Einleitung

(Christoph Pöppe, Monika Wierse)

Wie das Wasser an einem Schiff vorbeiströmt, die Luft an einem Flugzeug, die Verbrennungsgase durch einen Motor oder eine Turbine: Das kann man alles ausrechnen; die dafür wesentlichen physikalischen Gesetze sind bekannt und als Formeln ausdrückbar. Es gelingt ja auch, einen Flugzeugflügel erst mit dem Computer zu berechnen und dann zu bauen – und siehe da, er fliegt, wie er soll.

Aber von den Formeln für die physikalischen Gesetze bis zum richtig entworfenen Flügel ist es ein weiter Weg. In unserem Kurs haben wir diesen Weg von Anfang bis Ende mitverfolgt – manchmal etwas hastig, weil es wirklich ein langer Weg ist.

Am Anfang steht etwas, das uns aus der Schule gerade noch geläufig ist: der Begriff der Ableitung. Natalja Deng hat unser Gedächtnis kurz und treffsicher aufgefrischt und uns gleich noch ein Lieblingswerkzeug der Differential- und Integralrechner vorgestellt, von dem auch die Computer-Rechner (die „numerischen Mathematiker“) ausgiebig Gebrauch machen: die Taylorreihe.

Christoph Grothaus hat uns an mehreren Beispielen erläutert, wie sich der Ableitungsbegriff einigermaßen zwingend aus einer physikalischen Beschreibung von Naturvorgängen ergibt. Zunächst geht es nur um Ableitungen nach einer einzigen unabhängigen Variablen, der Zeit. Die physikalischen Gesetze nehmen die Form von Gleichungen an, in denen eine unbekannte Funktion und ihre Ableitung vorkommen: Das sind gewöhnliche Differentialgleichungen. Christoph hat uns auch vorgeführt, wie man einige (einfache) von ihnen mit Papier und Bleistift lösen kann.

Dieser himmlische Zustand endet recht bald, wenn die Gleichungen auch nur ein bisschen komplizierter werden. Man muss zu irdischen Mitteln greifen: diskretisieren. Lisa Huber hat uns erklärt, wie das geht, und Christoph Pöppe hat das zur allgemeinen Erheiterung in Begriffen von himmlischer Seligkeit und irdischer Mühsal, Sünde und Vergebung interpretiert. Florian Conrad hat uns den Unterschied zwischen dem Pfad der Tugend und dem computerberechneten Irrweg an einem Beispiel vorgerechnet.

Immerhin: Ewige Verdammnis ist kein unvermeidbares Schicksal. Das Paradies existiert – und ist eindeutig bestimmt (Karin Heitzmann). Das kann man beweisen, und Christoph Pöppe konnte es nicht lassen, den Kursteilnehmern dieses Prachtstück harter Mathematik vorzuführen, samt den typischen Gedankengängen der Analysis, die beim ersten Mal doch recht fremdartig anmuten. Sebastian Tivig hat dann wirklich einmal etwas programmiert: Räuber-Beute-Modelle, und uns dadurch einige theoretisch gewonnene Sätze mit dem Computer bestätigt.

Damit verließen wir die noch einigermaßen übersichtliche Welt der gewöhnlichen Differentialgleichungen und wandten uns den Funktionen zu, die zugleich von mehreren Veränderlichen abhängen: der Zeit und/oder

einer oder mehreren Ortskoordinaten. Die Ouvertüre mit der Einführung der Begriffe kam wieder von Natalja. Wenn Ableitungen der unbekannten Funktion nach verschiedenen Variablen zusammen in einer Gleichung vorkommen, spricht man von einer partiellen Differentialgleichung (obgleich nicht die Gleichung partiell ist, sondern allenfalls die Ableitungen). Der Zoo dieser Tierchen ist so unübersichtlich, dass man sich im Allgemeinen darauf beschränkt, sich einige besonders charakteristische Exemplare anzusehen. Das hat Petra Kersting für uns getan. Jörn-Thorsten Paßmann hat uns dann vorgerechnet, wie die Kunst des Diskretisierens, die wir an den gewöhnlichen Differentialgleichungen erlernt hatten, auf sie anzuwenden ist, und Arne Schneck hat uns einen vollkommen anderen Ansatz vorgeführt, der andere und häufig bessere Diskretisierungen liefert: die finiten Elemente. So oder so: Was dem Computer nach der theoretischen Vorarbeit zu tun bleibt, ist große lineare Gleichungssysteme zu lösen. Bloß nicht exakt lösen, sondern nur ungefähr; das geht schneller und wird genauer (Christian Moldenhauer)!

Dann ging es verschärft auf die Strömungsprobleme zu. Wenn das strömende Medium ein Gas ist, können mangels innerer Reibung Phänomene wie Stoßwellen (der Überschallknall) auftreten, die einem bei der näherungsweisen (Computer-)Berechnung mächtig zu schaffen machen und deswegen theoretisch genauer anzuschauen sind (Christine Rogg). Und bis man die Gleichung hergeleitet hat, die Strömungsphänomene wirklich hinreichend korrekt beschreibt, die Navier-Stokes-Gleichung, vergeht eine ganze Weile (Matthias Klotz).

Silja Kinnebrock hat uns eine Klasse von Verfahren vorgestellt, die gerade für Strömungsprobleme im Allgemeinen und die Navier-Stokes-Gleichung im Besonderen den Rechenaufwand dramatisch verringern: die Mehrgitterverfahren. Und selbst die helfen nicht, wenn es um Wirbelphänomene auf kleinem und kleinstem Raum geht, so klein, dass die größten Computer mit dem Diskretisieren nicht nachkommen. Für Turbulenzen muss man sich etwas Neues ausdenken (Sabine Schamberg).

Und wenn man schließlich die Lösung in Form von Abermillionen Zahlen im Computer stecken hat, möchte man sich von ihr ein Bild machen. Wenn die Strömung (wie meistens) dreidimensional ist, dann ist es eine Kunst für sich, das Wesentliche (was immer das ist) auf einem zweidimensionalen Bildschirm vor Augen zu führen (Bastian Katz).

Zum Schluss hat uns Monika Wierse Ergebnisse aus ihrer Arbeit vorgeführt, in die alle bei uns diskutierten Weisheiten (und noch viel mehr) eingeflossen sind.

Für Strömungen um Fluggeräte ist eigentlich das Deutsche Forschungszentrum für Luft- und Raumfahrt (DLR) die richtige Adresse. Monika kennt einen Menschen, der dort praktische Probleme am Computer rechnet, und wir fuhren – gemeinsam mit dem Astronomie-Kurs – zur DLR nach Köln-Porz. Aber ach!, die Leute, die uns dazu hätten Auskunft geben können, waren in Urlaub. So haben wir etwas über Raumfahrt und Infrarot-Flugzeug-Astronomie erfahren (auch ganz nett), haben den Windkanal, in dem die Strömungen im Experiment gemessen statt berechnet werden, wenigstens von außen gesehen und einiges über ihn erfahren – und bekamen ganz unerwartet doch noch eine Zugabe: Herr Georg Hertkorn erforscht Verkehrsstaus mit Hilfe eines diskretisierten Modells. Das ist ja das Schöne an der Mathematik: Ein wandernder Verkehrsstau und eine Überschall-Stoßwelle sind, von einem hinreichend abstrakten Standpunkt aus betrachtet, eigentlich dasselbe. So konnten wir von Herrn Hertkorn doch noch etwas über unser Thema erfahren.

Da der Stoff der Vorträge nicht einfach war, gab es zwischendurch häufig Meditationspäuschen: Schweigeminuten, in denen man über das soeben Gehörte nachdenken konnte, und dann zuweilen Erläuterungen – zuweilen ausufernde – von Kursleiter Christoph. Einige von ihnen hat er in die vorliegende Dokumentation eingearbeitet und, wie diesen Absatz, durch Kursivschrift gekennzeichnet.

Grundlegendes zur Differential- und Integralrechnung

(Natalja Deng)

Definition der Ableitung

Gegeben sei eine Funktion $y = f(x)$. Die Ableitung von $f(x)$ ist definiert als

$$f'(x) = \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Geometrisch gibt $f'(x_0)$ die Steigung der Tangente an den Graphen von $f(x)$ im Punkt (x_0, y_0) an ($y_0 = f(x_0)$).

Taylor-Reihen

Gesucht wird eine Möglichkeit, die Funktion $f(x)$ als Potenzreihe darzustellen. Dazu gehen wir rückwärts vor:

Wenn eine Reihendarstellung

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + \dots$$

existiert, dann darf man sie gliedweise differenzieren:

$$f'(x) = a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + \dots; \quad f'(0) = a_1.$$

$$f''(x) = 2a_2 + 2 \cdot 3a_3 x + 3 \cdot 4a_4 x^2 + \dots; \quad f''(0) = 2a_2; \quad a_2 = \frac{1}{2!} f''(0).$$

$$f'''(x) = 2 \cdot 3a_3 + 2 \cdot 3 \cdot 4a_4 x + \dots; \quad f'''(0) = 2 \cdot 3a_3; \quad a_3 = \frac{1}{3!} f'''(0).$$

Somit ist $a_k = \frac{1}{k!} f^{(k)}(0)$. Einsetzen in die Ausgangsreihe ergibt

$$f(x) = f(0) + \frac{f'(0)}{1!} x + \frac{f''(0)}{2!} x^2 + \frac{f'''(0)}{3!} x^3 + \dots = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

($f^{(n)}(x)$ bezeichnet die n-te Ableitung von f an der Stelle x). Manchmal ist es notwendig, nicht $f(x)$ nach Potenzen von x an der Stelle $x = 0$, sondern $f(x + h)$ nach Potenzen von h an der Stelle x zu entwickeln. Man erhält die allgemeine Form der Reihe von Taylor:

$$f(x + h) = f(x) + \frac{f'(x)}{1!} h + \frac{f''(x)}{2!} h^2 + \frac{f'''(x)}{3!} h^3 + \dots = \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!} h^n.$$

Das ist eine unglaublich dreiste Behauptung! Schaut euch zum Beispiel die Taylorreihe in der ersten Form ($f(x) = f(0) + \dots$) mal scharf an. Auf der linken Seite steht $f(x)$, und auf der rechten kommen f und seine Ableitungen nur im Punkt $x = 0$ vor. Das ist höchst merkwürdig. Nehmen wir an, x sei die Zeit, und wir können irgendeine Größe f samt allen ihren Ableitungen zum Zeitpunkt $x = 0$ bestimmen. Dann könnten wir sie für jeden beliebigen Zeitpunkt $x > 0$ in der Zukunft berechnen! Das kann nicht gut sein. Eine Funktion ist eine Vorschrift, die jedem x genau ein $f(x)$ zuordnet – keine weiteren Einschränkungen. Wer eine Funktion definiert, darf ihr für $x = 1$ einen beliebigen Wert geben und ist nicht daran gebunden, was er in einer Umgebung von $x = 0$ festgelegt hat. Aber $f(0)$, $f'(0)$ und so weiter hängen nur davon ab, was sich in einer beliebig kleinen Umgebung von $x = 0$ abspielt. Wie kann man daraus erschließen, was die Funktion bei $x = 1$ macht? Im Allgemeinen gar nicht.

Aber unter welchen Umständen existiert diese Reihe überhaupt? Nach der Taylorschen Formel gilt:

$$f(x + h) = f(x) + \frac{f'(x)}{1!} h + \frac{f''(x)}{2!} h^2 + \frac{f'''(x)}{3!} h^3 + \dots + \frac{f^{(n)}}{n!} h^n + R_n$$

mit $R_n = \frac{f^{(n+1)}(\tilde{x}) h^{(n+1)}}{(n+1)!}$ für ein (nicht näher bekanntes) $\tilde{x}$ zwischen x und $x + h$ (Lagrange'sches Restglied). Die Taylor-Reihe stellt $f(x)$ nun für genau die Werte von x dar, für die $\lim_{n \to \infty} R_n(x) = 0$.

Ach so. Also: Damit diese wundersame Zukunftsvorhersage funktioniert, muss f nicht nur unendlich oft differenzierbar sein (diese höheren Ableitungen müssen alle existieren, und zwar auf dem ganzen Intervall), es muss auch noch das Restglied R_n für $n \to \infty$ gegen 0 gehen. Das Schöne ist: Dieses Wunder findet überraschend häufig statt. Seine Voraussetzungen treffen auf alle Funktionen zu, mit denen man in der Mathematik üblicherweise umgeht, und sie treffen (meistens) auf die Funktionen zu, für die wir uns hier verschärft interessieren: Lösungen von Differentialgleichungen (siehe die nächsten Beiträge). Das ist nicht so verwunderlich: Differentialgleichungen lösen ist dasselbe wie die Zukunft vorhersagen, und das tut eine Taylorreihe auch (wenn sie konvergiert).

Die typische Situation ist folgende: Wir kennen f nicht (f ist gesucht), und wir haben nur Informationen über f und einige seiner Ableitungen in einem Punkt (zum Beispiel $x = 0$). Was können wir daraus für $f(x)$ für $x > 0$ schließen? Solange x klein ist, ist x^n noch viel kleiner; für große n tut das $n!$ im Nenner das Seinige, um das Restglied klein zu machen; wir machen also keinen großen Fehler, wenn wir das Restglied einfach vernachlässigen, und können $f(x)$ aus den restlichen (berechenbaren) Termen ziemlich genau bestimmen – vorausgesetzt, der letzte Bestandteil des Restglieds, $f^{(n+1)}(\tilde{x})$, hält sich in Grenzen. Den kennen wir nämlich meistens nicht, können ihn allenfalls abschätzen.

Also: Die Taylorreihe verschafft uns eine gewisse unscharfe Information über das Verhalten einer Funktion in der Nähe eines Punktes, in dem wir über sie Bescheid wissen. Wie unscharf, wissen wir im Prinzip auch, in der Praxis meistens nicht, weil wir die Funktion nicht kennen (und wenn wir sie kennen würden, müssten wir uns nicht mit der Taylorreihe rumquälen). Diese (unvermeidliche) Unklarheit vererbt sich auf die Aussagen, die wir später mit Hilfe der Taylorreihe herleiten.

Einige Reihenentwicklungen an der Stelle $x = 0$:

$$\sin(x) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \dots$$

$$\cos(x) = \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots$$

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \dots$$

$$\ln(1+x) = \sum_{n=1}^{\infty}(-1)^{n-1}\frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \dots, -1 < x \leq 1$$

$$\arctan(x) = \sum_{n=1}^{\infty}(-1)^{n-1}\frac{x^{2n-1}}{2n-1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \dots, -1 \leq x \leq 1$$

Definition des Integrals

Gegeben sei eine Funktion $f(x)$, die in $x_0 \leq x \leq x_n$ stetig sei. Man unterteilt das Intervall $[x_0, x_n]$ in n Teilintervalle durch die Punkte $x_1, x_2, x_3, \dots, x_{n-1}$. In jedem dieser Teilintervalle sei ein Punkt $\tilde{x}_j$ ausgewählt und $f(\tilde{x}_j)$ sein Funktionswert. Man multipliziert nun jeweils diesen Funktionswert mit der Länge des Teilintervalls $\Delta x_k = x_k - x_{k-1}$ (die nicht konstant sein muss) und bildet die Summe:

$$f(\tilde{x}_1)(x_1 - x_0) + f(\tilde{x}_2)(x_2 - x_1) + f(\tilde{x}_3)(x_3 - x_2) + \dots + f(\tilde{x}_n)(x_n - x_{n-1}) = \sum_{k=1}^{n} f(\tilde{x}_k)\Delta x_k.$$

Geometrisch bedeutet diese Summe eine Annäherung der Fläche unter der Kurve von $f(x)$ durch Rechtecke. Existiert der Grenzwert

$$\lim_{\substack{\Delta x \to 0 \\ n \to \infty}} \sum_{k=1}^{n} f(\tilde{x}_k)\Delta x_k$$

(und hängt er nicht von der Intervallzerlegung ab), so wird er das bestimmte Integral von $f(x)$ zwischen a und b,

$$\int_a^b f(x)dx$$

genannt. Dabei heißt $f(x)$ der Integrand, x Integrationsvariable, $[a, b]$ Integrationsbereich und a und b Integrationsgrenzen. Es stellt sich heraus, dass in der Integralrechnung die umgekehrte Grundaufgabe vorliegt wie in der Differentialrechnung, nämlich das Auffinden einer Funktion $F(x)$, deren Ableitung gleich $f(x)$ ist, einer sogenannten Stammfunktion von $f(x)$. Da die Konstanten beim Differenzieren verschwinden, gibt es unendlich viele verschiedene Stammfunktionen zu einer Funktion $f(x)$, die sich nur um eine Integrationskonstante C unterscheiden. Ihre Graphen gehen durch Verschiebungen in y-Richtung auseinander hervor. Die Menge aller Stammfunktionen zu einer Funktion wird unbestimmtes Integral von f,

$$\int f(x)dx\ ,$$

genannt. Durch eine Anfangsbedingung kann die Integrationskonstante festgelegt und eine bestimmte Funktion, Integralfunktion genannt, herausgegriffen werden:

$$I(x) = \int_a^x f(x)dx = F(x) - F(a).$$

Das bestimmte Integral ist nun ein bestimmter Wert einer solchen Funktion und bedeutet geometrisch die Festlegung des variablen rechten Randes durch einen bestimmten Punkt b.

Einige Grundintegrale:

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C$$

$$\int \frac{dx}{x} = \ln|x| + C$$

$$\int \sin x dx = -\cos x + C$$

$$\int \cos x dx = \sin x + C$$

$$\int e^x dx = e^x + C$$

Zwei Freunde sitzen in der Kneipe und erregen sich über die mathematische Unbildung der Menschen – vor allem Alfred. Bruno hält dagegen, so schlimm sei es doch gar nicht, und gewisse mathematische Kenntnisse seien in der Allgemeinheit durchaus verbreitet. Als Alfred austreten muss, winkt Bruno die Kellnerin herbei: „Ich will meinem Freund einen Streich spielen. Ich werde Sie vor seinen Ohren fragen, wieviel $\int x dx$ ist, und Sie antworten einfach $x^2/2$.“ Kaum kommt Alfred vom Klo zurück, ruft Bruno, um seine Behauptung zu „beweisen“, die Kellnerin herbei, fragt: „Wieviel ist $\int x dx$?“, die Kellnerin antwortet „$x^2/2$“, dem Alfred klappt der Unterkiefer runter, da wendet sich die Kellnerin zum Gehen und sagt über die Schulter weg „plus C“.

Nie die Integrationskonstante vergessen!

Partielle Integration

Die Produktregel der Differentiation lautet

$$\frac{d(uv)}{dx} = \frac{du}{dx}v + u\frac{dv}{dx}$$

oder auch

$$(uv)' = u'v + uv'.$$

Durch Integrieren erhält man

$$uv + C = \int u'v dx + \int uv' dx$$

oder

$$\int uv' dx = uv - \int uv' dx.$$

Diese Formel kann angewendet werden, wenn der Integrand das Produkt zweier Terme ist, von denen der eine leicht integriert und der andere leicht differenziert werden kann. Partielles Integrieren bringt im Allgemeinen nur dann etwas ein, wenn uv' leichter zu integrieren ist als $u'v$. Dabei ist die Wahl von u und v' entscheidend für den Erfolg des Verfahrens.

Numerische Integration

Man erhält bereits einige Näherungsformeln für bestimmte Integrale, wenn man statt des Grenzwerts eine der Größen berechnet, die gegen diesen Grenzwert konvergieren:

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} f(x_k)\Delta x_k \quad \text{(Wert am rechten Rand des Teilintervalls) oder}$$

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} f(x_{k-1})\Delta x_k \quad \text{(Wert am linken Rand) oder}$$

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} f(\frac{x_k + x_{k-1}}{2})\Delta x_k \quad \text{(Wert in der Intervallmitte)}$$

Eine weitere Formel ergibt sich, wenn man statt der Rechtecke Trapeze verwendet:

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} \frac{1}{2}(f(x_k) + f(x_{k-1}))\Delta x_k.$$

Wenn alle Δx_k gleich sind, lässt sich das umformen zu

$$\left(\frac{1}{2}f(a) + \sum_{k=1}^{n-1} f(x_k) + \frac{1}{2}f(b)\right)\Delta x.$$

Oder man approximiert den Integranden nicht durch einen Streckenzug (wie bei der Trapezregel), sondern durch Parabelstücke. Daraus ergibt sich die Simpson-Regel (wieder mit konstanten Δx, n muss gerade sein):

$$\int_a^b f(x)dx \approx \frac{\Delta x}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \ldots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)).$$

Literatur:

Leupold, Conrad, Völkel: Analysis für Ingenieure. Harri Deutsch, Frankfurt a. M. und Zürich.

Ayres: Differential- und Integralrechnung. Schaum's Outline Series/McGraw-Hill Book Company.

Spiegel: Vektoranalysis. Schaum's Outline Series/McGraw-Hill Book Company.

Dallmann, Elster: Einführung in die höhere Mathematik 1. Vieweg, Braunschweig.

Brauch, Dreyer, Haacke: Mathematik für Ingenieure. B. G. Teubner, Stuttgart.

Papula: Mathematik für Ingenieure und Naturwissenschaftler, Band 2. Vieweg, Braunschweig.

Beispiele einfacher Differentialgleichungen aus der Natur

(Christoph Grothaus)

Nach den Übungen im Differenzieren und Integrieren tauchte die Frage auf: Wozu brauchen wir das? Die Antwort ist folgende: In der Natur werden viele Vorgänge durch mehr oder minder komplizierte Differentialgleichungen beschrieben. Dies liegt zu großen Teilen daran, dass ein universelles Gesetz, das **2. Newtonsche Gesetz**, fast überall auftaucht. Dieses Gesetz verbindet eine Kraft F, die auf einen Massenpunkt wirkt, mit seiner Masse m und der 2. Ableitung seines Ortes, der Beschleunigung a. Es lautet $F = m \cdot a$.

1. Bewegungsgleichung

Wie fällt ein Stein, wenn man ihn loslässt? Hier gilt das Newtonsche Gesetz in einer speziellen Version: $F = m{\cdot}g$. Dabei ist g die (in Erdnähe konstante) Erdbeschleunigung ($g \approx 9,81 m/s^2$). Gesucht ist eine Gleichung $x(t)$, die den Ort angibt, an dem sich der fallende Stein zu einer bestimmten Zeit befindet.

Allgemein gilt:

Weg: $x(t)$

Geschwindigkeit: $v = x'(t)$ 1. Ableitung des Weges nach der Zeit

Beschleunigung: $a = x''(t)$ 2. Ableitung des Weges nach der Zeit

Daraus ergibt sich umgekehrt, dass man $x(t)$ durch Integrieren erhält, wenn a bekannt ist.

Da $g[= a]$ als konstant angenommen wird, gilt: $g(t)$ =const. Durch Integrieren nach t ergibt sich: $v(t) = g{\cdot}t{+}C_1$. Nochmaliges Integrieren liefert: $x(t) = \frac{1}{2}g \cdot t^2 + C_1 \cdot t + C_2$.

Somit hat man eine Lösung $x(t)$ gefunden, in der noch die Integrationskonstanten C_1 und C_2 enthalten sind. Die Interpretation dieser Konstanten ergibt sich aus der Physik: Wurde der Stein nicht aus der Ruhe losgelassen, sondern hatte eine bestimmte Anfangsgeschwindigkeit v_0, so kann man aus der Gleichung für $v(t)$ und der Anfangsbedingung $v(0) = v_0$ die Integrationskonstante C_1 ermitteln. Für $t = 0$ liefert die Gleichung für $v(t)$: $v(0) = g \cdot 0 + C_1 = C_1$. Daraus ergibt sich: $C_1 = v_0$. Genauso verfährt man mit C_2 und der Anfangsbedingung $x(0) = x_0$. Es ergibt sich: $C_2 = x_0$. In endgültiger Fassung heißt die Lösung der Differentialgleichung also:

$$x(t) = \frac{1}{2}g \cdot t^2 + v_0 \cdot t + x_0\,.$$

Diese Lösung ist sehr vereinfacht, da sie nur gerade herunterfallende Steine berücksichtigt, also außer Acht lässt, dass x, g und v eigentlich Vektoren $\vec{x}, \vec{g}, \vec{v}$ sind.

Die Lösung war gar nicht schwer, denn in der Gleichung $x'' = g$ kommt die unbekannte Funktion $x(t)$ nur einmal vor, als zweite Ableitung. Dies liegt daran, dass in Erdnähe die Anziehungskraft und damit die Beschleunigung konstant sind. Bei nicht konstanter Kraft gilt für den Weg folgende Differentialgleichung 2. Ordnung, die man so allgemein gar nicht lösen kann:

$$m \cdot x''(t) = F(x(t))$$

2. Radioaktiver Zerfall

Das Ausgangsproblem ist: Es ist eine Menge Q radioaktiven Materials zum Zeitpunkt $t = 0$ gegeben, gesucht ist eine Funktion $Q(t)$, die die zum Zeitpunkt t noch vorhandene Menge Q angibt. Aus physikalischen Beobachtungen und theoretischen Annahmen weiß man, dass die Rate, mit der das radioaktive Material zerfällt, direkt proportional zur Menge des noch vorhandenen Materials ist. Daraus ergibt sich folgende Differentialgleichung 1. Ordnung:

$$\frac{dQ}{dt} = -r \cdot Q$$

Die Proportionalitätskonstante $r \quad (r > 0)$ ist die für jedes radioaktive Material unterschiedliche Zerfallsrate. Diese Differentialgleichung soll nach Q aufgelöst werden. Der folgende Lösungsweg ist mathematisch nicht korrekt, führt jedoch zum richtigen Ergebnis und ist weitaus anschaulicher als der mathematisch korrekte Weg:

$\frac{dQ}{dt} = -r \cdot Q$ Separation der Variablen: alles mit Q auf die eine Seite räumen, alles mit t auf die andere

$\Leftrightarrow \frac{dQ}{Q} = -r \cdot dt$ auf beide Seiten das Integralzeichen anwenden

$\Leftrightarrow \int \frac{1}{Q} dQ = -r \cdot \int dt$

$\Leftrightarrow \ln(|Q|) = -r \cdot t + C$

$\Leftrightarrow Q = e^{-r \cdot t + C}$

$\Leftrightarrow Q = e^{C} \cdot e^{-r \cdot t}$

Zusammen mit der Anfangsbedingung $Q(0) = Q_0$ ergibt sich für den konstanten Term $e^C = Q_0$. Somit lautet die endgültige Form der Lösung

$$Q(t) = Q_0 \cdot e^{-r \cdot t}$$

Etwas schwieriger wird das Lösen der Differentialgleichung unter der Annahme, dass ständig neues radioaktives Material mit der konstanten Rate k (Dimension von k ist Masse / Zeit) zugeführt wird. Diese Annahme ist nicht realitätsfern, ein Beispiel wäre, die Menge radioaktiven Materials im Abwasser eines Kernkraftwerks zu berechnen, nachdem die Anfangsmenge zerfallen ist und nur noch täglich eine geringe Menge hinzukommt. Wir ändern die Differentialgleichung für den normalen Zerfall wie folgt ab:

$$\frac{dQ}{dt} = -r \cdot Q + k$$

Die Zerfallsrate r, die Zuführung k und die Anfangsbedingung $Q(0) = Q_0$ seien bekannt.

