

# Kapitel 1

## Zahlen

### Verständnisfragen

#### Sachfragen

1. Wie lauten die Peano-Axiome?
2. Wie können die Zahlensysteme der natürlichen, ganzen, rationalen, reellen und komplexen Zahlen charakterisiert werden?
3. Wie ist der Absolutbetrag einer Zahl definiert? Welche Eigenschaften besitzt er?
4. Was versteht man unter dem Logarithmus einer Zahl zu einer Basis  $a$ ?
5. Beschreiben Sie den Unterschied zwischen einer reellen und einer komplexen Zahl!
6. Was versteht man unter Polarkoordinaten einer komplexen Zahl?
7. Beschreiben Sie die Multiplikation und Division komplexer Zahlen in der Gauß'schen Zahlenebene!
8. Was ist eine Einheitswurzel?
9. Beschreiben Sie den Begriff der Indextransformation bei Summen und Produkten!
10. Beschreiben Sie die Analogie zwischen Summen und Produkten und Schleifen in einer Programmiersprache!
11. Beschreiben Sie die Zahlensysteme mit den Basen  $b = 10$ ,  $b = 2$  und  $b = 16$ !
12. Warum kann das Konstruktionsverfahren für die Dezimalbruchdarstellung einer reellen Zahl nie zu der Situation führen, dass ab irgendeiner Nachkommastelle alle Koeffizienten  $d_i$  Null werden?
13. Wie wird die Darstellung einer gegebenen Dezimalzahl zu einer Basis  $b$  berechnet?
14. Was versteht man unter Festkomma-Darstellung? Wie werden die Grundrechenarten realisiert?
15. Was versteht man unter Gleitkomma-Darstellung?
16. Beschreiben Sie ein Modell der Gleitkomma-Arithmetik für Addition und Multiplikation!

## Methodenfragen

1. Eine allgemeine Potenz  $a^x$  mit Hilfe des natürlichen Logarithmus ausdrücken können.
2. Den Logarithmus zu einer allgemeinen Basis  $a$  mit Hilfe des natürlichen Logarithmus ausdrücken können.
3. Die Rechengesetze für die Logarithmen anwenden können.
4. Die Arithmetik mit komplexen Zahlen durchführen können.
5. Zwischen der kartesischen Darstellung und der Polarkoordinatendarstellung komplexer Zahlen umrechnen können.
6. Eigenschaften komplexer Zahlen in der Gauß'schen Zahlenebene visualisieren können.
7. Allgemeine Summen und Produkte berechnen können.
8. Indextransformationen ausführen können.
9. Die Fakultät berechnen können.
10. Die Darstellung einer gegebenen ganzen oder rationalen Dezimalzahl in einem Stellenwertsystem aufstellen können.
11. Zwischen zwei verschiedenen Zahlensystemen umrechnen können; insbesondere zwischen Dezimalsystem, Dualsystem und Hexadezimalsystem.
12. Rationale und reelle Zahlen in Fest- und Gleitkommazahlen umwandeln können.
13. Die Kenngrößen eines Gleitkommasystems beschreiben können.
14. Gleitkommaarithmetik durchführen können.
15. Matrizen addieren und transponieren können.
16. Matrizen multiplizieren können, mindestens unter Zuhilfenahme des Falk'schen Schemas.

## Übungsaufgaben

1. Weisen Sie nach, dass  $\log_{10}(2)$  irrational ist!

*Lösung:*

Angenommen, es gibt zwei natürliche Zahlen  $m, n$  mit

$$\log_{10}(2) = \frac{m}{n},$$

dann gilt nach Definition des Logarithmus

$$2^n = 10^m.$$

Das ist aber ein offensichtlicher Widerspruch, denn die rechte Seite dieser Gleichung ist durch 5 teilbar, die linke Seite aber nicht!

2. Weisen Sie nach, dass die Näherung  $\text{ld } x \approx \ln x + \log_{10} x$  einen Fehler von weniger als 1% aufweist!

*Lösung:*

Es ist  $\ln(2)^{-1} \approx 1,442\,695\,040\,888\,963$ , deshalb gilt  $\text{ld}(x) \approx 1,442\,695\,040\,888\,963 \cdot \ln(x)$ . Mit dem gleichen Argument gilt  $\log_{10}(x) \approx 1,434\,294\,481\,903\,251 \cdot \ln(x)$ . Der relative Fehler der Darstellung ist dann ungefähr gegeben durch

$$\frac{1,442\,695 - 1,434\,294}{1,442\,695} \approx 0,582\%.$$

3. Implementieren Sie in der Programmiersprache Ihrer Wahl die Näherung aus Aufgabe 2 und  $\text{ld}(x) = \frac{\ln(x)}{\ln(2)}$ . Vergleichen Sie die damit erzielten Ergebnisse!

