dies geht i.a. schief!

Beispiel 9.3 (Stapelsuchproblem) Thorsten durchsucht 7 gleichgroße Stapel von CDs nach einer ganz bestimmten. Die Wahrscheinlichkeit, dass die CD überhaupt in einem Stapel vorhanden ist, sei 0.8. Es wurden bereits 6 Stapel erfolglos durchsucht. Wie groß ist die Wahrscheinlichkeit, die CD im 7. Stapel zu finden?

Wir legen zunächst die Ereignisse fest:

$$A_i \mathrel{\widehat{=}} \text{“CD ist im } i\text{-ten Stapel”},$$

wobei $P(A_1) = P(A_2) = \cdots = P(A_7) = p$ gilt.

Aus

$$P(A_1 \cup A_2 \cup \ldots \cup A_7) = P(A_1) + P(A_2) + \ldots + P(A_7) = 7p \stackrel{!}{=} 0.8$$

folgt $p = \frac{0.8}{7} = 0.114$ und daher

$$\begin{aligned} P(A_7|\overline{A_1 \cup A_2 \cup \ldots \cup A_6}) &= P(A_7|\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_6) = \frac{P(\overbrace{A_7 \cap \overline{A}_1 \cap \ldots \cap \overline{A}_6}^{=A_7})}{P(\overline{A}_1 \cap \ldots \cap \overline{A}_6)} \\ &= \frac{P(A_7)}{1 - P(A_1 \cup \ldots \cup A_6)} = \frac{p}{1-6p} = 0.3636. \end{aligned}$$

△

9.2 Multiplikationsregeln

Durch Umstellen der die bedingte Wahrscheinlichkeit $P(A|B)$ definierenden Quotientenbeziehung (9.1) erhält man eine einfache Multiplikationsregel:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A). \tag{9.2}$$

Beispiel 9.4 70% der Studenten eines Jahrgangs schließen das Fach Mathematik wenigstens mit der Note 3 ab. Unter diesen Studenten erreichen 25% sogar eine der Noten 1 oder 2. Mit welcher Wahrscheinlichkeit schließt ein beliebig ausgewählter Student das Fach mit 1 oder 2 ab?

Um diese Frage zu beantworten, sei

$$\begin{aligned} A &\mathrel{\widehat{=}} \text{“Student schließt das Fach Mathematik mit 1 oder 2 ab”}, \\ B &\mathrel{\widehat{=}} \text{“Student schließt das Fach Mathematik mit 1,2 oder 3 ab”}. \end{aligned}$$

Gemäß (9.2) folgt dann

$$P(A) = P(A \cap B) = \underbrace{P(A|B)}_{=0.25} \cdot \underbrace{P(B)}_{=0.7} = 0.175.$$

△

Satz 9.5 (Erweiterte Multiplikationsregel) Es seien $A_1, A_2, \ldots, A_n$ Ereignisse aus dem Ereignisfeld $\mathcal{A}$ zu einer festen Zufallssituation, wobei $P(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$ gelte. Dann ist

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \ldots \cap A_{n-1}).$$

Beweis. Die Behauptung folgt durch vollständige Induktion aus (9.2), weil

$$\begin{aligned} P(A_1 \cap \ldots \cap A_n) &= P(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot P(A_1 \cap \ldots \cap A_{n-1}), \\ P(A_1 \cap \ldots \cap A_{n-1}) &= P(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}) \cdot P(A_1 \cap \ldots \cap A_{n-2}), \\ &\vdots \\ P(A_1 \cap A_2) &= P(A_2|A_1) \cdot P(A_1). \end{aligned}$$

□

Satz 9.6 (Satz von der totalen Wahrscheinlichkeit) Es seien $B_1, B_2, \ldots$ eine Zerlegung von Ω, das heißt

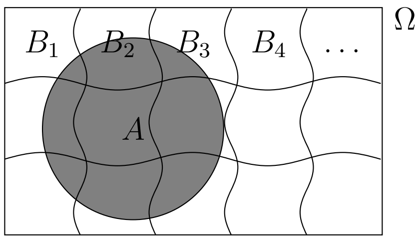
1. $\bigcup_{i=1}^{\infty} B_i = \Omega$,
2. $B_i \cap B_j = \emptyset$ für $i \neq j$.

Dann gilt für ein beliebiges Ereignis A die Formel

$$P(A) = P(A|B_1) \cdot P(B) + P(A|B_2) \cdot P(B_2) + \ldots = \sum_{i=1}^{\infty} P(A|B_i) \cdot P(B_i).$$

Beweis. Da die B_i eine Zerlegung von Ω bilden, gilt

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^{\infty} B_i \right) = \bigcup_{i=1}^{\infty} (A \cap B_i).$$



Wegen der paarweisen Unvereinbarkeit von $A \cap B_i$ für alle $i = 1, 2, \ldots$ folgt hieraus aber

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) \overset{(9.2)}{=} \sum_{i=1}^{\infty} P(A|B_i) \cdot P(B_i).$$

□

Satz 9.7 (Bayes) Unter den Voraussetzungen von Satz 9.6 gilt für $P(A) > 0$ die **Bayessche Formel**

$$P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{P(A)} = \frac{P(A|B_j) \cdot P(B_j)}{\sum_{i=1}^{\infty} P(A|B_i) \cdot P(B_i)}, \qquad j = 1, 2, \ldots .$$

Beweis. Aus (9.2) folgt

$$P(A \cap B_j) = P(A|B_j) \cdot P(B_j) = P(B_j|A) \cdot P(A),$$

dies bedeutet

$$P(B_j|A) = \frac{P(A|B_j) \cdot P(B_j)}{P(A)}.$$

Mit dem Satz von der totalen Wahrscheinlichkeit folgt dann auch die letzte Gleichung. □

Beispiel 9.8 (Pannenhilfe-Problem) Die Statistik der ambulanten Pannenhelfer des Automobilclubs ADAC wies für ein Jahr aus, dass bei vorgefundenen Schäden im Bereich der Motorausfälle folgende Schadenstypverteilung zu verzeichnen war:

50% Störungen an der Zündanlage,
30% Störungen an der Kraftstoffzufuhr,
20% andere Störungen.

Der Pannenhelfer konnte vor Ort den Schaden beheben

in 50% aller Fälle bei Störungen der Zündanlage,
in 30% aller Fälle bei Störungen der Kraftstoffzufuhr,
in 5% aller Fälle bei sonstigen Störungen.

Wir wollen zwei Fragen beantworten:

1. In wieviel Prozent der Fälle konnte bei Motorausfällen vor Ort geholfen werden?
2. Wie wahrscheinlich sind die drei Schadenstypen, wenn geholfen werden konnte?

zu 1) Die Ereignisse

$$\begin{aligned} B_1 &\mathrel{\widehat{=}} \text{“Fehler an der Zündanlage”}, \\ B_2 &\mathrel{\widehat{=}} \text{“Fehler and der Kraftstoffzufuhr”}, \\ B_3 &\mathrel{\widehat{=}} \text{“sonstiger Fehler”} \end{aligned}$$

bilden eine Zerlegung von Ω, wobei

$$P(B_1) = 0.5, \quad P(B_2) = 0.3, \quad P(B_3) = 0.2$$

gilt. Damit erhalten wir für das Ereignis

$$A \mathrel{\widehat{=}} \text{“Schaden konnte vor Ort behoben werden”}$$

gemäß dem Satz von der totalen Wahrscheinlichkeit

$$P(A) = \underbrace{P(A|B_1)}_{=0.5} \cdot P(B_1) + \underbrace{P(A|B_2)}_{=0.3} \cdot P(B_2) + \underbrace{P(A|B_3)}_{=0.05} \cdot P(B_3) = 0.35.$$

Es konnte also in 35% aller Fälle vor Ort geholfen werden.

zu 2) Mit der Bayesschen Formel erhalten wir

$$P(B_1|A) = \frac{P(A|B_1) \cdot P(B_1)}{P(A)} = 0.714,$$
$$P(B_2|A) = \frac{P(A|B_2) \cdot P(B_2)}{P(A)} = 0.257,$$
$$P(B_3|A) = \frac{P(A|B_3) \cdot P(B_3)}{P(A)} = 0.029.$$

△

9.3 Stochastische Unabhängigkeit

Definition 9.9 Zwei Ereignisse $A, B \in \mathcal{A}$ heißen **stochastisch unabhängig**, wenn gilt

$$P(A \cap B) = P(A) \cdot P(B).$$

Bemerkung: Stochastisch unabhängig bedeutet $P(A|B) = P(A)$ und $P(B|A) = P(B)$, das heißt, der Zufallscharakter der Ereignisse A und B beeinflusst sich nicht gegenseitig.

Satz 9.10 Falls A und B stochastisch unabhängig sind, so sind auch die Ereignisse $\overline{A}$ und $\overline{B}$, A und $\overline{B}$, sowie $\overline{A}$ und B stochastisch unabhängig.

Beweis. Wegen

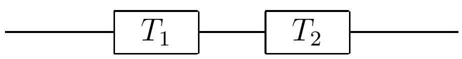
$$\begin{aligned} P(A \cap \overline{B}) &= P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) \\ &= P(A) \cdot \big(1 - P(B)\big) = P(A) \cdot P(\overline{B}) \end{aligned}$$

sind mit A und B auch A und $\overline{B}$ stochastisch unabhängig. Analog zeigt man die verbliebenen Fälle. □

Wichtige Anwendung: Auf der stochastischen Unabhängigkeit beruhen die meisten Aussagen der Zuverlässigkeitstheorie.

Beispiel 9.11 (Serien- und Parallelschaltung von Bauteilen) Ein Gerät bestehe aus 2 Bauteilen T_1 und T_2, bei denen unabhängig voneinander Defekte auftreten können. Die stochastisch unabhängigen Ereignisse A_1 mit $P(A_1) = p_1$ und A_2 mit $P(A_2) = p_2$ treten auf, wenn die Bauteile T_1 und T_2 funktionieren.

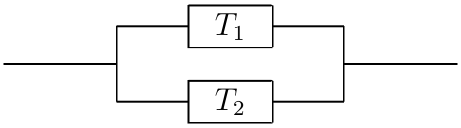
1. *Serienschaltung:*



Eine Serienschaltung funktioniert, wenn sowohl T_1 als auch T_2 funktionieren, das heißt

$$P(A_1 \cap A_2) \overset{\substack{\text{stoch.} \\ \text{unabh.}}}{=} P(A_1) \cdot P(A_2).$$

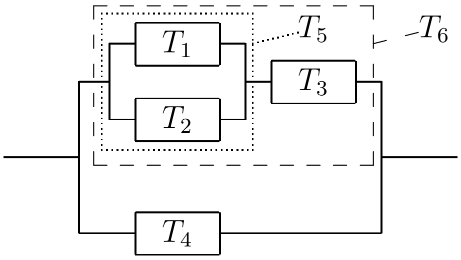
2. *Parallelschaltung:*



Eine Parallelschaltung funktioniert, wenn T_1 oder T_2 funktionieren, das heißt

$$P(A_1 \cup A_2) = 1 - P(\overline{A_1 \cup A_2}) = 1 - P(\overline{A}_1 \cap \overline{A}_2) \overset{\substack{\text{stoch.}\\\text{unabh.}}}{=} 1 - P(\overline{A}_1) \cdot P(\overline{A}_2)$$
$$= 1 - (1 - p_1)(1 - p_2).$$

3. Kombinationen von Reihen- und Parallelschaltungen kann man durch geeignetes Zusammenfassen behandeln:



△

Wir betrachten nun Unabhängigkeitsfragen für n Ereignisse $A_1, A_2, \ldots, A_n \in \mathcal{A}$.

Definition 9.12 Die Ereignisse $A_1, \ldots, A_n$ heißen **vollständig stochastisch unabhängig**, wenn für jede Auswahl von $m \le n$ Ereignissen gilt:

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \ldots \cdot P(A_{i_m}).$$

Bemerkung: Es genügt nicht, nur die Gleichheit

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i)$$

zu fordern! Insbesondere können die Ereignisse $A_1, \ldots, A_n$ *paarweise* stochastisch unabhängig sein, ohne dass sie *vollständig* stochastisch unabhängig sind.

Beispiel 9.13 Sei $\Omega = \{1, 2, 3, 4\}$ mit

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = \frac{1}{4}.$$

Die Ereignisse $A = \{1, 2\}$, $B = \{1, 3\}$ und $C = \{2, 3\}$ besitzen die Wahrscheinlichkeit

$$P(A) = P(B) = P(C) = \frac{1}{2}.$$

Wegen

$$\begin{aligned} P(A \cap B) &= P(\{1\}) = \frac{1}{4} = P(A) \cdot P(B), \\ P(A \cap C) &= P(\{2\}) = \frac{1}{4} = P(A) \cdot P(C), \\ P(B \cap C) &= P(\{3\}) = \frac{1}{4} = P(B) \cdot P(C) \end{aligned}$$

sind A, B, C *paarweise* stochastisch unabhängig. Jedoch ergibt sich

$$P(A \cap B \cap C) = P(\emptyset) = 0 \neq P(A) \cdot P(B) \cdot P(C) = \frac{1}{8},$$

das heißt, es liegt keine *vollständige* stochastische Unabhängigkeit vor! △

Wichtige Anwendung der vollständigen stochastischen Unabhängigkeit: Sind die Ereignisse $A_1, A_2, \ldots, A_n$ vollständig stochastisch unabhängig, so gilt

$$\begin{aligned} P(A_1 \cup A_2 \cup \ldots \cup A_n) &= 1 - P(\overline{A}_1 \cap \ldots \cap \overline{A}_n) \\ &= 1 - P(\overline{A}_1) \cdot P(\overline{A}_2) \cdots P(\overline{A}_n) \\ &= 1 - \big(1 - P(A_1)\big) \cdot \big(1 - P(A_2)\big) \cdots \big(1 - P(A_n)\big). \end{aligned}$$

9.4 Produktexperimente

Wir nehmen an, wir kennen schon Modelle $(\Omega_1, \mathcal{A}_1, P_1), \ldots, (\Omega_n, \mathcal{A}_n, P_n)$ und wollen nun ein Modell für das Experiment konstruieren, welches in der unabhängigen Hintereinanderausführung dieser Teilexperimente besteht.

Definition und Satz 9.14 Werden die Wahrscheinlichkeitsräume $(\Omega_i, \mathcal{A}_i, P_i)$, $i = 1, 2, \ldots, n$, **unabhängig gekoppelt**, so entsteht das **Produkt** der Wahrscheinlichkeitsräume $(\Omega, \mathcal{A}, P)$ gemäß

$$\begin{aligned} \Omega &= \Omega_1 \times \cdots \times \Omega_n &&= \{(\omega_1, \ldots, \omega_n) \colon \omega_i \in \Omega_i\}, \\ \mathcal{A} &= \mathcal{A}_1 \times \cdots \times \mathcal{A}_n &&= \{A_1 \times \cdots \times A_n \colon A_i \in \mathcal{A}_i\}, \\ P &= P_1 \otimes \cdots \otimes P_n. \end{aligned}$$

Ist $X_i(\boldsymbol{\omega})$ die i-te Koordinate von $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$, so wird in Ω das Ereignis, dass sich im i-ten Teilexperiment $A_i \subseteq \Omega_i$ ereignet, durch $\{\boldsymbol{\omega} \in \Omega : X_i(\boldsymbol{\omega}) \in A_i\}$ beschrieben, kurz $\{X_i \in A_i\}$. Es ist

$$A = A_i \times \cdots \times A_n = \bigcap_{i=1}^{n} \{X_i \in A_i\}$$

das Ereignis, das sich für alle $i = 1, \ldots, n$ im i-ten Teilexperiment A_i ereignet. Unter dem *Produktmaß* P ist die Wahrscheinlichkeit dafür

$$P\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = P(A) = P_1(A_1) \cdot P_2(A_2) \cdots P_n(A_n).$$

Hält man $1 \leq k \leq n$ fest und setzt $A_i := \Omega_i$ für $i \neq k$, so folgt

$$\bigcap_{i=1}^{n} \{X_i \in A_i\} = \{X_k \in A_k\},$$

und weiter

$$P(\{X_k \in A_k\}) = P_k(A_k).$$

Dies entspricht der selbstverständlichen Forderung, dass die Wahrscheinlichkeit dafür, dass im k-ten Teilexperiment $A_k \subset \Omega_k$ eintritt, mit der Wahrscheinlichkeit übereinstimmen soll, die A_k im k-ten Teilmodell $(\Omega_k, \mathcal{A}_k, P_k)$ besitzt. Aus der Rechnung folgt insbesondere

$$P\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} P(\{X_i \in A_i\}).$$

Da hierin beliebig viele $A_i = \Omega_i$ gesetzt werden können, gilt

$$P\left(\bigcap_{i \in I} \{X_i \in A_i\}\right) = \prod_{i \in \mathcal{I}} P(\{X_i \in A_i\}), \quad \mathcal{I} \subseteq \{1, 2, \ldots, n\}.$$

Das Modell hat also wirklich die geforderte Eigenschaft, dass darin Ereignisse, die etwas über die Ausgänge verschiedener Teilexperimente aussagen, unabhängig sind.

10. Diskrete Verteilungen

10.1 Zufallsgrößen

Vielfach sind die Ergebnisse von zufälligen Versuchen von Natur aus Zahlenwerte. Häufig möchte man aber auch in Fällen, wo dies nicht der Fall ist, Zahlenwerte zur Charakterisierung der Ergebnisse von Zufallssituationen verwenden. Dies geschieht mit Hilfe von *Zufallsgrößen* X, wobei jedem Ergebnis $\omega \in \Omega$ eine reelle Zahl $X(\omega)$ als Wert der Zufallsgröße zugeordnet wird.

Beispiele 10.1 (Zufallsgrößen)

1. Beim Werfen von zwei Würfeln gilt $\Omega = \{(i,j) : i,j \in \{1,2,\dots,6\}\}$, wobei i und j die gewürfelte Augenzahlen des ersten bzw. zweiten Würfels bezeichnen. Die Augensumme ergibt sich dann als $X(\omega) = i + j$.
2. Für n betrachtete Glühlampen bezeichne ω_i die zufällige Lebensdauer der i-ten Lampe in Stunden. Dann ist $\Omega = \{(\omega_1,\dots,\omega_n) : \omega_i \geq 0\,\forall i\}$ die Ergebnismenge. Sowohl $X(\omega) = \omega_k$ (Lebensdauer der k-ten Lampe), als auch $X(\omega) = \frac{1}{n}(\omega_1 + \omega_2 + \dots + \omega_n)$ (mittlere Lebensdauer von n Lampen) stellen Zufallsgrößen dar.

$\triangle$

Definition 10.2 Sei $(\Omega, \mathcal{A}, \mathcal{P})$ ein Wahrscheinlichkeitsraum zu einer festen Zufallssituation. Dann heißt die Abbildung $X : \Omega \to \mathbb{R}$ **Zufallsgröße** oder **Zufallsvariable** über $(\Omega, \mathcal{A}, \mathcal{P})$, wenn

$$\{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{A}$$

für alle Intervalle I der reellen Achse.

Der Begriff Zufallsvariable ist, obwohl er häufig benutzt wird, etwas irreführend, denn strenggenommen ist eine Zufallsgröße eine *Abbildung*.

Bemerkung: Für die Wahrscheinlichkeit $P(\{\omega \in \Omega : X(\omega) \in I\})$ schreiben wir verkürzt $P(X \in I)$.

Definition 10.3 Eine Zufallsgröße heißt **diskret**, wenn sie nur endlich oder abzählbar unendlich viele Werte annehmen kann.

Die erste Zufallsgröße aus Beispiel 10.1 ist eine diskrete Zufallsgröße. Bei der zweiten handelt es sich um eine stetige Zufallsgröße, auf die wir im nächsten Kapitel eingehen werden.

Definition 10.4 Ist X eine diskrete Zufallsgröße mit den Werten $x_1, x_2, \ldots$, so bezeichnet die Zuordnung

$$p_i := P(X = x_i), \qquad i = 1, 2, \ldots$$

die **Wahrscheinlichkeitsfunktion** der diskreten Zufallsgröße X.

Bemerkung: Eine diskrete Zufallsgröße X wird vollständig durch ihre Wahrscheinlichkeitsfunktion charakterisiert.

Beispiel 10.5 (Werfen mit einem Würfel) Die endliche Zahl der Ereignisse erlaubt eine Darstellung der Wahrscheinlichkeitsfunktion als Tabelle:

i	1	2	3	4	5	6
x_i	1	2	3	4	5	6
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

△

Eigenschaften diskreter Verteilungen:

1. $0 \leq p_i \leq 1$
2. $\sum\limits_i p_i = 1$
3. $P(a \leq X \leq b) = \sum\limits_{a \leq x_i \leq b} p_i$

Beispiel 10.6 (Geometrische Verteilung) Ein Automat sei so eingerichtet, dass er sofort anhält, sobald ein defektes Teil produziert wird. Die Wahrscheinlichkeit dafür, dass ein defektes Teil erzeugt wird, sei p. Die Ausfälle (Defekte) sind von Teil zu Teil unabhängig. Die diskrete Zufallsgröße

$$X \mathrel{\widehat{=}} \text{“Anzahl der produzierten einwandfreien Teile”}$$

besitzt die *geometrische Verteilung*. Dabei gilt für das Ereignis

$$A_i \mathrel{\widehat{=}} \text{“}i\text{-tes produziertes Teil defekt”}$$

$P(A_i) = p$ und $P(\overline{A_i}) = 1 - p$. Dies liefert die Wahrscheinlichkeitsfunktion für X:

$$\begin{aligned}
P(X = 0) &= P(A_1) &&= p, \\
P(X = 1) &= P(\overline{A_1} \cap A_2) &&= (1-p) \cdot p, \\
P(X = 2) &= P(\overline{A_1} \cap \overline{A_2} \cap A_3) &&= (1-p)^2 \cdot p, \\
&\ \vdots \\
P(X = i) &= P(\overline{A_1} \cap \ldots \cap \overline{A_i} \cap A_{i+1}) &&= (1-p)^i \cdot p, \qquad i = 0, 1, 2, \ldots.
\end{aligned}$$

Offensichtlich gilt $0 \leq p \cdot (1-p)^i \leq 1$ für alle $i = 0, 1, 2, \ldots$ und

$$\sum_{i=0}^{\infty} (1-p)^i \cdot p = p \cdot \sum_{i=0}^{\infty} (1-p)^i = p \cdot \frac{1}{1-(1-p)} = p \cdot \frac{1}{p} = 1.$$

△

10.2 Verteilungsfunktion

Definition 10.7 Es sei X eine Zufallsgröße. Dann heißt die zu X gehörige Funktion

$$F(x) := P(X \leq x)$$

Verteilungsfunktion von X. Für diskrete Zufallsgrößen X mit den Werten $x_1, x_2, \ldots$ gilt

$$F(x) = \sum_{x_i \leq x} p_i.$$

Eigenschaften der Verteilungsfunktion:

1. $\lim\limits_{x \to -\infty} F(x) = 0$
2. $\lim\limits_{x \to \infty} F(x) = 1$
3. $F(x)$ ist monoton wachsend (nicht notwendigerweise streng), das heißt
$$x < y \Rightarrow F(x) \leq F(y).$$
4. F ist rechtsseitig stetig, das heißt
$$\lim_{y \searrow x} F(y) = F(x).$$

Bemerkung: Bei diskreten Zufallsgrößen ist die Verteilungsfunktion immer eine reine Treppenfunktion. Die Punkte x_i kennzeichnen die Sprungpunkte und die Werte p_i die dazugehörige Sprunghöhe. Rechtsseitige Stetigkeit bedeutet, dass der Funktionswert an der Sprungstelle dem rechten Teil des Funktionsgraphen zugeordnet wird.

10.3 Erwartungswert

Wir nutzen zur Erklärung das Maschinenmodell der geometrischen Verteilung. Die Maschine wird dabei n mal gestartet mit dem folgenden Ergebnis:

$$\begin{aligned}
h_n(0) &\mathrel{\widehat{=}} \text{“Anzahl der Versuche, die 0 einwandfreie Teile liefern”},\\
h_n(1) &\mathrel{\widehat{=}} \text{“Anzahl der Versuche, die 1 einwandfreie Teile liefern”},\\
&\vdots\\
h_n(i) &\mathrel{\widehat{=}} \text{“Anzahl der Versuche, die i einwandfreie Teile liefern”}.\\
&\vdots
\end{aligned}$$

Wir fragen, wieviel funktionstüchtige Teile die Maschine im Mittel auswirft. Es sind

$$\sum_{i=0}^{\infty} \frac{h_n(i)}{n} \cdot i = \sum_{i=0}^{\infty} H_n(i) \cdot i \overset{n \to \infty}{\longrightarrow} \sum_{i=0}^{\infty} P(X = i) \cdot i$$

funktionstüchtige Teile.

Für die geometrische Verteilung ergibt sich beispielsweise

$$\begin{aligned}\sum_{i=0}^{\infty} P(X=i)\cdot i &= \sum_{i=0}^{\infty} p(1-p)^i \cdot i = p(1-p)\sum_{i=0}^{\infty} i(1-p)^{i-1}\\ &= p\cdot(1-p)\underbrace{\frac{1}{(1-(1-p))^2}}_{=\frac{1}{p^2}} = \frac{1-p}{p},\end{aligned}$$

wobei wir die Summenformel

$$\sum_{i=0}^{\infty} i x^{i-1} = \frac{1}{(1-x)^2} \tag{10.1}$$

benutzt haben, die sich durch Differentation nach x aus der bekannten Identität

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$$

ergibt.

Definition 10.8 Sei $p_i = P(X = x_i)$ die Wahrscheinlichkeitsfunktion einer diskreten Zufallsgröße X. Dann wird

$$E(X) = \sum_i x_i p_i = \sum_i x_i P(X = x_i) \tag{10.2}$$

Erwartungswert oder **Mittelwert** der Zufallsgröße X genannt, falls gilt $\sum_i |x_i| p_i < \infty$. Konvergiert die Reihe (10.2) nicht absolut, so existiert kein Erwartungswert.

Bemerkung: Es gilt die Identität

$$E(X) = \sum_{\omega\in\Omega} P(\{\omega\})\cdot X(\omega).$$

Beispiel 10.9 (Würfeln mit einem Würfel) Für die Zufallsgröße

$$X \mathrel{\widehat{=}} \text{“gewürfelte Augenzahl”}$$

ergibt sich

$$E(X) = \sum_{i=1}^{6} i\cdot\frac{1}{6} = 3.5.$$

△

Eigenschaften des Erwartungswertes

1. Falls $E(X)$ existiert, so gilt für alle $a, b \in \mathbb{R}$

$$E(aX+b) = a\cdot E(X) + b. \tag{10.3}$$

Beweis. Es gilt

$$E(aX+b)=\sum_i (a\cdot x_i+b)p_i = a\cdot \underbrace{\sum_i x_i p_i}_{=E(X)} + b\cdot \underbrace{\sum_i p_i}_{=1} = a\cdot E(X)+b.$$

□

2. Falls $E(X)$ und $E(Y)$ existieren, so gilt

$$E(X+Y)=E(X)+E(Y). \tag{10.4}$$

Beweis. Nach Bemerkung 10.3 folgt

$$\begin{aligned} E(X+Y) &= \sum_{\omega\in\Omega}\big(X(\omega)+Y(\omega)\big)\cdot P(\{\omega\}) \\ &= \sum_{\omega\in\Omega} X(\omega)\cdot P(\{\omega\}) + \sum_{\omega\in\Omega} Y(\omega)\cdot P(\{\omega\}) \\ &= E(X)+E(Y). \end{aligned}$$

□

3. Falls $E(X)$ und $E(Y)$ existieren, gilt

$$X\le Y \Rightarrow E(X)\le E(Y).$$

Beweis. Die Behauptung folgt wieder mit Bemerkung 10.3:

$$E(X)=\sum_{\omega\in\Omega}\underbrace{X(\omega)}_{\le Y(\omega)}\cdot P(\{\omega\}) \le \sum_{\omega\in\Omega} Y(\omega)\cdot P(\{\omega\}) = E(Y).$$

□

Bemerkung: Die Beziehungen (10.3) und (10.4) implizieren die Linearität des Erwartungswerts, dies bedeutet

$$E(aX+bY)=aE(X)+bE(Y). \tag{10.5}$$

Neben $E(X)$ kann auch der Erwartungswert von Funktionen einer Zufallsgröße X betrachtet werden, z.B. $g(X)=\sin(X)$ oder $g(X)=X^2$.

Satz 10.10 Sei $p_i = P(X=x_i)$ die Wahrscheinlichkeitsfunktion der diskreten Zufallsgröße X. Falls $E(X)$ und $E\big(g(X)\big)$ existieren, gilt

$$E\big(g(X)\big)=\sum_i g(x_i)\cdot p_i = \sum_i g(x_i)P(X=x_i).$$

Beweis. Für die Zufallsgröße $Y := g(X)$ folgt

$$\begin{aligned} E(Y) &= \sum_i y_i P(Y = y_i) \\ &= \sum_i \sum_{j:g(x_j)=y_i} g(x_j) \underbrace{P\big(g(X) = g(x_j)\big)}_{=P(X=x_j)} \\ &= \sum_k g(x_k) \cdot P(X = x_k). \end{aligned}$$

□

10.4 Varianz

Definition 10.11 Existiert $E(X^2)$, so heißt die Größe $\sigma^2 = V(X)$ mit

$$V(X) := E\big([X - E(X)]^2\big)$$

Varianz oder **Streuung** der Zufallsgröße X. Die Wurzel $\sigma := \sqrt{V(X)}$ heißt **Standardabweichung** der Zufallsgröße X.

Bemerkung:

1. Die Varianz $V(X)$ ist die mittlere quadratische Abweichung einer Zufallsgröße X von ihrem Erwartungswert $E(X)$.
2. Wegen
$$|X| \le 1 + X^2$$
folgt aus der Existenz von $E(X^2)$ auch die Existenz von $E(X)$.

Lemma 10.12 Für eine Zufallsgröße X gilt mit $a, b \in \mathbb{R}$

$$V(aX + b) = a^2 V(X).$$

Beweis. Gemäß der Definition der Varianz folgt

$$\begin{aligned} V(aX + b) &= E\big([aX + b - \underbrace{E(aX + b)}_{=aE(X)+b}]^2\big) = E\big([aX - aE(X)]^2\big) \\ &= a^2 E\big([X - E(X)]^2\big) = a^2 V(X), \end{aligned}$$

das ist die Behauptung. □

Satz 10.13 Es gilt

$$V(X) = E(X^2) - E^2(X).$$

Beweis. Wir haben

$$\begin{aligned} V(X) &= E\big([X - E(X)]^2\big) = E\big(X^2 - 2XE(X) + E^2(X)\big) \\ &\overset{(10.5)}{=} E(X^2) - 2 \cdot \underbrace{E(X) \cdot E(X)}_{=E^2(X)} + E^2(X) = E(X^2) - E^2(X). \end{aligned}$$

□

Beispiel 10.14 (Varianz der geometrischen Verteilung) Wir berechnen die Varianz $V(X)$ für die geometrische Verteilung auf Grundlage der Formeln

$$p_i = p \cdot (1-p)^i, \qquad E(X) = \frac{1-p}{p}, \qquad V(X) = E(X^2) - E^2(X).$$

Hierzu benötigen wir die folgende Summenformel

$$\sum_{i=0}^{\infty} i(i-1)x^{i-2} = \frac{2}{(1-x)^3},$$

die durch Differentiation nach x aus (10.1) folgt. Damit erhalten wir

$$\begin{aligned} E(X^2) &= \sum_{i=0}^{\infty} i^2 p(1-p)^i \\ &= \sum_{i=0}^{\infty} i(i-1) \underbrace{p(1-p)^i}_{=p(1-p)^2(1-p)^{i-2}} + \underbrace{\sum_{i=0}^{\infty} ip(1-p)^i}_{=E(X)} \\ &= p \cdot \underbrace{\frac{2}{(1-(1-p))^3}}_{=2/p^3} (1-p)^2 + \frac{1-p}{p} \\ &= \frac{2(1-p)^2}{p^2} + \frac{1-p}{p} \end{aligned}$$

und weiter

$$V(X) = E(X^2) - E^2(X) = \frac{(1-p)^2}{p^2} + \frac{1-p}{p}.$$

△

Satz 10.15 (Tschebyscheffsche Ungleichung) Für alle $\varepsilon > 0$ gilt

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$$

oder anders geschrieben

$$P(|X - E(X)| \geq k \cdot \sigma) \leq \frac{1}{k^2}, \qquad k > 0.$$

Beweis. Für eine diskrete Zufallsgröße ergibt sich die Behauptung aus

$$\begin{aligned} V(X) &= \sum_i \big(x_i - E(X)\big)^2 p_i \geq \sum_{i:|x_i - E(X)| \geq \varepsilon} \underbrace{\big(x_i - E(X)\big)^2}_{\geq \varepsilon^2} p_i \\ &\geq \varepsilon^2 \sum_{i:|x_i - E(X)| \geq \varepsilon} p_i = \varepsilon^2 \cdot P(|X - E(X)| \geq \varepsilon) \end{aligned}$$

durch Auflösen nach $P(|X - E(X)| \geq \varepsilon)$. Der Beweis im Falle stetiger Zufallsgrößen wird später nachgeliefert. □

Bemerkung: Wie der Beweis zeigt, gilt die Tschebyscheffsche Ungleichung auch im Falle strenger Ungleichheitszeichen.

