Cosinus und Sinus numerisch effizient annähern.

H.R. Schneebeli

Version vom 1. März 2018

Zusammenfassung

Wie lassen sich Funktionswerte für Sinus oder Cosinus rasch und zuverlässig numerisch annähern? Wir betrachten gleichförmige Kreisbewegungen auf dem Einheitskreis in $\mathbb{C}$ und finden Sinus und Cosinus als Lösungen einer Differentialgleichung. Die beiden Funktionen lassen sich in einem hybriden Verfahren gleichzeitig numerisch approximieren. Das Eulerverfahren wird auf diesen Fall angepasst. Das Ziel ist ein numerisches Verfahren, das stabil ist und rasch zum Ziel führt: Funktionswerte mit mindestens 12 gültigen Dezimalstellen.

Stichworte Komplexe Zahlen, allgemeine binomische Formel, Taylorentwicklung, Exponentialfunktion in $\mathbb{C}$, Differentialgleichungen als Vektorfelder, Eulerverfahren. Verhalten von Grenzwerten unter stetigen Funktionen. Programmierumgebung mit IEEE double precision Arithmetik [z.B. Matlab™, Octave, programmierbare Grafikrechner oder CAS-Rechner]

Ziele Die Funktionen Cosinus und Sinus im Intervall $[0, \pi/2]$ bei einer vorgegebenen maximal zulässigen Toleranz zum Idealwert numerisch effizient annähern.

1 Cosinus und Sinus: Definition, Eigenschaften, Berechnung.

In diesem Abschnitt werden die Funktionen Cosinus und Sinus geometrisch definiert, kinematisch interpretiert und analytisch beschrieben und Schlüsseleigenschaften festgestellt. Der analytischen Funktionsbeschreibung wird die numerische Approximation der Funktionswerte gegenübergestellt.

Die Funktionen Cosinus und Sinus lassen sich in einem ebenen kartesischen Koordinatensystem gleichzeitig definieren. Das kartesische Koordinatenpaar $(\cos(t)|\sin(t))$ beschreibt den Punkt mit den Polarkoordinaten $(1, \angle t)$. Dabei wird angenommen, dass die Ebene orientiert sei und Winkel t im Sinne dieser Orientierung abgetragen werden. Es werden durchwegs Winkel im Bogenmass benutzt, denn für $t \approx 0$ gelten dann die Näherungen $\sin(t) \approx t \approx \tan(t)$.

Die Definition von Cosinus und Sinus lässt sich in kinematischer Einkleidung etwas kompakter fassen: Die Parameterdarstellung

$$\vec{c} : t \mapsto \overrightarrow{OP}(t) := \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix}$$

beschreibt einen Punkt $P(t)$, der sich auf dem Einheitskreis im positiven Drehsinn mit Winkelgeschwindigkeit $\omega = 1$ bewegt und den Punkt $(1|0)$ zur Zeit 0 passiert.

Da der Kreisradius konstant gleich 1 ist, gilt $\|\vec{c}(t)\|^2 = \vec{c}(t) \cdot \vec{c}(t) = 1$ für alle t. Daher gilt für die Ableitung

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\vec{c}(t)\|^2 = 2 \cdot \vec{c}(t) \cdot \vec{c}\,'(t) = 0 \quad \text{für alle } t.$$

Weil das Skalarprodukt immer Null ergibt, stehen der Vektor $\vec{c}$ und seine Ableitung $\vec{c}\,'$ immer senkrecht. Sie bilden (in dieser Reihenfolge) ein Paar von positiv orientierten Vektoren. Da der Betrag der Geschwindigkeit eines Punktes, der auf dem Einheitskreis mit Winkelgeschwindigkeit $\omega = 1$ rotiert, selbst gleich 1 ist, gilt $\|\vec{c}\,'(t)\| = 1$ für alle t. Es folgt

$$\vec{c}\,'(t) = \begin{bmatrix} -\sin(t) \\ \cos(t) \end{bmatrix}.$$

Aus der kinematischen Beschreibung lassen sich unmittelbar die Ableitungsregeln ablesen:

$$\text{Für alle } t \text{ gilt:} \quad \cos'(t) = -\sin(t) \quad \text{und} \quad \sin'(t) = \cos(t)$$

Exakte Funktionswerte lassen sich für einige ausgewählte Werte von t elementar angeben, etwa $\sin(0) = \cos(\pi/2) = 0$ und $\cos(0) = \sin(\pi/2) = 1$ oder $\cos(\pi/4) = \sin(\pi/4) = 1/\sqrt{2}$. Aber was ist beispielsweise der exakte Wert von $\cos(1)$?

Die Analysis hat für beliebige t eine formale Antwort auf diese Frage:

$$\cos(t) := \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\, t^{2n} \qquad \sin(t) := \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}\, t^{2n+1}$$

Die unendlichen Reihen beinhalten Grenzwertbestimmungen. Diese bilden ein Hindernis für exakte Antworten mit numerischen Methoden, die ja nur endlich vielen Operationen gestatten.

Begnügt man sich mit Näherungen innerhalb einer gegebenen Toleranz $\tau > 0$, so ist folgendes Vorgehen klassisch. Die beiden genannten Taylorreihen lassen sich im Prinzip für jedes t durch Taylorpolynome genügend hohen Grades beliebig gut approximieren. Die Abweichungen zwischen den Taylorpolynomen und der zugehörigen Funktion sind analytisch abschätzbar.

Die konkrete Frage nach der Berechnung von $\cos(1)$ lautet dann: Für welchen Grad $2r$ ist

$$T_{2r}(1) := \sum_{n=0}^{r} \frac{(-1)^n}{(2n)!}$$

eine Näherung, die $|T_{2r}(1) - \cos(1)| < \tau$ erfüllt? Das ist keine leichte Frage, denn der exakte Wert $\cos(1)$ ist eine rein formale Grösse, deren numerischer Wert immer ungenau bestimmbar bleibt.

Im Beispiel lässt sich die Berechnung von $T_{2r}(1)$ mit rationalen Zahlen exakt ausführen. Setzt man die Toleranz $\tau := 10^{-16}$, so reicht der Grad $2r = 20$. Daraus ergibt sich als Antwort die rationale Näherung

$$\cos(1) \approx \frac{1314502564969066301}{2432902008176640000}$$

Sie ist in dieser Gestalt kaum lesbar oder verständlich. Ein Taschenrechner verwandelt die Antwort in eine Dezimalzahl $\cos(1) \approx 0.54030230586814$, verschluckt dabei aber typischerweise die 15. und alle folgenden Ziffern.

Aufgaben

1. Es sei $S_{19} : x \mapsto \sum_{r=1}^{19} s_r \cdot x^r$ die Taylorentwicklung für die Sinusfunktion bis zum Grade 19. Welches ist der exakte Wert von $S_{19}(1)$? Welche 14-stellige Dezimalzahl ergibt sich aus daraus?

2. Angenommen, die Funktion $c : t \mapsto c(t)$ ist nur definiert für $0 \leq t \leq \pi/4$ und es gilt auf diesem Definitionsbereich $c(t) = \cos(t)$.

   (a) Es sei $0 \leq t \leq \pi/4$. Wie lässt sich $\sin(t)$ aus $c(t)$ bestimmen?

   (b) Wie lassen sich für beliebige $T \in \mathbb{R}$ $\cos(T)$ und $\sin(T)$ mit der Hilfsfunktion c bestimmen?

2 Cosinus, Sinus und eine Differentialgleichung

Statt eines einzigen Punktes lassen wir alle Punkte der Ebene gleichzeitig um den Nullpunkt kreisen. Damit wird die Bewegung auf dem Einheitskreis eingebettet in eine Drehbewegung der ganzen Ebene, die mit konstanter Winkelgeschwindigkeit 1 um den Nullpunkt rotiert. Wir benötigen diese Einbettung zwar nur in der Nähe des Einheitskreises, aber es kostet keine Mühe, gleich die ganze Ebene einzubeziehen.