*Lösung:*

Hier eine C++-Version:

```
const double ln2 = log(2.0);
double logapprox(double x)
{
    return log(x) + log10(x);
}
double ld(double x)
{
    return log(x)/ln2;
}
```

Hier eine Java-Version:

```
static private double ln2 = Math.log(2.0);

static private double approx(double x)
{
    return Math.log(x) + Math.log10(x);
}

static private double ld(double x)
{
    return Math.log(x)/ln2;
}
```

Und hier ein Ergebnis (Pentium III, g++ unter Cygwin):

x-Wert	$\ln(x) + \log_{10}(x)$	$\ln(x) / \ln(2)$
0.5	-0.994177	-1
0.6	-0.732674	-0.736966
0.7	-0.511577	-0.514573
0.8	-0.320054	-0.321928
0.9	-0.151118	-0.152003
1	-1.59239e-16	-1.60171e-16
1.1	0.136703	0.137504
1.2	0.261503	0.263034
1.3	0.376308	0.378512
1.4	0.4826	0.485427
1.5	0.581556	0.584963
1.6	0.674124	0.678072
1.7	0.761077	0.765535
1.8	0.843059	0.847997
1.9	0.920607	0.925999
2	0.994177	1
2.1	1.06416	1.07039
2.2	1.13088	1.1375

2.3	1.19464	1.20163
2.4	1.25568	1.26303
2.5	1.31423	1.32193

4. Stellen Sie die folgenden Zahlen und die dazu konjugiert komplexen Zahlen in der Gauß'schen Zahlenebene dar und berechnen Sie  $|z|$  und  $|\bar{z}|$ :  $z_1 = 3 + 4i$ ,  $z_2 = 1 - 2i$ ,  $z_3 = -2i$ ,  $z_4$  mit  $|z_4| = 5$ ,  $\operatorname{Re}(z_4) = 3$ .

*Lösung:*

Es ist immer  $|z| = |\bar{z}|$ ;  $|z_1| = 5$ ,  $|z_2| = \sqrt{5}$ ,  $|z_3| = 2$ .

$z_4$  liegt auf dem Kreis um den Ursprung mit Radius 5, und auf der Linie  $\operatorname{Re}(z) = 3$ , Dann sind die beiden Schnittpunkte gegeben durch den Imaginärteil  $3^2 + y^2 = 25$ ; damit ergeben sich die beiden Lösungen  $\operatorname{Im}(z_4) = \pm 4$ . Eine Lösung ist also  $z_1$ , die andere gegeben durch  $z_4 = 3 - 4i$ .

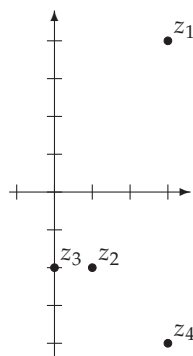


Abbildung 1.1: Lösung zu Aufgabe 5

5. Berechnen Sie die folgenden Summen, Produkte und Quotienten:  $z_1 = (1 + 2i) + (3 + 4i)$ ,  $z_2 = (8 - 5i) + (3 + 9i)$ ,  $z_3 = (3 + 7i)(5 + 6i)$ ,  $z_4 = (7 - i)^2$ ,  $z_5 = \frac{1}{2+i}$ ,  $z_6 = \frac{1}{3-5i}$ ,  $z_7 = (5 + 3i)^{-2}$ .

*Lösung:*

$$z_1 = 4 + 6i, z_2 = 11 + 4i, z_3 = -27 + 53i, z_4 = 48 - 14i, z_5 = \frac{2}{5} - \frac{1}{5}i, z_6 = \frac{3}{34} + \frac{5}{34}i, z_7 = \frac{4}{289} - \frac{15}{578}i.$$

6. Bestimmen Sie die Polarkoordinatendarstellung von  $(1 + i)^{\frac{1}{2}}$ ,  $(-8 + i8\sqrt{3})^{\frac{1}{4}}$  und  $(i)^{\frac{1}{3}}$ .