Substitution: Sei $k = -r \cdot z$. Dann ist $\frac{dQ}{dt} = -r \cdot Q - r \cdot z \Leftrightarrow \frac{dQ}{dt} = -r \cdot (Q + z)$

Substitution: Sei $p = Q + z$. Dann gilt $\frac{dp}{dt} = \frac{dQ}{dt}$, da beim Differenzieren nach t das konstante Glied z wegfällt. Daraus ergibt sich die umformulierte Differentialgleichung

$$\frac{dp}{dt} = -r \cdot p\,,$$

deren Lösung uns schon vom einfachen radioaktiven Zerfall bekannt ist: $p = e^C \cdot e^{-r \cdot t}$. Resubstitution: $p = Q + z$ $\Rightarrow Q + z = e^C \cdot e^{-r \cdot t} \Leftrightarrow Q = e^C \cdot e^{-r \cdot t} - z$. Resubstitution: $k = -r \cdot z \Leftrightarrow z = -\frac{k}{r}$

$$Q = e^C \cdot e^{-r \cdot t} + \frac{k}{r}$$

Anhand dieser Gleichung, die noch die Integrationskonstante C enthält und deshalb noch unbestimmt ist, lassen sich folgende Aussagen treffen: Der Term $e^C \cdot e^{-r \cdot t}$ konvergiert gegen 0 und ist nach einiger Zeit zu vernachlässigen. Das Niveau, auf das sich die Menge radioaktiven Materials nach einer gewissen Zeit einpendelt, ergibt sich also aus dem Term $\frac{k}{r}$, der Zuleitung pro Zeitschritt geteilt durch die Zerfallsrate. Mit der Anfangsbedingung ergibt sich für den konstanten Term e^C durch Einsetzen: $e^C = Q_0 - \frac{k}{r}$. Setzt man dies in obige Gleichung ein, so erhält man

$$Q = Q_0 \cdot e^{-r \cdot t} - \frac{k}{r}(e^{-r \cdot t} - 1)$$

3. Zinsrechnung

Der Vermehrung eines Grundkapitals durch Zahlung von Zinsen liegt dasselbe Grundprinzip zugrunde wie dem radioaktiven Zerfall: Der Zuwachs des Kapitals ist direkt proportional zum momentanen Kapital. Die Differentialgleichung lautet also $\frac{dS}{dt} = r \cdot S \quad (r > 0)$, der Unterschied zum radioaktiven Zerfall besteht lediglich im Vorzeichen von r. Dementsprechend lautet die Gleichung aufgelöst nach S folgendermaßen: $S = S_0 \cdot e^{r \cdot t}$. In unserer Vorstellung funktioniert dieses Berechnungsmodell einwandfrei, es lassen sich einfach weitere Rückschlüsse ziehen, und der Graph verläuft glatt. In der Realität berechnen die Banken die Zinsen jedoch nicht **kontinuierlich**, sondern in **diskreten** Schritten von einem Jahr, einem Halbjahr, einem Quartal oder sonstigen Zeitschritten. Die Gleichungen für diese Berechnungsweisen sehen so aus:

$$\text{1-mal jährlich} : S(t) = S_0 \cdot (1 + r)^t$$
$$\text{2-mal jährlich} : S(t) = S_0 \cdot (1 + \frac{r}{2})^{2t}$$
$$\text{m-mal jährlich} : S(t) = S_0 \cdot (1 + \frac{r}{m})^{mt}$$
$$(t \text{ ist gemessen in Jahren})$$

Diese Funktionen sind wesentlich unhandlicher als die Exponentialfunktion und haben einen weiteren Nachteil: Sie machen Sprünge. Ihr Graph verläuft für jeweils einen Zeitschritt horizontal und springt dann auf die nächste Stufe. Ihre Werte weichen von denen der kontinuierlichen Funktion ab, dieser Effekt ist jedoch (für kleine Verzinsungszeiträume) vernachlässigbar. Je größer m ist, desto mehr nähert sich das Ergebnis dem genauen Wert an, und es lässt sich zeigen:

$$\lim_{m \to \infty} S_0 \cdot (1 + \frac{r}{m})^{mt} = S_0 \cdot e^{r \cdot t} \quad .$$

An diesem Beispiel wird deutlich, dass ein schönes mathematisches Modell (stetig, differenzierbar) die Realität nur unzureichend widerspiegelt, da die Realität sich nicht immer einfach modellieren lässt.

Bei genauerer physikalischer Betrachtung erkennt man, dass auch die mathematische Beschreibung des radioaktiven Zerfalls ungenau ist. Die Moleküle zerfallen nicht irgendwie kontinuierlich, sondern in einzelnen diskreten Akten. Wenn die Molekülzahl ausreichend klein ist, ist das exponentielle Berechnungsmodell ungenau.

Dieses Problem tritt bei vielen Modellen auf, und zwar häufig in der umgekehrten Weise: Ein kontinuierlicher Vorgang wird diskret modelliert, z. B. bei der Simulation von Bewegungen am Computer, da der Computer nur in diskreten Schritten rechnen kann.

Literatur: Boyce, DiPrima: Gewöhnliche Differentialgleichungen. Spektrum Akademischer Verlag, Heidelberg; S. 52–66

Numerische Lösung von Differentialgleichungen

(Lisa Huber, Christoph Pöppe)

Bisher konnten wir unsere Differentialgleichungen noch analytisch lösen. Die weitaus meisten Probleme sind dafür jedoch zu kompliziert. Deshalb stellen wir im Folgenden einige numerische Verfahren vor.

Wir betrachten ein Anfangswertproblem:

$$y'(t) = \frac{dy(t)}{dt} = f(t, y(t))$$

$$y(t_0) = y_0$$

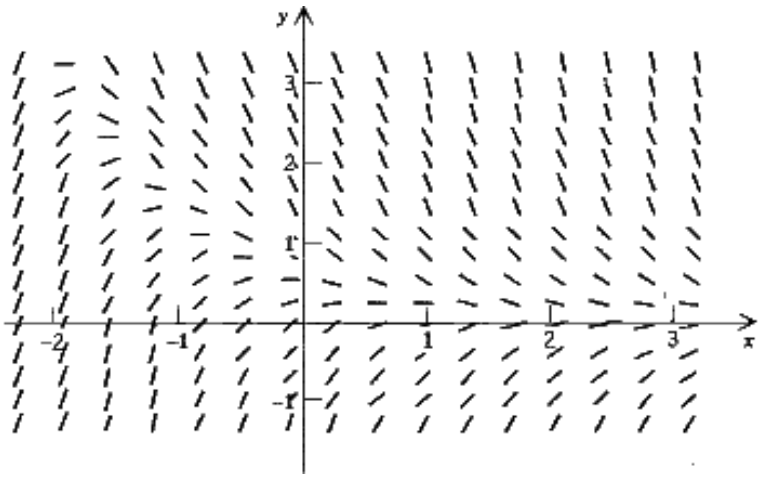
bei dem wir davon ausgehen, dass es eine eindeutige Lösung besitzt (siehe den Beitrag von Karin Heitzmann). Wir nennen die eindeutige Lösung $\Phi(t)$. Mit y_n bezeichnen wir die angenäherten Werte von $\Phi(t_n)$ an den Stellen $t_n = t_0 + nh, h = t_{n+1} - t_n, n = 0, 1, 2, \ldots$.

Hier gilt es höllisch aufzupassen! Diese Differentialgleichung, so wie sie dasteht, ist zunächst mal eine Wunschvorstellung, eine Forderung an das y. Das y haben wir aber erstmal noch nicht. Da steht zwar $y'(t) = f(t, y(t))$, aber deswegen ist f noch lange nicht die Ableitung von y. Das gilt erst, wenn y die Lösung ist, und dann sind wir sowieso fertig. Solange wir die Lösung nicht haben, ist f – na ja, eben gar nix mit y. Es ist die rechte Seite der Differentialgleichung; einen schöneren Namen gibt es nicht dafür. Physikalisch gesprochen: f ist das (bekannte) Naturgesetz, und y ist das (vorläufig unbekannte) Verhalten eines Systems, das diesem

Naturgesetz folgt und zum Zeitpunkt t_0 im Zustand y_0 ist. Um Lösung und Weiß-noch-nicht-ob-Lösung säuberlich auseinanderzuhalten, schreiben wir Φ für ersteres und y für letzteres. Die meisten Lehrbücher machen diese Unterscheidung (und die Praktiker verschlampern sie, weil denen klar ist, was jeweils gemeint ist).

Übrigens muss in der Differentialgleichung das y (bzw. die Lösung Φ) nicht unbedingt eine zahlenwertige Funktion sein. Alle jetzt folgenden Überlegungen funktionieren genauso, wenn die unbekannte Funktion ein Vektor ist, das heißt aus mehreren Komponenten besteht. Die rechte Seite f bildet dann einen Vektor y samt einem Skalar t auf einen Vektor $f(t, y)$ ab. Physikalisch bedeutet das, dass der Zustand des Systems nicht nur durch eine Zahlengröße beschrieben wird, sondern durch mehrere – was der bei weitem interessantere Fall ist. Diese mehreren Größen – die Komponenten von y – können zum Beispiel die Koordinaten eines oder mehrerer Massenpunkte sein, die durch Kräfte aufeinander einwirken. Anders ausgedrückt: Man hat mehrere unbekannte Funktionen, die auch noch voneinander abhängig sind: ein System von Differentialgleichungen. Eine Gleichung höherer Ordnung kann auf eine einfache Weise in ein System von Gleichungen 1. Ordnung umgewandelt werden. Das angegebene Anfangswertproblem ist also keineswegs so speziell, wie es aussieht, sondern bereits so ziemlich das allgemeinste, was es gibt. Für die bildliche Veranschaulichung, die jetzt kommt, muss man sich allerdings das y doch wieder als einen Skalar vorstellen.

An einem geeigneten Richtungsfeld (hier abgebildet ist das Richtungsfeld für $y' = e^{-t} - 2y$) können wir uns die Vielfalt der Lösungen veranschaulichen und die Problematik der Näherung betrachten.



Die rechte Seite f der Differentialgleichung sagt uns zu jedem Punkt (t, y) in der (t, y)-Ebene, welche Steigung $y'(t)$ die Lösungskurve haben müsste (nämlich $f(t, y)$), wenn sie durch den Punkt (t, y) verliefe. (Ob sie wirklich durch diesen Punkt verläuft, wissen wir noch nicht.) Anders ausgedrückt: Wir suchen eine Kurve $y(t)$, die in jedem ihrer Punkte die für diesen Punkt vorgeschriebene Tangente $f(t, y(t))$ hat.

Durch jeden Punkt der (t, y)-Ebene verläuft genau eine solche Lösungskurve, vorausgesetzt, die rechte Seite f ist nicht zu exotisch. Das sagt uns der Existenz- und Eindeutigkeitssatz (siehe unten). Lösungskurven können sich nicht schneiden; sonst gäbe es im Schnittpunkt zwei verschiedene Steigungen. Es gibt aber nur einen Wert von $f(t, y)$. Das hindert nicht, dass Lösungskurven auseinander- oder zusammenlaufen. Aus der großen Schar der Lösungskurven (der möglichen Verhaltensweisen des Systems) suchen wir diejenige, die durch den Punkt (t_0, y_0) verläuft (zum Zeitpunkt t_0 im Zustand y_0 ist).

Es ist ein bisschen wie Autofahren auf einer unendlich vielspurigen Autobahn: Immer schön auf der Spur bleiben, das heißt die Richtung halten, die an der Stelle, wo du gerade bist, angesagt ist. Wenn du nicht aufpasst, machst du einen unwillkürlichen Spurwechsel. Auf einer echten Autobahn kracht es dann meistens. Auf einem Richtungsfeld ist es – na ja, eben ein Fehler.

Somit stellt sich uns vor allem die Aufgabe, einen solchen Fehler möglichst klein zu halten.

Das einfache Eulersche Verfahren

Das erste numerische Verfahren wurde von Euler entwickelt. Der erste Schritt besteht darin, dass wir mit unserem bekannten, exakten Wert y_0 durch Einsetzen in die rechte Seite der Differentialgleichung die Steigung in diesem Punkt bekommen. Mit deren Hilfe konstruieren wir die Tangente an die Kurve im Punkt $y_0 = \Phi(t_0)$. Entlang dieser Tangente $f(t_0, y_0)$ bewegen wir uns um eine konstante Schrittweite h weiter bis zu dem Punkt y_1. Das soll ein Näherungswert für $\Phi(t_1)$ sein: $y_1 = y_0 + hf(t_0, y_0)$. Jetzt setzen wir wieder den errechneten Wert in die rechte Seite der Differentialgleichung ein, bekommen einen Steigungswert, konstruieren die Tangente und erlangen den nächsten Punkt y_2. Allgemein können wir schreiben:

$$y_{n+1} = y_n + hf(t_n, y_n)$$

mit $n = 0, 1, 2, 3, \ldots$. Je kleiner hierbei die Schrittweite, desto geringer ist die Abweichung der Näherung von der exakten Lösung, desto höher allerdings der Rechenaufwand.

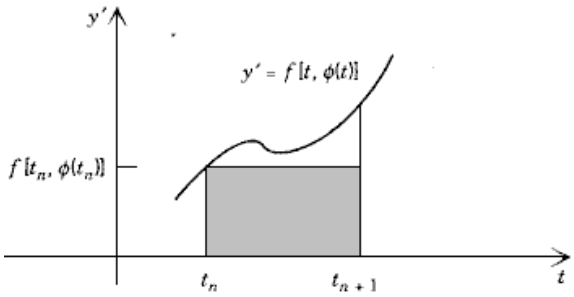
Eulerverfahren ist wie ein total übermüdeter Autofahrer: Fährt los, nickt ein, schreckt hoch, stellt das Steuer richtig für die Spur, auf der er gerade ist, pennt wieder eine Sekunde, schreckt hoch, und so weiter. Der hält

natürlich um so besser die Spur, je öfter er die Augen aufmacht. Einmal Augen aufmachen ist in diesem Bild aber soviel wie ein kompletter Rechenschritt auf dem Computer – egal wieviel Zeit zwischen zwei „Augenblicken“ verstreicht. Deswegen ist für das Rechnen mit dem Computer unpraktikabel, was auf der Autobahn eine gute Idee ist: immer die Augen aufzuhalten.

Es ist auch möglich, das Eulersche Verfahren mathematisch über das Integral über $\Phi'(t)$ herzuleiten:

$$\Phi(t_{n+1}) - \Phi(t_n) = \int_{t_n}^{t_{n+1}} \Phi'(t)dt = \int_{t_n}^{t_{n+1}} f[t, \Phi(t)]dt$$

$$\Rightarrow \Phi(t_{n+1}) = \Phi(t_n) + \int_{t_n}^{t_{n+1}} f[t, \Phi(t)]dt$$



Beim Euler-Verfahren ersetzen wir den Integranden durch seinen Wert an der Stelle $t = t_n$, eine Konstante, und können ihn somit aus dem Integral herausziehen und über die Konstante 1 integrieren:

$$\Phi(t_{n+1}) \approx \Phi(t_n) + f[t_n, \Phi(t_n)](t_{n+1} - t_n)$$

Ersetzt man jetzt $\Phi(t_n)$ durch unsere Näherungswerte y_n, so ergibt sich wiederum das Euler-Verfahren.

Auch mit der Taylor-Reihe um den Punkt t_n (ausgewertet in $t_n + h = t_{n+1}$), von der wir natürlich ausgehen, dass sie existiert, lässt sich das Eulersche Verfahren finden.

$$\Phi(t_n + h) = \Phi(t_n) + \Phi'(t_n)h + \Phi''(\tilde{t_n})h^2$$

$$\Leftrightarrow \Phi(t_{n+1}) = \Phi(t_n) + f[t_n, \Phi(t_n)]h + \Phi''(\tilde{t_n})h^2$$

($\tilde{t_n}$ ist wieder eine unbekannte Zwischenstelle zwischen t_n und t_{n+1}). Wenn wir nämlich nach dem 2. Glied abbrechen, d. h. das Restglied nicht berücksichtigen, und für $\Phi(t_{n+1}), \Phi(t_n)$ wieder die Näherungswerte y_{n+1}, y_n verwenden, so ergibt sich hieraus ebenfalls das Eulersche Verfahren. Uns muss jedoch bewusst sein, dass dieses Verfahren Fehler macht, die wir im Folgenden betrachten wollen.

Fehlerabschätzung:

Nennen wir den globalen Approximationsfehler E_n. Das ist der Unterschied zwischen der exakten Lösung und der angenäherten zum Zeitpunkt t_n: $E_n = \Phi(t_n) - y_n$.

Jetzt kommt die Sache mit der Sünde, die euch so erheitert hat. Und zwar: Die exakte Lösungskurve ist der Pfad der Tugend. Aber den kennt man nur im Himmel. Auf Erden sind wir allzumal Sünder, das heißt wir weichen vom Pfad der Tugend ab. Das ist unvermeidlich, wenn man in diskreten Zeitschritten rechnet, und anders können wir es nicht – wenn wir die Lösung nicht auf analytischem Wege finden. Beim Euler-Verfahren geht man halt ein Stück die Tangente entlang statt die Lösungskurve; das ist die Sünde, die man in jedem Schritt aufs Neue begeht. Allgemein verwendet man bei jedem Schritt nur eine Näherungsformel zur Lösungsbestimmung, denn eine exakte Formel gibt es nicht. Außerdem war man aber schon – abgesehen vom allerersten Schritt, als man sich noch im Stande der Unschuld befand – auf der falschen Lösungskurve. Das ist die Folge der früheren Fehltritte: die Erbsünde eben.

Versuchen wir, die Gesamtsünde E_n in Neusünde und Erbsünde zu zerlegen. Dazu bilden wir die Taylor-Reihe von $\Phi(t)$ um t_n mit Restglied 2. Ordnung:

$$\Phi(t_n + h) = \Phi(t_n) + \Phi'(t_n)h + \Phi''(\tilde{t_n})h^2$$

mit einem gewissen $\tilde{t_n}$ zwischen t_n und $t_n + h$. Das formen wir wie vorher um und subtrahieren die Formel des Eulerschen Verfahrens. Für die Gesamtsünde zur Zeit t_{n+1} ergibt sich

$$E_{n+1} = \Phi(t_{n+1}) - y_{n+1} = \Phi(t_n) - y_n + h\{f[t_n, \Phi(t_n)] - f[t_n, y_n]\} + \Phi''(\tilde{t_n})h^2$$

An dieser ganzen Sünde ist neu nur der letzte Term; denn wenn wir im n-ten Schritt noch auf dem Pfad der Tugend wären, wäre $y_n = \Phi(t_n)$, und die beiden ersten Summanden würden wegfallen. Also ist die Neusünde – offizieller Name: der lokale Diskretisierungsfehler –

$$e_{n+1} = \Phi(t_{n+1}) - y_{n+1} = \frac{1}{2}\Phi''(\tilde{t_n})h^2,$$

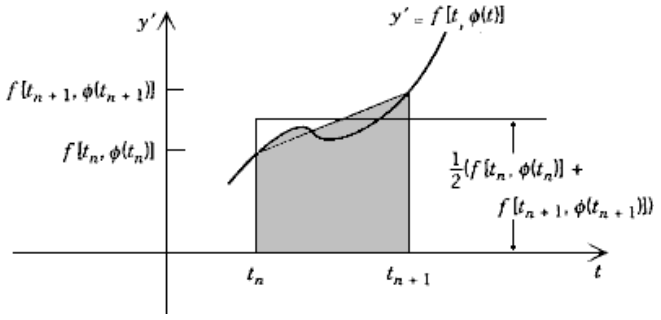
das ist proportional zu h^2 und der 2. Ableitung von $\Phi(t)$.

Was haben wir davon? Φ kennen wir nicht (sonst hätten wir die Lösung und könnten uns das Sündigen sparen) und $\tilde{t}_n$ schon gar nicht. Wenn wir die Sünde nun schon nicht ausrechnen können, wollen wir sie wenigstens in Grenzen halten. Das gelingt auch häufig, denn meistens ist $\Phi''(t)$ beschränkt: $|\Phi''(t_n)| \leq M$ mit einem gewissen M, das wir zwar meistens auch nicht kennen, das aber wenigstens nicht von h abhängt. Dann ist auch der lokale Diskretisierungsfehler begrenzt: $|e_{n+1}| \leq \frac{Mh^2}{2}$.

Also: Wenn wir die Schrittweite halbieren, begehen wir nur noch ein Viertel der Neusünde – pro Schritt. Um zum selben Ziel zu gelangen, müssen wir jetzt aber doppelt so viele Schritte machen. Allgemein: Wenn wir n Schritte benötigen, um von t_0 zu $\bar{t} = t_0 + nh$ zu gelangen, und bei jedem Schritt der Fehler höchstens $Mh^2/2$ beträgt, ist der Fehler nach n Schritten höchstens $nMh^2/2$. Da $n = (\bar{t} - t_0)/h$ ist, erhalten wir daraus als Schranke für die Summe aller Neusünden $(\bar{t} - t_0)Mh/2$. Der Fehler wird also nicht größer als eine Konstante mal h, d. h. er lässt sich durch Übergang zu kleineren Schrittweiten beliebig klein machen.

Diese Argumentation unterstellt, die Fehler jedes einzelnen Schrittes würden sich einfach aufaddieren, berücksichtigt also nicht die Fehlerfortpflanzungseffekte. (Wie würde man die nennen? Erbs-Erbsünde in Analogie zu Zinseszins?) Durch eine etwas kompliziertere Argumentation lässt sich jedoch zeigen, dass der Fehler E_n tatsächlich abschätzbar ist durch h mal eine gewisse Konstante; nur ist die Konstante etwas größer als die oben angegebene $(\bar{t} - t_0)M/2$.

Das verbesserte Eulersche Verfahren (Heun-Formel)



Beim einfachen Eulerschen Verfahren wird ein Integral durch eine Rechtecksfläche angenähert (siehe oben). Die Höhe des Rechtecks ist dabei der Funktionswert $f[t_n, \Phi(t_n)]$ am linken Rand des zu untersuchenden Intervalls. Das verbesserte Eulersche Verfahren approximiert den Integranden genauer, und zwar durch eine Trapezfläche, indem man den Mittelwert zwischen den Eckpunktwerten $\left(f[t_n, \Phi(t_n)] + f[t_{n+1}, \Phi(t_{n+1})]\right)/2$ nimmt. Wieder ersetzen wir die exakten durch die Näherungswerte: $\Phi(t_{n+1}), \Phi(t_n)$ durch y_{n+1}, y_n. Daraufhin erhalten wir $y_{n+1} = y_n + (h/2)\left(f[t_n, y_n] + f[t_{n+1}, y_{n+1}]\right)$. Da die Unbekannte y_{n+1} jedoch in der rechten Seite der Gleichung vorkommt, ersetzen wir sie hier durch den Wert den wir mittels der Euler-Formel erhalten. Folglich gilt:

$$y_{n+1} = y_n + \frac{h}{2}\left(f[t_n, y_n] + f[t_n + h, y_n + hf(t_n, y_n)]\right)$$

Das ist das Euler-Verfahren mit Beichte! Wir machen zuerst einen Fehltritt, das heißt, wir berechnen einen vorläufigen Wert für y_{n+1} nach dem Euler-Verfahren. Im Lichte der Erfahrung, die wir dadurch gewonnen haben, bringen wir eine Korrektur an und machen dadurch unseren Fehler wenigstens zum Teil wieder gut. Allgemein heißen Verfahren dieses Typs (deren es viele gibt) Prädiktor-Korrektor-Verfahren.

Es lässt sich beweisen, dass der lokale Diskretisierungsfehler bei diesem Verfahren durch eine Konstante mal h^3 und der globale auf einem endlichem Intervall durch eine Konstante mal h^2 beschränkt ist, d. h. die Fehlergrenzen werden enger. Allerdings erfordert eine ausreichend komplizierte Funktion f einen beträchtlichen Rechenaufwand. Für das Heun-Verfahren müssen wir f doppelt so oft auswerten wie für das einfache Verfahren; oder: Das Heun-Verfahren ist so aufwendig wie ein Euler-Verfahren mit der halben Schrittweite. Es bringt allerdings mehr an Genauigkeit.

Überhaupt sollte kein schiefes Bild entstehen! Das Euler-Verfahren ist zwar das einzige, das wir uns etwas ausführlicher ankucken; aber verglichen mit anderen, tatsächlich eingesetzten Verfahren ist es so elend schlecht, dass es nur als abschreckendes Beispiel taugt. Eigentlich hat es den Namen des großen Leonhard Euler (eines der größten Mathematiker überhaupt) gar nicht verdient. Es ist nur nützlich, weil es so einfach zu verstehen ist und trotzdem bereits die wesentlichen Eigenschaften aller Verfahren für Anfangswertprobleme zeigt.

Eine wesentliche Eigenschaft ist: Wenn es um die Qualität eines Verfahrens geht, kommt es auf das genaue Ausmaß der Sünde gar nicht besonders an. Man macht sich in der Regel auch nicht die Mühe, die Konstante M (bzw. die etwas größere, die eigentlich korrekt ist) auszurechnen, selbst wenn man die Daten zur Verfügung hat. Das einzige, was bei der ganzen Fehlerabschätzerei am Ende interessiert, ist der Exponent an dem h. Das ist die Ordnung des Verfahrens. Das Euler-Verfahren hat Ordnung 1, das Heun-Verfahren Ordnung 2. Eine Halbierung der Schrittweite vermindert den zu befürchtenden Fehler bei einem Verfahren 1. Ordnung auf die Hälfte, bei 2. Ordnung auf ein Viertel, bei 3. Ordnung auf ein Achtel. Je höher die Ordnung des Verfahrens,

desto besser das Ergebnis (desto höher im Allgemeinen auch der Aufwand).

Es ist eine gute Idee, dasselbe Problem mit zwei verschiedenen Schrittweiten zu rechnen. Der Unterschied in den Ergebnissen gibt einem – ohne dass man theoretischen Aufwand treiben müsste – einen Hinweis darauf, wie groß das M in diesem Fall ist. Daraus kann man wiederum ausrechnen, wie groß man das h wählen muss, damit der globale Fehler unter einer vorgegebenen Schranke bleibt. Und das ist die Standard-Forderung an jedes gute Verfahren.

Immer nur mit entsetzlich kleiner Schrittweite h die Zeitachse lang zu tippeln bringt nichts – nur Rechenaufwand: Durch die guten Werke allein wirst du nicht selig. Die Kunst besteht darin, die Schritte so groß zu nehmen, wie es der globalen Genauigkeitsforderung gerade noch vereinbar ist. Das kann zu verschiedenen Zeiten verschieden sein: Wenn das System in heftiger Bewegung ist (scharfe Kurve auf der gedachten Autobahn), muss man sehr häufig hinkucken, damit man alle Einzelheiten mitkriegt. Auf langen Geraden, wo nahezu nichts passiert, genügt ab und zu ein verschlafenes Blinzeln. Hochentwickelte Verfahren beherrschen die Kunst, die Schrittweite den Verhältnissen automatisch anzupassen.

Wie sehen andere Verfahren aus? Ein paar Stichworte müssen genügen. Man kann die Taylor-Entwicklung weiter treiben als nur bis zum zweiten Glied, muss dann aber auch Werte aus früheren Zeitschritten mitverwenden, weil sonst keine eindeutigen Formeln zustandekommen. Das läuft darauf hinaus, dass man die Lösungskurve nicht durch die Tangente (die in diesem Punkt anschmiegsamste Gerade) nähert, sondern durch die anschmiegsamste Parabel, kubische Parabel und so weiter. Das sind die Mehrschrittverfahren höherer Ordnung. Oder man treibt das Spiel aus Sünde und Beichte mehrmals hintereinander über gewisse Teil-Zeitschritte. Das läuft auf die sogenannten Runge-Kutta-Verfahren hinaus. Oder man rechnet von t_0 bis $\bar{t}$ mit verschiedenen Schrittweiten h_1, h_2, $h_3 \ldots$, bekommt verschiedene Werte für $\Phi(\bar{t})$ in Abhängigkeit vom jeweiligen Wert von h und extrapoliert daraus einen Wert für die (himmlische) Schrittweite $h = 0$. Das sind die Extrapolationsverfahren. Alle diese Verfahren kann man mit automatischer Schrittweitensteuerung ausstatten; das funktioniert am elegantesten bei den Extrapolationsverfahren.

Literatur: W. E. Boyce/R. C. DiPrima: Gewöhnliche Differentialgleichungen

Numerische Lösung nach dem Eulerverfahren am Beispiel *dy(t)/dt=ay(t)*

(Florian Conrad)

Ein Test-Beispiel für das Euler-Verfahren sollte zum Zwecke des Vergleichs so einfach wie möglich sein und eine bekannte Lösung haben. Deshalb verwende ich hier die DGL $y'(t) = ay(t)$ mit der Anfangsbedingung $y(0) = y_0$, deren Lösung $\Phi(t) = y_0 \cdot e^{at}$ ist.