10.5 Schwaches Gesetz der großen Zahlen

Definition 10.16 Eine Folge von Zufallsgrößen $X_1, X_2, \ldots, X_n$ heißt **stochastisch unabhängig**, wenn sich der zufällige Charakter aller beteiligten Zufallsgrößen gegenseitig nicht beeinflusst, das heißt, wenn gilt

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n).$$

Bemerkung: Im Gegensatz zu Ereignissen (vgl. Definition 9.12) ist hier die Definition ausreichend. Die Unabhängigkeit von $X_1, X_2, \ldots, X_n$ impliziert aufgrund von

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \\ &\quad = \sum_{x_n} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &\quad = \sum_{x_n} P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_n = x_n) \\ &\quad = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_{n-1} = x_{n-1}) \underbrace{\sum_{x_n} P(X_n = x_n)}_{=1} \end{aligned}$$

auch die Unabhängigkeit von $X_1, X_2, \ldots, X_{n-1}$. Rekursiv erhält man hieraus

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdots P(X_m = x_m)$$

für alle $m \leq n$. Weil aber die Bezeichnung der X_i beliebig ist, folgt sogar für alle $m \leq n$ die Beziehung

$$\begin{aligned} &P(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \ldots, X_{i_m} = x_{i_m}) \\ &\qquad\qquad = P(X_{i_1} = x_{i_1}) \cdot P(X_{i_2} = x_{i_2}) \cdots P(X_{i_m} = x_{i_m}). \end{aligned}$$

Lemma 10.17 Die Zufallsgrößen $X_1, X_2, \ldots, X_n$ seien stochastisch unabhängig und $E(X_1), E(X_2), \ldots, E(X_n)$ mögen existieren. Dann gilt

$$E(X_1 \cdot X_2 \cdots X_n) = E(X_1) \cdot E(X_2) \cdots E(X_n).$$

Beweis. Die Behauptung folgt aus

$$\begin{aligned} E(X_1 \cdot X_2 \cdots X_n) &= \sum_{x_1,\ldots,x_n} x_1 \cdot x_2 \cdots x_n \cdot \underbrace{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}_{=P(X_1=x_1)\cdot P(X_2=x_2)\cdots P(X_n=x_n)} \\ &= \underbrace{\left(\sum_{x_1} x_1 \cdot P(X_1 = x_1)\right)}_{=E(X_1)} \cdots \underbrace{\left(\sum_{x_n} x_n \cdot P(X_n = x_n)\right)}_{=E(X_n)}. \end{aligned}$$

□

Satz 10.18 Gegeben seien die Zufallsgrößen $X_1, X_2, \ldots, X_n$ mit endlichen Erwartungswerten und Varianzen. Sind $X_1, X_2, \ldots, X_n$ unabhängig, so gilt für alle $a_1, \ldots, a_n \in \mathbb{R}$

$$V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \ldots + a_n^2V(X_n).$$

Beweis. Nach Satz 10.13 folgt

$$\begin{aligned} V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = E\big((a_1X_1 + a_2X_2 + \ldots + a_nX_n)^2\big) \\ - E^2(a_1X_1 + a_2X_2 + \ldots + a_nX_n). \end{aligned}$$

Durch Ausmultiplizieren unter Berücksichtigung der Linearität des Erwartungswertes erhalten wir

$$\begin{aligned} &V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) \\ &\quad = E\left(a_1^2X_1^2 + a_2^2X_2^2 + \ldots + a_n^2X_n^2 + \sum_{i\neq j} a_ia_jX_iX_j\right) \\ &\qquad - \big(a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n)\big)^2 \\ &\quad = a_1^2E(X_1^2) + a_2^2E(X_2^2) + \ldots + a_n^2E(X_n^2) + \sum_{i\neq j} a_ia_jE(X_i)E(X_j) \\ &\qquad - \left(a_1^2E^2(X_1) + a_2^2E^2(X_2) + \ldots + a_n^2E^2(X_n) + \sum_{i\neq j} a_ia_jE(X_i)E(X_j)\right) \\ &\quad = a_1^2\underbrace{\big(E(X_1^2) - E^2(X_1)\big)}_{=V(X_1)} + a_2^2\underbrace{\big(E(X_2^2) - E^2(X_2)\big)}_{=V(X_2)} + \ldots + a_n^2\underbrace{\big(E(X_n^2) - E^2(X_n)\big)}_{=V(X_n)} \\ &\quad = a_1^2V(X_1) + a_2^2V(X_2) + \ldots + a_n^2V(X_n). \end{aligned}$$

□

Satz 10.19 (Schwaches Gesetz der großen Zahlen) Es sei $X_1, X_2, \dots, X_n$ eine Folge von unabhängigen Zufallsgrößen mit

$$E(X_i) = \mu, \qquad V(X_i) = \sigma_i^2 \leq M < \infty, \qquad i = 1, 2, \dots, n.$$

Dann gilt für alle $\varepsilon > 0$

$$P\left(\left|\frac{1}{n}(X_1 + X_2 + \dots + X_n) - \mu\right| \geq \varepsilon\right) \leq \frac{M}{\varepsilon^2 n} \overset{n\to\infty}{\longrightarrow} 0.$$

Beweis. Setzen wir

$$\overline{X}_n := \frac{1}{n}(X_1 + X_2 + \dots + X_n),$$

dann gilt

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

und

$$V(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \leq \frac{1}{n^2}\sum_{i=1}^{n} M = \frac{M}{n}.$$

Die Behauptung folgt nun sofort aus der Tschebyscheffschen Ungleichung. □

Beispiel 10.20 (Konvergenz der relativen Häufigkeit) Gegeben sei eine Zufallssituation mit dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$. Wir betrachten ein festes, zufälliges Ereignis $A \in \mathcal{A}$ und realisieren die Zufallssituation in n Versuchen. Definieren wir für $i = 1, 2, \dots, n$ die Zufallsgrößen

$$X_i = \begin{cases} 1, & A \text{ tritt im } i\text{-ten Versuch ein,} \\ 0, & A \text{ tritt im } i\text{-ten Versuch nicht ein,} \end{cases}$$

so gilt

$$E(X_i) = P(A) := p, \qquad V(X_i) = p(1-p) \leq \frac{1}{4}.$$

Das schwache Gesetzt der großen Zahlen liefert für die relative Häufigkeit $H_n(A) = \frac{1}{n}\sum_{i=1}^{n} X_i$ die Abschätzung

$$P(|H_n(A) - p| \geq \varepsilon) \leq \frac{1}{4\varepsilon^2 n} \overset{n\to\infty}{\longrightarrow} 0.$$

△

10.6 Binomialverteilung

Wir betrachten eine Zufallssituation mit dem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und ein festes Ereignis $A \in \mathcal{A}$. Ein dieser Situation entsprechender Versuch wird (unabhängig) n-mal wiederholt. Dazu sei

$$X_i = \begin{cases} 1, & \text{wenn } A \text{ im } i\text{-ten Versuch eintritt,} \\ 0, & \text{wenn } A \text{ im } i\text{-ten Versuch nicht eintritt,} \end{cases}$$

und

$$P(A) = P(X_i = 1) := p, \qquad P(\overline{A}) = P(X_i = 0) = 1 - p.$$

Für ein Tupel $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ mit $\sum_{i=1}^n x_i = k$ gilt

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) = p^k(1-p)^{n-k}.$$

Da es insgesamt $\binom{n}{k}$ Tupel mit $\sum_{i=1}^n x_i = k$ gibt, folgt für die Zufallsgröße

$$\overline{X}_n = X_1 + X_2 + \cdots + X_n,$$

dass

$$\begin{aligned} P(\overline{X}_n = k) &= P\left(\left\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n : \sum_{i=1}^n x_i = k\right\}\right) \\ &= \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

Insbesondere gilt

$$E(\overline{X}_n) = \sum_{i=1}^n \underbrace{E(X_i)}_{=p} = np, \qquad V(\overline{X}_n) = \sum_{i=1}^n \underbrace{V(X_i)}_{=p(1-p)} = np(1-p).$$

Definition 10.21 (Binomialverteilung) Eine Zufallsgröße X mit der Wahrscheinlichkeitsfunktion

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n$$

heißt **binomialverteilt** mit den Parametern n (Zahl der Freiheitsgrade) und p (Fehlerrate), kurz

$$X \sim \mathrm{Bin}(n, p).$$

Beispiel 10.22 (Werfen eines idealen Würfels) Ein idealer Würfel wird $n = 20$ mal geworfen. Mit welcher Wahrscheinlichkeit werden mindestens zwei Sechsen gewürfelt? Dazu sei

$$A \widehat{=} \text{“Es wird eine 6 gewürfelt”},$$

wobei $p = P(A) = 1/6$, und

$$X \widehat{=} \text{“Anzahl der geworfenen Sechsen bei } n = 20 \text{ Würfen”}.$$

Wegen $X \sim \mathrm{Bin}(n, p) = \mathrm{Bin}(20, 1/6)$ folgt

$$\begin{aligned} P(X \geq 2) &= 1 - P(\underbrace{X < 2}_{\widehat{=}\, X \leq 1}) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \binom{20}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{20} - \binom{20}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{9}\right)^{19} \\ &= 0.8696. \end{aligned}$$

Um die Frage zu beantworten, mit welcher Wahrscheinlichkeit die tatsächliche bei $n = 20$ Würfen erzielte Anzahl von Sechsen um mehr als 3 vom Mittelwert $E(X) = n \cdot p = 20 \cdot 1/6 = 3.\overline{3}$ abweicht, kann man die Tschebyscheffsche Ungleichung

$$P(|X - E(X)| > 3) < \frac{V(X)}{9}$$

benutzen. Für unser Beispiel erhalten wir

$$V(X) = n \cdot p \cdot (1 - p) = 20 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{25}{9}.$$

Die Tschebbyscheffsche Ungleichung sichert mit

$$P(|X - 3.\overline{3}| > 3) < \frac{25}{9 \cdot 9} = 0.31,$$

dass in höchstens 31% aller Fälle eine solch große Abweichung der gewürfelten Anzahl von Sechsen zum Mittelwert auftritt. Summiert man aber exakt alle Wahrscheinlichkeiten der Fälle mit einer so großen Abweichung, so ergibt sich:

$$P(X = 0) + P(X = 7) + P(X = 8) + \ldots + P(X = 20) = 0.0632.$$

Es sind also in Wirklichkeit nur reichlich 6% aller Fälle, in denen eine Abweichung von mehr als 3 vom Mittelwert auftritt. Die Tschebbyscheffsche Ungleichung liefert insofern eine recht grobe Abschätzung. $\triangle$

10.7 Poisson-Verteilung

Als Referenzmodell für eine Poisson-verteilte Zufallsgröße wollen wir eine Telefonzelle betrachten: Innerhalb eines Zeitintervalls der Länge t kommen X_t Anrufe an. Dabei ist die Zufallsgröße X_t *Poisson-verteilt*, wenn die folgenden drei *Poissonschen Annahmen* gelten:

1. *Stationarität:* Die Wahrscheinlichkeit für das Eintreten einer bestimmten Anzahl von Ereignissen im betrachteten Zeitintervall hängt nur von der Länge t des Zeitintervalls ab, nicht aber von seiner konkreten Lage auf der Zeitachse.

2. *Homogenität:* Die Ereignisfolge sei ohne Nachwirkungen, das heißt, die Anzahl von Ereignissen im Zeitintervall $[t_0, t_1]$ hat keinen Einfluss auf die Anzahl von Ereignissen in einem späteren Zeitintervall $[t_2, t_3]$ mit $t_2 > t_1$.

3. *Ordinarität:* Die Ereignisse treten für hinreichend kleine Zeitintervalle einzeln auf, d.h. für Δt genügend klein gilt entweder $X_{\Delta t} = 0$ oder $X_{\Delta t} = 1$. Weiter gelte $P(X_{\Delta t} = 1) = \mu \cdot \Delta t$. Der Parameter μ mit $0 < \mu < \infty$ heißt *Intensität*.

Definition 10.23 (Poisson-Verteilung) Eine Zufallsgröße X_t, welche der Wahrscheinlichkeitsfunktion

$$P(X_t = k) = \frac{(\mu t)^k}{k!} e^{-\mu t}, \qquad k = 0, 1, 2, \ldots \tag{10.6}$$

genügt, heißt **Poisson-verteilt**. Oft wird $\lambda = \mu \cdot t$ gesetzt (λ ist der Parameter der Poisson-Verteilung) und $X_t \sim \pi_\lambda = \pi_{\mu t}$ geschrieben.

Begründung von Formel (10.6): Wir teilen das Zeitintervall der Länge t in n gleichlange, hinreichend kleine Teilintervalle $\Delta t = t/n$ mit $P(X_{\Delta t} = 1) = \mu \cdot \Delta t$ auf. Dann liefert die Binomialverteilung

$$P(X_t = k) = \binom{n}{k} (\mu \cdot \Delta t)^k (1 - \mu \Delta t)^{n-k},$$

da in k aus n Teilintervallen ein Anruf eingeht, während in $n-k$ Teilintervallen kein Anruf registriert wird. Es folgt:

$$\begin{aligned} P(X_t = k) &= \frac{n!}{k!(n-k)!} \left(\frac{\mu t}{n}\right)^k \left(1 - \frac{\mu t}{n}\right)^{n-k} \\ &= \frac{(\mu t)^k}{k!} \underbrace{\left(1 - \frac{\mu t}{n}\right)^n}_{\xrightarrow[n\to\infty]{} e^{-\mu t}} \underbrace{\frac{\overbrace{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}^{k \text{ Faktoren}}}{n^k}}_{\xrightarrow[n\to\infty]{} 1} \underbrace{\frac{1}{\left(1 - \frac{\mu t}{n}\right)^k}}_{\xrightarrow[n\to\infty]{} 1} \\ &\xrightarrow{n\to\infty} \frac{(\mu t)^k}{k!} e^{-\mu t}. \end{aligned}$$

Dies zeigt insbesondere, dass die Poisson-Verteilung als Grenzverteilung der Binomialverteilung im Fall $p_n \cdot n \to \lambda$ und $n \to \infty$ aufgefasst werden kann.

Beispiel 10.24 (Telefonzentrale) Wir suchen für eine Telefonzentrale die Wahrscheinlichkeit dafür, dass innerhalb einer Viertelstunde wenigstens 3 und höchstens 7 Anrufe ankommen. Dabei sei die gewählte Zeiteinheit "Minuten" und die entsprechende Intensität $\mu = 1/3$. Es gilt dann mit $\lambda = \mu \cdot t = 1/3 \cdot 15 = 5$ für die Zufallsgröße X_t, welche die Anzahl der ankommenden Anrufe charakterisiert

$$X_t \sim \pi_{\mu t} = \pi_5,$$

also

$$P(3 \leq X_t \leq 7) = \sum_{k=3}^{7} \frac{5^k}{k!} e^{-5} = 0.742.$$

$\triangle$

Berechnung von $E(X_t)$ im Falle $X_t \sim \pi_\lambda$: Es gilt

$$E(X_t) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \overset{j:=k-1}{=} \lambda e^{-\lambda} \cdot \underbrace{\sum_{j=0}^{\infty} \frac{\lambda^j}{j!}}_{=e^\lambda} = \lambda.$$

Bemerkung: Für eine Poisson-verteilte Zufallsgröße $X_t \sim \pi_{\mu t}$ gibt die Intensität μ mit $\mu = E(X_t)/t$ die mittlere Anzahl von auftretenden Ereignissen pro Zeiteinheit an.

Berechnung von $V(X_t)$ im Falle von $X_t \sim \pi_{\mu t}$: Mit

$$
\begin{aligned}
E(X_t^2) &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \left(\sum_{k=0}^{\infty} k \cdot (k-1) \frac{\lambda^k}{k!} e^{-\lambda} \right) + \underbrace{\left(\sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \right)}_{=E(X_t)} \\
&= \left(\lambda^2 e^{-\lambda} \sum_{k=-2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \right) + \lambda \\
&\overset{j:=k-2}{=} \lambda^2 e^{-\lambda} \underbrace{\left(\sum_{j=0}^{\infty} \frac{\lambda^i}{j!} \right)}_{=e^{\lambda}} + \lambda \\
&= \lambda^2 + \lambda
\end{aligned}
$$

folgt

$$V(X) = E(X_t^2) - E^2(X_t) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

10.8 Hypergeometrische Verteilung

Als Referenzmodell zur hypergeometrischen Verteilung wird das schon aus der Kombinatorik bekannte Urnenmodell herangezogen. Eine Urne enthalte N Kugeln, von denen M schwarz sind. Der Rest der Kugeln sei weiß. Wir ziehen ohne Zurücklegen n Kugeln aus der Urne und zählen die dabei gezogenen schwarzen Kugeln, das heißt

$$X \widehat{=} \text{“Anzahl der entnommenen schwarzen Kugeln”}.$$

Es gilt

$$P(X = m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}, \qquad m = 0, 1, 2, \ldots, \min\{n, M\}. \tag{10.7}$$

Definition 10.25 Eine Zufallsgröße X mit der Wahrscheinlichkeitsfunktion (10.7) heißt **hypergeometrisch verteilt** mit den Parametern n, N, M, kurz

$$X \sim H(n, N, M).$$

Berechnung von $E(X)$ im Falle $X \sim H(n,N,M)$: Es gilt

$$
\begin{aligned}
E(X) &= \sum_{m=0}^{\min\{n,M\}} m \frac{\binom{M}{m} \cdot \binom{N-M}{n-m}}{\binom{N}{n}} \\
&= \sum_{m=1}^{\min\{n,M\}} m \cdot \frac{M!}{m!(M-m)!} \cdot \frac{(N-M)!}{(n-m)!((N-M)-(n-m))!} \cdot \frac{n!(N-n)!}{N!} \\
&= n \cdot \frac{M}{N} \sum_{m=1}^{\min\{n,M\}} \frac{(M-1)!}{(m-1)!(M-1)-(m-1))!} \\
&\qquad \cdot \frac{((N-1)-(M-1))!}{((n-1)-(m-1))!(((N-1)-(M-1))-((n-1)-(m-1)))!} \\
&\qquad \cdot \frac{(n-1)!((N-1)-(n-1))}{(N-1)!} \\
&= n \cdot \frac{M}{N} \sum_{m=1}^{\min\{n,M\}} \frac{\binom{M-1}{m-1}\binom{(N-1)-(M-1)}{(n-1)-(m-1)}}{\binom{N-1}{n-1}}.
\end{aligned}
$$

Hierin erfüllt der letzte Term

$$
\sum_{m=1}^{\min\{n,M\}} \frac{\binom{M-1}{m-1}\binom{(N-1)-(M-1)}{(n-1)-(m-1)}}{\binom{N-1}{n-1}} \overset{j:=m-1}{=} n \cdot \frac{M}{N} \sum_{j=0}^{\min\{n-1,M-1\}} \frac{\binom{M-1}{j}\binom{(N-1)-(M-1)}{(n-1)-j}}{\binom{N-1}{n-1}} = 1,
$$

da dies der Aufsummation der Wahrscheinlichkeitsfunktion einer Zufallsgröße $Y \sim H(n-1, N-1, M-1)$ entspricht. Damit erhalten wir schließlich

$$
E(X) = n \cdot \frac{M}{N}.
$$

Berechnung von $V(X)$ im Falle $X \sim H(n,N,M)$: Ähnlich zum Erwartungswert ergibt sich

$$
\begin{aligned}
E(X^2) &= \sum_{m=0}^{\min\{n,M\}} m^2 \frac{\binom{M}{m} \cdot \binom{N-M}{n-m}}{\binom{N}{n}} \\
&= \sum_{m=0}^{\min\{n,M\}} m(m-1) \frac{\binom{M}{m} \cdot \binom{N-M}{n-m}}{\binom{N}{n}} + E(X) \\
&= n(n-1) \cdot \frac{M(M-1)}{N(N-1)} \underbrace{\sum_{m=2}^{\min\{n,M\}} \frac{\binom{M-2}{m-2}\binom{(N-2)-(M-2)}{(n-2)-(m-2)}}{\binom{N-2}{n-2}}}_{=1} + E(X) \\
&= n(n-1) \cdot \frac{M(M-1)}{N(N-1)} + n \cdot \frac{M}{N},
\end{aligned}
$$

womit nach kleinerer Rechnung das Ergebnis folgt

$$
\begin{aligned}
V(X) &= E(X^2) - E^2(X) \\
&= n(n-1) \cdot \frac{M(M-1)}{N(N-1)} + n \cdot \frac{M}{N} - n^2 \cdot \frac{M^2}{N^2} \\
&= \frac{nM(M-N)(n-N)}{N^2(N-1)} \\
&= n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}.
\end{aligned}
$$

11. Stetige Verteilungen

11.1 Dichtefunktion

Definition 11.1 (Stetige Zufallsgröße) Eine Zufallsgröße X heißt **stetig**, wenn eine reelle integrierbare Funktion $f(x)$ existiert, so dass

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

gilt. Dabei heißt $f(x)$ die **Dichtefunktion** der Zufallsgröße X.

Eigenschaften der Dichtefunktion

1. Die Dichtefunktion ist nichtnegativ, dies bedeutet

$$f(x) \geq 0 \quad \text{für alle} \quad x \in \mathbb{R}.$$

2. Es gilt

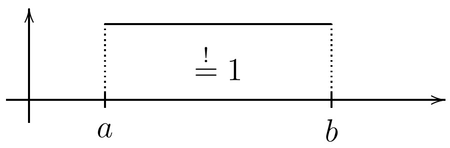
$$P(-\infty < X < \infty) = P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Das Integral ist dabei im Sinne von Riemann oder Lebesgue zu verstehen. Als Funktionen $f(x)$ treten vorzugsweise stetige oder stückweise stetige Funktionen auf, die gegebenenfalls auch schwache Polstellen besitzen dürfen. Wegen

$$P(a \leq X \leq a) = P(X = a) = \int_a^a f(x)\,\mathrm{d}x = 0$$

ist die Wahrscheinlichkeit, dass X genau einen festen Wert annimmt, immer gleich Null.

Beispiel 11.2 (Gleichverteilung) Bei der Gleichverteilung auf dem Intervall $[a, b]$ sind die Werte von X gleichwahrscheinlich über das Intervall $[a, b]$ verteilt. Außerhalb des Intervalls $[a, b]$ kann X keine Werte annehmen.



Damit ergibt sich für die Dichtefunktion

$$f(x) = \begin{cases} C, & x \in [a, b], \\ 0, & \text{sonst.} \end{cases}$$

Die Normierungseigenschaft der Dichtefunktion impliziert

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_a^b C\,\mathrm{d}x = C(b-a),$$

das heißt, der konkrete Wert der Konstanten ist

$$C = \frac{1}{b-a}.$$

△

11.2 Erwartungswert und Varianz

Wir wollen den Erwartungswert einer stetigen Zufallsgröße X mittels Approximation einführen. Dazu definieren wir für $n \in \mathbb{N}$ die diskrete Zufallsgröße X_n mit den Werten k/n und der Wahrscheinlichkeitsfunktion

$$P(X_n = k/n) = P\left(\frac{k}{n} \leq X < \frac{k+1}{n}\right), \quad k \in \mathbb{Z}.$$

Insbesondere ist dann

$$X_n \leq X \leq X_n + \frac{1}{n}$$

und

$$|X_n - X_m| \leq |X_n - X| + |X - X_m| \leq \frac{1}{n} + \frac{1}{m}.$$

Existiert $E(X_n)$ für ein n, so existiert folglich $E(X_n)$ für alle $n \in \mathbb{N}$ und es gilt

$$|E(X_n) - E(X_m)| \leq \frac{1}{n} + \frac{1}{m},$$

dies bedeutet, $E(X_n)$ ist eine Cauchy-Folge. Wir sagen, dass $E(X)$ existiert, falls $E(X_1)$ existiert, und setzen

$$E(X) := \lim_{n\to\infty} E(X_n).$$

Satz 11.3 Es sei X eine stetige Zufallsgröße mit Dichtefunktion f und $g : \mathbb{R} \to \mathbb{R}$ eine stetige Funktion. Dann existiert $E\big(g(X)\big)$ genau dann, wenn

$$I := \int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty,$$

und in diesem Fall gilt

$$E\big(g(X)\big) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x.$$

Insbesondere ist

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

falls $\int_{-\infty}^{\infty} |x|f(x)\,\mathrm{d}x < \infty$.

Beweis. Aufgrund der Stetigkeit von g existiert zu jedem $\varepsilon > 0$ eine strikt monoton wachsende Folge $\{x_n\}_{n\in\mathbb{Z}}$ mit $x_n \to -\infty$ für $n \to -\infty$ und $x_n \to \infty$ für $n \to \infty$, so dass

$$|g(x) - g(x_n)| < \varepsilon \quad \text{für} \quad x_n \leq x \leq x_{n+1}.$$

Setzen wir $g_\varepsilon(x) := g(x_n)$ für $x_n \leq x < x_{n+1}$, so gilt

$$|g_\varepsilon(x) - g(x)| < \varepsilon \tag{11.1}$$

und

$$E\big(g_\varepsilon(X)\big) = \sum_{n=-\infty}^{\infty} g(x_n) P(x_n \leq X < x_{n+1}) = \sum_{n=-\infty}^{\infty} g(x_n) \int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x.$$

Hierin konvergiert die letzte Summe genau dann absolut, wenn I endlich ist, denn

$$\begin{aligned}\sum_{n=-\infty}^{\infty} |g(x_n)| \int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x &\leq \int_{-\infty}^{\infty} |g(x)| f(x)\,\mathrm{d}x + \sum_{n=-\infty}^{\infty} \int_{x_n}^{x_{n+1}} \underbrace{|g(x_n) - g(x)|}_{<\varepsilon} f(x)\,\mathrm{d}x \\ &< I + \varepsilon\end{aligned}$$

und

$$\begin{aligned}\sum_{n=-\infty}^{\infty} |g(x_n)| \int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x &\geq \int_{-\infty}^{\infty} |g(x)| f(x)\,\mathrm{d}x - \sum_{n=-\infty}^{\infty} \int_{x_n}^{x_{n+1}} \underbrace{|g(x_n) - g(x)|}_{<\varepsilon} f(x)\,\mathrm{d}x \\ &> I - \varepsilon.\end{aligned}$$

Insbesondere ist wegen (11.1)

$$\begin{aligned}&\left| E\big(g(X)\big) - \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x \right| \\ &\qquad \leq \underbrace{\big| E\big(g(X)\big) - E\big(g_\varepsilon(X)\big)\big|}_{=|E(g(X)-g_\varepsilon(X))|} + \left| E\big(g_\varepsilon(X)\big) - \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x \right| < 2\varepsilon,\end{aligned}$$

woraus die Behauptung folgt. □

Definition 11.4 Sei X eine stetige Zufallsgröße mit Dichtefunktion f und

$$\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x < \infty.$$

Dann ist die **Varianz** von X definiert als

$$V(X) = E\big([X - E(X)]^2\big) = E(X^2) - E^2(X).$$

Beispiel 11.5 (Gleichverteilung (Fortsetzung von Beispiel 11.2)) Die Dichtefunktion der Gleichverteilung lautet

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b], \\ 0, & \text{sonst.} \end{cases}$$

Für den Erwartungswert folgt daher

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \int_a^b x\frac{1}{b-a}\,\mathrm{d}x = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

Weiter folgt mit

$$E(X^2) = \int_{-\infty}^{\infty} x^2\frac{1}{b-a}\,\mathrm{d}x = \int_a^b x^2\frac{1}{b-a}\,\mathrm{d}x = \frac{x^3}{3(b-a)}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

für die Varianz

$$\begin{aligned} V(X) &= E(X^2) - E^2(X) \\ &= \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4a^2+4ab+4b^2-3(a^2+2ab+b^2)}{12} \\ &= \frac{a^2-2ab+b^2}{12} \\ &= \frac{(a-b)^2}{12}. \end{aligned}$$

△

Satz 11.6 Auch für die stetigen Zufallsgrößen gilt die Tschebyscheffsche Ungleichung

$$P(|X-E(X)| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}, \qquad \varepsilon > 0.$$

Beweis. Analog zum Beweis von Satz 10.15 folgt die Behauptung durch Umstellen von

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} \big(x-E(X)\big)^2 f(x)\,\mathrm{d}x \\ &\geq \int_{|x-E(X)|\geq\varepsilon} \underbrace{\big(x-E(X)\big)^2}_{\geq\varepsilon^2} f(x)\,\mathrm{d}x \\ &\geq \varepsilon^2 \int_{|x-E(X)|\geq\varepsilon} f(x)\,\mathrm{d}x \\ &= \varepsilon^2 P(|X-E(X)| \geq \varepsilon). \end{aligned}$$

□

11.3 Verteilungsfunktion

Definition 11.7 (Verteilungsfunktion) Die reelle Funktion

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

heißt **Verteilungsfunktion** der stetigen Zufallsgröße X.

Eigenschaften der Verteilungsfunktion

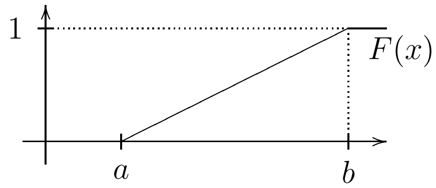
1. $\lim\limits_{x\to-\infty} F(X) = 0$
2. $\lim\limits_{x\to\infty} F(X) = 1$
3. $P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$
4. $F(x)$ ist monoton wachsend (nicht notwendigerweise streng), das heißt
$$x < y \Rightarrow F(x) \leq F(y).$$
5. $F(x)$ ist eine *stetige* Funktion für alle $x \in \mathbb{R}$. Dies bedeutet, dass wir im Gegensatz zur Verteilungsfunktion einer diskreten Zufallsgröße links- und rechtsseitige Stetigkeit haben.
6. Falls $f(x)$ in $y \in \mathbb{R}$ stetig ist, so ist $F(x)$ in y differenzierbar und es gilt
$$F'(y) = f(y).$$

Beispiel 11.8 (Gleichverteilung (Fortsetzung von Beispiel 11.2)) Wir berechnen nun die Verteilungsfunktion für die Gleichverteilung. Sei $x \in [a, b]$, dann gilt

$$F(X) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_a^x \frac{1}{b-a}\,\mathrm{d}t = \frac{t}{b-a}\Big|_a^x = \frac{x-a}{b-a}.$$

Folglich ergibt sich für die Verteilungsfunktion

$$F(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x \leq b, \\ 1, & x > b. \end{cases}$$



△

11.4 Exponentialverteilung

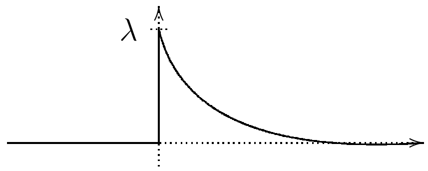
Definition 11.9 Eine stetige Zufallsgröße mit der Dichtefunktion

$$f(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0. \end{cases}$$

heißt **exponentialverteilt** mit dem Parameter $\lambda > 0$, kurz

$$X \sim \mathrm{Ex}(\lambda).$$

Die Dichtefunktion $f(x)$ besitzt folgendes Aussehen:



Satz 11.10 Die Verteilungsfunktion einer exponentialverteilten Zufallsgröße lautet

$$F(X) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Beweis. Im Falle $x < 0$ ist nichts zu zeigen. Sei also $x \geq 0$, dann folgt

$$F(X) = \int_0^x \lambda \cdot e^{-\lambda t} \, \mathrm{d}t = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x}.$$

□

Zusammenhang zwischen Exponential- und Poisson-Verteilung: Wir betrachten wieder die Telefonzentrale, wobei

$\mu \mathrel{\widehat{=}}$ "Intensität",

$X_t \mathrel{\widehat{=}}$ "Anzahl der Anrufe im Zeitintervall der Länge t",

$T \mathrel{\widehat{=}}$ "Zeitabstand zwischen 2 Anrufen"

bezeichne. Wir wissen bereits, dass

$$X_t \sim \pi_{\mu t}$$

gilt. Im nachfolgenden Satz wird gezeigt, dass T exponentialverteilt ist mit μ als Parameter, das heißt

$$T \sim \mathrm{Ex}(\mu).$$

Satz 11.11 Die Zufallsgröße X_t zähle das Eintreten eines Ereignisses A innerhalb eines Zeitintervalls der Länge t. Die Zeit T, die nach dem Eintreten von A bis zum nächsten Eintreten verstreicht, ist exponentialverteilt

$$T \sim \mathrm{Ex}(\mu),$$

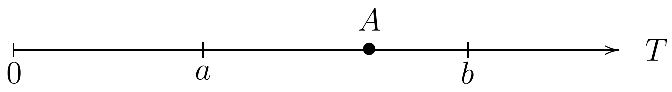
falls X_t Poisson-verteilt ist mit der Intensität μ, das heißt

$$X_t \sim \pi_{\mu t}.$$

Beweis. Wir wollen zeigen, dass

$$P(a \leq T \leq b) = \int_a^b \mu \cdot e^{-\mu t} \, \mathrm{d}t$$

gilt. Hierzu beachte man, dass zum Zeitpunkt a das Ereignis A noch nicht, zum Zeitpunkt b jedoch *mindestens* einmal eingetreten ist (vergleiche Skizze).



Dies bedeutet,

$$X_a = 0, \quad \text{und} \quad X_b \geq 1.$$

Aus $[0, a] \subseteq [0, b]$ folgt

$$\begin{aligned} P(T \in [a, b]) = P(a \leq T \leq b) = P(X_a = 0) - P(X_b = 0) &= e^{-\mu a} - e^{-\mu b} \\ &= \int_a^b \mu e^{-\mu t} \,\mathrm{d}t. \end{aligned}$$

□

Beispiel 11.12 (Telefonzentrale) In einer Telefonzentrale kommen im Mittel pro Stunde 20 Anrufe an. Gesucht ist die Wahrscheinlichkeit dafür, dass 3 bis 6 Minuten zwischen zwei Anrufen vergehen. Legt man Minuten als verbindliche Zeiteinheit fest, so folgt aus

$$E(X_{60}) = \mu \cdot 60 \overset{!}{=} 20,$$

dass $\mu = \frac{1}{3}$ gilt. T sei die Zufallsgröße, welche die Zeit zwischen 2 Anrufen misst. Dann gilt

$$\begin{aligned} P(3 \leq T \leq 6) &= F(6) - F(3) \\ &= (1 - e^{-\mu \cdot 6}) - (1 - e^{-\mu \cdot 3}) \\ &= -e^{-2} + e^{-1} \\ &= 0.2325. \end{aligned}$$

△

Erwartungswert und Streuung der Exponentialverteilung: Gemäß Definition des Erwartungswertes gilt

$$E(X) = \int_0^\infty x \cdot \lambda e^{-\lambda x} \,\mathrm{d}x.$$

Substituieren wir $t = \lambda x$ so folgt

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \int_0^\infty \underset{u}{t} \cdot \underset{v'}{e^{-t}} \,\mathrm{d}t \\ &= \frac{1}{\lambda} \left\{ \underset{u}{t} \cdot \underset{v}{(-e^{-t})} \Big|_0^\infty - \int_0^\infty \underset{u'}{1} \cdot \underset{v}{(-e^{-t})} \,\mathrm{d}t \right\} \\ &= \frac{1}{\lambda} \left\{ -0 + 0 - e^{-t} \Big|_0^\infty \right\} \\ &= \frac{1}{\lambda} \{0 + 1\} = \frac{1}{\lambda}. \end{aligned}$$

In ähnlicher Weise folgt

$$\begin{aligned}
E(X^2) &= \int_0^\infty x^2 \cdot \lambda e^{-\lambda x}\,\mathrm{d}x \\
&= \frac{1}{\lambda^2}\int_0^\infty \underset{u}{t^2}\,\underset{v'}{e^{-t}}\,\mathrm{d}t \\
&= \frac{1}{\lambda^2}\left\{\underset{u}{t^2}\cdot \underset{v}{(-e^{-t})}\Big|_0^\infty - \int_0^\infty \underset{u'}{2t}\cdot \underset{v}{(-e^{-t})}\,\mathrm{d}t\right\} \\
&= \frac{1}{\lambda}\{-0+0+2\} = \frac{2}{\lambda^2},
\end{aligned}$$

womit sich

$$V(X) = E(X^2) - E^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

ergibt.