*Lösung:*

Die beiden Wurzeln für  $(1 + i)^{\frac{1}{2}}$ :

$$z_1 = \sqrt[4]{2}(\cos 22,5^\circ + i \sin 22,5^\circ)$$

$$z_2 = \sqrt[4]{2}(\cos 202,5^\circ + i \sin 202,5^\circ)$$

Die vier Wurzeln für  $(-8 + i8\sqrt{3})^{\frac{1}{4}}$ :

$$z_1 = 2(\cos 30^\circ + i \sin 30^\circ)$$

$$z_2 = 2(\cos 120^\circ + i \sin 120^\circ)$$

$$z_3 = 2(\cos 210^\circ + i \sin 210^\circ)$$

$$z_4 = 2(\cos 300^\circ + i \sin 300^\circ)$$

Die drei Wurzeln für  $i^{\frac{1}{3}}$ :

$$z_1 = \cos 30^\circ + i \sin 30^\circ$$

$$z_2 = \cos 150^\circ + i \sin 150^\circ$$

$$z_3 = \cos 270^\circ + i \sin 270^\circ$$

7. Berechnen Sie die 5-ten Einheitswurzeln und stellen Sie diese in der komplexen Zahlenebene dar!

*Lösung:*

$$z_i = \cos\left((i-1)\frac{2\pi}{5}\right) + i \sin\left((i-1)\frac{2\pi}{5}\right), 1 \leq i \leq 5.$$

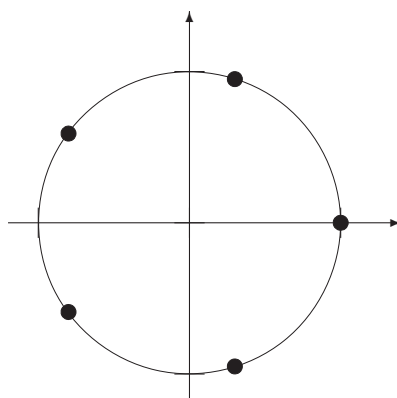


Abbildung 1.2: Lösung zu Aufgabe 7

8. Berechnen Sie mit Hilfe eines Programms die Fakultäten  $6!$ ,  $9!$ ,  $10!$  und  $12!$  und vergleichen Sie die Näherungen mit der Stirling'schen Formel! Vergleichen Sie insbesondere den relativen Fehler mit  $\frac{1}{12n}$ .

*Lösung:*

Die Fakultät als rekursive Funktion:

```
double fakultaet(int n)
{
    if (n==1)
        return 1;
    else
        return n * fakultaet(n-1);
}

double stirling(int n)
{
    return sqrt(2 * M_PI * n) * pow(n/M_E, n);
}

double relativ(int n)
{
    double fak = fakultaet(n);
    return (fak - stirling(n))/fak;
}
```

Und hier die Fakultäten, die Näherung durch die Stirling'sche Formel und der Vergleich des relativen Fehlers mit  $\frac{1}{12n}$ :

```

n! fuer n=6 ist 720
Die Näherung durch die Stirling'sche Formel 710.078
Der relative Fehler: 0.0137803
1/(12n) im Vergleich dazu: 0.0138889

n! fuer n=9 ist 362880
Die Näherung durch die Stirling'sche Formel 359537
Der relative Fehler: 0.00921276
1/(12n) im Vergleich dazu: 0.00925926

n! fuer n=10 ist 3.6288e+06
Die Näherung durch die Stirling'sche Formel 3.5987e+06
Der relative Fehler: 0.00829596
1/(12n) im Vergleich dazu: 0.00833333

n! fuer n=12 ist 4.79002e+08
Die Näherung durch die Stirling'sche Formel 4.75687e+08
Der relative Fehler: 0.00691879
1/(12n) im Vergleich dazu: 0.00694444

```

9. Berechnen Sie für die Zahlen  $x_1 = 5, x_2 = 2, x_3 = 1, x_4 = 2, y_1 = 1, y_2 = 4, y_3 = 3$  und  $y_4 = 1$

$$\sum_{i=1}^4 x_i, \prod_{i=1}^4 x_i, \sum_{i=1}^4 x_i y_i \text{ und } \prod_{i=1}^4 x_i y_i.$$

*Lösung:*

$$\sum_{i=1}^4 x_i = 10, \prod_{i=1}^4 x_i = 20, \sum_{i=1}^4 x_i y_i = 18 \text{ und } \prod_{i=1}^4 x_i y_i = 240.$$

10. Berechnen Sie

$$\sum_{i=1}^{10} \frac{1}{i}, \sum_{i=1}^{10} (i+3) \text{ und } \prod_{i=1}^{10} (6i-2).$$

*Lösung:*

$$\sum_{i=1}^{10} \frac{1}{i} = \frac{7381}{2520} \approx 2,928\,968, \sum_{i=1}^{10} (i+3) = 85, \prod_{i=1}^{10} (6i-2) = 74\,385\,581\,670\,400.$$