Beim Euler-Verfahren wird die t-Achse in diskrete Zeitschritte $t_0, t_1, \ldots$ eingeteilt, die voneinander den gleichen Abstand h haben. Es sei $y_k = y(t_k)$ für $k = 1, 2, 3, \ldots$, wobei $y(t)$ die angenäherte Lösung sei, die aber wiederum nur an den Stellen t_k definiert ist. Wir finden y_1, indem wir in (t_0, y_0) die Tangente mit der Steigung $f(t_0, y_0) = ay_0$ anlegen. y_1 wird auf den Wert dieser Tangente an der Stelle t_1 festgelegt. Um nun weitere y_{k+1} zu finden, wird immer wieder das jeweils bekannte y_k in die rechte Seite der DGL eingesetzt. Das heißt, man legt an die Lösungskurve durch (t_k, y_k) (die schon lange nicht mehr die richtige sein muss) die Tangente und läuft an ihr entlang bis t_{k+1}. Der dort erreichte Wert gilt als Näherung y_{k+1} für $\Phi(t_{k+1})$. Dieses Verfahren heißt explizites Euler-Verfahren (siehe den vorigen Beitrag).

Die Lösung an einer festen Stelle $T > 0$ kann man mit verschiedenen Schrittweiten annähern: Man setze $h = T/n$, wobei n die Anzahl der t-Schritte ist, die man bei der Annäherung macht.

Für die angenäherte Lösung in diesem Beispiel gilt also $y_{k+1} = y_k + hay_k = (1+ah)y_k$. Daraus folgt per Induktion $y_k = y_0 \cdot (1+ah)^k$ (für $k \in N$). Es ist $y(T) = y_n = y_0 \cdot (1 + aT/n)^n$. Es ist $\lim_{n\to\infty} y_0 \cdot (1 + aT/n)^n = y_0 \cdot e^{aT}$, d. h. die Annäherung wird um so genauer, je kleiner man die Schrittweite wählt. Unangenehme Nebenwirkungen hat hier (vor allem bei $a < 0$) die Wahl einer Schrittweite h mit $ah \leq -1$, da die genäherte Lösung in einem solchen Fall entweder einfach nur 0 ist oder gar alterniert.

Eine weitere Möglichkeit der Annäherung bietet das implizite Euler-Verfahren, bei dem, um y_{k+1} zu berechnen, nicht y_k sondern y_{k+1} in die rechte Seite der DGL (zwecks Tangentenermittlung) eingesetzt wird. (Dazu muss man eine Gleichung für y_{k+1} lösen. Das ist in diesem Beispiel einfach, im Allgemeinen aber zu kompliziert, um den Aufwand zu rechtfertigen.)

Es ist dann $y_{k+1} = y_k + h(ay_{k+1}) \Rightarrow y_{k+1} = y_k/(1 - ah)$. Mittels Induktion ergibt sich $y_k = y_0/(1 - ah)^k = y_0/(1 - aT/n)^k$, wobei T und n wie oben definiert seien. Auch hier geht die genäherte Lösung an der Stelle T für $n \to \infty$ (d. h. $h \to 0$) wieder gegen die Exponentialfunktion, es ist also eine beliebig gute Annäherung

durch eine entsprechend kleine Schrittweite möglich:

$$\lim_{n\to\infty} y_n = \frac{y_0}{e^{-aT}} = y_0 \cdot e^{aT}$$

Einem großen Problem steht man beim impliziten Euler-Verfahren für $a > 0$ gegenüber, wenn man eine Schrittweite h wählt, für die $ah = 1$ (Division durch 0!) oder $ah > 1$ (Alternieren des Bruchs im Nenner!).

Wegen all dieser Probleme sei im Folgenden $|ah| < 1$. Außerdem beschränke ich mich auf $ah \neq 0$, was kaum schadet, da sowieso $h > 0$ und bei $a = 0$ die DGL eine Funktion Φ mit konstantem Wert für alle t ergäbe.

Durch Ausprobieren erkennt man (jedenfalls beim Beispiel der Exponentialfunktion): Die eine genäherte Lösung liegt immer oberhalb, die andere immer unterhalb des Traumergebnisses. Entscheidend ist dabei der Faktor, um den sich bei dem jeweiligen Verfahren bei jedem Schritt der Wert verändert. Ist dieser kleiner (größer) als e^{ah}, welches der Faktor ist, um den sich die Exponentialfunktion in einem Schritt verändert, verhält sich die Näherung ebenso (oder bei negativem y_0 umgekehrt) zur e-Funktion. Es soll nun die besagte Eigenschaft der jeweiligen Faktoren bewiesen werden.

Zu zeigen ist $1 + ah \leq e^{ah}$ und $1/(1 - ah) \geq e^{ah}$.

Sei $u \in R$ mit $u \neq 0$ und $|u| < 1$.

Die Taylor-Entwicklung von e^u um 0 ist $e^u = 1 + u + u^2 \cdot (e^{\tilde{u}})''/2$ mit einem $\tilde{u}$ zwischen 0 und u. Auf $\tilde{u}$ kommt es nicht so genau an, denn die zweite Ableitung $d^2e^{cx}/dx^2 = c^2e^{cx}$ der Exponentialfunktion e^{cx} ist ohnehin positiv.

Es folgt $1 + u \leq 1 + u + u^2 \cdot (e^{\tilde{u}})''/2 = e^u$

Für $u = ah$ ergibt sich $1 + ah \leq e^{ah}$, für $u = -ah$ gilt $1 - ah \leq e^{-ah}$, woraus nach Umkehrung, die wegen $|ah| < 1$ problemlos durchführbar ist, $1/(1 - ah) \geq e^{ah}$ folgt. qed.

Angesichts dieser Tatsache wäre es vielleicht besser, gleich ein Verfahren zu verwenden, das zwischen explizitem und implizitem Euler-Verfahren liegt. *(Das ist die Idee des Heun-Verfahrens – siehe den vorigen Beitrag –; nur macht man beim Heun-Verfahren das Spiel zwischen Sünde und Beichte, weil man im Allgemeinen eben nicht nach y_{k+1} auflösen kann.)* Dazu verwendet man für die Steigung des Geradenstückes, das zur Ermittlung von y_{k+1} gebraucht wird, den Wert, der zwischen der rechten Seite der DGL für y_k und der für y_{k+1} liegt:

$$y_{k+1} = y_k + h \cdot \frac{1}{2}(ay_k + ay_{k+1}) \iff y_{k+1}(1 - \frac{1}{2}ah) = y_k(1 + \frac{1}{2}ah) \iff y_{k+1} = y_k\frac{1 + \frac{1}{2}ah}{1 - \frac{1}{2}ah}$$

Die Ungleichungen $1 + ah \leq \frac{1+\frac{1}{2}ah}{1-\frac{1}{2}ah}$ und $\frac{1}{1-ah} \geq \frac{1+\frac{1}{2}ah}{1-\frac{1}{2}ah}$ sind erfüllt (problemlos beweisbar). Dieses Verfahren liefert also zumindest bei diesem Beispiel Werte, die zwischen denen des expliziten und denen des impliziten Euler-Verfahrens liegen. Dadurch ist auch der Nachweis der Konvergenz gegen e^{aT} bei $h \to 0$ geliefert.

Existenz und Eindeutigkeit von Lösungen

(Karin Heitzmann)

1. Existenz

Gegeben ist ein Anfangswertproblem bestehend aus einer Differentialgleichung erster Ordnung in der Form $\vec{y}' = \vec{f}(t, \vec{y})$ und einer Anfangsbedingung $\vec{y}(0) = y_0$.

(Wir arbeiten hier mit Vektoren, da mit ihrer Hilfe Systeme beliebiger Ordnung als Differentialgleichung erster Ordnung ausgedrückt werden können.)

Wir setzen voraus, dass die rechte Seite $\vec{f}$der Differentialgleichung stetig ist: „Natura non facit saltus (die Natur macht keine Sprünge)“.

Es kann sein, dass die Lösung eines Anfangswertproblems nicht ewig lebt. Sie kann nach endlicher Zeit „explodieren“, d. h. gegen $\pm\infty$ streben.

Beispiel (aber nur skalar): $y' = y^2 \qquad y(0) = a$

Es gibt zwei Wege, um die richtige Lösung zu erhalten:

a) das „Schmuddelverfahren“ der Physiker: b) die analytisch korrekte Methode:

$$\begin{aligned} \frac{dy}{dt} &= y^2 \\ \frac{dy}{y^2} &= dt \\ y^{-2} \cdot dy &= dt \end{aligned} \qquad \begin{aligned} y' &= y^2 \\ \frac{y'}{y^2} &= 1 \\ [-y^{-1}]' &= 1 \end{aligned}$$

$$y = -\frac{1}{t + C}$$

Durch Einsetzen der Anfangsbedingung erhalten wir für die Integrationskonstante den Wert $C = -\frac{1}{a}$ und bekommen somit die korrekte Lösung $y(t) = \frac{a}{1-at}$ für unser Anfangswertproblem. Diese Lösung existiert jedoch nur für $t < \frac{1}{a}$, da die rechte Seite der Differentialgleichung mit dem geforderten Anfangswert nur in diesem Bereich stetig ist: Wenn y (für $t \to 1/a$) gegen unendlich geht, tut die rechte Seite der Differentialgleichung das auch.

Aber: Wenn eine Lösung $\vec{y}$ vorzeitig stirbt, dann nur durch Explosion. Es kann nicht sein, dass $\vec{y}(t)$ zu einer endlichen Zeit einfach verendet, d. h. zwar beschränkt bleibt, aber nicht mehr fortsetzbar ist. Vielmehr ist $\vec{y}$ fortsetzbar bis zum Rand des Definitionsbereichs von $\vec{f}$. (Der Definitionsberetch von $\vec{f}$ ist normalerweise $R \times R^n$, d. h. $\vec{f}$ ist für alle Werte von t und $\vec{y}$ definiert.) Das heißt üblicherweise für $t \to \infty$, denn bei $t \to \infty$ hört der Definitionsbereich auf. Er hört aber auch für $y \to \infty$ auf, und das ist der Grund, warum die Lösung im Beispiel vorzeitig verstirbt.

Eindeutigkeit

Meistens weiß man noch mehr: Eine Lösung existiert nicht nur, sie ist auch eindeutig bestimmt.

Existenz- und Eindeutigkeitssatz:
Sei das Anfangswertproblem $\vec{y}' = \vec{f}(t, \vec{y}), \vec{y}(t_0) = \vec{y}_0$ gegeben. $\vec{f}$ sei lokal lipschitzstetig im zweiten Argument ($\vec{y}$). Dann existiert eine eindeutige Lösung des Anfangswertproblems (d. h. jedem t wird genau ein $\vec{y}$ zugeordnet) bis zum Rand des Definitionsbereiches.

$\vec{f}$ heißt lipschitzstetig im zweiten Argument, wenn es eine Konstante L gibt, so dass gilt: $|\vec{f}(t, \vec{y}_1) - \vec{f}(t, \vec{y}_2)| \leq L|\vec{y}_1 - \vec{y}_2|$ oder auch: $\frac{|\vec{f}(t,\vec{y}_1) - \vec{f}(t,\vec{y}_2)|}{|\vec{y}_1 - \vec{y}_2|} \leq L$ für alle $\vec{y}_1, \vec{y}_2$ aus einem beliebigen Intervall.

D. h. wenn f nach y differenzierbar ist, und die Ableitung ist beschränkt, dann ist f lipschitzstetig.

Quelle:

Deuflhard/Bornemann: Numerische Mathematik II (De Gruyter); Kapitel 2.1 und 2.2

Zum Beweis des Existenz- und Eindeutigkeitssatzes

(Christoph Pöppe)

Dieser unscheinbare Satz ist von ungeheurer philosophischer Bedeutung. $\vec{y}$ ist der Zustand eines physikalischen Systems, $\vec{f}$ ist das Naturgesetz, das die Veränderung dieses Zustands mit der Zeit beschreibt. Dann sagt der Satz: Wenn zu einem bestimmten Zeitpunkt t_0 der Zustand des Systems bestimmt ist und die Gesetze bekannt sind, die seine zeitliche Entwicklung bestimmen, dann ist der Zustand des Systems für alle Zeiten bestimmt (determiniert). Das System kann sehr wohl die ganze Welt sein, die Naturgesetze kennt man einigermaßen vollständig (jedenfalls waren die Wissenschaftsgläubigen des 19. Jahrhunderts davon überzeugt), und sie sind so ziemlich alle lipschitzstetig; also: Wenn ich den Zustand der Welt zu einem bestimmten Zeitpunkt kenne, kann ich ihn (ausreichende Rechenfähigkeiten vorausgesetzt) für alle Zukunft vorhersagen. (Für alle Vergangenheit übrigens auch; aber das beeindruckt einen nicht so.)

Aus diesem Determinismus läßt sich zweierlei herleiten. Erstens eine ungeheure Arroganz: Allein durch geeignetes Setzen der Anfangsbedingungen können wir das Verhalten eines Systems für alle Zeiten nach unseren Wünschen bestimmen. Jedes überhaupt konstruierbare System ist wie ein Uhrwerk: Wir müssen es nur in Gang setzen, und es läuft für alle Zeiten so, wie wir wollen (und es vorher ausgerechnet haben). Wir sind die Beherrscher der Materie! Andererseits ein ungeheurer Fatalismus: Die ganze Welt ist nichts weiter als ein solches Uhrwerk. Sie läuft einfach ab, und wenn wir glauben, dass wir mit unserem Willen etwas an ihrem Ablauf ändern könnten, so ist das pure Illusion. Wir handeln stets nur so, wie es uns von Anfang an vorherbestimmt ist.

Beide Vorstellungen haben vor allem im 19. Jahrhundert eine sehr große Rolle gespielt. Das hat inzwischen mächtig nachgelassen. Es ist klar geworden, dass die Voraussetzungen des Eindeutigkeitssatzes eigentlich nie erfüllt sind. Wir kennen den Zustand der Welt nie genau, sondern immer nur ungefähr. Also ist die Vorhersage auch nicht genau, sondern ebenfalls nur ungefähr, und zwar häufig viel schwammiger als die Anfangswerte; selbst wenn wir die Anfangswerte ziemlich genau kennen, kann die Qualität der Vorhersage beliebig schlecht sein, wie man bereits am Wetterbericht sieht. Man braucht noch nicht einmal die Quantenmechanik samt ihrer Unbestimmtheitsrelation aus der Kiste zu holen; klassisches (deterministisches) Chaos reicht bereits aus, dem klassischen Determinismus die Zähne zu ziehen. Aber das ist eine ganz andere Geschichte.

Gleichwohl: In (wichtigen) idealisierten Situationen entfaltet der Existenz- und Eindeutigkeitssatz seinen vollen Glanz. Wegen seiner großen Bedeutung und weil es ein richtig schönes Stück Mathematik ist, möchte ich euch seinen Beweis vorführen.

Bei diskreter statt kontinuierlicher Zeit wäre der Eindeutigkeitsbeweis übrigens trivial. Dann gäbe es nur diskrete Zeitpunkte t_0, t_1, $t_2 \ldots$, und das Naturgesetz wäre von der Form $\vec{y}(t_{n+1}) = \vec{f}(t_n, \vec{y}(t_n))$. Wenn man $\vec{y}(t_0)$ kennt und will $\vec{y}(t_N)$ wissen – nichts einfacher als das! Wende einfach N-mal $\vec{f}$ auf $\vec{y}(t_0)$ an. Da kann nichts schiefgehen; Funktionen sind immer eindeutig.

Aber jetzt zum Beweis. Der erste Schritt: In dem Anfangswertproblem $\vec{y}' = \vec{f}(t, \vec{y}), \vec{y}(t_0) = y_0$ integriere ich beide Seiten von t_0 bis zu einem gewissen t und bringe $\vec{y}(t_0)$ auf die rechte Seite:

$$\vec{y}(t) = \vec{y}(t_0) + \int_{t_0}^{t} f(s, \vec{y}(s))ds = \vec{y}_0 + \int_{t_0}^{t} f(s, \vec{y}(s))ds$$

Das bringt mich einer Lösung noch nicht wesentlich näher, denn die Unbekannte $\vec{y}$ steht nicht nur auf der linken Seite, sondern auch noch rechts unterm Integral. Aber man könnte ja mit einer (beliebig bescheuerten) Näherung für $\vec{y}(t)$ anfangen, nennen wir sie $\vec{y}_1(t)$. Die setze ich rechts ein und kriege durch Anwenden der Formel eine neue Näherung $\vec{y}_2(t) = \vec{y}_0 + \int_{t_0}^{t} f(s, \vec{y}_1(s))ds$. Allgemein so weiter (Rekursionsformel):

$$\vec{y}_{k+1}(t) = \vec{y}_0 + \int_{t_0}^{t} f(s, \vec{y}_k(s))ds$$

Dadurch kriege ich eine Folge von Funktionen. Wenn die konvergiert, habe ich gewonnen. Denn dann definiere ich mir $\vec{y}(t) = \lim_{k\to\infty} y_k(t)$, bilde auf beiden Seiten der Rekursionsgleichung den Grenzwert $k \to \infty$ und erhalte $\vec{y}(t) = \vec{y}_0 + \int_{t_0}^{t} f(s, \vec{y}(s))ds$, was zu beweisen war. Nur: Wie kommt die Folge der Funktionen dazu, zu konvergieren? Die gute Nachricht ist: Sie tut es, vielleicht nur für t in der Nähe von t_0, aber das macht nichts. Auf diese Weise komme ich vielleicht von t_0 bis zu einem gewissen t_1, dann habe ich halt in t_1 wieder ein Anfangswertproblem, durch Anwenden desselben Satzes hangele ich mich von t_1 weiter bis zu einem t_2 und so weiter bis zum Rand des Definitionsbereiches – vorausgesetzt, die Hangeldistanzen werden nicht beliebig klein.

Wie beweist man, dass sie konvergiert? Mit dem Banachschen Fixpunktsatz. Der fällt jetzt, wie in der Mathematik häufig üblich, erstmal vom Himmel, und später sehen wir erst, wofür er gut ist. Kleinen Moment Geduld bitte!

Banachscher Fixpunktsatz

Sei $F : X \to X$ eine kontrahierende Abbildung von einem vollständigen normierten Raum X in sich. Dann hat F genau einen Fixpunkt, das heißt, es gibt genau ein $x \in X$ mit der Eigenschaft $F(x) = x$.

Aus was für Elementen dieser Raum X besteht, wollen wir im Moment gar nicht so genau wissen. Nennen wir sie Punkte und stellen uns von mir aus Punkte im gewöhnlichen anschaulichen Raum vor. Unter diesen Punkten muss nur eine sehr zarte Andeutung von Geometrie gelten: Man muss in sinnvoller Weise vom Abstand zweier Punkte reden können. Weil wir's so allgemein gar nicht brauchen, gehen wir gleich davon aus, dass man die Punkte unseres Raumes addieren und mit einer Zahl multiplizieren kann wie die (Orts-) Vektoren des gewöhnlichen Raums (man sagt: X ist ein Vektorraum) und dass es eine Norm gibt. Das ist die Verallgemeinerung der Länge eines gewöhnlichen Vektors. Man schreibt $\|x\|$ für die Norm von x. Der Abstand von x und y ist $\|x - y\|$. Genau besehen ist eine Norm auch nur eine Funktion von X nach R^+ (die Norm von x ist stets ≥ 0) mit einigen naheliegenden Eigenschaften: $\|x\| = 0$ genau dann, wenn $x = 0$, $\|ax\| = |a|\|x\|$ für alle Zahlen a und $\|x + y\| \leq \|x\| + \|y\|$ (die Dreiecksungleichung). Wir haben die Freiheit, Normen nach unseren Bedürfnissen zu definieren, und werden von dieser Freiheit Gebrauch machen.

Man sagt, x sei der Grenzwert einer Folge x_n von Punkten aus X, wenn $\lim_{n\to\infty} \|x_n - x\| = 0$ ist. Der Raum X heißt vollständig, wenn in ihm jede Cauchyfolge (das heißt jede Folge, deren Glieder ab einer gewissen

Nummer beliebig kleinen Abstand voneinander haben) konvergiert. Den Unterschied zwischen vollständig und nicht vollständig sieht man schon bei den Zahlen: Die rationalen Zahlen sind kein vollständiger Raum, denn es gibt Folgen rationaler Zahlen, die eigentlich konvergieren, aber eben nur gegen eine irrationale Zahl. Deswegen muss man ja die rationalen Zahlen zu den reellen Zahlen vervollständigen.

Was heißt schließlich kontrahierend? $F : X \to X$ ist kontrahierend, wenn für alle $x, y \in X$ gilt $\|F(x) - F(y)\| \leq \alpha\|x - y\|$ mit einer Konstante $\alpha < 1$. Die Bilder der Abbildung F (sprich: $F(x)$ und $F(y)$) liegen also mindestens um den Faktor α näher beieinander als die Urbilder x und y. F verkleinert Abstände!

Nach dieser länglichen Begriffserklärung ist der Beweis des Banachschen Fixpunktsatzes überraschend einfach: Wähle einen beliebigen Punkt $x_1 \in X$ und bilde die Folge $x_2 = F(x_1)$, $x_3 = F(x_2)$, allgemein $x_{k+1} = F(x_k)$. Dann ist die Folge x_k eine Cauchyfolge, denn die Abstände aufeinanderfolgender Glieder gehen (mindestens) so schnell gegen 0 wie die Glieder einer geometrischen Folge mit dem Quotienten α: Wegen der Kontraktionseigenschaft ist $\|F(x_{k+1}) - F(x_k)\| \leq \alpha\|F(x_k) - F(x_{k-1})\| \leq \alpha^2\|F(x_{k-1}) - F(x_{k-2})\| \ldots \leq \alpha^{k-1}\|F(x_2) - F(x_1)\|$. Also hat die Folge x_k einen Grenzwert. Nennen wir ihn x, gehen in der Rekursionsgleichung $x_{k+1} = F(x_k)$ zum Grenzwert über und erhalten $x = F(x)$, was zu beweisen war.

Der Punkt x ist auch eindeutig bestimmt. Nehmen wir an, es gäbe zwei Fixpunkte x und y, also $F(x) = x$ und $F(y) = y$. Dann ist $\|F(x)-F(y)\| \leq \alpha\|x-y\|$ wegen der Kontraktionseigenschaft, zugleich aber $\|F(x)-F(y)\| = \|x - y\|$ wegen der Fixpunkteigenschaft. Da $\alpha < 1$ ist, kann das nur sein, wenn $\|x - y\| = 0$, also $x = y$ ist. Fertig!

Dieser Beweis ist nur deswegen so einfach, weil er so abstrakt ist. Wir denken nur in Punkten und stellen uns darunter innerlich Punkte auf einer Geraden vor oder allenfalls Punkte im anschaulichen (dreidimensionalen) Raum, also im Wesentlichen reelle Zahlen oder Vektoren aus drei (vier, fünf, endlich vielen) Zahlen. Seine richtige Schlagkraft entfaltet der Satz aber erst, wenn wir für die Punkte kompliziertere Dinge einsetzen – komplette Funktionen. Von diesem Prinzip, eine Funktion als einen Punkt aufzufassen, lebt ein ganzes Fachgebiet der Mathematik, die Funktionalanalysis.

Es bleibt jetzt, das ursprüngliche Problem, den Existenz- und Eindeutigkeitssatz, geeignet zurechtzulegen. Ein gezielter Schlag mit dem Hammer des Banachschen Fixpunktsatzes, und wumm! unser Problem löst sich in Wohlgefallen auf.

Damit die Voraussetzungen des Fixpunktsatzes erfüllt sind, brauchen wir eine kontrahierende Abbildung F, die auf einem vollständigen normierten Raum definiert ist. Unsere Abbildung wirkt auf Funktionen und ist wie folgt definiert:

$$F(\vec{y})(t) = \vec{y}_0 + \int_0^t f(s, \vec{y}(s))ds$$

Was ist das für eine merkwürdige Schreibweise mit den zwei Klammern hintereinander? Gewöhnungsbedürftig, aber vollkommen korrekt. Eine Funktion (oder Abbildung; die beiden Wörter bedeuten dasselbe) ist eine Vorschrift, die zu jedem Element aus einer Menge A sagt, welches Element einer Menge B daraus werden soll. Normalerweise schreibt man das als $f(x) = \ldots$ (irgendein Ausdruck, der x enthält); dabei ist x ein unbestimmtes Element aus A und $f(x)$ das Element aus B, das f aus x macht. Bei der Funktion $\vec{y}$ ist das noch ganz einfach: A ist das Intervall von t_0 bis t_1: $A = [t_0, t_1]$; für den (End-)Zeitpunkt t_1 sind noch Bedingungen einzuhalten, deren Einzelheiten ich euch hier erspare. B ist ein Raum von Vektoren, und wenn man $\vec{y}$ definieren will, schreibt man eine Formel für $\vec{y}(t)$ auf, wobei man für t ein beliebiges Element aus A einsetzen darf. (Aufpassen: $\vec{y}$ ist eine Funktion, $\vec{y}(t)$ ist eine Zahl. Die muss man auseinanderhalten.)

Aber jetzt das F! Die Abbildung F ist definiert auf einem Raum A stetiger (vektorwertiger) Funktionen (das ist ein anderes A als gerade eben!), sie bildet ab in denselben Raum ($B = A$), und man definiert sie, indem man sagt, was F aus einem beliebigen Element $\vec{y} \in A$ macht. Na ja, ein Element $F(\vec{y}) \in A$. Aber $F(\vec{y})$ ist seinerseits eine Funktion; um die zu definieren, muss man sagen, was sie aus einem beliebigen $t \in R$ macht. Also muss man ansagen, was $F(\vec{y})(t)$ sein soll, für beliebige $\vec{y}$ und t. Bitte sehr ...

Wie ist das mit der Normierung? Für Räume stetiger Funktionen gibt es so etwas wie eine Standardnorm, die für die meisten Zwecke gerade die richtige ist. Man nimmt die größte Abweichung von der Null: $\|\vec{y}\| := \sup_{t \in R} |\vec{y}(t)|$ (Statt $\sup$ wie „Supremum“ darf man sich meistens „Maximum“ denken. Ersparen wir uns den feinen Unterschied.) Das passt ganz gut zu unserer Vorstellung: Der Unterschied zwischen zwei stetigen Funktionen ($\|\vec{y} - \vec{z}\|$) wird gemessen an der Stelle, wo sie am meisten voneinander abweichen. Wenn eine Folge von Funktionen $\vec{y}_n$ gegen eine Funktion $\vec{y}$ konvergieren soll, dann muss (ab einem gewissen n_0, das übliche Konvergenzritual) der Abstand $|\vec{y}_n(t) - \vec{y}(t)|$ für alle t kleiner sein als ϵ. (Dann und nur dann ist nämlich das Supremum der Abstände kleiner als ϵ.)

Für unsere Zwecke brauchen wir eine speziell angepasste Norm. Wenn nämlich eine vorgebliche Lösung unseres Anfangswertproblems nicht ganz genau stimmt, dann macht sich das in der Nähe unseres Anfangs-

zeitpunkts t_0 nicht so besonders bemerkbar, aber später dafür umso doller. Das ist unvermeidlich: Schon wenn man am Anfangswert ein bisschen wackelt, kann das Verhalten des Systems für große Zeiten völlig anders aussehen. Kleine Ursachen haben (möglicherweise) große Wirkungen – aber erst nach einer Weile! Für kleine Zeiten gilt das Prinzip der stetigen Abhängigkeit von den Daten, und das heißt insbesondere: Kleine Ursachen haben kleine Wirkungen. Damit man in dem großen Kuddelmuddel überhaupt noch etwas sehen kann, braucht man eine kurzsichtige Norm, das heißt eine, die für große Zeiten weniger genau hinkuckt als für kleine. Hier ist sie:

$$\|\vec{y}\| := \sup_{t_0<t<t_1} e^{-L(t-t_0)}|\vec{y}(t)|$$

Wir dämpfen also die Kapriolen, die $\vec{y}$ für große t schlagen könnte, mit dem Faktor $e^{-L(t-t_0)}$. Dabei ist L die Lipschitzkonstante von f.