Exponentialverteilung als Lebensdauer-Verteilung: Wartezeiten, Reparaturzeiten und die Lebensdauer von Bauelementen können als exponentialverteilt angenommen werden. Allerdings ist dabei zu beachten, dass *keine* Alterungseffekte modelliert werden können, wie folgende Überlegung zeigt:

Es sei $X \sim \mathrm{Ex}(\lambda)$ die Lebensdauer eines Bauelements. Dann gilt

$$\begin{aligned}
P(X \le x+y \mid X \ge x) &= \frac{P(x \le X \le x+y)}{P(X \ge x)} \\
&= \frac{F(x+y) - F(x)}{1 - F(x)} \\
&= \frac{(1 - e^{-\lambda(x+y)}) - (1 - e^{-\lambda x})}{1 - (1 - e^{-\lambda x})} \\
&= \frac{e^{-\lambda x} - e^{-\lambda(x+y)}}{e^{-\lambda x}} \\
&= \frac{e^{-\lambda x} - e^{-\lambda x}e^{-\lambda y}}{e^{-\lambda x}} \\
&= 1 - e^{-\lambda y} = F(y) = P(X \le y).
\end{aligned}$$

Dies bedeutet, dass die Wahrscheinlichkeit dafür, dass ein intaktes Bauteil in den nächsten y Zeiteinheiten kaputtgeht, unabhängig von dessen Alter X immer gleich ist.

11.5 Normalverteilung

Die *Normalverteilung*, die auch als *Gaußsche Verteilung* bezeichnet wird, ist die wohl bekannteste stetige Verteilung. Ihre Dichtefunktion ist die allseits beliebte *Gaußsche Glockenkurve* .

Definition 11.13 (Normalverteilung) Eine stetige Zufallsgröße heißt **normalverteilt** mit den Parametern μ und $\sigma^2 > 0$, wenn ihre Dichtefunktion der Gleichung

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

genügt. Wir schreiben dann auch kurz

$$X \sim N(\mu, \sigma^2).$$

Die Verteilungsfunktion der $N(\mu, \sigma^2)$-Normalverteilung

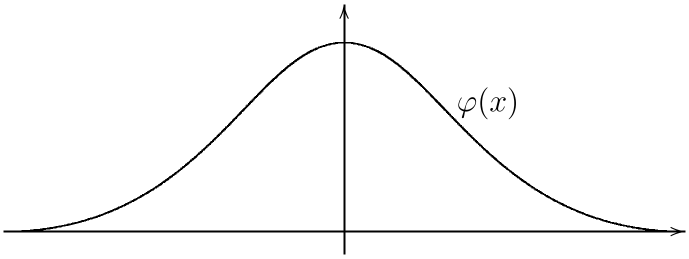
$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}t$$

ist i.a. nicht geschlossen auswertbar. Daher nutzt man Techniken der *Standardisierung*, um wenigstens im Spezialfall die tabellierten Werte der Verteilungsfunktion nutzen zu können.

Definition 11.14 (Standardisiert normalverteilt) Im Falle $X \sim N(0,1)$ bezeichnet man die Zufallsgröße X als **standardisiert normalverteilt**.

Für die Dichtefunktion einer standardisiert normalverteilten Zufallsgröße X gilt

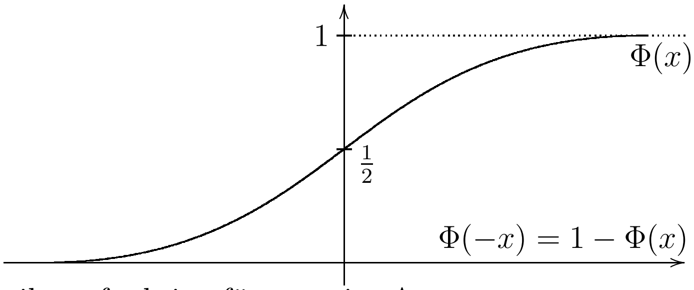
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty.$$



Die zugehörige Verteilungsfunktion

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,\mathrm{d}t$$

ist für $x \geq 0$ tabelliert.



Für die Berechnung der Werte der Verteilungsfunktion für negative Argumente nutzt man den aus der Symmetrie der Verteilung resultierenden Zusammenhang

$$\Phi(x) = 1 - \Phi(x).$$

Satz 11.15 Die standardisiert normalverteilte Zufallsgröße $X \sim N(0,1)$ besitzt den Erwartungswert $E(X) = 0$ und die Varianz $V(X) = 1$.

Beweis. Wegen $x\varphi(x) = -\varphi'(x)$ folgt

$$\int_{-\infty}^{\infty} |x|\varphi(x)\,\mathrm{d}x = 2\lim_{C\to\infty} \int_{C}^{C} \varphi'(x)\,\mathrm{d}x = 2\lim_{C\to\infty} \big(\varphi(0) - \varphi(C)\big) = 2\varphi(0) < \infty$$

und daher ist (die Funktion $x\varphi(x)$ ist punktsymmetrisch bezüglich 0)

$$E(X) = \int_{-\infty}^{\infty} x\varphi(x)\,\mathrm{d}x = 0.$$

Weiter ergibt sich mit partieller Integration

$$V(X) = E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot x e^{-\frac{x^2}{2}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \Bigg[\underbrace{-xe^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty}}_{=0} + \underbrace{\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x}_{=\sqrt{2\pi}} \Bigg] = 1.$$

Das letzte Integral rechnet man dabei wie folgt aus:

$$\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x \right)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y.$$

Verwenden wir nun Polarkoordinaten

$$x = r\cos\varphi, \qquad y = r\sin\varphi,$$

so ergibt sich schließlich die gesuchte Identität

$$\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x \right)^2 = \int_0^{2\pi}\int_0^{\infty} r \cdot e^{-\frac{r^2}{2}}\,\mathrm{d}r\,\mathrm{d}\varphi = \int_0^{2\pi} -e^{-\frac{x^2}{2}} \Big|_0^{\infty} \mathrm{d}\varphi = \int_0^{2\pi} 1\,\mathrm{d}\varphi = 2\pi.$$

□

Bemerkung: Die lineare Transformation

$$Y = \frac{X - E(X)}{\sqrt{V(X)}}$$

einer Zufallsgröße X heißt *Standardisierung* von X. Dabei gilt

$$E(Y) = E\left(\frac{X - E(X)}{\sqrt{V(X)}} \right) = \frac{1}{\sqrt{V(X)}} \big(E(X) - E(X) \big) = 0$$

und

$$V(Y) = V\left(\frac{X - E(X)}{\sqrt{V(X)}} \right) = \frac{1}{V(X)} V(X) = 1.$$

Satz 11.16 Falls $X \sim N(\mu, \sigma^2)$, so gilt $\big(X - E(X)\big)/\sqrt{V(X)} \sim N(0,1)$. Insbesondere gilt $E(X) = \mu$ und $V(X) = \sigma^2$.

Beweis. Sei $Y \sim N(0,1)$, dann folgt für $X := \sigma Y - \mu$ dass

$$P(a \le X \le b) = P(a \le \sigma Y - \mu \le b) = P\left(\frac{b-\mu}{\sigma} \le Y \le \frac{b-\mu}{\sigma} \right) = \int_{\frac{b-\mu}{\sigma}}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\,\mathrm{d}x.$$

Mit der Substitution

$$x := \sigma y + \mu$$

ergibt sich hieraus

$$P(a \leq X \leq b) = \int_b^a \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}y,$$

das heißt, $X \sim N(\mu, \sigma^2)$. Schließlich erhalten wir

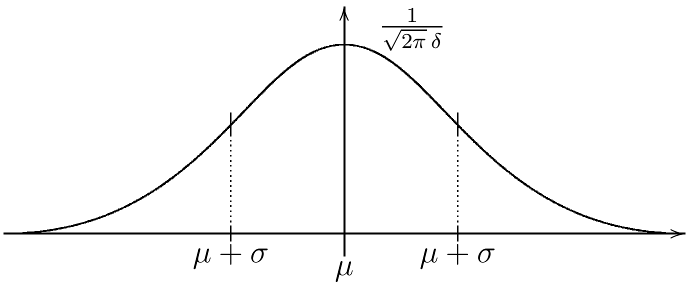
$$E(X) = E(\sigma Y + \mu) = \sigma E(Y) + \mu = \mu$$

und

$$V(X) = V(\sigma Y + \mu) = \sigma^2 V(Y) = \sigma^2.$$

□

Interpretation der Parameter μ und σ^2: Die Dichtefunktion der Normalverteilung ist eine Glockenkurve, die durch den Lageparameter μ und den Formparameter $\sigma^2 > 0$ charakterisiert wird.



Der Parameter μ gibt die Lage der Symmetrieachse an, während σ^2 die Breite der Glocke bestimmt. Die Glocke hat ihr Maximum in μ mit dem Wert $\frac{1}{\sqrt{2\pi}\sigma}$ und besitzt in $\mu \pm \sigma$ je einen Wendepunkt. Ist σ^2 groß, so erhalten wir eine breitgezogene Glockenkurve, für kleines σ^2 ergibt sich hingegen eine nadelförmige Glockenkurve.

Wahrscheinlichkeitsberechnung bei normalverteilten Zufallsgrößen: Die praktischen Berechnungen zur Normalverteilung erfordern in der Regel die Bestimmung von Wahrscheinlichkeiten des Typs $P(a \leq X \leq b)$ für $X \sim N(\mu, \sigma^2)$. Durch Ausnutzung des Standardisierungsgedankens führt man dies auf die Berechnung einer Differenz zweier Werte der Verteilungsfunktion Φ der Standardnormalverteilung zurück:

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \underbrace{\frac{X-\mu}{\sigma}}_{=:Y\sim N(0,1)} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Als normalverteilt können Zufallsgrößen angesehen werden, die durch Überlagerung einer großen Anzahl von Einflüssen entstehen, wobei jede Einflussgröße nur einen im Verhältnis zur Gesamtsumme unbedeutenden Beitrag liefert. Beispiele normalverteilter Zufallsgrößen sind:

- zufällige Beobachtungs- oder Messfehler,
- zufällige Abweichungen vom Nennmaß bei der Fertigung von Werkstücken,
- Effekte beim Prozess der Brownschen Molekularbewegung.

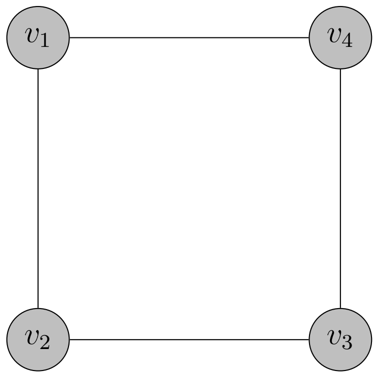
Beispiel 11.17 (Fertigungstoleranzen) Ein Werkstück besitzt die gewünschte Qualität, wenn die Abweichung eines bestimmten Maßes vom entsprechendem Nennmaß dem Betrage nach nicht größer als $3.6mm$ ist. Der Herstellungsprozess sei so beschaffen, dass dieses Maß als normalverteilte Zufallsgröße angesehen werden kann, wobei der Erwartungswert mit dem Nennmaß übereinstimmt. Weiter sei $\sigma = 3mm$ bekannt. Wieviel Prozent der Werkstücke einer Serie werden durchschnittlich mit gewünschter Qualität produziert?

Es sei
$$X \widehat{=} \text{“zufällige Abweichung vom Nennmaß”},$$
dann gilt
$$X \sim N(0,9).$$
Wegen
$$\begin{aligned} P(|X| \leq 3.6) &= P\Bigg(\underbrace{\frac{|X|}{3}}_{\sim N(0,1)} \leq \frac{3.6}{3}\Bigg) = \Phi(1.2) - \Phi(-1.2) = \Phi(1.2) - \big(1 - \Phi(1.2)\big) \\ &= 2\Phi(1.2) - 1 = 0.88493 \cdot 2 - 1 = 0.76983 \end{aligned}$$
genügen duchschnittlich etwa 77% aller Werkstücke den Qualitätsansprüchen. $\triangle$

12. Markov-Ketten

12.1 Grundlagen

Beispiel 12.1 (Simple Random Walk) Wir betrachten einen “Random Walker” in einer sehr kleinen Stadt, die nur vier Straßen und vier Straßenecken besitzt:



Zum Zeitpunkt 0 steht der Random Walker an der Ecke v_1. Zum Zeitpunkt 1 wirft er eine Münze und geht nach v_2 bzw. v_4, je nachdem ob er Zahl oder Kopf geworfen hat. Zum Zeitpunkt 2 wirft er wieder eine Münze, um zu entscheiden zu welcher nächstgelegenenen Straßenecke er nun geht. Diese Prozedur wird für die Zeitschritte $3, 4, \dots$ iteriert.
Für jedes $n \in \mathbb{N}_0$ sei

$$X_n \mathrel{\widehat{=}} \text{“Straßenecke zum Zeitpunkt } n\text{”}.$$

Dann gilt

$$P(X_0 = v_1) = 1$$

und

$$P(X_1 = v_2) = \frac{1}{2}, \qquad P(X_1 = v_4) = \frac{1}{2}.$$

Um die Wahrscheinlichkeitsverteilung für X_n mit $n \geq 2$ zu berechnen, ist es vorteilhaft, bedingte Wahrscheinlichkeiten zu betrachten. Angenommen, der Random Walker steht zum Zeitpunkt n an der Ecke v_2. Dann gilt

$$P(X_{n+1} = v_1 | X_n = v_2) = \frac{1}{2}, \qquad P(X_{n+1} = v_3 | X_n = v_2) = \frac{1}{2}.$$

Diese Wahrscheinlichkeiten ändern sich nicht, wenn wir die gesamte Vergangenheit in Betracht ziehen:

$$P(X_{n+1} = v_1 | X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = v_2) = \frac{1}{2}$$

und

$$P(X_{n+1} = v_3 | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = v_2) = \frac{1}{2}$$

für jede Wahl möglicher $i_0, i_1, \ldots, i_{n-1}$. Insbesondere ist die Wahrscheinlichkeit $P(X_{n+1} = v_i | X_n = v_j)$ unabhängig vom Zeitpunkt n. △

Definition 12.2 Es sei S eine endliche oder abzählbar unendliche Menge. Dann bildet die Folge $X_0, X_1, \ldots$ von Zufallsgrößen eine **Markov-Kette**, falls die **Markov-Eigenschaft** gilt: Für alle $n \in \mathbb{N}_0$ und für $s_0, s_1, \ldots, s_{n+1} \in S$ mit

$$P(X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) > 0$$

gilt

$$P(X_{n+1} = s_{n+1} | X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) = P(X_{n+1} = s_{n+1} | X_n = s_n).$$

Die Menge S wird **Zustandsraum** genannt.

Definition 12.3 Es sei $\mathbf{P} = [p_{i,j}]_{i,j=1}^k$ eine $(k \times k)$-Matrix. Eine Markov-Kette $(X_0, X_1, \ldots)$ mit endlichem Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ heißt **homogene Markov-Kette** mit **Übergangsmatrix P**, falls für alle $n \in \mathbb{N}_0$ und $i, j \in \{1, 2, \ldots, k\}$ gilt

$$P(X_{n+1} = s_j | X_n = s_i) = p_{i,j}.$$

Die Einträge von **P** heißen **Übergangswahrscheinlichkeiten**.

Bemerkung: Die Übergangsmatrix ist eine *stochastische Matrix*, das heißt, es gilt

$$p_{i,j} \geq 0 \quad \text{für alle} \quad i, j \in \{1, 2, \ldots, k\}$$

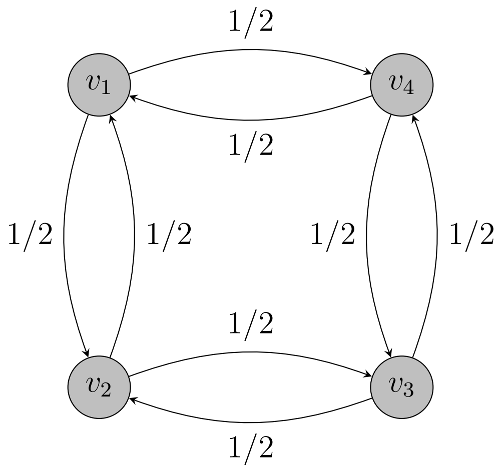
und

$$\sum_{j=1}^{k} p_{i,j} = 1 \quad \text{für alle} \quad i \in \{1, 2, \ldots, k\}.$$

Für unser Einführungsbeispiel ergibt sich beispielsweise die Übergangsmatrix

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

Neben der Übergangsmatrix ist der *Übergangsgraph* eine weitere sehr hilfreiche Darstellung von Markov-Ketten. Im Falle des Simple Random Walks erhalten wir:



Die Wahrscheinlichkeitsverteilung von X_n wollen wir nun als Zeilenvektor auffassen, also

$$\boldsymbol{\mu}^{(n)} = \left[\mu_1^{(n)}, \mu_2^{(n)}, \ldots, \mu_k^{(n)}\right] = \left[P(X_n = s_1), P(X_n = s_2), \ldots, P(X_n = s_k)\right].$$

Die Wahrscheinlichkeitsverteilung $\boldsymbol{\mu}^{(0)}$ von X_0 heißt auch *Startverteilung*.

Satz 12.4 Für die homogene Markov-Kette $(X_0, X_1, \ldots)$ mit endlichem Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$, Startverteilung $\boldsymbol{\mu}^{(0)}$ und Übergangsmatrix $\mathbf{P}$ gilt für jedes $n \in \mathbb{N}$

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{\mu}^{(0)}\mathbf{P}^n.$$

Beweis. Im Fall $n = 1$ gilt für alle $j \in \{1, 2, \ldots, k\}$

$$\begin{aligned}
\mu_j^{(1)} = P(X_1 = s_j) &= \sum_{i=1}^{k} P(X_0 = s_i, X_1 = s_j) \\
&= \sum_{i=1}^{k} \underbrace{P(X_0 = s_i)}_{=\mu_i^{(0)}} \cdot \underbrace{P(X_1 = s_j | X_0 = s_i)}_{=p_{i,j}} \\
&= \sum_{i=1}^{k} \mu_i^{(0)} p_{i,j} = [\boldsymbol{\mu}^{(0)}\mathbf{P}]_j.
\end{aligned}$$

Mit vollständiger Induktion folgt der Schritt $n \mapsto n + 1$:

$$\begin{aligned}
\mu_j^{(n+1)} = P(X_{n+1} = s_j) &= \sum_{i=1}^{k} P(X_n = s_i, X_{n+1} = s_j) \\
&= \sum_{i=1}^{k} \underbrace{P(X_n = s_i)}_{=\mu_i^{(n)}} \cdot \underbrace{P(X_{n+1} = s_j | X_n = s_i)}_{=p_{i,j}} \\
&= \sum_{i=1}^{k} \mu_i^{(n)} p_{i,j} = [\boldsymbol{\mu}^{(n)}\mathbf{P}]_j \overset{\text{Induktions-}}{\underset{\text{annahme}}{=}} [\boldsymbol{\mu}^{(0)}\mathbf{P}^{n+1}]_j.
\end{aligned}$$

□

Beispiel 12.5 (Bonner Wettermodell) Es wird gelegentlich behauptet, dass die beste Wettervorhersage diejenige ist, bei der das Wetter morgen dem heutigen entspricht. Ist diese Behauptung richtig, so ist es naheliegend, das Wetter als Markov-Kette zu modellieren. Wir nehmen einfachheitshalber an, dass es nur die Zustände s_1 = "Regen" und s_2 = "Sonne" gibt. Vorausgesetzt unsere Wettervorhersage ist mit der Wahrscheinlichkeit $p \in [0, 1]$ richtig, so folgt die Übergangsmatrix

$$\mathbf{P} = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}.$$

△

Beispiel 12.6 (Londoner Wettermodell) Wir haben im letzten Beispiel eine perfekte Symmetrie zwischen "Regen" und "Sonne". Dies mag in Bonn richtig sein, sicherlich jedoch nicht in London. Dort dürfte die Übergangsmatrix

$$\mathbf{P} = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}.$$

mit $q < p$ realistischer sein. △

Definition 12.7 Es sei $\mathbf{P}^{(1)} = [p_{i,j}^{(1)}]_{i,j=1}^k, \mathbf{P}^{(2)} = [p_{i,j}^{(2)}]_{i,j=1}^k, \ldots$ eine Folge von stochastischen $(k \times k)$-Matrizen. Eine Markov-Kette $(X_0, X_1, \ldots)$ mit endlichem Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ heißt **inhomogene Markov-Kette** mit Übergangsmatrizen $\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots$, falls für alle $n \in \mathbb{N}_0$ und $i, j \in \{1, 2, \ldots, k\}$ gilt

$$P(X_{n+1} = s_i | X_n = s_j) = p_{i,j}^{(n+1)}.$$

Beispiel 12.8 (verfeinertes Bonner Wettermodell) Wir verfeinern unser Bonner Wettermodell, indem wir zusätzlich zwischen "Winter" und "Sommer" unterscheiden. Hierzu erweitern wir den Zustandsraum um den Wert s_3 = "Schnee" und unterscheiden zwischen Sommer (Mai–Oktober) und Winter (November–April). Realistisch wäre beispielsweise die inhomogene Markov-Kette mit

$$\mathbf{P}_{\text{Sommer}} = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0.25 & 0.75 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

und

$$\mathbf{P}_{\text{Winter}} = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.15 & 0.7 & 0.15 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}.$$

Im Sommer entspricht das verfeinerte Modell dem alten Modell im Falle $p = 0.75$ außer eventuellem Tauwetter am 1. Mai. △

Satz 12.9 Für die inhomogene Markov-Kette $(X_0, X_1, \ldots)$ mit endlichem Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$, Startverteilung $\boldsymbol{\mu}^{(0)}$ und Übergangsmatrizen $\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots$ gilt für jedes $n \in \mathbb{N}$

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{\mu}^{(0)} \mathbf{P}^{(1)} \mathbf{P}^{(2)} \cdots \mathbf{P}^{(n)}.$$

Beweis. Die Aussage wird ähnlich zum Beweis von Satz 12.4 gezeigt. □

12.2 Irreduzible und aperiodische Markov-Ketten

Wir wollen uns im folgenden auf homogene Markov-Ketten beschränken. Dann ist die Wahrscheinlichkeit

$$P(X_{m+n} = s_j | X_m = s_i) = [\mathbf{P}^n]_{i,j}$$

unabhängig von m.

Definition 12.10 Es sei $(X_0, X_1, \ldots)$ eine homogene Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Wir sagen, der Zustand s_i **führt** zum Zustand s_j, kurz $s_i \to s_j$, falls ein $n > 0$ existiert, so dass

$$P(X_{m+n} = s_j | X_m = s_i) > 0.$$

Führt s_i zu s_j und s_j zu s_i, so sagen wir die Zustände s_i und s_j **kommunizieren**, kurz $s_i \leftrightarrow s_j$.

Aus

$$\mathbf{P}^{m+n} = \mathbf{P}^m \cdot \mathbf{P}^n$$

folgt sofort

$$\begin{aligned} P(X_{m+n} = s_j | X_0 = s_i) = [\mathbf{P}^{(m+n)}]_{i,j} &= \sum_{\ell=1}^{k} [\mathbf{P}^m]_{i,\ell} [\mathbf{P}^n]_{\ell,j} \\ &= \sum_{\ell=1}^{k} P(X_{m+n} = s_j | X_m = s_\ell) P(X_m = s_\ell | X_0 = s_i). \end{aligned}$$

Damit haben wir die Chapman-Kolmogorov-Gleichung bewiesen:

Satz 12.11 (Chapman-Kolmogorov-Gleichung) Ist $(X_1, X_2, \ldots)$ eine homogene Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und $\ell < m < n$, dann gilt für alle $s_h, s_i \in S$

$$P(X_n = s_j | X_\ell = s_h) = \sum_{i=1}^{k} P(X_n = s_j | X_m = s_i) P(X_m = s_i | X_\ell = s_h). \tag{12.1}$$

Bemerkung: Wie man sich leicht überlegt, gilt die Chapman-Kolmogorov-Gleichung auch im Fall inhomogener Markov-Ketten.

Aus (12.1) folgt nun für $\ell < m < n$ und $i \in \{1, 2, \ldots, k\}$, dass

$$P(X_n = s_j | X_\ell = s_h) \geq P(X_n = s_j | X_m = s_i) P(X_m = s_i | X_\ell = s_h),$$

beziehungsweise

$$s_h \to s_i \wedge s_i \to s_j \Rightarrow s_h \to s_j,$$

dies bedeutet, die Relation "$\to$" ist transitiv.

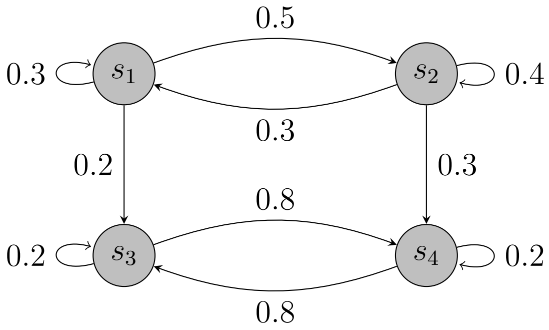
Definition 12.12 Eine homogene Markov-Kette $(X_0, X_1, \dots)$ mit Zustandsraum $S = \{s_1, s_2, \dots, s_k\}$ heißt **irreduzibel**, falls für alle $s_i, s_j \in S$ gilt $s_i \leftrightarrow s_j$, andernfalls heißt die Markov-Kette **reduzibel**.

Bemerkung: Eine homogene Markov-Kette ist genau dann irreduzibel, wenn der zugehörige Übergangsgraph stark zusammenhängend ist.

Beispiel 12.13 Betrachte die Markov-Kette $(X_0, X_1, \dots)$ mit Zustandsraum $S = \{s_1, s_2, \dots, s_k\}$ und Übergangsmatrix

$$\mathbf{P} = \begin{bmatrix} 0.3 & 0.5 & 0.2 & 0 \\ 0.3 & 0.4 & 0 & 0.3 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0.8 & 0.2 \end{bmatrix}.$$

Der zugehörige Übergangsgraph ist:



Falls wir mit s_3 oder s_4 starten, verhält sich die Markov-Kette genauso wie die Markov-Kette mit Zustandsraum $\{s_3, s_4\}$ und Übergangsmatrix

$$\begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}.$$

$\triangle$

Das Beispiel illustriert: Ist eine Markov-Kette reduzibel, dann kann die Analyse ihres Langzeitverhaltens auf eine oder mehrere Markov-Ketten mit kleinerem Zustandsraum reduziert werden.

Bemerkung: Eine Markov-Kette ist genau dann reduzibel, falls eine Permutationsmatrix $\mathbf{T} \in \mathbb{R}^{k \times k}$ und quadratische Matrizen $\mathbf{A}, \mathbf{C}$ existieren, so dass für ihre Übergangsmatrix gilt

$$\mathbf{TPT}^\star = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}.$$

Definition 12.14 Für einen Zustand $s_i \in S$ mit $s_i \to s_i$ heißt der größte gemeinsame Teiler der potenziellen Rückkehrzeiten

$$d(s_i) := \mathrm{ggT}\{n \geq 1 : [\mathbf{P}^n]_{i,i} > 0\}$$

die **Periode** des Zustands. Gilt nicht $s_i \to s_i$, so setzen wir $d(s_i) := \infty$. Zustände mit $d(s_i) = 1$ heißen **aperiodisch**. Die Markov-Kette heißt **aperiodisch**, wenn alle Zustände aperiodisch sind, ansonsten heißt sie **periodisch**.

Beispiel 12.15 Im Random Walk aus Beispiel 12.1 gilt $[\mathbf{P}^n]_{i,i} > 0$ nur für $n = 2, 4, 6, \ldots$. Daher ist $d(s_i) = 2$ für alle $i = 1, 2, 3, 4$ und folglich die Markov-Kette periodisch. Hingegen gilt für $p \neq 0$ beim Bonner Wettermodell (Beispiel 12.5) $[\mathbf{P}^n]_{i,i} > 0$ für alle $n \in \mathbb{N}$, das heißt, die zugehörige Markov-Kette ist aperiodisch. △

Wir benötigen folgendes Resultat aus der elementaren Zahlentheorie:

Lemma 12.16 Gegeben sei die Menge $A = \{a_1, a_2, \ldots\} \subseteq \mathbb{N}$ mit den Eigenschaften

1. $\operatorname{ggT}(a_1, a_2, \ldots) = 1$,
2. $a, b \in A \Rightarrow a + b \in A$.

Dann existiert eine Zahl $N < \infty$, so dass $n \in A$ gilt für alle $n \geq N$.

Beweis. *i.* Wir zeigen zunächst, dass für beliebige Zahlen $a, b \in \mathbb{N}$ Zahlen $x, y \in \mathbb{N}$ existieren mit

$$ax - by = \operatorname{ggT}(a, b).$$

Hierzu sei ohne Einschränkung der Allgemeinheit $\operatorname{ggT}(a, b) = 1$, denn sonst betrachte einfach $a/\operatorname{ggT}(a, b)$ und $b/\operatorname{ggT}(a, b)$. Definieren wir die $b-1$ Zahlen

$$\begin{aligned} z_1 &:= a \bmod b, \\ z_2 &:= 2a \bmod b, \\ &\ \ \vdots \\ z_{b-1} &:= (b-1)a \bmod b, \end{aligned}$$

so gilt $0 \leq z_i < b$ für alle i. Weiter gilt $z_i \neq 0$ für alle i, denn sonst gäbe es ein $p \in \mathbb{N}$ mit $a = pb$, das heißt, b teilt a im Widerspruch zu $\operatorname{ggT}(a, b) = 1$. Ist hingegen $z_i \neq 1$ für alle i, so muss es $0 < k < \ell < b$ geben mit

$$ka \bmod b = z_k = z_\ell = \ell a \bmod b.$$

Dann ist aber $(\ell - k)a \bmod b = 0$, das heißt, b teilt $(\ell - k)a$, was wegen $\operatorname{ggT}(a, b) = 1$ auf den Widerspruch b teilt $0 \neq \ell - k < b$ führt. Folglich gibt es ein $0 < k < b$ mit $z_k = ka \bmod b = 1$, oder anders ausgedrückt,

$$ka - \ell b = 1.$$

ii. Aussage *i* kann mittels vollständiger Induktion auf K Zahlen verallgemeinert werden. Denn gibt es $K-1$ Zahlen $n_1, n_2, \ldots, n_{K-1} \in \mathbb{Z}$ mit

$$b = \sum_{k=1}^{K-1} n_k a_k = \operatorname{ggT}(a_1, a_2, \ldots, a_{K-1}),$$

so existieren nach Aussage *i* $x, y \in \mathbb{N}$ mit

$$bx - a_K y = \operatorname{ggT}(b, a_K) = \operatorname{ggT}(a_1, a_2, \ldots, a_K).$$

Sezten wir $\widetilde{n}_k := x n_k \in \mathbb{Z}$ für alle $0 < k < K$ und $\widetilde{n}_K := -y \in \mathbb{Z}$, so folgt

$$\sum_{k=1}^{K} \widetilde{n}_k a_k = \operatorname{ggT}(a_1, a_2, \ldots, a_K).$$

iii. Da jedes $a \in A$ eine endliche Primfaktorzerlegung besitzt, gibt es ein $K < \infty$ mit

$$\operatorname{ggT}(a_1, a_2, \dots, a_K) = 1.$$

Gemäß Aussage *ii* existieren also Zahlen $n_1, n_2, \dots, n_K \in \mathbb{Z}$ mit

$$\sum_{k=1}^{K} n_k a_k = 1.$$

Setzen wir

$$L := \max\{|n_1|, |n_2|, \dots, |n_K|\}$$

und

$$N := La_1(a_1 + a_2 + \cdots + a_K),$$

dann gibt es zu jedem $n \geq N$ eine eindeutige Zerlegung

$$n = N + ka_1 + \ell$$

mit $k \geq 0$ und $0 \leq \ell < a_1$. Nun gilt aber

$$n = La_1(a_1 + a_2 + \cdots + a_K) + ka_1 + \ell \sum_{k=1}^{K} n_k a_k = \sum_{k=1}^{K} m_k a_k$$

mit nichtnegativen, ganzzahligen Koeffizienten

$$\begin{aligned} m_1 &= La_1 + k + n_1 \geq 0, \\ m_2 &= La_2 + n_2 \geq 0, \\ &\vdots \\ m_K &= La_K + n_K \geq 0, \end{aligned}$$

was wegen der Abgeschlossenheit von A bezüglich der Addition schließlich die Behauptung liefert. □

Satz 12.17 Gegeben sei eine aperiodische Markov-Kette $(X_0, X_1, \dots)$ mit Zustandsraum $S = \{s_1, s_2, \dots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Dann existiert ein $N < \infty$, so dass $[\mathbf{P}^n]_{i,i} > 0$ für alle $i \in \{1, 2, \dots, k\}$ und $n \geq N$.

Beweis. Für $s_i \in S$ setze

$$A_i := \{n \geq 1 : [\mathbf{P}^n]_{i,j} > 0\},$$

das heißt, A_i ist die Menge der Rückkehrzeiten zum Zustand s_i. Da gemäß Voraussetzung die Markov-Kette aperiodisch ist, ist $\operatorname{ggT}(A_i) = 1$. Weiter bedeutet $m, n \in A_i$ dass

$$P(X_m = s_i | X_0 = s_i) > 0, \qquad P(X_{m+n} = s_i | X_m = s_i) > 0,$$

und gemäß (12.1) ergibt sich

$$P(X_{m+n} = s_i | X_0 = s_i) \geq P(X_{m+n} = s_i | X_m = s_i) P(X_m = s_i | X_0 = s_i) > 0,$$

beziehungsweise $m + n \in A_i$. Damit lässt sich Lemma 12.16 anwenden und es existiert ein $N_i < \infty$ derart, dass $n \in A_i$ für alle $n \geq N_i$. Setzen wir $N = \max\{N_1, N_2, \dots, N_k\}$ erhalten wir schließlich das Behauptete. □

Korollar 12.18 Sei $(X_0, X_1, \ldots)$ eine aperiodische, irreduzible Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Dann existiert ein $M < \infty$, so dass $[\mathbf{P}^m]_{i,j} > 0$ für alle $i, j \in \{1, 2, \ldots, k\}$ und $m \geq M$.