Wir betrachten nun die Ortsvektoren in der Ebene als Funktionen der Zeit $\vec{x} : t \mapsto \vec{x(t)}$. Jede dieser Funktionen parametrisiert die Bewegung eines Punktes längs einer Stromlinie. Diese wird festgelegt durch einen Startpunkt $\vec{x}(0)$ und die Bedingung, dass sich jeder Punkt um den Nullpunkt im positiven Drehsinn mit Winkelgeschwindigkeit 1 bewegt.

Zu jedem Bahnpunkt X gehört ein Geschwindigkeitsvektor $\vec{x}\,'$. Gilt $\vec{x}(t) = \overrightarrow{OX}$, so steht der Geschwindigkeitsvektor $\vec{x}\,'(t)$ auf $\vec{x}(t)$ senkrecht. Die Koordinaten der Punkte sind nun Funktionen der Zeit $x : t \mapsto x(t)$ und $y : t \mapsto y(t)$. Falls $x(0) = 1$ und $y(0) = 0$ gilt, so kennen wir schon eine Lösung: Für alle t ist $x(t) = \cos(t)$ und $y(t) = \sin(t)$. Da der Abstand jedes Punktes zum Nullpunkt konstant bleibt, gilt jederzeit $x(t)^2 + y(t)^2 = r^2$. Durch Ableiten nach der Zeit folgt die Orthogonalitätsbedingung für die Koordinatenfunktionen. Sie lautet $x \cdot x' + y \cdot y' = 0$. Da der Betrag der Geschwindigkeit auf jeder Stromlinie konstant ist, folgt, dass die gesamte Bewegung durch die *Differentialgleichung*

$$\begin{aligned} x' &= -y \\ y' &= x \end{aligned}$$

beschrieben wird.

Diese Differentialgleichung beschreibt das Geschwindigkeitsfeld einer Ebene, die sich mit konstanter Winkelgeschwindigkeit um den Ursprung dreht. Jeder der Punkte bewegt sich mit der Strömung in je 2π Zeiteinheiten einmal über seine ganze Bahn. Die Abbildung

$$R : t \mapsto \begin{bmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

beschreibt diese Drehung formal. Der Punkt, der zur Zeit $t = 0$ beim Punkt $(1|0)$ war, befindet sich zur Zeit t dann an der Stelle $(\cos(t)|\sin(t))$. Wenn es also gelingt, gute numerische Näherungslösungen für das Anfangswertproblem

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} -y \\ x \end{bmatrix} \quad \text{mit} \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

zu erzeugen, so ist ein Verfahren gefunden, das gleichzeitig die Funktionswerte von Cosinus und Sinus numerisch approximiert. Von einer numerischen Näherung wird erwartet, dass sie zuverlässig eine vorgeschriebene Genauigkeit erzielt. Eine zuverlässige Näherung gilt als umso besser, je geringer ihr Bedarf an Leistung der Hardware (Rechenzeit, Speicher oder Operationen) ist.

3 Eine elegantere Notation mit komplexen Zahlen

Das Ziel des Abschnittes ist eine Vereinfachung der Denkweise, Sprechweise und Notation. Anstelle der Koordinatenebene $\mathbb{R}^2$ mit Vektoren und Matrizen werden die komplexen Zahlen $\mathbb{C}$ treten. Es ist sinnvoll, von einer Programmierumgebung auszugehen, die den Datentyp 'komplexe Zahl' in Fliesskommadarstellung mit den üblichen arithmetischen Operationen aus $\mathbb{C}$ anbietet.

Wir benutzen die folgende Potenzreihendarstellung als Definition für die Exponentialfunktion im Komplexen.

$$\exp(z) := \sum_{n=0}^{\infty} \frac{1}{n!} z^n$$

Die Exponentialreihe ist für alle komplexen Zahlen z definiert und konvergiert in ganz $\mathbb{C}$.

Durch Vergleich der Potenzreihen der drei Funktionen erkennt man die Eulerrelation

$$\exp(i \cdot t) = \cos(t) + i \cdot \sin(t)\,,$$

die für alle reellen t gilt.

Insbesondere ist $|\exp(i{\cdot}t)| = 1$ für alle $t \in \mathbb{R}$. Die Multiplikation einer beliebigen komplexen Zahl c mit $\exp(i \cdot t)$ bewirkt die Drehung von c um den Nullpunkt mit dem Drehwinkel t [rad].

Die Differentialgleichung im Komplexen

Im wesentlichen sagt die Differentialgleichung

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} -y \\ x \end{bmatrix} \quad \text{mit} \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} := \begin{bmatrix} a \\ b \end{bmatrix},$$

dass der Geschwindigkeitsvektor an einem bestimmten Ort aus dem Ortsvektor durch eine Drehung um $\pi/2$ im positiven Drehsinn hervorgeht.

In der komplexen Schreibweise werden die Stromlinien durch Funktionen $z : t \mapsto z(t)$ erzeugt, die $\mathbb{R}$ nach $\mathbb{C}$ abbilden und die Bedingung

$$\frac{\mathrm{d}z}{\mathrm{d}t} = i \cdot z \quad \text{mit } z(0) = a + i \cdot b$$

erfüllen.

Ihre formale Lösung lautet $z : t \mapsto z(0) \cdot \exp(i \cdot t)$, da für alle Konstanten $c \in \mathbb{C}$ nach der Kettenregel gilt

$$\frac{\mathrm{d}}{\mathrm{d}t} \exp(c \cdot t) = c \cdot \exp(c \cdot t)$$

Aufgaben

3. Wie ergibt sich aus den Bedingungen $\frac{\mathrm{d}f}{\mathrm{d}z} = f$ und $f(0) = 1$ eine Potenzreihendarstellung für f von der Art $f(z) = \sum_{r=0}^{\infty} f_r \cdot z^r$?

4. Wie lässt sich nachweisen, dass für jedes $c \in \mathbb{C}$ die Funktion $f_c : z \mapsto \exp(c \cdot z)$ die Differentialgleichung $f' = c \cdot f$ mit der Anfangsbedingung $f(0) = 1$ löst?

5. Begründen oder widerlegen Sie folgende Behauptung: Die Abbildung

$$R : t \mapsto \begin{bmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

beschreibt, wie $\begin{bmatrix} a \\ b \end{bmatrix}$ um $O \in \mathbb{R}^2$ um den Winkel t gedreht wird. Es ist die reelle Schreibweise für die komplexe Darstellung

$$r : t \mapsto \exp(i \cdot t) \cdot (a + i \cdot b)$$

6. Warum gilt für alle natürlichen Zahlen n und alle reellen t: $\exp(i \cdot t) = \exp(i \cdot t/n)^n$?

4 Die Differentialgleichung und das Verfahren von Euler

Das Anfangswertproblem $z'(t) = i \cdot z(t)$ und $z(0) := 1$ soll numerisch behandelt werden.

Das Verfahren von Euler benutzt den Differenzenquotient als Näherung für die Ableitung. Für $\Delta t \neq 0$, aber $\Delta t \approx 0$ gilt

$$\frac{z(\Delta t) - z(0)}{\Delta t} \approx \frac{\mathrm{d}z}{\mathrm{d}t}(0) = i.$$

Daraus folgt die Näherung

$$\tilde{z}(\Delta t) := z(0) + \frac{\mathrm{d}z}{\mathrm{d}t}(0) \cdot \Delta t = 1 + i \cdot \Delta t \approx z(\Delta t)$$

Diese Näherung ist bei exakter Arithmetik umso besser, je näher Δt bei 0 liegt.

Das Verfahren lässt sich wie folgt plausibel machen. Die Differentialgleichung beschreibt mit $z'(t) = i \cdot z(t)$ den Geschwindigkeitsvektor einer Strömung an der Stelle $z(t)$. Wenn nun ein Punkt zur Zeit 0 mit der Strömung an der Stelle 1 vorbeischwimmt, so hat er in diesem Modell die Momentangeschwindigkeit i. Es ist ganz korrekt, wenn auch ungewohnt, dass die Geschwindigkeit der ebenen Strömung eine komplexe Zahl ist. Der Physik ist es gleichgültig, in welcher Sprache über *Vektoren* gesprochen wird.

In einem kleinen Zeitintervall der Grösse Δt wird der Punkt mit der Strömung um $i \cdot \Delta t$ weiter getrieben und befindet sich nun (näherungsweise!) an der Stelle $\tilde{z}(\Delta t) = 1 + i \cdot \Delta t$.