11. Wenden Sie das Divisionsverfahren auf die natürliche Zahl  $n = 674$  für die Basen  $b = 5$  und  $b = 2$  an!

*Lösung:*

Für  $b = 5$  ist  $k = 5$ , denn es gilt  $5^4 = 625 \leq 674 < 5^5$ . Das Divisionsverfahren ergibt die Darstellung  $(10144)_5$ :

$$674 = 1 \cdot 5^4 + 49$$

$$49 = 0 \cdot 5^3 + 49$$

$$49 = 1 \cdot 5^2 + 24$$

$$24 = 4 \cdot 5^1 + 4$$

$$4 = 4 \cdot 5^0.$$

Für  $b = 2$  ist  $k = 10$ , das Divisionsverfahren

$$674 = 1 \cdot 2^9 + 162$$

$$162 = 0 \cdot 2^8 + 162$$

$$162 = 1 \cdot 2^7 + 34$$

$$34 = 0 \cdot 2^6 + 34$$

$$34 = 1 \cdot 2^5 + 2$$

$$2 = 0 \cdot 2^4 + 2$$

$$2 = 0 \cdot 2^3 + 2$$

$$2 = 0 \cdot 2^2 + 2$$

$$2 = 1 \cdot 2^1 + 0$$

$$0 = 0 \cdot 2^0.$$

führt zum Ergebnis  $674 = (1010100010)_2$ .

12. Wandeln Sie die Zahl  $(745)_8$  in das System zur Basis  $b = 3$  um!

*Lösung:*

Die Zahl  $(745)_8$  ist als Dezimalzahl gegeben durch  $7 \cdot 64 + 4 \cdot 8 + 5 = 485$ . Zur Basis 3 ist die Darstellung dann  $(122222)_3$  wegen

$$485 = 1 \cdot 3^5 + 242$$

$$242 = 2 \cdot 3^4 + 80$$

$$80 = 2 \cdot 3^3 + 26$$

$$26 = 2 \cdot 3^2 + 8$$

$$8 = 2 \cdot 3^1 + 2$$

$$2 = 2 \cdot 3^0.$$

13. Formal kann mit Darstellungen zu einer Basis  $b$  gerechnet werden, wie wir das vom Dezimalsystem gewohnt sind. Prüfen Sie dies an Hand der Beispiele  $(543)_6 + (242)_6 = (1225)_6$ ,  $(213)_6 - (132)_6 = (41)_6$  und  $(153)_6 \cdot (23)_6 = (4443)_6$  nach!

*Lösung:*

Mit Hilfe des Stellenwertsystems ergibt sich beispielsweise für die Addition

$$\begin{aligned} (543)_6 + (242)_6 &= 5 \cdot 6^2 + 4 \cdot 6 + 3 \cdot 1 + 2 \cdot 6^2 + 4 \cdot 6 + 2 \cdot 1 \\ &= 1 \cdot 6^3 + 2 \cdot 6^2 + 2 \cdot 6^1 + 4 \cdot 1 \\ &= (1225)_6 \end{aligned}$$

Es ist  $8 = (12)_6$  und deshalb

$$\begin{array}{r} ( 5 \ 4 \ 3 )_6 \\ + ( 2 \ 4 \ 2 )_6 \\ \hline (1 \ 2 \ 2 \ 5)_6 \end{array}$$

wenn Sie alle Berechnungen und Überträge wie in der Schule gelernt zur Basis 6 durchführen.

Für die Subtraktion erhalten Sie

$$\begin{array}{r} (2 \ 1 \ 3)_6 \\ - (1 \ 3 \ 2)_6 \\ \hline (0 \ 4 \ 1)_6 \end{array}$$

und für die Multiplikation

$$\begin{array}{r} (153)_6 \cdot (23)_6 \\ \hline (350)_6 \\ (543)_6 \\ \hline (4443)_6 \end{array}$$

