A kaleidoscopic excursion into numerical calculations of differential equations

E. van Groesen
Faculty of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract

In this expository paper a sketch is given of some basic problems that arise when differential equations are discretised and calculated numerically. Some analytic techniques are demonstrated for the investigation of these numerical schemes. Numerous pictures illustrate the ideas and the formulae. References to available software packages are given.

For the logistic equation the time asymptotic behaviour depends on the chosen time step, showing period doubling from a correct attraction to the stable equilibrium solution to chaotic behaviour. For the linear harmonic oscillator energy conservation for large times is investigated.

Besides these ordinary differential equations, several aspects of linear and nonlinear wave equations, described by partial differential equations, will also be considered. A finite difference method is demonstrated for a linear wave equation. For the Korteweg-de Vries equation, an equation that combines dispersive and nonlinear effects in wave propagation, Fourier truncations are studied. It is shown that the basic Hamiltonian structure is preserved (implying energy conservation), as well as a continuous translation symmetry (implying conservation of horizontal momentum). As a consequence, travelling waves - i.e. wave profiles that translate with constant speed undeformed in shape - are present in truncations of any order. Numerical evidence, as well as a complete analytical proof, is given.

A preliminary version of a software package WAVEPACK, aimed to familiarize the unexperienced user with many basic concepts from wave theory (dispersion, groupvelocity, mode analysis, etc.) and to perform actual calculations in an easy way, is available upon request.

Keywords: model consistent discretisations, logistic equation, harmonic oscillator, wave equations, Korteweg - de Vries equation, travelling waves.

1991 Mathematics Subject Classification: 65-02, 35Q20.

1 Introduction

This expository paper has several goals. It is written by an applied analyst who occasionally feels the need to use the computer, for instance to produce pictures or to get an idea of the behaviour of some set of differential equations. Mostly I use simple standard software packages, many of which are available nowadays and operate on PC's; incidentally I need the help of numerical analysts to produce my own simple programmes. All the results and pictures in this paper were obtained in either way, and can be reproduced by anyone who is interested.

When performing such calculations, this goes along with a considerable amount of scepticism from my part: do I get (from the computer) what I really wanted, or is the calculated phenomenon a consequence of some unforeseen property of the numerical scheme that is used for the computations. (Even if "you get what you see", it may not be true that what you see is what you should see.) My experience from simple examples is that the latter thing often happens. This explains my scepticism, and at the same time increases my respect for scientists who are able to calculate turbulent flows, the flow around wings, the atmospheric conditions for weather forecasts, to mention just a few of the most difficult areas of application for numerical calculations. Of course, scientists involved are very much aware of the many pitfalls, and try to cope with them as good as possible. Because, despite the difficulties, numerical calculations are extremely useful, as shown by the results of these applications.

My aim is to demonstrate with much simpler examples some of the fundamental problems that are encountered. And how they sometimes (partly) can be resolved by exploiting some specific knowledge one has about the problem to be calculated. All examples deal with differential equations, both ordinary and partial differential equations. We will look at these problems, and their discretisation, with the eye of an analyst, and use some basic analytical techniques for their investigation. The text should be understandable for advanced students (math, physics, engineering) who are willing to put enough effort into the subject to grasp the relevance of the reasoning and the concepts used. Even if a few of the more difficult concepts are not understood completely, this should not hinder to follow the main line of the rest of the text.

In the remainder of this introduction we will make some of the general ideas more explicit.

Continuous variables

As stated, we will deal with problems connected to numerical computations. Such computations were in fact the first and primary aim computers were designed for: to make large calculations to solve problems that are encountered in science and in technical applications. A fundamental problem arises because of the discrepancy between the natural, simplest, way to describe these phenomena in mathematical models, and the kind of problems a computer can handle: it is the difference between a continuous and a discrete description. Let us look at this in somewhat more detail.

Physics, Chemistry, and the natural sciences in general, provide the basic laws that describe the phenomena of nature. They do so, more or less accurate depending on the particular phenomenon or situation under investigation.

In many cases, these laws of nature are formulated in variables describing the "state" of the system. These variables are considered, or idealised, to be continuous. For instance, to describe a fluid in a certain spatial domain, it is in many instances convenient to pretend as if the fluid does not consist of individual molecules, but as if the important quantities like mass and velocity are smoothed out over the whole domain. Therefore the continuum description uses concepts like mass density (mass per volume) and fluid velocity (summed velocity of the particles per volume) that are defined at *each* point by a limiting procedure over ever decreasing volumes around the point under consideration. Usually, a continuum description leads to equations that can be formulated very compactly in mathematical terms, often by a set of (partial) differential equations.

In dynamical problems, whether continuous or discrete in space, the evolution of the system is described with the time, which is a continuous variable by its nature. Hence all these problems are of a continuous nature.

Calculations for such "continuous" problems can be performed by a computer provided they are reformulated in algorithms which break the computations into a few elementary operations that the hardware can handle. Basically, a computer can work only with a finite number of quantities. Although that number is quite large in modern computers, the number is essentially finite. Therefore, the computer requires that the variables and the equations are "discretised".

Computers may be a recent invention, the idea of replacing continuous quantities by discrete approximations is quite old and lies at the heart and at the beginning of mathematics.

Calculating the circumference of a circle by approximating it by better and better fitting polygons for which the length is easily calculated, is an example from ancient times. In fact, just as the continuum hypothesis in the physical description of fluids, it could equally well be said that the continuous circle is an idealisation of a polygone-like construction. Other examples are Riemann sums for calculating integrals, the difference quotient for the derivative etc. And again, it can well be said that the mathematical definitions of integral and derivative are idealisations, limiting quantities, of the discrete concepts. As a final example, remember the representation of periodic functions on an interval by their Fourier series. Since there are infinitely many different base functions, the sine- and cosine functions, the set of periodic functions is called an infinite dimensional function space. *Infinite* to distinguish from a finite dimensional vector space for which a finite number of base vectors suffice to form each element by linear combinations. By considering a truncation of the series to some order, the infinite dimensional function space reduces (is "discretised") to a finite dimensional one which consists of vectors for which the elements are the finite number of Fourier coefficients. The other

way around would be to start with finite sums of harmonic functions, and then consider linear combinations of more and more base functions, leading from finite dimensional spaces to an infinite dimensional one in the limit.

In the following, we will consider some of the basic questions that are related to the process of discretising. Starting with given quantities and equations of a continuous nature, we will address the question which properties a discretisation should have, and how these requirements are related to the actual purpose the discretisation should serve.

Approximation

Considering functions defined on some interval $I \subset \mathbb{R}$, a function is defined once its value is specified at each point of I . This requires an infinity of data for each function. So, a nontrivial (linear) space of functions, say U , is infinitely dimensional. An important part of *approximation theory* deals with the question how, and how well, a given continuous function can be approximated if only a finite number of data are given. The simplest example is to define a "grid", i.e. a partition of the interval, to take the value of the function u at the grid points, and to specify a prescription how the approximation will look like in between. Linear, quadratic, cubic or higher order interpolation between grid points are well known examples. Another discretisation is with finite Fourier truncations mentioned above, i.e. representing a periodic function with only a finite sum of harmonic functions. In each of these methods, a quantity h can be introduced, for instance the (maximal) mesh width of the grid, or the inverse of the number of Fourier modes. Such an approximation will be denoted by u_h and is an element from a finite dimensional subspace U_h of U . In approximation theory convergence and other properties are studied. Sensibly speaking, the approximation procedure should be such that the smaller h is taken, the better u_h will "look like" the original function u :

$$u_h \rightarrow u \text{ as } h \rightarrow 0. \quad (1)$$

Specifying the norm in which sense the convergence holds, gives information about the quality of the approximation.

Remark. Finite dimensional linear spaces are closed for the usual linear operations: addition and scalar multiplication. One of the major problems in numerical calculations, where one often has to deal with nonlinearities, is that this is not the case for multiplication! The product of two piecewise linear functions is not piecewise linear; the product of two finite Fourier truncations contains in general more Fourier modes than the order of truncation. All these properties are well known, especially when one has in mind the same properties for decimal point arithmetic. Indeed, when dealing with numbers with only one decimal products may have two decimals (e.g. $0.1 \times 0.1 = 0.01$, which would be considered to be zero when approximating numbers with only one decimal).

