1 Einführung und Begriffe

1.1 **Mathematische Modellbildung und numerische Simulation** am Beispiel eines Wasserkreislaufs

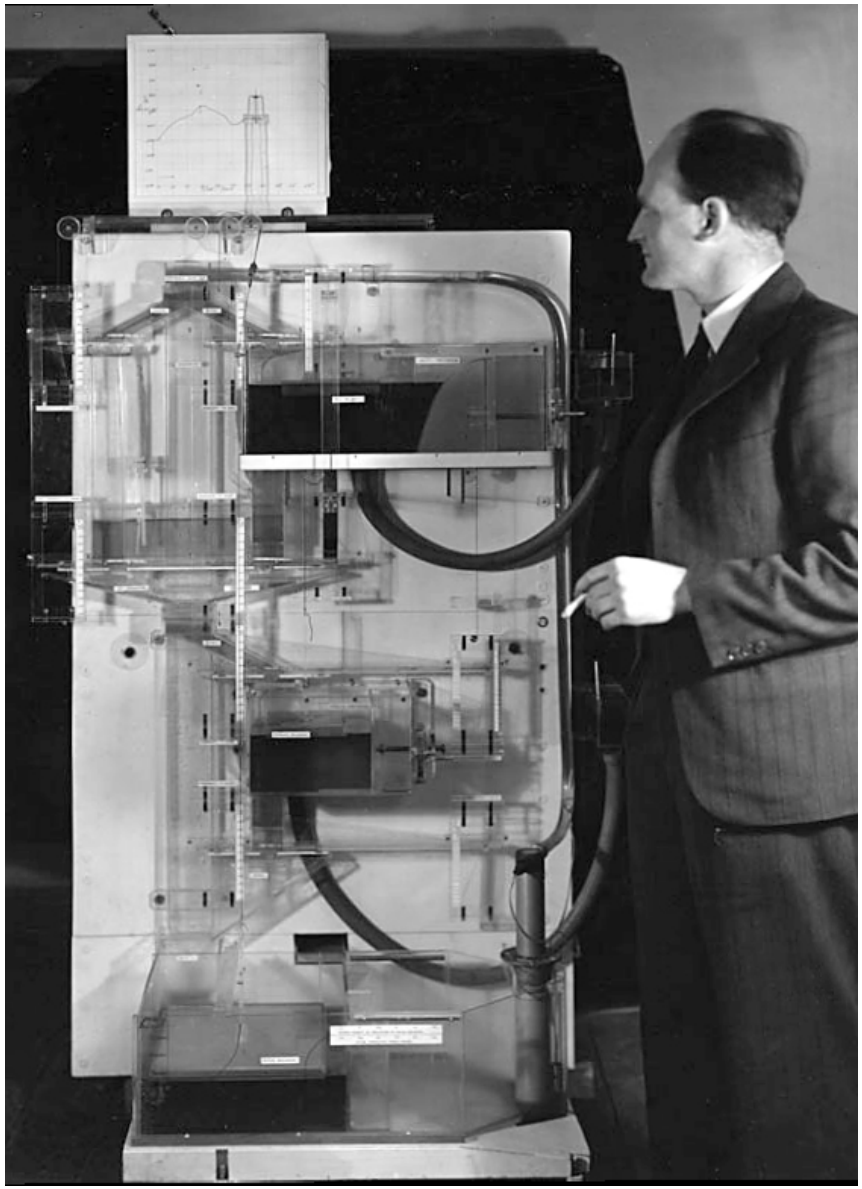
1.2 **Linearisierung und Iterationsverfahren** am Beispiel des Newton-Verfahrens

1.3 **Diskretisierung und Stabilität** am Beispiel der Wärmeleitungsgleichung

1.1 Mathematische Modellbildung und numerische Simulation am Beispiel eines Wasserkreislaufs

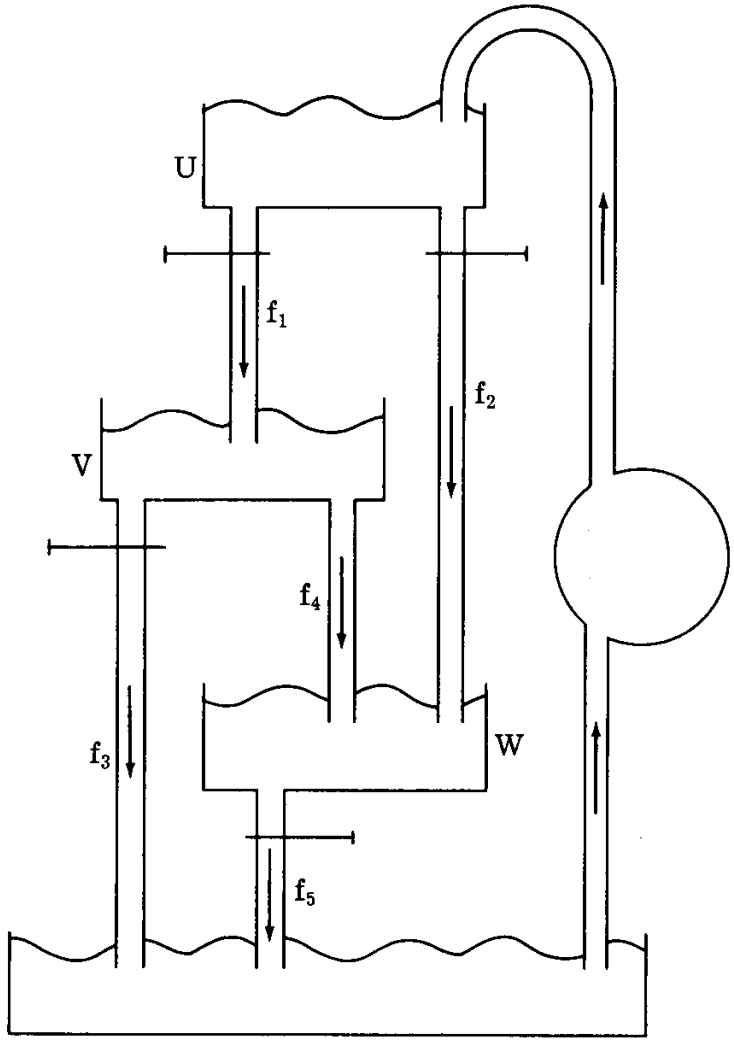
„Simulation ist die Nachbildung eines dynamischen Prozesses in einem Modell, um zu Erkenntnissen zu gelangen, die auf die Wirklichkeit übertragbar sind“ (VDI-Richtlinie 3633).

Zum einen ist die rechnerische Simulation dann unumgänglich, wenn reale Experimente mit den Untersuchungsobjekten undurchführbar sind: Denken Sie etwa an die Entstehung von Galaxien oder an Untersuchungsobjekte, die erst geplant sind, also real noch gar nicht existieren. Aber auch wenn reale Experimente möglich sind, ist es oft kostengünstiger und ressourcenschonender, stattdessen numerische Simulationen einzusetzen.



Modelle müssen nicht notwendig rechnerisch/mathematisch sein.

A. W. Philips, (London School of Economics) mit dem Analogcomputer MONIAC zur Modellierung ökonomischer Vorgänge („hydraulic macroeconomics“)



vgl. Davis & Hersch: *Descartes' Dream*. Dover, 2005

A. Physikalische Grundlagen.

[EVANGELISTA TORRICELLI (1608–1647)]:

Abflussgeschwindigkeit $v = \sqrt{2gh}$, $g = 9.81$ (Gravitationsbeschleunigung),
$h =$ Höhe des Wasserspiegels.