Beweis. Nach Satz 12.17 existiert ein $N < \infty$, so dass $[\mathbf{P}^n]_{i,i} > 0$ für alle $n \geq N$. Für beliebige, feste $s_i, s_j \in S$ folgt aus der Irreduzibilität, dass ein $n_{i,j}$ existiert mit $[\mathbf{P}^{n_{i,j}}]_{i,j} > 0$. Setze $M_{i,j} := N + n_{i,j}$, dann gilt für alle $m \geq M$ (beachte: $m - n_{i,j} \geq N$)

$$P(X_m = s_j|X_0 = s_i) \overset{(12.1)}{\geq} P(X_m = s_j|X_{m-n_{i,j}} = s_i)P(X_{m-n_{i,j}} = s_i|X_0 = s_i) > 0,$$

Für $M = \max_{i,j}\{M_{i,j}\}$ folgt nun die Behauptung. □

12.3 Stationäre Verteilungen

Definition 12.19 Gegeben sei $(X_0, X_1, \ldots)$ eine Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und $X_0 = s_i \in S$. Dann heißt

$$T_{i,j} = \min\{n > 0 : X_n = s_j\}$$

die **Treffzeit** von s_i mit $s_j \in S$. $T_{i,j}$ ist also die Zeit, die nach Start in s_i vergeht bis erstmals s_j besucht wird. Ist $P(X_n = s_j|X_0 = s_i) = 0$ für alle $n \in \mathbb{N}$, so setzen wir $T_{i,j} = \infty$.

Lemma 12.20 Sei $(X_0, X_1, \ldots)$ eine aperiodische, irreduzible Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Dann gilt

$$P(T_{i,j} < \infty) = 1$$

und

$$E(T_{i,j}) < \infty.$$

Beweis. Nach Korollar 12.18 existiert ein $M < \infty$ mit $[\mathbf{P}^M]_{i,j} > 0$ für alle $i, j \in \{1, 2, \ldots, k\}$. Setze

$$\alpha := \min\left\{[\mathbf{P}^M]_{i,j} : i, j \in \{1, 2, \ldots, k\}\right\} > 0$$

und wähle $s_i, s_j \in S$ beliebig. Sartet die Markov-Kette in s_i, dann folgt

$$P(T_{i,j} > M) \leq P(X_M \neq s_j|X_0 = s_i) \leq 1 - \alpha,$$

und, da

$$P(X_{2M} = s_j|T_{i,j} > M) = \sum_{\substack{k=1\\ i \neq k}}^{K} P(X_{2M} = s_j|X_M = s_k) \geq \alpha,$$

auch

$$\begin{aligned} P(T_{i,j} > 2M) &= P(T_{i,j} > M)P(T_{i,j} > 2M|T_{i,j} > M) \\ &\leq \underbrace{P(T_{i,j} > M)}_{\leq 1-\alpha}\underbrace{P(X_{2M} \neq s_j|T_{i,j} > M)}_{\leq 1-\alpha} \leq (1-\alpha)^2. \end{aligned}$$

Für beliebiges $\ell \in \mathbb{N}$ ergibt sich

$$\begin{aligned}P(T_{i,j} > \ell M) &= P\big(T_{i,j} > (\ell-1)M\big) P\big(T_{i,j} > \ell M | T_{i,j} > (\ell-1)M\big) \\ &\leq \underbrace{P(T_{i,j} > (\ell-1)M)}_{\leq (1-\alpha)^{\ell-1}} \underbrace{P\big(X_{\ell M} \neq s_j | T_{i,j} > (\ell-1)M\big)}_{\leq 1-\alpha} \leq (1-\alpha)^\ell \xrightarrow{\ell\to\infty} 0,\end{aligned}$$

dies bedeutet, $P(T_{i,j} = \infty) = 0$. Weiter folgt

$$\begin{aligned}E(T_{i,j}) &= \sum_{n=1}^{\infty} nP(T_{i,j} = n) = \sum_{n=1}^{\infty}\sum_{m=1}^{n} P(T_{i,j} = n) = \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} P(T_{i,j} = n) \\ &= \sum_{m=0}^{\infty} P(T_{i,j} > m) = \sum_{\ell=0}^{\infty} \sum_{m=\ell M}^{(\ell+1)M-1} \underbrace{P(T_{i,j} > m)}_{\leq P(T_{i,j} > \ell M) \leq (1-\alpha)^\ell} \\ &\leq M \sum_{\ell=0}^{\infty} (1-\alpha)^\ell = \frac{M}{1-(1-\alpha)} = \frac{M}{\alpha} < \infty.\end{aligned}$$

□

Definition 12.21 Sei $(X_0, X_1, \ldots)$ eine Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Ein Zeilenvektor $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_k]$ heißt **stationäre Verteilung** der Markov-Kette, falls gilt

1. $\pi_i \geq 0$ für alle $i \in \{1, 2, \ldots, k\}$ und $\sum_{i=1}^{k} \pi_i = 1$, und
2. $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, das heißt $\sum_{i=1}^{k} \pi_i p_{i,j} = \pi_j$ für alle $j \in \{1, 2. \ldots, k\}$.

Bemerkung: Eine stationäre Verteilung $\boldsymbol{\pi}$ ist also ein Linkseigenvektor von $\mathbf{P}$ zum Eigenwert 1.

Satz 12.22 (Existenz einer stationären Verteilung) Für eine aperiodische, irreduzible Markov-Kette $(X_0, X_1, \ldots)$ mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$ existiert mindestens eine stationäre Verteilung.

Beweis. Wir nehmen an, die Markov-Kette startet in s_1 und setzen für jedes $i \in \{1, 2, \ldots, k\}$

$$\rho_i := \sum_{n=0}^{\infty} P(X_n = s_i, T_{1,1} > n) \leq E(T_{1,1}) < \infty.$$

Wir wählen

$$\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_k] = \left[\frac{\rho_1}{E(T_{1,1})}, \frac{\rho_2}{E(T_{1,1})}, \ldots, \frac{\rho_k}{E(T_{1,1})}\right]$$

und zeigen zunächst $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. Für $j \neq 1$ gilt

$$\begin{aligned}
\rho_j &= \sum_{n=0}^{\infty} P(X_n \in s_j, T_{1,1} > n) \\
&\overset{j\neq 1}{=} \sum_{n=1}^{\infty} P(X_n = s_j, T_{1,1} > n) \\
&\overset{j\neq 1}{=} \sum_{n=1}^{\infty} P(X_n = s_j, T_{1,1} > n-1) \\
&= \sum_{n=1}^{\infty} \sum_{i=1}^{k} P(X_{n-1} = s_i, X_n = s_j, T_{1,1} > n-1) \\
&\overset{j\neq 1}{=} \sum_{n=1}^{\infty} \sum_{i=1}^{k} P(X_{n-1} = s_i, T_{1,1} > n-1) P(X_n = s_j | X_{n-1} = s_i, T_{1,1} > n-1).
\end{aligned}$$

Da das Ereignis $\{T_{1,1} > n-1\}$ einzig bestimmt ist durch $X_0, X_1, \ldots, X_{n-1}$, folgt

$$P(X_n = s_j | X_{n-1} = s_i, T_{1,1} > n-1) = P(X_n = s_j | X_{n-1} = s_i) = p_{i,j}$$

und weiter

$$\begin{aligned}
\rho_j &= \sum_{n=1}^{\infty} \sum_{i=1}^{k} p_{i,j} P(X_{n-1} = s_i, T_{1,1} > n-1) \\
&= \sum_{i=1}^{k} p_{i,j} \sum_{n=1}^{\infty} P(X_{n-1} = s_i, T_{1,1} > n-1) \\
&= \sum_{i=1}^{k} p_{i,j} \sum_{m=0}^{\infty} P(X_m = s_i, T_{1,1} > m) \\
&= \sum_{i=1}^{k} \rho_i p_{i,j},
\end{aligned}$$

beziehungsweise

$$\pi_j = \frac{\rho_j}{E(T_{1,1})} = \sum_{i=1}^{k} \frac{\rho_i}{E(T_{1,1})} p_{i,j} = \sum_{i=1}^{k} \pi_i p_{i,j}.$$

Für $j = 1$ ergibt sich

$$\begin{aligned}
\rho_1 = 1 = P(T_{1,1} < \infty) &= \sum_{n=1}^{\infty} P(T_{1,1} = n) \\
&= \sum_{n=1}^{\infty} \sum_{i=1}^{k} P(X_{n-1} = s_i, T_{1,1} = n) \\
&= \sum_{n=1}^{\infty} \sum_{i=1}^{k} P(X_{n-1} = s_i, X_n = s_1, T_{1,1} > n-1) P(X_n = s_1 | X_{n-1} = s_i, T_{1,1} > n-1).
\end{aligned}$$

Verwenden wir wieder, dass das Ereignis $\{T_{1,1} > n-1\}$ nur von $X_0, X_1, \dots, X_{n-1}$ abhängt, so folgt

$$\begin{aligned}
\rho_1 &= \sum_{i=1}^{k} p_{i,1} \sum_{n=1}^{\infty} P(X_{n-1} = s_i, T_{1,1} > n-1) \\
&= \sum_{i=1}^{k} p_{i,1} \sum_{m=0}^{\infty} P(X_m = s_i, T_{1,1} > m) \\
&= \sum_{i=1}^{k} \rho_i p_{i,1},
\end{aligned}$$

das heißt

$$\pi_1 = \frac{\rho_1}{E(T_{1,1})} = \sum_{i=1}^{k} \frac{\rho_i}{E(T_{1,1})} p_{i,1} = \sum_{i=1}^{k} \pi_i p_{i,1}.$$

Damit ist $\boldsymbol{\pi}$ ein Linkseigenvektor von $\mathbf{P}$ zum Eigenwert 1. Es verbleibt zu zeigen, dass gilt $\sum_{i=1}^{k} \pi_i = 1$. Dies folgt jedoch sofort aus

$$\begin{aligned}
E(T_{1,1}) &= \sum_{n=0}^{\infty} P(T_{1,1} > n) \\
&= \sum_{n=0}^{\infty} \sum_{i=1}^{k} P(X_n = s_i, T_{1,1} > n) \\
&= \sum_{i=1}^{k} \sum_{n=0}^{\infty} P(X_n = s_i, T_{1,1} > n) \\
&= \sum_{i=1}^{k} \rho_i.
\end{aligned}$$

$\square$

Bemerkung: Die stationäre Verteilung besitzt sogar die Eigenschaft, dass alle $\pi_i > 0$ sind. Denn wären etwa $\pi_1, \pi_2, \dots, \pi_\ell = 0$ und $\pi_{\ell+1}, \pi_{\ell+2}, \dots, \pi_k \neq 0$, so folgt $p_{i,j} = 0$ für alle $i, j \in \{\ell+1, \ell+2, \dots, k\}$. Die Markov-Kette wäre also reduzibel.

Satz 12.23 (Asymptotik) Sei $(X_0, X_1, \dots)$ eine aperiodische, irreduzible Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \dots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Dann gilt für jede Startverteilung $\boldsymbol{\mu}^{(0)}$ und jede stationäre Verteilung $\boldsymbol{\pi}$, dass

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{\mu}^{(0)} \mathbf{P}^n \stackrel{n\to\infty}{\longrightarrow} \boldsymbol{\pi}.$$

Beweis. Es genügt die Behauptung für alle $\boldsymbol{\mu}^{(0)} = \mathbf{e}_i^\star$, $i \in \{1, 2, \dots, k\}$ zu zeigen. Seien hierzu $(X_0, X_1, \dots)$ und $(Y_0, Y_1, \dots)$ die Markov-Ketten, welche in $X_0 = s_i$ bzw. mit Startverteilung $\boldsymbol{\pi}$ starten. Beide Markov-Ketten sind unabhängig voneinander und beeinflussen sich nicht gegenseitig.

Wir betrachten auch eine dritte Markov-Kette $(Z_1, Z_2, \dots)$ welche aus den Paaren $Z_\ell = (X_\ell, Y_\ell)$ besteht. Ihr Zustandsraum ist also $S \times S$ und die Übergangswahrscheinlichkeiten sind

$$q_{(m,m'),(n,n')} = p_{m,n}p_{m',n'}.$$

Die Irreduzibilität impliziert, dass ein $N < \infty$ existiert mit $[\mathbf{P}^N]_{m,n} > 0$ für alle $m, n \in \{1, 2, \ldots, k\}$. Damit ist aber auch $[\mathbf{P}^N]_{i,k}[\mathbf{P}^N]_{j,n} > 0$, und folglich auch die Markov-Kette $(Z_0, Z_1, \ldots)$ irreduzibel. Offensichtlich ist sie auch aperiodisch, so dass aus Lemma 12.20 für die Zufallsgröße

$$T := \min\{n > 0 : X_n = Y_n\}$$

folgt

$$\begin{aligned} P(T = \infty) &= \sum_{j=0}^{k} P(T = \infty | Y_0 = s_j) P(Y_0 = s_j) \\ &= \sum_{j=0}^{k} P\big(T_{(i,j),(1,1)} = \infty, T_{(i,j),(2,2)} = \infty, \ldots, T_{(i,j),(k,k)} = \infty\big) P(Y_0 = s_j) \\ &= 0, \end{aligned}$$

das heißt $P(T < \infty) = 1$.

Nun ist T die Zeit, die vergeht, bis das Verhalten der Markov-Ketten $(X_0, X_1, \ldots)$ und $(Y_0, Y_1, \ldots)$ übereinstimmt, denn es für alle $n \geq T$ alle $j \in \{1, 2, \ldots, k\}$ gilt

$$P(X_n = s_j | n \geq T) = P(Y_n = s_j | n \geq T).$$

Multiplikation mit $P(n \geq T)$ liefert dann

$$P(X_n = s_j, n \geq T) = P(Y_n = s_j, n \geq T). \tag{12.2}$$

Wir haben

$$\pi_j = P(Y_n = s_j) = P(Y_n = s_j, n \geq T) + \underbrace{P(Y_n = s_j, n < T)}_{\leq P(n<T) \overset{n\to\infty}{\longrightarrow} 0},$$

dies bedeutet,

$$P(Y_n = s_j, n \geq T) \overset{n\to\infty}{\longrightarrow} \pi_j.$$

Aus (12.2) folgt damit

$$P(X_n = s_j, n \geq T) \overset{n\to\infty}{\longrightarrow} \pi_j.$$

Wegen

$$P(X_n = s_j) = \underbrace{P(X_n = s_j, n \geq T)}_{\overset{n\to\infty}{\longrightarrow} \pi_j} + \underbrace{P(X_n = s_j, n < T)}_{\leq P(n<T) \overset{n\to\infty}{\longrightarrow} 0} \overset{n\to\infty}{\longrightarrow} \pi_j$$

folgt schließlich die Behauptung. □

Satz 12.24 (Eindeutigkeit der stationären Verteilung) Für eine aperiodische, irreduzible Markov-Kette $(X_0, X_1, \ldots)$ mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$ existiert genau eine stationäre Verteilung.

Beweis. Es seien $\boldsymbol{\pi}_1$ und $\boldsymbol{\pi}_2$ zwei stationäre Verteilungen. Nach Satz 12.23 folgt

$$\|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\| = \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\mathbf{P}^n\| \overset{n\to\infty}{\longrightarrow} \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_1\| = 0,$$

das heißt $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$. □

Definition 12.25 Sei $(X_0, X_1, \ldots)$ eine aperiodische, irreduzible Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Eine Verteilung $\boldsymbol{\pi}$ heißt **reversibel**, falls für alle $i, j \in \{1, 2, \ldots, k\}$ gilt

$$\pi_i p_{i,j} = \pi_j p_{j,i}.$$

Satz 12.26 Sei $(X_0, X_1, \ldots)$ eine Markov-Kette mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$. Ist die Verteilung $\boldsymbol{\pi}$ reversibel, dann ist sie auch stationär.

Beweis. Ist $\boldsymbol{\pi}$ reversibel, so folgt für jedes $j \in \{1, 2, \ldots, k\}$

$$\sum_{i=1}^{k} \pi_i p_{i,j} = \sum_{i=1}^{k} \pi_j p_{j,i} = \pi_j \underbrace{\sum_{i=1}^{k} p_{j,i}}_{=1} = \pi_j,$$

mit anderen Worten,

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}.$$

□

12.4 Markov-Ketten-Monte-Carlo-Verfahren

Um eine Markov-Kette $(X_0, X_1, \ldots)$ mit Zustandsraum $S \in \{s_1, s_2, \ldots, s_k\}$ und Übergangsmatrix $\mathbf{P}$ numerisch zu simulieren, benötigen wir eine *Startfunktion* und eine *Updatefunktion*.

Zu gegebener Startverteilung

$$\boldsymbol{\mu}^{(0)} = [\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}],$$

wählen wir die Startfunktion

$$\Psi(x) = \begin{cases} s_1, & x \in [0, \mu_1^{(0)}), \\ s_2, & x \in [\mu_1^{(0)}, \mu_1^{(0)} + \mu_2^{(0)}), \\ & \vdots \\ s_j, & x \in \left[\sum_{\ell=1}^{j-1} \mu_\ell^{(0)}, \sum_{\ell=1}^{j} \mu_\ell^{(0)}\right), \\ & \vdots \\ s_k, & x \in \left[\sum_{\ell=1}^{k-1} \mu_\ell^{(0)}, 1\right]. \end{cases} \tag{12.3}$$

Für eine auf $[0, 1]$ gleichverteilte Zufallsgröße Z gilt dann

$$P\big(\Psi(Z) = s_i\big) = P\left(\sum_{\ell=1}^{i-1} \mu_\ell^{(0)} \leq Z < \sum_{\ell=1}^{i} \mu_\ell^{(0)}\right) = \mu_i^{(0)},$$

dies bedeutet, $X_0 = \Psi(Z)$ besitzt die gewünschte Startverteilung $\boldsymbol{\mu}^{(0)}$.

Die Updatefunktion liefert für $n \in \mathbb{N}$ den Schritt

$$\boldsymbol{\mu}^{(n)} \mapsto \boldsymbol{\mu}^{(n+1)} = \boldsymbol{\mu}^{(n)}\mathbf{P},$$

das heißt, für gegebenes $X_n = s_i$ die Verteilung

$$P(X_{n+1} = s_j | X_n = s_i) = p_{i,j}, \quad j \in \{1, 2, \ldots, k\}.$$

Sie ist gegeben durch

$$\Phi(x, s_i) = \begin{cases} s_1, & x \in [0, p_{i,1}), \\ s_2, & x \in [p_{i,1}, p_{i,1} + p_{i,2}), \\ & \vdots \\ s_j, & x \in \big[\sum_{\ell=1}^{j-1} p_{i,\ell}, \sum_{\ell=1}^{j} p_{i,\ell}\big), \\ & \vdots \\ s_k, & x \in \big[\sum_{\ell=1}^{k-1} p_{i,\ell}, 1\big]. \end{cases} \tag{12.4}$$

Wie man sich leicht überlegt, folgt für eine auf $[0, 1]$-gleichverteilte Zufallsgröße Z, dass

$$P(X_{n+1} = s_j | X_n = s_i) = P\big(\Phi(Z, s_i) = s_j\big) = p_{i,j}.$$

Bemerkung: Im Fall einer inhomogenen Markov-Kette ist diese Funktion zeitabhängig; wir haben dann für jedes $n \in \mathbb{N}$ eine andere Updatefunktion $\Phi^{(n)}(x, s_i)$.

Wir erhalten schließlich den folgenden Algorithmus:

Algorithmus 12.27 (Monte-Carlo-Simulation von Markov-Ketten)
input: Startverteilung $\boldsymbol{\mu}^{(0)}$ und Übergangsmatrix $\mathbf{P}$
output: Realisierung $(x_0, x_1, \ldots)$ der Markov-Kette $(X_0, X_1, \ldots)$
① Initialisierung: für eine Zufallszahl $r \in [0, 1]$ bestimme $x_0 = \Psi(r)$ gemäß (12.3)
② für alle $n \in \mathbb{N}$: für eine Zufallszahl $r \in [0, 1]$ bestimme $x_n = \Phi(r, x_{n-1})$ gemäß (12.4)

Wir wollen uns nun dem Problem zuwenden, eine Markov-Kette zu konstruieren, die eine vorgegebene stationäre Verteilung $\boldsymbol{\pi}$ besitzt. Dies ist vor allem dann interessant, falls der Zustandsraum sehr groß ist und man nur die stationäre Verteilung $\boldsymbol{\pi}$ samplen will.

Dazu sei $G = (S, E)$ ein ungerichteter, zusammenhängender Graph mit dem Zustandsraum als Knotenmenge. Den Knotengrad bezeichnen wir abkürzend mit $d_i := |\operatorname{post}(s_i)|$. Betrachte die Übergangsmatrix $\mathbf{P} = [p_{i,j}]_{i,j=1}^k$ mit

$$p_{i,j} = \begin{cases} \frac{1}{d_i} \min\big\{\frac{\pi_j d_i}{\pi_i d_j}, 1\big\}, & \text{falls } (s_i, s_j) \in E, \\ 1 - \sum_{s_j \in \operatorname{post}(s_i)} \frac{1}{d_i} \min\big\{\frac{\pi_j d_i}{\pi_i d_j}, 1\big\}, & \text{falls } i = j, \\ 0, & \text{sonst}. \end{cases} \tag{12.5}$$

Diese Übergangsmatrix gehört zum folgenden Übergangsmechanismus: Angenommen es ist $X_n = s_i$, dann bestimme gleichverteilt einen Nachbarn $s_j \in \operatorname{post}(s_i)$ (jeder Nachbar wird mit Wahrscheinlichkeit $1/d_i$ ausgewählt). Setze

$$X_{n+1} = \begin{cases} s_j \text{ mit Wahrscheinlichkeit } \min\big\{\frac{\pi_j d_i}{\pi_i d_j}, 1\big\}, \\ s_i \text{ mit Wahrscheinlichkeit } 1 - \min\big\{\frac{\pi_j d_i}{\pi_i d_j}, 1\big\}. \end{cases}$$

Satz 12.28 Die Markov-Kette $(X_0, X_1, \ldots)$ mit Zustandsraum $S = \{s_1, s_2, \ldots, s_k\}$ und Übergangwahrscheinlichkeiten (12.5) ist irreduzibel und besitzt die stationäre Verteilung $\boldsymbol{\pi}$.

Beweis. Da der ungerichtete Graph $G = (S, E)$ zusammenhängend ist, ist der Übergangsgraph der Markov-Kette stark zusammenhängend (denn jede Kante ist vorwärts und rückwärts vorhanden) und somit die Markov-Kette irreduzibel.

Wir wollen nun zeigen, dass die Verteilung $\boldsymbol{\pi}$ reversibel ist, also für alle $i, j \in \{1, 2, \ldots, k\}$ gilt

$$\pi_i p_{i,j} = \pi_j p_{j,i}.$$

Im Fall $i = j$ oder $(s_i, s_j) \notin E$ ist dies klar. Sei also $i \neq j$ und $(s_i, s_j) \in E$. Gilt $\frac{\pi_j d_i}{\pi_i d_j} \geq 1$, dann folgt

$$\pi_i p_{i,j} = \pi_i \frac{1}{d_i} \underbrace{\min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}}_{=1} = \frac{\pi_i}{d_i} = \pi_j \frac{1}{d_j} \underbrace{\min\left\{\frac{\pi_i d_j}{\pi_j d_i}, 1\right\}}_{=\frac{\pi_i d_j}{\pi_j d_i}} = \pi_j p_{j,i}.$$

Ist hingegen $\frac{\pi_j d_i}{\pi_i d_j} < 1$, so ergibt sich

$$\pi_i p_{i,j} = \pi_i \frac{1}{d_i} \underbrace{\min\left\{\frac{\pi_j d_i}{\pi_i d_j}, 1\right\}}_{=\frac{\pi_j d_i}{\pi_i d_j}} = \frac{\pi_i}{d_i} = \pi_j \frac{1}{d_j} \underbrace{\min\left\{\frac{\pi_i d_j}{\pi_j d_i}, 1\right\}}_{=1} = \pi_j p_{j,i}.$$

Gemäß Satz 12.26 ist folglich $\boldsymbol{\pi}$ die stationäre Verteilung der Markov-Kette. □

Die Simulation läuft sehr schnell, wenn der Graph dünn ist, das heißt, der Knotengrad klein ist. Dies bedeutet insbesondere, dass die Übergangsmatrix $\mathbf{P}$ dünnbesetzt ist, also viele Nulleinträge besitzt.

Algorithmus 12.29 (Metropolis-Algorithmus)

input: ungerichteter, zusammenhängender Graph $G = (S, E)$ und Startknoten $s \in S = \{1, 2, \ldots, k\}$

output: Realisierung $(x_0, x_1, \ldots)$ der Markov-Kette $(X_0, X_1, \ldots)$

① Initialisierung: setzte $x_0 := s$ und $n := 0$

② bestimme gleichverteilt einen Nachbarknoten $j \in \text{post}(x_n)$

③ setze $x_{n+1} := j$ mit Wahrscheinlichkeit $\min\left\{\frac{\pi_j d_{x_n}}{\pi_{x_n} d_j}, 1\right\}$
oder $x_{n+1} := x_n$ mit Wahrscheinlichkeit $1 - \min\left\{\frac{\pi_j d_{x_n}}{\pi_{x_n} d_j}, 1\right\}$

④ erhöhe $n := n + 1$ und gehe nach ②

13. Polynominterpolation

13.1 Lagrange-Interpolation

Im folgenden bezeichne

$$\Pi_n = \left\{ \sum_{i=0}^{n} a_i x^i : a_0, a_1, \ldots, a_n \in \mathbb{K} \right\}$$

den Raum der Polynome bis zum Grad n.

Lagrangesche Interpolationsaufgabe: Gegeben seien $n+1$ paarweise verschiedene Knoten $x_0, x_1, \ldots, x_n$ sowie $n{+}1$ Werte $y_0, y_1, \ldots, y_n$, kurz $n{+}1$ Stützstellen (x_0, y_0), (x_1, y_1), $\ldots, (x_n, y_n)$. Gesucht ist ein Polynom $p \in \Pi_n$ mit

$$p(x_i) \stackrel{!}{=} y_i, \qquad i = 0, 1, \ldots, n. \tag{13.1}$$

Ohne Einschränkung der Allgemeinheit gelte $x_0 < x_1 < \ldots < x_n$.

Definition 13.1 Wir bezeichnen als **Knotenpolynom**

$$w(x) = \prod_{j=0}^{n} (x - x_j) \in \Pi_{n+1}$$

und als **Lagrange-Grundpolynome** die Polynome

$$L_i(x) = \frac{w(x)}{(x - x_i) w'(x_i)} = \prod_{j=0,\, j \neq i}^{n} \frac{x - x_j}{x_i - x_j} \in \Pi_n.$$

Beachte: Es gilt offensichtlich

$$L_i(x_j) = \delta_{i,j}. \tag{13.2}$$

Satz 13.2 Die Interpolationsaufgabe (13.1) hat genau eine Lösung

$$p = \sum_{i=0}^{n} y_i L_i.$$

Beweis. Wegen (13.2) gilt

$$p(x_j) = \sum_{i=0}^{n} y_i L_i(x_j) = y_j,$$

also (13.1). Damit ist die Existenz einer Lösung gesichert. Seien $p, q \in \Pi_n$ zwei Lösungen der Interpolationsaufgabe (13.1). Dann folgt aus

$$(p-q)(x_i) = 0, \qquad i = 0, 1, \ldots, n,$$

dass das Polynom $p - q \in \Pi_n$ $n+1$ Nullstellen besitzt. Daraus folgt jedoch $p \equiv q$. □

Satz 13.3 Sei $f \in C^{n+1}$ und $p \in \Pi_n$ das Interpolationspolynom zu den Knoten $\{x_i\}_{i=0}^n$ und Werten $\{y_i = f(x_i)\}_{i=0}^n$. Dann existiert zu jedem x ein $\xi \in I := \operatorname{conv}\{x, x_0, x_1, \ldots, x_n\}$ mit

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x). \tag{13.3}$$

Beweis. Gilt $x = x_i$ für ein $0 \le i \le n$, so ist (13.3) gültig mit $\xi = x$. Sei also $x \notin \{x_i\}_{i=0}^m$. Dann hat

$$h(t) := f(t) - p(t) - \frac{w(t)}{w(x)}\big(f(x) - p(x)\big)$$

Nullstellen in $x_0, x_1, \ldots, x_n$ sowie für $t = x$. Damit enthält jedes der $n+1$ Teilintervalle zwischen diesen Nullstellen nach dem Satz von Rolle eine Nullstelle $\tau_i^{(1)}$ von h', $i = 1, 2, \ldots, n+1$. So fortfahrend erhält man $n, n-1, \ldots$ Nullstellen $\tau_i^{(k)}$ von $h^{(k)}$, $k = 2, 3, \ldots$ im Intervall I. Insbesondere ergibt sich eine Nullstelle $\xi = \tau_1^{(n+1)}$ von $h^{(n+1)}$ in I. Wegen $p \in \Pi_n$ ist $p^{(n+1)} \equiv 0$, und wegen

$$w^{(n+1)}(t) = \left(\frac{d}{dt}\right)^{n+1} (t^{n+1} + \ldots) \equiv (n+1)!,$$

folgt hieraus

$$0 = h^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(n+1)!}{w(x)}\big(f(x) - p(x)\big),$$

beziehungsweise

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} w(x).$$

□

Achtung: Als Verfahren der *numerischen Approximation* ist die Polynominterpolation nur mit großen Einschränkungen brauchbar, also etwa bei wenigen Interpolationspunkten, dies heißt, kleinem Polynomgrad. Andernfalls hat man das Problem großer Oszillationen. Ein berühmtes Beispiel von Runge (1901) besagt, dass die Polynome p_{2m}, die die Funktion $f(x) = 1/(1+25x^2)$ in den Punkten $\pm i/m$ $(i = 0, 1, \ldots, m)$ interpolieren, auf $[-1, 1]$ nicht punktweise gegen f konvergieren.

Bemerkung: Bei der *Hermite-Interpolation* werden einzelne Gitterknoten mehrfach zugelassen. Tritt beispielsweise der Knoten x_i k-mal auf, so werden neben dem Funktionswert y_i an der Stelle $x = x_i$ noch die ersten $k-1$ Ableitungen des Interpolationspolynoms vorgeschrieben:

$$p(x_i) = y_i,\ p'(x_i) = y_i',\ \ldots,\ p^{(k-1)}(x_i) = y_i^{(k-1)}.$$

Auch diese Interpolationsaufgabe ist eindeutig lösbar und Satz 13.3 gilt entsprechend: Im Knotenpolynom w treten mehrfache Knoten x_i dann mit entsprechender Vielfachheit auf.

13.2 Neville-Schema

Die Bestimmung des Interpolationspolynoms mit Hilfe Lagrangescher Grundpolynome ist instabil und teuer. Will man das Interpolationspolynom nur an einigen wenigen Stellen auswerten, so bietet sich das Neville-Schema an.

Definition 13.4 Zu den $n+1$ Stützstellen $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ bezeichne $f_{i,i+1,\ldots,i+j}$, $0 \le i \le i+j \le n$, dasjenige (eindeutig bestimmte) Polynom vom Grad $\le j$ mit der Eigenschaft

$$f_{i,i+1,\ldots,i+j}(x_k) = y_k, \qquad k = i, i+1, \ldots, i+j. \tag{13.4}$$

Satz 13.5 Für die Interpolationspolynome $f_{i,i+1,\ldots,i+j}$ mit $0 \le i \le i+j \le n$ aus (13.4) gilt die Rekursionsformel

$$f_i(x) \equiv y_i, \quad j = 0,$$

$$f_{i,i+1,\ldots,i+j}(x) = \frac{(x-x_i)f_{i+1,i+2,\ldots,i+j}(x) - (x-x_{i+j})f_{i,i+1,\ldots,i+j-1}(x)}{x_{i+j}-x_i}, \quad j \ge 1. \tag{13.5}$$

Beweis. Den Beweis führen wir mittels vollständiger Induktion über j. Für $j = 0$ ist die Aussage für alle i wegen $f_i \in \Pi_0$ und $f_i(x_i) = y_i$ offensichtlich richtig. Für den Induktionsschritt $j-1 \mapsto j$ bezeichne $q(x)$ die rechte Seite von (13.5), und $q = f_{i,i+1,\ldots,i+j}$ ist dann nachzuweisen, was im folgenden geschieht. Es gilt $f_{i+1,i+2,\ldots,i+j}, f_{i,i+1,\ldots,i+j-1} \in \Pi_{j-1}$ und demnach $q \in \Pi_j$. Weiter gilt

$$q(x_i) = \frac{0 - (x_i - x_{i+j})f_{i,i+1,\ldots,i+j-1}(x_i)}{x_{i+j}-x_i} = \frac{0 - (x_i - x_{i+j})y_i}{x_{i+j}-x_i} = y_i,$$

$$q(x_{i+j}) = \frac{(x_{i+j}-x_i)f_{i+1,i+2,\ldots,i+j}(x) - 0}{x_{i+j}-x_i} = \frac{(x_{i+j}-x_i)y_{i+j} - 0}{x_{i+j}-x_i} = y_{i+j},$$

und für $k = i+1, i+2, \ldots, i+j-1$ gilt

$$\begin{aligned} q(x_k) &= \frac{(x_k - x_i)f_{i+1,i+2,\ldots,i+j}(x_k) - (x_k - x_{i+j})f_{i,i+1,\ldots,i+j-1}(x_k)}{x_{i+j}-x_i} \\ &= \frac{(x_k - x_i)y_k - (x_k - x_{i+j})y_k}{x_{i+j}-x_i} = \frac{-x_i y_k + x_{i+j}y_k}{x_{i+j}-x_i} = y_k. \end{aligned}$$

Aufgrund der Eindeutigkeit des Interpolationspolynoms gilt daher notwendigerweise die Identität $q = f_{i,i+1,\ldots,i+j}$. $\square$

Die sich für die Werte $f_{i,i+1,\ldots,i+j}(x)$ aus der Rekursionsformel (13.5) ergebenden Abhängigkeiten sind im nachfolgendem *Neville-Schema* dargestellt:

$$
\begin{array}{ccccc}
f_0(x) = y_0 & & & & \\
f_1(x) = y_1 & f_{0,1}(x) & & & \\
f_2(x) = y_2 & f_{1,2}(x) & f_{0,1,2}(x) & & \\
\vdots & \vdots & \ddots & & \\
f_{n-1}(x) = y_{n-1} & f_{n-2,n-1}(x) & \cdots & f_{0,1,\ldots,n-1}(x) & \\
f_n(x) = y_n & f_{n-1,n}(x) & \cdots & f_{1,2,\ldots,n}(x) & f_{0,1,\ldots,n}(x)
\end{array}
$$

Die Einträge lassen sich spaltenweise jeweils von oben nach unten berechnen. Das resultierende Verfahren wird zur Auswertung des Interpolationspolynoms an einzelnen Stellen x verwendet. Wie man leicht nachzählt, fallen dabei jeweils $3n^2/2 + \mathcal{O}(n)$ Multiplikationen an.