Numerical convergence

Now assume that we have some "continuous system". Symbolically it can be

written like

$$C(u) = 0$$

where C is some operation applied to the function u . For instance, this equation can be an algebraic or differential equation. Let C_h be some "discretisation" of C , i.e. an operation defined on a space of approximations u_h , and consider the equation

$$C_h(u_h) = 0.$$

Then it is said, roughly speaking, that the discretisation procedure is *numerically convergent* if solutions of the discretised equation tend to the solution of the equation for vanishing h :

$$\text{if } C_h(u_h) = 0 \text{ and } h \rightarrow 0, \text{ then } u_h \rightarrow u \text{ with } C(u) = 0. \quad (2)$$

Just as is the case in approximation theory with condition (1), numerical convergence expressed by (2) is a natural requirement for any discretisation. Many investigations in this area concentrate on proving the numerical convergence for a given discretisation (using such concepts as numerical consistency and - stability). The norm(s) in which the convergence holds provides, just as for functions, a measure for the quality of the discretisation.

Model consistency

In approximation theory, when attention is restricted to functions with a specific property (such as the degree of differentiability, or a monotonicity or convexity property for instance), one may want to restrict to approximations with the same properties. The same could be asked when discretising equations: if for the given continuous system some property is known to hold (possibly only for a limited set of solutions), one would like to find numerically convergent discretisations which possess the same property. Although the formulation here is rather vague, the question is quite relevant, as shall be explained in the rest of this contribution. Clearly, what is required in order to achieve this aim, is some knowledge of the original system that is essential (and analytically manageable enough) to preserve under discretisation.

In this paper we will deal with dynamical systems, and show that properties such as (time-) asymptotic behaviour, conservation of energy, and the appearance of travelling waves can be inherited by suitably constructed discretisations, at a fixed h .

We will refer to discretisations which retain such a specific property as *model consistent discretisations*, since such discretisations will "model" some essential property of the continuous equation. Compared to multi-purpose discretisation methods, the construction of model consistent discretisations requires additional analytical insights.

Furthermore, there is an essential difference with the concept of numerical convergence, which is a property of the discretisation for $h \rightarrow 0$. Model consistency,

on the other hand, requires that a special property of the continuous equation is present in the discretisation at a *fixed* h .

2 Logistic equation

The *logistic differential equation* is the first order differential equation for a scalar function x of time t : $x = x(t)$ given by

$$\dot{x} = x(1 - x) \quad (3)$$

To give just one interpretation of this equation, when the variable x denotes the size of some population, and $x(t)$ is the size at time t , this equation describes the restricted growth of the population. The variables have been normalised in such a way that possible constants can be taken to be 1.

Given a certain *initial value*, i.e. the value of x at some initial time (which may be taken without restriction to be $t = 0$), say $x(0) = x_0$, with x_0 a given number, the solution of (3) can be written down explicitly (using the method of separation of variables). The qualitative behaviour of the solutions is depicted in Fig.1. (For the interpretation of x as the size of a population, negative values of x are not realistic and could be discarded.)

Now consider the most direct method to discretise (3). The unknown function $x(t)$ which is sought to satisfy the differential equation, is replaced by function values at gridpoints. For the first grid point we take $t = 0$, and next points a distance h apart. Then we look for a discretisation of (3), i.e. a procedure to calculate values $\hat{x}_n$ which are an approximation for $x(nh)$.

To that end, the simplest way is to replace the derivative in (3) by the (forward) difference quotient. The expression that arises is known as the *Euler discretisation*:

$$\hat{x}_{n+1} - \hat{x}_n = h\hat{x}_n(1 - \hat{x}_n). \quad (4)$$

This equation can be rewritten in a simpler way by introducing a normalised variable $y_n = a\hat{x}_n$ with $a = \frac{h}{1+h}$. Then the way how y_{n+1} follows from y_n can be written concisely: with $y = y_n$ and then writing y' for the next value: $y' = y_{n+1}$, the relation is given by

$$y' = \mu y(1 - y) \text{ with } \mu = 1 + h. \quad (5)$$

This map $y \rightarrow y'$ is known as the *logistic map*, and can be depicted by plotting y versus y' as in Fig.2.

The fixed points of this map are $y = 0$ and the more interesting one, which we will call $\bar{y}$, is given by $\bar{y} = \frac{\mu-1}{\mu} = \frac{h}{1+h}$. These fixed points $y = 0$ and $\bar{y}$ correspond to the equilibrium solutions $x(t) = 0$ and $x(t) = 1$ respectively of the logistic differential equation.

Applying the logistic map several times to a given initial condition, produces a sequence of points y_n , or $\hat{x}_n$ which should be approximants for $x(nh)$ with $x(t)$ the

solution of (3). In between the points $\hat{x}_n$ we imagine the function to be approximated by linear interpolation. It is intuitively clear that the smaller the value of h , the better this piecewise linear function will approximate the solution of the continuous differential equation on compact time intervals. (A mathematical proof of this intuitive idea will be omitted here.) This means that the Euler discretisation scheme is numerically convergent.

Note, however, that even for the relatively large value of h in Fig.2, and despite the difference with the trajectories depicted in Fig.1, the asymptotic behaviour is the same: the solution $\hat{x}_n$ tends to the stable equilibrium $x = 1$ for increasing n , or, equivalently, the sequence y_n converges to the nontrivial fixed point $\bar{y}$ of the discrete map (5).

Increasing further the value of h , a qualitative change in the asymptotic behaviour appears as h crosses the value 2. For $h > 2$, i.e. $\mu > 3$, the fixed point $\bar{y}$ is no longer stable: arbitrary points, even those starting arbitrarily close, are no longer attracted to $\bar{y}$. What actually happens, depends on the precise value of h : for $2 < h < 2.44$ there appears a so-called attracting period 2 solution of (5): the values $\hat{x}_n$ tend to switch between two fixed values around $x = 1$, see Fig.3. For $2.44 < h$ an attracting period 4 solution appears, and the period doubling process is repeated until for $h > 2.57$ chaotic behaviour appears: the successive values $\hat{x}_n$ jump from values above 1 to values below 1 in a "random, chaotic way".

Remark. The logistic map serves as the prototype map in bifurcation theory to show the appearance of chaos through period doubling process. The bifurcation diagram is shown in Fig.4; interested readers are referred to the literature, e.g. Guckenheimer & Holmes (1983), Devaney (1986), Schuster (1987), and Broer e.a. (1991).

From these observations, we can make some valuable conclusions explicit. Firstly, the Euler discretisation method is numerically convergent. More than that, although the discretisation is rather poor initially for stepsizes h not close to 0, for each $h < 2$ the asymptotic behaviour of the solutions is well captured by the discretisation. Hence it can be said that for each fixed time step $h < 2$ the Euler discretisation is "model consistent" as far as the asymptotic behaviour is concerned. It may also be remarked that even in this rather simple case, a detailed study of the relation between the differential equation and its discretisation is rather involved. This should make one modest in expectations when dealing with more complicated situations.

Concerning the appearance of chaos it may be argued that this only appears for unrealistically large values of h , namely for $h > 2$. Indeed, for the isolated problem at hand this is true. But as we shall see later, in more complicated equations with several quantities, it may very well be the case that a time step which is small for one quantity is large for another one. Then a very good approximation of one quantity may be accompanied by a (numerically enforced) chaotic behaviour of another one, which may influence the complete result. See also the Epilogue.

Figure 1:

Solution curves for the logistic d.e. $\dot{x} = x(1-x)$, plotted x versus time t horizontally. There are two *equilibrium solutions*: $x(t) = 0$, and $x(t) = 1$ for all time. For initial data x_0 between 0 and 1, the solution exists for all time, and tends to 1 for $t \rightarrow \infty$ and to 0 for $t \rightarrow -\infty$. For initial value $x_0 > 1$ the solution exists for all positive times, and tends to 1 for $t \rightarrow \infty$. For $x_0 < 0$ the solution blows up to $-\infty$ for a finite positive time. Because the equilibrium 1 attracts all solutions starting sufficiently close, this equilibrium is called *stable*. The equilibrium solution 0 is *unstable*.