Wenn wir zwei Zeitschritte der Grösse Δt warten, so befindet sich der Punkt angenähert an der Stelle $1 + 2i \cdot \Delta t$, falls wir nur die Information im Startpunkt benutzen. Aber wenn $2\Delta t$ zerlegt wird als $\Delta t + \Delta t$ und die Näherung $\tilde{z}(\Delta t) \approx 1 + i \cdot \Delta t$ für den nächsten Schritt benutzt wird, so ergibt sich eine bessere Näherung:

$$z(2\Delta t) \approx \tilde{z}(\Delta t) + i \cdot \tilde{z}(\Delta t) \cdot \Delta t = 1 + i \cdot \Delta t + i \cdot \Delta t + i^2 \cdot (\Delta t)^2 = (1 + i \cdot \Delta t)^2 = \tilde{z}(\Delta t)^2$$

Natürlich lässt sich dieser Gedankengang beliebig fortsetzen. Wenn für eine beliebige natürliche Zahl $n > 0$ der Zeitschritt $\Delta t := t/n$ eingeführt wird, so erwarten wir als Verallgemeinerung die Beziehung

$$z(t) = z(n \cdot \Delta t) = z(\Delta t)^n \approx \tilde{z}(\Delta t)^n = (1 + i \cdot \frac{t}{n})^n$$

Wir betrachten kurz das Verhalten einer geometrischen Folge $g : r \mapsto q^r$ mit komplexem q. Multiplikation mit q erzeugt in $\mathbb{C}$ eine Drehstreckung, dabei ist $|q|$ der Streckfaktor und $\arg(q)$ der Drehwinkel. Folglich ordnen sich die Werte $g(0) = 1,\ g(1) = q,\ g(2) = q^2, \ldots$ in $\mathbb{C}$ auf einer Spirale an, wobei der Polarwinkel zwischen $g(r)$ und $g(r+1)$ immer gleich $\arg(q)$ ist. Für $0 < |q| < 1$ strebt die Spirale ins Innere des Einheitskreises, für $|q| = 1$ liegen alle $g(r)$ auf dem Einheitskreis und für $|q| > 1$ öffnet sich die Spirale und strebt exponentiell schnell nach aussen.

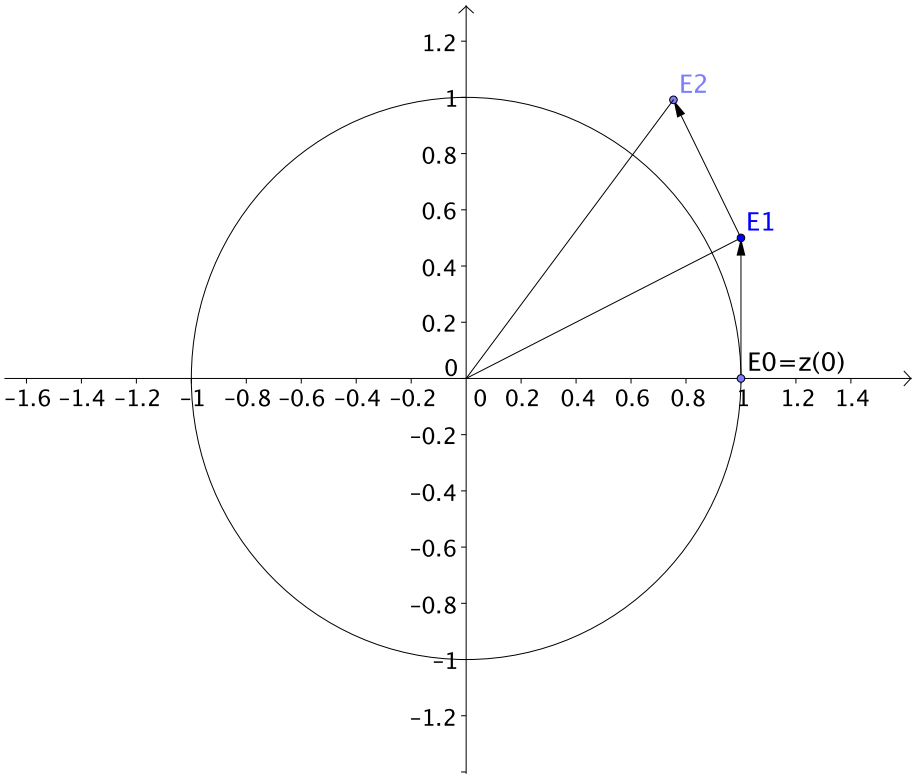

Abbildung 1: Prinzipskizze: Eulernäherung E_r, [dargestellt ist $\tilde{z}(0.5)^r$ und $E_2 \approx \exp(i)$]

Für $t \neq 0$ ist $|1 + i \cdot t/n| = \sqrt{1 + (t/n)^2} > 1$. Folglich winden sich die Näherungen, welche das Eulerverfahren beim Lösen der Differentialgleichung erzeugt, immer vom Einheitskreis nach aussen. Jede der Näherungen ist prinzipiell falsch. Es ist nun wesentlich, dass n auch in $\Delta t := t/n$ auftritt und $|(1 + i \cdot t/n)|^n$ mit $n \to \infty$ selbst gegen 1 strebt, das heisst, dass die Punkte der Folge $\tilde{z} : n \mapsto (1 + i \cdot t/n)^n$ sich von aussen dem Einheitskreis beliebig gut annähern. [Vgl. Aufgabe 12]

Euler hat einen Schritt mehr gewagt. Er *definiert* allgemein für $z \in \mathbb{C}$:

$$\exp(z) := \lim_{n \to \infty} (1 + \frac{z}{n})^n$$

Die Exponentialfunktion ist damit als Grenzwert einer Folge von Polynomen $(1 + z/n)^n$ mit $z \in \mathbb{C}$ und $n \to \infty$ dargestellt.

Wir verfolgen hier nur den Spezialfall $\exp(i \cdot t) := \lim_{n\to\infty}(1 + i \cdot t/n)^n$.

Diese Definition lässt naive numerischen Berechnungen scheitern. Betrachten wir den Fall $t = 1$ und $n := 10^{1000}$. Als Fliesskommazahl wird $1 + i \cdot 10^{-1000} \approx 1.0$ gerundet. Damit ergibt sich die total falsche Näherung $\tilde{z}(1) \approx 1$ statt $z(1) \approx \exp(i) \approx 0.5403 + i \cdot 0.84147$.

Schon besser ist die Näherung $(1+i/n)^n \approx 1+n\cdot i/n = 1+i$, die sich aus der binomischen Formel ergibt.

Aufgaben

7. Stellen Sie die komplexen Zahlen $(1 + i/4)^n$ für $n := 1, 2, 3, 4$ in der komplexen Ebene dar. Sind diese Zahlen mit Zirkel und Lineal aus 1 und i exakt konstruierbar? Begründen Sie Ihre Antwort.

8. Wir betrachten das Anfangswertproblem $z'(t) = i \cdot z(t)$ und $z(0) := z_0$. Begründen Sie exakt, weshalb das Eulerverfahren die Näherung $z(n \cdot \Delta t) \approx z_0 \cdot \tilde{z}(\Delta t)^n$ liefert.

9. Erstellen Sie [mit Hilfe des Rechners] eine Tabelle, welche für jedes n aus der Liste $\{1, 10, 100, \ldots, 10^5\}$ die Zahlen $(1 + i/n)^n$ und $|1 + i/n|^n$ in Dezimaldarstellung mit 6 Ziffern angibt.

   Was zeigen die Ergebnisse? Vergleichen Sie mit dem Rechnerwert $\exp(i)$.

10. Für $\exp(i \cdot t)$ wurden zwei analytische Definitionen erwähnt:

$$\exp(i \cdot t) := \sum_{r=0}^{\infty} \frac{i^r t^r}{r!} \qquad \text{und} \qquad \exp(i \cdot t) := \lim_{n\to\infty} (1 + i\,\frac{t}{n})^n$$

    Was zeigt sich beim Vergleich des Summanden $i^r t^r/r!$ in der Potenzreihe mit jenem, der die Potenz t^r wenn die Faktorisierung $(1 + i \cdot t/n)^n$ ausmultipliziert wird? Was geschieht, wenn n grösser und grösser wird?