Beispiel 13.6 Man betrachte folgende Stützpunkte

i	0	1	2
x_i	0	1	3
y_i	1	3	2

Für $x = 2$ ergibt die Auswertung des Interpolationspolynoms gemäß dem Neville-Schema:

$$
\begin{array}{lll}
f_0(2) = y_0 = 1 & & \\
f_1(2) = y_1 = 3 & f_{0,1}(2) = \frac{(2-0)\cdot 3-(2-1)\cdot 1}{1-0} = 5 & \\
f_2(2) = y_2 = 2 & f_{1,2}(2) = \frac{(2-1)\cdot 2-(2-3)\cdot 3}{3-1} = \frac{5}{2} & f_{0,1,2}(2) = \frac{(2-0)\cdot\frac{5}{2}-(2-3)\cdot 5}{3-0} = \frac{10}{3}
\end{array}
$$

△

13.3 Newtonsche Interpolationsformel

Definition 13.7 Zu gegebenen paarweise verschiedenen $n+1$ Knoten $x_0, x_1, \ldots, x_n \in \mathbb{K}$ sind die $n+1$ **Newtonsche Basispolynome** folgendermaßen erklärt:

$$1, \quad x - x_0, \quad (x-x_0)(x-x_1), \quad \ldots, \quad (x-x_0)(x-x_1)\cdots(x-x_n) \tag{13.6}$$

Das gesuchte interpolierende Polynom $p \in \Pi_n$ soll nun als Linearkombination der Newtonschen Basispolynome dargestellt werden, also in der Form

$$p(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots + a_n(x-x_0)(x-x_1)\cdots(x-x_{n-1})$$

mit noch zu bestimmenden Koeffizienten $a_0, a_1, \ldots, a_n \in \mathbb{K}$. Sind diese Koeffizienten erst einmal bestimmt, so kann für jede Zahl $x = \xi$ das Polynom $p(x)$ mit dem *Horner-Schema*

$$p(\xi) = \Big(\cdots\Big(\big(a_n(\xi - x_{n-1}) + a_{n-1}\big)(\xi - x_{n-2}) + a_{n-2}\Big)(\xi - x_{n-3}) + \cdots + a_1\Big)(\xi - x_0) + a_0$$

ausgewertet werden, wobei die (insgesamt $3n$) Operationen von links nach rechts auszuführen sind.

Bemerkung: Die Koeffizienten $a_0, a_1, \ldots, a_n \in \mathbb{K}$ können aus den Interpolationsbedingungen

$$\begin{aligned}
y_0 &= p(x_0) = a_0, \\
y_1 &= p(x_1) = a_0 + a_1(x_1 - x_0), \\
y_2 &= p(x_2) = a_0 + a_1(x_1 - x_0) + a_2(x_2 - x_1)(x_1 - x_0), \\
&\vdots
\end{aligned}$$

gewonnen werden, wobei bei einer naiven Berechnung allerdings $n^3/6 + \mathcal{O}(n^2)$ Multiplikationen anfallen, wie man sich leicht überlegt. Im folgenden soll eine günstigere Vorgehensweise vorgestellt werden, die eine Berechnung dieser Koeffizienten mit $\mathcal{O}(n^2)$ Operationen ermöglicht.

Definition 13.8 Zu gegebenen Stützstellen $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ sind die **dividierten Differenzen** erklärt gemäß

$$\begin{aligned}
f[x_i] &\equiv y_i, \qquad i = 0, 1, 2, \ldots, n, \\
f[x_i, x_{i+1}, \ldots, x_{i+j}] &= \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+j}] - f[x_i, x_{i+1}, \ldots, x_{i+j-1}]}{x_{i+j} - x_i}, \\
& \qquad 0 \le i < i + j \le n.
\end{aligned} \tag{13.7}$$

Für die Berechnung aller dividierten Differenzen zu den Stützstellen $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ sind lediglich $n(n+1)/2$ Multiplikationen erforderlich. Die Abhängigkeiten zwischen den dividierten Differenzen sind im folgenden Schema dargestellt:

$$\begin{array}{ccccccccc}
f[x_0] = y_0 & \searrow & & & & & & & \\
f[x_1] = y_1 & \longrightarrow & f[x_0, x_1] & \searrow & & & & & \\
 & \searrow & & & & & & & \\
f[x_2] = y_2 & \longrightarrow & f[x_1, x_2] & \longrightarrow & f[x_0, x_1, x_2] & & & & \\
\vdots & & \vdots & & \ddots & & & & \\
f[x_{n-1}] = y_{n-1} & \longrightarrow & f[x_{n-2}, x_{n-1}] & \longrightarrow \cdots \longrightarrow & f[x_0, \ldots, x_{n-1}] & \searrow & & & \\
 & \searrow & & \searrow \quad \searrow & & & & & \\
f[x_n] = y_n & \longrightarrow & f[x_{n-1}, x_n] & \longrightarrow \cdots \longrightarrow & f[x_1, \ldots, x_n] & \longrightarrow & f[x_0, \ldots, x_n] & &
\end{array}$$

Satz 13.9 (Newtonsche Interpolationsformel) Für das Interpolationspolynom $p \in \Pi_n$ zu gegebenen Stützstellen $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ gilt

$$\begin{aligned} p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned} \tag{13.8}$$

Beweis. Wir bemerken zunächst, dass der führende Koeffizient des Polynoms $p(x)$ aus (13.8) durch $f[x_0, x_1, \ldots, x_n]$ gegeben ist, das heißt

$$p(x) = f[x_0, x_1, \ldots, x_n]x^n + \ldots .$$

Wir wollen den Satz mit vollständiger Induktion beweisen. Im Fall $n = 0$ ist die Aussage klar. Wir nehmen nun an, dass die Darstellung (13.8) für $n - 1$ gilt. Für die Interpolationspolynome $f_{0,1,\ldots,n-1}, f_{1,2,\ldots,n} \in \Pi_{n-1}$ folgt dann nach Induktionsannahme

$$\begin{aligned} f_{0,1,\ldots,n-1}(x) &= f[x_0, x_1, \ldots, x_{n-1}]x^{n-1} + \ldots, \\ f_{1,2,\ldots,n}(x) &= f[x_1, x_2, \ldots, x_n]x^{n-1} + \ldots . \end{aligned}$$

Das Interpolationspolynom $f_{0,1,\ldots,n} \in \Pi_n$ lässt sich schreiben als

$$f_{0,1,\ldots,n}(x) = c(x - x_0)(x - x_1) \cdots (x - x_{n-1}) + q(x), \qquad c \in \mathbb{K}, \qquad q \in \Pi_{n-1}.$$

Können wir zeigen, dass $c = f[x_0, x_1, \ldots, x_n]$ und $q = f_{0,1,\ldots,n-1}$ gilt, dann ergibt sich die Behauptung wegen $p = f_{0,1,\ldots,n}$. Aus

$$y_i = f_{0,1,\ldots,n}(x_i) = c\,\underbrace{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{n-1})}_{=0} + q(x_i), \qquad i = 0, 1, 2, \ldots, n - 1,$$

folgt tatsächlich $q = f_{0,1,\ldots,n-1}$. Gemäß (13.5) gilt ferner

$$\begin{aligned} f_{0,1,\ldots,n}(x) &= \frac{(x - x_0)f_{1,2,\ldots,n}(x) - (x - x_n)f_{0,1,\ldots,n-1}(x)}{x_n - x_0} \\ &= \frac{(x - x_0)(f[x_1, \ldots, x_n]x^{n-1} + \ldots) - (x - x_n)(f[x_0, \ldots, x_{n-1}]x^{n-1} + \ldots)}{x_n - x_0} \\ &= \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0} x^n + \ldots, \end{aligned}$$

dies bedeutet

$$c = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0} \overset{(13.7)}{=} f[x_0, x_1, \ldots, x_n].$$

Damit ist der Satz bewiesen. □

Beispiel 13.10 Zu den Stützpunkten aus Beispiel 13.6 wollen wir das entsprechende Interpolationspolynom in der Newton-Basis bestimmen. Es gilt

$$\begin{array}{lll} f[x_0] = y_0 = 1 & & \\ f[x_1] = y_1 = 3 & f[x_0, x_1] = \frac{3-1}{1-0} = 2 & \\ f[x_2] = y_2 = 2 & f[x_1, x_2] = \frac{2-3}{3-1} = -\frac{1}{2} & f[x_0, x_1, x_2] = \frac{-\frac{1}{2}-2}{3-0} = -\frac{5}{6} \end{array}$$

Damit ergibt sich

$$p(x) = 1 + 2x - \frac{5}{6}x(x-1).$$

△

14. Trigonometrische Interpolation

14.1 Theoretische Grundlagen

Die trigonometrische Interpolation benutzt man zur Analyse periodischer Funktionen $f(x) = f(x + 2\pi)$. Gegeben seien n Stützstellen $(x_0, y_0), (x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$, wobei wir der Einfachheit halber sogar annehmen wollen, dass die Knoten $\{x_i\}_{i=0}^{n-1}$ äquidistant verteilt sind, das heißt

$$x_k = \frac{2\pi k}{n}, \qquad k = 0, 1, \ldots, n-1.$$

Da die trignometrischen Funktionen $\cos(kx)$ und $\sin(kx)$ für ganzzahliges k alle die Periode 2π besitzen, liegt es nahe, geeignete Linearkombinationen von n dieser Funktionen zur Interpolation der n Stützpunkte zu verwenden. Es stellt sich als zweckmäßig heraus, einen trigonometrischen Ausdruck der Form

$$\begin{aligned} q(x) &= \frac{a_0}{2} + \sum_{\ell=1}^{m} \big(a_\ell \cos(\ell x) + b_\ell \sin(\ell x)\big), && n = 2m+1 \\ q(x) &= \frac{a_0}{2} + \sum_{\ell=1}^{m-1} \big(a_\ell \cos(\ell x) + b_\ell \sin(\ell x)\big) + \frac{a_m}{2}\cos(mx), && n = 2m \end{aligned} \tag{14.1}$$

zu nehmen.

Die Formeln werden bei komplexer Rechnung durchsichtiger:

Trigonometrische Interpolationsaufgabe: Gesucht ist ein *trigonometrisches Polynom* vom Grad n

$$p(x) = \beta_0 + \beta_1 e^{ix} + \beta_2 e^{2ix} + \cdots + \beta_{n-1} e^{(n-1)ix} \tag{14.2}$$

mit

$$p(x_k) \stackrel{!}{=} y_k, \qquad k = 0, 1, \ldots, n-1. \tag{14.3}$$

Bemerkung: Dass (14.2) äquivalent zu den Ausdrücken (14.1) ist, sieht man leicht aus der Moivreschen Formel

$$e^{i\ell x} = \cos(\ell x) + i \sin(\ell x)$$

und der Definition der x_k,

$$e^{-i\ell x_k} = e^{-2\pi i \ell k/n} = e^{2\pi i(n-\ell)k/n} = e^{i(n-\ell)x_k},$$

so dass

$$\cos(\ell x_k) = \frac{e^{i\ell x_k} + e^{i(n-\ell)x_k}}{2}, \qquad \sin(\ell x_k) = \frac{e^{i\ell x_k} - e^{i(n-\ell)x_k}}{2i}. \tag{14.4}$$

Ersetzt man in (14.1) für $x = x_k$ die Terme $\cos(\ell x_k)$ und $\sin(\ell x_k)$ durch (14.4) und ordnet man die Summanden nach Potenzen von e^{ix_k} um, findet man für $n = 2m + 1$ den Zusammenhang

$$\begin{aligned} &\beta_0 = \frac{a_0}{2}, &&\beta_k = \frac{1}{2}(a_k - ib_k), &&\beta_{n-k} = \frac{1}{2}(a_k + ib_k), &&k = 1, 2, \ldots, m, \\ &a_0 = 2\beta_0, &&a_\ell = \beta_\ell + \beta_{n-\ell}, &&b_\ell = i(\beta_\ell - \beta_{n-\ell}), &&\ell = 1, 2, \ldots, m, \end{aligned}$$

und für $n = 2m$

$$\begin{aligned} &\beta_0 = \frac{a_0}{2}, &&\beta_k = \frac{1}{2}(a_k - ib_k), &&\beta_{n-k} = \frac{1}{2}(a_k + ib_k), &&k = 1, 2, \ldots, m-1, &&\beta_m = \frac{a_m}{2}, \\ &a_0 = 2\beta_0, &&a_\ell = \beta_\ell + \beta_{n-\ell}, &&b_\ell = i(\beta_\ell - \beta_{n-\ell}), &&\ell = 1, 2, \ldots, m-1, &&a_m = 2\beta_m. \end{aligned}$$

Beachte: Es gilt entsprechend der Herleitung $p(x_k) = q(x_k)$, $k = 0, 1, 2, \ldots, n-1$, im allgemeinen jedoch nicht $p(x) = q(x)$ für $x \neq x_k$. Damit sind p und q nur in dem Sinne äquivalent, dass aus der Darstellung (14.2) sofort die Darstellung (14.1) folgt, und umgekehrt.

Satz 14.1 Zu beliebigen Werten $y_k \in \mathbb{K}$ gibt es genau ein trigonometrisches Polynom, das die Interpolationsaufgabe (14.3) löst.

Beweis. Substituieren wir in (14.2) $\omega = e^{ix}$ und $\omega_k = e^{ix_k}$, so stellen wir fest, dass die Interpolationsaufgabe (14.3) äquivalent ist zu: suche

$$r(\omega) = \sum_{\ell=0}^{n-1} \beta_\ell \omega^\ell \in \Pi_{n-1}$$

mit $r(\omega_k) \overset{!}{=} y_k$, $k = 0, 1, \ldots, n-1$. Diese Aufgabe ist aber gemäß Satz 13.2 eindeutig lösbar. □

Die Koeffizienten β_ℓ können explizit angegeben werden. Eine wesentliche Rolle spielt dabei die n-te komplexe Einheitswurzel

$$\omega_n := e^{2\pi i/n}.$$

Hierfür gelten insbesondere die Rechenregeln

$$\overline{\omega_n^k} = \overline{e^{2\pi ik/n}} = e^{-2\pi ik/n} = \omega_n^{-k}$$

und

$$\omega_n^k \omega_n^\ell = e^{2\pi ik/n} e^{2\pi i\ell/n} = e^{2\pi i(k+\ell)/n} = \omega_n^{k+\ell}.$$

Lemma 14.2 Die Vektoren

$$\boldsymbol{\omega}^k := \begin{bmatrix} \omega_n^0 \\ \omega_n^k \\ \omega_n^{2k} \\ \vdots \\ \omega_n^{(n-1)k} \end{bmatrix}, \qquad k = 0, 1, \ldots, n-1$$

bilden eine Othogonalbasis im $\mathbb{C}^n$

$$(\boldsymbol{\omega}^\ell)^\star \boldsymbol{\omega}^k = \begin{cases} n & \text{für } k = \ell, \\ 0 & \text{für } k \neq \ell. \end{cases}$$

Beweis. Es gilt

$$(\boldsymbol{\omega}^\ell)^\star \boldsymbol{\omega}^k = \sum_{m=0}^{n-1} \overline{\omega_n^{\ell m}} \omega_n^{km} = \sum_{m=0}^{n-1} \omega_n^{-\ell m} \omega_n^{km} = \sum_{m=0}^{n-1} \omega_n^{m(k-\ell)}.$$

Im Falle $k = \ell$ ergibt sich

$$(\boldsymbol{\omega}^\ell)^\star \boldsymbol{\omega}^k = \sum_{m=0}^{n-1} \omega_n^0 = \sum_{m=0}^{n-1} 1 = n.$$

Ist $k \neq \ell$, so folgt mit der Summenformel für die geometrische Reihe

$$(\boldsymbol{\omega}^\ell)^\star \boldsymbol{\omega}^k = \sum_{m=0}^{n-1} e^{2\pi i m(k-\ell)/n} = \frac{e^{2\pi i n(k-\ell)/n} - 1}{e^{2\pi i(k-\ell)/n} - 1} = \frac{1-1}{e^{2\pi i(k-\ell)/n} - 1} = 0.$$

□

Satz 14.3 Das trigonometrische Polynom $p(x) = \sum_{\ell=0}^{n-1} \beta_\ell e^{i\ell x}$ erfüllt die Interpolationsbedingung (14.3) genau dann, wenn

$$\beta_\ell = \frac{1}{n} \sum_{m=0}^{n-1} y_m \omega_n^{-m\ell} = \frac{1}{n} \sum_{m=0}^{n-1} y_m e^{-2\pi i \ell m/n}, \qquad \ell = 0, 1, \ldots, n-1.$$

Beweis. Für den Vektor

$$\mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} \in \mathbb{K}^n$$

folgt aus (14.3)

$$\mathbf{y} = \sum_{k=0}^{n-1} \beta_k \boldsymbol{\omega}^k,$$

so dass

$$\sum_{m=0}^{n-1} y_m \omega_n^{-m\ell} = (\boldsymbol{\omega}^\ell)^\star \mathbf{y} = (\boldsymbol{\omega}^\ell)^\star \left(\sum_{k=0}^{n-1} \beta_k \boldsymbol{\omega}^k \right) = \sum_{k=0}^{n-1} \beta_k \underbrace{(\boldsymbol{\omega}^\ell)^\star \boldsymbol{\omega}^k}_{= n\delta_{k,\ell}} = n\beta_\ell.$$

□

Wir kehren nun zu den trigonometrischen Ausdrücken (14.1) zurück.

Satz 14.4 Die trigonometrischen Ausdrücke (14.1) genügen den Interpolationsbedingungen $q(x_k) = y_k$, $k = 0, 1, 2, \ldots, n-1$, genau dann, wenn für ihre Koeffizienten gilt

$$a_\ell = \frac{2}{n}\sum_{m=0}^{n-1} y_m \cos(m x_\ell) = \frac{2}{n}\sum_{m=0}^{n-1} y_m \cos\left(\frac{2\pi k m}{n}\right),$$
$$b_\ell = \frac{2}{n}\sum_{m=0}^{n-1} y_m \sin(m x_\ell) = \frac{2}{n}\sum_{m=0}^{n-1} y_m \sin\left(\frac{2\pi k m}{n}\right).$$

Beweis. Für die Koeffizienten a_ℓ, b_ℓ gilt

$$\begin{aligned} a_\ell = \beta_\ell + \beta_{n-\ell} &= \frac{1}{n}\sum_{m=0}^{n-1} y_m(\omega_n^{-\ell m} + \omega_n^{(\ell-n)m}) = \frac{1}{n}\sum_{m=0}^{n-1} y_m(e^{-2\pi i\ell m/n} + e^{2\pi i(\ell-n)m/n}) \\ &= \frac{2}{n}\sum_{m=0}^{n-1} y_m \frac{e^{-2\pi i\ell m/n} + e^{2\pi i\ell m/n}}{2} = \frac{2}{n}\sum_{m=0}^{n-1} y_m \cos(m x_\ell) \end{aligned}$$

und

$$\begin{aligned} b_\ell = i(\beta_\ell - \beta_{n-\ell}) &= \frac{i}{n}\sum_{m=0}^{n-1} y_m(\omega_n^{-\ell m} - \omega_n^{(\ell-n)m}) = \frac{i}{n}\sum_{m=0}^{n-1} y_m(e^{-2\pi i\ell m/n} - e^{2\pi i(\ell-n)m/n}) \\ &= -\frac{2}{n}\sum_{m=0}^{n-1} y_m \frac{e^{-2\pi i\ell m/n} - e^{2\pi i\ell m/n}}{2i} = \frac{2}{n}\sum_{m=0}^{n-1} y_m \sin(m x_\ell). \end{aligned}$$

□

14.2 Schnelle Fourier-Transformation

Schreiben wir

$$\mathbf{T}_n = \left[\, \boldsymbol{\omega}^0 \,\middle|\, \boldsymbol{\omega}^1 \,\middle|\, \boldsymbol{\omega}^2 \,\middle|\, \ldots \,\middle|\, \boldsymbol{\omega}^{n-1} \,\right] = \begin{bmatrix} \omega_n^0 & \omega_n^0 & \omega_n^0 & \ldots & \omega_n^0 \\ \omega_n^0 & \omega_n^1 & \omega_n^2 & \ldots & \omega_n^{n-1} \\ \omega_n^0 & \omega_n^2 & \omega_n^4 & \ldots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ \omega_n^0 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \ldots & \omega_n^{(n-1)(n-1)} \end{bmatrix}$$

und

$$\mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{bmatrix},$$

so bedeutet Satz 14.3, dass

$$\boldsymbol{\beta} = \frac{1}{n}\mathbf{T}_n^\star \mathbf{y}.$$

Da ferner Lemma 14.2 die Beziehung

$$\mathbf{T}_n^\star \mathbf{T}_n = n\mathbf{I}$$

impliziert, folgt umgekehrt

$$\mathbf{y} = \mathbf{T}_n \boldsymbol{\beta}.$$

Definition 14.5 Die Abbildung

$$\mathbf{y} \mapsto \boldsymbol{\beta} = \frac{1}{n} \mathbf{T}_n^{\star} \mathbf{y}$$

wird **diskrete Fourier-Transformation** genannt. Ihre Umkehrung

$$\boldsymbol{\beta} \mapsto \mathbf{y} = \mathbf{T}_n \boldsymbol{\beta}$$

heißt **Fourier-Synthese**.

Bemerkung: Da $\mathbf{T}_n$ symmetrisch ist, folgt

$$\boldsymbol{\beta} = \frac{1}{n} \mathbf{T}_n^{\star} \mathbf{y} = \frac{1}{n} \overline{\mathbf{T}_n} \mathbf{y} = \frac{1}{n} \overline{\mathbf{T}_n \overline{\mathbf{y}}}.$$

Da folglich die Fourier-Transformation mit Hilfe der Fourier-Synthese berechenbar ist, werden wir uns im folgenden auf letztere beschränken.

Bemerkung: Die Fourier-Synthese entspricht der Auswertung eines trigonometrischen Polynoms $p(x)$ an den Stellen $x_k = 2\pi k/n$, denn

$$y_k = p(x_k) = \sum_{\ell=0}^{n-1} \beta_\ell e^{2\pi i k \ell / n} = \sum_{\ell=0}^{n-1} \beta_\ell \omega_n^{k\ell}, \qquad k = 0, 1, 2, \ldots, n-1.$$

Bei naiver Anwendung von $\mathbf{T}_n$ sind n^2 Multiplikationen auszuführen. Die *schnelle Fourier-Transformation* (auch *FFT* für *Fast Fourier Transform* genannt) entwickelt von Cooley und Tukey (1965), benötigt jedoch nur $\frac{1}{2} n \log_2 n$ Multiplikationen. Anhand eines Beispiels wollen wir die Vorgehensweise bei der schnellen Fourier-Transformation motivieren.

Beispiel 14.6 Es sei $n = 8$, dann berechnet sich die Fourier-Synthese gemäß

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_7 \end{bmatrix} = \begin{bmatrix} \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 \\ \omega_8^0 & \omega_8^1 & \omega_8^2 & \omega_8^3 & \omega_8^4 & \omega_8^5 & \omega_8^6 & \omega_8^7 \\ \omega_8^0 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8^0 & \omega_8^2 & \omega_8^4 & \omega_8^6 \\ \omega_8^0 & \omega_8^3 & \omega_8^6 & \omega_8^1 & \omega_8^4 & \omega_8^7 & \omega_8^2 & \omega_8^5 \\ \omega_8^0 & \omega_8^4 & \omega_8^0 & \omega_8^4 & \omega_8^0 & \omega_8^4 & \omega_8^0 & \omega_8^4 \\ \omega_8^0 & \omega_8^5 & \omega_8^2 & \omega_8^7 & \omega_8^4 & \omega_8^1 & \omega_8^6 & \omega_8^3 \\ \omega_8^0 & \omega_8^6 & \omega_8^4 & \omega_8^2 & \omega_8^0 & \omega_8^6 & \omega_8^4 & \omega_8^2 \\ \omega_8^0 & \omega_8^7 & \omega_8^6 & \omega_8^5 & \omega_8^4 & \omega_8^3 & \omega_8^2 & \omega_8^1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_7 \end{bmatrix}$$

Wir ordnen nun die rechte Seite nach geraden und ungeraden Indizes:

$$\left[\begin{array}{c} y_0 \\ y_1 \\ y_2 \\ y_3 \\ \hline y_4 \\ y_5 \\ y_6 \\ y_7 \end{array}\right] = \left[\begin{array}{cccc|cccc} \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 \\ \omega_8^0 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8^1 & \omega_8^3 & \omega_8^5 & \omega_8^7 \\ \omega_8^0 & \omega_8^4 & \omega_8^0 & \omega_8^4 & \omega_8^2 & \omega_8^6 & \omega_8^2 & \omega_8^6 \\ \omega_8^0 & \omega_8^6 & \omega_8^4 & \omega_8^2 & \omega_8^3 & \omega_8^1 & \omega_8^7 & \omega_8^5 \\ \hline \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^0 & \omega_8^4 & \omega_8^4 & \omega_8^4 & \omega_8^4 \\ \omega_8^0 & \omega_8^2 & \omega_8^4 & \omega_8^6 & \omega_8^5 & \omega_8^7 & \omega_8^1 & \omega_8^3 \\ \omega_8^0 & \omega_8^4 & \omega_8^0 & \omega_8^4 & \omega_8^6 & \omega_8^2 & \omega_8^6 & \omega_8^2 \\ \omega_8^0 & \omega_8^6 & \omega_8^4 & \omega_8^2 & \omega_8^7 & \omega_8^5 & \omega_8^3 & \omega_8^1 \end{array}\right] \left[\begin{array}{c} \beta_0 \\ \beta_2 \\ \beta_4 \\ \beta_6 \\ \hline \beta_1 \\ \beta_3 \\ \beta_5 \\ \beta_7 \end{array}\right].$$

Nun gilt $\omega_8^4 = e^{-\pi i} = -1$ und $\omega_8^{2k} = e^{-2\pi i k/4} = \omega_4^k$. Mit $\mathbf{D}_4 = \operatorname{diag}(\omega_8^0, \omega_8^1, \omega_8^2, \omega_8^3)$ können wir daher schreiben

$$\left[\begin{array}{c} y_0 \\ y_1 \\ y_2 \\ y_3 \\ \hline y_4 \\ y_5 \\ y_6 \\ y_7 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{T}_4 & \mathbf{D}_4\mathbf{T}_4 \\ \hline \mathbf{T}_4 & -\mathbf{D}_4\mathbf{T}_4 \end{array}\right] \left[\begin{array}{c} \beta_0 \\ \beta_2 \\ \beta_4 \\ \beta_6 \\ \hline \beta_1 \\ \beta_3 \\ \beta_5 \\ \beta_7 \end{array}\right].$$

Setzen wir

$$\mathbf{c} = \mathbf{T}_4 \begin{bmatrix} \beta_0 \\ \beta_2 \\ \beta_4 \\ \beta_6 \end{bmatrix}, \qquad \mathbf{d} = \mathbf{D}_4\mathbf{T}_4 \begin{bmatrix} \beta_1 \\ \beta_3 \\ \beta_5 \\ \beta_7 \end{bmatrix},$$

so folgt

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \mathbf{c} + \mathbf{d}, \qquad \begin{bmatrix} y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \mathbf{c} - \mathbf{d}.$$

Dies bedeutet, die Fourier-Synthese für $n = 8$ Unbekannte lässt sich aus der Fourier-Synthese für $n/2 = 4$ Unbekannte zusammensetzen. △

Satz 14.7 Zu gegebenem $\boldsymbol{\beta} \in \mathbb{C}^{2n}$ sei

$$\mathbf{D}_n = \operatorname{diag}(\omega_{2n}^0, \omega_{2n}^1, \ldots, \omega_{2n}^{n-1}), \qquad \mathbf{c} = \mathbf{T}_n \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_{2n-2} \end{bmatrix}, \qquad \mathbf{d} = \mathbf{D}_n\mathbf{T}_n \begin{bmatrix} \beta_1 \\ \beta_3 \\ \vdots \\ \beta_{2n-1} \end{bmatrix}.$$

Dann gilt für $\mathbf{y} = \mathbf{T}_{2n}\boldsymbol{\beta} \in \mathbb{C}^{2n}$ die Beziehung

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix} = \mathbf{c} + \mathbf{d}, \qquad \begin{bmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_{2n-1} \end{bmatrix} = \mathbf{c} - \mathbf{d}.$$

Beweis. Nachrechnen liefert für $\ell = 0, 1, 2, \ldots, n-1$

$$\begin{aligned} y_\ell &= \sum_{k=0}^{n-1} \beta_{2k}\omega_n^{k\ell} + \omega_{2n}^{\ell} \sum_{k=0}^{n-1} \beta_{2k+1}\omega_n^{k\ell} = \sum_{k=0}^{n-1} \beta_{2k}\omega_{2n}^{2k\ell} + \omega_{2n}^{\ell} \sum_{k=0}^{n-1} \beta_{2k+1}\omega_{2n}^{2k\ell} \\ &= \sum_{k=0}^{n-1} \beta_{2k}\omega_{2n}^{2k\ell} + \sum_{k=0}^{n-1} \beta_{2k+1}\omega_{2n}^{(2k+1)\ell} = \sum_{k=0}^{2n-1} \beta_k\omega_{2n}^{k\ell} \end{aligned}$$

und für $\ell = n, n+1, \ldots, 2n-1$

$$y_\ell = \sum_{k=0}^{n-1} \beta_{2k}\omega_n^{k\ell} - \omega_{2n}^{\ell-n} \sum_{k=0}^{n-1} \beta_{2k+1}\omega_n^{k\ell} = \sum_{k=0}^{n-1} \beta_{2k}\omega_n^{k\ell} + \omega_{2n}^{\ell} \sum_{k=0}^{n-1} \beta_{2k+1}\omega_n^{k\ell} = \sum_{k=0}^{2n-1} \beta_k\omega_{2n}^{k\ell}.$$

□

Zur Berechnung von $\mathbf{y} = \mathbf{T}_{2n}\boldsymbol{\beta} \in \mathbb{C}^{2n}$ werden demnach nur die Vektoren $\mathbf{c}, \mathbf{d} \in \mathbb{C}^n$ benötigt. Dies entspricht wieder der Divide-and-Conquer-Vorgehensweise, da die ursprüngliche Aufgabe für $2n$ Unbekannte durch geschicktes Unterteilen in zwei Unterprobleme mit nur noch n Unbekannten zerlegt wurde. Die vollständige schnelle Fourier-Synthese erhalten wir, wenn wir $\mathbf{c}$ und $\mathbf{d}$ rekursiv mit Hilfe des selben Algorithmus berechnen. Dazu seien $j \in \mathbb{N}$ und $n = 2^j$ eine Zweierpotenz.

Satz 14.8 Der Aufwand zur Bewältigung der vollständigen schnellen Fourier-Synthese lässt sich abschätzen durch $n/2 \log_2 n$ Multiplikationen.

Beweis. Wir führen den Beweis vermittels Induktion über j. Für den Induktionsanfang $j = 1$ ist keine weitere Unterteilung möglich und wir müssen die volle (2×2)-Matrix anwenden

$$\mathbf{T}_2\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \boldsymbol{\beta}.$$

Der Aufwand hierfür lässt sich durch $1/2 \cdot 2 \log_2 2 = 1$ Multiplikation abschätzen. Wir wollen nun annehmen, dass die Behauptung für j richtig ist, wir also nicht mehr als $1/2 \cdot 2^j \log_2 2^j = 1/2 \cdot 2^j j$ Multiplikationen zur Berechnung von $\mathbf{T}_{2^j}\boldsymbol{\beta}$ benötigen. Der Induktionsschritt $j \mapsto j+1$ ergibt sich nun (vgl. Satz 14.7) wie folgt:

$$2 \cdot \frac{1}{2} j 2^j + 2^j = (j+1)2^j = \frac{1}{2}(j+1)2^{j+1}.$$

□

Bemerkung: Die Reduktion der Komplexität von $\mathcal{O}(n^2)$ auf $\mathcal{O}(n \log_2 n)$ bewirkte eine technische Revolution in der Signalverarbeitung. Erst mit ihrer Hilfe war die digitale Signalverarbeitung möglich.

14.3 Zirkulante Matrizen

Definition 14.9 Eine Matrix $\mathbf{C} \in \mathbb{C}^{n\times n}$ heißt **zirkulant**, falls ein $\mathbf{c} = [c_0, c_1, c_2, \ldots, c_{n-1}]^T \in \mathbb{C}^n$ existiert, so dass

$$\mathbf{C} = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-2} & c_{n-1} \\ c_{n-1} & c_0 & c_1 & & c_{n-1} & c_0 \\ c_{n-2} & c_{n-1} & c_0 & & c_0 & c_1 \\ \vdots & & & \ddots & & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_{n-1} & c_0 \end{bmatrix}.$$

Satz 14.10 Sei $\mathbf{C} \in \mathbb{C}^{n\times n}$ eine zirkulante Matrix und

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n-1} \end{bmatrix} = \mathbf{T}_n \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{bmatrix}.$$

Dann gilt für $\mathbf{D} = \operatorname{diag}(\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_{n-1})$

$$\mathbf{C} = \frac{1}{n}\mathbf{T}_n\mathbf{D}\mathbf{T}_n^\star,$$

dies bedeutet, die $(\lambda_k, \boldsymbol{\omega}^k)$ sind die Eigenpaare von $\mathbf{C}$.