Figure 2:

In the left picture the piecewise linear function is drawn obtained from Euler's discretisation method, using $h = 1.95$, for 40 iterations. In the picture at the right, the same data are plotted in the normalised variable y as a map.

Figure 3:

As Fig.2, but now for $h = 2.3$, for which a fast convergence to a period two solution of the map is seen. The corresponding piecewise linear curve in x vs. t will not converge asymptotically to the equilibrium solution $x(t) = 1$ of the logistic differential equation.

Figure 4:

The bifurcation diagram for the logistic map (5) with horizontally the bifurcation parameter μ . Note the period doubling process, starting at $\mu = 3$, after which the fixed point is no longer stable: for $\mu > 3$, corresponding to stepsize $h > 2$, the discretisation (5) has a different asymptotic behaviour than the continuous logistic differential equation. The wonderful, complicated scaling processes in the chaotic area of this picture is explained in the relevant literature.

3 Harmonic Oscillator

Consider another well known simple differential equation, now of second order, for the scalar function $q(t)$:

$$\ddot{q} = -\omega^2 q. \quad (6)$$

This is the *harmonic oscillator*, which has as general solution the harmonic functions with frequency ω :

$$q(t) = A \sin(\omega t) + B \cos(\omega t).$$

The constants A and B can be adjusted to satisfy given initial data for q and its first derivative $\dot{q}$.

Written as a system of two first order equations, (6) becomes

$$\begin{aligned} \dot{q} &= \omega p \\ \dot{p} &= -\omega q \end{aligned} \quad (7)$$

Introducing the two-vector $z = (q, p)$, this is more concisely written as

$$\dot{z} = \omega J z \quad (8)$$

where J is the matrix: $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. This matrix is skew-symmetric and satisfies $J^{-1} = -J$, and is called a *symplectic* matrix. It is possible to scale down the coefficient ω in (7) by introducing a new variable $\tau = \omega t$. Writing $\hat{z}(\tau) = z(t)$, the equation becomes

$$\frac{d}{d\tau} \hat{z} = J \hat{z}. \quad (9)$$

For the rest of this section we will simply write z for $\hat{z}$, i.e. consider (8) with $\omega = 1$.

When performing numerical calculations, this scaling should be kept in mind. In fact, in view of the sinusoidal solution, it is clear that at least two time steps are required per period to capture the least required information about such a solution. Hence the maximum time step h should satisfy $\omega h \leq 2\pi/2$, i.e.

$$h \leq \frac{\pi}{\omega}. \quad (10)$$

For later reference we note that the equations above are a Hamiltonian system. Indeed, introducing the *total energy* H as the Hamiltonian:

$$H(z) = \frac{1}{2}|z|^2 = \frac{1}{2}(q^2 + p^2), \quad (11)$$

the equations can be written like

$$\dot{z} = J \nabla H(z), \quad \text{equivalently} \quad \begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial q} \end{aligned} \quad (12)$$

Equations of this form are called Hamilton equations for a system with Hamiltonian H . One particular property of such equations is that the total energy is conserved. Indeed, as a consequence of the skew symmetry of J , it holds for solutions of (12) :

$$\frac{d}{dt}H(z(t)) = z \cdot Jz = 0.$$

Fig.5 summarises the most important properties of the solutions of the system. The *phase portrait* is obtained by plotting $q(t)$ versus $p(t)$ upon eliminating the time t . The trajectories are circles centered around the equilibrium solution $q = 0, p = 0$. This equilibrium is called a *center*.

In every point in the phase plane, the direction field is tangent to the trajectory (circle) through the point. This immediately explains why the Euler discretisation method from the foregoing section will produce solutions that diverge away from circles. See Fig.6. This divergence is present for all values of the stepsize h , of course less prominent for small values of h .

When long time calculations are to be made for which the energy should be conserved, the choice for a higher order method such as Runge Kutta may seem to be obvious. Indeed, RK4 is a fourth order method and conserves the energy much better than the first order Euler scheme. However, RK4 is a scheme that is somewhat dissipative, so that despite the higher order accuracy, for large times the energy will decrease anyhow (see Fig.10).

There are several ways to retain energy conservation in calculations exactly. We will describe two essentially different approaches. The first one is of a numerical nature and consists of modifying a given discretisation method. The second one is more analytical and modifies the given differential equation in such a way that, (almost) independent of the discretisation method to be used, the calculations have the desired property of energy conservation.

3.1 Modification of the discretisation

When a given discretisation method can be studied in detail, the reason of the change in the energy can be investigated, and one may be able to improve energy conservation in a direct way. This can only be done for simple schemes. In this subsection we will demonstrate the idea to Euler's scheme considered in the preceding section.

Euler's scheme for $\dot{z} = Jz$ can be written as the map

$$z \rightarrow z' \text{ given by } z' = (I + hJ)z, \quad (13)$$

where I denotes the identity matrix and h is the stepsize. The energy E' after one step is easily found to be related to the previous energy E as:

$$E' = \frac{1}{2}|z'|^2 = \frac{1}{2}(1 + h^2)|z|^2 = (1 + h^2)E \quad (14)$$

Figure 5:

The harmonic oscillator. UL: The upper left picture is a plot of the direction field. UR: Some phase curves are shown in the upper right picture, motivating the name 'center' for the equilibrium at the origin. LL: In the lower left picture the component q is plotted versus time, showing the harmonic behaviour. LR: The lower right picture shows the energy versus time, constant for the solutions shown in the phaseplane.

Figure 6:

Euler discretisation of the harmonic oscillator, with stepsize $h = 0.05$. UL: A trajectory diverges quickly, even for this small value of h . UR: The energy increases monotonically for the calculated solution. LL: The component q versus time shows an increase in amplitude. LR: The error as a function of time. The error is defined as the distance in the phase plane between the calculated solution and the exact solution. Except from the divergence (energy increase), this error is mainly due to *phase error*. See also the next figures.

This explains the monotone increase of the energy as depicted in Fig.5. To modify the scheme such that it conserves energy, artificial dissipation has to be introduced to compensate for the energy increase. This can easily be done by introducing a dissipation in the discretisation (13), the strenght of which is controlled by a parameter α , the value of which has to be determined suitably. For the map

$$z \rightarrow z' \text{ given by } z' = ((1 - \alpha h)I + hJ)z, \quad (15)$$

the energy change in one step is given by

$$E' = \gamma E \text{ with } \gamma = (1 - \alpha h)^2 + h^2. \quad (16)$$

Hence the scheme (15) is energy conserving for the choice

$$\alpha = \frac{1 - \sqrt{1 - h^2}}{h} \approx \frac{1}{2}h. \quad (17)$$

This energy conserving scheme can be interpreted in several ways. One way is to note that the modified discretisation (15) is in fact Eulers method with stepsize h for the modified differential equation

$$\dot{z} = Jz - \alpha z. \quad (18)$$

This differential equation is dissipative with an exponential decrease of the energy: $\frac{d}{dt}H(z) = -2\alpha H(z)$. So, this dissipation compensates for the energy increasing effect of the Euler scheme.

Remark. A second interpretation of (15) is more interesting. For the chosen value of α the discretisation scheme can be formulated, after some calculation, like

$$z' - z = \bar{h}J\left(\frac{1}{2}(z + z')\right) \quad (19)$$

where $\bar{h}$ is related to h by $h = \bar{h}/(1 + \bar{h}^2/4)$. This is a simple example of a *symplectic* discretisation scheme. Such schemes get much attention nowadays and are especially designed to retain an important property of classical Hamiltonian systems. This property is that the time evolution is a symplectic map (implying, for instance, that the Hamiltonian flow is volume preserving in the phase space). The last scheme has this property too since the defining map is symplectic, which means that $z' \cdot Jz' = z \cdot Jz$, as can easily be verified. Energy conservation, however, is not guaranteed for symplectic schemes. The conservation property in this particular example is a consequence of the additional property that the equations are linear.)