14. Berechnen Sie die Entwicklung zur Basis  $b = 7$  von  $\frac{1}{5}$  und von  $\frac{2}{5}$  zur Basis 2!

*Lösung:*

$\frac{1}{5} = 0,2 = (0, \overline{1254})_7$ , denn es ist  $1 = 0 \cdot 5 + 1$  und

$$7 = 1 \cdot 5 + 2$$

$$14 = 2 \cdot 5 + 4$$

$$28 = 5 \cdot 5 + 3$$

$$21 = 4 \cdot 5 + 1.$$

Es ist  $\frac{2}{5} = 0,2 = (0, \overline{0110})_2$  wegen  $2 = 0 \cdot 5 + 2$  und

$$4 = 0 \cdot 5 + 4$$

$$8 = 1 \cdot 5 + 3$$

$$6 = 1 \cdot 5 + 1$$

15. Die Darstellung im Zweierkomplement wird in den Mikroprozessoren von Intel verwendet, mit den Längen  $N = 16, 32$  und  $32$  werden damit die Datentypen *short integer*, *word integer* und *long* realisiert. Darüber hinaus gibt es die Datentypen *unsigned short integer*, *unsigned word integer* und *unsigned long*. Geben Sie die minimal und maximal darstellbaren ganzen Zahlen für diese 6 Datentypen an!

*Lösung:*

Die Lösung finden Sie in Tabelle 1.1.

**Tabelle 1.1:** Lösung von Aufgabe 15

Datentyp	$i_{min}$	$i_{max}$
unsigned short int	0	65 535
short int	-32 768	32 767
unsigned int	0	4 294 967 295
int	-2 147 483 648	2 147 483 647
unsigned long int	0	4 294 967 295
long int	-2 147 483 648	2 147 483 647

16. Wandeln Sie die Dezimalzahlen  $1, 0, -0,5, 6, 625$  und  $3\,456$  in das IEC single-Format um!

*Lösung:*

Die Zahl  $1, 0$  ist gegeben durch  $1_2 \cdot 2^0$ . Also ist die Mantisse gegeben durch  $00\,000\,000\,000\,000\,000\,000\,000$ , die Charakteristik ist gegeben durch  $127 + 0 = 127$ , also durch  $01\,111\,111$ .

$0,5$  ist gegeben durch  $(0,1)_2 \cdot 2^0$ . Durch Normalisieren erhält man den Dualbruch  $(1,0)_2 \cdot 2^{-1}$ . Die Mantisse ist also  $00\,000\,000\,000\,000\,000\,000\,000$ , die Charakteristik  $127 - 1 = 126 = (01\,111\,110)_2$ .

$6,625$  entspricht in Dualdarstellung  $(110,101)_2 \cdot 2^0 = (1,10101)_2 \cdot 2^2$ . Dann ist die Mantisse gegeben durch  $m = 01\,010\,100\,000\,000\,000\,000\,000$ ; die Charakteristik ist  $127 + 2 = 129 = (10\,000\,001)_2$ .

3456 ist als Dualzahl gegeben durch  $(110\ 110\ 000\ 000)_2 \cdot 2^0 = (1,101\ 100\ 000\ 00)_2 \cdot 2^{11}$ . Die Mantisse ist dann gegeben durch  $m = (10\ 110\ 000\ 000\ 000\ 000\ 000\ 000)_2$ , die Charakteristik durch  $c = 127 + 11 = 138 = (10\ 001\ 010)_2$ .

17. Welche Dezimalzahlen werden im IEC single-Format durch die Charakteristik  $c_1 = (10\ 000\ 000)_2$  und die Mantisse  $m_1 = (10\ 000\ 000\ 000\ 000\ 000\ 000\ 000)_2$  beziehungsweise  $c_2 = (10\ 100\ 000)_2$ ,  $m_2 = (00\ 100\ 000\ 000\ 000\ 000\ 000\ 000)_2$  dargestellt?

*Lösung:*

Die erste Bitfolge stellt  $x_1 = 3$ , denn die Charakteristik entspricht der Zahl 128, dadurch ergibt sich der Exponent 1. Die Mantisse entspricht  $(1,1)_2 = 1,5$ . Insgesamt ergibt sich  $x_1 = 1,5 \cdot 2^1 = 3$ .

Die zweite Zahl  $x_2$  hat den Exponenten 33, denn es ist  $(10\ 100\ 000)_2 = 160$ . Die Mantisse entspricht  $(1,001)_2 = 1,125$ . Dann ist  $x_2 = 1,125 \cdot 2^{33}$ .

18. Wie sieht die größte und kleinste darstellbare Gleitkommazahl in einem normalisierten System mit den Parametern  $b, p, e_{min}$  und  $e_{max}$  aus? Welche Zahlen ergeben sich für die IEC-Grundformate?

*Lösung:*

Die kleinste Zahl ist

$$-1, \underbrace{00 \dots 00}_{(p-1)\text{-mal}} 1 \cdot b^{e_{min}},$$

die größte ist

$$1, \underbrace{(b-1)(b-1) \dots (b-1)(b-1)}_{p\text{-mal}} \cdot b^{e_{max}}.$$

Die kleinste Zahl für IEC single ist  $-1, \underbrace{00 \dots 00}_{22\text{-mal}} 1 \cdot 2^{-125}$ , die größte  $1, \underbrace{11 \dots 11}_{23\text{-mal}} \cdot 2^{128}$ .