Beweis. Setzen wir $c_{n+i} := c_i$, so gilt für alle $j, k = 0, 1, 2, \ldots, n-1$

$$\begin{aligned} \left[\mathbf{C}\boldsymbol{\omega}^k\right]_j &= \sum_{i=0}^{n-1} c_{i-j}\omega_n^{ik} = \omega_n^{jk} \sum_{i=0}^{n-1} c_{i-j}\omega_n^{(i-j)k} \\ &= \omega_n^{jk}\left[\sum_{i=-j}^{-1} \underbrace{c_i\omega_n^{ik}}_{=c_{n+i}\omega_n^{(n+i)k}} + \sum_{i=0}^{n-1-j} c_i\omega_n^{ik}\right] \\ &= \omega_n^{jk} \underbrace{\sum_{i=0}^{n-1} c_i\omega_n^{ik}}_{=\lambda_k}. \end{aligned}$$

Damit sind genau alle $(\lambda_k, \boldsymbol{\omega}^k)$ Eigenpaare von $\mathbf{C}$, dies bedeutet,

$$\mathbf{C}\mathbf{T}_n = \mathbf{T}_n\mathbf{D}.$$

$\square$

Eine zirkulante Matrix ist folglich durch die Fourier-Transformation leicht diagonalisierbar. Daher ist ein lineares Gleichungssystem $\mathbf{C}\mathbf{x} = \mathbf{b}$ mit Hilfe der schnellen Fouriertransformation gemäß

$$\mathbf{x} = \mathbf{C}^{-1}\mathbf{b} = n\mathbf{T}_n^{-\star}\mathbf{D}^{-1}\mathbf{T}_n^{-1}\mathbf{b} = \frac{1}{n}\mathbf{T}_n\mathbf{D}^{-1}\mathbf{T}_n^\star\mathbf{b}$$

lösbar mit Aufwand $\mathcal{O}(n \log n)$. Ebenso können Matrix-Vektor-Multiplikation in $\mathcal{O}(n \log n)$ Operationen durchgeführt werden:

$$\mathbf{Cx} = \frac{1}{n}\mathbf{T}_n\mathbf{D}\mathbf{T}_n^{\star}\mathbf{x}.$$

15. Splines

15.1 Spline-Räume

Wie wir in Kapitel 13 bereits erwähnt haben, ist die Polynominterpolation kein zufriedenstellendes numerisches Verfahren, wenn der Polynomgrad groß wird. Als Ausweg bietet es sich an, Polynome niederen Grades "aneinanderzuhängen".

Definition 15.1 Sei $\triangle = \{x_0, x_1, \ldots, x_n\}$ ein Gitter von $n+1$ paarweise verschiedenen Knoten mit $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$. Ein **Spline** vom Grad m ist eine $(m-1)$-mal stetig differenzierbare Funktion s, die auf jedem Intervall $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, n$, mit einem Polynom $s_i \in \Pi_m$ übereinstimmt. Den Raum der Splines vom Grad m bezüglich $\triangle$ bezeichnen wir mit $S_m(\triangle)$. Splines vom Grad 1,2, bzw. 3 werden auch linear, quadratisch, bzw. kubisch genannt.

Offensichtlich ist $S_m(\triangle)$ ein linearer Vektorraum, und es gilt

$$\Pi_m \subset S_m(\triangle).$$

Splines vom Grad ≥ 3 werden vom Auge als "glatt" empfunden. Deshalb werden kubische Splines besonders oft verwendet.

Wir betrachten zunächst den Fall linearer Splines.

Satz 15.2 Zu gegebenen Daten $y_0, y_1, \ldots, y_n$ existiert genau ein linearer Spline $s \in S_1(\triangle)$ mit

$$s(x_i) = y_i, \qquad i = 0, 1, \ldots, n.$$

Der Vektorraum $S_1(\triangle)$ besitzt die Dimension $\dim S_1(\triangle) = n + 1$.

Beweis. Zu den Daten $y_0, y_1, \ldots, y_n$ konstruieren wir einen linearen Spline durch

$$s(x) := \frac{x_{i+1} - x}{x_{i+1} - x_i} y_i + \frac{x - x_i}{x_{i+1} - x_i} y_{i+1}, \qquad x \in [x_i, x_{i+1}].$$

Die Eindeutigkeit dieses Splines folgt aus der Tatsache, dass jede lineare Teilfunktion durch die beiden Randpunkte (x_i, y_i) und (x_{i+1}, y_{i+1}) eindeutig festgelegt ist.

Damit ist die lineare Abbildung

$$A : S_1(\triangle) \to \mathbb{K}^{n+1}, \qquad s \mapsto \begin{bmatrix} s(x_0) \\ s(x_1) \\ \vdots \\ s(x_n) \end{bmatrix}.$$

bijektiv, womit alle Aussagen des Satzes bewiesen sind. □

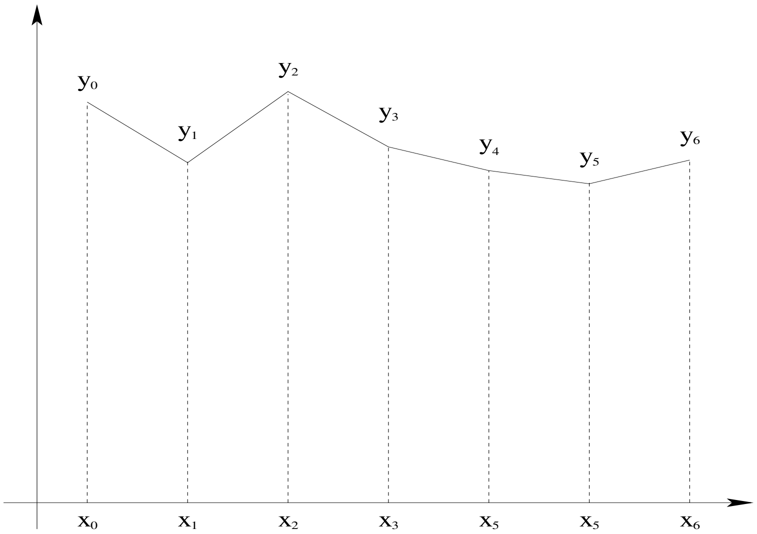


Abbildung 15.1: Linearer Spline

Satz 15.3 $S_m(\triangle)$ ist ein $(n+m)$-dimensionaler Vektorraum.

Beweis. Für $m = 1$ haben wir die Behauptung im vorhergehenden Satz bereits bewiesen. Sei also $m > 1$ und sei $s_0, s_1, \ldots, s_n$ eine Basis von $S_1(\triangle)$. Weiterhin sei σ_i eine beliebige $(m-1)$-te Stammfunktion von s_i, $i = 0, 1, \ldots, n$. Dann gehören sowohl σ_i als auch die Monome $x^0, x^1, \ldots, x^{m-2}$ zu $S_m(\triangle)$. Wir zeigen, dass

$$\{\sigma_0, \sigma_1, \ldots, \sigma_n\} \cup \{x^0, x^1, \ldots, x^{m-2}\}$$

eine Basis von $S_m(\triangle)$ ist.
Ist $s \in S_m(\triangle)$, so gilt $s^{(m-1)} \in S_1(\triangle)$ und deshalb

$$s^{(m-1)}(x) = \sum_{i=0}^{n} c_i s_i(x)$$

für gewisse Koeffizienten $c_i \in \mathbb{K}$. Daraus folgt aber, dass

$$s(x) = \sum_{i=0}^{n} c_i \sigma_i(x) + \sum_{i=0}^{m-2} d_i x^i$$

mit Koeffizienten $d_i \in \mathbb{K}$. Also lässt sich jedes $s \in S_m(\triangle)$ als Linearkombination der Basisvektoren darstellen. Zum Nachweis der linearen Unabhängigkeit nehmen wir an, es sei

$$\sum_{i=0}^{n} c_i \sigma_i(x) + \sum_{i=0}^{m-2} d_i x^i \equiv 0$$

für Zahlen $c_0, c_1, \ldots, c_n, d_0, d_1, \ldots, d_{m-2} \in \mathbb{K}$. Differenzieren wir diese Gleichung $(m-1)$-mal, dann folgt

$$\sum_{i=0}^{n} c_i s_i(x) \equiv 0.$$

Da $s_0, s_1, \ldots, s_n$ eine Basis von $S_1(\triangle)$ ist, folgt hieraus $c_0 = c_1 = \cdots = c_n = 0$. Aus

$$\sum_{i=0}^{m-2} d_i x^i \equiv 0$$

folgt schließlich auch $d_0 = d_1 = \cdots = d_{m-2} = 0$. □

15.2 Kubische Splines

Lineare Splines über $\triangle$ haben genau $n+1$ Freiheitsgrade, weshalb das Interpolationsproblem eindeutig lösbar ist. Da die Dimension von $S_m(\triangle)$ für $m > 1$ größer als $n+1$ ist, müssen wir zusätzliche Bedingungen an den interpolieren Spline stellen, um Eindeutigkeit zu erzwingen. Für ungerades m benötigen wir eine gerade Anzahl von Zusatzbedingungen, die wir symmetrisch an den Intervallenden verteilen. Obwohl die folgenden Sätze leicht auf beliebige m übertragen werden können, beschränken wir uns der Übersichtlichkeit halber auf den wichtigen Fall der kubischen Splines.

Kubische Spline-Interpolation: Gegeben seien ein Gitter $\triangle = \{x_0, x_1, \ldots, x_n\}$ über $[a, b]$ mit $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ sowie $n+1$ Werte $y_0, y_1, \ldots, y_n \in \mathbb{K}$. Gesucht ist ein kubischer Spline $s \in S_3(\triangle)$ mit

$$s(x_i) \stackrel{!}{=} y_i, \qquad i = 0, 1, \ldots, n, \tag{15.1}$$

der zusätzlich einer der folgenden Randbedingungen

$$\begin{aligned} s''(a) &= s''(b) = 0 && \text{(natürliche Randbedingungen)} \\ s'(a) &= y_0', \quad s'(b) = y_n' && \text{(Hermite-Randbedingungen)} \\ s'(a) &= s'(b), \quad s''(a) = s''(b) && \text{(periodische Randbedingungen)} \end{aligned} \tag{15.2}$$

genügt. Dabei gelte im zweiten Fall $y_0', y_n' \in \mathbb{K}$ sowie im dritten $s(a) = y_0 = y_n = s(b)$.

Bemerkung: Interpolierende kubische Splines haben eine interessante Optimalitätseigenschaft, die die "Glattheit" betrifft. Für eine Funktion $y : [a, b] \to \mathbb{K}$ ist

$$\kappa(x) := \frac{y''(x)}{\big(1 + y'(x)^2\big)^{3/2}}$$

die *Krümmung* der Kurve $\big(x, y(x)\big)$ für gegebenes Argument x. Beschreibt etwa $y(x)$ die Lage einer dünnen Holzlatte, so misst

$$E = \int_a^b |\kappa(x)|^2 \, \mathrm{d}x = \int_a^b \left| \frac{y''(x)}{\big(1 + y'(x)^2\big)^{3/2}} \right|^2 \mathrm{d}x$$

die *Biegeenergie* der Latte. Aufgrund des Hamiltonschen Prinzips stellt sich die Latte so ein, dass E minimal wird. Für kleine Auslenkungen $y'(x)$ gilt

$$E \approx \int_a^b |y''(x)|^2 \, \mathrm{d}x =: \|y''\|_{L^2}^2.$$

Wie wir gleich sehen werden, beschreibt daher der kubische Spline näherungsweise die Form dieser Holzlatte, falls sie an den Stellen (x_i, y_i), $i = 0, 1, \ldots, n$, fixiert wird und an den Enden

- lose und deshalb gerade ist: $s''(x) = 0$ für $x \le a$ und $x \ge b$ (natürliche Randbedingungen),
- in eine bestimmte Richtung fest eingespannt ist (Hermite-Randbedingungen), oder
- zusammengeklebt ist (periodische Randbedingungen).

Satz 15.4 Der kubische Spline $s \in S_3(\triangle)$ interpoliere die Punkte (x_i, y_i), $i = 0, 1, \ldots, n$, und erfülle eine der Randbedingungen aus (15.2). Die Funktion $g \in C^2([a, b])$ sei irgendeine andere interpolierende Funktion, die den selben Randbedingungen genügt. Dann gilt

$$\|g''\|_{L^2}^2 = \|s''\|_{L^2}^2 + \|g'' - s''\|_{L^2}^2, \tag{15.3}$$

insbesondere also

$$\|s''\|_{L^2} \le \|g''\|_{L^2}.$$

Beweis. Es gilt

$$\|g''\|_{L^2}^2 = \|s'' + (g'' - s'')\|_{L^2}^2 = \|s''\|_{L^2}^2 + \|g'' - s''\|_{L^2}^2 + 2\operatorname{Re}\left(\int_a^b s''\overline{(g'' - s'')}\,\mathrm{d}x\right).$$

Wir müssen also zeigen, dass hierin der letzte Term verschwindet.
Da s und g nach Voraussetzung die selben Randbedingungen erfüllen, gilt die Identität

$$s''(a)\overline{\big(g'(a) - s'(a)\big)} = s''(b)\overline{\big(g'(b) - s'(b)\big)}. \tag{15.4}$$

Daher folgt mit partieller Integration

$$\begin{aligned}
\int_a^b s''\overline{(g'' - s'')}\,\mathrm{d}x &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} s''\overline{(g'' - s'')}\,\mathrm{d}x \\
&= \sum_{i=1}^n \left\{ s''\overline{(g' - s')}\Big|_{x_{i-1}}^{x_i} + \int_{x_{i-1}}^{x_i} s'''\overline{(g' - s')}\,\mathrm{d}x \right\}.
\end{aligned}$$

Da s''' im Innern von $[x_{i-1}, x_i]$ konstant ist, etwa gleich c_i, ergibt sich also

$$\begin{aligned}
\int_a^b s''\overline{(g'' - s'')}\,\mathrm{d}x &= \sum_{i=1}^n s''\overline{(g' - s')}\Big|_{x_{i-1}}^{x_i} - \sum_{i=1}^n c_i \int_{x_{i-1}}^{x_i} \overline{g' - s'}\,\mathrm{d}x \\
&= \underbrace{s''\overline{(g' - s')}\Big|_a^b}_{= 0 \text{ wegen } (15.4)} - \sum_{i=1}^n c_i \underbrace{\overline{(g - s)}\Big|_{x_{i-1}}^{x_i}}_{=0} = 0,
\end{aligned}$$

da g und s die gleichen Daten interpolieren. □

Satz 15.5 Die durch (15.1) und (15.2) bestimmte kubische Spline-Interpolationsaufgabe ist eindeutig lösbar.

Beweis. Wegen $\dim S_3(\triangle) = n+3$ ist die lineare Abbildung

$$A_H : S_3(\triangle) \to \mathbb{K}^{n+3}, \qquad s \mapsto \begin{bmatrix} s(x_0) \\ s(x_1) \\ \vdots \\ s(x_n) \\ s'(x_0) \\ s'(x_n) \end{bmatrix}$$

surjektiv. Damit besitzt das Hermitesche Interpolationsproblem eine Lösung.
Im Falle natürlicher Randbedingungen betrachten wir die lineare Abbildung

$$A_N : \{S_3(\triangle) : s''(a) = s''(b) = 0\} \to \mathbb{K}^{n+1}, \qquad s \mapsto \begin{bmatrix} s(x_0) \\ s(x_1) \\ \vdots \\ s(x_n) \end{bmatrix}.$$

Da $\{S_3(\triangle) : s''(a) = s''(b) = 0\}$ als Nullraum der surjektiven Abbildung

$$S_3(\triangle) \to \mathbb{K}^2, \qquad s \mapsto \begin{bmatrix} s''(a) \\ s''(b) \end{bmatrix}$$

die Dimension $n+1$ besitzt, ist auch A_N surjektiv und folglich die Interpolationsaufgabe lösbar.
Analog argumentiert man im Falle periodischer Randbedingungen, um zu zeigen, dass

$$A_P : \{S_3(\triangle) : s'(a) = s'(b) \wedge s''(a) = s''(b)\} \to \mathbb{K}^{n+1}, \qquad s \mapsto \begin{bmatrix} s(x_0) \\ s(x_1) \\ \vdots \\ s(x_n) \end{bmatrix}.$$

surjektiv ist.
Seien nun $s, \sigma \in S_3(\triangle)$ zwei kubische Splines mit (15.1) und (15.2). Aus Satz 15.4 und (15.3) folgt dann

$$\|s'' - \sigma''\|_{L^2}^2 = \|s''\|_{L^2}^2 - \|\sigma''\|_{L^2}^2 = 0,$$

und deshalb gilt $\sigma - s \in \Pi_1$. Die Interpolationsbedingungen (15.1) implizieren insbesondere

$$(\sigma - s)(a) = (\sigma - s)(b) = 0,$$

woraus sich $\sigma = s$ ergibt. □

15.3 B–Splines

Wir führen nun eine Basis in den Spline-Räumen $S_m(\triangle)$ ein, die zur numerischen Berechnung interpolierender Splines genutzt werden kann. Dabei wollen wir uns der Einfachheit halber auf äquidistante Gitter beschränken.

Satz 15.6 Die durch

$$B_0(x) := \begin{cases} 1, & |x| \le 0.5, \\ 0, & |x| > 0.5, \end{cases}$$

und

$$B_{m+1}(x) := \int_{x-1/2}^{x+1/2} B_m(t)\,\mathrm{d}t, \qquad x \in \mathbb{R}, \qquad m = 0, 1, 2, \dots \tag{15.5}$$

rekursiv definierten Funktionen sind Splines vom Grad m auf dem Gitter

$$\triangle_m := \left\{ i - \frac{m+1}{2} : i = 0, 1, \dots, m+1 \right\}.$$

Sie heißen *B–Splines* vom Grad m, sind nichtnegativ und erfüllen $B_m(x) = 0$ für $|x| > (m+1)/2$.

Beweis. Wir beweisen die Behauptung durch Induktion nach m. Da für $m = 0$ die Aussage klar ist, nehmen wir an, dass die Behauptung für $m \ge 0$ erfüllt ist. Dann folgt aus (15.5), dass $B_{m+1}(x) \ge 0$ für alle x, da $B_m(x) \ge 0$ für alle x, und $B_{m+1}(x) = 0$ für alle $|x| > (m+2)/2$, da $B_m(x) = 0$ für alle $|x| > (m+1)/2$. Außerdem ist

$$B'_{m+1}(x) = B_m\left(x + \frac{1}{2}\right) - B_m\left(x - \frac{1}{2}\right).$$

Aufgrund der Induktionsannahme ist daher $B'_{m+1}(x)$ überall mindestens $(m-1)$-mal stetig differenzierbar und auf allen Intervallen $[x - 1/2, x + 1/2]$ mit $x \in \triangle_m$ ein Polynom vom Grad $\le m$. Daher ist B_{m+1} ein Spline vom Grad $m + 1$. $\square$

Die ersten B–Splines sind gegeben durch

$$B_1(x) = \begin{cases} 1 - |x|, & |x| \le 1, \\ 0, & |x| > 1, \end{cases}$$

$$B_2(x) = \frac{1}{2} \begin{cases} 2 - (|x| - 0.5)^2 - (|x| + 0.5)^2, & |x| \le 0.5, \\ (|x| - 1.5)^2, & 0.5 < |x| \le 1.5, \\ 0, & |x| > 1.5, \end{cases}$$

$$B_3(x) = \frac{1}{6} \begin{cases} (2 - |x|)^3 - 4(1 - |x|)^3, & |x| \le 1, \\ (2 - |x|)^3, & 1 < |x| \le 2, \\ 0, & |x| > 2, \end{cases}$$

vergleiche Abbildung 15.2.

Satz 15.7 Für $m = 0, 1, 2, \dots$ sind die B–Splines

$$B_m(\cdot - i), \qquad i = 0, 1, \dots, m$$

linear unabhängig auf dem Intervall $I_m := [(m-1)/2, (m+1)/2]$.

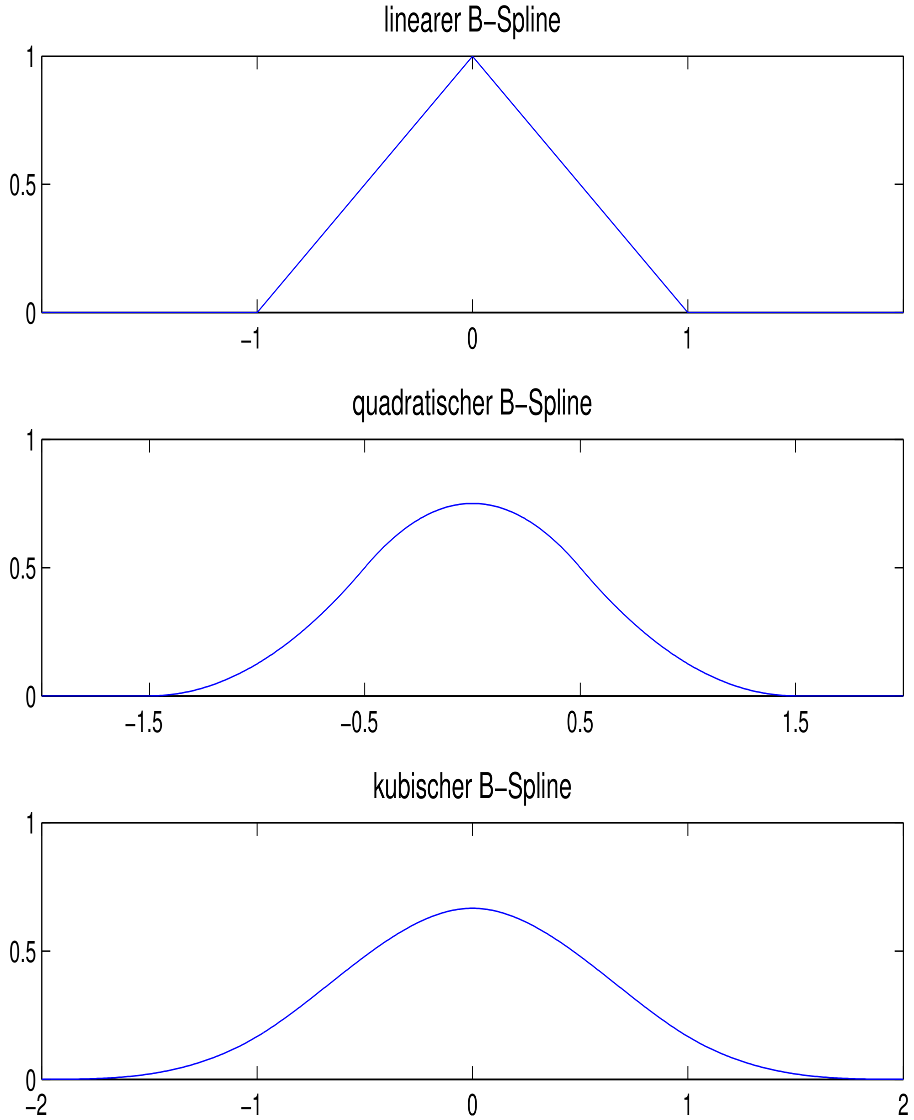


Abbildung 15.2: Lineare, quadratische und kubische B–Splines.

Beweis. Für $m = 0$ ist die Behauptung offensichtlich. Wir nehmen also an, dass die Behauptung wahr ist für $m-1$. Zu zeigen ist, dass

$$\sum_{i=0}^{m} c_i B_m(x-i) = 0, \qquad x \in I_m \tag{15.6}$$

nur gilt, falls $c_0 = c_1 = \cdots = c_m = 0$. Differenzieren von (15.6) liefert

$$\sum_{i=0}^{m} c_i \left\{ B_{m-1}\left(x - i + \frac{1}{2}\right) - B_{m-1}\left(x - i - \frac{1}{2}\right) \right\} = 0, \qquad x \in I_m.$$

Wegen $B_{m-1}(x) = 0$ für alle $|x| \geq m/2$ folgt

$$B_{m-1}\left(x + \frac{1}{2}\right) = B_{m-1}\left(x - m - \frac{1}{2}\right) = 0, \qquad x \in I_m.$$

Daher können wir die letzte Gleichung wie folgt umformen:

$$\sum_{i=1}^{m}(c_i - c_{i-1})B_{m-1}\left(x - i + \frac{1}{2}\right) = 0, \qquad x \in I_m.$$

Aus der Induktionsannahme folgt nun $c_i = c_{i-1}$ für alle $i = 1, 2, \ldots, m$, dies bedeutet $c_0 = c_1 = \cdots = c_m =: c$. Daher gilt nach (15.6)

$$c\sum_{i=0}^{m} B_m(x - i) = 0, \qquad x \in I_m.$$

Durch Integration dieser Gleichung über dem Interval I_m erhalten wir

$$c\int_{(m-1)/2}^{(m+1)/2}\sum_{i=0}^{m} B_m(x-i)\,\mathrm{d}x = c\sum_{i=0}^{m}\int_{(m-1)/2-i}^{(m+1)/2-i} B_m(x)\,\mathrm{d}x = c\int_{-(m+1)/2}^{(m+1)/2} B_m(x)\,\mathrm{d}x = 0.$$

Dies impliziert schließlich $c = 0$, da B_m positiv ist. □

Korollar 15.8 Sei $\triangle = \{x_0, x_1, \ldots, x_n\}$ ein äquidistantes Gitter mit Gitterweite $h > 0$, das heißt $x_i := x_0 + hi$, und sei $m = 2\ell - 1$ mit $\ell \in \mathbb{N}$. Dann bilden die B–Splines

$$B_{m,i}(x) := B_m\left(\frac{x - x_i}{h}\right), \qquad x \in [x_0, x_n]$$

für $i = 1 - \ell, 2 - \ell, \ldots, n + \ell - 1$ eine Basis von $S_m(\triangle)$.

Beweis. Nach Satz 15.6 liegen die Funktionen $B_{m,i}$ alle in $S_m(\triangle)$. Nach Satz 15.3 müssen wir daher lediglich zeigen, dass sie linear unabhängig sind. Dies folgt aber durch Anwendung von Satz 15.7 auf jedes Teilintervall. □

Mit Hilfe von B–Splines können interpolierende Splines effizient berechnet werden. Wir demonstrieren dies am Beipiel linearer und kubischer Splines:

Beispiel 15.9 Der lineare Spline

$$s(x) = \sum_{i=0}^{n} y_i B_1\left(\frac{x - x_i}{h}\right), \qquad x \in [x_0, x_n]$$

interpoliert die Werte (x_i, y_i), $i = 0, 1, \ldots, n$. △

Beispiel 15.10 Es gilt

$$B_3(0) = \frac{2}{3}, \quad B_3(\pm 1) = \frac{1}{6}, \quad B_3'(0) = 0, \quad B_3'(\pm 1) = \mp\frac{1}{2}, \quad B_3''(0) = -2, \quad B_3''(\pm 1) = 1.$$

Daher erfüllt der kubische Spline

$$s(x) = \sum_{i=-1}^{n+1} c_i B_3\left(\frac{x - x_i}{h}\right), \qquad x \in [x_0, x_n]$$

genau dann die Interpolationsbedingungen (15.1), falls nachfolgendes Gleichungssystem erfüllt ist:

1. natürliche Randbedingungen:

$$\begin{bmatrix} 1 & -2 & 1 & & & \\ 1/6 & 2/3 & 1/6 & & & \\ & 1/6 & 2/3 & 1/6 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1/6 & 2/3 & 1/6 \\ & & & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \\ \vdots \\ c_n \\ c_{n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ y_0 \\ y_1 \\ \vdots \\ y_n \\ 0 \end{bmatrix}$$

2. Hermite-Randbedingungen:

$$\begin{bmatrix} -1/2 & 0 & 1/2 & & & \\ 1/6 & 2/3 & 1/6 & & & \\ & 1/6 & 2/3 & 1/6 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1/6 & 2/3 & 1/6 \\ & & & -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \\ \vdots \\ c_n \\ c_{n+1} \end{bmatrix} = \begin{bmatrix} hy_0' \\ y_0 \\ y_1 \\ \vdots \\ y_n \\ hy_n' \end{bmatrix}$$

3. periodischer Randbedingungen: unter der Voraussetzung $y_0 = y_n$ können wir hier aufgrund der Periodizität B_{i-1} identifizieren mit B_{i+n-1}, $i = 0, 1, 2$, dies bedeutet $c_{i+n-1} = c_{i-1}$ für $i = 0, 1, 2$. Folglich gilt

$$\begin{bmatrix} 1/6 & 2/3 & 1/6 & & & \\ & 1/6 & 2/3 & 1/6 & & \\ & & & \ddots & & \\ & & & 1/6 & 2/3 & 1/6 \\ 1/6 & & & & 1/6 & 2/3 \\ 2/3 & 1/6 & & & & 1/6 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ \vdots \\ c_{n-2} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}$$

△

15.4 Interpolationsfehler

Es sei $\triangle = \{x_0, x_1, \ldots, x_n\}$ ein Gitter mit $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$. Mit $h_i := x_{i+1} - x_i$, $i = 0, 1, \ldots, n-1$, bezeichnet $h = \max_{i=0,1,\ldots,n-1} h_i$ die Gitterweite. Wir zeigen zunächst, dass die lineare Splineinterpolation eine Approximation zweiter Ordnung liefert. Dazu führen wir den Interpolationsprojektor

$$L_1 : C([a, b]) \to S_1(\triangle), \qquad f \mapsto s$$

ein, wobei $s \in S_1(\triangle)$ gemäß Satz 15.2 durch die Interpolationsbedingungen $s(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$, eindeutig bestimmt ist.

Satz 15.11 Für $f \in C^2([a, b])$ gilt

$$\|f - L_1 f\|_{L^2} \leq \frac{h^2}{2} \|f''\|_{L^2}. \tag{15.7}$$

Beweis. Die Funktion $g := f - L_1 f$ besitzt die Nullstellen $x_0, x_1, \ldots, x_n$. Daher gilt

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} |g(x)|^2 \,\mathrm{d}x &= \int_{x_i}^{x_{i+1}} \left| \int_{x_i}^{x} g'(t) \cdot 1 \,\mathrm{d}t \right|^2 \mathrm{d}x \\
&\overset{\text{CSU}}{\leq} \int_{x_i}^{x_{i+1}} \left(\int_{x_i}^{x} 1 \,\mathrm{d}t \right) \left(\int_{x_i}^{x} |g'(t)|^2 \,\mathrm{d}t \right) \mathrm{d}x \\
&= \int_{x_i}^{x_{i+1}} (x - x_i) \left(\int_{x_i}^{x} |g'(t)|^2 \,\mathrm{d}t \right) \mathrm{d}x \\
&\leq \left(\int_{x_i}^{x_{i+1}} |g'(t)|^2 \,\mathrm{d}t \right) \int_{x_i}^{x_{i+1}} (x - x_i) \,\mathrm{d}x \\
&= \frac{h_i^2}{2} \int_{x_i}^{x_{i+1}} |g'(t)|^2 \,\mathrm{d}t.
\end{aligned}$$

Durch Summation über i erhalten wir die Abschätzung

$$\|f - L_1 f\|_{L^2} \leq \frac{h}{\sqrt{2}} \|(f - L_1 f)'\|_{L^2}. \tag{15.8}$$

Weiter liefert partielle Integration

$$\begin{aligned}
\|(f - L_1 f)'\|_{L^2}^2 = \sum_{i=0}^{n-1} \Bigg\{ & \underbrace{(f - L_1 f)'(x)\overline{(f - L_1 f)(x)}\Big|_{x_i}^{x_{i+1}}}_{=0} \\
& - \int_{x_i}^{x_{i+1}} (f - L_1 f)''(x) \overline{(f - L_1 f)(x)} \,\mathrm{d}x \Bigg\} \\
= -\sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} & f''(x) \overline{(f - L_1 f)(x)} \,\mathrm{d}x.
\end{aligned}$$

Mit Hilfe der Cauchy-Schwarzschen Ungleichung und (15.8) folgt hieraus

$$\|(f - L_1 f)'\|_{L^2}^2 \leq \|f - L_1 f\|_{L^2} \|f''\|_{L^2} \leq \frac{h}{\sqrt{2}} \|(f - L_1 f)'\|_{L^2} \|f''\|_{L^2}.$$

Division durch $\|(f - L_1 f)'\|_{L^2}$ liefert

$$\|(f - L_1 f)'\|_{L^2} \leq \frac{h}{\sqrt{2}} \|f''\|_{L^2},$$

was in Kombination mit (15.8) die Behauptung ergibt. □

Wir führen nun den kubischen Spline-Interpolationsprojektor

$$L_3 : C([a, b]) \to S_3(\triangle), \qquad f \mapsto s$$

ein, der durch die Interpolationsbedingungen $s(x_i) = y_i$, $i = 0, 1, \ldots, n$, und eine der Randbedingungen aus (15.2) eindeutig bestimmt ist.

Satz 15.12 Für $f \in C^4([a,b])$ gilt

$$\|f - L_3 f\|_{L^2} \leq \frac{h^4}{4} \|f^{(4)}\|_{L^2}.$$

Beweis. Da gilt $L_1(f - L_3 f) = 0$, erhalten wir mit (15.7)

$$\|f - L_3 f\|_{L^2} = \|(f - L_3 f) - L_1(f - L_3 f)\|_{L^2} \leq \frac{h^2}{2} \|f'' - (L_3 f)''\|_{L^2}.$$

Wir wählen ein $s \in S_3(\triangle)$ mit $s'' = L_1(f'')$ und definieren $g := f - s$. Dabei können wir s so wählen, dass auch die Randbedingungen erfüllt sind, da bei zweimaligem Integrieren von $L_1(f'')$ zwei Konstanten frei wählbar sind.
Es gilt

$$\begin{aligned}
\|f'' - (L_3 f)''\|_{L^2} &= \|f'' - s'' - (L_3 f)'' + s''\|_{L^2} \overset{L_3 s = s}{=} \|g'' - (L_3 g)''\|_{L^2} \\
&\leq \|g'' - (L_3 g)''\|_{L^2} + \|(L_3 g)''\|_{L^2} \overset{(15.3)}{=} \|g''\|_{L^2} \\
&= \|f'' - s''\|_{L^2}^2 = \|f'' - L_1(f'')\|_{L^2}^2
\end{aligned}$$

und eine erneute Anwendung von (15.7) auf f'' ergibt

$$\|f'' - (L_3 f)''\|_{L^2} \leq \|f'' - L_1 f''\|_{L^2} \leq \frac{h^2}{2} \|f^{(4)}\|_{L^2}.$$

Durch Kombination mit der ersten Ungleichung erhalten wir die Behauptung. □

16. Numerische Quadratur

16.1 Trapezregel

Gegenstand dieses Kapitels ist die numerische Approximation bestimmter Integrale

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

die nicht in geschlossener Form durch Angabe einer Stammfunktion integriert werden können.