3.2 Modification of the differential equation

In this subsection we consider a completely different method to improve energy conservation in numerical schemes. Instead of looking into the properties of the schemes itself, we will modify the differential equation itself. The modification will be such that each discretisation will more successfully conserve the energy for solutions that have the same initial energy (the discretisation will be different

Figure 7:

For the modified equation (20) the direction field is shown in the left picture. Trajectories starting outside the level set E_0 are shown to be attracted to this set in the right picture.

Figure 8:

Euler discretisation of the modified differential equation (20) for time step $h = 0.2$ and parameter $a = 0.05$. The initial data has energy $E_0 = 12.5$, and the asymptotic value is approximately 13.5, in agreement with (23). UL: trajectory of solution starting at E_0 , LL: the q - component vs time, UR: the evolution of the energy for the calculated solution, LR: the error of the solution.

for different values of the initial energy). For given initial data x_0 let E_0 be the corresponding value of the energy. For the analytical solution of (12) the energy is conserved $H(x(t)) = E_0$ for all time, but not for a solution from numerical calculations, as is seen from (14). Therefore we will modify the differential equation in such a way that the level set E_0 becomes attractive: trajectories that have deviated from this level set (by numerical errors) will be forced back to this level set. To that aim consider for some suitably chosen parameter α the modified equation:

$$\dot{z} = J\nabla H(z) - \alpha(H(z) - E_0)\nabla H(z), \quad (20)$$

explicitly

$$\dot{z} = Jz - \alpha(H(z) - E_0)z.$$

The additional term $-\alpha(H(z) - E_0)\nabla H(z)$ in the equation acts like a driving force with the desired properties. This can be seen by considering the evolution of the energy which is given by

$$\frac{d}{dt}H(z) = -\alpha(H(z) - E_0)|\nabla H(z)|^2. \quad (21)$$

For the case under consideration of a quadratic H this becomes the logistic equation for H :

$$\frac{d}{dt}\left(\frac{H(z)}{E_0}\right) = 2\alpha E_0\left(\frac{H(z)}{E_0}\right)\left(1 - \frac{H(z)}{E_0}\right) \quad (22)$$

From this it follows that the analytical solutions of (20) are attracted to the level set E_0 . In Fig.7 this property is displayed pictorially: the direction field clearly shows the presence of the modification (compare with Fig.5).

Now discretisations of the modified differential equation will preserve the energy much better. This is shown in Fig.8 for the Euler discretisation and in Fig.9 for the fourth order Runge Kutta method. Note that for this simple example the modification makes it possible to take much larger time steps than would be possible without the modification. The conservation of energy in large time calculations is shown in Fig.10.

Remark As we can see from the numerical calculations, the energy is well conserved, although the asymptotic value E_∞ differs from the initial value E_0 that had to be retained. This is due to the net result of two effects in these calculations: the dissipative effect from the modified equation (when $\alpha > 0$), and the property of the numerical scheme: slightly dissipative for RK4, and energy increasing for Euler.

To understand some of the results, consider once again the Euler method. A simple calculation shows that the energy changes per step as in (16) but now with a modified coefficient:

$$E' = \gamma E \text{ with } \gamma = (1 - \hat{\alpha}h)^2 + h^2,$$

Figure 9:

Runge-Kutta 4-th order discretisation of the modified differential equation (20) for time step $h = 0.2$ and parameter $a = 0.03$. Same display as in Fig.8, but note the different scale along the energy axis (multiplied by 2) in the LL picture.

Figure 10:

Calculated energy during large time calculations with large time step $h = .04$ for Euler (upper) and Runge Kutta (lower). Note the different asymptotic values E_∞ for Euler depending on the choice of the parameter $a = \alpha$, agreeing with (23). For $\alpha \geq .2$ Euler's scheme becomes unstable. For $\alpha = 0$ RK4 is seen to be dissipative. For $\alpha = .05$, calculations with RK4 are indistinguishable in this picture from the exact value at the axis.

with

$$\hat{\alpha} = \alpha(H(z) - E_0).$$

Hence, in the pictures the asymptotic energy follows with (17):

$$\alpha(E_\infty - E_0) \approx \frac{1}{2}h. \quad (23)$$

Remark. Despite the fact that we succeed to control the energy of the oscillation, at a value more or less close to the desired value, this has hardly any effect on the error of the calculated solution itself. Compare the error in Figs.6,8 and 9. The reason is that the error, defined as the distance in phase plane between the exact and the calculated solution, is mainly caused by an error in phase. So, even when we are able to keep the calculated solution well on the circle with the correct radius (determining the energy), the place at this circle at a certain time will be different from the exact position. When the calculated solution is “out of phase” with the exact one, the maximal error is found, being twice the radius. This phase error depends on the frequency, i.e. on the value ω in (8). When dealing with only one oscillator, this phase error can be suppressed by using a special discretisation that is based on knowledge of the exact solution. When two or more oscillators have to be calculated at the same time, as will be the case in the next section, this phase error cannot be suppressed when the frequencies are different.

4 Wave equations

In this section we will consider a much more difficult example than in the previous sections. Firstly, the equation will be a partial differential equation instead of an ordinary differential equation (a relation between a function of more than one variable and its partial derivatives). Secondly, different from the foregoing example, some of the equations to consider will be nonlinear, the importance of which will be explained. To start, however, we consider the simplest possible equations.

4.1 Linear first order wave equations

Let u be a scalar function of two variables, the time t and a spatial coordinate x : $u = u(x, t)$. Partial differentiation with respect to t will be denoted by ∂_t , and partial differentiation with respect to x as ∂_x or with a subscript x as in u_x or u_{xx} for the first and second partial derivative respectively. Consider for a given constant c_0 the following equation:

$$\partial_t u = -c_0 \partial_x u \quad (24)$$

This partial differential equation is of first order in time, and we can write down the solutions explicitly: for arbitrary function f the function

$$u(x, t) = f(x - c_0 t) \quad (25)$$

satisfies (24). In fact, (25) is the unique solution of (24) which satisfies the *initial condition*

$$u(x, 0) = f(x).$$

With the interpretation of u as the height of the free surface of a layer of water which extends in the x -direction, measured from some horizontal reference level, it will be clear why equation (24) is called a *wave equation*, and the solution (25) a *uniformly travelling wave*, see Fig.11.

In the x vs. t -plane the solution can be found from the observation that on lines $x - c_0 t = \text{constant}$ the function u satisfying (24) is constant. Stated differently, on the so-called *characteristics* the partial differential equation reduces to an ordinary differential equation:

$$\frac{d}{dt}u(x(t), t) = 0 \text{ on } \frac{d}{dt}x(t) = c_0.$$

For a numerical integration of the partial differential equation (and more difficult ones for which such characteristics can be found), this property can be used advantageously. Indeed, a simple *finite difference method* can be constructed by taking Euler discretisations in both the x and t -variable, for t the forward-Euler method with step Δt , and for x the backward difference with step Δx . Then the information about the value of u can be transported without distortion provided the quotient of the stepsizes is chosen appropriate. This is shown in Fig.12. Note that if the quotient $\frac{\Delta x}{\Delta t}$ is chosen different from the speed c_0 , this finite difference scheme will deform the solution in an unacceptable way, as in Fig.13, or the scheme will be unstable.

The following wave equation shares some, but not all, properties of (24):

$$\partial_t u = -c_0 \partial_x (u + \alpha u_{xx}). \quad (26)$$

Special solutions, so-called *monochromatic solutions*, can be found of the following form. For each value of the *wave number* k , the (real and imaginary part of the) function

$$\varphi_k(x, t) = \exp(i(kx - \omega t)) \quad (27)$$

is a solution provided the *frequency* ω is related to k according to

$$\omega = \omega(k) = c_0(k - \alpha k^3). \quad (28)$$

Clearly, each monochromatic solution is a travelling wave solution, periodic in time and in space, propagating with velocity

$$v_{\text{phase}}(k) = \frac{\omega(k)}{k},$$

the so-called phase velocity. For $\alpha \neq 0$ this propagation speed depends on the wave number. This has the effect that a superposition of two monochromatic solutions

Figure 11:

With f the wave shape at some initial time $t = 0$, the corresponding solution $u(x, t)$ of (24) denotes the wave shape at later times. This solution $u(x, t) = f(x - c_0 t)$ is a *uniformly travelling wave*: the wave shape f is translated undisturbed in shape with constant speed c_0 (in the positive x -direction if c_0 is positive).