11. Angenommen, $e(n) := (1 + i/n)^n$ sei für eine noch wählbare Zahl $n > 1000$ zu berechnen. Warum ist $n := 1024$ eine kluge Wahl und wie gelingt es, $e(1024)$ mit bloss 9 Multiplikationen in $\mathbb{C}$ zu berechnen? Berechnen Sie $e(1024)$ mit einem kurzen selbst geschriebenen Programm. Wie lautet die Antwort?

12. Begründen oder widerlegen Sie folgende Behauptungen: Für jedes natürliche $n > 0$ und beliebige reelle t gilt:

    (a) $|1 + i \cdot t/n|^2 = (1 + i \cdot t/n) \cdot (1 - i \cdot t/n)$

    (b) $|1 + i \cdot t/n| = |1 - i \cdot t/n|$

    (c) $|(1 + i \cdot t/n)^n|^2 = (1 + i \cdot t/n)^n \cdot (1 - i \cdot t/n)^n = (1 + (t/n)^2)^n$

    (d) $\lim_{n\to\infty}(1 - t^2/n^2)^n = \lim_{n\to\infty}((1 + t/n)^n \cdot (1 - t/n)^n) = \exp(t) \cdot \exp(-t) = 1$

    (e) $\lim_{n\to\infty}(1 + (t/n)^2)^n = \exp(i \cdot t) \cdot \exp(-i \cdot t) = 1$

5 Besser, genauer, schneller

5.1 Der erste Schritt: Analysis

Das Eulerverfahren ist ein Beispiel für einen einfachen *numerischen Differentialgleichungslöser*. Das Wort 'Löser' ist in diesem Zusammenhang etwas übertrieben, denn das Eulerverfahren liefert endlich viele Näherungswerte $\{\tilde{z}(r\cdot\Delta t)|r = 1, \ldots, n\}$ statt der Parameterdarstellung einer glatten Bahnkurve, welche die Einsetzprobe in der Differentialgleichung besteht. Unser Ziel ist, ein schultaugliches und einsichtiges numerisches Verfahren zu entwickeln, um präzise numerische Approximationen $\tilde{z}(t) \approx \exp(i \cdot t)$ mit einem Computer zuverlässig und und mit geringem Aufwand, gemessen an Speicherbedarf und Rechenzeit, zu erzeugen.

Euler hat sein Verfahren *universell* konzipiert, das heisst, dass es unter den vielfältigsten Bedingungen nutzbar sein sollte. Unser Beispiel hat besondere Eigenschaften: Alle Bahnen sind Kreise um den Nullpunkt, die in 2π Zeiteinheiten einmal bei konstanter Winkelgeschwindigkeit abgefahren werden. Daraus ergeben sich Vorteile gegenüber dem allgemeinen Problem, für welches Euler's Verfahren konzipiert wurde. Wir werden sie suchen und nutzen.

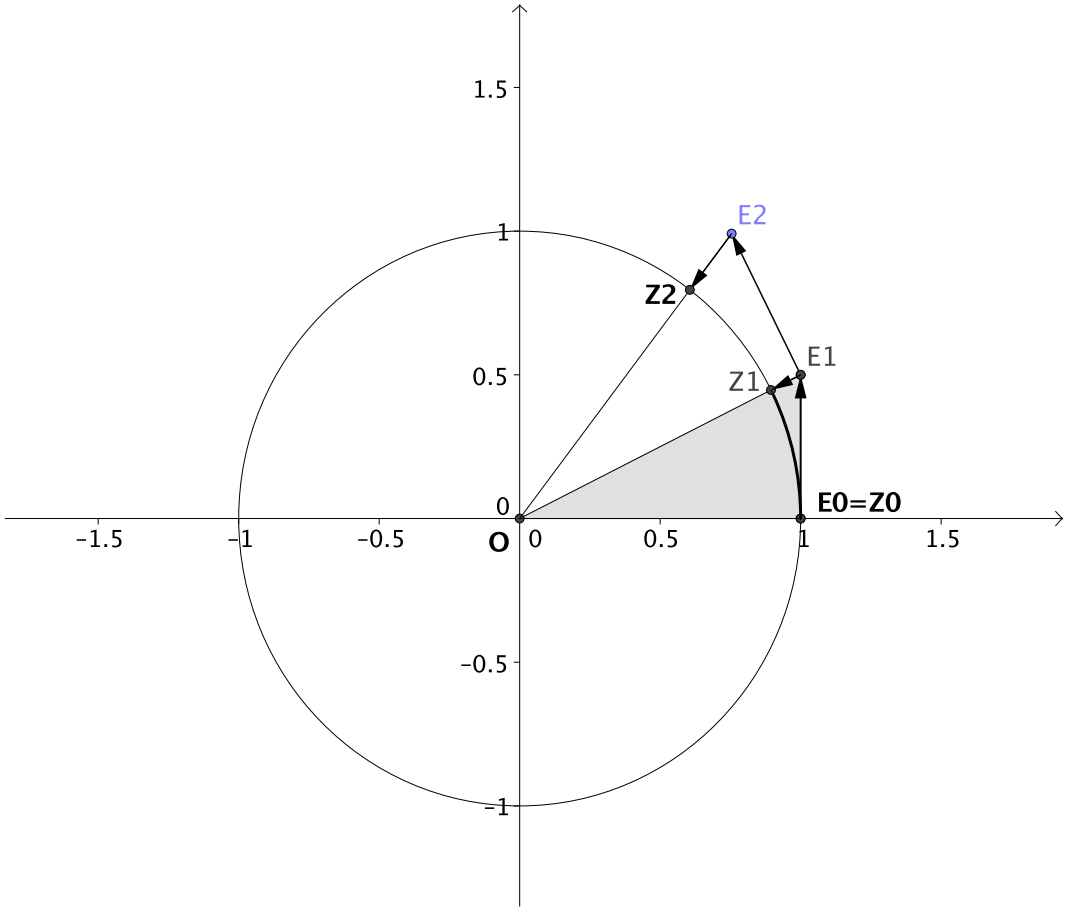

Abbildung 2: Prinzipskizze zu den Korrekturen an die Eulernäherung

Die Lösungskurve von $z'(t) = i\cdot z(t)$ mit $z(0) := 1$ ist der Einheitskreis. Die Differentialgleichung schreibt aber auch die Geschwindigkeit $z'(t)$ in jedem Zeitpunkt und den Startpunkt 1 für $t = 0$ vor. Damit ist eine besondere *Parameterdarstellung* der Kreisbahn festgelegt. Die

Lösung des Problems ist – im Gegensatz zur Gleichung $x^2 + y^2 = 1$ – nicht bloss eine Punktmenge, sondern eine mathematisch beschriebene gleichförmige Kreisbewegung $z : t \mapsto z(t)$

Verbesserungen Exakte Lösungen zum Anfangswertproblem $z'(t) = i \cdot z(t)$ und $z(0) = 1$ müssen zwei Kriterien erfüllen:

1. *Ortsbedingung:* Alle Lösungen liegen auf dem Einheitskreis.
2. *Zeitbedingung:* Für den Parameterwert t zur Lösung $z(t)$ gilt $t = \arg(z(t))$ [vgl Abb. 2]

Am richtigen Ort: Die Normierung $\check{z}(\Delta t) := \tilde{z}(\Delta t)/|\tilde{z}(\Delta t)|$ zwingt die Eulernäherung nach dem ersten Zeitschritt $\tilde{z}(\Delta t)$ auf den Einheitskreis. In Abb. 2 entspricht E_0 dem Startwert, E_1 der ersten Eulernäherung nach einem Schritt und Z_1 der Korrektur durch Normierung.