Das einfachste und oft auch schon hinreichend gute Beispiel ist die sogenannte *Trapezregel*

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\big(f(a)+f(b)\big). \tag{16.1}$$

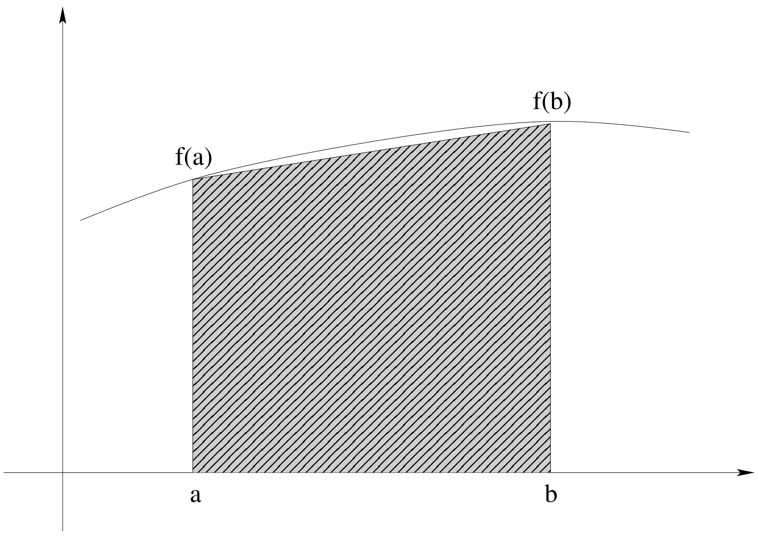


Abbildung 16.1: Geometrische Interpretation der Trapezregel.

Natürlich hat (16.1) in der Regel einen (beliebig großen) festen Fehler. Daher zerlegt man in der Praxis das Intervall $[a, b]$ in n gleichgroße Teilintervalle und wendet (16.1) auf jedes Teilintervall an. Dieses Verfahren nennt man die *zusammengesetzte Trapezregel*:

$$\begin{aligned} &a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b, \qquad x_i = a + ih, \qquad h = \frac{b-a}{n}, \\ &T_n[f] := \sum_{i=1}^{n} \frac{x_i - x_{i-1}}{2}\big(f(x_i) + f(x_{i-1})\big) = \frac{h}{2}f(a) + h\sum_{i=1}^{n-1} f(x_i) + \frac{h}{2}f(b). \end{aligned} \tag{16.2}$$

Man macht sich leicht mit der Definition des Riemann-Integrals klar, dass $T_n[f] \to I[f]$ für $n \to \infty$, falls f über $[a,b]$ Riemann-integrierbar ist.

Unter Zusatzannahmen kann folgende Fehlerabschätzung bewiesen werden:

Satz 16.1 Sei $f \in C^2([a,b])$. Dann gilt mit $h = (b-a)/n$

$$|I[f] - T_n[f]| \leq \frac{b-a}{12} h^2 M_2$$

mit $M_2 = \max_{a \leq x \leq b} |f''(x)|$.

Beweis. Betrachte zunächst die Näherung (16.1), also $n = 1$. Wie aus Abbildung 16.1 deutlich wird, ist

$$T_1[f] = \frac{b-a}{2}\big(f(a) + f(b)\big) = \int_a^b p(x)\,\mathrm{d}x,$$

wobei p das lineare Polynom

$$p(x) = f(a) + \frac{x-a}{b-a}\big(f(b) - f(a)\big)$$

ist, also insbesondere f in den Punkten a und b interpoliert. Gemäß Satz 13.3 folgt

$$|f(x) - p(x)| = \left|\frac{f''(\xi)}{2}(x-a)(x-b)\right| \leq \frac{M_2}{2}(x-a)(b-x).$$

Folglich ist

$$\begin{aligned}|I[f] - T_1[f]| &= \left|\int_a^b \big(f(x) - p(x)\big)\,\mathrm{d}x\right| \leq \frac{M_2}{2}\int_a^b (x-a)(b-x)\,\mathrm{d}x \\ &\overset{x:=a+(b-a)t}{=} \frac{M_2}{2}(b-a)^3 \int_0^1 t(1-t)\,\mathrm{d}t = \frac{M_2}{2}(b-a)^3\left(\frac{t^2}{2} - \frac{t^3}{3}\right)\Big|_0^1 = \frac{M_2}{12}(b-a)^3.\end{aligned}$$

Angewandt auf (16.2) ergibt sich

$$\begin{aligned}|I[f] - T_n[f]| &\leq \sum_{i=1}^n \left|\int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x - \frac{h}{2}\big(f(x_{i-1}) + f(x_i)\big)\right| \\ &\leq \sum_{i=1}^n \frac{1}{12} h^3 M_2 = n\frac{1}{12}h^3 M_2 = \frac{b-a}{12}h^2 M_2.\end{aligned}$$

□

Im folgenden betrachten wir noch andere Methoden zur Berechnung von Integralen

$$I[f] = \int_a^b f(x)\,\mathrm{d}x.$$

Definition 16.2 Wir sprechen von einer **Quadraturformel**

$$Q[f] = \sum_{i=1}^m w_i f(x_i)$$

mit **Knoten** x_i und **Gewichten** w_i, wenn die Wahl von m, $\{x_i\}$ und $\{w_i\}$ fix ist. Unter der dazugehörigen **zusammengesetzten Quadraturformel** $Q_n[f]$ verstehen wir dann die Unterteilung von $[a,b]$ in n gleich große Teilintervalle, in denen jeweils die Quadraturformel angewandt wird.

Im Zusammenhang mit zusammengesetzten Quadraturformeln steht dabei die Asymptotik $n \to \infty$ im Vordergrund.
Um die qualitativen Merkmale einer (zusammengesetzten) Quadraturformel beschreiben zu können, führen wir folgende Definitionen ein:

Definition 16.3 (a) Eine Quadraturformel $Q[f]$ hat **Exaktheitsgrad** q, falls

$$Q[p] = I[p] \quad \forall p \in \Pi_q.$$

(b) Eine zusammengesetzte Quadraturformel konvergiert gegen $I[f]$ mit der **Ordnung** s, falls

$$|Q_n[f] - I[f]| = \mathcal{O}(n^{-s}), \quad n \to \infty.$$

Beachte: Für den Exaktheitsgrad reicht es, eine Basis von Π_q zu untersuchen, da sowohl $Q[\cdot]$ wie $I[\cdot]$ lineare Abbildungen sind.

Beispiel 16.4 Die Trapezregel hat Exaktheitsgrad $q = 1$. Die zusammengesetzte Trapezregel konvergiert mit Ordnung $s = 2$. △

16.2 Newton-Cotes-Formeln

Mit Hilfe der Polynominterpolation lassen sich leicht Quadraturformeln für $I[f]$ mit beliebigen Exaktheitsgrad q angeben. Seien $x_0 < x_1 < \cdots < x_m$ $m+1$ vorgegebene Knoten in $[a,b]$ und

$$w_i := \int_a^b L_i(x)\,\mathrm{d}x \tag{16.3}$$

das Integral des i-ten zugehörigen Lagrange-Grundpolynoms. Dann gilt:

Proposition 16.5 Die Quadraturformel Q mit Knoten $\{x_i\}$ und Gewichten $\{w_i\}$ gemäß (16.3) hat den Exaktheitsgrad $q = m$.

Beweis. Sei $p \in \Pi_m$. Offensichtlich interpoliert p sich selbst. Wegen der Eindeutigkeit des Interpolationspolynoms gilt daher

$$p(x) = \sum_{i=0}^{m} p(x_i) L_i(x)$$

(vgl. Satz 13.2). Daraus folgt

$$\begin{aligned} I[p] &= \int_a^b p(x)\,\mathrm{d}x = \int_a^b \sum_{i=0}^m p(x_i) L_i(x)\,\mathrm{d}x \\ &= \sum_{i=0}^m p(x_i) \int_a^b L_i(x)\,\mathrm{d}x \overset{(16.3)}{=} \sum_{i=0}^m w_i p(x_i) = Q[p], \end{aligned}$$

was zu zeigen war. □

Außerdem erhalten wir aus Satz 13.3 folgende Fehlerabschätzung:

$$|I[f] - Q[f]| \leq \frac{M_{m+1}}{(m+1)!} \int_a^b w(x)\,\mathrm{d}x. \tag{16.4}$$

Hierbei ist $M_{m+1} = \max_{a \leq x \leq b} |f^{(m+1)}(x)|$ und $w(x)$ das Knotenpolynom aus Definition 13.1.

Beispiel 16.6 Beschränken wir uns auf den Fall äquidistanter Knoten $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$, dann erhalten wir die *Newton-Cotes-Formeln*. Im Fall $m = 1$ erhält man speziell die Trapezregel (16.1). Für $m = 2$ ergibt sich die *Simpson-Regel*

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\Big(f(a) + 4f\Big(\frac{a+b}{2}\Big) + f(b)\Big).$$

△

Beachte: Aus $1 \in \Pi_0$ folgt

$$\int_a^b 1\,\mathrm{d}x = \sum_{i=0}^m w_i \underbrace{f(x_i)}_{=1} = \sum_{i=0}^m w_i.$$

Daher gilt immer

$$\sum_{i=0}^m w_i = b - a$$

für die Formeln mit Exaktheitsgrad $q \geq 0$ (insbesondere Newton-Cotes-Formeln).

Bemerkung: Da die Polynominterpolation nur bedingt eine gute Approximation an f liefert (vgl. Kapitel 13), macht es im allgemeinen nur wenig Sinn den Approximationsparameter m hochzuschrauben. Zudem treten ab $m = 7$ negative Gewichte und dadurch unter Umständen Stabilitätsverluste selbst bei positiven Integranden auf. Die Gewichte der Newton-Cotes-Formeln für $m \leq 6$ sind in nachfolgender Tabelle zusammengefasst:

m	Gewichte	$Q[f] - \int_0^1 f(x)\,\mathrm{d}x$	Name
1	$\frac{1}{2}$ $\frac{1}{2}$	$\frac{1}{12}h^2 M_2$	Trapezregel
2	$\frac{1}{6}$ $\frac{4}{6}$ $\frac{1}{6}$	$\frac{1}{90}h^4 M_4$	Simpson-Regel
3	$\frac{1}{8}$ $\frac{3}{8}$ $\frac{3}{8}$ $\frac{1}{8}$	$\frac{3}{80}h^4 M_4$	3/8-Regel
4	$\frac{7}{90}$ $\frac{32}{90}$ $\frac{12}{90}$ $\frac{32}{90}$ $\frac{7}{90}$	$\frac{8}{945}h^6 M_6$	Milne-Regel
5	$\frac{19}{288}$ $\frac{75}{288}$ $\frac{50}{288}$ $\frac{50}{288}$ $\frac{75}{288}$ $\frac{19}{288}$	$\frac{275}{12096}h^6 M_6$	—
6	$\frac{41}{840}$ $\frac{216}{840}$ $\frac{27}{840}$ $\frac{272}{840}$ $\frac{27}{840}$ $\frac{216}{840}$ $\frac{41}{840}$	$\frac{9}{1400}h^8 M_8$	Weddle-Regel

Um genauere Approximationen an ein Integral zu erhalten, wird man also häufig das Intervall unterteilen — wie bereits in Abschnitt 16.1 gesehen — und die zugehörige zusammengesetzte Newton-Cotes-Formel verwenden.
Im Fall $m = 2$ erhält man so die *zusammengesetzte Simpson-Regel*

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}\{f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{2n-2}) + 4f(x_{2n-1}) + f(b)\}$$

mit $x_i = a + ih$, $i = 1, 2, \ldots, 2n-1$, und $h = (b-a)/(2n)$.
Wir benötigen das folgende Hilfsresultat:

Lemma 16.7 Sei $Q[f] = \sum_{i=0}^m w_i f(x_i)$ eine Quadraturformel für $\int_a^b f(x)\,\mathrm{d}x$ mit zu $(a+b)/2$ symmetrischen Knoten und Gewichten. Ist $Q[p] = I[p]$ für alle Polynome $p \in \Pi_{2q}$, dann hat $Q[\cdot]$ mindestens den Exaktheitsgrad $2q+1$.

Beweis. Betrachte die Basis

$$\left\{x^0, x^1, \ldots, x^{2q}, \left(x - \frac{a+b}{2}\right)^{2q+1}\right\}$$

von Π_{2q+1}. Nach Voraussetzung ist $Q[x^i] = I[x^i]$ für alle $i = 0, 1, \ldots, 2q$. Ferner ist

$$Q\left[\left(x - \tfrac{a+b}{2}\right)^{2q+1}\right] = 0,$$

da die Knoten und Gewichte symmetrisch zu $(a+b)/2$ liegen und $\big(x - (a+b)/2\big)^{2q+1}$ punktsymmetrisch zum Punkt $\big((a+b)/2, 0\big)$ ist. Andererseits ist auch

$$\begin{aligned} I\left[\left(x - \tfrac{a+b}{2}\right)^{2q+1}\right] &= \frac{1}{2q+2}\left(x - \frac{a+b}{2}\right)^{2q+2}\Bigg|_a^b \\ &= \frac{1}{2q+2}\left\{\left(\frac{b-a}{2}\right)^{2q+2} - \left(\frac{a-b}{2}\right)^{2q+2}\right\} \\ &= 0. \end{aligned}$$

Folglich ist $Q[p] = I[p]$ für alle $p \in \Pi_{2q+1}$. □

Offensichtlich hat die Simpson-Regel Exaktheitsgrad 3 (und nicht 2)! Es gilt:

Satz 16.8 Sei $f \in C^4([a,b])$. Dann gilt für die zusammengesetze Simpson-Regel $S_n[f]$ der Fehler

$$|I[f] - S_n[f]| \leq \frac{b-a}{180} h^4 M_4, \quad h = \frac{b-a}{2n},$$

mit $M_4 = \max_{a \leq x \leq b} |f^{(4)}(x)|$.

Beweis. Wir interpolieren f in jedem der n Teilintervalle $[c,d]$ durch ein Polynom p dritten Grades mit Stützstellen c, d und $(c+d)/2$ (letztere doppelt). Da die Quadraturformel für Polynome dritten Grades exakt ist, verbleibt das Integral über $f-p$ abzuschätzen. Dazu verwenden wir die Fehlerdarstellung (13.3) mit dem Knotenpolynom

$$w(x) = (x-c)(x-d)\Big(x - \frac{c+d}{2}\Big)^2$$

(vergleiche letzte Bemerkung von Abschnitt 13.1).

Durch Integration ergibt sich

$$\begin{aligned}
\int_c^d |f(x) - p(x)| \, \mathrm{d}x &\leq \frac{M_4}{4!} \Big(\frac{d-c}{2}\Big)^5 \int_{-1}^1 t^2(1-t^2) \, \mathrm{d}t \\
&\leq \frac{M_4}{4!} \Big(\frac{d-c}{2}\Big)^5 \Big(\frac{1}{3}t^3 - \frac{1}{5}t^5\Big)\Big|_{-1}^1 \\
&= \frac{M_4}{24} \Big(\frac{d-c}{2}\Big)^5 \frac{4}{15} \\
&= \frac{d-c}{180} \Big(\frac{b-a}{2n}\Big)^4 M_4.
\end{aligned}$$

Aufsummation ergibt die gewünschte Behauptung. □

Beachte: Hier werden doppelt so viele Funktionswerte benötigt wie für die zusammengesetzte Trapezregel!

17. Iterative Lösungsverfahren

17.1 Fixpunktiterationen

In der Praxis tritt oft das Problem auf, eine nichtlineare Gleichung oder gar ein System von nichtlinearen Gleichungen lösen zu müssen. Während wir für lineare Gleichungssysteme Verfahren in Kapitel 7 kennengelernt haben, mit denen man in endlich vielen Schritten eine Lösung erhält, ist dies im allgemeinen bei nichtlinearen Gleichungssystemen nicht möglich. Es werden deshalb fast immer *iterative* Verfahren angewendet, bei denen eine Folge von Approximationen konstruiert wird, die gegen die gesuchte Lösung konvergiert.

Definition 17.1 Eine Abbildung Φ heißt **Selbstabbildung** von $D \subset \mathbb{K}^n$, falls $\Phi : D \to D$. Gilt zusätzlich

$$\|\Phi(\mathbf{y}) - \Phi(\mathbf{z})\| \leq L\|\mathbf{y} - \mathbf{z}\| \qquad \forall \mathbf{y}, \mathbf{z} \in D$$

für ein $L < 1$, so heißt Φ **kontrahierend**.

Definition 17.2 Sei $K \subset \mathbb{K}^n$ eine abgeschlossene Menge und Φ eine Selbstabbildung von K. Ein Punkt $\mathbf{x} \in K$ heißt ein **Fixpunkt** von Φ, falls dieser der **Fixpunktgleichung**

$$\mathbf{x} = \Phi(\mathbf{x})$$

genügt.

Nichtlineare Gleichungen werden zumeist als Fixpunktgleichung umgeformt, weil zu ihrer Lösung folgendes iteratives Verfahren naheliegt: für einen geeigneten Startwert $\mathbf{x}_0$ definiert man die Folge $\{\mathbf{x}_i\}_{i\in\mathbb{N}_0}$ durch

$$\mathbf{x}_{i+1} = \Phi(\mathbf{x}_i), \qquad i = 0, 1, 2, \ldots. \tag{17.1}$$

Eine solche Iteration heißt *Fixpunktiteration*.

Satz 17.3 (Banachscher Fixpunktsatz) Sei $K \subset \mathbb{K}^n$ eine abgeschlossene Menge. Ferner sei Φ eine kontrahierende Selbstabbildung von K. Dann existiert genau ein Fixpunkt $\mathbf{x} \in K$ und für jeden Startwert $\mathbf{x}_0 \in K$ konvergiert die durch die Iterationsvorschrift (17.1) definierte Folge $\{\mathbf{x}_i\}_{i\in\mathbb{N}_0}$ gegen diesen Fixpunkt. Ferner gelten die folgenden Fehlabschätzungen:

(i) $\|\mathbf{x} - \mathbf{x}_i\| \leq L\|\mathbf{x} - \mathbf{x}_{i-1}\|$ "Monotonie"

(ii) $\|\mathbf{x} - \mathbf{x}_i\| \leq \dfrac{L^i}{1-L}\|\mathbf{x}_1 - \mathbf{x}_0\|$ "a-piori-Schranke"

(iii) $\|\mathbf{x} - \mathbf{x}_i\| \leq \dfrac{L}{1-L}\|\mathbf{x}_i - \mathbf{x}_{i-1}\|$ "a-posteriori-Schranke"

Beweis. Sei $\mathbf{x}_0 \in K$ beliebig. Wir zeigen zunächst, dass die Folge $\{\mathbf{x}_i\}_{i\in\mathbb{N}_0}$ eine Cauchy-Folge ist. Da $\Phi(\mathbf{z}) \in K$ für beliebige $\mathbf{z} \in K$ ist, gilt $\mathbf{x}_i \in K$ für alle $i \in \mathbb{N}_0$. Für beliebiges $i \in \mathbb{N}$ folgt die Abschätzung

$$\begin{aligned}\|\mathbf{x}_{i+1} - \mathbf{x}_i\| &= \quad \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_{i-1})\| \le L\,\|\mathbf{x}_i - \mathbf{x}_{i-1}\| \\ &= L\,\|\Phi(\mathbf{x}_{i-1}) - \Phi(\mathbf{x}_{i-2})\| \le L^2\|\mathbf{x}_{i-1} - \mathbf{x}_{i-2}\| \\ &= L^2\|\Phi(\mathbf{x}_{i-2}) - \Phi(\mathbf{x}_{i-3})\| \le L^3\|\mathbf{x}_{i-2} - \mathbf{x}_{i-3}\| \\ &= \ldots \le L^i\|\mathbf{x}_1 - \mathbf{x}_0\|.\end{aligned}$$

Eingesetzt in die Dreiecksungleichung

$$\|\mathbf{x}_i - \mathbf{x}_j\| \le \sum_{k=i}^{j-1} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|$$

liefert dies

$$\begin{aligned}\|\mathbf{x}_i - \mathbf{x}_j\| &\le \sum_{k=i}^{j-1} L^k\|\mathbf{x}_1 - \mathbf{x}_0\| = \|\mathbf{x}_1 - \mathbf{x}_0\|\sum_{k=i}^{j-1} L^k = \|\mathbf{x}_1 - \mathbf{x}_0\|L^i\sum_{k=0}^{j-i-1} L^k \\ &= \|\mathbf{x}_1 - \mathbf{x}_0\|L^i\frac{1 - L^{j-i}}{1-L} \le \|\mathbf{x}_1 - \mathbf{x}_0\|\frac{L^i}{1-L} \overset{i\to\infty}{\longrightarrow} 0.\end{aligned}$$

Dies bedeutet, zu jedem $\varepsilon > 0$ existiert ein $N \in \mathbb{N}$, so dass für alle $N < i < j$

$$\|\mathbf{x}_i - \mathbf{x}_j\| \le \varepsilon.$$

Demnach ist $\{\mathbf{x}_i\}_{i\in\mathbb{N}_0}$ eine Cauchy-Folge. Da $K \subset \mathbb{K}^n$ vollständig ist, existiert das Grenzelement

$$\mathbf{x} = \lim_{i\to\infty} \mathbf{x}_i \in K.$$

Ferner ist Φ nach Voraussetzung Lipschitz-stetig, also insbesondere stetig auf K, woraus

$$\mathbf{x} = \lim_{i\to\infty} \mathbf{x}_{i+1} = \lim_{i\to\infty} \Phi(\mathbf{x}_i) = \Phi\Big(\lim_{i\to\infty} \mathbf{x}_i\Big) = \Phi(\mathbf{x})$$

folgt, das heißt, $\mathbf{x} \in K$ ist Fixpunkt von Φ.
Sei $\boldsymbol{\xi} \in K$ ein weiterer Fixpunkt von Φ, dann gilt

$$0 \le \|\boldsymbol{\xi} - \mathbf{x}\| = \|\Phi(\boldsymbol{\xi}) - \Phi(\mathbf{x})\| \le L\|\boldsymbol{\xi} - \mathbf{x}\|,$$

und daher $\|\boldsymbol{\xi} - \mathbf{x}\| = 0$, das heißt, der Fixpunkt ist eindeutig.
Die Monotonie folgt gemäß

$$0 \le \|\mathbf{x} - \mathbf{x}_i\| = \|\mathbf{x} - \Phi(\mathbf{x}_{i-1})\| \le L\|\mathbf{x} - \mathbf{x}_{i-1}\|.$$

Um die Fehlerabschätzungen zu zeigen, wenden wir noch die Dreiecksungleichung an

$$\|\mathbf{x} - \mathbf{x}_i\| \le L\|\mathbf{x} - \mathbf{x}_i + \mathbf{x}_i - \mathbf{x}_{i-1}\| \le L\|\mathbf{x} - \mathbf{x}_i\| + L\|\mathbf{x}_i - \mathbf{x}_{i-1}\|.$$

Hieraus ergibt sich dann

$$\|\mathbf{x} - \mathbf{x}_i\| \le \frac{L}{1-L}\|\mathbf{x}_i - \mathbf{x}_{i-1}\| \le \frac{L^2}{1-L}\|\mathbf{x}_{i-1} - \mathbf{x}_{i-2}\| \le \ldots \le \frac{L^i}{1-L}\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

$\square$

Beispiel 17.4 Jede Lösung des nichtlinearen Gleichungssystems

$$\begin{aligned} x &= 0.7\sin x + 0.2\cos y \\ y &= 0.7\cos x - 0.2\sin y \end{aligned}$$

ist ein Fixpunkt der Abbildung $f : \mathbb{R}^2 \to \mathbb{R}^2$ mit

$$\Phi(x,y) := \begin{bmatrix} 0.7\sin x + 0.2\cos y \\ 0.7\cos x - 0.2\sin y \end{bmatrix}.$$

Für die Jacobimatrix Φ' von Φ,

$$\Phi'(x,y) = \begin{bmatrix} 0.7\cos x & -0.2\sin y \\ -0.7\sin x & -0.2\cos y \end{bmatrix},$$

folgt

$$\begin{aligned} L &:= \|\Phi'(x,y)\|_F \\ &= \sqrt{0.49\cos^2 x + 0.04\sin^2 y + 0.49\sin^2 x + 0.04\cos^2 y} \\ &= \sqrt{0.53} \approx 0.728. \end{aligned}$$

Also ist $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ kontrahierend, und besitzt nach dem Banachschen Fixpunktsatz in $\mathbb{R}^2$ genau einen Fixpunkt $(\xi,\eta)^\star$. Die Folge

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} := \Phi(x_i, y_i)$$

konvergiert für jeden Startvektor $(x,y)^\star \in \mathbb{R}^2$ gegen $(\xi,\eta)^\star$.
Wir wählen $x_0 = y_0 = 0$ und verlangen

$$\left\| \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \xi \\ \eta \end{bmatrix} \right\|_2 \le 10^{-4}.$$

Unter Verwendung der a-priori-Fehlerabschätzung

$$\left\| \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \xi \\ \eta \end{bmatrix} \right\|_2 \le \frac{L^i}{1-L} \left\| \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} - \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \right\|_2$$

ergibt sich $i \ge 33$ als hinreichende Bedingung. Die folgende Tabelle enthält neben x_i und y_i die Schranken

$$e_i := \frac{L}{1-L} \left\| \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} x_{i-1} \\ y_{i-1} \end{bmatrix} \right\|_2$$

aus der a-posteriori-Fehlerabschätzung.

i	x_i	y_i	e_i
1	0.200000	0.700000	1.9
2	0.292037	0.557203	0.45
3	0.371280	0.564599	0.21
4	0.422927	0.545289	0.14
5	0.458297	0.534591	$0.99 \cdot 10^{-1}$
$\vdots$			
10	0.518525	0.511306	$0.13 \cdot 10^{-1}$
$\vdots$			
20	0.526420	0.507964	$0.16 \cdot 10^{-3}$
21	0.526456	0.507948	$0.11 \cdot 10^{-3}$
22	0.526480	0.507938	$0.69 \cdot 10^{-4}$
$\vdots$			
33	0.526522	0.507920	$0.57 \cdot 10^{-6}$

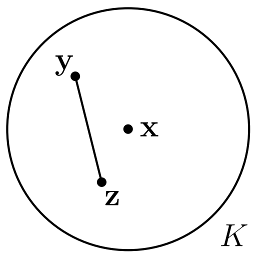
Offensichtlich wird die Fehlerschranke $\varepsilon = 10^{-4}$ bereits nach $i = 22$ Iterationen unterschritten. $\triangle$

Der Banachsche Fixpunktsatz garantiert die Existenz eines Fixpunktes und die Konvergenz der Fixpunktiteration. In der Praxis ist die Existenz eines Fixpunktes meist bekannt (beispielsweise durch graphische Betrachtungen) und sogar seine ungefähre Lage. In dieser Situation besteht die Relevanz des Banachschen Fixpunktsatzes darin, ein relativ einfach zu überprüfendes Kriterium für die Konvergenz der Fixpunktiteration zu liefern:

Korollar 17.5 Sei $D \subset \mathbb{K}^n$ eine offene Menge und $\Phi : D \to \mathbb{K}^n$ stetig differenzierbar, und $\mathbf{x} \in D$ ein Fixpunkt von Φ. Ferner sei $\|\cdot\|_V$ eine Norm in $\mathbb{K}^n$ und $\|\cdot\|_M$ eine verträgliche Matrixnorm, für die $\|\Phi'(\mathbf{x})\|_M < 1$. Dann gibt es ein $\varepsilon > 0$ derart, dass für jedes $\mathbf{x}_0$ mit $\|\mathbf{x} - \mathbf{x}_0\|_V \leq \varepsilon$ die Folge $\{\mathbf{x}_i\}_{i\in\mathbb{N}_0}$, definiert durch die Fixpunktiteration $\mathbf{x}_{i+1} = \Phi(\mathbf{x}_i)$, gegen $\mathbf{x}$ konvergiert.

Beweis. Sei $\delta := 1 - \|\Phi'(\mathbf{x})\|_M > 0$. Wegen der Stetigkeit von Φ' lässt sich dann ein $\varepsilon > 0$ finden, so dass $\|\Phi'(\mathbf{y})\|_M \leq 1 - \delta/2$ für alle $\mathbf{y} \in K := \{\mathbf{z} \in \mathbb{K}^n : \|\mathbf{x} - \mathbf{z}\|_V \leq \varepsilon\} \subset D$. Für alle $\mathbf{y}, \mathbf{z} \in K$ folgt nach dem Mittelwertsatz der Integralrechnung.

$$\Phi(\mathbf{y}) - \Phi(\mathbf{z}) = \int_0^1 \Phi'\big((1-t)\mathbf{y} + t\mathbf{z}\big)(\mathbf{y} - \mathbf{z})\,\mathrm{d}t.$$



Da mit $\mathbf{y}$ und $\mathbf{z}$ auch deren Verbindungsstrecke

$$[\mathbf{y}, \mathbf{z}] := \{(1-t)\mathbf{y} + t\mathbf{z} : t \in [0,1]\}$$

in K liegt, folgt

$$\|\Phi(\mathbf{y}) - \Phi(\mathbf{z})\|_V \leq \int_0^1 \underbrace{\big\|\Phi'\big((1-t)\mathbf{y} + t\mathbf{z}\big)\big\|_M}_{\leq (1-\frac{\delta}{2})} \|\mathbf{y} - \mathbf{z}\|_V \,\mathrm{d}t \leq \left(1 - \frac{\delta}{2}\right)\|\mathbf{y} - \mathbf{z}\|_V.$$

Dies bedeutet, Φ ist eine Kontraktion auf der abgeschlossenen Menge K. Die Behauptung des Korollars folgt nun aus dem Banachschen Fixpunktsatz, wenn wir zeigen können, dass

$\Phi(\mathbf{z}) \in K$ für alle $\mathbf{z} \in K$. Sei hierzu $\mathbf{z} \in K$ beliebig. Dann gilt wegen $\mathbf{x} = \Phi(\mathbf{x})$

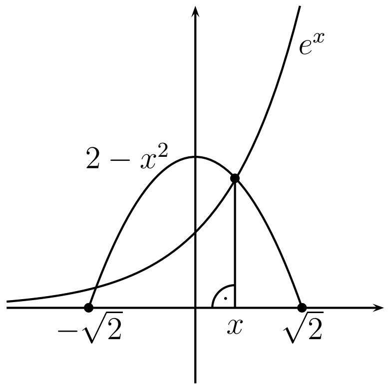
$$\begin{aligned}\|\mathbf{x} - \Phi(\mathbf{z})\|_V = \|\Phi(\mathbf{x}) - \Phi(\mathbf{z})\|_V &\leq \int_0^1 \underbrace{\big\|\Phi'\big((1-t)\mathbf{x} + t\mathbf{z}\big)\big\|_M}_{\leq (1-\frac{\delta}{2})} \underbrace{\|\mathbf{x} - \mathbf{z}\|_V}_{\leq \varepsilon} \, \mathrm{d}t \\ &\leq \left(1 - \frac{\delta}{2}\right)\varepsilon \leq \varepsilon,\end{aligned}$$

wobei wir wieder $[\mathbf{x}, \mathbf{z}] \subset K$ verwendet haben. □

Beispiel 17.6 Gesucht werde eine Lösung der nichtlinearen Gleichung

$$2 - x^2 - e^x = 0.$$

Durch graphische Überlegungen sieht man, dass es genau eine positive Lösung $x \approx 0.5$ gibt:



Für $x > 0$ kann die Gleichung in verschiedener Weise in eine Fixpunktgleichung umgeformt werden:

$$x = \sqrt{2 - e^x} =: \Phi_1(x), \qquad x = \ln(2 - x^2) =: \Phi_2(x),$$

Die Fixpunktiterationen, die auf diesen beiden Iterationsfunktionen basieren, verhalten sich aber unterschiedlich, wenn man mit $x_0 = 0.5$ startet, wie nachfolgende Tabelle zeigt:

i	$x_{i+1} = \Phi_1(x_i)$	$x_{i+1} = \Phi_2(x_i)$
0	0.592687716508341	0.559615787935423
1	0.437214425050104	0.522851128605001
2	0.672020792350124	0.546169619063046
3	0.204473907097276	0.531627015197373
4	0.879272743474883	0.540795632739194
5	Abbruch $(2 - e^{0.87} < 0)$	0.535053787215218
6		0.538664955236433
7		0.536399837485597
8		0.537823020842571
9		0.536929765486145

Die Fixpunktiteration $x_{i+1} = \Phi_1(x_i)$ konvergiert nicht (sie bricht sogar ab), während die Iteration $x_{i+1} = \Phi_2(x_i)$ gegen den korrekten Wert $x = 0.5372744491738\ldots$ konvergiert. Korollar 17.5 erklärt das unterschiedliche Verhalten: Am Fixpunkt gilt

$$\Phi_1'(x) \approx -1.59, \qquad \Phi_2'(x) \approx -0.62.$$

Wir erwarten deshalb Konvergenz der auf Φ_2 basierenden Iteration für hinreichend nahe am Fixpunkt gelegene Startwerte. Für die auf Φ_1 basierende Iteration ist Korollar 17.5 nicht anwendbar. Man kann sogar zeigen, dass für stetig differenzierbare Iterationsfunktionen Φ die Bedingung $|\Phi'(x)| > 1$ zu Divergenz der Fixpunktiteration führt. △

Um einen Vergleich von verschiedenen Fixpunktverfahren zu ermöglichen, will man die Konvergenzgeschwindigkeit messen können. Dazu führen wir den Begriff der Konvergenzordnung ein.

Definition 17.7 Sei $\{\varepsilon_k\}_{k\in\mathbb{N}_0}$ eine Folge positiver reeller Zahlen mit $\varepsilon_k \to 0$ für $k \to \infty$. Wir sagen, dass die Konvergenz (mindestens) die Ordnung $p \geq 1$ hat, wenn ein $C > 0$ existiert, so dass

$$\varepsilon_{k+1} \leq C\varepsilon_k^p.$$

Ist $p = 1$, so fordert man zusätzlich, dass $C < 1$!

Beispiel 17.8

- Der Fall $p = 1$ ist von besonderer Bedeutung und uns bereits im Zusammenhang mit dem Banachschen Fixpunktsatz 17.3 begegnet. In diesem Fall spricht man von *linearer Konvergenz*.