IEC double berechnet man analog.

19. Weisen Sie nach, dass der Abstand zwischen 1.0 und  $\frac{1}{b}$  durch  $\frac{\epsilon_M}{b}$  gegeben ist!

*Lösung:*

$\frac{1}{b}$  ist gegeben durch  $0,10 \dots 0 \cdot b^0$ . Die Zahlen zwischen  $\frac{1}{b}$  und 1 haben den Abstand  $0,0 \dots 01 \cdot b^0 = b^{-p}$ . Dies entspricht

$$\frac{\epsilon_M}{b} = \frac{b^{1-p}}{b}.$$

20. Definieren Sie für Festkommazahlen mit Skalierungsfaktor  $s$  und Verschiebung  $V$  eine Multiplikation, sodass das Produkt  $x_1 \odot x_2$  der beiden ganzen Zahlen  $x_1, x_2$  dem Produkt der durch sie dargestellten reellen Zahlen entspricht!

*Lösung:*

$f(x_1 \odot x_2) = s(x_1 \odot x_2) + V = f(x_1)f(x_2) = (sx_1 + V)(sx_2 + V)$ . Durch Auflösen erhält man die Darstellung  $x_1 \odot x_2 = sx_1x_2 + V(x_1 + x_2) + \frac{1}{s}V(V-1)$ .

21. Berechnen Sie Summe und Differenz für  $x_1 = 100, x_2 = 99,7$  in einem Gleitkommasystem mit  $b = 10$  und  $p = 2$ .

*Lösung:*

Durch Angleichen der Exponenten ist die Summe gegeben durch  $0,1997 \cdot 10^3 = 0,2 \cdot 10^3 = 200$ .

Die Differenz ist gegeben durch  $0,0003 \cdot 10^3 = 0,3 \cdot 10^0 = 0,3$ .

22. Berechnen Sie im Gleitkommazahlensystem mit  $b = 10, p = 4$  die Summe  $0,4462 \cdot 10^{-4} + 0,2413 \cdot 10^{-3} + 0,1234 \cdot 10^0$  von links nach rechts und von rechts nach links. Notieren Sie die jeweils bei jeder Addition auftretenden Rundungsfehler!

*Lösung:*

Von links nach rechts ergibt sich als erste Summe  $0,2859 \cdot 10^{-3}$ . Durch Angleichen der Exponenten ist dann die Summe gegeben durch  $0,12368592 \cdot 10^0$ , durch Runden ergibt sich dann das Endergebnis  $0,1237 \cdot 10^0$ .

Von rechts nach links: Die erste Summe berechnet sich als  $0,1236413 \cdot 10^0$ , durch Runden ergibt sich  $0,1236 \cdot 10^0$ . Die Addition von  $0,4462 \cdot 10^{-4}$  ergibt nach Runden  $0,1236 \cdot 10^0$ .

Summen sollten immer der Größe nach geordnet berechnet werden; beginnend mit den kleinsten Summanden!

23. Das arithmetische Mittel zweier Zahlen ist definiert als  $\frac{x_1+x_2}{2}$ . Berechnen Sie mit Gleitkommazahlen mit  $b = 10, p = 2$  für  $x_1 = 0.99, x_2 = 0.98$  das arithmetische Mittel. Betrachten Sie den absoluten und den relativen Fehler!

Alternativ ergibt sich das arithmetische Mittel durch  $x_2 + (x_1 - x_2)/2$ . Berechnen Sie das arithmetische Mittel durch diese Formel und vergleichen Sie die Ergebnisse!

*Lösung:*

Für die Summe  $x_1 + x_2$  ergibt sich nach Runden das Ergebnis  $0,2 \cdot 10^1$ . Multiplikation mit  $0,5 \cdot 10^0$  ergibt dann das Endergebnis  $0,1 \cdot 10^1 = 1$ .

Für die Alternative wird zuerst die Differenz  $x_1 - x_2$ , dafür ergibt sich  $0,01 \cdot 10^0$ . Multiplikation dieser Zahl mit  $0,5 \cdot 10^0$  ergibt  $0,01 \cdot 10^0$ ; als Näherung für das arithmetische Mittel insgesamt dann  $0,99 \cdot 10^0$ .