Abflussrate als Funktion des im Behälter befindlichen Wasservolumens V (falls es sich um einen Zylinder mit Grundfläche A handelt)

$$f = a\sqrt{2gV/A} = c\sqrt{V} \quad \text{mit} \quad c := a\sqrt{2g/A}\,.$$

Der Parameter c kann über a variiert werden, wenn der Abfluss einen Hahn besitzt. Wir sprechen von einem Steuerungsparameter.

B. Mathematisches Modell.

$U(t), V(t), W(t), R(t)$: Wassermengen zur Zeit t in den Behältern.

$f_1, \ldots, f_5$: Abflussfunktionen mit den Steuerungsparametern $c_1, \ldots, c_5$.

$p = p(t)$: „Pumpenfunktion“

Änderungsraten der Wasservolumina: Zuflüsse weniger Abflüsse, d.h.

$$\begin{aligned} U'(t) &= p(t) - f_1(U(t)) - f_2(U(t)) \\ V'(t) &= f_1(U(t)) - f_3(V(t)) - f_4(V(t)) \\ W'(t) &= f_2(U(t)) + f_4(V(t)) - f_5(W(t)) \\ R'(t) &= f_3(V(t)) + f_5(W(t)) - p(t). \end{aligned} \tag{1.1}$$

Diese Gleichungen, sog. (gewöhnliche) Differentialgleichungen, heißen die Kontinuitätsgleichungen unseres Systems.

Anfangszustand: Wassermengen in Behältern zu einem festen Zeitpunkt, etwa für $t = 0$.

Das Verhalten unseres Systems ist für alle Zeiten $t > 0$ durch die obigen Differentialgleichungen eindeutig bestimmt. Die Aufgabe, eine Lösung des Systems (1.1) zu bestimmen, welche gegebene Anfangsbedingungen erfüllt, nennt man ein Anfangswertaufgabe.

Addiert man alle Gleichungen, so ergibt sich

$$U'(t) + V'(t) + W'(t) + R'(t) = 0,$$

ein globales Erhaltungsprinzip, welches besagt, dass sich die Gesamtwassermenge in unserer Apperatur nicht verändert. (Es handelt sich hier um ein geschlossenes System.)

C. Algorithmus.

Anfangswertaufgaben lassen sich nur in Ausnahmefällen geschlossen lösen (reine Mathematik: in unserem Fall gibt es genau eine Lösung).

Aufgabe der Numerik: Bereitstellung von Näherungslösungen.

Idee: Wir betrachten die Gleichungen nicht mehr für jeden beliebigen Zeitpunkt, sondern nur noch zu bestimmten diskreten Zeitpunkten, etwa für $t_0 = 0$, $t_1 = 1, \ldots$ (Diskretisierung).

$U_n := U(t_n), \ldots, R_n := R(t_n)$ (Volumina, die sich zum Zeitpunkt t_n in den Behältern $U, \ldots, R$ befinden).

Änderungsrate $U'(t_n)$ wird durch $U(t_{n+1}) - U(t_n) = U_{n+1} - U_n$ approximiert (wir nähern hier eine Tangentensteigung durch eine Sekantensteigung an).

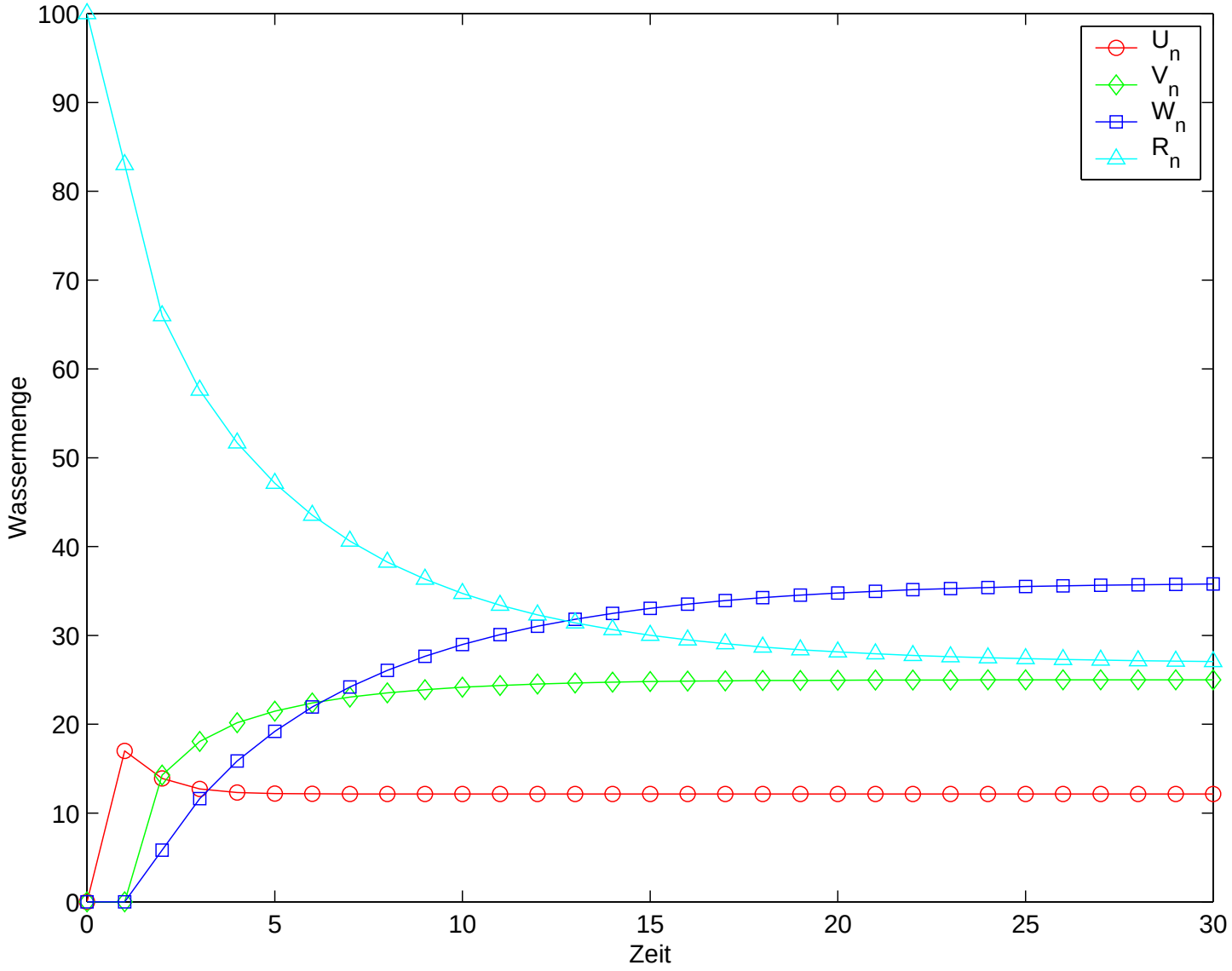
$$\begin{aligned} U_{n+1} &= U_n + p_n - f_1(U_n) - f_2(U_n) \\ V_{n+1} &= V_n + f_1(U_n) - f_3(V_n) - f_4(V_n) \\ W_{n+1} &= W_n + f_2(U_n) + f_4(V_n) - f_5(W_n) \\ R_{n+1} &= R_n + f_3(V_n) + f_5(W_n) - p_n. \end{aligned} \tag{1.2}$$

Diese vier Gleichungen heißen die diskreten Kontinuitätsgleichungen unserer Kreislaufs (System von vier Differenzengleichungen).