- Der Fall $p = 2$ wird uns im Abschnitt 17.3 begegenen, im Zusammenhang mit dem *Newton-Verfahren*. Hier spricht man von *quadratischer Konvergenz*.

△

Bei nichtlinearen Problemen ist es wichtig, zwischen lokaler und globaler Konvergenz zu unterscheiden:

Definition 17.9 Ein Iterationsverfahren mit Iterierten $\mathbf{x}_i \in \mathbb{K}^n$ heißt **lokal konvergent** gegen $\mathbf{x} \in \mathbb{K}^n$, falls eine Umgebung $U \subset \mathbb{K}^n$ um $\mathbf{x} \in U$ existiert, so dass

$$\mathbf{x}_i \xrightarrow{i\to\infty} \mathbf{x} \qquad \forall \mathbf{x}_0 \in U.$$

Man spricht von **globaler Konvergenz**, falls $U = \mathbb{K}^n$.

17.2 Iterationsverfahren für lineare Gleichungssyteme

Wenn die Matrizen sehr groß sind, verbieten sich direkte Löser zur Lösung eines linearen Gleichungssystems $\mathbf{A}\mathbf{x} = \mathbf{b}$ wegen ihres $\mathcal{O}(n^3)$-Aufwands. Zudem sind die großen, in der Praxis auftretenden Systeme ($n \gtrsim 10^5$) meist dünn besetzt, das heißt, nur wenige ($\lesssim 10$) Einträge in jeder Zeile sind ungleich Null. Während die Matrix eines solchen Problems noch gut in den Speicher passen mag, trifft dies für die L- und R-Faktoren in der Gaußelimination in der Regel nicht mehr zu ("fill-in"). In solchen Fällen behilft man sich gerne mit Iterationsverfahren.

Am einfachsten ist hierbei vermutlich das *Gesamtschrittverfahren* oder *Jacobi-Verfahren*:

Algorithmus 17.10 (Gesamtschrittverfahren)

input: Matrix $\mathbf{A} = [a_{i,j}] \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} = [b_i] \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}^{(0)} = [x_i^{(0)}] \in \mathbb{K}^n$

output: Folge von Iterierten $\{\mathbf{x}^{(k)}\}_{k\in\mathbb{N}}$ mit $\mathbf{x}^{(k)} = [x_i^{(k)}] \in \mathbb{K}^n$

① Initialisierung: $k = 0$

② für $i = 1, 2, \ldots, n$ setze

$$x_i^{(k+1)} := \frac{1}{a_{i,i}}\Big(b_i - \sum_{j\neq i} a_{i,j}x_j^{(k)}\Big)$$

③ erhöhe $k := k + 1$ und gehe nach ②

Vorausgesetzt werden muss offensichtlich, dass $a_{i,i} \neq 0$ ist für alle $i = 1, 2, \ldots, n$. Die Frage nach der Konvergenz ist dadurch jedoch nicht beantwortet. Sicher ist nur, dass bei vorliegender Konvergenz die Iterierten $\mathbf{x}^{(k)}$ gegen eine Lösung von $\mathbf{Ax} = \mathbf{b}$ konvergieren.

Das Gesamtschrittverfahren lässt sich auch in Matrixnotation formulieren. Dazu zerlegen wir (aus historischen Gründen mit "–")

$$\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$$

in eine Diagonal- und in *strikte* linke untere und rechte obere Dreiecksmatrizen. Dann ist

$$\mathbf{x}^{(k+1)} = \mathbf{D}^{-1}\big(\mathbf{b} + (\mathbf{L} + \mathbf{R})\mathbf{x}^{(k)}\big). \tag{17.2}$$

Beim *Einzelschritt-* oder *Gauß-Seidel-Verfahren* verwendet man in ② bereits alle berechneten Komponenten von $\mathbf{x}^{(k+1)}$:

Algorithmus 17.11 (Einzelschrittverfahren)

input: Matrix $\mathbf{A} = [a_{i,j}] \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} = [b_i] \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}^{(0)} = [x_i^{(0)}] \in \mathbb{K}^n$

output: Folge von Iterierten $\{\mathbf{x}^{(k)}\}_{k\in\mathbb{N}}$ mit $\mathbf{x}^{(k)} = [x_i^{(k)}] \in \mathbb{K}^n$

① Initialisierung: $k = 0$

② für $i = 1, 2, \ldots, n$ setze

$$x_i^{(k+1)} := \frac{1}{a_{i,i}}\Big(b_i - \sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{j>i} a_{i,j}x_j^{(k)}\Big)$$

③ erhöhe $k := k + 1$ und gehe nach ②

Entsprechend zu (17.2) erhält man die Matrixformulierung, indem man alle Komponenten von $\mathbf{x}^{(k+1)}$ in ② auf die linke Seite bringt. Dann folgt

$$a_{i,i}x_i^{(k+1)} + \sum_{j<i} a_{i,j}x_j^{(k+1)} = b_i - \sum_{j>i} a_{i,j}x_j^{(k)}, \qquad i = 1, 2, \ldots, n,$$

das heißt, $\mathbf{x}^{(k+1)}$ ergibt sich durch Auflösen des Dreiecksystems

$$(\mathbf{D} - \mathbf{L})\mathbf{x}^{(k+1)} = \mathbf{b} + \mathbf{R}\mathbf{x}^{(k)},$$

also

$$\mathbf{x}^{(k+1)} = (\mathbf{D}-\mathbf{L})^{-1}\big(\mathbf{b}+\mathbf{R}\mathbf{x}^{(k)}\big). \tag{17.3}$$

Verschiedene, mehr oder weniger praktikable, Konvergenzkriterien für das Gesamtschrittverfahren (17.2) und das Einzelschrittverfahren (17.3) sind in der Literatur bekannt. Wir wollen uns auf nachfolgendes, einfach nachzuweisendes Kriterium beschränken.

Definition 17.12 Eine Matrix $\mathbf{A}$ heißt **strikt diagonaldominant**, falls

$$|a_{i,i}| > \sum_{j\neq i} |a_{i,j}| \quad \text{für alle} \quad i = 1,2,\ldots,n.$$

Satz 17.13 Ist $\mathbf{A}$ strikt diagonaldominant, dann konvergieren Gesamt- und Einzelschrittverfahren für jeden Startvektor $\mathbf{x}^{(0)} \in \mathbb{K}^n$ gegen die eindeutige Lösung von $\mathbf{A}\mathbf{x} = \mathbf{b}$.

Beweis. Da sowohl Gesamt-, als auch Einzelschrittverfahren, Fixpunktiterationen $\mathbf{x}^{(k+1)} = \Phi(\mathbf{x}^{(k)})$ sind, wobei die Abbildung Φ sogar affin ist, können wir den Banachschen Fixpunktsatz 17.3 anwenden. Da wir $K = \mathbb{K}^n$ wählen können, müssen wir lediglich noch zeigen, dass Φ eine Kontraktion ist.

Zunächst betrachten wir das Gesamtschrittverfahren. Zu zeigen ist, dass ein $L < 1$ existiert mit

$$\|\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})\mathbf{w}\| \le L\|\mathbf{w}\| \tag{17.4}$$

für alle $\mathbf{w} \in \mathbb{K}^n$. Aus der strikten Diagonaldominanz von $\mathbf{A}$ folgt

$$\|\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})\|_\infty = \max_{1\le i\le n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{i,j}|}{|a_{i,i}|} < 1,$$

das ist (17.4) mit $\|\cdot\| = \|\cdot\|_\infty$ und $L = \|\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})\|_\infty$.

Für das Einzelschrittverfahren ist der Beweis komplizierter. Wieder verwenden wir ∞-Norm und müssen entsprechend zu (17.4) nachweisen, dass

$$\max_{\|\mathbf{x}\|_\infty=1} \|(\mathbf{D}-\mathbf{L})^{-1}\mathbf{R}\mathbf{x}\|_\infty = \|(\mathbf{D}-\mathbf{L})^{-1}\mathbf{R}\|_\infty < 1. \tag{17.5}$$

Sei also $\|\mathbf{x}\|_\infty = 1$ und $L < 1$ wie zuvor definiert. Die einzelnen Komponenten y_i von $\mathbf{y} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{R}\mathbf{x}$ ergeben sich gemäß Algorithmus 17.11 aus

$$y_i := \frac{1}{a_{i,i}}\bigg(-\sum_{j<i} a_{i,j}y_j - \sum_{j>i} a_{i,j}x_j\bigg). \tag{17.6}$$

Wir zeigen induktiv, dass $|y_i| \le L < 1$ für alle $i = 1,2,\ldots,n$ gilt. Hierzu schätzen wir in (17.6) $|y_i|$ mit der Dreiecksungleichung und der Induktionsannahme ab (der Fall $i = 1$ ist klar):

$$\begin{aligned}|y_i| &\le \frac{1}{|a_{i,i}|}\Big(\sum_{j<i}|a_{i,j}||y_j| + \sum_{j>i}|a_{i,j}||x_j|\Big)\\ &\le \frac{1}{|a_{i,i}|}\Big(\sum_{j<i}|a_{i,j}|L + \sum_{j>i}|a_{i,j}|\|\mathbf{x}\|_\infty\Big)\\ &\le \frac{1}{|a_{i,i}|}\Big(\sum_{j<i}|a_{i,j}| + \sum_{j>i}|a_{i,j}|\Big) \le L.\end{aligned}$$

Hieraus folgt $\|\mathbf{y}\|_\infty \le L$ und somit (17.5). □

17.3 Newton-Verfahren

Das *Newton-Verfahren* und seine Varianten sind wohl die wichtigsten Verfahren zum Lösen von nichtlinearen Gleichungen.

Die Funktion $f : \mathbb{K}^n \to \mathbb{K}^n$ sei differenzierbar. Wir wollen die Nullstelle $\mathbf{x} \in \mathbb{K}^n$ der nichtlinearen Gleichung

$$f(\mathbf{x}) = \mathbf{0}$$

finden. Ist $\mathbf{x}_0 \in \mathbb{K}^n$ ein Nährerungswert an diese Lösung, dann approximieren wir

$$f(\mathbf{x}) = \mathbf{0} \approx f(\mathbf{x}_0) + f'(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

Falls $\big(f'(\mathbf{x}_0)\big)^{-1} \in \mathbb{K}^{n\times n}$ existiert, so folgt

$$\mathbf{x} \approx \mathbf{x}_0 - \big(f'(\mathbf{x}_0)\big)^{-1} f(\mathbf{x}_0).$$

Setzen wir

$$\mathbf{x}_1 := \mathbf{x}_0 - \big(f'(\mathbf{x}_0)\big)^{-1} f(\mathbf{x}_0)$$

so ist $\mathbf{x}_1$ möglicherweise eine bessere Näherung an $\mathbf{x}$. Dies ist in einer geeigneten Umgebung von $\mathbf{x}$ wahrscheinlich der Fall. Daher ist es naheliegend, das folgende Iterationsverfahren

$$\mathbf{x}_{i+1} := \mathbf{x}_i - \big(f'(\mathbf{x}_i)\big)^{-1} f(\mathbf{x}_i), \qquad i = 0, 1, 2, \ldots \tag{17.7}$$

zum Auffinden einer Nullstelle zu verwenden. Dies ist das bekannte Newton-Verfahren. Es ist von der Form (17.1), das heißt eine Fixpunktiteration. Wann und wie schnell dieses Verfahren konvergiert, beantwortet der nächste Satz.

Satz 17.14 Sei $D \subset \mathbb{K}^n$ offen und konvex, $\|\cdot\|_V$ eine Norm in $\mathbb{K}^n$ und $\|\cdot\|_M$ eine verträgliche Matrixnorm. Ferner sei $f : D \to \mathbb{K}^n$ eine stetig differenzierbare Funktion mit invertierbarer Jacobimatrix $f'(\mathbf{z})$ mit

$$\big\|\big(f'(\mathbf{z})\big)^{-1}\big\|_M \le \alpha$$

für alle $\mathbf{z} \in D$. Zusätzlich sei $f'(\mathbf{z})$ auf D Lipschitz-stetig mit der Konstanten β,

$$\|f'(\mathbf{y}) - f'(\mathbf{z})\|_M \le \beta\|\mathbf{y} - \mathbf{z}\|_V, \qquad \mathbf{y}, \mathbf{z} \in D.$$

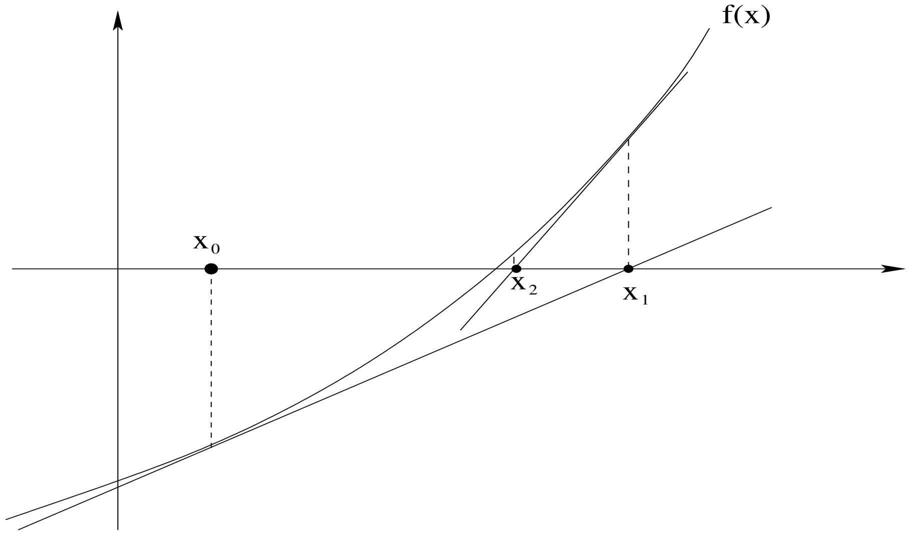


Abbildung 17.1: Geometrische Interpretation des Newton-Verfahrens.

Der Punkt $\mathbf{x} \in D$ sei eine Nullstelle, das heißt $f(\mathbf{x}) = \mathbf{0}$, und $\mathbf{x}_0$ eine Startnäherung mit

$$\mathbf{x}_0 \in K := \{\mathbf{z} \in \mathbb{K}^n : \|\mathbf{z} - \mathbf{x}\|_V \leq \gamma\},$$

wobei γ hinreichend klein sei, so dass $K \subset D$ und

$$\gamma \leq \frac{2}{\alpha\beta}.$$

Dann bleibt die durch das Newton-Verfahren (17.7) definierte Folge $\{\mathbf{x}_i\}_{i \in \mathbb{N}_0}$ innerhalb der Kugel K und konvergiert quadratisch gegen $\mathbf{x}$, das heißt

$$\|\mathbf{x}_{i+1} - \mathbf{x}\|_V \leq \frac{\alpha\beta}{2}\|\mathbf{x}_i - \mathbf{x}\|_V^2, \qquad i = 0, 1, 2, \ldots.$$

Beweis. Wegen (17.7) und $f(\mathbf{x}) = \mathbf{0}$ hat man für $\mathbf{x}_i \in D$

$$\begin{aligned}
\mathbf{x}_{i+1} - \mathbf{x} &= \mathbf{x}_i - \big(f'(\mathbf{x}_i)\big)^{-1} f(\mathbf{x}_i) - \mathbf{x} \\
&= \mathbf{x}_i - \mathbf{x} - \big(f'(\mathbf{x}_i)\big)^{-1}\big[f(\mathbf{x}_i) - f(\mathbf{x})\big] \\
&= \big(f'(\mathbf{x}_i)\big)^{-1}\big[f(\mathbf{x}) - f(\mathbf{x}_i) - f'(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i)\big].
\end{aligned}$$

Daher gilt

$$\begin{aligned}
\|\mathbf{x}_{i+1} - \mathbf{x}\|_V &\leq \underbrace{\Big\|\big(f'(\mathbf{x}_i)\big)^{-1}\Big\|_M}_{\leq\alpha} \big\|f(\mathbf{x}) - f(\mathbf{x}_i) - f'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_i)\big\|_V \\
&\leq \alpha \big\|f(\mathbf{x}) - f(\mathbf{x}_i) - f'(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_i)\big\|_V. \qquad (17.8)
\end{aligned}$$

Letzter Term ist nun abzuschätzen. Dazu setzen wir für $\mathbf{y}, \mathbf{z} \in D$

$$g(t) = f\big((1-t)\mathbf{y} + t\mathbf{z}\big)$$

und bemerken, dass $g(0) = f(\mathbf{y})$ und $g(1) = f(\mathbf{z})$ gilt. Nach Voraussetzung ist g differenzierbar und nach der Kettenregel folgt

$$g'(t) = f'\big((1-t)\mathbf{y} + t\mathbf{z}\big)(\mathbf{z} - \mathbf{y}).$$

Also ist wegen der Lipschitz-Stetigkeit

$$\begin{aligned}\|g'(t) - g'(0)\|_V &= \big\|\big[f'\big((1-t)\mathbf{y} + t\mathbf{z}\big) - f'(\mathbf{y})\big](\mathbf{z} - \mathbf{y})\big\|_V \\ &\le \underbrace{\|f'\big((1-t)\mathbf{y} + t\mathbf{z}\big) - f'(\mathbf{y})\|_M}_{\le \beta t \|\mathbf{z}-\mathbf{y}\|_V} \|\mathbf{z} - \mathbf{y}\|_V \\ &\le \beta t \|\mathbf{z} - \mathbf{y}\|_V^2.\end{aligned}$$

Mit

$$f(\mathbf{z}) - f(\mathbf{y}) - f'(\mathbf{y})(\mathbf{z} - \mathbf{y}) = g(1) - g(0) - g'(0) = \int_0^1 g'(t) - g'(0)\,\mathrm{d}t$$

folgt hieraus

$$\|f(\mathbf{z}) - f(\mathbf{y}) - f'(\mathbf{y})(\mathbf{z} - \mathbf{y})\|_V \le \beta \|\mathbf{z} - \mathbf{y}\|_V^2 \int_0^1 t\,\mathrm{d}t = \frac{\beta}{2}\|\mathbf{z} - \mathbf{y}\|_V^2.$$

Wegen (17.8) erhalten wir daher die quadratische Konvergenzrate

$$\|\mathbf{x}_{i+1} - \mathbf{x}\|_V \le \frac{\alpha\beta}{2}\|\mathbf{x}_i - \mathbf{x}\|_V^2. \tag{17.9}$$

Es verbleibt zu zeigen, dass für alle i die Ungleichung $\|\mathbf{x} - \mathbf{x}_i\|_V \le \gamma$ gilt. Da dies nach Voraussetzung für $i = 0$ gilt, bietet sich vollständige Induktion an. Der Induktionschritt ergibt sich aus (17.9) gemäß

$$\|\mathbf{x}_{i+1} - \mathbf{x}\|_V \le \underbrace{\frac{\alpha\beta}{2}\|\mathbf{x}_i - \mathbf{x}\|_V}_{\le 1} \underbrace{\|\mathbf{x}_i - \mathbf{x}\|_V}_{\le \gamma} \le \gamma.$$

□

Beachte: Die Konvergenz des Newton-Verfahrens ist im allgemeinen nur lokal!

Bemerkung: Im Newton-Verfahren wird die inverse Jacobimatrix nicht berechnet, sondern die Iterationsvorschrift (17.7) wie folgt modifiziert:

1. Löse das lineare Gleichungssystem $f'(\mathbf{x}_i)\triangle\mathbf{x}_i = -f(\mathbf{x}_i)$ und
2. setze $\mathbf{x}_{i+1} = \mathbf{x}_i + \triangle\mathbf{x}_i$.

Zur Lösung des linearen Gleichungssystems kann man beispielsweise die LR-Zerlegung verwenden.

Beispiel 17.15 In Beispiel 17.4 haben wir die Lösung $(\xi, \eta)^\star$ des nichtlinearen Gleichungssystems

$$\begin{aligned} x &= 0.7\sin x + 0.2\cos y \\ y &= 0.7\cos x - 0.2\sin y \end{aligned}$$

mit dem Banachschen Fixpunktsatz berechnet. Zum Vergleich soll $(\xi, \eta)^\star$ auch mit dem Newton-Verfahren approximiert werden. Dazu wird das Fixpunktproblem als Nullstellenaufgabe einer Funktion f formuliert, nämlich

$$f(x, y) := \begin{bmatrix} x - 0.7 \sin x - 0.2 \cos y \\ y - 0.7 \cos x + 0.2 \sin y \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Das Newtonsche Iterationsverfahren lautet dann

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} := \begin{bmatrix} x_i \\ y_i \end{bmatrix} - \left(f'(x_i, y_i)\right)^{-1} f(x_i, y_i) \in \mathbb{R}^2.$$

Hierbei ist

$$f'(x, y) = \begin{bmatrix} 1 - 0.7 \cos x & 0.2 \sin y \\ 0.7 \sin x & 1 + 0.2 \cos y \end{bmatrix}$$

mit Determinante

$$\det f'(x, y) = 1 - 0.7 \cos x + 0.2 \cos y - 0.14 \cos x \cos y - 0.14 \sin x \sin y.$$

Folglich gilt

$$\left(f'(x, y)\right)^{-1} = \frac{1}{\det f'(x, y)} \begin{bmatrix} 1 + 0.2 \cos y & -0.2 \sin y \\ -0.7 \sin x & 1 - 0.7 \cos x \end{bmatrix}.$$

Aus der untenstehenden Tabelle erkennt man, dass x_i und y_i ab $i = 4$ auf sechs Stellen genau sind (Startwerte $x_0 = y_0 = 0$). Es gilt sogar $|x_i - \xi| < 5 \cdot 10^{-13}$ für $i \geq 5$ und $|y_i - \eta| < 5 \cdot 10^{-13}$ für $i \geq 6$.

	Banachscher Fixpunktsatz		Newton-Verfahren	
i	x_i	y_i	x_i	y_i
1	0.200000	0.700000	0.666667	0.583333
2	0.292037	0.557203	0.536240	0.508849
3	0.371280	0.564599	0.526562	0.507932
4	0.422927	0.545289	0.526523	0.507920
$\vdots$			$\vdots$	$\vdots$
22	0.526480	0.507938	$\vdots$	$\vdots$
$\vdots$			$\vdots$	$\vdots$
33	0.526522	0.507920	0.526523	0.507920

Exakte Lösung ist $\xi = 0.526522621917$ und $\eta = 0.507919719037$. $\triangle$

17.4 Verfahren der konjugierten Gradienten

Das Verfahren der konjugierten Gradienten von Hestenes und Stiefel (1952), welches auch als cg-Verfahren (von engl.: conjugate gradient method) bekannt ist, ist wohl das effektivste Verfahren zur Lösung großer linearer Gleichungssysteme $\mathbf{Ax} = \mathbf{b}$ mit hermitescher und positiv definiter Matrix $\mathbf{A}$.

Definition 17.16 Ist $\mathbf{A} \in \mathbb{K}^{n\times n}$ hermitesch und positiv definit, dann definiert

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^\star \mathbf{A}\mathbf{y}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{K}^n$$

ein Skalarprodukt. Die induzierte Norm

$$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}}} = \sqrt{\mathbf{x}^\star \mathbf{A}\mathbf{x}}$$

wird **Energienorm** bezüglich $\mathbf{A}$ genannt. Zwei Vektoren $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ heißen **konjugiert** bezüglich $\mathbf{A}$, falls

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^\star \mathbf{A}\mathbf{y} = 0.$$

Lemma 17.17 Seien $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ bezüglich $\mathbf{A}$ konjugierte Richtungen. Für jedes $\mathbf{x}_0 \in \mathbb{K}^n$ liefert die durch

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{d}_i$$

mit

$$\alpha_i = \frac{\mathbf{d}_i^\star \mathbf{r}_i}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}, \qquad \mathbf{r}_i = \mathbf{b}_i - \mathbf{A}\mathbf{x}_i$$

für $i \geq 0$ erzeugte Folge nach (höchstens) n Schritten die Lösung $\mathbf{x}_n = \mathbf{A}^{-1}\mathbf{b}$.

Beweis. Mit dem Ansatz

$$\mathbf{x} - \mathbf{x}_0 = \sum_{j=0}^{n-1} \alpha_j \mathbf{d}_j$$

erhalten wir wegen den Orthogonalitätsrelationen

$$\mathbf{d}_i^\star \mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{d}_i^\star \mathbf{A}\left(\sum_{j=0}^{n-1} \alpha_j \mathbf{d}_j\right) = \sum_{j=0}^{n-1} \alpha_j \underbrace{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_j}_{=0 \text{ falls } i \neq j} = \alpha_i \mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i$$

die Beziehung

$$\alpha_i = \frac{\mathbf{d}_i^\star \mathbf{A}(\mathbf{x} - \mathbf{x}_0)}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}.$$

Weil $\mathbf{d}_i$ zu den anderen Richtungen konjugiert ist, gilt

$$\mathbf{d}_i^\star \mathbf{A}(\mathbf{x}_i - \mathbf{x}_0) = \mathbf{d}_i^\star \mathbf{A}\left(\sum_{j=0}^{i-1} \alpha_j \mathbf{d}_j\right) = \sum_{j=0}^{i-1} \alpha_j \underbrace{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_j}_{=0} = 0.$$

Deshalb ist

$$\alpha_i = \frac{\mathbf{d}_i^\star (\overbrace{\mathbf{A}\mathbf{x}}^{=\mathbf{b}} - \mathbf{A}\mathbf{x}_i + \mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_0)}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i} = \frac{\mathbf{d}_i^\star(\mathbf{b} - \mathbf{A}\mathbf{x}_i)}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i} + \underbrace{\frac{\mathbf{d}_i^\star \mathbf{A}(\mathbf{x}_i - \mathbf{x}_0)}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}}_{=0} = \frac{\mathbf{d}_i^\star \mathbf{r}_i}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}.$$

□

Bemerkung: Der Vektor $\mathbf{r}_i = \mathbf{b} - \mathbf{A}\mathbf{x}_i$ wird *Residuum* genannt. Seine Norm $\|\mathbf{r}_i\|_2$ ist ein Maß für den Fehler im i-ten Schritt. Gilt $\|\mathbf{r}_i\|_2 = 0$, so stimmt $\mathbf{x}_i$ mit der Lösung $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ überein.

Beim Verfahren der konjugierten Gradienten werden die Richtungen $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ nicht von vornherein gewählt, sondern aus dem jeweils aktuellen Residuum $\mathbf{r}_i$ durch Addition einer Korrektur ermittelt. Das Verfahren der konjugierten Gradienten lautet:

Algorithmus 17.18 (cg-Verfahren)

input: Matrix $\mathbf{A} \in \mathbb{K}^{n\times n}$, rechte Seite $\mathbf{b} \in \mathbb{K}^n$ und Startnäherung $\mathbf{x}_0 \in \mathbb{K}^n$

output: Folge von Iterierten $\{\mathbf{x}_k\}_{k>0}$

① Initialisierung: setze $\mathbf{d}_0 = \mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$ und $i := 0$

② berechne

$$\alpha_i := \frac{\mathbf{d}_i^\star \mathbf{r}_i}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i} \tag{17.10}$$

$$\mathbf{x}_{i+1} := \mathbf{x}_i + \alpha_i \mathbf{d}_i \tag{17.11}$$

$$\mathbf{r}_{i+1} := \mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{d}_i \tag{17.12}$$

$$\beta_i := \frac{\mathbf{d}_i^\star \mathbf{A}\mathbf{r}_{i+1}}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i} \tag{17.13}$$

$$\mathbf{d}_{i+1} := \mathbf{r}_{i+1} - \beta_i \mathbf{d}_i \tag{17.14}$$

③ falls $\|\mathbf{r}_{i+1}\|_2 \neq 0$ erhöhe $i := i+1$ und gehe nach ②

Satz 17.19 Solange $\mathbf{r}_i \neq \mathbf{0}$ ist, gelten die folgenden Aussagen:

1. Es ist $\mathbf{d}_j^\star \mathbf{r}_i = 0$ für alle $j < i$.
2. Es gilt $\mathbf{r}_j^\star \mathbf{r}_i = 0$ für alle $j < i$.
3. Die Vektoren $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_i$ sind paarweise konjugiert.

Beweis. Wir bemerken zunächst, dass gilt

$$\mathbf{d}_i^\star \mathbf{r}_{i+1} \overset{(17.12)}{=} \mathbf{d}_i^\star(\mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{d}_i) = \mathbf{d}_i^\star \mathbf{r}_i - \underbrace{\alpha_i}_{\overset{(17.10)}{=} \frac{\mathbf{d}_i^\star \mathbf{r}_i}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}} \mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i = 0. \tag{17.15}$$

Wir wollen nun den Beweis vermittels Induktion führen.

Im Falle $i = 1$ folgen die ersten beiden Aussagen direkt aus (17.15) während sich die dritte wegen

$$\mathbf{d}_0^\star \mathbf{A}\mathbf{d}_1 \overset{(17.14)}{=} \mathbf{d}_0^\star \mathbf{A}\mathbf{r}_1 - \underbrace{\beta_i}_{\overset{(17.13)}{=} \frac{\mathbf{d}_0^\star \mathbf{A}\mathbf{r}_1}{\mathbf{d}_0^\star \mathbf{A}\mathbf{d}_0}} \mathbf{d}_0^\star \mathbf{A}\mathbf{d}_0 = 0$$

ergibt.

Für den Induktionsschritt $i \mapsto i+1$ nehmen wir an, dass alle drei Aussagen für i gelten. Nun ergibt sich die erste Aussage für $i+1$ und $j = i$ wieder aus (17.15) während sie für $j < i$ mit Hilfe der Induktionsannahme aus

$$\mathbf{d}_j^\star \mathbf{r}_{i+1} \overset{(17.12)}{=} \mathbf{d}_j^\star(\mathbf{r}_i - \alpha_i \mathbf{A}\mathbf{d}_i) = \underbrace{\mathbf{d}_j^\star \mathbf{r}_i}_{=0} - \alpha_i \underbrace{\mathbf{d}_j^\star \mathbf{A}\mathbf{d}_i}_{=0} = 0$$

folgt. Wegen

$$\mathbf{r}_j \overset{(17.14)}{=} \mathbf{d}_j + \beta_{j-1}\mathbf{d}_{j-1}, \qquad 1 \le j \le i$$

ergibt sich die zweite Aussage direkt aus der ersten.
Weiter sind $\mathbf{d}_i$ und $\mathbf{d}_{i+1}$ konjugiert, da

$$\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_{i+1} \overset{(17.14)}{=} \mathbf{d}_i^\star \mathbf{A}\mathbf{r}_{i+1} - \underbrace{\beta_i}_{\overset{(17.13)}{=} \frac{\mathbf{d}_i^\star \mathbf{A}\mathbf{r}_{i+1}}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}} \mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i = 0.$$

Für $\mathbf{d}_j$ mit $j < i$ folgt aufgrund der Induktionsannahme

$$\mathbf{d}_j^\star \mathbf{A}\mathbf{d}_{i+1} \overset{(17.14)}{=} \mathbf{d}_j^\star \mathbf{A}(\mathbf{r}_{i+1} - \beta_i \mathbf{d}_i) = \mathbf{d}_j^\star \mathbf{A}\mathbf{r}_{i+1} - \beta_i \underbrace{\mathbf{d}_j^\star \mathbf{A}\mathbf{d}_i}_{=0} = \mathbf{d}_j^\star \mathbf{A}\mathbf{r}_{i+1},$$

und damit wegen der schon bewiesenen zweiten Aussage

$$\overline{\alpha}_j \mathbf{d}_j^\star \mathbf{A}\mathbf{d}_{i+1} = \overline{\alpha}_j \mathbf{d}_j^\star \mathbf{A}\mathbf{r}_{i+1} \overset{(17.12)}{=} (\mathbf{r}_{j+1} - \mathbf{r}_j)^\star \mathbf{r}_{i+1} = \mathbf{r}_{j+1}^\star \mathbf{r}_{i+1} - \mathbf{r}_j^\star \mathbf{r}_{i+1} = 0.$$

Es bleibt nur noch zu zeigen, dass α_j nicht Null werden kann. Angenommen $\alpha_j = 0$, dann folgt

$$\mathbf{d}_j^\star \mathbf{r}_j = 0,$$

und wegen (17.14) ergibt sich

$$0 = (\mathbf{r}_j - \beta_{j-1}\mathbf{d}_{j-1})^\star \mathbf{r}_j = \mathbf{r}_j^\star \mathbf{r}_j - \overline{\beta_{j-1}} \underbrace{\mathbf{d}_{j-1}^\star \mathbf{r}_j}_{=0} = \|\mathbf{r}_j\|_2^2.$$

Dies steht jedoch im Widerspruch zur Voraussetzung $\mathbf{r}_j \neq 0$. □

Bemerkung:

1. Äquivalent zu (17.10) und (17.13), aber effizienter und numerisch stabiler, hat sich die Wahl
$$\alpha_i = \frac{\mathbf{r}_i^\star \mathbf{r}_i}{\mathbf{d}_i^\star \mathbf{A}\mathbf{d}_i}, \qquad \beta_i = -\frac{\mathbf{r}_{i+1}^\star \mathbf{r}_{i+1}}{\mathbf{r}_i^\star \mathbf{r}_i}$$
erwiesen.

2. Das Verfahren der konjugierten Gradienten wird generell als Iterationsverfahren verwendet, das heißt, man bricht die Iteration ab, falls $\|\mathbf{r}_i\|_2$ klein ist. Pro Iterationsschritt wird nur eine Matrix-Vektor-Multiplikation benötigt. Die Konvergenz des Verfahrens hängt dabei stark von der Kondition der Matrix ab. Genauer, es gilt die Fehlerabschätzung
$$\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{A}} \leq \left(\frac{\sqrt{\operatorname{cond}_2 \mathbf{A}} - 1}{\sqrt{\operatorname{cond}_2 \mathbf{A}} + 1} \right)^i \|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{A}},$$
siehe z.B. J. Stoer und R. Bulirsch *Numerische Mathematik II*.

3. Es bezeichne $\mathcal{K}_i$ den sogenannten *Krylov-Raum*
$$\mathcal{K}_i := \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \mathbf{A}^2\mathbf{r}_0, \ldots \mathbf{A}^i\mathbf{r}_0\} = \operatorname{span}\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_i\}.$$
Man kann zeigen, dass die Iterierte $\mathbf{x}_i$ des i-ten Schritts die Funktion
$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\star \mathbf{A}\mathbf{x} - \operatorname{Re}(\mathbf{b}^\star \mathbf{x})$$
unter allen $\mathbf{x} \in \mathcal{K}_i$ minimiert.

Index