Und bezüglich dieser Norm soll F eine Kontraktion sein? Rechnen wir's aus. Wir müssen zeigen, dass $\|F(\vec{y}) - F(\vec{z})\| \leq \alpha\|\vec{y}-\vec{z}\|$ mit einem noch zu bestimmenden $\alpha < 1$. Dazu schauen wir erst nach, wie wir $F(\vec{y})(t)-F(\vec{z})(t)$ abschätzen können:

$$\begin{aligned}|F(\vec{y})(t) - F(\vec{z})(t)| &= |\int_{t_0}^{t}(f(s,\vec{y}(s)) - f(s,\vec{z}(s)))ds| \\ &\leq \int_{t_0}^{t}|f(s,\vec{y}(s)) - f(s,\vec{z}(s))|ds\end{aligned}$$

(wenn man den Betrag ins Integral reinzieht, wird es höchstens größer)

$$\leq \int_{t_0}^{t} L|\vec{y}(s) - \vec{z}(s)|ds \quad \text{wegen der Lipschitzstetigkeit}$$

so, jetzt so einen Faktor reinfummeln, damit man die Sache mit unserer Exotennorm schreiben kann

$$= \int_{t_0}^{t} Le^{L(s-t_0)}\left\{e^{-L(s-t_0)}|\vec{y}(s) - \vec{z}(s)|\right\}ds$$

wenn man das Maximum über alle s von dem nimmt, was in der geschweiften Klammer steht, kommt die Norm von $\vec{y} - \vec{z}$ raus. Die ziehen wir ausm Integral raus, denn sie hängt nicht mehr von s ab:

$$\begin{aligned}&\leq \|\vec{y} - \vec{z}\|\int_{t_0}^{t} Le^{L(s-t_0)}ds \\ &= \|\vec{y} - \vec{z}\|(e^{L(t-t_0)} - 1)\end{aligned}$$

Nach dieser Vorarbeit ist der Rest nicht mehr schwer:

$$\begin{aligned}\|F(\vec{y}) - F(\vec{z})\| &= \sup_{t_0<t<t_1} e^{-L(t-t_0)}|F(\vec{y})(t) - F(\vec{z})(t)| \\ &\leq \|\vec{y} - \vec{z}\| \sup_{t_0<t<t_1} e^{-L(t-t_0)}(e^{L(t-t_0)} - 1) \\ &= \|\vec{y} - \vec{z}\| \sup_{t_0<t<t_1} (1 - e^{-L(t-t_0)}) \\ &= \|\vec{y} - \vec{z}\|(1 - e^{-L(t_1-t_0)})\end{aligned}$$

und fertig! F ist bezüglich der Exotennorm kontrahierend mit dem Faktor $\alpha = (1-e^{-L(t_1-t_0)})$, und der ist kleiner als 1.

Eigentlich müsste man jetzt noch beweisen, dass F einen Funktionenraum in sich abbildet, und die Vollständigkeit: dass ein Grenzwert stetiger Funktionen bezüglich der kurzsichtigen Norm wieder eine stetige Funktion ist. Diese Dreckarbeit muss zwar sein, aber wir ersparen sie uns hier.

Der Beweis ist übrigens konstruktiv (na ja, was die Mathematiker so konstruktiv nennen): Im Prinzip kann man durch das Iterationsverfahren, dessen Konvergenz den Beweis liefert, Anfangswertprobleme lösen. Im Prinzip heißt: wenn man das Integral in jedem Iterationsschritt explizit berechnen kann und wenn man den Grenzwert der Funktionenfolge, die sich dabei ergibt, rauskriegt, dann hat man die Lösung des Anfangswertproblems. Wer möchte: $y' = ay$, $y(0) = y_0$ ist ein gutes Übungsbeispiel. Wir wissen auch schon, was rauskommt. Als Startpunkt der Iteration empfiehlt sich $y_1(t) = y_0$ oder etwas ähnlich Einfaches. Viel Spaß!

Lösung von gewöhnlichen Differentialgleichungen am Computer

(Sebastian Tivig)

Das Ziel: Zu unserem Leidwesen gibt es immer noch viele Differentialgleichungen, die zwar für uns einen sehr großen Wert haben, aber nicht durch eine geschlossene Formel lösbar sind. Man kann die Lösung nur annähern. Dafür ist es aber sehr wichtig, dass wir elektronische Hilfe haben, sprich den Computer. Bei den Millionen von Rechnungen, die anfallen, um beispielsweise das Räuber-Beute-Modell (s. u.) zu simulieren, brauchen wir so nicht mehr ein Leben lang per Hand zu rechnen, sondern erhalten die Ergebnisse in wenigen Sekunden. Selbst komplizierte Modelle lassen sich so noch relativ leicht simulieren. Sehen wir uns nun einige dieser Modelle genauer an.

Das Räuber-Beute-Modell: Zwei Populationen leben alleine in einem geschlossenen, isolierten Ökosystem. Dabei ernährt sich die Population y (sagen wir die Füchse) ausschließlich von der Population x (sagen wir den Hasen). Wir wollen nun wissen, wie viele Hasen und wie viele Füchse in unserem System zu einem beliebigen Zeitpunkt t leben. Dabei gehen wir von folgenden drei Annahmen aus: Erstens vermehren sich die Hasen proportional zu ihrer eigenen Anzahl, wenn es keine Füchse gibt. Zweitens sterben die Füchse aus, wenn es keine Hasen gibt. Und drittens ist die Wahrscheinlichkeit, dass ein Fuchs einen Hasen trifft, um so größer, je mehr es von diesen Tieren gibt, also proportional zur Größe der beiden Populationen x und y. In Formeln gefasst bedeutet dies für die Variation der Populationsgrößen pro Zeitschritt:

$$\frac{dx}{dt} = ax - \alpha xy$$

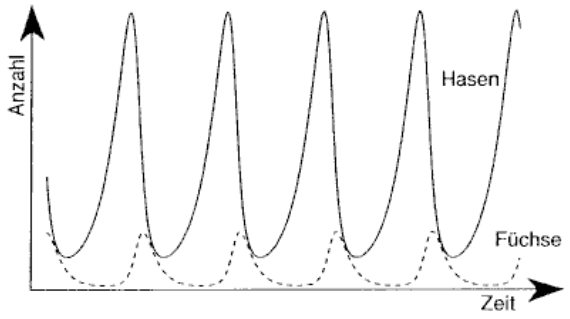
$$\frac{dy}{dt} = -cy + \gamma xy$$

Dabei bezeichnen wir hier mit a und c die Wachstumsraten der beiden Populationen, mit α die Wahrscheinlichkeit, dass bei einem Treffen von Füchsen und Hasen der Fuchs auch tatsächlich den Hasen frißt, und mit γ dieselbe Wahrscheinlichkeit mal der Anzahl an Füchsen, die ein Hase ernähren kann. Der Term ax entspricht dem natürlichen Zuwachs an Hasen, αxy ist die Menge der von Füchsen gefressenen Hasen. $-cy$ ist die Abnahmerate (Geburtenrate minus Todesrate) der Füchse, wenn es keine Hasen gäbe, und γxy steht für die Anzahl der Füchse, die Hasen fressen können und dadurch überleben. Dieses Gleichungssystem ist nicht explizit nach der Zeit lösbar, sondern muß angenähert werden. Nehmen wir dafür das Eulerverfahren, so erhalten wir:

$$x(t + \Delta t) = x(t) + \Delta t(ax(t) - \alpha x(t)y(t))$$

$$y(t + \Delta t) = y(t) + \Delta t(-cy(t) + \gamma x(t)y(t))$$

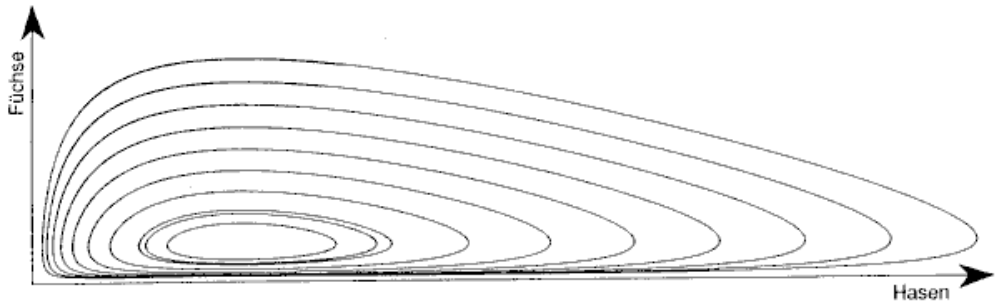
(Statt Δt haben die Vortragenden bisher h geschrieben.) Dieses Modell wurde auch schon angewandt und in der Realität bestätigt. Ein Beispiel für das Resultat, das uns ein Computer gibt, sehen wir in nebenstehendem Bild, wo die Größe der beiden Populationen gegen die Zeit abgetragen ist.



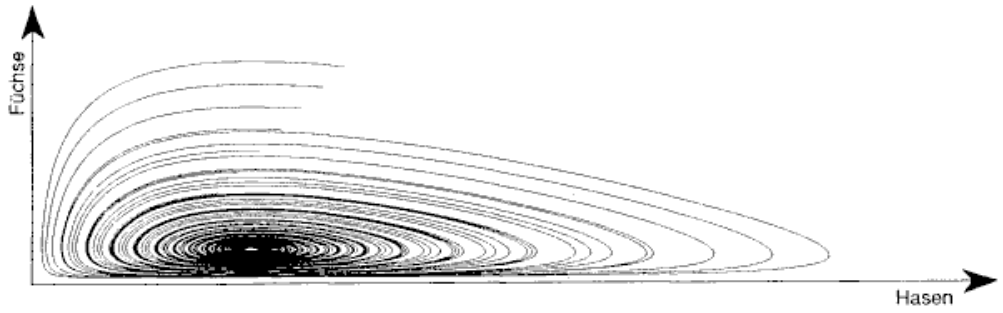
Jedoch ist die Annahme, dass die Hasenpopulation proportional zu ihrer Größe wächst, unrealistisch. Es würde sich ein exponentielles Wachstum ergeben, so dass die Größe der Population irgendwann explodieren würde: Ohne Füchse würden die Hasenanzahlen über jede Grenze wachsen. Um deshalb näher an die Realität zu kommen, gehen wir lieber davon aus, dass sich die Populationen gemäß den Gesetzen des logistischen Wachstum vermehren. Das heißt, dass wir eine obere Populationsgrenze einführen, die Tragekapazität (*carrying capacity*) unseres Systems. Die Populationen werden sich nun asymptotisch dieser Grenze nähern. Wir können so Faktoren wie z. B. eine begrenzte Weidekapazität für die Hasen oder einen begrenzten Lebensraum für die Füchse simulieren. Modifizieren wir nun unser Räuber-Beute-Modell entsprechend, so erhalten wir ein neues Gleichungssystem:

$$\frac{dx}{dt} = a(1 - \frac{x}{K})x - \alpha xy$$

$$\frac{dy}{dt} = -cy + \gamma xy - \mu y^2$$



Dabei ist K hier die oben genannte Tragfähigkeit des Ökosystems für die Hasen, das heißt die Populationsgröße die sich in Abwesenheit der Füchse als Gleichgewicht einstellen würde. In dieser Situation hätten wir $x = K$, dadurch würde dx/dt gleich null und $x(t)$ konstant werden. Mit μ bezeichnen wir hier die Wahrscheinlichkeit, dass es bei einem Treffen von zwei Füchsen zu einem Revierkampf kommt, bei dem einer der Füchse getötet wird. Was für Ergebnisse wir aus diesen beiden Varianten bekommen, sieht man am besten in den beiden Phasendiagrammen (oben und nächste Seite), wo wir die Hasenpopulation auf der x-Achse gegen die Fuchspopulation auf der y-Achse abgetragen haben. Gezeigt werden auf den Diagrammen zehn verschiedene



Kurven, bei denen nur die Startwerte variieren, und zwar um je 10 Einheiten auf den beiden Achsen. Das einfache Modell (siehe oben) ist periodisch und kreist um den Gleichgewichtspunkt, erreicht diesen aber nie. Diese Periode kann man berechnen, sie ist $\frac{2\pi}{\sqrt{ac}}$. In dem erweiterten Modell (oben) trifft dies nicht mehr zu. Alle Kurven streben dem Gleichgewichtspunkt zu, erreichen diesen aber nicht mehr in endlicher Zeit.

Konkurrierende Spezies: Es gibt noch ein anderes Modell aus der Biologie, das einen gewissen Reiz hat. Nehmen wir wieder an, dass wir ein geschlossenes, isoliertes Ökosystem haben, in dem nur zwei Populationen x und y (Karnickel und Hasen) leben. Diesmal aber gehen wir davon aus, dass sie sich um denselben Futtervorrat streiten. Die oberste Vermehrungsgrenze ist hier bestimmt durch die Menge an Nahrung, die das Ökosystem liefern kann. Die beiden Populationen vermehren sich also nach den Gesetzen des logistischen Wachstums. Gehen wir einmal davon aus, dass die Hasen $s_x x(t)$ Nahrung brauchen und die Karnickel $s_y y(t)$, wobei $x(t)$ die Anzahl der Hasen und $y(t)$ die Anzahl der Karnickel zum Zeitpunkt t ist. Der gesamte Nahrungsverbrauch ist also $N(t) = s_x x(t) + s_y y(t)$. Er variiert mit der Anzahl der Hasen und der Karnickel. Daraus erhalten wir dann das Gleichungssystem:

$$\frac{dx}{dt} = a(1 - \frac{N(t)}{K_x})x(t)$$

$$\frac{dy}{dt} = b(1 - \frac{N(t)}{K_y})y(t)$$

Hier sind a und b die Wachstumsraten der beiden Populationen, K_x und K_y sind ihre Bevölkerungsgrenzen. Setzen wir nun N in das obige System ein, so erhalten wir:

$$\frac{dx}{dt} = a(1 - \frac{s_x x(t) + s_y y(t)}{K_x})x(t)$$

$$\frac{dy}{dt} = b(1 - \frac{s_x x(t) + s_y y(t)}{K_y})y(t)$$

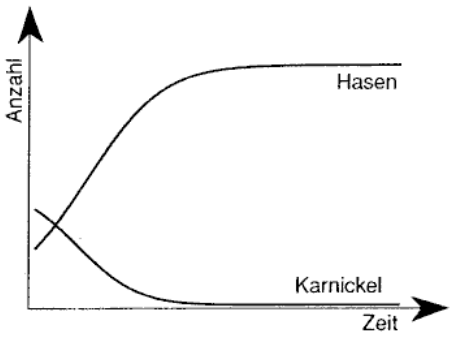
Wollen wir nun dieses System mit dem Computer simulieren, so müssen wir zum Beispiel nur noch das Eulerverfahren einsetzen und finden

$$x(t+\Delta t) = x(t) + a\Delta t(1 - \frac{s_x x(t) + s_y y(t)}{K_x})x(t)$$

und

$$y(t+\Delta t) = y(t) + b\Delta t(1 - \frac{s_x x(t) + s_y y(t)}{K_y})y(t).$$

Im Bild rechts sehen wir die Größe der beiden Populationen gegen die Zeit aufgetragen. Man sieht, dass die eine Population ausstirbt und sich die andere asymptotisch der maximalen Bevölkerungszahl nähert. Diese Tatsache, dass immer nur eine der beiden Populationen überleben kann, ist als Exklusionsprinzip von Volterra bekannt.



Das Modellieren: Verglichen mit obigen einfachen Modellen sind echte Ökosysteme hoffnungslos kompliziert. Allenfalls auf einsamen Inseln kann man noch ein solches Verhalten annähernd finden. Trotz allem sind die theoretisch beschriebenen Effekte auch in komplizierten Situationen wirksam, wie das Räuber-Beute-Modell gezeigt hat. Die einfache Variante davon wurde experimentell bestätigt. Das zeigt, dass wir durchaus berechtigt sind zu vereinfachen, man muß nur wissen, wo man dies machen kann, ohne die Realitätsnähe des Modells zu beeinträchtigen. Auch kann es passieren, dass ein Modell, das sehr gut erscheint, sich als ungenauer erweist als eine Vereinfachung desselben. Die beiden Varianten des Räuber-Beute-Modells sind ein gutes Beispiel dafür.

Funktionen von mehreren Veränderlichen

(Natalja Deng)

Unter einer Funktion von zwei unabhängigen Veränderlichen versteht man eine Vorschrift, die jedem geordneten Zahlenpaar (x, y) aus einer Menge D genau ein Element aus einer Menge W zuordnet:

$$f : (x, y) \mapsto z$$

oder

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

Analog gelangt man zu Funktionen von n unabhängigen Veränderlichen:

$$y = f(x_1, x_2, x_3, \ldots, x_n)$$

Eine Funktion von zwei unabhängigen Veränderlichen lässt sich als Fläche im 3-dimensionalen Raum deuten; der Funktionswert besitzt dann die Bedeutung einer Höhenkoordinate und der Definitionsbereich die eines Gebietes in der (x, y)-Ebene. Eine andere Darstellung ist das Höhenliniendiagramm. Hierbei werden alle „Linien gleicher Höhe“ (das heißt gleichen f-Wertes) in die (x, y)-Ebene projiziert.

Partielle Ableitungen

(Die folgenden Beziehungen gelten entsprechend für Funktionen von mehr als zwei Veränderlichen; nur die geometrische Veranschaulichung entfällt.) Unter der ersten partiellen Ableitung einer Funktion $f(x, y) = z$ versteht man den Grenzwert:

$$f_x(x, y) = \frac{\partial z}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} \quad .$$

Die partielle Ableitung nach y wird entsprechend definiert. Geometrisch bedeuten die partiellen Ableitungen den Anstieg der Flächentangente in der jeweiligen Koordinatenrichtung. Man kann sich dabei vorstellen, dass eine Ebene $y = \text{const.}$ mit der Funktionsfläche zum Schnitt gebracht wird. Mit dieser Veranschaulichung im Einklang steht die Technik des partiellen Differenzierens: In der Funktionsgleichung werden alle Variablen bis auf die, nach der differenziert wird, als konstant angesehen. Dann werden die bekannten Ableitungsregeln für „gewöhnliche“ Funktionen angewendet. Die partiellen Differentialoperatoren $\frac{\partial}{\partial x}$ und $\frac{\partial}{\partial y}$ erzeugen aus einer Funktion durch ihr „Einwirken“ die partiellen Ableitungen:

$$\frac{\partial}{\partial x}[f(x, y)] = f_x(x, y); \frac{\partial}{\partial y}[f(x, y)] = f_y(x, y) \quad .$$

Werden die partiellen Ableitungen erster Ordnung, die ja wieder Funktionen von x und y sind, wiederum partiell differenziert, so entstehen höhere partielle Ableitungen, z. B.

$$\frac{\partial}{\partial x}(\frac{\partial f}{\partial x}(x, y)) = \frac{\partial^2 f}{\partial x^2}(x, y) = f_{xx}(x, y); \frac{\partial}{\partial y}(\frac{\partial f}{\partial x})(x, y) = f_{xy}(x, y); \frac{\partial}{\partial x}(\frac{\partial f^2}{\partial y^2})(x, y) = f_{yyx}(x, y) \quad .$$

Dabei darf die Reihenfolge der gemischten partiellen Ableitungen in den meisten Fällen vertauscht werden, nämlich genau dann, wenn die Funktion und ihre partiellen Ableitungen stetig sind.

An die Stelle der Kurventangente bei Funktionen einer Veränderlichen tritt bei Funktionen mehrerer Veränderlicher die Tangentialebene in einem Punkt (x, y). Sie berührt die Funktionsfläche in diesem Punkt und enthält sämtliche in diesem Punkt an die Fläche angelegten Tangenten. Der Zuwachs der Höhenkoordinate entlang der Tangentialebene wird beschrieben durch das vollständige oder totale Differential

$$dz = \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy \ ,$$

das, wie bei gewöhnlichen Funktionen, bei kleinem Zuwachs in den unabhängigen Veränderlichen eine gute Näherung für Δz, die Änderung des Funktionswertes, darstellt.

Ableitung zusammengesetzter Funktionen

Wenn in $z = f(x, y)$ die Variablen x und y selbst Funktionen von Parametern r und s sind, so dass $z = f(x(r, s), y(r, s))$, dann gilt

$$\frac{\partial z}{\partial r} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial r}$$

und

$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial s} \quad .$$

Hängen speziell x und y nur von einem Parameter r ab, so gilt

$$\frac{\partial z}{\partial r} = \frac{\partial z}{\partial x}\frac{dx}{dr} + \frac{\partial z}{\partial y}\frac{dy}{dr} \quad .$$

Entsprechendes gilt für Funktionen mit mehr Veränderlichen und/oder mehr Parametern.

Gerade d's schreibt man, wenn es – in dem jeweiligen Kontext – nur eine einzige unabhängige Variable gibt; krumme sind in dem Falle nicht falsch. Aber krumme muss man schreiben, wenn es mehr als eine unabhängige Variable gibt. Das ist eine Vorsichtsmaßnahme gegen die Schlampigkeit der Physiker! Wenn die sowas wie $\frac{dz}{dx}\frac{dx}{dt}$ sehen, dann kürzen die hastdunichtgesehen mit dx. Den Mathematiker schaudert's, aber in dem Fall kommt sogar das Richtige raus (Kettenregel). In Fällen wie dem obigen dagegen wäre das Kürzen genau falsch. Das krumme Schwänzchen vom ∂ ist also ein Widerhaken, der das ∂ daran hindern soll, aus Versehen durch den Bruchstrich zu flutschen.

Vektoren

Vektoren werden auf unterschiedliche Arten dargestellt. Wenn A_1, A_2 und A_3 die Komponenten eines Vektors $\vec{A}$ in x-, y-, z-Richtung sind, so ist

$$\vec{A} = (A_1, A_2, A_3) = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} .$$

Es gilt

$$\vec{A} + \vec{B} = (A_1 + B_1, A_2 + B_2, A_3 + B_3)$$

$$m\vec{A} = (mA_1, mA_2, mA_3)$$

Länge oder Betrag: $|\vec{A}| = \sqrt{A_1^2 + A_2^2 + A_3^2}$

Man kann einen Vektor also als Zeilen- oder als Spaltenvektor schreiben. Im Zusammenhang mit Matrizen (siehe den Beitrag von Christian Moldenhauer) muss man der Korrektheit halber die Spaltenschreibweise verwenden.

Das Skalarprodukt von zwei Vektoren ist definiert als $\vec{A} \cdot \vec{B} = A_1B_1 + A_2B_2 + A_3B_3$. Es gilt $\vec{A} \cdot \vec{B} = |\vec{A}| \cdot |\vec{B}| \cdot \cos\alpha$, wobei α der Zwischenwinkel der von Pfeilen repräsentierten Vektoren ist. Das Skalarprodukt ist kommutativ. Ist das Skalarprodukt von zwei Vektoren null, und ist keiner von ihnen der Nullvektor, so stehen sie senkrecht aufeinander.

Eine Funktion, die jedem n-Tupel $(x_1, x_2, \ldots, x_n)$ eine Zahl zuordnet, heißt skalare Funktion (von dieser Art waren alle Funktionen, die wir bisher betrachtet haben). Eine skalare Funktion $\Phi(x, y, z)$, die jedem Raumpunkt (x, y, z) eine Zahl (z. B. Temperatur) zuordnet, heißt ein skalares Feld. Entsprechend heißt eine Funktion, die jedem Raumpunkt mehr als eine Zahl (sprich: einen Vektor) zuordnet, ein Vektorfeld. Typische Beispiele sind Kraftfelder oder Magnetfelder (Vektor der magnetischen Feldstärke).

Der Nabla-Operator ist definiert als

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} .$$

Er ist ein Differentialoperator, der formal wie ein Vektor behandelt wird. Mit ihm definiert man den Gradienten eines skalaren Feldes $\Phi(x, y, z)$:

$$\nabla\Phi = \operatorname{grad}\Phi = \begin{pmatrix} \frac{\partial\Phi}{\partial x} \\ \frac{\partial\Phi}{\partial y} \\ \frac{\partial\Phi}{\partial z} \end{pmatrix} .$$

Der Gradient von Φ ist also der Vektor, der die partiellen Ableitungen von Φ als Komponenten hat, und definiert ein Vektorfeld. Der Gradient zeigt in jedem Punkt in Richtung des steilsten Anstieges und steht senkrecht auf einer Höhenlinie.

Ist $\vec{V}(x, y, z)$ eine vektorwertige Funktion, so definiert man

$$\nabla \cdot \vec{V} = \operatorname{div} \vec{V} = \frac{\partial V_1}{\partial x} + \frac{\partial V_2}{\partial y} + \frac{\partial V_3}{\partial z}$$

als die Divergenz von $\vec{V}$. Ein Beispiel für die Divergenz eines Vektorfeldes findet sich in den Maxwell'schen Gleichungen: die Ladungsdichte ist die Divergenz des elektrischen Feldes.

$$\Delta = \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

wird Laplace'scher Operator genannt.

Mehrdimensionale Integration

Die Herleitung des Zweifachintegrals erfolgt analog zu der des einfachen bestimmten Integrals. Anstelle der Berechnung eines Flächeninhaltes tritt die eines Volumens, die Teilintervalle werden zu Teilgebieten ΔA_k. Man erhält

$$\iint f(x,y)dA = \lim_{\substack{n\to\infty \\ \Delta A_k \to 0}} \sum_{k=1}^{n} f(x_k, y_k)\Delta A_k.$$

Die Berechnung eines solchen Doppelintegrals erfolgt in zwei gewöhnlichen Integrationsschritten. Wenn etwa die Fläche A, über die integriert werden soll, gleich dem Gebiet zwischen den Graphen zweier Funktionen $f_u(x)$ (unten) und $f_o(x)$ (oben) ist, berechnet man

$$\iint f(x,y)dA = \int_{x=a}^{b} \int_{y=f_u(x)}^{y=f_o(x)} f(x,y)dydx.$$

Im ersten Schritt (Auswertung des inneren Integrals) wird x als konstant angesehen und nach y integriert. Da die Integrationsgrenzen variabel sind, erhält man als Ergebnis eine von x abhängige Funktion. Diese wird im zweiten Schritt in den Grenzen von a bis b nach x integriert. Die Definition von mehrdimensionalen Integralen erfolgt entsprechend; ebenso ihre Berechnung durch mehrere gewöhnliche Integrationen.

Literatur: wie in Nataljas erstem Beitrag

Partielle Differentialgleichungen

(Petra Kersting)

Partielle Differentialgleichungen:

Das sind Gleichungen, in denen die gesuchten Funktionen u von mehr als einer Variablen abhängen und partielle Ableitungen auftreten. Die *Ordnung einer partiellen Differentialgleichung* gibt die höchste darin vorkommende (partielle) Ableitung an. Bei einer *linearen partiellen Differentialgleichung* treten die Funktionen und ihre Ableitungen linear auf (mit Koeffizienten, die nur von den unabhängigen Veränderlichen $x_1, x_2, \ldots, x_n$ abhängen).

Beispiel: $a_1 u_{xx} + a_2 u_{xy} + a_3 u_{yy} + a_4 u_x + a_5 u_y + a_6 u + a_7 = 0$ mit $a_k = a_k(x, y)$

Den Hauptteil einer partiellen Differentialgleichung stellen die Terme dar, die die höchste Ableitung enthalten. In einer *quasilinearen partiellen Differentialgleichung* ist der Hauptteil linear.

Beispiele von partiellen Differentialgleichungen

Die meisten für die Anwendungen interessanten partiellen Differentialgleichungen sind von zweiter Ordnung und quasilinear. Es gibt im Wesentlichen drei verschiedene Typen; von ihnen stelle ich jeweils den einfachsten (und charakteristischsten) Vertreter vor.

Elliptische Differentialgleichungen: die Poissongleichung

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = f(x,y), \quad (x,y) \in R$$

Dabei sei R ein Gebiet in der (x,y)-Ebene mit dem Rand S. Randbedingung: $u(x,y) = g(x,y)$ auf S (Dirichlet-Randbedingung).

Zum Beispiel ist u die (unbekannte) Temperatur einer Membran, Stahlplatte o. ä.; g ist die bekannte Temperatur am Rand der Platte und f eine über das Gebiet verteilte Wärmequelle. Die Lösung der Gleichung ist der Gleichgewichtszustand des Systems; sie ist eindeutig bestimmt.

Hyperbolische Differentialgleichungen: die Wellengleichung

$$\frac{\alpha^2 \partial^2 u(x,t)}{\partial x^2} = \frac{\partial^2 u(x,t)}{\partial t^2} \text{ für } 0 < x < l; 0 < t$$

mit den Anfangsbedingungen $u(x,0) = f(x)$ (Anfangslage zum Zeitpunkt $t = 0$), $\frac{\partial u(x,0)}{\partial t} = g(x)$ (Anfangsgeschwindigkeit) und den Randbedingungen $u(0,t) = g_1(t) = 0$ und $u(l,t) = g_2(t) = 0$.