Leider zur falschen Zeit. Da die Startgeschwindigkeit $z'(0) = i \cdot z(0) = i$ den Betrag 1 hat, der im Intervall Δt konstant bleibt, gilt $\overline{E_0\,E_1} = |\tilde{z}(\Delta t) - z(0)| = |i \cdot \Delta t| = |\Delta t|$. Die Grösse Δt spielt eine Doppelrolle, einmal als Zeitdauer, einmal als Produkt von Zeitdauer und Einheitsgeschwindigkeit $\Delta t \cdot 1$, also als Streckenlänge. Der Winkel $\tau = \arg(Z_1)$ entspricht in der exakten Lösung der Bogenlänge auf dem Einheitskreis zwischen $E_0 = Z_0$ und Z_1 [vgl. Abb. 2]. Der Punkt Z_1 mit Polarwinkel τ [rad] wird in dieser Konstruktion zur Zeit Δt erreicht. Nun ist aber $\tau = \arctan \Delta t < \Delta t$. Für die mittlere Winkelgeschwindigkeit gilt $|\bar{\omega}| = \tau/\Delta t < 1$. Die Vorgabe ist jedoch eine konstante Winkelgeschwindigkeit $\omega = 1$. Die Bahn ist richtig, aber die Parametrisierung stimmt noch nicht.

Zeitkorrektur: Die Gleichheit $\tau = \Delta t$ wird erreicht mit einem verlängerten Eulerschritt von $\tan(\Delta t) > \Delta t > 0$. Dabei streckt sich die Bogenlänge τ auf Δt, wie das Dreieck $O\,E_0E_1$ mit dem Winkel τ [rad] bei O in Abb. 2 zeigt. Wir definieren

$$\hat{z}(\Delta t) := \check{z}(\tan(\Delta t)) \quad \text{im Bereich } |\Delta t| \le \pi/4$$

Dann bewegt sich $\hat{z}$ auf dem Einheitskreisbogen mit der Winkelgeschwindigkeit $\omega = 1$ in diskreten Zeitschritten der Grösse Δt und erreicht seine Positionen zur rechten Zeit:

$$\arg(\hat{z}(\Delta t)) = \arg(\tilde{z}(\tan(\Delta t)) = \Delta t$$

Formal lässt sich nun die Definition auf die ganzzahligen Vielfachen von Δt fortsetzen:

$$\hat{z}(n \cdot \Delta t) := \hat{z}(\Delta t)^n \quad \text{und} \quad \hat{z}(\Delta t)^n = \frac{\tilde{z}(\tan(\Delta t))^n}{|\tilde{z}(\tan(\Delta t))^n|} = \left(\frac{\tilde{z}(\tan(\Delta t))}{|\tilde{z}(\tan(\Delta t)|}\right)^n$$

Bemerkenswert ist, dass *ein einziger modifizierter Eulerschritt genügt.* Alle übrigen Bahnpunkte werden als *geometrische Folge auf dem Einheitskreis rein algebraisch definiert.* Die numerische Approximation von $\tan(\Delta t)$ im ersten Schritt bedeutet einen Mehraufwand, aber die Hauptlast auf 'sehr viele ganz kleine' Eulerschritte entfällt.

Der modifizierte Eulerschritt entspricht einer Zentralprojektion zum Zentrum O auf den Einheitskreis von dessen Tangente in (0|1) aus.

Kosmetik Die Normierung $\hat{z}(\Delta t) := \tilde{z}(\tan(\Delta t))/|\tilde{z}(\tan(\Delta t))|$ beinhaltet die Berechnung einer Quadratwurzel im Ausdruck $|\tilde{z}(\Delta t)| = \sqrt{1+\tan(\Delta t)^2}$.

Diese lässt sich vermeiden durch eine neue Parametrisierung mit dem Parameter $\frac{1}{2}\Delta t$:

$$\hat{z}(\Delta t) = \left(\hat{z}(\frac{1}{2}\Delta t)\right)^2 = \frac{(1+i\cdot\tan(\frac{1}{2}\Delta t))^2}{1+\tan(\frac{1}{2}\Delta t)^2} = \frac{1-\tan(\frac{1}{2}\Delta t)^2+2\cdot i\cdot\tan(\frac{1}{2}\Delta t)}{1+\tan(\frac{1}{2}\Delta t)^2}$$

Randbemerkung Die Abbildung

$$\mu : \mathbb{Q} \to \mathbb{C} \quad \text{mit} \quad \mu : s \mapsto \frac{1-s^2}{1+s^2} + i\cdot\frac{2s}{1+s^2}$$

erzeugt nur Punkte auf dem Einheitskreis mit lauter rationalen kartesischen Koordinaten. Weil μ stetig ist, lässt es sich von $\mathbb{Q}$ auf ganz $\mathbb{R}$ stetig erweitern. Damit liegt das Bild $\mu(\mathbb{Q})$ dicht im Bild $\mu\mathbb{R}$, das den Einheitskreis mit Ausnahme von $(-1|0)$ bedeckt. Das heisst, dass sich alle Punkte auf dem Einheitskreis beliebig gut annähern lassen mit Punkten aus dem Einheitskreis deren Koordinaten rational sind.

Das gleiche Argument gilt in unserem Zusammenhang für die Bilder jener $\Delta t/2$, für welche $s := \tan(\Delta t/2)$ rational ausfällt. [Der Wertebereich von $\tan(\Delta t/2)$ ist ganz $\mathbb{R}$, wenn alle Δt mit $|\Delta t/2| < \pi/2$ zugelassen sind.]

5.2 Der modifizierte Eulerschritt

Wir wissen schon, dass $\hat{z}(\tan(\frac{1}{2}\Delta t))^2 = \exp(i\Delta t)$ gilt. Aber diese Einsicht tauscht bloss das Problem der Berechnung von $\cos(\alpha)+i\sin(\alpha)$ gegen die Berechnung von $\tan(\frac{1}{2}\alpha)$ aus. Nicht ganz, denn wir benötigen bloss Tangenswerte zu kleinen Winkeln oder wir erzeugen eine Hilfstabelle, die nur einmal berechnet und gespeichert wird.

Im Bereich kleiner Winkel $h \approx 0$ bieten Taylorpolynome zur Tangensfunktion eine Option, um rasch gute Näherungen zu finden. Wir verwenden

$$T_7(h) := h + \frac{1}{3}h^3 + \frac{2}{15}h^5 + \frac{17}{315}h^7 \approx \tan(h)$$

Für $|h| < 10^{-2}$ und in exakter Arithmetik liegen die Beträge der Abweichung der Taylornäherung von den exakten Tangenswerten unter 10^{-17}. Das genügt für die beabsichtigte Anwendung.

5.3 Viele Schritte vom Kleinen zum Grossen mit Algebra

Alles ist jetzt so eingerichtet, dass wir im Kleinen (d.h. lokal) eine sehr gute Näherung $\hat{z}(\frac{1}{2}\Delta t)$ kennen. Die Übersetzung 'vom Kleinen ins Grosse' gelingt im Prinzip mit Algebra dank einer besonderen algebraischen Eigenschaft der Exponentialfunktion: $\hat{z}(t) = \hat{z}(n\cdot\Delta t) = (\hat{z}(\Delta t))^n$. Allerdings ist jeder Rechenschritt mit Fliesskommazahlen eine Quelle möglicher Rundungsfehler. So kommt auch noch Numerik ins Spiel.

Schnell potenzieren: Angenommen, es sei c^N für eine natürliche Zahl N und eine beliebige Zahl $c \in \mathbb{C}$ zu berechnen. Die Berechnung von c^2 benötigt eine einzige Multiplikation, die Berechnung von c^3 und jene von c^4 ist mit 2 Produkten zu schaffen, wenn die Zwischenergebnisse verwendet werden. Wählt man $N := 2^r$ und $r \in \mathbb{N}$, so schafft man die Berechnung von c^N mit genau r Multiplikationen.