Figure 12:

Simple finite difference method for (27); forward Euler in time (step Δt) and backward Euler in space (step Δx). On a grid in the x, t -plane determined by the stepsizes, the value at a point P is determined from the values at A and B at a previous time. From $\frac{u_P - u_B}{\Delta t} = -c_0 \frac{u_B - u_A}{\Delta x}$ the weight of these contributions follow as $u_P = (1 - \sigma)u_B + \sigma u_A$, with $\sigma = c_0 \frac{\Delta t}{\Delta x}$. For $0 < \sigma < 1$ this is a convex combination (linear interpolation) giving rise to a dissipative scheme, while for $\sigma > 1$, u_P is obtained from an extrapolation, leading to an unstable scheme. Only for $\sigma = 1$ the correct value at P is found from that at A .

Figure 13:

Distortion from finite difference method for the simple equation (24). Shown are the calculated wave profiles at different times. The bottom line represents the initial profile: a block function. The line above that is the evolution after 20 time steps, and so on for 400 time steps. For this calculation $\sigma = 0.35$ (see Fig.12), and the dissipation of the block function is clearly visible.

Figure 14:

Dispersion. Shown is the linear superposition of two waves in a frame of reference moving with the velocity of the largest wave. The smaller wave has relative amplitude $\frac{1}{3}$ and a three times larger velocity.

(two "modes") with different k will not behave as a uniformly travelling wave: the two modes will translate with different velocity and the total wave profile is the sum of the two shapes which translate with different speed. This effect is called *dispersion*, and the relation (28) is the dispersion relation. Equations such as (26) with this property are therefore called dispersive wave equations. See Fig.14.

The monochromatic solutions form a basis with which the general solution of (26) can be found using Fourier theory as follows. If an initial wave form is a given function f , either decaying sufficiently fast at infinity, or being a periodic function with, say, period 2π , its Fourier transform can be determined:

$$f(x) = \int_{-\infty}^{+\infty} \hat{f}(k) \exp(ikx) dk, \text{ respectively } f(x) = \sum_{k=-\infty}^{+\infty} \hat{f}(k) \exp(ikx).$$

Then the solution of (26) is the corresponding superposition of monochromatic modes with the same spectral function $\hat{f}(k)$:

$$u(x, t) = \int_{-\infty}^{+\infty} \hat{f}(k) \varphi_k(x, t) dk, \text{ respectively } u(x, t) = \sum_{k=-\infty}^{+\infty} \hat{f}(k) \varphi_k(x, t).$$

So the analytic solution of the initial value problem for (26) can be written down explicitly using Fourier theory.

The numerical calculation of (26) is much more difficult than that of (24) when finite difference schemes are considered, since for $\alpha \neq 0$ no characteristics can be used. Replacing the spatial derivative with some difference quotient, either central-, backward - or forward difference, errors are introduced which will deform the wave shapes from the analytic solution.

When dealing with the periodic case, i.e. restricting to functions which are periodic (with period 2π for simplicity), an alternative way is to approximate the solution at each time by a truncated Fourier series, with Fourier coefficients depending on time and to be determined. Writing (using real quantities again, instead of the complex coefficients above) for functions with mean value zero

$$u(x, t) = \sum_{k=1}^N a_k(t) \cos(kx) + b_k(t) \sin(kx), \quad (29)$$

substitution into equation (26) leads to

$$\sum_{k=1}^N \left((\dot{a}_k - \omega(k)b_k) \cos(kx) + (\dot{b}_k + \omega(k)a_k) \sin(kx) \right) = 0.$$

The dynamic equations for the coefficients follow easily: for $1 \leq k \leq N$

$$\begin{aligned} \dot{a}_k &= \omega(k)b_k \\ \dot{b}_k &= -\omega(k)a_k. \end{aligned} \quad (30)$$

This shows that the partial differential equation has been reduced to a set of ordinary differential equation for the coefficients, and we are left with a generalisation of the problem considered in the foregoing section. Indeed, for each k the equation in a_k and b_k is nothing but that of a harmonic oscillator, with frequency $\omega(k)$, and the time integration can be performed as discussed previously.

Remark Note, however, that each mode oscillates with its own frequency $\omega(k)$. Hence, for a reliable numerical calculation of the k -th mode the time step h_k should satisfy according to (10): $h_k < \pi \cdot 1/\omega(k)$. Taking $N = 3$ as the highest mode, for which $\omega = -24$ has the largest absolute value, the condition (10) gives for the maximal timestep h

$$h < \pi \cdot 1/\omega_{max} \approx .13 \text{ for } N = 3. \quad (31)$$

The results of Fig.16,17 and 18 indicate that (for the more difficult problem treated there) $h = .05$ is not sufficient, but that $h = .005$ (and in fact also $h = .01$) gives satisfactory results.

Hamiltonian formulation

Introducing the $2N$ -vector $z = (a_1, b_1; \dots; a_N, b_N)$, the equations can be written more concisely in various ways. For reasons to become clear shortly, we rewrite them as:

$$\dot{z} = -c_0 D \nabla H(z). \quad (32)$$

Here D is the $2N \times 2N$ -block-diagonal matrix corresponding to the differential operator ∂_x for the truncated functions:

$$D = \text{diag}(J, 2 \cdot J, \dots, N \cdot J), \text{ with } J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (33)$$

The function H is given by

$$H(z) = H((a_1, b_1; \dots; a_N, b_N)) = \sum_{k=1}^N \frac{1}{2} (1 - \alpha k^2) (a_k^2 + b_k^2). \quad (34)$$

The point we want to make in writing the equations as in (32), is that this set of equations is of the form of a Hamiltonian system, akin to (12). Now the symplectic matrix J is replaced by $c_0 D$, which, although not symplectic anymore, is skew-symmetric: $D^* = -D$. This suffices to call (32) a *generalised Hamiltonian system*. More precisely, the equation is of the form of a *Poisson system*, a notion that does not have to be explained here in detail. The most appealing fact is that for equation (32) the Hamiltonian H is conserved, just as for classical Hamiltonian systems, as a consequence of the skew-symmetry of D :

$$\frac{d}{dt} H(z) = \nabla H(z) \cdot \partial_t z = -c_0 \nabla H(z) \cdot D \nabla H(z) = 0. \quad (35)$$

Since in this example H is just the sum of the energies of all harmonic oscillators, H is the *total energy* of the system, and (35) expresses that this total energy remains constant during the evolution.

In the next subsection it is shown that the same holds true for the more difficult case of non-linear equations, for which the modes are no longer uncoupled as in (30), but are coupled through non-linear terms.