Zum Beispiel ist $u(x,t)$ die Auslenkung eines elastischen Bandes der Länge l, das bei $x = 0$ und $x = l$ zwischen zwei Halterungen eingespannt ist. (Dämpfung wird vernachlässigt.)

Parabolische Differentialgleichungen: die Wärmeleitungsgleichung

$$\frac{\partial u(x,t)}{\partial t} - \frac{\alpha^2 \partial^2 u(x,t)}{\partial x^2} = 0, \quad x \in [0,l], \quad t \geq R \tag{1}$$

mit den Anfangsbedingungen $u(x,0) = f(x),\ x \in [0,l]$ und den Randbedingungen $u(0,t) = g_1$ und $u(l,t) = g_2$ für alle t.

Die Gleichung beschreibt zum Beispiel die Wärmeleitung in einem Stab der Länge l, der in jedem Querschnittselement eine einheitliche Temperatur aufweist. Der Stab ist an seinen Seitenflächen vollständig isoliert. Die Konstante α lässt sich über die wärmeleitenden Eigenschaften des Stab-Materials bestimmen. Dabei ist $u(x,t)$ die Temperatur an dem Ort x zur Zeit t. Die Anfangsbedingung stellt hier die anfängliche Temperaturverteilung in dem Stab dar. Die Randbedingungen legen fest, dass die Enden bei den konstanten Temperaturen $u(0,t) = g_1$ und $u(l,t) = g_2$ gehalten werden.

Bestimmen der Lösung der Wärmeleitungsgleichung

Wir beschränken uns auf den einfacheren Fall, dass die Randbedingungen $u(0,t) = u(1,t) = 0$ sind. Man kann den Separationsansatz $u(x,t) = v(x) \cdot w(t)$ verwenden, d. h. man sucht eine spezielle Lösung, die sich in ein Produkt von Funktionen zerlegen lässt, die nicht von allen unabhängigen Variablen abhängen. Durch Einsetzen in die Wärmeleitungsgleichung ergibt sich $v(x) \cdot w'(t) - \alpha^2 \cdot v''(x) \cdot w(t) = 0 \Longleftrightarrow v(x) \cdot w'(t) = \alpha^2 \cdot v''(x) \cdot w(t)$.

Unter der Voraussetzung: $v(x) \neq 0$ und $w(t) \neq 0$ erhält man

$$\frac{w'(t)}{w(t)} = \frac{\alpha^2 \cdot v''(x)}{v(x)} \ .$$

Da die linke Seite von x und die rechte Seite von t unabhängig, beide Seiten allerdings gleich sind, müssen beide Seiten gleich einer Konstanten C sein:

$$\frac{w'(t)}{w(t)} = C \quad \text{und} \quad \frac{\alpha^2 \cdot v''(x)}{v(x)} = C$$

$$w'(t) = C \cdot w(t) \tag{2}$$
$$\alpha^2 \cdot v''(x) = C \cdot v(x) \tag{3}$$

Es liegen nun zwei gewöhnliche Differentialgleichungen vor.

Gleichung (2) lösen wir, wie Christoph Grothaus uns das vorgemacht hat. Es kommt heraus

$$w(t) = e^{C \cdot t} \cdot C_1 \ . \tag{4}$$

Für die Gleichung (3) machen wir den Ansatz $v(x) = e^{\lambda \cdot x}$ und erhalten $v'(x) = \lambda \cdot e^{\lambda \cdot x}$ und $v''(x) = \lambda^2 \cdot e^{\lambda \cdot x}$. Durch Einsetzen von $v''(x)$ und $v(x)$ in die Gleichung (3) erhalten wir $\lambda^2 \cdot e^{\lambda \cdot x} = \frac{C}{\alpha^2} \cdot e^{\lambda \cdot x}$ mit den Lösungen $\lambda_1 = \sqrt{\frac{C}{\alpha^2}}$ und $\lambda_2 = -\lambda_1 = -\sqrt{\frac{C}{\alpha^2}}$. Im Folgenden schreibe ich der Einfachheit zuliebe λ statt λ_1.

Nach dem Superpositionsprinzip ist die Summe zweier Lösungen wiederum eine Lösung der Gleichung, so dass nun folgender allgemeiner Ansatz für die Lösung der Gleichung (3) gewählt werden kann: $v(x) = c_1 \cdot e^{\lambda \cdot x} + c_2 \cdot e^{-\lambda \cdot x}$. Durch Einsetzen in den Separationsansatz ergibt sich $u(x,t) = (c_1 \cdot e^{\lambda \cdot x} + c_2 \cdot e^{-\lambda \cdot x}) \cdot w(x)$.

Durch Einarbeiten der Rand- und Anfangsbedingungen $u(0,t) = 0$ und $u(1,t) = 0$ lassen sich die Konstanten bestimmen: $u(0,t) = (c_1 + c_2) \cdot w(t) = 0 \Rightarrow c_1 + c_2 = 0$, denn $w(t) \neq 0$, $c_1 = -c_2$; $u(1,t) = v(1) \cdot w(t) = 0 \Rightarrow -c_2(e^{\lambda} - e^{-\lambda}) = 0 \Rightarrow e^{\lambda} = e^{-\lambda}$, denn $c_2 \neq 0$ (der Fall $c_2 = 0$ ist langweilig) $\Rightarrow e^{2\lambda} = 1$.

Jetzt müssen wir die Formel $e^{i \cdot \phi} = \cos(\phi) + i \cdot \sin(\phi)$ einsetzen. $e^{2\lambda} = 1$ kann nur sein, wenn λ imaginär ist. Ich schreibe $\lambda = ia$ mit reellem a. Darum gilt $1 = e^{2\lambda} = e^{2ia} = \cos(2a) + i\sin(2a)$. Dann muss $2a = 2k\pi$ mit $k \in Z$ (ganze Zahlen) sein.

Alles zusammen eingesetzt ergibt $C = k^2\pi^2 i^2 \alpha^2 = -k^2\pi^2\alpha^2$ und $v(x) = -c_2 \cdot (e^{k\pi ix} - e^{-k\pi ix}) = -c_2 \cdot (e^{iax} - e^{-iax}) = -c_2 \cdot 2i \sin ax = c(k) \cdot \sin(k\pi x)$ (es sind unendlich viele Lösungen, da $k \in Z$ beliebig ist).

Wie passt man die Vielfalt von Lösungen an die Anfangsbedingung an? Jede (einigermaßen anständige) Funktion f läßt sich in eine Fourierreihe entwickeln:

$$f(x) = \sum_{k=0}^{\infty} a_k \cdot \sin(k\pi x) \quad \text{mit } \; a_k \in R$$

Man erhält nun endlich die Lösung der Wärmeleitungsgleichung (1):

$$u(x,t) = v(x) \cdot w(t) = \sum_{k=0}^{\infty} a_k \cdot \sin(k\pi x) \cdot e^{-k^2\pi^2\alpha^2 \cdot t}$$

Die Integrationskonstante C_1 aus Gleichung (4) stellt sich als entbehrlich heraus. Die a_k erledigen die Sache mit dem konstanten Faktor schon.

Energieintegral – Nachweis der Eindeutigkeit

Um zu zeigen, dass diese Lösung eindeutig ist, nimmt man an, es gebe zwei verschiedene Lösungen $u_1(x,t)$ und $u_2(x,t)$. Ziel ist es dabei, zu zeigen, dass die Differenz $u_1 - u_2 = 0$ ist, denn daraus würde folgen: $u_1 = u_2$.

Weil u_1 Lösung ist, gilt $u_{1t} - \alpha^2 u_{1xx} = 0$ mit den Bedingungen $u_1(0,t) = g_1, u_1(1,t) = g_2, u_1(x,0) = f(x)$, entsprechend muss gelten: $u_{2t} - \alpha^2 u_{2xx} = 0$ mit den selben Bedingungen. Man definiert: $u(x,t) := u_1(x,t) - u_2(x,t)$. Durch Subtrahieren der beiden angenommenen Lösungen erhält man :

$$u_{1t} - \alpha^2 u_{1xx} - u_{2t} + \alpha^2 u_{2xx} = u_t - \alpha^2 u_{xx} = 0 \tag{5}$$

mit den Bedingungen: $u(0,t) = 0, u(1,t) = 0, u(x,0) = 0$. Daraus soll folgen: $u(x,t) = 0$ für alle x und t.

Ich multipliziere (5) mit u, forme um und erhalte $u_t u = \alpha^2 u_{xx} u$, integriere von 0 bis 1: $\int_0^1 u_t u dx = \alpha^2 \int_0^1 u_{xx} u dx$

Die linke Seite ist $= \int_0^1 \frac{\partial}{\partial t} \frac{u^2}{2} dx = \frac{d}{dt} \int_0^1 \frac{u^2}{2} dx$ (wir dürfen Differentiation und Integration vertauschen).

Auf die rechte Seite wenden wir partielle Integration an: $\alpha^2 \int_0^1 u_{xx} u dx = \alpha^2 (u_x u)\Big|_0^1 - \alpha^2 \int_0^1 u_x^2 dx = -\alpha^2 \int_0^1 u_x^2 dx$, denn $(u_x u)\Big|_0^1 = 0$, da $u(0,t) = u(1,t) = 0$.

Nun fasse ich alles zusammen:

$$I(t) := \int_0^1 u^2(x,t) dx \tag{a}$$

ist immer ≥ 0, denn der Integrand u^2 ist ≥ 0.

$$I'(t) = 2\frac{d}{dt} \int_0^1 \frac{1}{2} u^2(x,t) dx = -2\alpha^2 \int_0^1 u_x^2(x,t) dx \leq 0 \; , \tag{b}$$

denn der Integrand u_x^2 ist ≥ 0, und

$$I(0) = \int_0^1 u^2(x,0)dx = 0\ , \qquad (c)$$

weil der Anfangswert $= 0$ ist.

Also: I ist anfangs 0 (c), ist stets ≥ 0 (a) und kann höchstens weniger werden (b). Also muss $I = 0$ sein. Dann muss auch u für alle t und $x = 0$ sein, sonst würde u^2 einen Beitrag zu dem Integral liefern.

Bei einem Schwingungsproblem ist $\int_0^1 u_x^2 dx$ die gesamte potentielle Energie des Zustandes. Daher kommt der Name Energieintegral.

Literaturangaben:

J. Douglas Faires, Richard L. Burden: Numerische Methoden. Spektrum Akademischer Verlag 1994, S. 474ff.

Wieland Richter: Partielle Differentialgleichungen. Spektrum Akademischer Verlag 1995, S. 61f., 147f.

Finite-Differenzen-Verfahren zum Lösen partieller Differentialgleichungen

(Jörn-Thorsten Paßmann)

Elliptische Gleichungen

Gegeben sei eine elliptische Differentialgleichung, die sogenannte Poisson-Gleichung. Sie hat die Form

$$\frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y)$$

und sei definiert auf einem rechteckigen Gebiet

$$R = \{(x,y) | a < x < b, c < y < d\}$$

mit der Randbedingung (S bezeichnet den Rand des Gebietes)

$$u(x,y) = g(x,y) \quad \text{für } (x,y) \in S\ .$$

Die Differentialgleichung allein ist nicht eindeutig lösbar; für ein komplettes Problem müssen Rand- und/oder Anfangswertbedingungen hinzukommen. Das hier gegebene Problem ist eindeutig lösbar, sofern f und g auf ihrem Definitionsbereich stetig sind.

Ein Beispiel für ein physikalisches Problem, das durch diese Gleichung beschrieben wird, ist die stationäre Temperaturverteilung in einer dünnen rechteckigen Metallplatte. Die Platte wird durch die x-Koordinaten a und b nach links und rechts und die y-Koordinaten c und d nach unten und oben begrenzt. In diesem Fall gilt $f(x,y) = 0$ (keine äußere Wärmequelle), was die Gleichung zur sogenannten Laplace-Gleichung vereinfacht. g ist eine bekannte Funktion, die die Temperaturen beschreibt, die am Rand herrschen. $u(x,y)$ ist die gesuchte Temperatur an einer bestimmten Stelle der Platte.

Ziel soll sein, eine Poisson-Gleichung numerisch zu lösen, d. h. die Werte der Funktion u an möglichst dicht verteilten diskreten Punkten näherungsweise zu berechnen.

Diese Punkte werden z. B. ausgewählt, indem das Gebiet R mit einem Netz überzogen wird. Man legt eine Anzahl n von Zerlegungen in x-Richtung und m in y-Richtung fest und definiert die Schrittweiten $h := \frac{b-a}{n}$ und $k := \frac{d-c}{m}$. Sodann zieht man horizontale und vertikale Gitternetzlinien durch die Gitterpunkte (x_i, y_j) mit $x_i := a + ih$ und $y_j := c + jk$.

Es sollen nun die Werte $u(x_i, y_j)$ durch die Werte von u in umliegenden Punkten näherungsweise ausgedrückt werden. Dadurch entsteht ein Gleichungssystem mit den Werten in den Gitterpunkten als Unbekannten. Setzt man die Rand- beziehungsweise Anfangswerte ein, kann man es lösen und erhält Näherungen der gesuchten Funktionswerte in allen Gitterpunkten.

Dazu werden zunächst die zweiten partiellen Ableitungen an der Stelle (x_i, y_j) mit Hilfe der Taylor-Reihe durch die Werte und Ableitungen von u in (x_i, y_j) und umliegenden Punkten ausgedrückt. Die zweite partielle Ableitung nach x ergibt sich folgendermaßen:

$$u(x_{i+1}, y_j) = u(x_i, y_j) + \frac{h}{1!}\frac{\partial u}{\partial x}(x_i, y_j) + \frac{h^2}{2!}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) + \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3}(x_i, y_j) + \frac{h^4}{4!}\frac{\partial^4 u}{\partial x^4}(\xi_i^{(1)}, y_j) \quad \text{mit } \xi_i^{(1)} \in [x_i, x_{i+1}]$$

und

$$u(x_{i-1},y_j) = u(x_i,y_j) - \frac{h}{1!}\frac{\partial u}{\partial x}(x_i,y_j) + \frac{h^2}{2!}\frac{\partial^2 u}{\partial x^2}(x_i,y_j) - \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3}(x_i,y_j) + \frac{h^4}{4!}\frac{\partial^4 u}{\partial x^4}(\xi_i^{(2)},y_j) \quad \text{mit } \xi_i^{(2)} \in [x_{i-1},x_i]$$

Die Gleichungen addiert man und fasst die Restglieder mit $O(h^4)$ zusammen:

$$\begin{aligned} u(x_{i+1},y_j) + u(x_{i-1},y_j) &= 2u(x_i,y_j) + \frac{2h^2}{2!}\frac{\partial^2 u}{\partial x^2}(x_i,y_j) + O(h^4) \\ \Leftrightarrow \frac{\partial^2 u}{\partial x^2}(x_i,y_j) &= \frac{u(x_{i-1},y_j) - 2u(x_i,y_j) + u(x_{i+1},y_j)}{h^2} - O(h^2) \end{aligned} \tag{1}$$

Entsprechend ergibt sich für die zweite partielle Ableitung nach y

$$\frac{\partial^2 u}{\partial y^2}(x_i,y_j) = \frac{u(x_i,y_{j-1}) - 2u(x_i,y_j) + u(x_i,y_{j+1})}{k^2} - O(k^2)\ .$$

Einsetzen in die Poisson-Gleichung führt zu

$$\frac{u(x_{i-1},y_j) - 2u(x_i,y_j) + u(x_{i+1},y_j)}{h^2} + \frac{u(x_i,y_{j-1}) - 2u(x_i,y_j) + u(x_i,y_{j+1})}{k^2} = f(x_i,y_j) + O(h^2+k^2)$$

Anschließend wird die Formel diskretisiert, d. h. für die korrekten Werte $u(x_i,y_j)$ werden die approximierten (genäherten) Werte ω_{ij} eingesetzt, und das Restglied wird vernachlässigt. Das führt zur zentralen Differenzenmethode mit der Formel

$$2\left[\left(\frac{h}{k}\right)^2 + 1\right]\omega_{ij} - (\omega_{i+1,j} + \omega_{i-1,j}) - \left(\frac{h}{k}\right)^2(\omega_{i,j+1} + \omega_{i,j-1}) = -h^2 f(x_i,y_j)\ .$$

Der durch das Vernachlässigen des Restgliedes entstandene Fehler liegt in der Größenordnung von h^2 und k^2 und kann durch kleine Schrittweiten klein gehalten werden.

Es empfiehlt sich, jetzt die inneren Gitterpunkte (x_i,y_j) von links nach rechts und von oben nach unten durchzunummerieren, ebenso wie ihre genäherten Funktionswerte: $P_1 = (x_1,y_{m-1})$ mit dem Näherungswert ω_1, $P_2 = (x_2,y_{m-1})$ mit ω_2 und so weiter. Nach der zentralen Differenzenmethode lässt sich für jeden Punkt eine Gleichung aufstellen, die den Wert ω durch die umliegenden Werte ausdrückt. Durch Einsetzen der Randbedingungen erhält man ein Gleichungssystem, in dem nur die inneren Gitterpunkte als Unbekannte vorkommen. In diese Gleichungen werden auch alle dem Wert nicht benachbarten Werte eingefügt und mit dem Koeffizienten 0 versehen, so dass das Gleichungssystem so viele Gleichungen und so viele Variablen enthält, wie es innere Gitterpunkte gibt. Das so entstandene Gleichungssystem lässt sich als Matrix schreiben und kann schließlich z. B. mit einem der Verfahren, die Christian Moldenhauer in seinem Beitrag darstellt, aufgelöst werden.

Parabolische Gleichungen

Aufsteigendes Differenzenverfahren

Bei der sogenannten Wärme- oder Diffusionsgleichung handelt es sich um eine parabolische Gleichung. Sie hat die Form

$$\frac{\partial u}{\partial t}(x,t) = \alpha^2\frac{\partial^2 u}{\partial x^2}(x,t) \quad \text{für } 0 < x < l \text{ und } t > 0$$

mit den Anfangs- und Randbedingungen:

$$u(0,t) = u(l,t) = 0 \text{ für } t > 0 \text{ und } u(x,0) = f(x) \text{ für } 0 \le x \le l$$

Zur numerischen Lösung dieses Problems wird prinzipiell genauso wie im vorigen Abschnitt beschrieben vorgegangen. Die Näherungslösung für $\frac{\partial^2 u}{\partial x^2}$ (Gleichung (1)) kann (ersetze y durch t) übernommen und in die Wärmegleichung eingesetzt werden.

Die Taylorreihe von u um (x_i,t_j)

$$u(x_i,t_{j+1}) = u(x_i,t_j) + \frac{k}{1!}\frac{\partial u}{\partial t}(x_i,t_j) + \frac{k^2}{2!}\frac{\partial^2 u}{\partial t^2}(x_i,\mu_j) \quad \text{mit } \mu_j \in [t_j,t_{j+1}] \tag{2}$$

lässt sich nach der ersten partiellen Ableitung nach t auflösen:

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} - \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) \tag{3}$$

Einsetzen dieses Terms in die Wärmegleichung und Diskretisieren führen zum aufsteigenden Differenzenverfahren

$$\frac{\omega_{i,j+1} - \omega_{ij}}{k} - \alpha^2 \frac{\omega_{i+1,j} - 2\omega_{ij} + \omega_{i-1,j}}{h^2} = 0$$
$$\Leftrightarrow \omega_{i,j+1} = \left(1 - \frac{2\alpha^2 k}{h^2}\right)\omega_{ij} + \alpha^2 \frac{k}{h^2}(\omega_{i+1,j} + \omega_{i-1,j}) \; .$$

Es handelt sich hierbei um ein explizites Verfahren, da aus den für den ersten Zeitschritt durch die Anfangswertbedingung gegebenen Werten direkt die Werte für den nächsten Zeitschritt berechnet werden können und so weiter. Daher erfordert das Verfahren kein rechenintensives Lösen eines Gleichungssystems.

Die Nachteile liegen zum einen in der Fehlerordnung: Wegen des k im Restglied hängt die Größenordnung von h^2 und k ab, was einen sehr kleinen Zeitschritt erfordert, will man hinreichend genaue Ergebnisse erhalten. Zum anderen gilt es, das Stabilitätskriterium $\alpha^2 \frac{k}{h^2} \leq \frac{1}{2}$ zu erfüllen, da sich ansonsten völlig fehlerhafte und unbrauchbare Werte ergeben.

Absteigendes Differenzenverfahren

In Gleichung (3) wird $\frac{\partial u}{\partial t}(x_i, t_j)$ mit Hilfe von $u(x_i, t_{j+1})$, dem Funktionswert im folgenden Zeitschritt, ausgedrückt. Berechnet man die erste partielle Ableitung nach der Zeit jedoch unter Verwendung des Wertes im vorangehenden Zeitschritt, $u(x_i, t_j)$, so erhält man statt der Gleichung (3) folgenden Ausdruck:

$$\frac{\partial u}{\partial t}(x_i, t_j) = \frac{u(x_i, t_j) - u(x_i, t_{j-1})}{k} + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_i, \mu_j) \quad \text{mit } \mu_j \in (t_{j-1}, t_j) \tag{4}$$

Setzt man dies statt (3) in die Wärmegleichung ein, so erhält man nach Diskretisieren das absteigende Differenzenverfahren

$$\frac{\omega_{ij} - \omega_{i,j-1}}{k} - \alpha^2 \frac{\omega_{i+1,j} - 2\omega_{ij} + \omega_{i-1,j}}{h^2} = 0 \quad .$$

Es hat ebenfalls den Fehler $O(h^2 + k)$, weist jedoch nicht die Stabilitätsprobleme des aufsteigenden Verfahrens auf. Da es ein implizites Verfahren ist, führt es zu einem Gleichungssystem, das gelöst werden muss.

Methode von Crank-Nicolson

Durch Addieren der Gleichungen des aufsteigenden Verfahrens im Schritt j,

$$\frac{\omega_{i,j+1} - \omega_{ij}}{k} - \alpha^2 \frac{\omega_{i+1,j} - 2\omega_{ij} + \omega_{i-1,j}}{h^2} = 0 \quad ,$$

und des absteigenden Verfahrens im Schritt $j + 1$,

$$\frac{\omega_{i,j+1} - \omega_{ij}}{k} - \alpha^2 \frac{\omega_{i+1,j+1} - 2\omega_{i,j+1} + \omega_{i-1,j+1}}{h^2} = 0 \quad ,$$

entsteht die Gleichung der Methode von Crank-Nicolson:

$$\frac{\omega_{i,j+1} - \omega_{ij}}{k} - \frac{\alpha^2}{2}\left[\frac{\omega_{i+1,j} - 2\omega_{ij} + \omega_{i-1,j}}{h^2} + \frac{\omega_{i+1,j+1} - 2\omega_{i,j+1} + \omega_{i-1,j+1}}{h^2}\right] = 0$$

Da die Restglieder der Gleichungen (3) und (4) in etwa den gleichen Betrag, aber unterschiedliche Vorzeichen besitzen, heben sie sich gegenseitig auf. Folglich weist dieses Verfahren den günstigen Fehler $O(h^2 + k^2)$ auf.

Wie ist das? Das aufsteigende Verfahren setzt das (diskretisierte) u_{xx} von heute (Zeitschritt j) in Beziehung mit der Differenz zwischen morgen (Zeitschritt $j + 1$) und heute. Das absteigende Verfahren setzt das u_{xx} von morgen in Beziehung mit der Differenz zwischen morgen und heute. Gehört die Differenz zwischen morgen und heute eher zu morgen oder zu heute? Na ja, sie gehört zu beiden Zeitpunkten gleich schlecht, nämlich mit dem Fehler $O(k)$. Nur: Daran können wir nichts tun. Wenn wir u_t diskretisieren wollen und haben nur die Werte von u in den Zeitpunkten t_j und t_{j+1} zur Verfügung, gibt es keine bessere Näherung als den Differenzenquotienten. Interessant ist es, die Frage andersrum zu stellen: Für welchen Zeitpunkt t liefert der Differenzenquotient $(u(t_{j+1}) - u(t_j))/k$ die beste Näherung an $u_t(t)$? Antwort: für den Zeitpunkt genau in der Mitte, $t = (t_j + t_{j+1})/2$, „heute um Mitternacht". Dann nämlich läppern sich in der Taylorreihe die Glieder erster Ordnung genau weg, und es bleibt ein Fehler zweiter Ordnung übrig. Wenn das so ist, müssen wir natürlich auch das genäherte u_{xx} heute um Mitternacht auswerten. Wie macht man das, mit den Gitterpunkten, die wir zur Verfügung haben? Man nimmt den Mittelwert zwischen heute und morgen – was sonst? Und siehe da – es entsteht das Verfahren von Crank-Nicolson.

Literatur: J. Douglas Faires, Richard L. Burden, Numerische Methoden, Spektrum Akademischer Verlag 1994

Finite Elemente

(Arne Schneck)

Finite-Elemente-Methoden (FEM) sind eine Möglichkeit, gewöhnliche und insbesondere partielle Differentialgleichungen numerisch zu lösen. Anders als beispielsweise bei den Finite-Differenzen-Methoden (FDM) wird jedoch nicht von der Differentialgleichung selbst ausgegangen. Für den FEM-Ansatz muss die Differentialgleichung vielmehr zunächst umgeformt werden in ein sogenanntes Variationsproblem.

Unter einem Variationsproblem versteht man die Aufgabe, die Funktion zu finden, die einen bestimmten Ausdruck extremal werden lässt. Anschaulich kann man zum Beispiel den Gleichgewichtszustand einer Konstruktion wie etwa eines Gebäudedachs auf zwei Arten beschreiben. Bei einer Differentialgleichung betrachtet man alle Kräfte, die auf einen Punkt wirken. Wenn die Gleichgewichtskonfiguration erreicht ist, muss die Summe aller Kräfte gleich Null sein, andernfalls würde sich die Struktur bewegen. Man kann diesen Zustand aber auch anders ausdrücken, und zwar als Variationsaufgabe („Energiefunktional“): Der Gleichgewichtszustand ist erreicht, wenn in der Struktur die minimale potentielle Energie steckt; überschüssige Energie würde sich zum Beispiel in Bewegung verwandeln (die vielleicht zum Einsturz des Gebäudes führen könnte).

Mathematisch betrachtet ist es relativ einfach, ein vorgegebenes Variationsproblem in eine Differentialgleichung umzuformen. Der umgekehrte Weg, zu einer vorhandenen Differentialgleichung das entsprechende Variationsproblem zu finden, gestaltet sich jedoch äußerst schwierig. Hier soll daher nur an einem Beispiel gezeigt werden, wie man von einem Variationsproblem auf die entsprechende Differentialgleichung kommt.