Aufgaben

13. Wie lässt sich $T_7(h)$ mit möglichst wenigen Multiplikationen berechnen?

14. Begründen oder widerlegen Sie folgende Aussagen: Für alle $|\Delta t| < \frac{\pi}{2}$ gilt:

    (a) $\exp(i\Delta t) = \hat{z}(\tan(\Delta t))$
    (b) $\exp(i\Delta t) = (\hat{z}(\frac{1}{2}\tan(\Delta t)))^2$
    (c) $\exp(i\Delta t) = (\hat{z}(\tan(\frac{1}{2}\Delta t)))^2$
    (d) $\exp(i\arctan(\Delta t)) = \hat{z}(\Delta t)$
    (e) $\exp(i\Delta t) = (\hat{z}(\frac{1}{2}\Delta t))^2$
    (f) $\exp(i \cdot 10^{-6}) = \hat{z}(10^{-6})$

15. Begründen oder widerlegen Sie: Die folgenden Identitäten gelten für alle reellen τ.

$$\text{(a)} \quad \cos(\tau) = \frac{1-\tan(\frac{1}{2}\tau)^2}{1+\tan(\frac{1}{2}\tau)^2} \qquad\qquad \text{(b)} \quad \sin(\tau) = \frac{2\cdot\tan(\frac{1}{2}\tau)}{1+\tan(\frac{1}{2}\tau)^2}$$

16. Beweisen Sie die Behauptung: Angenommen, $u > 0$ und $v > 0$ seien rationale Zahlen mit $u^2 + v^2 = 1$. Dann gibt es pythagoreische Trippel p, q, r mit $p^2 + q^2 = r^2$ und einen Winkel α mit $\tan(\alpha) = q/p$, $u = \cos(\alpha) = p/r$, $v = \sin(\alpha) = q/r$.

6 Ein Programm zum Experimentieren

Das folgende Programm in Pseudocode berechnet eine Näherung für $\exp(i \cdot t)$. Neben t wird ein weiterer Parameter $r \in \mathbb{N}$ verwendet. Die Zahl r steuert die Feinheit der Diskretisierung, indem $h := t \cdot 2^{-r}$ berechnet wird. Sie beeinflusst die Genauigkeit des Ergebnisses über die Diskretisierungsfehler und über die Rundungsfehler. Ferner hängt der Rechenaufwand etwas von der Wahl von r ab.

Im Listing ist die Berechnung des Taylorpolynoms $T_7(h)$ nach dem Hornerschema erkennbar. Vor der Ausgabe wird das Ergebnis nochmals auf Länge 1 normiert. Mit dem Näherungswert y wird auch die Differenz zum Ergebnis der Berechnung mit der Standardfunktion des Rechners ausgegeben. Mit i wird die imaginäre Einheit aufgerufen, j ist ein Schleifenzähler. Die Programmiersprache unterstützt arithmetische Operationen und numerische Approximation analytischer Standardfunktionen in $\mathbb{C}$.

```
expi(r,t)
Func
Local h, hh, j, tah, ttah, y
t/2^r → h
h*h → hh
h*(1+hh/3*(1+hh/5*(2+hh/21*17))) → tah
tah*tah → ttah
(1-ttah+2*tah*i)/(1+ttah) → y
For j, 1, r-1
    y*y → y
EndFor
y/abs(y) → y
Return {y, abs(y), y-exp(i*t)}
```

Dieses Programm ist zum Experimentieren gedacht. Zum Beispiel lässt sich der Einfluss der Diskretisierung durch die Wahl von r studieren. Die für die Taylorformel relevante Grösse ist h, welches die Rolle von $\frac{1}{2}\Delta t$ spielt. Wie beeinflusst die Wahl von r die Genauigkeit der Approximation? In diese Fragestellung geht die Tatsache ein, dass der Rechner mit jeder arithmetischen Operation potentiell rundet und Informationen verwischt. Der Qualitätsvergleich mit $y - \exp(i * t)$ ist etwas naiv, weil als Standard die Rechnerfunktion exp anstelle der mathematisch exakten Exponentialfunktion benutzt wird. Für die Beurteilung von Feinheiten reicht dieses Kriterium nicht.

Entscheidend ist, wie gross h bei der Berechnung der Taylornäherung für $\tan(h)$ ist. Nach einer Faustregel lässt sich abschätzen, dass $|h| < 10^{-2}$ sein soll, damit die Taylornäherung gut genug wird. Für $|t| < 1$ folgt bei exakter Arithmetik $r \leq 8$ [d.h. höchstens 7 iterierte Winkelverdoppelungen der Art `y*y→y` in der For-Schleife]. Gelegentlich zeigt sich noch bessere Übereinstimmung mit den Standardfunktionen der SW für $r = 9$. Im Idealfall wird r abhängig von t bestimmt.

Es ist sinnvoll, die Eingaben für t zu beschränken zum Beispiel auf $|x| \leq \pi/4$. Das ist keine wesentliche Einschränkung, wenn man die Identitäten $\sin(\pi/4 + u) = \cos(\pi/4 - u)$ und die Periodizität der Kreisbewegung ausnutzt.

Das Programm lässt sich sogar noch etwas vereinfachen. Wenn man es im kleinen Definitionsbereich $0 \leq t \leq \pi/4$ anwendet, kann in der Zeile

```
(1-hh+2*h*i)/(1+hh) → y
```

die Normierung weggelassen werden. Der einfachere Befehl

```
(1-hh+2*h*i) → y
```

könnte genügen, weil noch eine Normierung nach der Vorschleife vorhanden ist. Sie ist nicht überflüssig, auch wenn y normiert wird.

Weitere Überlegungen sind nötig, um den Definitionsbereich von Cosinus und Sinus auf eine ganze Periode der Länge 2π auszudehnen. Von da erscheint eine Erweiterung auf 'alle' Rechnerzahlen in unmittelbarer Reichweite. Die Periodizitätsbeziehung $f(x) = f(x + 2n\pi)$ gilt jedoch nur bei exakter Arithmetik für alle $n \in \mathbb{Z}$. Beispielsweise sind mit Fliesskommazahlen und 16-stelliger Mantisse die Zahlen 10^{20} und $10^{20} + 2\pi$ in der Rechnernäherung nicht mehr unterscheidbar.

Aufgaben

17. Gilt die Beziehung $\sin(\pi/4 + u) = \cos(\pi/4 - u)$ für alle reellen u? Begründen Sie die Antwort.

18. Angenommen, es soll $\cos(1)$ angenähert werden mit einem Verfahren, das nur Eingaben $|x| < \pi/4$ akzeptiert. Wie ist das möglich, obwohl $1 > \pi/4$ gilt?

19. Begründen oder widerlegen Sie:
    Das Programm `expi(r,x)` liefert für die Eingaben $r = 6$ und $x = 1.0$ ein Ergebnis $a + ib$, so, dass $|a - \cos(1.0)| + |b - \sin(1.0)| < 10^{-13}$ gilt, wobei hier cos() und sin() die auf Ihrem Rechner implementierten Standardfunktionen bezeichnen.

20. Welche Vorteile ergeben sich, wenn Cosinus und Sinus gleichzeitig berechnet werden? Welches sind die Nachteile?

21. Es sei r eine natürliche Zahl. Wie erfolgt die Berechnung von $x/2^r$ im Binärsystem?

7 Anhang

7.1 Bemerkungen und Literaturhinweise

- Wenn Speicherplatz billig ist und kurze Rechenzeiten wichtig sind, so werden Funktionen auf einem Intervall als Listen von Abtastwerten einmal aufbereitet und abgespeichert. Die Abtastung ist so fein gestaltet, dass lineare Interpolation ausreicht, um die Lücken bei Bedarf ohne Genauigkeitseinbusse zu füllen. Diese Option ist bei den trigonometrischen Funktionen und ihren Umkehrungen gegeben. Das alte Konzept der Funktionstafeln wird dank der Speicherkapazität und der Verwendung von Caches angepasst und potenziert.

- Ein interessantes Verfahren ist von Jack E. Volder im September 1959 veröffentlicht worden:

  Volder, J.E., The CORDIC trigonometric computing technique. IRE Trans. Electronic Computers, 8, September, 330-334 (1959).

  Es ist für eine rudimentäre Rechnertechnologie so optimiert, dass Multiplikationen vermieden werden und im wesentlichen binäre shift und add Operationen benutzt werden. Die Methode wurde auf das Dezimalsystem angepasst und in manchen Taschenrechnern realisiert, weil sie sich universell einsetzen lässt zur Approximation von exp, log, der trigonometrischen oder hyperbolischen Funktionen und deren Umkehrungen.

- Volder schildert die Umstände der Erfindung von CORDIC in der Zeit ab 1956 in einem untechnischen und historischen Bericht:

  Jack E. Volder, The Birth of CORDIC, Journal of VLSI Signal Processing 25, 101-105, 2000, Kluwer Academic Publishers.