It may also be remarked here that the appearance of the Hamiltonian structure in the equations for the Fourier coefficients is not accidental. In fact, this structure is inherited from the continuous partial differential equation (26), which is also in the form of a generalised Hamiltonian system. The operator $c_0 \partial_x$ is a skew-symmetric operator (using the usual L_2 -innerproduct for periodic functions), and the expression between brackets $u + \alpha u_{xx}$ can be seen as the "derivative" (the variational derivative in this case) of the functional which is the total energy $\bar{H}(u)$ of the continuous system. This energy is now a functional, given by

$$\bar{H}(u) = \int_{-\pi}^{\pi} \left(\frac{1}{2} u^2 - \frac{1}{2} \alpha u_x^2 \right) dx, \quad (36)$$

and is also constant in time for solutions of (26) (which can be verified directly). There is no need to go any further in all these details here. But just note the optical resemblance of the equation (26) with the Hamiltonian form (32) when we write the variational derivative of $\bar{H}$ as $\delta \bar{H}(u)$ ¹:

$$\partial_t = -c_0 \partial_x \delta \bar{H}(u). \quad (37)$$

Also, note that inserting the Fourier truncation (29) of u into the expression for $\bar{H}(u)$, performing the integrations (here, using Parseval's identity), the expression (34) results for H : H is nothing but the truncation of the energy $\bar{H}$ of the continuous system:

$$H(z) = H(a_1, \dots, a_N, b_1, \dots, b_N) = \bar{H} \left(\sum_{k=1}^N a_k(t) \cos(kx) + b_k(t) \sin(kx) \right). \quad (38)$$

Remark: In general the Hamiltonian structure will be destroyed when using finite difference schemes for (26), unless very special precautions are taken. For one thing this means that the energy will not be conserved. The Fourier truncation,

¹For functionals defined on the set of periodic functions, like $\bar{H}$, the *variational derivative* is the generalization of the gradient of a function of a finite number of variables. In fact, it is defined as that function $\delta \bar{H}(u)$ such that for each other periodic function η it holds that

$$\int_{-\pi}^{\pi} \delta \bar{H}(u), \eta dx = \frac{d}{d\epsilon} \bar{H}(u + \epsilon \eta) |_{\epsilon=0}.$$

For the given expression for $\bar{H}$ it holds $\bar{H}(u + \epsilon \eta) = \bar{H}(u) + \epsilon \int_{-\pi}^{\pi} (u\eta - \alpha u_x \eta_x) dx + O(\epsilon^2)$. Integrating by parts the order ϵ term gives $\int_{-\pi}^{\pi} (u\eta - \alpha u_x \eta_x) dx = \int_{-\pi}^{\pi} (u + \alpha u_{xx}) \eta dx$ which shows that $\delta \bar{H}(u) = u + \alpha u_{xx}$.

however, has retained the Hamiltonian structure of the original continuous system: the discretisation correctly *models* the basic structure, and hence the conservation property, of the original problem.

4.2 Nonlinear equation

For the linear equation considered above, each Fourier mode behaves independently from the other modes: the equations for a_k and b_k only contains these coefficients, no coefficients a_m or b_m with $m \neq k$. This is a consequence of the linearity of the original equation. When dealing with nonlinear equations, this does not hold any longer: the various Fourier modes will interact, causing a much richer (more difficult) dynamical behaviour.

As an example, we will consider the **Korteweg-de Vries equation**. This partial differential equation describes many physical phenomena in which dispersive and nonlinear effects are present in the same order of magnitude. For instance, the equation was derived originally to describe the evolution of surface waves on a layer of fluid, in the approximation of waves running (mainly) in one direction only.²

The equation is given by

$$\partial_t u = -c_0 \partial_x (u + \alpha u_{xx} + 3\beta u^2), \quad (39)$$

or in a normalised form, which will be used in the following, by³

$$\partial_t u = -\partial_x (u_{xx} + 3u^2). \quad (40)$$

This equation has several very special properties and has given rise to very far-going new developments in the theory of partial differential equations (complete integrability). At this place we will only touch briefly on some of these (less far reaching) properties.

The first observation is that, just as the linear equation (26), this equation has a Hamiltonian structure, i.e. is of the form of equation (37). Now the energy $\bar{H}(u)$ contains a non-quadratic term. For (40) the energy is given by

$$\bar{H}(u) = \int_{-\pi}^{\pi} \left(\frac{1}{2} u_x^2 + u^3 \right) dx. \quad (41)$$

The variational derivative of $\bar{H}(u)$ turns out to produce precisely the term in brackets in (40): $\delta \bar{H}(u) = u_{xx} + 3u^2$, and so the equation is indeed of the form

²The validity of this equation is restricted to "rather low, rather long" waves, for which dispersive and nonlinear terms are of comparable order, but small compared to the basic equation (24). This means that in reality the coefficients α and β in (39) should be small.

³When transforming to a frame of reference moving with velocity c_0 , the equation (39) becomes: $\partial_t u = -c_0 \partial_x (\alpha u_{xx} + 3\beta u^2)$. Then, by scaling the independent variables x and t and the variable u in an appropriate way, it is possible to arrive at the equation (40) in the scaled variables.

(37) of a Hamiltonian system.

This may lead immediately to the idea how consistent discretisations for (39) can be found using Fourier truncation as before. Indeed, based on this Hamiltonian structure, we could exploit the observation made for the linear equation. That is, substitution of the truncation for u in $\tilde{H}(u)$ will produce a function of the Fourier coefficients, the truncated energy as in (38). Explicitly, after some tedious calculations (involving the integration of the product of combinations of three sine and cosine functions from the cubic term in the Hamiltonian), there results

$$H(z) = H_2(z) + \beta H_3(z), \quad (42)$$

where H_2 is the quadratic part as in (34), and H_3 is from the cubic terms.

The Hamiltonian being determined in this way, we take as discretisation of (39) simply the Hamiltonian equations given by (32). For each k they are of the form

$$\begin{aligned} \dot{a}_k &= -k \frac{\partial H}{\partial b_k} \\ \dot{b}_k &= k \frac{\partial H}{\partial a_k} \end{aligned} \quad (43)$$

For example, with $N = 3$, the Hamiltonian reads

$$H_3(z) = -1.5a_1^2a_2 + 1.5b_1^2a_2 - 3(a_1(b_2b_1 + a_3a_2 + b_2b_3) + (a_2b_3 - b_2a_3)b_1) \quad (44)$$

and the corresponding equations are:

$$\begin{aligned} \dot{a}_1 &= -b_1 - 3(a_3b_2 + a_2b_1 - a_1b_2 - a_2b_3) \\ \dot{b}_1 &= a_1 - 3(a_3a_2 + a_2a_1 + b_1b_2 + b_2b_3) \\ \dot{a}_2 &= -8b_2 - 6(a_3b_1 - a_1b_1 - a_1b_3) \\ \dot{b}_2 &= 8a_2 + 6(-a_3a_1 - 0.5a_1^2 + 0.5b_1^2 - b_1b_3) \\ \dot{a}_3 &= -27b_3 + 9(a_2b_1 + a_1b_2) \\ \dot{b}_3 &= 27a_3 + 9(-a_2a_1 + b_1b_2). \end{aligned} \quad (45)$$

Equations like those for the coefficients of the Fourier modes are usually referred to as mode equations. They can be interpreted to describe three *coupled nonlinear oscillators*, the coupling in this case being a consequence of the nonlinearity. Clearly, higher order Fourier truncations will lead to equally more mode equations, more coupled oscillators, all in interaction with (most of) the others.

Mode interaction

The equations (45) describe the three-mode interaction: due to the nonlinear terms, which result from H_3 in the Hamiltonian, the modes are coupled. For instance, the change in a_1 is determined not only by b_1 but also by the components of the other modes a_2, a_3, b_2, b_3 as is seen from the equations (45). See Fig.15. It will be observed that the interaction process is quite complicated. Many small scale features are visible in the trajectory of the first mode, more clearly visible in

Figure 15:

Mode interaction for the 3-mode system (44). Starting with an initial condition for which the second mode is zero at $t = 0$, this mode is excited because of the nonlinear terms in (44). For initial data $(a_1, b_1; a_2, b_2; a_3, b_3) = (1, 0; 0, 0; 1, 0)$ the figure shows: UL: The evolution of the first mode in the a_1, b_1 plane of the six-dimensional phase space (for about a three times longer time period than the other pictures in this figure). UR: The evolution of the strength of the first mode $\sqrt{a_1^2 + b_1^2}$, which is not constant because of interaction with the other modes. LL: The evolution of the second mode (starting at zero) in the a_2, b_2 plane of the phase space. LR: The evolution of the strength of the second mode $\sqrt{a_2^2 + b_2^2}$, showing the excitation from and interaction with the other modes.

Figure 16:

Mode interaction for the 3-mode system (44). Starting with the same initial data as in Fig.15, the trajectory of the first mode is shown, using two different timesteps. In the upper picture the timestep is $h = .005$, the same as used to produce Fig.15 and 17. The lower picture has been calculated, for the same initial condition, with timestep $h = .05$, the same as used to produce Fig.18. Then the trajectory is transferred in the *opposite* direction! Moreover, the fine scale structure showing the presence of another time scale in the upper picture, is lost in the lower picture. The dependence on the timestep is drastically shown in this picture.