Betrachten wir dazu die folgende gewöhnliche Differentialgleichung mit Randbedingungen:

$$-\frac{\partial}{\partial x}\big(p(x)\frac{\partial y}{\partial x}\big) + q(x)\cdot y(x) = f(x), \qquad y(0) = y(1) = 0 \tag{1}$$

das sogenannte Sturm-Liouville-Problem. Im Folgenden soll gezeigt werden, dass das Minimum u des Funktionals

$$I[u] = \int_0^1 \Big(p(x)[u'(x)]^2 + q(x)[u(x)]^2 - 2f(x)u(x) \Big)\, dx \tag{2}$$

die Differentialgleichung (1) löst. Wenn u ein solches Minimum ist, dann muss für beliebige Funktionen $\phi(x)$ (die die Randbedingugen erfüllen und einmal stetig differenzierbar sind) gelten: $I[u] \leq I[u+\epsilon\phi]$, mit $\epsilon \in R$. Nimmt man u und ϕ als feste Funktionen, bleibt nur noch ϵ als freie Veränderliche übrig. Wenn $I[u+\epsilon\phi]$ minimal sein soll, muss seine Ableitung an der Stelle $\epsilon = 0$ gleich null sein. Dies führt auf

$$\begin{aligned}
0 &= \frac{d}{d\epsilon} I[u(x) + \epsilon\cdot\phi(x)]\,\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\int_0^1 \Big(p(x)\Big(u'(x)+\epsilon\phi'(x)\Big)^2 + q(x)\Big(u(x)+\epsilon\phi(x)\Big)^2 - 2f(x)\Big(u(x)+\epsilon\phi(x)\Big)\Big)\, dx\,\Big|_{\epsilon=0} \\
&= \int_0^1 \Big(2p(x)\Big(u'(x)+\epsilon\phi'(x)\Big)\cdot\phi'(x) + 2q(x)\Big(u(x)+\epsilon\phi(x)\Big)\cdot\phi(x) - 2f(x)\phi(x)\Big)\, dx\,\Big|_{\epsilon=0} \\
&= 2\int_0^1 \Big(p(x)u'(x)\phi'(x) + q(x)u(x)\phi(x) - f(x)\phi(x)\Big)\, dx
\end{aligned}$$

Wegen $\int_0^1 \big(p(x)u'(x)\big)\cdot\phi'(x)\,dx = \big[p(x)u'(x)\cdot\phi(x)\big]_0^1 - \int_0^1 \big(p(x)u'(x)\big)'\phi(x)\,dx$ und $\big[p(x)u'(x)\cdot\phi(x)\big]_0^1 = 0$ (da $\phi(x)$ die Randbedingungen erfüllt und somit $\phi(0)=\phi(1)=0$ ist) ergibt sich

$$\int_0^1 \Big(-\big(p(x)u'(x)\big)' + q(x)u(x) - f(x)\Big)\cdot\phi(x)\,dx = 0$$

für beliebige Funktionen $\phi(x)$, welche die Randbedingungen erfüllen. Damit das Integral aber für alle $\phi(x)$ gleich Null ist, muss $-\big(p(x)u'(x)\big)' + q(x)u(x) - f(x) = 0$ sein. Damit ist also bewiesen, dass die Funktion $u(x)$, die (2) minimiert, genau die Gleichung (1) löst. *Andersrum stimmt es nicht unbedingt. Minimum und Nullstelle der ersten Ableitung sind schon bei gewöhnlichen Funktionen nicht dasselbe, und bei solch komplizierten Funktionalen erst recht nicht. Aber meistens interessiert man sich unter den möglicherweise vielen Lösungen von (1) genau für die, die (2) minimiert – weil nämlich das ursprüngliche Problem auf Minimierung einer potentiellen Energie hinauslief.*

Der FEM-Ansatz geht nun direkt vom Variationsproblem aus. Er versucht allerdings nicht, genau die richtige Funktion zu finden, die den Variationsausdruck minimiert, sondern er wählt eine Teilmenge von Funktionen aus, über die dann minimiert wird. Dazu wählt man aus dem Intervall $[0,1]$, über dem man approximiert, eine endliche Anzahl von Punkten x_0, x_1, …, x_{n+1} mit $0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = 1$ aus. $u(x)$ wird dann approximiert durch $\Phi(x) = \sum_{i=1}^{n} c_i \Phi_i(x)$. Die Funktionen $\Phi_i(x)$ sind dabei festgelegte, zum Beispiel stückweise lineare Basisfunktionen zu x_i, die nur auf einem kleinen Intervall ungleich null sind, und die Koeffizienten c_i sind die Näherungen für die Funktionswerte $u(x_i)$, die es zu finden gilt.

Die stückweise linearen Basisfunktionen $\Phi_i(x)$ werden definiert durch

$$\Phi_i(x) = \begin{cases} 0 & \text{für } 0 \le x \le x_{i-1} \\ \frac{x - x_{i-1}}{h_{i-1}} & \text{für } x_{i-1} < x \le x_i \\ \frac{x_{i+1} - x}{h_i} & \text{für } x_i < x \le x_{i+1} \\ 0 & \text{für } x_{i+1} < x \le 1 \end{cases},$$

wobei $h_i = x_{i+1} - x_i$ für alle $i = 0, 1, \ldots, n$ ist. Die Ableitungen $\Phi_i'(x)$ sind demnach

$$\Phi_i'(x) = \begin{cases} 0 & \text{für } 0 \le x \le x_{i-1} \\ \frac{1}{h_{i-1}} & \text{für } x_{i-1} < x \le x_i \\ -\frac{1}{h_i} & \text{für } x_i < x \le x_{i+1} \\ 0 & \text{für } x_{i+1} < x \le 1 \end{cases}.$$

Die Funktionen $\Phi_i(x)$ beschreiben somit kleine „Hütchenfunktionen“, die an der Spitze den Wert 1 annehmen. Multipliziert man die einzelnen $\Phi_i(x)$ mit den einzelnen Koeffizienten c_i und addiert sie, ergibt sich daraus für $\Phi(x)$ eine Art Streckenzug, also eine stückweise lineare Kurve. Dass die Koeffizienten c_i eine Approximation der Funktion u an x_i darstellen, ergibt sich aus der Definition von $\Phi(x)$.

Nun setzt man $\Phi(x)$ in das Funktional I in (2) ein:

$$I[\Phi] = I[\sum_{i=1}^{n} c_i \Phi_i(x)] = \int_0^1 \left(p(x) \left[\sum_{i=1}^{n} c_i \Phi_i'(x) \right]^2 + q(x) \left[\sum_{i=1}^{n} c_i \Phi_i(x) \right]^2 - 2 f(x) \sum_{i=1}^{n} c_i \Phi_i(x) \right) dx.$$

Damit ein Minimum auftritt, muss die Ableitung nach den einzelnen Koeffizienten c_j (die einzigen wirklich Veränderlichen, da die Basisfunktionen ja festgelegt sind) gleich Null sein:

$$0 = \frac{\partial I[\Phi]}{\partial c_j} = \int_0^1 \left(2p(x) \sum_{i=1}^{n} c_i \Phi_i'(x) \Phi_j'(x) + 2q(x) \sum_{i=1}^{n} c_i \Phi_i(x) \Phi_j(x) - 2 f(x) \Phi_j(x) \right) dx$$

Nach Umformen ergibt sich

$$\sum_{i=1}^{n} \left[\int_0^1 \Big(p(x) \Phi_i'(x) \Phi_j'(x) + q(x) \Phi_i(x) \Phi_j(x) \Big) \, dx \right] c_i = \int_0^1 f(x) \Phi_j(x) \, dx.$$

Definiert man nun

$$a_{ji} = \int_0^1 \Big(p(x) \Phi_i'(x) \Phi_j'(x) + q(x) \Phi_i(x) \Phi_j(x) \Big) \, dx$$

und

$$b_j = \int_0^1 f(x) \Phi_j(x) \, dx,$$

entsteht eine quadratische Matrix der Form $Ac = b$, die man dann für alle c_i lösen kann.

Auf Grund der speziellen Definition der Basisfunktionen gilt

$$\Phi_i(x) \Phi_j(x) = 0 \quad \text{und} \quad \Phi_i'(x) \Phi_j'(x) = 0,$$

falls $j \neq i$, $i - 1$ oder $i + 1$ ist. Daher reduziert sich die Matrix zu einer Tridiagonalmatrix, die relativ einfach zu lösen ist.

Für partielle Differentialgleichungen in 2 Raumdimensionen läuft das Verfahren ähnlich. Zunächst muss zu der partiellen Differentialgleichung ein entsprechender Variationsausdruck gefunden werden. Das Gebiet, über dem die Lösung approximiert werden soll, wird in Dreiecke unterteilt, und statt der zweidimensionalen „Hütchen“ kann man nun dreidimensionale „Hütchen“ (Pyramiden mit regelmäßigen oder unregelmäßigen Polygonen als Grundfläche) als Basisfunktionen verwenden. Dadurch entsteht wieder eine Matrix, die zwar nicht mehr tridiagonal ist, aber trotzdem zu einem großen Teil aus Nullen besteht.

Literatur:

J. D. Faires, R. L. Burden: Numerische Methoden, S. 458–470, 505–518. Heidelberg: Spektrum Akademischer Verlag (1994).

R. Wieland: Partielle Differentialgleichungen, S. 197–201. Heidelberg: Spektrum Akademischer Verlag (1995).

R. Rannacher, E. Stein: Finite Elemente. Spektrum der Wissenschaft 3/97.

Lösen von großen linearen Gleichungssystemen mit direkten und iterativen Methoden

(Christian Moldenhauer)

Um große lineare Gleichungssysteme der Form $A \cdot \vec{x} = \vec{b}$ zu lösen, wobei A eine $(n \cdot n)$-Matrix und $\vec{x}$ und $\vec{b}$ Vektoren sind, unterscheidet man zwei Lösungsansätze. Direkte Verfahren lösen das Gleichungssystem unter der Voraussetzung einer fehlerfreien Arithmetik nach endlich vielen Schritten exakt. Iterative Verfahren hingegen liefern auch dann nicht die exakte Lösung, wenn mit einer fehlerfreien Arithmetik gerechnet wird. Oft sind iterative Verfahren jedoch effektiver, da sie Rechenaufwand und Rundungsfehler reduzieren können. Mit iterativen Algorithmen ist es möglich, die Lösung bis auf eine beliebige Genauigkeit zu berechnen und dann den Algorithmus abzubrechen. Im Folgenden sollen hier der Gaußsche Algorithmus als Vertreter einer direkten und das Gauß-Seidel-Verfahren als Vertreter einer iterativen Methode vorgestellt werden.

Der Gaußsche Algorithmus

Gegeben ist das folgende Gleichungssystem aus den Gleichungen E_i

$$\begin{array}{lccccc} E_1: & a_{11}x_1 & +a_{12}x_2 & +\cdots & +a_{1n}x_n & = b_1 \\ E_2: & a_{21}x_1 & +a_{22}x_2 & +\cdots & +a_{2n}x_n & = b_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ E_n: & a_{n1}x_1 & +a_{n2}x_2 & +\cdots & +a_{nn}x_n & = b_n \end{array}$$

Zur Vereinfachung schreibt man dieses Geichungssystem in erweiterter Matrizenschreibweise

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{pmatrix}$$

Nun bildet man für die Zeilen $i = 1, \dots, n-1$ für alle Werte von $k = i+1, i+2, \dots, n$ einen Multiplikator $m_{ki} = a_{ki}/a_{ii}$. Durch Ausführen der Operation $(E_k - m_{ki}E_i)$ wird der Koeffizient a_{ki} gleich Null. Nach dem Bilden aller möglichen Multiplikatoren und der genannten Elimination aller Koeffizienten a_{ki} mit $i = 1, \dots, n-1$ und $k = i+1, i+2, \dots, n$ erhält man die folgende Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ 0 & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & a_{nn} & b_n \end{pmatrix}$$

Die a_{ij} und b_i sind andere als zuvor! Diese kleine Schlamperei ist branchenüblich. Jetzt ist es möglich, den unbekannten Vektor $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ zu berechnen. Durch Rückwärtssubstitution erhält man:

$$x_n = \frac{b_n}{a_{nn}}$$

$$x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}}$$

$$x_i = \frac{b_i - a_{i,n}x_n - a_{i,n-1}x_{n-1} - \ldots - a_{i,i+1}x_{i+1}}{a_{ii}} = \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}}$$

Das beschriebene Verfahren funktioniert jedoch nur unter der Voraussetzung $a_{ii} \neq 0$, da ansonsten der Multiplikator nicht definiert ist. Ist das Element $a_{ii} = 0$ (Pivotelement) null, durchsucht der Algorithmus die i-te

Spalte der Matrix von der i-ten Zeile an abwärts nach dem ersten Element ungleich null. Wird ein solches Element a_{ji} gefunden, vertauscht man die i-te und j-te Zeile der Matrix, und der Algorithmus wird wie vorher fortgesetzt. Existiert kein solches Element a_{ij}, so besitzt das Gleichungssystem keine eindeutige Lösung, und der Algorithmus bricht ab.

Abschließend sei noch eine Verbesserung des Gaußschen Algorithmus vorgestellt. Wie anfangs erwähnt, ist eine Lösung, die mit dem Gaußschen Algorithmus gefunden wurde, nur dann exakt, wenn mit einer exakten Arithmetik gerechnet wurde. Dies ist auf einem Computer nicht möglich. Hauptsächlich werden einmal entstandene Rundungsfehler durch Multiplikation mit dem Multiplikator fortgepflanzt. Ist also der Multiplikator sehr groß, so wachsen auch die Rundungsfehler stark. Es liegt damit nahe, mit kleinen Multiplikatoren zu rechnen, also a_{ii} möglichst groß zu wählen. Damit wird auch die Vergrößerung des Rundungsfehlers bei der Rücksubstitution verkleinert, wenn bei der Berechnung der x_i durch sehr kleine a_{ii} dividiert wird. Um ein möglichst großes a_{ii} zu erhalten, durchsucht der Algorithmus die i-te Spalte der Matrix von der i-ten Zeile an abwärts nach dem größtem Anfangselement, auch dann, wenn $a_{ii} \neq 0$ ist. Die Zeile mit dem größten Anfangselement vertauscht man anschließend mit der i-ten Zeile. Mit dieser Verbesserung ist es möglich, die meisten Gleichungssysteme hinreichend exakt zu lösen. Es können aber auch Situationen auftreten, in denen dieses Verfahren (Gaußscher Algorithmus mit Spaltenmaximumsstrategie) nicht funktioniert.

Eine weitere Verbesserung des Verfahrens besteht darin, die Zeilen der Matrix nicht einfach nach ihrem größten Anfangselement, sondern nach dem größten Anfangselement in Relation zu den übrigen Zeilenelementen zu durchsuchen. Hierzu definiert man sich einen Skalierungsfaktor $s_k = \max|a_{kj}|$ mit $k \geq i$ und $j = i + 1, i + 2, .., n$. Im nächsten Schritt berechnet man für jede Zeile den Quotienten aus Anfangselement der Zeile und zugehörigem Skalierungsfaktor. Im nächsten Schritt vertauscht das Verfahren die Zeile mit dem größten Quotienten mit der i-ten Zeile. Das Suchen der zu vertauschenden Zeile und das schlußendliche Vertauschen der i-ten Zeile mit der gefundenen Zeile nennt man Pivotierung. Da man die Anfangselemente der Zeilen relativ zu den übrigen Zeilenelementen betrachtet, heißt dieses verbesserte Verfahren Gaußscher Algorithmus mit relativer Spaltenmaximumsstrategie. Dieser Algorithmus löst die allermeisten Gleichungssysteme hinreichend exakt. Die Zeitkomplexität hierfür liegt in der Größenordnung $O(n^3)$: *Ein Vielfaches einer Zeile zu einer anderen zu addieren sind größenordnungsmäßig n Operationen. Das ist für $n(n-1)/2$ Elemente durchzuführen, nämlich alle unterhalb der Hauptdiagonalen (von links oben nach rechts unten). Das macht die 3 im Exponenten. Alle anderen Akte (Maximum berechnen, Zeilen vertauschen, Rückwärtssubstituieren) fallen demgegenüber nicht ins Gewicht.*

Das Gauß-Seidel-Verfahren

Gegeben ist ein lineares Gleichungssystem der Form $A \cdot \vec{x} = \vec{b}$, wobei A abermals eine $(n \cdot n)$-Matrix ist und $\vec{x}$ und $\vec{b}$ Vektoren sind. Um dieses Gleichungssystem iterativ zu lösen, überführt man es zunächst in die Form $\vec{x} = T \cdot \vec{x} + \vec{c}$. Hierzu zerlegt man A in drei verschiedene Matrizen.

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{nn} \end{pmatrix} - \begin{pmatrix} 0 & \dots & \dots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -a_{n1} & \dots & -a_{n,n-1} & 0 \end{pmatrix} - \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1,n} \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

$$A \qquad = \qquad D \qquad - \qquad L \qquad - \qquad U$$

Damit ist

$$\begin{aligned} (D - L - U) \cdot \vec{x} = A \cdot \vec{x} &= \vec{b} \\ \iff D \cdot \vec{x} &= (L + U) \cdot \vec{x} + \vec{b} \\ \iff \vec{x} &= D^{-1} \cdot (L + U) \cdot \vec{x} + D^{-1} \cdot \vec{b} \\ \iff \vec{x} &= T \cdot \vec{x} + \vec{c} \end{aligned}$$

mit $T = D^{-1} \cdot (L+U)$, $\vec{c} = D^{-1} \cdot \vec{b}$. ($D^{-1}$ anzuwenden ist sehr einfach: Man dividiere durch die Diagonalelemente. Die müssen dafür natürlich ungleich 0 sein.) Mit der Gleichung $\vec{x} = T \cdot \vec{x} + \vec{c}$ ist nun eine Vektorfolge $\vec{x}^{(n+1)} = T \cdot \vec{x}^{(n)} + \vec{c}$ bestimmt. Mit einem Startwert, z. B. $\vec{x}^{(0)} = \vec{c}$, ist es nun möglich, eine Fixpunktiteration durchzuführen, indem man $\vec{x}^{(1)}, \vec{x}^{(2)}, \dots, \vec{x}^{(n)}$ berechnet. Im Grenzfall für $n \to \infty$ soll $\vec{x}^{(n)} \to \vec{y}$ konvergieren, wobei $\vec{y}$ die exakte Lösung des Gleichungssystems sein soll. Dies ist dann der Fall, wenn T eine Kontraktion ist, und das ist dann der Fall, wenn der maximale Eigenwert von T betragsmäßig kleiner als eins ist.

Das beschriebene Verfahren läßt sich noch verbessern. Bisher berechnet man in jedem Schritt die Komponenten des nächsten Glieds der Vektorfolge $\vec{x}^{(k)}$. Anschließend setzt das Verfahren dieses neue Glied

in die Iterations- gleichung ein und berechnet das Nachfolgeglied $\vec{x}^{(k+1)}$. Die Verbesserung des Verfahrens besteht darin, daß nicht mehr alle Komponenten des neuen Glieds der Vektorfolge ausschließlich mit Hilfe der Komponenten des Vorgängerglieds berechnet werden, sondern die bereits berechneten Komponenten des neuen Glieds $\vec{x}^{(k+1)}$ zur Berechnung der übrigen Komponenten des neuen Vektors verwendet werden. Die beschriebene Verbesserung beschleunigt die Konvergenz des Verfahrens erheblich, teilweise um den Faktor 2.

Literatur: J. D. Faires, R. L. Burden: Numerische Methoden, Näherungsverfahren und ihre prakische Anwendung, Spektrum Akademischer Verlag.

Hyperbolische Differentialgleichungen und ihre numerische Behandlung

(Christine Rogg)

Charakteristiken und Unstetigkeiten

Wir beginnen mit einem Beispiel für eine nichtlineare hyperbolische Differentialgleichung:

$$\partial_t u(x,t) + \partial_x \frac{u^2(x,t)}{2} = 0 \ \ , u(x,0) = u_0(x) \tag{1}$$

Wir gehen davon aus, dass wir eine glatte, d. h. stetig differenzierbare Lösung $u \in C^1(R \times R^+)$ haben. Dann betrachten wir folgendes Anfangswertproblem für $\gamma(t)$:

$\gamma'(t) = u(\gamma(t),t)$, $\gamma(0) = a$, $\gamma \in C^1(]0,T[) \cap C^0([0,T])$

Dann gilt:

$$\frac{d}{dt}u(\gamma(t),t) = \gamma'(t)\partial_x u + \partial_t u = u(\gamma(t),t)\partial_x u(\gamma(t),t) + \partial_t u(\gamma(t),t) = \partial_x \frac{u^2}{2} + \partial_t u = 0.$$

Also ist $u(\gamma(t),t) = \gamma'(t) = \text{const}$. Demzufolge sind die Kurven von $\gamma(t)$ Geraden in der (t,x)-Ebene. Diese Geraden heißen Charakteristiken. Weil u entlang der Charakteristiken konstant ist, gilt: $\gamma'(t) = u(\gamma(t),t) = u(\gamma(0),0) = u_0(\gamma(0))$. Die Steigung ist also durch die Anfangswerte u_0 gegeben.

Jetzt kehren wir zu unserer Ursprungsgleichung (1) zurück. Ihre Anfangswerte seien folgendermaßen definiert:

$$u_0(x) = \begin{cases} 1 & \text{für } x \le -1 \\ 0 & \text{für } x(0) \ge 1 \\ \frac{-x}{2} + \frac{1}{2} & \text{für } -1 < x < 1 \end{cases}$$

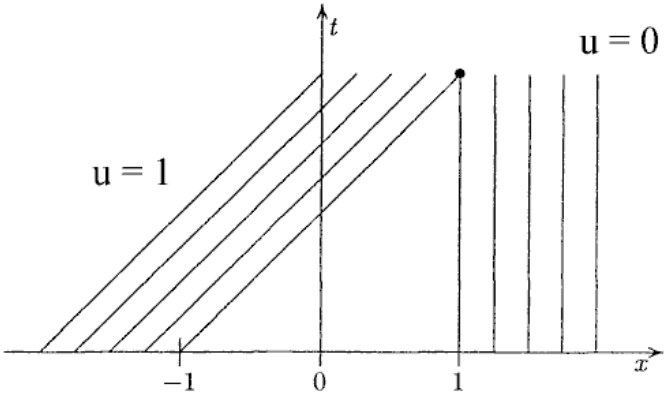
Daraus ergibt sich

$$u_0(\gamma(0)) = u(\gamma(t),t) = \gamma'(t) = \begin{cases} 1 & \text{für } \gamma(0) \le -1 \\ 0 & \text{für } \gamma(0) \ge 1 \\ \frac{-\gamma(0)}{2} + \frac{1}{2} & \text{für } -1 < \gamma(0) < 1 \end{cases}$$

Die Charakteristiken verlaufen dann wie in der Zeichnung rechts.

Bei der Wahl dieser Anfangsbedingungen erhalten wir also Charakteristiken, die sich schneiden können. Da die Lösungen der Gleichung (1) entlang der Charakteristiken konstant ist, bedeutet dies, dass die Lösung dort unstetig ist.

Weniger problematisch ist die Lösung linearer hyperbolischer Differentialgleichungen. Man betrachte beispielsweise



$$\partial_t u + a\partial_x u = 0, a = \text{const}. \text{ mit der Anfangsbedingung } u(x,0) = u_0(x)\ . \tag{2}$$

γ' ist in diesem Fall gleich a, d. h. die Steigung der Charakteristiken ist unabhängig von u_0 und damit konstant. Es kann also nicht zu Unstetigkeiten kommen, da alle Charakteristiken parallel zueinander sind.

Diese hyperbolische Differentialgleichung läßt sich als Transportgleichung auffassen, deren Lösung gegeben ist durch $u(x,t) = u_0(x - at)$. Dies bedeutet, dass die Lösung sich als Verschiebung der Anfangswerte mit der Geschwindigkeit a ergibt.

Charakteristiken lassen sich auch für die allgemeine hyperbolische Differentialgleichung $\partial_t u + \partial_x f(u) = 0$ bestimmen. Dabei soll gelten:

$t \geq 0,\ x \in R,\ u(x,0) = u_0(x),\ f \in C^1$ (d. h. einmal stetig differenzierbar). Es sei $\gamma'(t) = f'(u(\gamma(t),t))$ und $\gamma(0) = a$. Mit $u \in C^1$ gilt: $\frac{d}{dt}u(\gamma(t),t) = \partial_t u + \gamma'\partial_x u = \partial_t u + \partial_x f(u) = 0$. Demzufolge ist u konstant entlang der Charakteristik $\gamma(t)$, das gleiche gilt dann auch für $f'(u_0(a))$ und daher für $\gamma'(t)$.

Damit haben wir gezeigt, dass es bei hyperbolischen Differentialgleichungen zu Unstetigkeiten kommen kann, d. h. es existieren nicht immer globale Lösungen auf $[0,T]$. Dennoch ist es möglich, eine lokale Lösung in der Zeit zu finden, indem man den Endzeitpunkt T klein genug wählt.

Zur numerischen Lösung

Im Folgenden betrachten wir wieder $\partial_t u + a\partial_x u = 0$. Zunächst definieren wir einige Werte, die wir gleich verwenden werden:

$$u_k^n := u(k\Delta x, n\Delta t)\ ; \lambda := \frac{\Delta t}{\Delta x};\ Eu_k^n := u_{k+1}^n$$

Es folgt nach Taylor:

$$\frac{u_k^{n+1} - u_k^n}{\Delta t} + a\frac{u_{k+1}^n - u_k^n}{\Delta x} + O(\Delta x, \Delta t) = 0$$

Als numerisches Verfahren wird daher im Folgenden

$$u_k^{n+1} - u_k^n + a\lambda(u_{k+1}^n - u_k^n) = 0$$

verwendet. Dies lässt sich wie folgt umformulieren:

$$\begin{aligned} u_k^{n+1} &= u_k^n - a\lambda E u_k^n + a\lambda u_k^n \\ \Leftrightarrow u_k^{n+1} &= (1 + a\lambda - a\lambda E)u_k^n \\ \Rightarrow u_k^n &= (1 + a\lambda - a\lambda E)^n u_k^0 \end{aligned}$$

An diesem Beispiel will ich vorrechnen, dass bei unbedachter Diskretisierung großer Unfug herauskommen kann. Dafür benötigen wir den binomischen Lehrsatz. Dieser lautet:

$$(x+y)^n = \sum_{m=0}^{n} \binom{n}{m} x^{n-m}y^m \text{ mit } \binom{n}{m} = \frac{n!}{m!(n-m)!}$$

In unserem Fall sei $x = -a\lambda E$ und $y = 1 + a\lambda$. Der binomische Satz funktioniert nämlich auch noch, wenn eine der Beteiligten (in diesem Falle x) nicht eine Zahl, sondern ein linearer Operator ist. Damit ist

$$\begin{aligned} u_k^n &= \sum_{m=0}^{n} \binom{n}{m} (1+a\lambda)^m (-a\lambda E)^{(n-m)} u_k^0 \\ &= \sum_{m=0}^{n} \binom{n}{m} (1+a\lambda)^m (-a\lambda)^{(n-m)} u_{k+(n-m)}^0 \end{aligned}$$

Da $n \geq m$ ist, gilt $n - m \geq 0$. D. h., es werden nur Werte rechts von x_k verwendet, um u_k^n zu bestimmen. Die exakte Lösung aber lautet $u_0(x_k - at_n)$, das heißt, die echten Werte kommen von links. Somit erweist dieses Verfahren sich als ungeeignet. Je nachdem, ob die Steigung der Charakteristik positiv oder negativ ist, müssen wir also $\partial_x u$ unterschiedlich diskretisieren:

$$a > 0 : \partial_x u(x_k, t_n) \to \frac{u(x_k, t_n) - u(x_{k-1}, t_n)}{\Delta x}$$

$$a < 0 : \partial_x u(x_k, t_n) \to \frac{u(x_{k+1}, t_n) - u(x_k, t_n)}{\Delta x}$$

Diese Fallunterscheidung fällt bei konstantem a nicht sehr schwer, aber in der Regel haben wir Gleichungen der Form $\partial_t u + f'(u(x,t))\partial_x u = 0$. Da f' nun von u abhängig ist, kann es in Abhängigkeit von x und t das Vorzeichen ändern, weshalb dieses ständig überprüft werden muss.

Diese Vorgehensweise, immer in die Richtung zu gucken, aus der die Werte kommen, nennt man *upwinding* (Gegenwind-Diskretisierung).

Zur Stabilitätsbedingung

Dieses Mal betrachten wir die richtige Diskretisierung von (1) mit $a > 0$:

$$u_j^{n+1} = u_j^n - \lambda(u_j^n - u_{j-1}^n) \text{ mit } n \in N \text{ und } \lambda = a\frac{\Delta t}{\Delta x}$$

mit der Anfangsbedingung $u_j^0 = \begin{cases} 1 & \text{für } x \leq 0; \\ 0 & \text{für } x > 0. \end{cases}$

Dann lautet die Behauptung: $u_n^n = \lambda^n$ und $u_j^n = 0$ für $j \geq n+1$.

Beweisen wollen wir dies durch vollständige Induktion.

Induktionsanfang:
$n = 0 : \quad u_0^0 = 1$ und $u_1^0 = 1$ nach Definition der Anfangsbedingung

Induktionsschritt:
Unter der Voraussetzung, dass die Behauptung für n gilt, ist sie für $n+1$ zu beweisen.

$$u_{n+1}^{n+1} = u_{n+1}^n - \lambda(u_{n+1}^n - u_n^n) = u_{n+1}^n(1-\lambda) + \lambda u_n^n = 0 + \lambda \cdot \lambda^n = \lambda^{n+1}$$

Damit ist der erste Teil ($u_n^n = \lambda^n$) bewiesen. Ferner gilt für $j \geq n+2$:

$$u_j^{n+1} = u_j^n - \lambda(u_j^n - u_{j-1}^n) = u_j^n(1-\lambda) + \lambda u_{j-1}^n = 0$$

Damit ist auch der zweite Teil ($u_j^n = 0$ mit $j \geq n+1$) und somit die gesamte Behauptung bewiesen.