- Ein informativer Artikel über CORDIC ist: `http://www.wiete.com.au/journals/GJEE/Publish/vol8no3/Risse.pdf` [besucht 3.12.2017]

7.2 Bemerkungen zu den Aufgaben, Lösungsskizzen

1. Eine rationale Näherung für sin(1):

$$\sin(1) \approx \sum_{n=0}^{9} \frac{(-1)^n}{(2n+1)!} = \frac{102360822438075317}{121645100408832000} \approx 0.8414709848079$$

   Der Fehler in der Bruchdarstellung ist geringer als der nächste Term in der Taylorreihe $1/21! \approx 1.96 \cdot 10^{-20}$. Begründung: die Summanden haben wechselnde Vorzeichen und die Beträge der Summanden streben monoton gegen 0. Folglich ist der Approximationsfehler im Polynom vom Grad 19 höchstens so gross der Betrag des Summanden im Taylorpolynom der Ordnung 21, denn die Summanden gerader Ordnung treten bei der Sinusreihe nicht auf. Durch die Umrechnung der berechneten rationalen Näherung für sin(1) in die Dezimaldarstellung des Rechners mit 14 Ziffern wurden also mindestens 6 weitere korrekte Dezimalstellen unterdrückt.

2. (a) $\sin(t) = \sqrt{1 - (\cos(t))^2}$

   (b) Reduktion auf ein Grundintervall: Für jeden Winkel T gibt es einen Winkel s im Grundintervall $|s| \leq \pi$ mit $(\cos(T)|\sin(T)) = (\cos(s)|\sin(s))$ [Periodizität].

Durch wiederholte Spiegelungen, die den Einheitskreis auf sich abbilden, lässt sich der Achtelskreis $0 \leq t \leq \pi/4($ auf den Vollkreis ausdehnen:

Spiegelt man den Einheitskreis an der Winkelhalbierenden des ersten Quadranten, so bleiben $\pm(\cos(\pi/4)|\sin(\pi/4))$ fest und für alle Winkel s gilt $\sin(\pi/4+s) = \cos(\pi/4-s)$. Damit wird der Achtelskreisbogen mit Polarwinkel t im Bereich $0 \leq t \leq \pi/4$ auf den Viertelskreis im ersten Quadranten erweitert. Beim Spiegeln an der x-Achse bleibt der Cosinuswert erhalten und der Sinuswert wechselt das Vorzeichen, also $\sin(-s) = -\sin(s)$ und $\cos(-s) = \cos(s)$. Beim Spiegeln an der y-Achse bleibt der Sinuswert erhalten und der Cosinuswert wechselt das Vorzeichen, also $\sin(\pi/2-s) = \sin(\pi/2+s)$ und $\cos(\pi/2-s) = -\cos(\pi/2+s)$

3. Mit Induktion folgt, dass alle Ableitungen von f mit f identisch sind $\frac{\mathrm{d}^n}{\mathrm{d}z^n} f = f$. Da für alle Ableitungen gilt: $\frac{\mathrm{d}^n}{\mathrm{d}z^n} f(0) = 1$, ergibt sich die Darstellung als Potenzreihe nach Taylor
$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$
Diese Potenzreihe konvergiert für alle komplexen Zahlen und kann damit die Exponentialfunktion definieren. Für praktische Berechnungen ist sie i.a. ungeeignet.

4. Die Einsetzprobe ergibt $f_c'(z) = c \cdot \exp(c \cdot z) = c \cdot f_c(z)$ [Kettenregel] und $f_c(0) = 1$.

5. Die Behauptung ist richtig [Annahme: o.n. Koordinaten]. Zwei verschiedene Schreibweisen beschreiben denselben Sachverhalt, denn die Zuordnung $\varphi : \begin{bmatrix} a \\ b \end{bmatrix} \mapsto a + i \cdot b$ definiert eine bijektive Abbildung von $\mathbb{R}^2$ auf $\mathbb{C}$. Dem Produkt mit der Drehmatrix zum Drehwinkel t in $\mathbb{R}^2$ entspricht genau die Multiplikation mit $\exp(i \cdot t)$ im Komplexen.
Formal:
$$\begin{aligned}
\begin{bmatrix} a \\ b \end{bmatrix} \overset{\varphi}{\longmapsto} a + i \cdot b \overset{r}{\longmapsto} \exp(i \cdot t) \cdot (a + i \cdot b) &= r(t) \\
&= (\cos(t) + i \cdot \sin(t) \cdot (a + i \cdot b) \\
&= a \cdot \cos(t) - b \cdot \sin(t) + i \cdot (a \cdot \sin(t) + b \cdot \cos(t)) \\
\overset{\varphi^{-1}}{\longmapsto} \begin{bmatrix} a \cdot \cos(t) - b \cdot \sin(t) \\ a \cdot \sin(t) + b \cdot \cos(t) \end{bmatrix} &= \begin{bmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = R(t)
\end{aligned}$$

6. $|\exp(i \cdot t)| = |\exp(i \cdot t/n)| = |\exp(i \cdot t/n)^n|$ und $t = \arg(\exp(i \cdot t)$, $\arg(\exp(i \cdot t/n)^n = n \cdot t/n = t$

7. Ist die Zahl k aus der Einheitsstrecke mit Zirkel und Lineal konstruierbar, so ist auch $1 + i \cdot k \equiv (1|k)$ in $\mathbb{R}^2$ mit Zirkel und Lineal konstruierbar. Sind $z, w \in \mathbb{C}$ gegeben, so ist $z \cdot w \in \mathbb{C}$ konstruierbar, denn $|z \cdot w|$ ist als Produkt gegebener Längen mit dem Strahlensatz konstruierbar und $\arg(z \cdot w) = \arg(z) + \arg(w)$ ist als Addition von gegebenen Winkeln mit Zirkel und Lineal konstruierbar.

8. Beweis mit Induktion, bei beliebiger Konstante Δt:
Verankerung: $\tilde{z}(0 \cdot \Delta t) = z_0 \cdot 1$ gilt exakt.
Induktionsschritt:

Annahme $\tilde{z}(r \cdot \Delta t) = z_0 \cdot (1 + i \cdot \Delta t)^r$ [Ergebnis des Eulerverfahrens nach r Schritten der Grösse Δt.]

Dann liefert ein weiterer Eulerschritt der Schrittweite Δt ausgehend von $z_0 \cdot \tilde{z}(r \cdot \Delta t)$ die Näherung

$$\begin{aligned}\tilde{z}((r+1) \cdot \Delta t) &:= z_0 \cdot (\tilde{z}(r \cdot \Delta t) + i \cdot z_0 \cdot \tilde{z}(r \cdot \Delta t) \cdot \Delta t) \\ &= z_0 \cdot (1 + i \cdot \Delta t)^r + i \cdot \Delta t \cdot z_0 \cdot (1 + i \cdot \Delta t)^r \\ &= z_0 \cdot (1 + i \cdot \Delta t)^{r+1}\end{aligned}$$

9. numerische Näherungen für $\exp(i) \approx 0.5403023 + 0.841471i$

Tabelle 1: Näherungen für $(1 + i/n)^n$ und $|(1 + i/n)^n|$

n	$(1 + i/n)^n$	$\lvert(1 + i/n)^n\rvert$
1	$1+i$	1.41421
10	$0.570790+0.882508i$	1.05101
10^2	$0.543039+0.845671i$	1.00501
10^3	$0.540573+0.841892i$	1.00050
10^4	$0.540329+0.841513i$	1.00005
10^5	$0.540305+0.841475i$	1.000005

10. allgemeine binomische Formel:
$(1 + i \cdot t/n)^n = \sum_{r=0}^{n} \frac{n!}{r!(n-r)!} \cdot i^r \cdot \frac{t^r}{n^r} = \sum_{r=0}^{n} \frac{n!}{n^r(n-r)!} \cdot \frac{(i \cdot t)^r}{r!}$
Taylorpolynom, Grad n: $\sum_{r=0}^{n} \frac{(i \cdot t)^r}{r!}$
Da $n!/(n-r)! = n \cdot (n-1) \cdot \cdots \cdot (n-r+1)$, folgt für grosse n und festes r, dass für den Vorfaktor von t^r in der binomischen Formel gilt

$$\frac{n!}{n^r \cdot (n-r)!} = \frac{(n-1)}{n} \cdots \frac{(n-r+1)}{n} \approx 1$$

gilt und zwar innerhalb einer beliebigen positiven Toleranz, wenn nur n genügend gross gewählt wird. Das heisst, dass mit wachsendem n jedes beliebige Anfangsstück endlicher Länge ℓ der binomischen Entwicklung von $(1+i \cdot t/n)^n$ summandenweise gegen das entsprechnede Anfangsstück der Taylorreihe von $\exp(i \cdot t)$ strebt. Wegen der Konvergenz der beiden durch Grenzwerte bestimmten Ausdrücke, lässt sich das Anfangsstück so bestimmen, dass auch das bisher nicht näher betrachtete Reststück oberhalb von ℓ kleiner wird als jede vorgegebene positive Toleranz. Damit stimmen die beiden Grenzwerte überein.