Addition liefert globales Erhaltungsprinzip

$$U_{n+1} + V_{n+1} + W_{n+1} + R_{n+1} = U_n + V_n + W_n + R_n.$$

Legt man noch einen Anfangszustand fest (etwa $U_0 = V_0 = W_0 = 0$ sowie R_0 = Gesamtwassermenge = 100) und wählt geeignete Werte für die Parameter, etwa $c_1 = \sqrt{12}$, $c_2 = c_4 = \sqrt{2}$, $c_3 = 1$, $c_5 = 2$ sowie $p = 17$, so können wir unser Modell „laufen lassen“.


Wassermenge
Zeit
0
10
20
30
40
50
60
70
80
90
100
0
5
10
15
20
25
30
U_n
V_n
W_n
R_n

D. Gleichgewichtswerte.

Simulation: Jede der Größen U_n, V_n, W_n und R_n nähert sich mit zunehmendem n einem Gleichgewichtswert U_∞, V_∞, W_∞, R_∞, wenn wir die Steuerungsparameter nicht ändern.

Bestimme Gleichgewichtswerte ohne (zeitaufwendige) Simulation:

$$\begin{aligned} p &= f_1(U_\infty) + f_2(U_\infty) \\ f_1(U_\infty) &= f_3(V_\infty) + f_4(V_\infty) \\ f_5(W_\infty) &= f_2(U_\infty) + f_4(V_\infty). \end{aligned}$$

Die vierte Gleichung ($p = f_3(V_\infty) + f_5(W_\infty)$) ist redundant.

Hier – im Gegensatz zur „Praxis“ – Gleichungen einfach (Dreiecksform). Vorsicht: Die theoretisch ermittelten Gleichgewichtswerte können, aber müssen nicht im Fassungsbereich der Behälter liegen (Nebenbedingungen).

E. Steuerung.

Wesentliches Ziel von Simulationen: Optimierung des Systemverhaltens bzw. Entscheidungshilfen für die Steuerung des Systems.

In unserem Beispiel etwa: Wie muss man die Steuerungsparameter wählen, damit sich ein erwünschter (vorgegebener) Gleichgewichtszustand einstellt (auch hier: Nebenbedingungen, man kann z.B. die Hähne nicht beliebig weit öffnen).

Fixiert man p, so führt dies in unserem Fall zu drei Bedingungen für die fünf Parameter $c_1, \ldots, c_5$:

$$\begin{aligned} c_1 &= -c_2 + p/\sqrt{U_\infty} \\ c_4 &= -c_3 + c_1\sqrt{U_\infty}/\sqrt{V_\infty} \\ c_5 &= c_2\sqrt{U_\infty}/\sqrt{W_\infty} + c_4\sqrt{V_\infty}/\sqrt{W_\infty}. \end{aligned}$$

In realen Systemen ist ein solches Steuerungsproblem nicht explizit lösbar, man wird es nur näherungsweise und iterativ lösen können.

F. Kritik.

Realität $\rightarrow$ mathematisches Modell
$\rightarrow$ Algorithmus
$\rightarrow$ numerische Simulation der Realität.

Bei jedem dieser drei Übergänge haben wir Fehler begangen.

— Modellierungsfehler. Unser Modell setzt wirbelfreien Wasserfluss voraus; in der Realität werden sich aber Wirbel bilden. Die Torricellische Ausflussformel ist nur gültig, wenn sich die Spiegelhöhe langsam ändert und keine Druckdifferenz zwischen Spiegel und Austrittsöffnung besteht, Voraussetzungen, die in der Realität nicht immer erfüllt sind.

— Diskretisierungsfehler. Wir haben den stetigen Strom des Wassers durch „Durchschnittswerte“ (bez. Zeit und Raum) ersetzt.

— Rundungsfehler. Computer „rechnen falsch“.

1.2 Linearisierung und Iterationsverfahren am Beispiel des Newton-Verfahrens

Problem. *Bestimme die Nullstelle(n) einer Funktion*

$$f : \mathbb{R} \supseteq D \to \mathbb{R}, \qquad x \mapsto f(x),$$

bzw. die Lösung(en) der Gleichung

$$f(x) = 0, \qquad x \in D\ .$$

Konkreter: Bestimme $\sqrt{a}$, $a > 0$, d.h. die positive Nullstelle der Funktion

$$f : \mathbb{R} \to \mathbb{R}, \qquad x \mapsto x^2 - a,$$

mit Hilfe der Grundrechenarten.

Mathematischer Hintergrund.

[NIELS HENRIK ABEL (1802–1829)], [EVARISTE GALOIS (1811–1832)]:
Es ist unmöglich, die Nullstellen allgemeiner nichtlinearer Funktionen elementar zu berechnen.
Präziser: Die n Lösungen einer Gleichung der Form

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

können für $n > 4$ i.A. nicht mit Hilfe der Grundrechenarten und der Wurzelfunktionen durch die Koeffizienten dargestellt werden. Für $n = 2$:

$$x_{1,2} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0}}{2}.$$

Für $n = 3$ und $n = 4$ gibt es ähnliche (kompliziertere) Formeln.

D.h.: *Bei der Nullstellenbestimmung nichtlinearer Funktionen (oder, was dasselbe ist, bei der Lösung nichtlinearer Gleichungen) ist man so gut wie immer auf numerische Verfahren angewiesen!*

Angenommen, x_0 ist Näherung für $\sqrt{a}$ mit dem Fehler e:

$$\sqrt{a} = x_0 + e \quad \text{(z.B. } x_0 = a\text{)}.$$

Gesucht ist eine bessere Näherung x_1. Taylor-Entwicklung:

$$0 = f(\sqrt{a}) = f(x_0 + e) = \underbrace{f(x_0) + f'(x_0)\, e}_{\text{Taylor-Polynom}} + \frac{1}{2}\, f''(\xi)\, e^2$$

mit $\xi \in (x_0, \sqrt{a})$, falls $x_0 < \sqrt{a}$, bzw. $\xi \in (\sqrt{a}, x_0)$, falls $x_0 > \sqrt{a}$.

Man kann die Gleichung $0 = f(x_0) + f'(x_0)\, e + \frac{1}{2}\, f''(\xi)\, e^2$ nicht nach e auflösen (ξ ist unbekannt!). Ist e aber klein, so ist e^2 noch viel kleiner und wir vernachlässigen den Term mit dem Faktor e^2, d.h. wir betrachten die lineare Gleichung

$$0 = f(x_0) + f'(x_0)\, \tilde{e} \quad \text{mit Lösung} \quad \tilde{e} = -\frac{f(x_0)}{f'(x_0)} \quad \text{(falls } f'(x_0) \neq 0\text{)}.$$

Dann ist $x_1 := x_0 + \tilde{e} = x_0 - f(x_0)/f'(x_0)$ zwar keine Nullstelle von f, aber (hoffentlich) eine bessere Näherung für eine Nullstelle von f als x_0.