Figure 17:

Evolution of the wave profile $u(x, 0) = \cos x + \cos 3x$, corresponding to the initial data $(a_1, b_1; a_2, b_2; a_3, b_3) = (1, 0; 0, 0; 1, 0)$ of Figs.15 and 16. The horizontal axis has length two times the spatial period: $2 \times 2\pi$. Shown are the profiles at different times: starting at $t = 0$, the next profiles are at $t = \frac{1}{2}k$, with $1 \leq k \leq 14$, each one shifted vertically upwards over a fixed amount. The timestep $h = .005$ corresponds to the time step used to produce Fig.15 and the upper part of Fig.16.

Figure 18:

Evolution of the same wave profile as in Fig.17, but now calculated with timestep $h = .05$, corresponding to the timestep used to produce the lower part of Fig.16. Observe that all small scale features are quickly fading away, resulting in a profile that differs essentially from the one calculated in Fig.17.

the upper part of Fig.16. The corresponding evolution of the wave profile is given by the three term Fourier approximation (29) and is shown in Fig.17.

Comparing the lower part of Fig.16 with the upper part, and Fig.18 with Fig.17, shows that the calculated results depend very much on the timestep. Even for this simple three-mode model, the timestep has to be taken less than .01 in order to obtain results that capture the fine scale structures. When calculating real life problems, such as for instance atmospheric models for weather forecasts, or flows around wings, thousands and even millions of such coupled equations have to be calculated. The relatively simple example considered here illuminates some of the problems that are encountered there.

Travelling waves

Above we solved an initial value problem: given the profile at $t = 0$ we calculated the profile at later times. Now we will deal with a completely different problem. We will address the question whether there exists a *travelling wave*: a specific wave profile that is translated at constant speed undeformed in shape.

For a nonlinear equation as the one under consideration, that will be an exceptional solution, in contrast to the simple linear equations (24) for which each solution is of this form, or the linear dispersive equation (26) for which monochromatic solutions are of this kind.

We will investigate some aspects, both for the KdV equation itself and its N -mode truncation. Upon substitution in (39) the general form of a travelling wave, say a wave with profile $\hat{u}$ and velocity λ :

$$u(x, t) = \hat{u}(x - \lambda t), \quad (46)$$

there results an *ordinary* differential equation for the profile function $\hat{u}$ in which the velocity λ enters as a parameter. It can be shown that this equation admits (for appropriate values of the constants) 2π -periodic solutions. Since these solutions can be written down "explicitly" using the elliptic function that is known as the *cnoidal function*, these waves are called *cnoidal waves*.

Specifically, the equation for the function $\hat{u}(x)$ reads:

$$\lambda \partial_x \hat{u} = \partial_x (\hat{u}_{xx} + 3\hat{u}^2)$$

and can be integrated once, introducing a constant of integration σ :

$$\lambda \hat{u} = (\hat{u}_{xx} + 3\hat{u}^2) + \sigma. \quad (47)$$

This second order equation can be studied thoroughly. For instance, it can be brought to a Hamiltonian form again (with x replacing the time variable). Consequently, its Hamiltonian is constant (independent of x). This can also be found in a more direct way by multiplying equation (47) by u_x and integrating again with respect to x . The result is a first order equation for u of the form

$$u_x^2 = Q(u),$$

where $Q(u)$ is some cubic polynomial, depending on several constants, in particular on λ . Looking for a solution that is symmetric around $x = 0$ (without restriction of generality), this equation can be solved by separation as follows. If u_0 and u_π denote the value at $x = 0$ and $x = \pi$ respectively (the value at the crest and in the valley), which are chosen such that $Q(u) > 0$ for $u_\pi < u < u_0$, $Q(u_0) = Q(u_\pi) = 0$, then

$$\int_{u(x)}^{u_0} \frac{du}{\sqrt{Q(u)}} = x$$

describes implicitly the profile for $0 < x < \pi$. Periodicity requires that the coefficients are chosen such that $\int_{u_\pi}^{u_0} \frac{du}{\sqrt{Q(u)}} = \pi$. This still leaves a free parameter to adjust to the mean value of u , which can also be normalised to 0. See for more details e.g. Whitham, 1976.

Now that we know that the exact continuous equation has travelling waves, one could hope that its discretisation has such solutions too. In general that cannot be expected and the next statement is exceptional.

Proposition. *For each N , the N -mode Fourier truncation of the KdV-equation has exact travelling waves (2π -periodic in space).*

We will first present some "numerical evidence" for this statement, for $N = 3$, and then give the analytical arguments of the proof.

Numerical evidence.

It is *not* possible to find a periodic solution numerically by trial and error. We will show that the following set of data provide the desired travelling wave when $N = 3$. How these data are determined cannot be understood without referring to the analytic proof to follow. For the moment we just accept them as given. These data -for the resulting travelling wave velocity λ , and the Fourier coefficients- are as follows (the meaning and value of I will be explained later):

$$\begin{aligned} \lambda &= -3.65497558464, \\ \text{for } I = 2\pi, \quad \bar{a}_1 &= 1.69505557741, \quad \text{and} \quad \bar{b}_1 = \bar{b}_2 = \bar{b}_3 = 0 \\ \bar{a}_2 &= 0.97285831670, \\ \bar{a}_3 &= 0.42465666678, \end{aligned} \quad (48)$$

The Fourier coefficients are used as initial data for (45) and the resulting dynamics is calculated numerically with a small timestep. In the next figures this solution is depicted and interpreted in various ways. In Fig.19 the evolution of each of the three modes is shown in the phase space. For the specific initial value each mode behaves as an oscillator exhibiting a pure harmonic oscillation, without coupling with the other modes. This is shown in a different way in Fig.20.

The spatial wave profile corresponding to the Fourier coefficients of the special initial data is shown in the left picture of Fig.21. The dynamic evolution of the Fourier coefficients gives a profile that is shown in Fig.22. This picture gives optical evidence that the wave travels undisturbed in shape, at constant velocity, a

Figure 19:

For initial data given by (47), each mode behaves like a harmonic oscillator, with no visible sign of interaction. The time step used is $h = .01$.

Figure 20:

A more precise representation shows that the amplitudes of the modes actually do vary a little bit (in the fifth digit), as is seen from the UL and LL picture for this timestep $h = .01$ (scales are: 2.87321-2.87322 for the first mode, and .94645-.94646 for the second mode). The pictures at the right show again these modes, now calculated for the same initial data but with a smaller timestep: $h = .005$. Comparison of the pictures at the left and at the right indicates that the variation in mode amplitude is only caused by numerical errors.

Figure 21:

The wave profile obtained with the N -mode Fourier truncation, using as coefficients the data obtained from solving the constrained optimisation problem. In the picture at the left, calculated with $N = 3$, the low number of modes produce a wavy profile in the valley, which is not expected in reality. This wavy part disappears almost completely when more modes are taken into account: for $N = 6$ the profile, found by solving the optimisation problem with $N = 6$, is given in the right part of the picture.

Figure 22:

Evolution of the wave profile with $N = 3$ from Fig.21 according to the dynamical equations (44). As expected, the wave travels undisturbed in shape, at constant velocity. The calculations are performed with timestep $h = .005$, and the profiles are shown at intermediate time intervals $20h$, so for a total time interval of length 2. The shift during this time agrees with the given value of the velocity λ .

pure translation as claimed.

Analytic proof.

Of course, despite the numerical evidence given above -which may be more or less convincing depending on one's attitude- the actual proof of the statement should follow from analytic arguments. These arguments are in fact quite involved when presented without any background information, but they are a straightforward application of general ideas about *relative equilibria* in Hamiltonian systems. For the sake of this presentation we will state each of the relevant steps. Each partial statement can be verified directly (with some effort), thereby producing in total a complete proof of the proposition.

Step 1.