11. Der Term $c^{(2^n)}$ lässt sich iterativ mit n Multiplikationen auswerten. Beweis mit Induktion, siehe Abschnitt 5.3

12. Notation: $c := 1 + i \cdot t/n$, dann ergibt Konjugation $\bar{c} = 1 - i \cdot t/n$

(a) $|c|^2 = c \cdot \bar{c} = 1 + t^2/n^2$

(b) $|c| = |\bar{c}|$, Spiegelung lässt die Länge invariant.

(c) $|c^n|^2 = (c^n \cdot \bar{c}^{\,n}) = (c \cdot \bar{c})^n$

(d) Für alle reellen t gilt: $\lim_{n\to\infty}(1+t/n)^n = \exp(t), \quad \lim_{n\to\infty}(1-t/n)^n = \exp(-t)$. Da das Produkt zweier konvergenter Folgen gegen das Produkt der einzelnen Grenzwerte konvergiert, folgt

$$\exp(t) \cdot \exp(-t) = \exp(0) = 1 = \lim_{n\to\infty}(1 - t^2/n^2)^n$$

(e) Für alle reellen t gilt:
$\lim_{n\to\infty} c^n = \exp(i \cdot t)$ und $\lim_{n\to\infty} \bar{c}^{\,n} = \overline{\exp(i \cdot t)} = \exp(-i \cdot t)$.
Nun ist $\lim_{n\to\infty}(|c|^2)^n = \lim_{n\to\infty}(c \cdot \bar{c})^n = \exp(i \cdot t) \cdot \exp(-i \cdot t) = \exp(0) = 1$
Folgerung: $\lim_{n\to\infty}(|1 \pm i \cdot t/n|^2)^n = \lim_{n\to\infty}(1 + t^2/n^2)^n = 1$

Die Eulernäherungen $(1 + i \cdot t/n)^n$ streben für $n \to \infty$ und alle reellen t gegen Punkte auf dem Einheitskreis.

Bemerkung: [Ein anderer Zugang] Durch den Vergleich der Taylorreihen ergibt sich unmittelbar $\exp it = \cos(t) + i \cdot \sin(t)$. Damit ist gezeigt: $\exp(i \cdot t)$ liegt auf dem Einheitskreis, der Polarwinkel ist gleich t. Dann folgt $\exp(i \cdot (s + t) = \exp(i \cdot s) \cdot \exp(i \cdot t)$ ohne das Nebenergebnis $\lim_{n\to\infty}(1 + t^2/n^2)^n = 1$.

13. $T_7(h) := h + \frac{1}{3}\,h^3 + \frac{2}{15}\,h^5 + \frac{17}{315}\,h^7 = h \cdot (1 + \frac{q}{3} \cdot (1 + \frac{q}{5} \cdot (2 + q \cdot \frac{17}{21})))$ mit $q := h \cdot h$ [vgl. Hornerschema]

14. Notation: vgl Abb. 2.: $\Delta t := \angle(E_0\,O\,E_1) = \arctan(\overline{E_0\,E_1}) \Leftrightarrow \overline{E_0\,E_1} = \tan(\Delta t)$

(a) richtig, da $\tan(\Delta t) = \overline{E_0\,E_1}$

(b) falsch, da Seitenhalbierende auf einer Kathete nie gleich der Winkelhalbierenden ist.

(c) richtig, Regel von De Moivre für komplexe Zahl der Länge 1 anwenden.

(d) richtig, äquivalent zu (a)

(e) falsch, Widerspruch zu (b)

(f) falsch, Widerspruch zu (d), numerische Näherung mit Fehler der Grössenordnung $3 \cdot 10^{-19}$

15. Beide Aussagen sind in dieser Allgemeinheit falsch wegen der Polstellen von Tangens.

16. (a) Ist p, q, r pythagoreisches Trippel mit $p^2 + q^2 = r^2$, so sind
$\cos(\alpha) := p/r$, $\sin(\alpha) := q/r$ rationale Koordinaten eines Punktes auf dem Einheitskreis.

(b) Es sei $u^2 + v^2 = 1$, $u, v \in \mathbb{Q}$ und $u = u_1/u_2$, $v = v_1/v_2$ Quotienten ganzer Zahlen, $u_2 \neq 0$, $v_2 \neq 0$. Dann ist $u^2 + v^2 = \dfrac{u_1^2 \cdot v_2^2 + u_2^2 \cdot v_1^2}{u_2^2 \cdot v_2^2} = 1$ und $u_1 \cdot v_2$, $u_2 \cdot v_1$, $u_2 \cdot v_2$ bilden ein pythagoreisches Trippel.

17. Vgl. Aufgabe 2. Die Beziehung beschreibt eine Spiegelung der Punkte auf dem Einheitskreis an der Winkelhalbierenden des 1. und 3. Quadranten. Dabei werden die Vorzeichen der Winkel zur Richtung $\pi/4$ und die Koordinaten $x \leftrightarrow y$ vertauscht. Genau das wird in Aufgabe 14 behauptet.

18. $\pi/4 - u = 1$ definiert $u := \pi/4 - 1 < 0$, mit Nr 14 folgt $\cos(1) = \sin(\pi/2 - 1)$ mit $0 < \pi/2 - 1 < \pi/4$.

19. Mit einem Rechner und einer Programmiersprache, die IEEE doppelt genaue Arithmetik benutzt [zB MATLAB, Octave, CAS-Rechner], wurden die Aussagen verifiziert.

20. Das Grundintervall für die Berechnungen kann auf $[0, \pi/4]$ beschränkt bleiben und derselbe Prozess berechnet dort beide Funktionen. Dabei wird eine Arithmetik in $\mathbb{C}$ [bzw. Matrixoperationen] verwendet, was aufwendiger ist als Operationen mit reellen Zahlen. Ferner sind Zusatzmassnahmen nötig, um den Definitionsbereich mindestens auf den Vollkreis auszudehnen.

    Die Berücksichtigung der Periodizität ist auf grossen Definitionsbereichen nur mit formal exakter Arithmetik einfach. Das theoretische Verfahren versagt, wenn die Periodenlänge in der verwendeten Arithmetik nicht exakt berechenbar oder darstellbar ist.

    Beispiel π als Fliesskommazahl:
    Angenommen, für $X = 10^{40}$ sei $\sin(X)$ zu berechnen und wir verwenden $\tilde{\pi}$ als Näherung für π mit 20 gültigen Dezimalziffern $3.14\ldots$. Dann ist die ganze Zahl N, die am nächsten bei $X/(2\tilde{\pi})$ liegt, fast immer eindeutig und $|X - 2N\tilde{\pi}| \leq \tilde{\pi}$. Um wieviel unterscheiden sich $X/(2\pi)$ und $X/(2\tilde{\pi})$?
    Antwort $|X/(2\pi) - X/(2\tilde{\pi})| > 10^{18}$. Reduktion auf eine Grundperiode kann bei beschränkter Genauigkeit der Zahldarstellung nur beschränkt und angenähert angewandt werden.

21. Die Mantisse bleibt erhalten, nur der Zweierexponent wird um r verkleinert.
    (Bzw. Ziffernfolge bleibt, 'Dualpunkt' wird um r Stellen nach links verschoben, eventuelle Lücke vor der führenden Eins wird mit Ziffern 0 aufgefüllt.)