Auf die gleiche Weise gewinnt man aus x_1 eine neue Näherung x_2 usw. Man setzt ein Iterationsverfahren ein:

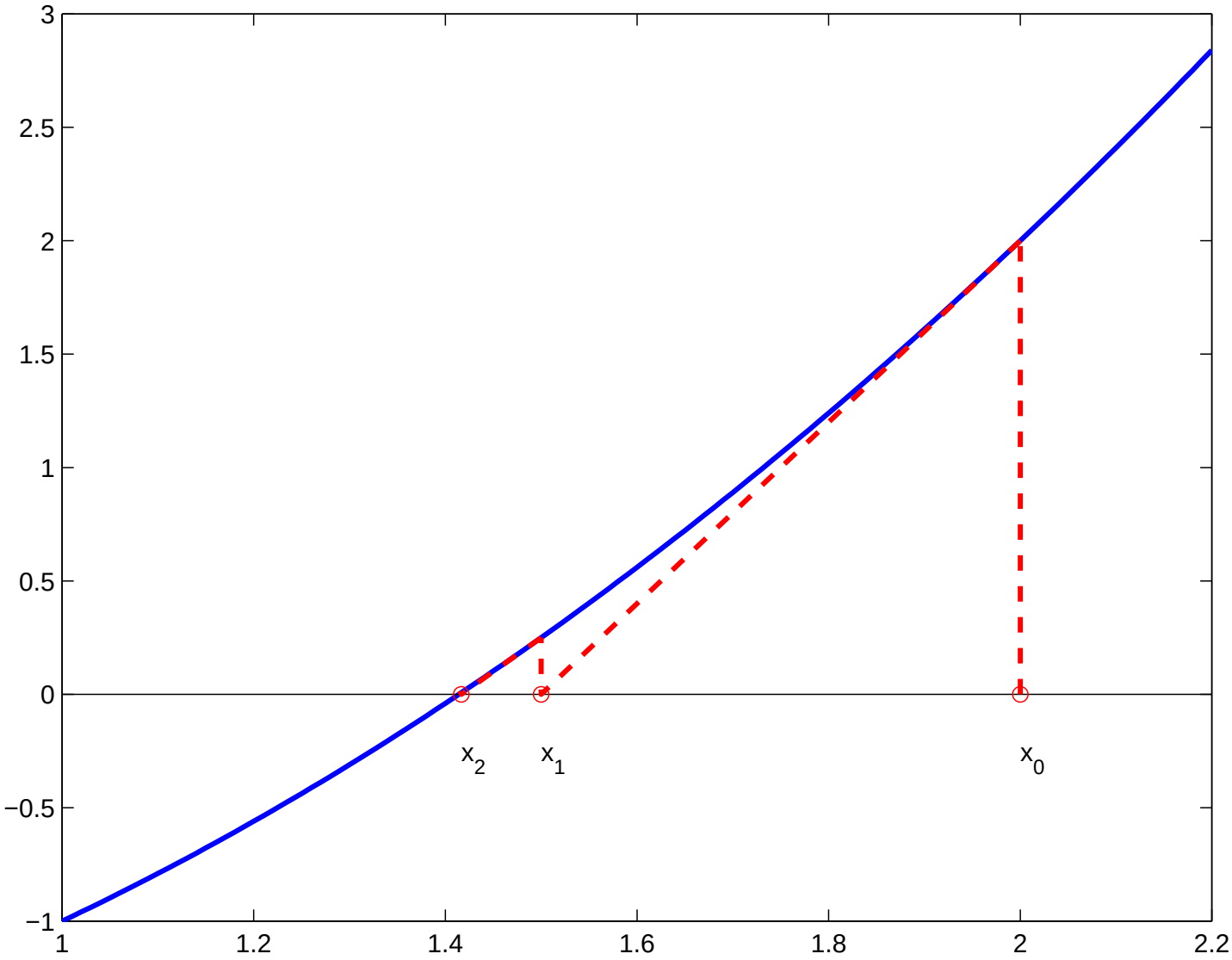
Wähle eine Ausgangsnäherung x_0 .

Für $m = 1, 2, \ldots$ iteriere gemäß

$$x_m := x_{m-1} - \frac{f(x_{m-1})}{f'(x_{m-1})}. \tag{1.3}$$

Für $f(x) = x^2 - a$ ergibt sich als Iterationsvorschrift

$$x_m := \frac{1}{2}\left(x_{m-1} + \frac{a}{x_{m-1}}\right) \qquad (m = 1, 2, \ldots). \tag{1.4}$$


3
2.5
2
1.5
1
0.5
0
−0.5
−1
1
1.2
1.4
1.6
1.8
2
2.2
x2
x1
x0

Man ersetzt das komplizierte Problem $f(x) = 0$ durch das lineare Problem $f(x_{m-1}) + f'(x_{m-1})\tilde{e} = 0$ und korrigiert $x_m = x_{m-1} + \tilde{e}$. Äquivalent: Wir betrachten die Tangente an den Graphen von f im Punkt $(x_{m-1}, f(x_{m-1}))$,

$$y = f(x_{m-1}) + f'(x_{m-1})\,(x - x_{m-1})\,,$$

und berechnen die Nullstelle x_m dieser linearen Funktion.

Die Idee der Linearisierung lässt sich also wie folgt beschreiben: Ersetze ein kompliziertes Problem durch ein benachbartes lineares Problem (bzw. durch eine Folge solcher Probleme). In unserem Beispiel wurde die komplizierte Gleichung $f(x) = 0$ durch eine Folge linearer Gleichungen, nämlich der Tangentengleichungen, ersetzt. Eine Linearisierung führt fast immer auf ein Iterationsverfahren, weil ein Korrekturschritt i.Allg. nicht ausreicht, um eine brauchbare Näherung für die Lösung des komplizierten Ausgangsproblems zu bestimmen.

Für $f(x) = x^2 - a$ mit $a = 2$:

x_0	x_1	x_2	x_3	x_4
2	1.5	1.41 ...	1.41421 ...	1.41421356237 ...

(Nur die korrekten Ziffern von x_2, x_3 und x_4 sind angeben.)

Es stellen sich folgende Fragen:

1. Konvergiert das Verfahren, d.h. gilt $\lim_{m\to\infty} x_m = \sqrt{a}$, für jede Wahl des Startwerts x_0?
   Offenbar nicht, z.B. für $x_0 = 0$ ist x_1 noch nicht einmal definiert.
2. Also, für welche x_0 konvergiert die Folge $\{x_m\}_{m\geq 0}$ gegen $\sqrt{a}$?
3. Wann bricht man das Verfahren ab? Schranken für den Abbruchfehler $|x_m - \sqrt{a}|$ sind erforderlich.

1.3 Diskretisierung und Stabilität am Beispiel der Wärmeleitungsgleichung

Problem. Die Temperatur

$$u(x,t), \qquad 0 \le x \le \pi,$$

in einem homogenen Stab mit konstantem Querschnitt habe zur Zeit $t = 0$ den Wert $u(x,0) = f(x)$. Der Stab sei wärmeisoliert – außer an den Rändern $x = 0$ und $x = \pi$, wo die Temperatur konstant auf $u(0,t) = u(\pi,t) = 0$ gehalten wird $(t > 0)$.

Bestimme die Wärmeverteilung $u(x,t^*)$, $0 \le x \le \pi$, *im Stab zur Zeit* $t^* > 0$.

Mathematisches Modell.

Energieerhaltungssatz und Fouriersches Gesetz („Wärme fließt in Richtung abfallender Temperatur und zwar umso intensiver, je größer die Temperaturdifferenzen sind“): Gesucht ist eine Funktion

$$u : [0, \pi] \times [0, \infty) \to \mathbb{R}, \quad (x, t) \mapsto u(x, t),$$

die die folgenden Eigenschaften besitzt:

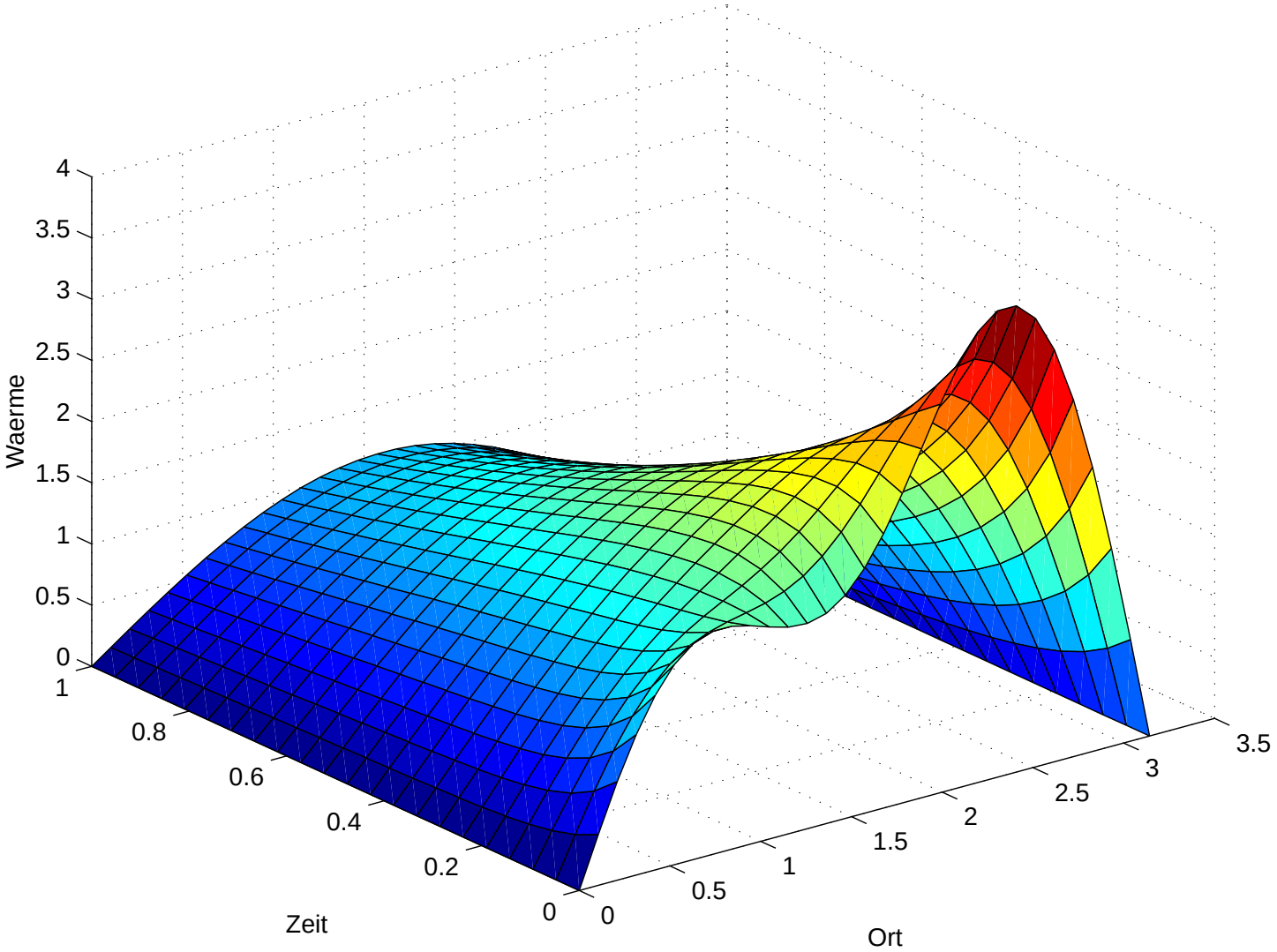
$$\frac{\partial u}{\partial t}(x, t) = \gamma^2 \frac{\partial^2 u}{\partial x^2}(x, t) \, , \qquad 0 < x < \pi, \quad t > 0 \tag{1.5a}$$

mit einer Materialkonstanten $\gamma (\equiv 1)$. Außerdem

$$u(x, 0) = f(x), \qquad 0 \le x \le \pi, \qquad \text{(Anfangsbedingung)} \tag{1.5b}$$

z.B. $f(x) = 3 \sin(x) - \sin(2x) + \sin(3x)$,

$$u(0, t) = u(\pi, t) = 0, \qquad t \ge 0, \qquad \text{(Randbedingungen)}. \tag{1.5c}$$


Waerme
4
3.5
3
2.5
2
1.5
1
0.5
0
Zeit
1
0.8
0.6
0.4
0.2
0
Ort
0
0.5
1
1.5
2
2.5
3
3.5

(1.5a) heißt Wärmeleitungsgleichung.

Für komplizierte Anfangs- und Randbedingungen oder ortsabhängige Materialkonstanten kann man die Lösung solcher Probleme nicht explizit angeben. Es lässt sich jedoch beweisen, dass (1.5a), (1.5b), (1.5c) auch dann ein sachgemäß gestelltes Problem (im Sinne von [JACQUES SALOMON HADAMARD (1865–1963)]) ist:

1. Es besitzt eine Lösung.
2. Diese ist eindeutig.
3. Sie hängt ferner stetig von den Daten (in diesem Fall den Anfangs- und Randbedingungen) ab!

Diskretisierung durch finite Differenzen: Bestimme u nur noch auf einem Gitter oder Netz

$$\Omega_{h,k} = \{(x_i, t_j) \,:\, x_i = ih \text{ für } i = 0, 1, \ldots n+1,\ t_j = jk \text{ für } j = 0, 1, \ldots\}\,.$$

Dabei sind $h := \pi/(n+1)$ bzw. $k > 0$ die Schrittweiten (Gitter- oder Netzweiten) in x- bzw. t-Richtung. Unsere Näherung für $u(x_i, t_j)$ werden wir mit $u_{i,j}$ bezeichnen.

In einem zweiten Schritt müssen wir die partiellen Ableitungen $\partial u/\partial t$ bzw. $\partial^2 u/\partial x^2$ aus (1.5a) durch Ausdrücke annähern, die wir auf dem Gitter bestimmen können.

Dazu betrachten wir zuerst eine Funktion in *einer* Variablen,

$$f : \mathbb{R} \supseteq I = [\alpha, \beta] \to \mathbb{R}\,, \quad x \mapsto f(x),$$

und nehmen an, dass f in $x_0 \in I$ differenzierbar ist. Weil

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

gilt, liegt es nahe, $f'(x_0)$ etwa durch eine der drei Formeln

$$\frac{f(x_0 + h) - f(x_0)}{h} \qquad \text{(Vorwärtsdifferenz)} \qquad (1.6)$$

$$\frac{f(x_0) - f(x_0 - h)}{h} \qquad \text{(Rückwärtsdifferenz)} \qquad (1.7)$$

$$\frac{f(x_0 + h) - f(x_0 - h)}{2h} \qquad \text{(zentrale Differenz)} \qquad (1.8)$$

anzunähern. Dabei soll die Schrittweite h natürlich „klein“ sein.

Die zweite Ableitung $f''(x_0)$ approximieren wir durch eine zentrale Differenz zweiter Ordnung

$$\begin{aligned} f''(x_0) &\sim \frac{f'(x_0+h)-f'(x_0)}{h} \\ &\sim \frac{\frac{f(x_0+h)-f(x_0)}{h} - \frac{f(x_0)-f(x_0-h)}{h}}{h} \\ &= \frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}. \end{aligned} \tag{1.9}$$

Diskretisierungsfehler:

a) Ist f in I zweimal stetig differenzierbar, so gilt

$$\frac{f(x_0+h)-f(x_0)}{h} = f'(x_0) + C_1 h$$

mit $|C_1| \leq \frac{1}{2} \max_{x\in I} |f''(x)|$ (analog für Rückwärtsdifferenz).

b) Ist f in I dreimal stetig differenzierbar, so gilt

$$\frac{f(x_0+h)-f(x_0-h)}{2h} = f'(x_0) + C_3 h^2$$

mit $|C_3| \leq \frac{1}{6} \max_{x\in I} |f'''(x)|$.

c) Ist f in I viermal stetig differenzierbar, so gilt

$$\frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2} = f''(x_0) + C_4 h^2$$

mit $|C_4| \leq \frac{1}{12} \max_{x\in I} |f^{(4)}(x)|$.