24. Gegeben sind die beiden Rekursionen  $f_0 = 1 - \frac{1}{e}, f_{k+1} = 1 - (k+1)f_k$  und  $g_k = \frac{1-g_{k+1}}{k+1}$  mit einem Startwert  $G_N, N > k$ . Die Rekursion für  $g_k$  ergibt sich durch Auflösen der Rekursion für  $f$  nach  $f_k$ . Berechnen Sie  $f_{30}$  mit dem Computer und vergleichen Sie das Ergebnis mit  $g_{30}$ . Dabei verwenden Sie für  $g_{50}$  einen beliebigen Startwert zwischen  $-10^{10}$  und  $10^{10}$ . Interpretieren Sie die Ergebnisse!

*Lösung:*

Bei der „Rückwärtsiteration“ wird der Fehler, der durch eine ungenaue Zahldarstellung entsteht mit  $\frac{1}{(50-30)!}$ , bei der Vorwärtsiteration mit  $30!$  multipliziert.

Ein Quelltext in C++:

```
// -----
// Implementierung der Integralrekursionen für Aufgabe 24, Kapitel Zahlen
// -----
#include <cmath>
#include "intrek.h"

double forwardIteration(unsigned int n)
{
    unsigned int counter;
    double iterationValue;

    iterationValue = (exp(1.0)-1.0)/exp(1.0);

    for (counter = 1; counter <= n; counter++)
        iterationValue = 1.0 - counter*iterationValue;
    return iterationValue;
}
```

```
double backwardIteration(unsigned int start, unsigned n, double istart)
{
    unsigned int counter;
    double iterationValue;

    iterationValue = istart;

    for (counter = start; counter > n; counter--)
        iterationValue = (1.0-iterationValue)/(double)(counter);
    return iterationValue;
}

// -----
//   Integralrekursion aus Aufgabe 24, Kapitel Zahlen
//   Hauptprogramm
// -----
#include <cmath>
#include <iostream.h>

#include "intrek.h"

int main(void)
{
    char vor='v', zwischen='n';
    int i, N, startN;
    double start=0.0;

    // Anzahl der Nachkommastellen in der Ausgabe hochsetzen!
    cout.precision(12);

    cout << "-----" << endl;
    cout << " Integralrekursion" << endl;
    cout << "-----" << endl;

    cout << "Vorwärts- oder Rückwärtsrekursion?" << endl;
    cout << "v = Vorwärts" << endl;
    cin >> vor;
    if (vor == 'v' || vor == 'V') {
        cout << "Welches n soll denn berechnet werden?" << endl;
        cin >> N;
        cout << "Sollen Zwischenergebnisse ausgegeben werden?" << endl;
        cout << "j/n" << endl;
        cin >> zwischen;
        cout << "Das Ergebnis der Vorwärtsrekursion" << endl;
        cout << "Der Startwert: " << (exp(1)-1.0)/exp(1.0) << endl;
        cout << endl;
        if (zwischen == 'j' || zwischen == 'J') {
            for (i=1; i<=N; i++)
                cout << "k= " << i << ": " << forwardIteration(i) << endl;
        }
        else
            cout << "I_" << N << forwardIteration(N) << endl;
    }
    else {
        cout << "Welches n soll denn berechnet werden?" << endl;
        cin >> N;
        cout << "Von welchem Startindex soll berechnet werden?" << endl;
    }
}
```

```

cin >> startN;
cout << "Welcher Startwert soll verwendet werden?" << endl;
cin >> start;
cout << "Sollen Zwischenergebnisse ausgegeben werden?" << endl;
cout << "j/n" << endl;
cin >> zwischen;
cout << "Das Ergebnis der Rückwärtsrekursion" << endl;
cout << "Der Startwert: " << start << endl;
cout << endl;
if (zwischen == 'j' || zwischen == 'J') {
    for (i=startN; i>= N; i--)
        cout << "k= " << i << ": " <<
            backwardIteration(startN, i, start) << endl;
}
else
    cout << "I_" << N << i" :" <<
        backwardIteration(startN, N, start) << endl;
}
return 0;
}

```

Hier die Ausgabe für die Vorwärts-Iteration:

```

Das Ergebnis der Vorwärts-Iteration
Der Startwert: 0.632120558829