Wie bereits erwähnt, lautet die exakte Lösung $u(x,t) = u_0(x-t)$, somit liegen die Werte von u zwischen 0 und 1. Das ist aber nur gegeben, wenn $\lambda \leq 1$ ist, denn sonst würde λ^n und damit u_n^n unbeschränkt wachsen. Wählt man λ aber zu klein, wird die Lösung sehr schnell verschmiert. Deshalb ist das Verhältnis $\lambda = a\frac{\Delta t}{\Delta x}$ bei komplizierten Funktionen mit Bedacht zu wählen. Die Beschränkung $a\frac{\Delta t}{\Delta x} \leq 1$ heißt Stabilitätsbedingung.

Abschließend lässt sich sagen, dass es bei hyperbolischen Differentialgleichungen zu Unstetigkeiten kommen kann, man das Differenzenverfahren sorgfältig auswählen muss (upwinding!) und außerdem eine Stabilitätsbedingung berücksichtigt werden muss.

Literatur: D. Kröner, Numerical Schemes for Conservation Laws, Wiley-Teubner, S. 1, 2, 15–17, 32, 33, 92

Herleitung der Navier-Stokes-Gleichungen

(Matthias Klotz)

Eine wichtige Stelle in der heutigen Physik nimmt die Beschreibung von dreidimensionalen reaktiven Strömungen in Flüssigkeiten und Gasgemischen ein. Das Ziel dieses Textes besteht im Wesentlichen darin, physikalische Gesetze, die derartige Vorgänge beschreiben, in Differentialgleichungen auszudrücken: in den sogenannten Navier-Stokes-Gleichungen. Dabei handelt es sich bei diesen Gesetzen um die drei Erhaltungsgleichungen für *Masse*, *Impuls* und *Energie*.

Im Folgenden stelle ich eine allgemeine Erhaltungsgleichung auf, um sie dann auf die Spezialfälle für jene drei physikalischen Größen anzuwenden. Dabei stellt man sich ein gasgefülltes Volumen als Kontinuum vor, was wegen der großen Anzahl der in ihm vorhandenen Moleküle auch gerechtfertigt ist. Dies ermöglicht den Umgang mit infinitesimalen Größen: Differentialquotienten und Integrale.

Man denke sich einen beliebig geformten Bereich Ω im dreidimensionalen Raum mit dem Volumen V und dessen Rand $\partial\Omega$ mit der Oberfläche S, der stückweise stetig differenzierbar sein soll. Zu jedem Zeitpunkt t lässt sich eine physikalische Größe $F(t)$ für Ω (z. B.: Gesamtmasse, Gesamtimpuls, ...) angeben. Man kann sie außerdem mit Hilfe der ihr zugehörigen Dichten $f(\vec{r},t)$ beschreiben (mit $\vec{r}$ = Ortsvektor), indem man nämlich über Ω integriert (mit Dichte ist nicht zwingend die Massendichte gemeint, sondern die Größe $F(t)$ pro Volumen):

$$F(t) = \int_\Omega f(\vec{r},t)dV$$

Zur Verbesserung der Lesbarkeit werde $f(\vec{r},t)$ nur noch als f dargestellt. Man vergesse also nicht, dass f von der jeweiligen Stelle im Bereich Ω und von der Zeit t abhängig ist. Die Änderung der Größe F mit der Zeit t wird durch folgenden Ausdruck beschrieben:

$$\frac{\partial F}{\partial t} = \int_\Omega \frac{\partial f}{\partial t}dV$$

Eine derartige Änderung kann durch drei verschiedene Prozesse stattfinden:

1.) Die Größe F strömt durch die Oberfläche in das Volumen hinein oder aus ihm heraus. Die Strömung selbst wird durch einen Vektor namens Stromdichte beschrieben: $\vec{\Phi}_f$ gibt an, wieviel von der Größe F pro Zeiteinheit durch eine Flächeneinheit strömt. Bei der Strömung durch die Oberfläche $\partial\Omega$ kommt es nur auf den Anteil von $\vec{\Phi}_f$ an, der senkrecht zur Oberfläche durchfließt; deswegen bildet man das Skalarprodukt $\vec{\Phi}_f \cdot \vec{n}$, wobei $\vec{n}$ die äußere Normale ist, d. h. der Vektor der Länge 1, der senkrecht auf der Oberfläche steht und nach außen zeigt. Dabei wird das Skalarprodukt negativ, wenn der Vektor $\vec{\Phi}_f$ auf die Oberfläche von Ω zeigt. Eine Veränderung der Dichte auf das Innere von Ω bezogen ist also durch $-\vec{\Phi}_f \cdot \vec{n} dS$ gegeben.

2.) Im Inneren des Gebietes Ω liegt eine Quelle (bzw. Senke) von F. So beschreibt man z. B. chemische Reaktionen: Bei der Verbrennung gibt es eine Senke für Sauerstoff und Brennstoff und eine Quelle für (u. a.) Kohlendioxid. Dargestellt wird das durch einen Quellterm q_f im Inneren des Volumenelementes, wobei q_f die pro Zeit t und Volumeneinheit gebildete Menge an F beschreibt.

3.) Fernwirkung (z. B. Gravitation oder Wärmestrahlung) beeinflussen von außerhalb das Innere von Ω, zu beschreiben durch einen Fernwirkungsterm s_f, der für die pro Zeit und Volumeneinheit gebildete Menge an F steht.

Für die einzelnen Prozesse ergibt sich jeweils die Änderung für F durch jeweilige Integration über die Oberfläche von $\partial\Omega$ bzw. über Ω selbst. Die gesamte Änderung von F pro Zeit t ergibt sich durch Summieren der einzelnen Integrale:

$$\frac{dF}{dt} = \int_\Omega \frac{\partial f}{\partial t} dV = -\int_{\partial\Omega} \vec{\Phi}_f \cdot \vec{n} dS + \int_\Omega q_f dV + \int_\Omega s_f dV$$

Man möchte das Integral über $\partial\Omega$ in ein Integral über Ω selbst umformen. Dazu nutzt man den Gauß'schen Integralsatz: Der Fluss von $\vec{\Phi}_f$ über $\partial\Omega$ ist gleich dem Volumenintegral von $\operatorname{div} \vec{\Phi}_f$ über Ω. *Das ist nichts Mystisches! Man drückt die Randwerte einer Funktion aus durch ein Integral über die Ableitung der Funktion im Inneren:* $f(b) - f(a) = \int_a^b f'(x) dx$*. Hauptsatz der Differential- und Integralrechnung! Dasselbe in mehreren Dimensionen ist natürlich ein bisschen komplizierter.* Hieraus folgt:

$$\int_\Omega \frac{\partial f}{\partial t} dV + \int_\Omega \operatorname{div} \vec{\Phi}_f dV = \int_\Omega q_f dV + \int_\Omega s_f dV$$

Das Gebiet Ω wurde zu Anfang unbestimmt gewählt. Wir haben hier also eine Gleichung für Integrale unabhängig vom Integrationsgebiet. Das ist jedoch nur dann möglich, wenn bereits die Integranden gleich sind:

$$\frac{\partial f}{\partial t} + \operatorname{div} \vec{\Phi}_f = q_f + s_f$$

Das ist die *allgemeine Erhaltungsgleichung*, die ich im Folgenden für Masse, Impuls und Energie an Stelle von F anwende.

Erhaltung der Gesamtmasse

Die Größe F ist in diesem Fall durch die Gesamtmasse m des Systems gegeben. Die entsprechende Dichte f_m ist die Massendichte ρ, die Massenstromdichte $\vec{\Phi}_m$ ergibt sich als Produkt aus der Strömungsgeschwindigkeit $\vec{v}$ und der Dichte ρ (denn je größer die Geschwindigkeit, desto mehr fließt pro Zeit t, entsprechend gilt dies für die Dichte). Masse kann weder durch Quellen im Inneren von Ω noch durch Fernwirkung entstehen. Somit sind q_m und s_m beide 0. Durch Einsetzen in die allgemeine Erhaltungsgleichung erhält man die *Massenerhaltungsgleichung*:

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \vec{v}) = 0$$

Erhaltung des Impulses

Die Größe F ist in diesem Fall durch den Gesamtimpuls des Systems gegeben. Die entsprechende Dichte $f_{m\vec{v}}$ ist die Impulsdichte $\rho\vec{v}$, die Impulsstromdichte $\vec{\Phi}_{m\vec{v}}$ (hier ein Tensor) ergibt sich zu $\vec{\Phi}_{m\vec{v}} = \rho\vec{v} \otimes \vec{v} + \bar{\bar{p}}$. Dabei bezeichnet $\rho\vec{v} \otimes \vec{v}$ den konvektiven Anteil und $\bar{\bar{p}}$ die Impulsänderung durch Druck- und Reibungskräfte (hydrostatischer und viskoser Anteil). Unter der Viskosität eines Gases bzw. einer Flüssigkeit hat man sich also seine Zähigkeit vorzustellen (die innere Reibung). Das dyadische Produkt $\vec{v} \otimes \vec{v}$ ergibt folgenden Tensor:

$$\vec{v} \otimes \vec{v} = \begin{pmatrix} v_1 v_1 & v_1 v_2 & v_1 v_3 \\ v_2 v_1 & v_2 v_2 & v_2 v_3 \\ v_3 v_1 & v_3 v_2 & v_3 v_3 \end{pmatrix} \quad \text{mit } \vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

Der Drucktensor $\bar{\bar{p}}$ ergibt sich empirisch aus einer großen Anzahl von Untersuchungen:

$$\bar{\bar{p}} = p\bar{\bar{E}} + \bar{\bar{\Pi}}$$

Dabei ist $\bar{\bar{E}}$ der Einheitstensor und p der hydrostatische Druck. Während $p\bar{\bar{E}}$ den hydrostatischen Anteil von $\bar{\bar{p}}$ beschreibt, tut $\bar{\bar{\Pi}}$ dies für den viskosen Anteil. Aus der kinetischen Theorie für verdünnte Gase ergibt sich weiterhin der Zusammenhang:

$$\bar{\bar{\Pi}} = -\mu[(\operatorname{grad}\vec{v}) + (\operatorname{grad}\vec{v})^T] + (\frac{2}{3}\mu - \kappa)\ (\operatorname{div}\vec{v})\ \bar{\bar{E}}$$

Während μ (my) für die mittlere dynamische Viskosität steht, bezeichnet κ (kappa) die Volumenviskosität. Für einatomige Gase gilt $\kappa = 0$.

Der Ausdruck $\operatorname{grad}\vec{v}$ ist ein Tensor:

$$\operatorname{grad}\vec{v} = \begin{pmatrix} \frac{\partial v_1}{\partial x} & \frac{\partial v_2}{\partial x} & \frac{\partial v_3}{\partial x} \\ \frac{\partial v_1}{\partial y} & \frac{\partial v_2}{\partial y} & \frac{\partial v_3}{\partial y} \\ \frac{\partial v_1}{\partial z} & \frac{\partial v_2}{\partial z} & \frac{\partial v_3}{\partial z} \end{pmatrix} \quad \text{mit } \vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

Der Term $(\operatorname{grad}\vec{v})^T$ ist der transponierte Tensor von $\operatorname{grad}\vec{v}$. Er entsteht durch eine Vertauschung von Zeilen und Spalten.

Der Quellterm $q_{m\vec{v}}$ ist 0, da im Inneren des Systems keine Kräfte entstehen, die eine Impulsänderung zur Folge hätten. Es existiert jedoch eine Fernwirkung $s_{m\vec{v}}$, nämlich die Gravitationskraft pro Volumen, was durch $\rho\vec{g}$ gegeben ist. Durch Einsetzen in die allgemeine Erhaltungsgleichung erhält man die *Impulserhaltungsgleichung*:

$$\frac{\partial\ (\rho\vec{v})}{\partial t} + \operatorname{div}(\rho\vec{v}\otimes\vec{v}) + \operatorname{div}\bar{\bar{p}} = \rho\vec{g}$$

Erhaltung der Energie

Die Größe F ist in diesem Fall durch die gesamte Energie des Systems gegeben, d. h. durch die Summe aus der potentiellen, der kinetischen und der inneren Energie (kinetische Energie: Schwerpunktsbewegung der Moleküle bzw. Translation; innere Energie: Schwingung und Rotation der einzelnen Moleküle). Die entsprechende Dichte f_e ist somit die Energiedichte ρe, wobei e für die jeweilige spezifische Gesamtenergie steht (d. h. Energie pro Masse). Da ρe für die Energie pro Volumen steht und da diese sich aus den oben genannten drei Energieformen zusammensetzt, gilt:

$$\rho e = \rho u + \frac{1}{2}\rho|\vec{v}|^2 + \rho G$$

Dabei steht u für die spezifische innere Energie, G für das Potential der Energie (potentielle Energie pro Masse). Die Energiestromdichte $\vec{\Phi}_e$ (hier wieder ein Vektor) ergibt sich zu $\vec{\Phi}_e = \rho e\vec{v} + \bar{\bar{p}}\vec{v} + \vec{j}_q$. Der Ausdruck $\rho e\vec{v}$ steht analog zu den vorherigen Erhaltungsgleichungen für den konvektiven Anteil, $\bar{\bar{p}}\vec{v}$ hingegen beschreibt die Reibungsarbeit, die aufgrund der Viskosität verrichtet wird. Die Wärmestromdichte ist durch $\vec{j}_q$ gegeben, was den Energietransport durch Wärmeleitung beschreibt. Sie ergibt sich zu $\vec{j}_q = -\lambda\operatorname{grad}T$, wobei λ für den Wärmeleitfähigkeitskoeffizient, T für die Temperatur steht. Der Vektor $\operatorname{grad}T$ zeigt in die Richtung, in der (in der Umgebung) die Temperatur am stärksten ansteigt. Ein Energietransport findet also in die entgegengesetzte Richtung statt, wodurch das Minuszeichen im Ausdruck $-\lambda\operatorname{grad}T$ zu Stande kommt. Der Quellterm q_e ist 0, da im Inneren des Systems keine Energie entsteht. Es existiert jedoch eine Fernwirkung s_e, die nämlich in der Wärmestrahlung liegt. Sie ist durch q_r gegeben, was die Wärmeproduktion pro Volumen und Zeit beschreibt. Es gilt also $[q_r] = Jm^{-3}s^{-1}$ (d. h. q_r hat die Dimension Joule pro Kubikmeter und Sekunde). Durch Einsetzen in die allgemeine Erhaltungsgleichung erhält man die *Energieerhaltungsgleichung*:

$$\frac{\partial(\rho e)}{\partial t} + \operatorname{div}(\rho e\vec{v} + \bar{\bar{p}}\vec{v} - \lambda\operatorname{grad}T) = q_r$$

Das Gleichungssystem für die Navier-Stokes-Gleichungen

Die nun drei aufgestellten Differentialgleichungen zum Berechnen von Strömungen (Masse-, Impuls- und Energieerhaltungsgleichung) lassen sich wie folgt in einem Gleichungssystem zusammenfassen:

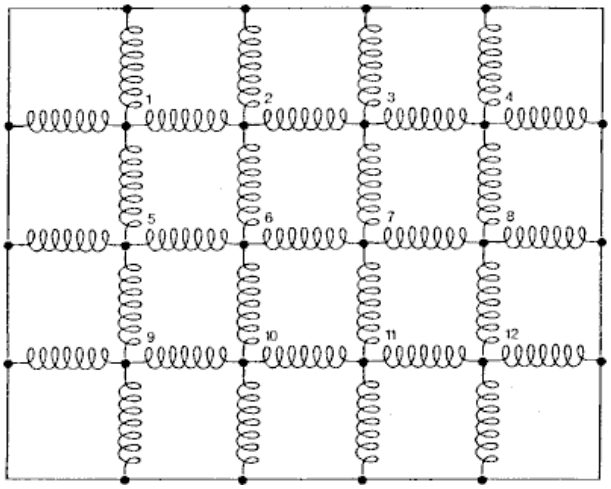
$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho\vec{v} \\ \rho e \end{pmatrix} + \operatorname{div}\begin{pmatrix} \rho\vec{v} \\ \rho\vec{v}\otimes\vec{v} + \bar{\bar{p}} \\ \rho e\vec{v} + \bar{\bar{p}}\vec{v} - \lambda\operatorname{grad}T \end{pmatrix} = \begin{pmatrix} 0 \\ \rho\vec{g} \\ q_r \end{pmatrix}$$

Literatur: Jürgen Warnatz und Ulrich Maas: Technische Verbrennung, Springer-Verlag Berlin Heidelberg 1993, S. 135–143: Kapitel 11

Mehrgitterverfahren

(Silja Kinnebrock)

Eine Vielzahl der Phänomene der mathematischen Physik, wie beispielsweise Schwingungen von Bauwerken, Wellen und Gewässern, lassen sich auf das Lösen großer, dünnbesetzter, linearer Gleichungssysteme zurückführen. Um diese möglichst schnell und einfach lösen zu können, sind Mehrgitterverfahren hilfreich. Sie bauen auf Iterationsverfahren auf und erweitern sie. So können Mehrgitterverfahren den Lösungsprozess um das Hundertfache beschleunigen.



Unser Übungsbeispiel ist, die Eigenschwingungen des Bodensees zu berechnen. *Die Wasseroberfläche ist eine sehr kontinuierliche Angelegenheit. Wir diskretisieren sie mit finiten Differenzen, wie Jörn-Thorsten uns das vorgerechnet hat; aber diesmal motivieren wir das Ganze physikalisch. Wir betrachten ein System, das „von Natur aus“ diskret ist: den Federrost eines Bettgestells. Dann lassen wir die Gitterweite gegen 0 gehen – wie man das beim physikalischen Modellieren von schwingenden Saiten und ähnlichen Dingen ohnehin macht.* Wir ersetzen die Auslenkung des Bodensees durch die Auslenkung eines Federrostes unter der Belastung durch den Schläfer. Wir stellen uns den Federrost als ein gleichmäßiges Rechtecksgitter aus Punkten vor, die mit ihren nächsten Nachbarn (rechts, links, oben, unten) durch Federn gekoppelt sind. Dabei stellt die Funktion $u(x, y)$, die Auslenkung des Federrostes gegen die Ebene des Bettrahmens dar. Nun gilt es die Kraft auf jeden einzelnen Punkt zu berechnen. Jede der an einem Punkt ansetzenden Federn trägt eine Kraft bei, die proportional der Differenz der Auslenkungen in ihren beiden Endpunkten ist. Es ergibt sich, wenn die Proportionalitätsfaktoren vernachlässigt werden:

$$F = (u_{\text{rechts}} - u) + (u_{\text{links}} - u) + (u_{\text{oben}} - u) + (u_{\text{unten}} - u) = u_{\text{rechts}} + u_{\text{links}} + u_{\text{oben}} + u_{\text{unten}} - 4u$$

Bei dieser Gleichung handelt es sich um eine Differenzengleichung. Für $F = 0$, also den Gleichgewichtszustand, ergibt sich:

$$u = (u_{\text{rechts}} + u_{\text{links}} + u_{\text{oben}} + u_{\text{unten}})/4$$

Man erkennt, dass die Auslenkung jedes Punktes das Mittel der Auslenkungen der vier Nachbarpunkte ist.

Als nächstes betrachten wir den Grenzfall eines unendlich fein gesponnenen Gitters, was über die Zwischenstufe

$$(1/h^2)(u_{\text{rechts}} - 2u + u_{\text{links}}) + (1/h^2)(u_{\text{oben}} - 2u + u_{\text{unten}})$$

(wobei h der Abstand zwischen zwei benachbarten Gitterpunkten ist) zum Laplace-Operator führt:

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

Der Laplace-Operator, der bereits vorher aufgetreten ist, dient nicht nur der mathematischen Modellierung von Schwingungs-, sondern auch von Diffusions- und Wärmeleitungsprozessen und ist auch in der Quantenmechanik (Schrödinger-Gleichung) sowie in der Elektrodynamik (Maxwell'sche Gleichungen) zu finden. Wir haben damit eine Differenzengleichung durch eine Differentialgleichung ersetzt und damit ein diskretes in ein kontinuierliches Problem verwandelt.

Für die numerische Berechnung muss man den umgekehrten Weg gehen und aus der Differentialgleichung wieder eine Differenzengleichung machen (diskretisieren). Dabei ist es günstig, wenn die Diskretisierung physikalisch interpretierbar ist wie in dem Beispiel mit dem Federrost. Das ist bei der Diskretisierung des Laplace-Operators mit zentralen Differenzen (vgl. den Beitrag von Jörn-Thorsten Paßmann) der Fall.

Zur Lösung der Differenzengleichungen benötigt man Verfahren zur Lösung großer, linearer und dünnbesetzer Gleichungssysteme. Zwar haben wir bereits das Gauß-Seidel-Verfahren sowie das Jacobi-Verfahren kennengelernt, allerdings verläuft die Konvergenz bei diesen Verfahren verhältnismäßig langsam; denn je größer n wird, desto mehr Schritte sind notwendig, um eine vorgegebene Genauigkeit zu erzielen. Physikalisch lässt sich dieser Sachverhalt durch den Wärmeausgleichsprozess plausibel machen: Wird ein Stab an beiden Enden auf verschiedenen Temperaturen konstant gehalten, so stellt sich nach einer gewissen Zeit eine stabile Temperaturverteilung ein. Bei diesem Vergleich steht die lokale Temperatur für die Auslenkung u, die Wärme,

die im Rahmen des Ausgleichsprozesses auf den Punkt zu oder von ihm wegfließt, repräsentiert die Korrektur durch ein numerisches Verfahren.

In jedem Zeitschritt findet ein Austausch von Wärme (oder auch Information) immer nur unter nächsten Nachbarn statt. Deswegen kann es sehr lange dauern, bis die Information über die Temperatur am Rande des Gebietes sich 100 Gitterpunkte weiter zu einem Punkt im Inneren des Gebietes herumgesprochen hat.

Die Konvergenzgeschwindigkeit eines Systems kann man zwar dadurch erhöhen, daß man die Korrektur jeweils mit einem geeigneten Faktor multipliziert (SOR-Verfahren), aber auch das bringt keine entscheidende Besserung; denn in erster Linie reduziert das Iterationsverfahren den Fehler nicht, sondern glättet ihn lediglich.

Hier setzt das Mehrgitterverfahren an, das durch gezieltes Wechseln von Gittern verschiedener Maschenweite eine enorme Konvergenzbeschleunigung erzielt. Man wendet das normale Iterationsverfahren zunächst im feinen Ausgangsgitter an und glättet in 2 bis 3 Schritten den Fehler. Die so erhaltenen Werte übertragen wir auf ein gröberes Gitter (Restriktion) und wiederholen das Iterationsverfahren. Das gröbere Gitter kann einfach die Teilmenge des feineren sein, bei der jede zweite Gitterzeile und -spalte gestrichen werden. Der Informationsaustausch findet jetzt über größere Entfernungen statt und ist dadurch effektiver. Außerdem geht auf dieser Stufe das Rechnen schneller, da die Anzahl der Gitterpunkte gesunken ist.

Wir restringieren wiederum und führen einige Iterationsschritte im gröbsten Gitter aus, wobei wir wieder einige Gitterpunkte unberücksichtigt lassen. Da die Anzahl der Variablen inzwischen ziemlich klein geworden ist, können wir die zugehörige Gleichung direkt (nicht iterativ) lösen. Die hier errechneten Werte werden beim Übergang in das mittlere Gitter (Prolongation) beibehalten. Die Werte der Gitterpunkte, die im gröbsten Gitter nicht berechnet worden sind, werden jetzt als Mittel der Nachbarpunkte berechnet. Ebenso verfährt man nach Übergang ins feinste Gitter, so dass letztendlich für alle Gitterpunkte ein approximierter Wert berechnet worden ist. Alle bisher durchgeführten Vorgänge auf den drei Gitterstufen bilden zusammen genau einen Schritt des Mehrgitterverfahrens. Typisch sind drei bis sieben Gitterstufen.

Das Mehrgitterverfahren ist wesentlich effizienter als das Gauß-Seidel- oder das Jacobi-Verfahren; denn aufgrund der flexiblen Gitterwahl ist ein besserer Informationsaustausch und damit eine schnellere Konvergenz gewährleistet. Auch ist der Mehraufwand nicht so groß, wie man anfangs vermutet; er liegt lediglich bei Faktor 3, da die Anzahl der zu berücksichtigenden Gitterpunkte bei allen Grobgittern zusammen meist kleiner als die gesamte Anzahl aller Gitterpunkte des feinsten Gitters ist.

Nun kommen wir wieder zurück auf die anfangs angesprochene Berechnung von Eigenschwingungen des Bodensees, die durch die sogenannten „Lange-Wellen-Gleichungen“ bzw. „Flachwassergleichungen“ beschrieben werden. Hierfür wurde ein eigens angepasstes Mehrgitterverfahren entwickelt, nach dem die Oberfläche des Sees so in Dreiecke zerlegt wird, dass die Uferlinie durch die entsprechenden Dreiecksseiten möglichst getreu wiedergegeben wird. Daraus entsteht ein verfeinertes Gitter, indem man jedes Dreieck unter Einführung neuer Gitterpunkte auf den Seitenmitten in vier kleinere teilt. Durch Wiederholungen dieser Verfahrensweise entsteht ein beliebig feines Gitter mit immer präziser dargestellter Uferlinie.

Die Praxis zeigt, dass Mehrgitterverfahren zur Verbesserung der Effektivität praxisnaher Berechnungen höchst effizient sind. Mit ihrer Hilfe können komplexe Probleme, die bisher Superrechnern vorbehalten waren, auf PC’s gelöst werden.

Literatur: G. Wittum: Mehrgitterverfahren, Spektrum der Wissenschaft, April 1990

Modellieren von Turbulenzen

(Sabine Schamberg)

Zum Bau von Flugzeugen ist es notwendig, Genaueres über Auftrieb, Drehmomente und den Luftwiderstand zu erfahren, wobei der Luftwiderstand besonders wichtig ist, da er einer der Hauptgründe für den Kraftstoffverbrauch ist. Luftwiderstand entsteht unter anderem wegen der in der Grenzschicht wenige cm um das Flugzeug herum herrschenden Turbulenz. Das ist eine unruhige, nicht laminare Strömung, die aus einzelnen kleinen Wirbeln, den sogenannten Eddies, besteht. Diese befördern die Luft an die Oberfläche des Flugzeugs. Dabei gibt sie einen Teil ihres Impulses an das Flugzeug ab, auf welches deshalb eine rückwärtsgerichtete Kraft wirkt. Diese Eddies bilden sich immer wieder neu, werden kleiner und dissipieren, d. h. sie werden durch innere Reibung in Wärme umgewandelt.

Um die Turbulenz zu verstehen und zu bestimmen, verwendet man eine Mischung aus numerischer Simulation und Windkanal-Tests. Erstere hat gewisse Vorteile, da Windkanal-Tests sehr teuer sind, weil man für jeden Versuch ein neues Modell braucht. Außerdem kommen Simulationen der Realität meistens näher als Versuche im Windkanal, weil der Windkanal Wände hat, an denen unerwünschte Turbulenzen entstehen können. Dazu kommt, dass im Windkanal keine Versuche zu Hyperschallflügen möglich sind. Trotzdem ist es wegen des

großen Rechenaufwands und der mangelhaften Theorie nicht möglich, ausschließlich numerische Simulationen zu verwenden.

Diese sieht theoretisch folgendermaßen aus: Ein gewisses Volumen in der Umgebung des gedachten Flugzeugs wird mit einem Punktgitter überzogen und die Zeit in diskrete Schritte eingeteilt. Danach werden die Gleichungen und die Anfangs- und Randbedingungen für den Computer umgeformt. Anschließend werden diese zusammen mit den Gitterkoordinaten in den Computer eingegeben. Daraus erhält man eine Lösung für den Druck und die Geschwindigkeit an jedem Gitterpunkt. Je mehr Punkte das Rechengitter hat, desto genauer ist die Lösung. Da die Navier-Stokes-Gleichungen nicht linear sind, entstehen Wirbel. Um die größten Wirbel zu erfassen, muss das Gitter sehr groß sein. Um die kleinsten Wirbel zu erfassen, muss das Gitter sehr fein sein. Deswegen braucht man entsetzlich viele Punkte. Erschwert wird die Berechnung durch die globale Abhängigkeit. Der Druck an einem Gitterpunkt ist von den Strömungen an einem anderen abhängig. Dieser Effekt kann sich weit gegen die Windrichtung fortpflanzen, weit entfernte Gitterpunkte müssen also auch berücksichtigt werden. Auch sind viele Eddies erheblich kleiner als die Gitterweite.