Analog bei partiellen Ableitungen, z.B.:

$$\frac{\partial u}{\partial t}(x,t) \approx \frac{u(x,t+k)-u(x,t)}{k},$$

$$\frac{\partial u}{\partial x}(x,t) \approx \frac{u(x+h,t)-u(x,t)}{h}$$

und

$$\frac{\partial^2 u}{\partial x^2}(x,t) \approx \frac{u(x+h,t)-2u(x,t)+u(x-h,t)}{h^2}.$$

Approximiere $\partial u/\partial t(x, t_0)$ durch eine Vorwärtsdifferenz und $\partial u^2/\partial x^2(x, t_0)$ durch eine zentrale Differenz zweiter Ordnung.

Für $n = 4$ ergibt sich dann

$$\begin{aligned}
\frac{u_{1,1} - u_{1,0}}{k} &= \frac{u_{0,0} - 2u_{1,0} + u_{2,0}}{h^2}, \\
\frac{u_{2,1} - u_{2,0}}{k} &= \frac{u_{1,0} - 2u_{2,0} + u_{3,0}}{h^2}, \\
\frac{u_{3,1} - u_{3,0}}{k} &= \frac{u_{2,0} - 2u_{3,0} + u_{4,0}}{h^2}, \\
\frac{u_{4,1} - u_{4,0}}{k} &= \frac{u_{3,0} - 2u_{4,0} + u_{5,0}}{h^2}.
\end{aligned}$$

Wir lösen diese Gleichungen nach $u_{i,1}$ auf und setzen $\tau := k/h^2$.
(Beachte $u_{0,0} = u_{5,0} = 0$, Randbedingungen!)

$$\begin{bmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \\ u_{4,1} \end{bmatrix} = \begin{bmatrix} u_{1,0} \\ u_{2,0} \\ u_{3,0} \\ u_{4,0} \end{bmatrix} + \tau \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} u_{1,0} \\ u_{2,0} \\ u_{3,0} \\ u_{4,0} \end{bmatrix}.$$

Alle Einträge auf der rechten Seite dieser Gleichung sind bekannt (Anfangsbedingung!), wir können also die Näherungswerte $u_{i,1}$ für die Zeitschicht $t = k$ bestimmen. Analog kann man danach aus den Werten $u_{i,1}$ die Werte $u_{i,2}$ für die Zeitschicht $t = 2k$ berechnen, usw.

Explizites Euler-Verfahren:

Ziel: Berechne Näherungen $u_{i,j}$ für die Lösung $u(ih, jk)$ von (1.5a), (1.5b), (1.5c), wobei $1 \leq i \leq n$, $1 \leq j \leq m$.

Bestimme $\boldsymbol{u}^{(0)} := [u_{1,0}, u_{2,0}, \ldots, u_{n,0}]^T = [f(x_1), f(x_2), \ldots, f(x_n)]^T$
aus der gegebenen Anfangsbedingung.

Für $j = 1, 2, \ldots, m$

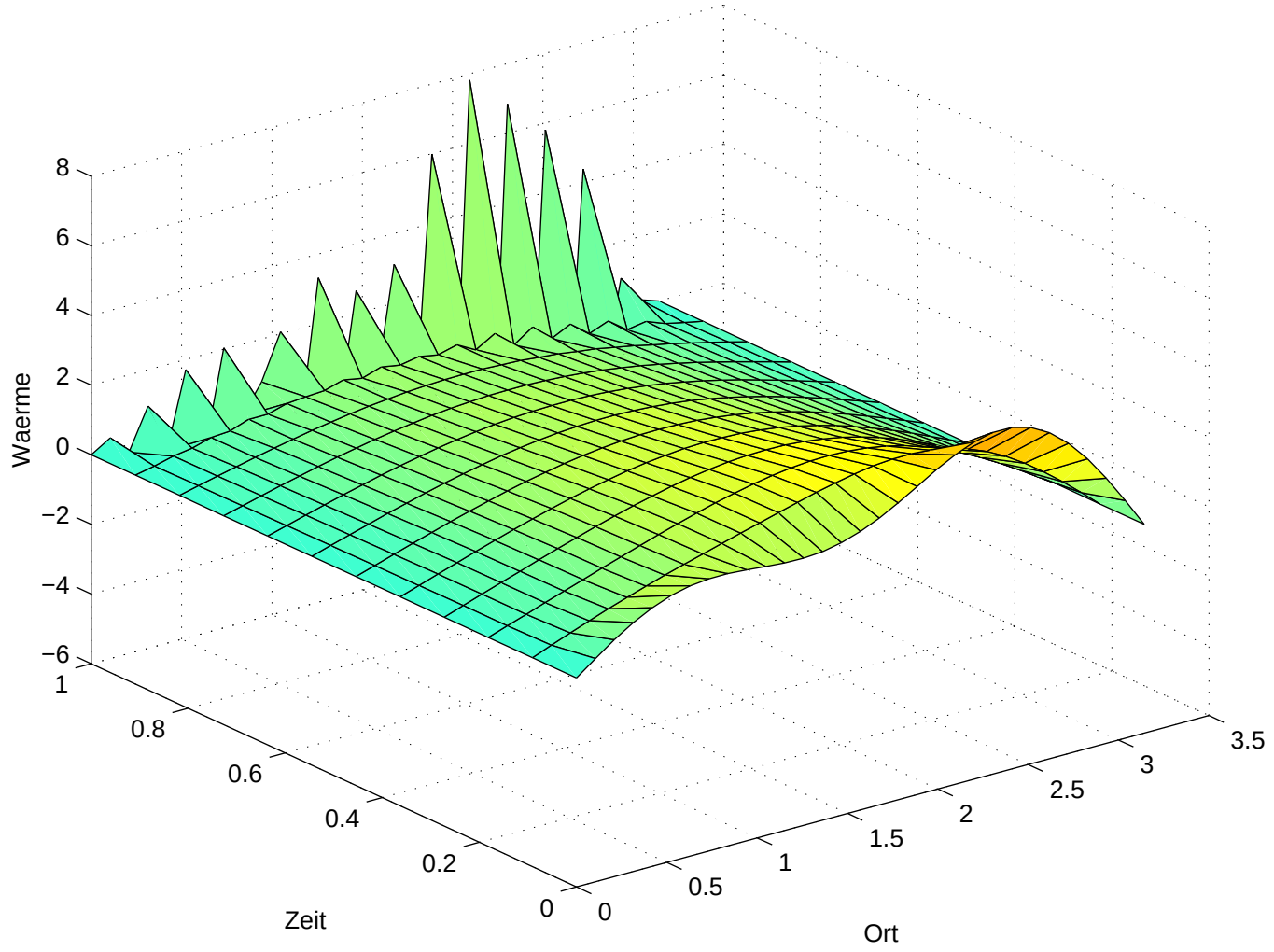
berechne $\boldsymbol{u}^{(j)} = [u_{1,j}, u_{2,j}, \ldots, u_{n,j}]^T$ durch

$$\boldsymbol{u}^{(j)} = [I + \tau A_h]\, \boldsymbol{u}^{(j-1)}. \tag{1.10}$$

Dabei bezeichnen

- I die Einheitsmatrix der Dimension $n \times n$,
- $\tau = k/h^2$,
- A_h die Tridiagonalmatrix $A_h = \text{tridiag}(1, -2, 1) \in \mathbb{R}^{n \times n}$ (s.o.) und
- $\boldsymbol{u}^{(j)}$ den Vektor, der die Näherungen für die Temperatur zur Zeit $t = jk$ enthält.