```

```

k= 1: 0.367879441171
k= 2: 0.264241117657
k= 3: 0.207276647029
k= 4: 0.170893411885
k= 5: 0.145532940573
k= 6: 0.126802356562
k= 7: 0.112383504069
k= 8: 0.100931967445
k= 9: 0.0916122929942
k= 10: 0.0838770700583
k= 11: 0.0773522293588
k= 12: 0.0717732476946
k= 13: 0.0669477799697
k= 14: 0.0627310804239
k= 15: 0.0590337936419
k= 16: 0.0554593017296
k= 17: 0.051918705973
k= 18: -0.0294536707515
k= 19: 1.55961974428
k= 20: -30.1923948856
k= 21: 635.040292597
k= 22: -13969.8864371
k= 23: 321308.388054
k= 24: -7711400.3133
k= 25: 192785008.833
k= 26: -5012410228.65
k= 27: 135335076174
k= 28: -3.78938213288e+12
k= 29: 1.09892081854e+14
k= 30: -3.29676245561e+15

```

Im Gegensatz dazu die Ergebnisse der Rückwärts-Iteration, beginnend bei einem beliebigen Startwert für  $I_{50}$  und einem Ergebnis für  $I_{30}$ :

```
Das Ergebnis der Rückwärtsrekursion
Der Startwert für n=50: 1000000
```

```
k= 50: 1000000
k= 49: -19999.98
k= 48: 408.183265306
k= 47: -8.48298469388
k= 46: 0.201765631785
k= 45: 0.0173529210482
k= 44: 0.0218366017545
k= 43: 0.0222309863238
k= 42: 0.0227388142715
k= 41: 0.0232681234697
k= 40: 0.0238227286959
k= 39: 0.0244044317826
k= 38: 0.0250152709799
k= 37: 0.0256574928689
k= 36: 0.0263335812738
k= 35: 0.0270462894091
k= 34: 0.0277986774455
k= 33: 0.0285941565457
k= 32: 0.0294365407107
k= 31: 0.0303301081028
k= 30: 0.0312796739322
```

25. Berechnen Sie die Produkte  $(AB)C$  und  $A(BC)$  für

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 2 & 1 \end{pmatrix}, B = \begin{pmatrix} 4 & -2 & 0 \\ -3 & -1 & -1 \\ 5 & 0 & 2 \end{pmatrix}, C = \begin{pmatrix} -1 & 2 \\ 4 & 0 \\ 3 & 1 \end{pmatrix}.$$

*Lösung:*

$(AB)C = A(BC)$  und

$$ABC = \begin{pmatrix} -29 & -3 \\ -7 & -2 \end{pmatrix}$$

26. Berechnen Sie  $A \cdot B$  mit dem Algorithmus von Strassen:

$$A = \begin{pmatrix} -1 & 2 & 1 & 5 \\ 0 & -3 & 4 & 2 \\ 1 & 5 & 6 & 1 \end{pmatrix}, B = \begin{pmatrix} 2 & 1 & 4 \\ 3 & 5 & 2 \\ 7 & -1 & 5 \\ 0 & 3 & -3 \end{pmatrix}$$

*Lösung:*

Um den Algorithmus von Strassen anzuwenden, müssen beide Matrizen durch Nullen zu  $4 \times 4$  Matrizen aufgefüllt werden.

Dann ergeben sich die folgenden Zwischenmatrizen:

$$\begin{aligned} W_1 &= \begin{pmatrix} 35 & 20 \\ 0 & -15 \end{pmatrix}, W_2 = \begin{pmatrix} 32 & 37 \\ 0 & 0 \end{pmatrix}, \\ W_3 &= \begin{pmatrix} 11 & 0 \\ -15 & 0 \end{pmatrix}, W_4 = \begin{pmatrix} 27 & -14 \\ 0 & 0 \end{pmatrix}, \\ W_5 &= \begin{pmatrix} -21 & 0 \\ 23 & 0 \end{pmatrix}, W_6 = \begin{pmatrix} 27 & 17 \\ 15 & 15 \end{pmatrix}, W_7 = \begin{pmatrix} -72 & 17 \\ 42 & 2 \end{pmatrix}. \end{aligned}$$

Als Endergebnis ergibt sich wie im Buch das Matrixprodukt

$$A \cdot B = \begin{pmatrix} 11 & 23 & -10 \\ 19 & -13 & 8 \\ 59 & 23 & 41 \end{pmatrix},$$

wenn man die überflüssigen Nullen wieder entfernt.

27. Für reelle Zahlen gilt die Kürzungsregel: Aus  $ab = ac$  und  $a \neq 0$  folgt  $b = c$ , und aus  $ad = 0$  folgt  $a = 0 \vee d = 0$ . Gelten diese Regeln für die Matrix-Arithmetik?

*Lösung:*

Für singuläre Matrizen  $A \neq 0$  können die Regeln verletzt werden!