Die Natur einer Strömung wird in kompakter Form durch die Reynoldszahl Re charakterisiert. Sie beschreibt das Verhältnis zwischen den Trägheits- und den Reibungskräften. Hierbei ist v die Strömungsgeschwindigkeit, l die Länge des umströmten Gegenstandes, wobei diese nicht genau bestimmt werden kann, ρ die Dichte und η die Viskosität, die vom Material abhängt.

$$Re = \frac{vl\rho}{\eta}$$

An dieser Gleichung erkennt man, dass Re sowohl proportional zu v als auch zu l ist. Außerdem sieht man deutlich, dass bei geringer Viskosität die Reynolds-Zahl sehr groß ist, also starke Turbulenz vorhanden ist. Das Größenverhältnis zwischen den größten und den kleinsten Eddies entspricht ungefähr $Re^{3/4}$. Somit ist die Anzahl der Gitterpunkte zur genauen Simulation ca. $Re^{9/4}$ (in drei Raumdimensionen). Da für jeden Gitterpunkt der Druck und die Geschwindigkeit berechnet werden sollen, ergibt dieses viel zu viele Daten, und es ist unmöglich, es mit heutigen kommerziell erhältlichen Computern zu berechnen.

Es gibt allerdings auch machbare Verfahren, um dieses Problem zu lösen. Das Ad-hoc-Modell basiert auf Experimenten und vereinfachten Annahmen über Turbulenzen. Es verwendet nur Mittelungen der großen Eddies, versucht aber auch die kleinen zu berücksichtigen. Bei dieser Mittelung ist jedoch zu beachten, dass man physikalische Gesetze wie Energie- und Impulserhaltung einhält. Trotzdem sind diese Simulationen nur so gut wie ihre Modelle. Um dieses Modell zu verbessern, verwendet man die direkte Simulation, die natürlich nur für kleine Reynolds-Zahlen möglich ist. Trotzdem liefert sie Ergebnisse über die Natur der Turbulenz. Die large eddie simulation (LES) verbindet die beiden genannten Modelle miteinander. Die großen Eddies werden direkt simuliert, die kleinen durch Mittelungsmodelle erfasst. Dieses Verfahren wird z. B. in der Meteorologie und zur Berechnung von Strömungen in Verbrennungsmotoren verwendet.

Die durch die Strömungsrechnung gewonnenen Erkenntnisse werden u. a. zur Verbesserung von Flugzeugen verwendet. Auf diesem Weg wurden auch Riblets entdeckt. Dies sind v-förmige Rillen auf der Oberfläche, die die Eddies daran hindern, ihren Impuls an das Flugzeug abzugeben, und so den Luftwiderstand und den Treibstoffverbrauch verringern. Dieses Prinzip ist auch in der Haut von Haien zu finden. Zu dem gleichen Ergebnis führen auch Sensoren und Verstellmechanismen auf der Haut des Flugzeugflügels, die auf Druck- und Strömungsgeschwindigkeitsänderung reagieren. Dieses Prinzip wird bei den Delphinen vermutet.

Durch die Berechnungen versteht man nun auch endlich, wie die kleinen Vertiefungen (Dimples) im Golfball wirken. Durch sie wird Turbulenz erzeugt, die Strömung legt sich enger an den Ball an, und der Staudruck wird verringert. Dies führt dazu, dass ein solcher Ball $2\frac{1}{2}$mal weiter fliegt als ein glatter. Ein weiteres Anwendungsgebiet ist die Optimierung der Formen von Turbinenschaufeln, Einlassschlitzen, Brennkammern und Triebwerksgondeln.

Literatur: Parviz Moin, John Kim: Modellieren von Turbulenz, Spektrum der Wissenschaft, Dezember 1997

Visualisierung von Strömungsgrößen berechnet mit numerischen Verfahren

(Bastian Katz)

Wir haben bisher Methoden zur Approximation sowohl von skalaren Größen wie Druck, Temperatur oder Dichte als auch von vektoriellen Größen wie z. B. dem Geschwindigkeitsfeld einer Strömung in einem Raum kennengelernt. Dazu wird das betrachtete Volumen durch ein dreidimensionales Gitter dargestellt – in unseren Beispielen zunächst mit einem strukturierten Gitter, d. h. der Raum wird in Quader zerlegt. Die gesuchten Größen werden dann an den Gitterpunkten, d. h. an den Ecken der Quader, mit Hilfe numerischer Verfahren angenähert. Um diese Strömungsdaten letztendlich interpretieren zu können, müssen sie jedoch noch in geeigneter Art visualisiert werden, was bedeutet, die Darstellung auf möglichst aussagekräftige

Informationen zu reduzieren, da niemand sich in einer solchen Menge von Daten auf dem dreidimensionalen Gitter zurechtfinden könnte, ganz abgesehen davon, dass dem Betrachter Informationen aus dem Inneren des betrachteten Volumens nicht mehr zugänglich wären, wenn man versuchen würde, alles darzustellen. Wir werden hier zwei Visualisierungsmethoden vorstellen: Schnitt- und Isoflächen mit Farbverläufen oder Vektorfeld und Particle Tracing, zu deutsch Teilchenverfolgung.

Schnitt- und Isoflächen

Modellierung der Fläche durch Dreiecke

Die naheliegendste Methode, Daten auf einem dreidimensionalen Gitter zu visualisieren, liegt darin, sich auf die Betrachtung von Flächen in diesem Raum zu beschränken. Auf solchen Flächen lassen sich dann ausgewählte Daten als Farbverläufe oder durch Vektorfelder darstellen. Mögliche Flächen sind zum einen Schnittflächen des Volumens mit Ebenen, Kugeln o. ä., zum anderen sogenannte Isoflächen, d. h. Flächen konstanten Funktionswertes. Am Beispiel eines Swimmingpools, dessen Temperaturverteilung dargestellt werden soll, gäbe es also die Möglichkeiten, entweder die Temperatur als Farbe auf einem Querschnitt durch den Pool darzustellen oder eine Isofläche einer bestimmten Temperatur zu berechnen, die die wärmeren von den kälteren Bereichen trennt. Für die Berechnung beider Flächen wird der sogenannte Marching Cubes Algorithm herangezogen, der darauf aufbaut, dass zunächst für jeden der Gitterpunkte ein Abstand d zur gesuchten Fläche berechnet wird, wobei das Vorzeichen von d angibt, auf welcher Seite der gesuchten Fläche sich der Punkt befindet. Im Falle einer Schnittebene berechnet sich der Abstand d eines Gitterpunktes $\vec{\chi}$ zur Ebene mit dem Hesse'schen Normaleneinheitsvektor $\vec{n^0}$ und dem Abstand m zum Ursprung nach $d = (\vec{\chi} \cdot \vec{n^0}) - m$, wobei Punkte mit negativem d auf der gleichen Seite der Ebene liegen wie der Ursprung. Bei Isoflächen ist d nicht Abstand im räumlichen Sinn, sondern bezieht sich auf die dargestellte Größe. Bei einem Punkt mit dem Wert D gilt als Abstand zur Isofläche des Wertes D_0 die Differenz $d = D - D_0$. Wurde so für jeden Punkt des Gitternetzes die Lage bezüglich der gesuchten Fläche geklärt (Punkte, die genau auf einer Fläche liegen, können wahlweise einem der beiden Fälle, oberhalb oder unterhalb, zugeordnet werden), wird nacheinander jeder Quader des Gitters einzeln daraufhin untersucht, ob er von der gesuchten Fläche geschnitten wird, was der Fall ist, wenn nicht alle Eckpunkte auf derselben Seite der Fläche liegen. Für diese Quader, die von der Fläche geschnitten werden, wird die jeweilige Schnittfläche berechnet. Und zwar berechnet man für jede Quaderkante, deren Endpunkte beiderseits der Fläche liegen, den Punkt auf der Kante, den die Fläche treffen müsste. Das geht mit linearer Interpolation. Man berechnet zunächst die Gewichtungsfaktoren für Punkte mit den Abstände d_1 und d_2:

$$w_1 = \frac{d_2}{d_2 - d_1} \quad , w_2 = \frac{d_1}{d_1 - d_2}$$

und daraus sowohl die Schnittpunkte $\vec{S}$ als auch skalare oder vektorielle Größen D auf den Schnittpunkten, die jeweils zwischen den betroffenen Eckpunkten der Quader interpoliert werden.

$$\vec{S} = \vec{P}_1 w_1 + \vec{P}_2 w_2 \ , \quad D = D_1 w_1 + D_2 w_2$$

Zwischen diesen Schnittpunkten auf den Kanten jedes Quaders wird nun eine Fläche aus Dreiecken modelliert, die die gesuchte Fläche approximiert. Zur Betrachtung der Schnittmöglichkeiten nutzt man dabei aus, dass sich die theoretisch 2^8 Möglichkeiten (Jeder Eckpunkt des Quaders kann entweder über oder unter der Fläche liegen) durch Berücksichtigung von Symmetrien auf 14 Fälle reduzieren lassen. Darunter befinden sich leider auch Schnitte, die nicht eindeutig, dafür aber auch vernachlässigbar sind, weil sie nur bei stark gekrümmten Flächen auftreten. Die Dreiecke aller Quader zusammen ergeben nun die gesuchte Fläche.

Isolinien

Neben dem Einfärben der Fläche bei der Ausgabe zur Visualisierung einer bestimmten Größe lassen sich auch Isolinien berechnen, also Mengen von Punkten gleichen Funktionswertes – dem zweidimensionalen Pendant der Isofläche – auf einer Fläche. Die Berechnung der Isolinien auf den Flächen erfolgt im Wesentlichen analog der der Flächen im Raum: Die Dreiecke, aus der sich die Fläche zusammensetzt, werden wieder einzeln daraufhin überprüft, ob und wie sie von einer Isolinie durchlaufen werden.

Vektorfelder

Gerade zur Visualisierung von Strömungsgrößen ist die Darstellung vektorieller Größen oft unerlässlich. Solche Größen wie z. B. die Strömungsgeschwindigkeit in einem bestimmten Punkt können in Form von an den Eckpunkten der Dreicke ansetzenden Pfeilen dargestellt werden. Es gibt jedoch auch die Möglichkeit, wenn die Flächennormale $\vec{n}^0$ bekannt ist, den Geschwindigkeitsvektor in einen zur Fläche normalen und einen tangentialen Anteil zu zerlegen. Der Betrag des normalen Anteils $\vec{\chi}_n = \vec{n}_0 \cdot \vec{v}$ kann dann als skalare Größe gespeichert und angezeigt werden, der zur Fläche tangentiale $\vec{\chi}_q = \vec{v} - \vec{\chi}_n = \vec{v} - |\vec{\chi}_n|\vec{n}^0$, die Querströmung, kann wieder durch Pfeile visualisiert werden. Bei der Darstellung von Querschnitten durch Rohre o. ä. lassen sich so Verwirbelungen besonders deutlich erkennen.

Particle Tracing

Ein anderer Ansatz zur Visualisierung von Strömungsdaten ist die Teilchenverfolgung oder Particle Tracing. Man denkt sich ein Teilchen an einer bestimmten Position im Volumen und möchte sehen, wie dieses Teilchen durch die Strömung transportiert wird. Der erkennbare Weg gilt zwar immer nur für ein Teilchen, das genau an einer bestimmten Stelle gestartet ist, lässt aber eindrucksvoll das Verhalten des Mediums deutlich werden und macht es dem Betrachter besonders leicht, z. B. Wirbel zu erkennen. Um die Teilchenbahn zu visualisieren, versucht das Modell, die Bahn eines Teilchens im Volumen für einen bestimmten Zeitraum durch einen Linienzug anzunähern. Soll mit $\vec{X}(\vec{x},t)$ der Weg des Teilchens beschrieben werden, der sich zum Zeitpunkt t in $\vec{x}$ befunden hat, so lautet das zugehörige Anfangswertproblem

$$\frac{d}{dt}\vec{X}(\vec{x},t) = \vec{u}(\vec{X}(\vec{x},t),t)$$
$$\vec{X}(\vec{x},0) = \vec{x}$$

Dabei ist $\vec{u}(\vec{x},t)$ die an den Gitterpunkten bekannte und dazwischen interpolierte Funktion der Strömungsgeschwindigkeit. Nach dem Euler-Verfahren ergibt sich dann als Iterationsvorschrift für die Annäherung einer Teilchenbahn in Zeitschritten h:

$$\vec{X}(\vec{x},t_{k+1}) = \vec{X}(\vec{x},t_k) + h\vec{u}(\vec{X}(\vec{x},t_k),t_k)$$

Da der Algorithmus bei der Approximation von $\vec{X}$ nur in den seltensten Fälle auf Gitterlinien oder -punkten landet, besteht ein großes Problem für den Algorithmus darin, herauszufinden, in welchem Gitterelement $\vec{X}$ liegt und zwischen welchen Gitterpunkten $\vec{u}$ interpoliert werden muss. Insbesondere bei unstrukturierten Gittern, die nicht nur aus Quadern, sondern aus einer Mischung vieler verschiedener dreidimensionaler Primitiven, wie Tetraeder, Pyramiden o. ä., zusammengesetzt ist, ist dieser Algorithmus mit hohem Aufwand verbunden.

Die Graphikpipeline

Nach all diesen Vorbereitungen stehen wir endlich am Anfang der sogenannten Graphikpipeline, einer Reihe von Operationen zur Ausgabe von 3D-Computergraphik. An ihrem Anfang steht die dreidimensionale geometrische Repräsentation des Darzustellenden – in unserem Fall der Dreiecke und Linienzüge – und an ihrem Ende die Ausgabe einer zweidimensionalen Graphik auf dem Bildschirm. Im Allgemeinen teilt man die Graphikpipeline in fünf Stationen ein, die jedoch nicht unbedingt, wie der Name suggeriert, sequentiell abgearbeitet werden, sondern je nach Verfahren auch parallel ablaufen.

Transformation
Der erste Schritt besteht aus der Umrechnung der dreidimensionalen Repräsentation, d. h. der Eckpunkte der Dreiecke und Linienzüge, in Bildschirmpunkte. Diese Transformation ist im Wesentlichen eine Kombination aus Drehungen und Verschiebungen, da der Betrachter sich im Raum frei bewegen und das Volumen aus jedem erdenklichen Winkel betrachten kann. Den Abschluss der Transformation bildet schließlich die Projektion, in den meisten Fällen eine Zentralprojektion.

Entfernung von Rückseiten
Im Gegensatz zu unseren Isoflächen, die von beiden Seiten aus betrachtet werden können, gibt es in der 3D-Graphik häufig Flächen, die nur von einer Seite aus sichtbar sind, weil sie z. B. einen geschlossenen Körper modellieren, den man nur von außen sehen kann. Um zu verhindern, dass für solche Dreiecke, die man quasi von hinten sieht, noch Berechnungen durchgeführt werden, überprüft man zunächst, ob die Normale einer Fläche vom Betrachter weg zeigt. In diesem Fall ignoriert man die Fläche im weiteren Verlauf. Da Normalen zu Dreiecken prinzipiell in beide Richtungen zeigen können, kann mittels der Reihenfolge der Angabe der Eckpunkte eines Dreiecks festgelegt werden, wie die richtige Normale ausgerechnet wird.

Entfernung verdeckter Oberflächen
Ein weiteres Problem der Visualisierung ergibt sich daraus, dass die Dreiecke einer Szene nacheinander dargestellt werden, dabei aber nicht unbedingt systematisch von weiter entfernten zu nahe gelegenen vorgegangen werden kann, weil eine Sortierung oft viel zu aufwendig wäre. Um zu verhindern, dass beim Zeichnen eines entfernten Dreiecks dichtere bereits auf dem Bildschirm sichtbare überschrieben werden, nutzt man aus, dass die Ausgabe auf dem Bildschirm gerastert ist, d. h. beim Zeichnen jedes Dreiecks werden jeweils alle Pixel (Bild- oder Rasterpunkte auf dem Bildschirm), die innerhalb des darzustellenden Dreiecks liegen, eingefärbt. Um ungewolltes Überzeichnen zu verhindern, merkt sich das Visualisierungsprogramm jetzt lediglich beim Setzen jedes Pixels, in welcher Tiefe es liegt, und kann beim nächsten Dreieck, das dieses Pixel einschließt, ein Überzeichnen entsprechend verhindern. Dieses Verfahren nennt sich Tiefenpuffer-Algorithmus, oder, weil die Tiefe oft auf der z-Achse liegt, z-Buffer-Algorithmus.

Beleuchtungsmodelle

Gerade bei einfarbigen Flächen, wie wir sie z. B. als Isoflächen verwenden, fehlen dem Betrachter des zweidimensionalen Bildes wichtige Informationen zur räumlichen Einordnung. Um dem Auge Plastizität vorzutäuschen, schattiert das Visualisierungsprogramm die Fläche entsprechend der Lage zu einer virtuellen Lichtquelle. Dabei gilt nach dem Lambert'schen Gesetz, dass die Helligkeit einer diffus reflektierenden (nicht spiegelnden, sondern matten) Fläche proportional dem Cosinus des Winkels zwischen der Flächennormalen und dem Vektor zur Lichtquelle ist. In der Praxis wird dieses Prinzip in drei Varianten umgesetzt: Im Flat Shading wird die Helligkeit für jedes Dreieck anhand seiner Normalen berechnet, daher haben benachbarte Flächen zum Teil deutliche Helligkeitsunterschiede. Beim Gouraud-Shading werden die Normalen in den Eckpunkten aus denen aller umliegender Dreiecke gemittelt, die Helligkeiten in diesen Punkten berechnet und beim Zeichnen der Dreiecke zwischen diesen Punkten interpoliert, was die Kanten glättet. Das aufwendigste Verfahren ist das sogenannte Phong-Shading, bei dem nicht die Helligkeiten, sondern die Normalen zwischen den Eckpunkten interpoliert werden und das Reflexionen erstaunlich genau simulieren lässt.

Rasterung

Der letzte Schritt der Visualisierungskette läuft eigentlich parallel zur Beleuchtungsberechnung und der Entfernung verdeckter Oberflächen, da diese beiden Schritte bereits auf einer Rasterung basieren. Hier versucht das Visualisierungsprogramm, die Dreiecke möglichst treffend für das Bildschirmraster zu diskretisieren. Diesen Vorgang kann man sich ungefähr so vorstellen wie den Versuch, auf einem Karopapier durch Ausmalen bestimmter Kästchen eine Gerade darzustellen; jede Gerade, die nicht horizontal oder vertikal liegt, wird eher zu einer Treppe als zu einer Geraden. Dieser Effekt nimmt mit zunehmender Auflösung des Bildschirms ab.

Literatur

Uwe Wössner: Implementierung von Visualisierungsmethoden für unstrukturierte Gitter in einer Multi-Block-Umgebung, Januar 1996, Stuttgart

F. H. Post, A. J. S. Hin (Eds.): Advances in Scientific Visualisation, 1992 Springer Verlag Berlin Heidelberg New York

K. D. Tönnies, H. U. Lemke, 3D-Computergrafische Darstellungen, 1994 R.Oldenbourg Verlag München Wien

Ergebnisse

(Monika Wierse)

Natürlich wäre es am schönsten gewesen, wenn wir mit den soeben erlernten Mitteln richtige Strömungsprobleme numerisch hätten lösen können. Leider sind die Dinge dafür dann doch zu kompliziert. Um aber trotzdem einen Einblick in aktuelle Problemstellungen zu geben, stelle ich einige Datensätze zu Strömungsproblemen vor. Die sehr zeitaufwendigen Rechnungen dazu wurden zum Teil auf millionenschweren Supercomputern durchgeführt. Die folgenden drei Beispiele stammen aus Forschungsprojekten der DaimlerChrysler AG (Bilder 1 und 2) und dem Institut für Thermische Strömungsmaschinen der Universität Stuttgart (Bild 3).

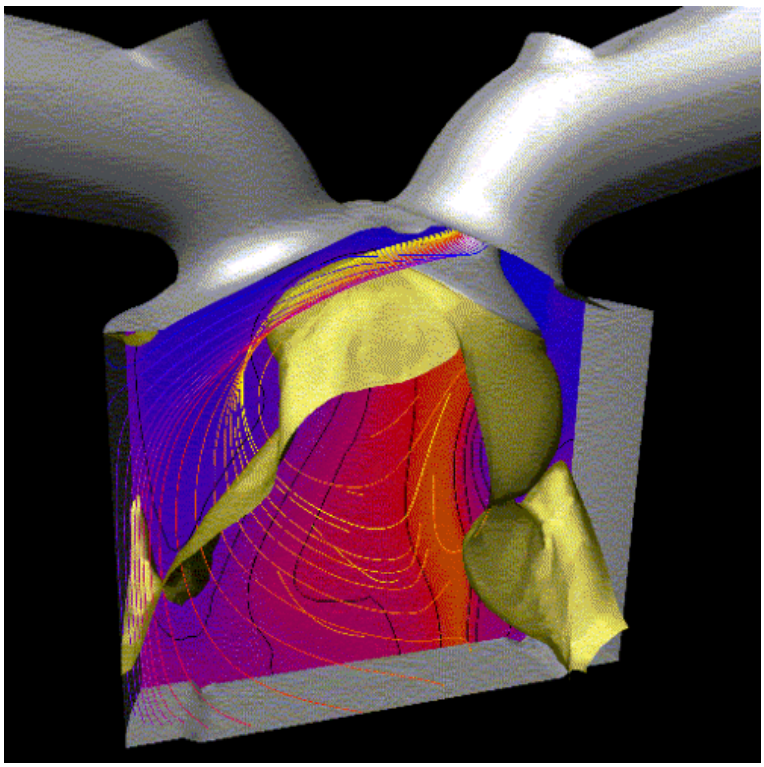


Bild 1 (rechts): Simuliert wurden die Strömungsverhältnisse in einem Zylinder eines 4-Takt-Motors. Gerechnet und abgebildet wird nur eine von zwei spiegelbildlichen Hälften (eigentlich ist es ein Vierventiler). Dies ist eine Möglichkeit, den Rechenaufwand kleiner zu halten. Man geht dann davon aus, dass die Strömungsverhältnisse symmetrisch sind.

Auf dem Bild ist das Einlassventil geöffnet, so dass frisches Gemisch in den Zylinder einströmen kann. Das

Auslassventil ist geschlossen, wir sehen also einen Zustand während des Ansaugtaktes. Die helle Fläche im Zylinder trennt frisches von bereits verbranntem Gemisch. Im Weiteren würde man untersuchen, ob das verbrannte Gemisch komplett ausgespült wird und wie gut das frische durchwirbelt wird. Die Striche (Partikelbahnen) zeigen Wege von Strömungspartikeln seit dem Öffnen des Einlassventils.

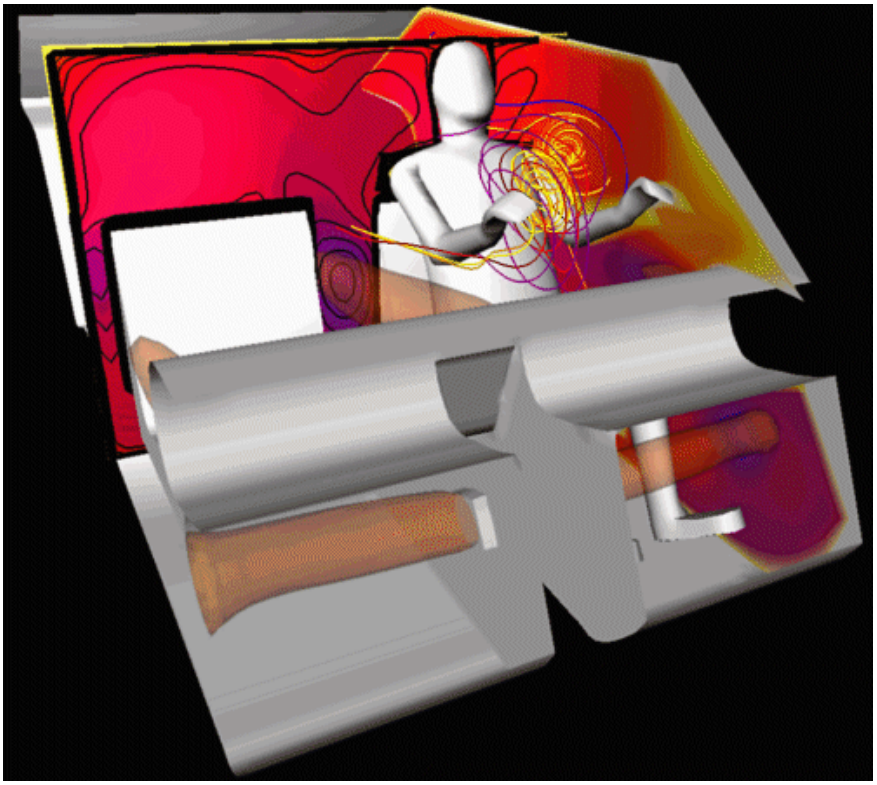


Bild 2: Berechnet wurden die Strömungsverhältnisse im Innenraum eines Fahrzeugs. Zu untersuchen war dabei, wie sich Temperaturveränderungen aus der Klimaanlage auf das Wohlbefinden des Fahrers auswirken. An den Partikelbahnen sieht man sehr schön, dass es vor dem Fahrer sehr turbulent zugeht.

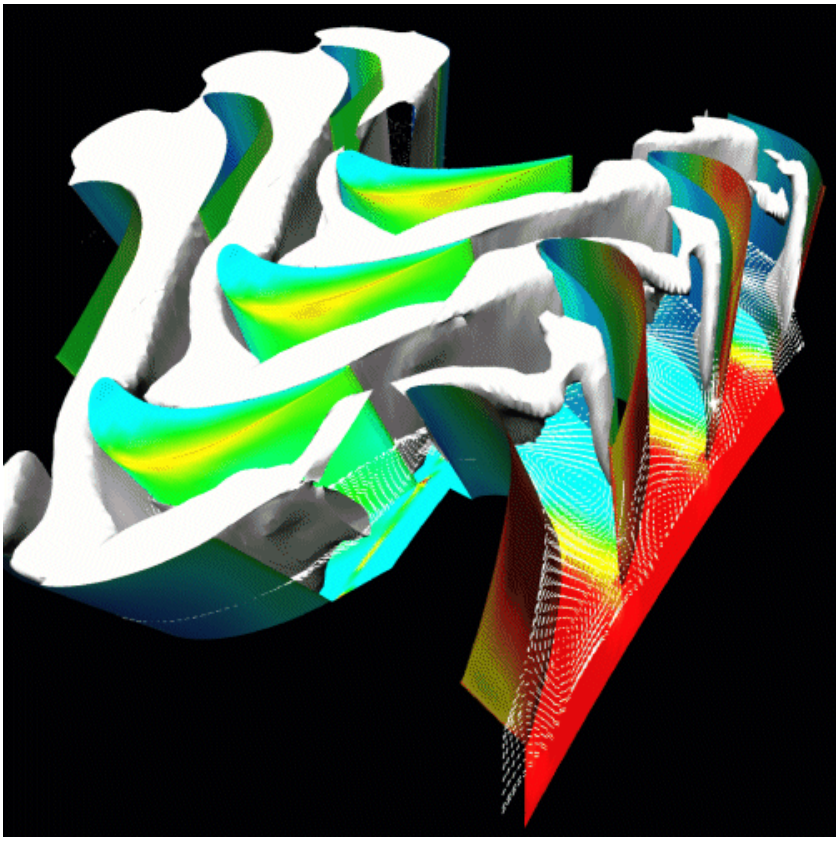


Bild 3: Die dunklen Flächen stellen (quer angeschnittene) Blätter von Schaufelrädern einer Turbine dar. Man sieht nur Teile von 3 Schaufelrädern, wobei die beiden äußeren feststehen (Statoren) und das mittlere sich bewegt (der Rotor, dies wird auch numerisch simuliert!). Die weißen Flächen sind Teile der Isofläche zu einem bestimmten Druckwert. Auf einer Schnittfläche durch das Berechnungsgebiet ist das Vektorfeld der Strömung dargestellt. Die Druckverhältnisse in einer Turbine müssen gut verstanden sein, damit die Turbinenblätter der Belastung durch das durchströmende Gas oder Wasser standhalten können.