Für $k = 1/11$, $h = \pi/30$:



Die folgende Tabelle zeigt, dass man für kleinere Werte von k sogar noch unsinnigere Werte erhält. Erst wenn die Zeitschrittweite k „winzig“ ist, ergeben sich brauchbare Näherungen.

k	$u(7h,1)$	$u(14h,1)$	$u(21h,1)$	$u(28h,1)$
1/11	3.7 10^{0}	3.1 10^{0}	-6.3 10^{0}	-1.3 10^{0}
1/50	-2.7 10^{23}	-3.7 10^{23}	8.5 10^{23}	-2.7 10^{23}
1/100	1.4 10^{24}	-6.1 10^{25}	1.1 10^{26}	-3.5 10^{25}
1/150	6.5 10^{6}	-1.2 10^{7}	1.2 10^{7}	-3.3 10^{6}
1/200	7.2 10^{-1}	1.1 10^{0}	9.1 10^{-1}	2.4 10^{-1}
1/250	7.2 10^{-1}	1.1 10^{0}	9.1 10^{-1}	2.4 10^{-1}
0	7.2 10^{-1}	1.1 10^{0}	9.1 10^{-1}	2.4 10^{-1}

($h = \pi/30$, $k = 0$ bedeutet hier, dass es sich bei den zugehörigigen u-Werten um die Funktionswerte der exakten Lösung handelt.)

Implizites Euler-Verfahren:

Einziger Unterschied zum expliziten Verfahren:

$$\frac{\partial u}{\partial t}(x,t) \approx \frac{u(x,t) - u(x,t-k)}{k}.$$

Nun ergibt sich für die Gitterpunkte auf der ersten Zeitschicht $t = k$ (im Spezialfall $n = 4$)

$$\begin{bmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \\ u_{4,1} \end{bmatrix} = \begin{bmatrix} u_{1,0} \\ u_{2,0} \\ u_{3,0} \\ u_{4,0} \end{bmatrix} + \tau \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \\ u_{4,1} \end{bmatrix}.$$

Die Unbekannten auf der neuen Zeitschicht sind hier implizit durch die Werte auf der alten Zeitschicht gegeben, nämlich als Lösung eines linearen Gleichungssystems. D.h. in jedem Zeitschritt des impliziten Euler-Verfahrens muss ein lineares Gleichungssystem gelöst werden.

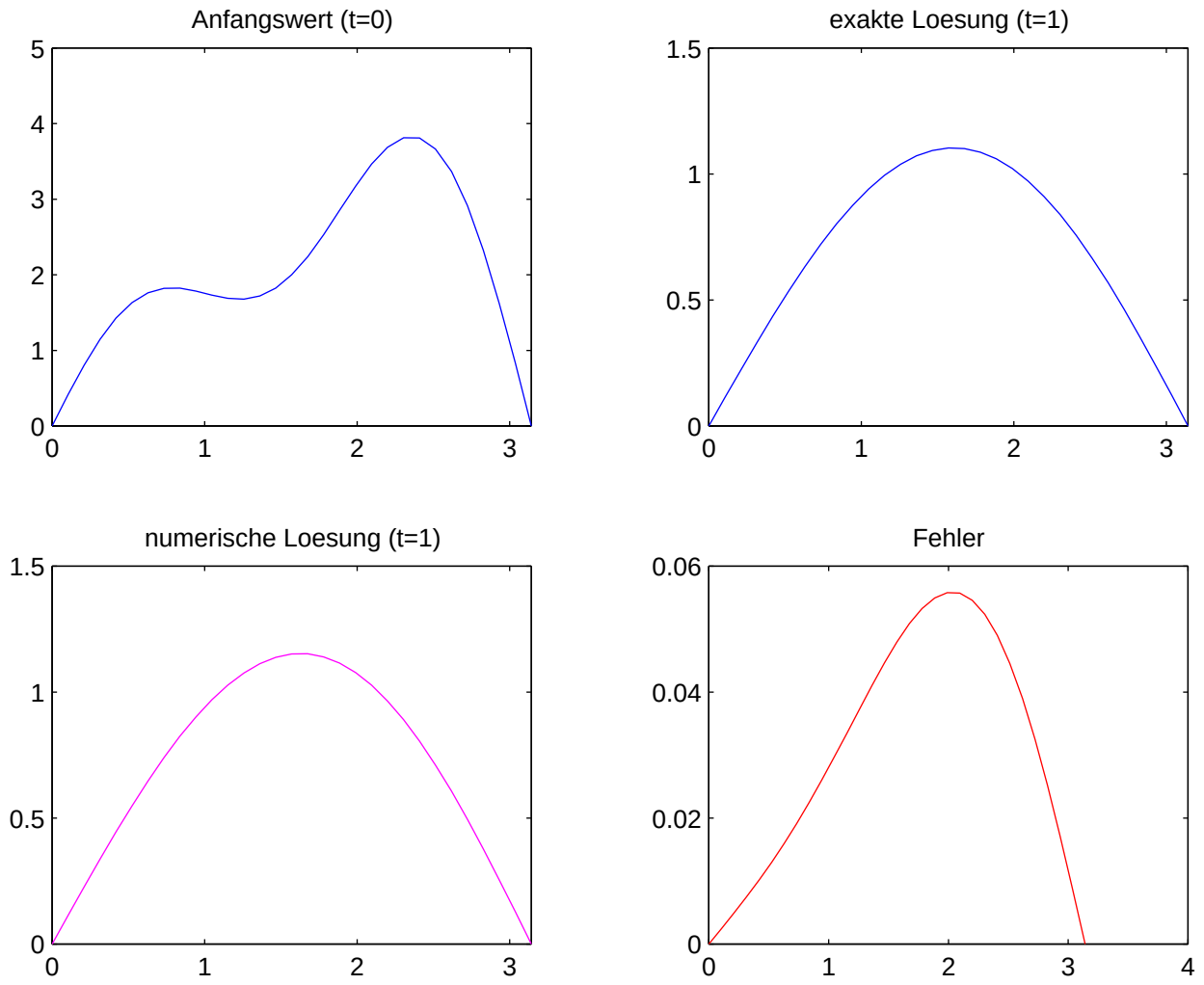
Bestimme $\boldsymbol{u}^{(0)} := [u_{1,0}, u_{2,0}, \ldots, u_{n,0}]^T = [f(x_1), f(x_2), \ldots, f(x_n)]^T$
aus der gegebenen Anfangsbedingung.

Für $j = 1, 2, \ldots, m$

berechne $\boldsymbol{u}^{(j)} = [u_{1,j}, u_{2,j}, \ldots, u_{n,j}]^T$ als Lösung von

$$(I - \tau A_h)\boldsymbol{u}^{(j)} = \boldsymbol{u}^{(j-1)}. \tag{1.11}$$

Für $k = 1/11$ und $h = \pi/30$ ergibt sich:


Anfangswert (t=0)
exakte Loesung (t=1)
numerische Loesung (t=1)
Fehler

Warum verhalten sich explizites und implizites Verfahren so unterschiedlich?

Bezeichnungen: Exakte Lösung für $t = jk$

$$\boldsymbol{u}_*^{(j)}(h,k) := [u(h,jk), u(2h,jk), \ldots, u(nh,jk)]^T \in \mathbb{R}^n.$$

Näherungslösung für $t = jk$

$$\boldsymbol{u}_{\text{Verf}}^{(j)}(h,k) \in \mathbb{R}^n \quad \text{mit Verf} \in \{\text{ex}, \text{im}\}.$$

Globaler Diskretisierungsfehler dieser Verfahren

$$\boldsymbol{e}_{\text{Verf}}^{(j)}(h,k) := \boldsymbol{u}_*^{(j)}(h,k) - \boldsymbol{u}_{\text{Verf}}^{(j)}(h,k).$$

Von einem „vernünftigen“ Verfahren wird man erwarten, dass der globale Diskretisierungsfehler gegen Null strebt, wenn die Schrittweiten klein werden,

$$\boldsymbol{e}^{(j)}(h,k) \to \mathbf{0} \quad \text{für } jk \text{ fixiert und } h,k \to 0.$$

Der lokale Diskretisierungsfehler der beiden Verfahren ist durch

$$\begin{aligned} \boldsymbol{d}_{\text{ex}}^{(j)}(h,k) &:= \boldsymbol{u}_*^{(j)}(h,k) - [I+\tau A_h]\boldsymbol{u}_*^{(j-1)}(h,k), \\ \boldsymbol{d}_{\text{im}}^{(j)}(h,k) &:= [I-\tau A_h]\boldsymbol{u}_*^{(j)}(h,k) - \boldsymbol{u}_*^{(j-1)}(h,k) \end{aligned}$$

($\tau = k/h^2$) erklärt. Er gibt an, wie gut die exakte Lösung die jeweilige Differenzenapproximation erfüllt (vgl. (1.10) und (1.11)).

Es gilt (C_1, C_2 unabhängig von h, k, j und ℓ)

$$\left|[\boldsymbol{d}_{\text{Verf}}^{(j)}(h,k)]_\ell\right| \le k(C_1 k + C_2 h^2) \qquad (\ell = 1,2,\ldots,n), \quad \text{Verf} \in \{\text{ex}, \text{im}\}.$$

Die lokalen Diskretisierungsfehler der beiden Verfahren sind qualitativ gleich. Insbesondere erfüllen sie

$$\boldsymbol{d}^{(j)}(h,k) \to \mathbf{0} \quad \text{für} \quad k,h \to 0$$

(solche Verfahren nennt man konsistent). Dass sich die globalen Diskretisierungsfehler trotzdem erheblich unterscheiden, liegt am unterschiedlichen Stabilitätsverhalten der beiden Algorithmen.

Entscheidend: Zusammenhang zwischen globalen und lokalen Diskretisierungsfehlern

$$\begin{aligned} \boldsymbol{e}_{\text{ex}}^{(j)}(h,k) &= [I+\tau A_h]\boldsymbol{e}_{\text{ex}}^{(j-1)}(h,k) + \boldsymbol{d}_{\text{ex}}^{(j)}(h,k), \\ [I-\tau A_h]\boldsymbol{e}_{\text{im}}^{(j)}(h,k) &= \boldsymbol{e}_{\text{im}}^{(j-1)}(h,k) + \boldsymbol{d}_{\text{im}}^{(j)}(h,k). \end{aligned}$$

Einheitliche Schreibweise:

$$\boldsymbol{e}^{(j)}(h,k) = B_{h,k}\,\boldsymbol{e}^{(j-1)}(h,k) + \boldsymbol{g}_{h,k}^{(j)} \tag{1.12}$$

mit der Fehlerfortpflanzungsmatrix

$$B_{h,k} := \begin{cases} I+\tau A_h & \text{für das explizite Euler-Verfahren,} \\ (I-\tau A_h)^{-1} & \text{für das implizite Euler-Verfahren} \end{cases} \tag{1.13}$$

und einem Vektor

$$\boldsymbol{g}_{h,k}^{(j)} := \begin{cases} \boldsymbol{d}_{\text{ex}}^{(j)}(h,k) & \text{(explizit),} \\ (I-\tau A_h)^{-1}\,\boldsymbol{d}_{\text{im}}^{(j)}(h,k) & \text{(implizit).} \end{cases} \tag{1.14}$$

Betrachten wir nun ganz abstrakt das Wachstumsverhalten einer Vektorfolge $\{\boldsymbol{e}^{(j)}\}_{j=0,1,\ldots}$, die rekursiv durch

$$\boldsymbol{e}^{(j)} := B\boldsymbol{e}^{(j-1)} + \boldsymbol{g}^{(j)} \quad (j = 1, 2, \ldots) \quad \text{mit } \boldsymbol{e}^{(0)} = \mathbf{0}$$

gegeben ist. Es gilt

$$\begin{aligned}
\boldsymbol{e}^{(1)} &= \boldsymbol{g}^{(1)}, \\
\boldsymbol{e}^{(2)} &= B\boldsymbol{e}^{(1)} + \boldsymbol{g}^{(2)} = B\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)}, \\
\boldsymbol{e}^{(3)} &= B\boldsymbol{e}^{(2)} + \boldsymbol{g}^{(3)} = B^2\boldsymbol{g}^{(1)} + B\boldsymbol{g}^{(2)} + \boldsymbol{g}^{(3)}, \\
\vdots &= \vdots \\
\boldsymbol{e}^{(j)} &= B\boldsymbol{e}^{(j-1)} + \boldsymbol{g}^{(j)} = \sum_{m=1}^{j} B^{j-m}\boldsymbol{g}^{(m)}.
\end{aligned}$$

Das bedeutet

$$\begin{aligned} \|\boldsymbol{e}^{(j)}\|_2 = \|\sum_{m=1}^{j} B^{j-m}\boldsymbol{g}^{(m)}\|_2 &\leq \sum_{m=1}^{j} \|B\|_2^{j-m}\|\boldsymbol{g}^{(m)}\|_2 \\ &\leq [\max_{1\leq m\leq j} \|\boldsymbol{g}^{(m)}\|_2] \sum_{m=1}^{j} \|B\|_2^{j-m}. \end{aligned} \tag{1.15}$$

Der erste Faktor $\max_{1\leq m\leq j} \|\boldsymbol{g}^{(m)}\|_2$ wird bei beiden Euler-Verfahren (wie bei allen konsistenten Differenzenschemata) beliebig klein, wenn h und k nur genügend klein gewählt sind. Der zweite Faktor,

$$\sum_{m=1}^{j} \|B\|_2^{j-m} = \begin{cases} j, & \text{falls } \|B\|_2 = 1, \\ (\|B\|_2^j - 1)/(\|B\|_2 - 1), & \text{falls } \|B\|_2 \neq 1 \end{cases},$$

ist beschränkt falls $\|B\|_2 < 1$ (nämlich durch $1/(1-\|B\|_2)$). Ist aber $\|B\|_2 \geq 1$, so wächst er über alle Schranken.

Wir nennen nun ein Differenzenverfahren stabil, wenn die Norm der zugehörigen Fehlerfortpflanzungsmatrix kleiner als 1 ist (und andernfalls instabil).

Mit (1.15) haben wir ein „Metatheorem“ der numerischen Mathematik bewiesen: Stabilität (dh. der zweite Faktor auf der rechten Seite von (1.15) ist beschränkt) und Konsistenz (dh. der erste Faktor strebt mit h und k gegen 0) eines Differenzenschemas implizieren, dass der globale Diskretisierungsfehler ebenfalls gegen 0 geht (für $h, k \to 0$) – man spricht dann von einem konvergenten Verfahren. Kürzer gefasst,

Stabilität + Konsistenz $\Rightarrow$ **Konvergenz**.

Für das explizite Euler-Verfahren gilt

$$\|B_{h,k}\|_2 = \|I + \tau A_h\|_2 < 1 \quad \text{genau dann, wenn} \quad \tau = \frac{k}{h^2} \leq \frac{1}{2}$$

(das explizite Euler-Verfahren ist nur bedingt stabil, d.h. unter der oben angegebenen Bedingung), während das implizite Euler-Verfahren unbedingt stabil ist (d.h. ohne Bedingungen an h und k),

$$\|B_{h,k}\|_2 = \|(I - \tau A_h)^{-1}\|_2 < 1 \quad \text{für alle } h \text{ und } k.$$