Define the following function of the Fourier coefficients:

$$I(z) := \frac{1}{2} \sum_{k=1}^{k=N} (a_k^2 + b_k^2), \quad (49)$$

the sum of the squared amplitudes of the modes. We will call this function the *momentum*, for reasons to become clear. This function is a constant of the motion:

$$\text{for solutions } z(t) \text{ of (32) it holds } \frac{d}{dt} I(z(t)) = 0.$$

Step 2.

With the energy H and I as two constants of the motion for (32), consider the following *constrained optimisation problem*: for given γ positive, find the maximiser of H on the level set of I :

$$\text{Min } \{H(z) | I(z) = \gamma\}. \quad (50)$$

There exist solutions of this minimization problem, (one of) which we will call $\bar{z}$ (there exist also other critical points). This minimizer satisfies the equation found from applying Lagrange's multiplier rule for constrained optimisation problems: for some "multiplier" $\lambda \in \mathbb{R}$, $\bar{z}$ satisfies

$$\nabla H(\bar{z}) = \lambda \nabla I(\bar{z}). \quad (51)$$

This minimization problem can be solved numerically (if necessary, by hand for $N = 3$). For $N = 3$ one solution is given by the data (approximated to a few digits) in (48), where λ is the multiplier and the Fourier coefficients form the vector $\bar{z}$, when for the value of the constrained is taken (rather arbitrarily) $\gamma = 2\pi$.

Step 3.

When a periodic function $u(x)$ is shifted over some distance ϵ say, the Fourier coefficients of $u(x + \epsilon)$ are related in a simple way to those of $u(x)$. This defines a mapping, which we will call T_ϵ , and which is determined from:

$$\text{if } z = (a_1, \dots, b_n) : u(x) = \sum (a_k \cos kx + b_k \sin kx),$$

then $T_\epsilon z$ follows from :

$$u(x + \epsilon) = \sum a_k \cos k(x + \epsilon) + b_k \sin k(x + \epsilon) \quad (52)$$

$$= \sum (a_k \cos k\epsilon + b_k \sin k\epsilon) \cos kx + (-a_k \sin k\epsilon + b_k \cos k\epsilon) \sin kx. \quad (53)$$

It can be verified that T_ϵ can also be interpreted in the following way: $T_\epsilon(z_0)$ is the solution starting at $t = 0$ at $z = z_0$ after time $t = \epsilon$ of the following equation:

$$\dot{z}(t) = -D\nabla I(\bar{z}(t)) = -Dz, \quad (54)$$

i.e. the same equation as (32), but now with the function I introduced above as the Hamiltonian. Hence $z(\epsilon) = T_\epsilon(z_0)$. For this reason T_ϵ is called the "flow" of the equation (54). It changes the Fourier coefficients in such a way that the function merely experiences a shift. It needs hardly any calculation to convince oneself that both the original Hamiltonian H and the momentum I are invariant for such a shift:

$$\text{for all } \epsilon: I(T_\epsilon z) = I(z), \text{ and } H(T_\epsilon z) = H(z). \quad (55)$$

(Just differentiate the expressions with respect to ϵ , and use the (essential!!) fact that I is a constant of the motion for (32)).

Step 4.

Now combine the facts found in Steps 2 and 3. Since $\bar{z}$ is a constrained minimizer, and since the values of both H and I do not change when applying T_ϵ , for each ϵ also $\bar{z}(\epsilon) := T_\epsilon \bar{z}$ is a minimizer and satisfies the same equation as $\bar{z}$ itself:

$$\nabla H(\bar{z}(\epsilon)) = \lambda \nabla I(\bar{z}(\epsilon)), \quad \text{for all } \epsilon. \quad (56)$$

(This explains why there are many more solutions of the minimization problem; for the solution given in (48) the sine-coefficients were taken identically zero from the onset, invoking the symmetry of the profile to be expected). Consequently, writing $\epsilon = c_0 \lambda t$, the function $t \rightarrow \bar{z}(t)$ satisfies the equation

$$\dot{\bar{z}}(t) = -c_0 \lambda D\nabla I(\bar{z}(t)) = -c_0 D\nabla H(\bar{z}(t)). \quad (57)$$

Consequently, $t \rightarrow \bar{z}(t)$ satisfies the required equation, and is at the same time merely a translation. This completes the proof of the proposition!

Remark. All steps above are equally true for the partial differential equation itself, and they are in fact a consequence of that. Since (39) has constant coefficients (independent of x), there is translation symmetry. Related to this, the functional $\bar{I}(u) = \int \frac{1}{2} u^2$ is a constant of the motion. The Fourier truncation of $\bar{I}$ is precisely the function I given by (49). Then with $\bar{H}$ given by (41) and $\bar{I}$ one can consider the optimisation problem akin to (50). With functional analytic tools it can be shown that a minimizer to this (infinite dimensional !!) minimization problem exists. This function has to satisfy the equation that replaces (51), which turns out to be precisely equation (47). (The constant σ is another multiplier from the restriction to functions with mean value zero.) Since the solution of (47) is the travelling wave profile, and λ its speed, there is a complete analogy with the Fourier truncation.

Remark. The key point for the existence of exact travelling waves in the discretised equations is that there are *two* constants of the motion, just as for the original equation (39). One, the energy H , is a result of the Hamiltonian character of the equations, a property that can also be obtained (with some effort) for finite difference discretisations. However, the existence of the other one, I in this case, is directly related to the fact that the continuous symmetry in the partial

differential equation (39) is preserved as a *continuous* symmetry in the discretised equations. Here this is owing to the fact that we used Fourier mode truncations. Finite difference methods will not preserve this continuous translation symmetry, and have no additional constant of the motion in general. Concluding, travelling waves can be retained after discretisation if analytical knowledge is exploited carefully.

5 Epilogue

Let us review the foregoing examples and see what can be learned from them. The first example -the logistic equation- showed the phenomenon of chaos. This was seen to appear for the simple Euler discretisation, and for large time steps as a consequence of the nonlinearity. However, for other discretisation schemes -all of which are in some way or another a modification of Euler's method- the same phenomenon can be expected to happen in general, possibly at larger time steps. For this simple example the time step had to be taken unrealistically large to observe the chaos. However, as is clear from the example on wave equations, when dealing with discretisations of partial differential equations, many (coupled, nonlinear) equations have to be solved, each of which has its own time scale. A time step which is small for some of them (as for the lowest modes in the Fourier truncation) may turn out to be large for others (the higher Fourier modes).

For the second example -the harmonic oscillator- we investigated energy conservation. Even when the energy was not conserved at the exact value, it was shown that an unbounded increase, or dissipation to a trivial state, can be avoided. The phase error made in the calculation of each mode cannot be overcome so easily when dealing with more modes of different frequency. Consequently, when the solution is obtained from a superposition of several calculated modes as in the example of the wave equation, the actual profile may be far off from the exact profile: in a pointwise norm, the error will be quite large. Although this may give some uneasy feelings, it seems an inevitable consequence of the numerical calculations, and calls for another measure to determine the quality of the calculation.

One such measure is the constancy of the constants of the motion, such as H and I above. Of course, this gives only a very rough indication since these are global quantities, taking contributions of all modes together. A more precise measure is the "spectrum" of the solutions, i.e. the amplitude of each mode (very often the quantity $a_k^2 + b_k^2$ is called the "energy" of the k -th mode, and one talks about the energy-spectrum. In the KdV-equation, this terminology is not consistent with the physical energy, which is approximated by H .) In general, this spectrum will vary in time because of (nonlinear) mode interactions. Then some time average can be taken. When compared with a more precise calculation (for instance the double number of modes), this gives some indication whether the individual modes, and their interactions, are calculated reliably or not. For the travelling waves considered in section 4, the spectrum is time independent and approximates the spectrum of the partial differential equation as good as possible (given the number

of modes) as a consequence of the optimization property.

Software.

All calculations were performed on a regular PC (with mathematical coprocessor). There are various simple software packages for investigating ordinary differential equations. We mention PHASER and PHASEPLANE (see the references; the software comes along with the book). DYNPAO is another package, developed by R.A. Posthumus, University of Groningen, and is available upon request. Figures 1,5,6,7,8,9,10,15,16,19, and 20 were prepared with Phaseplane, and Figures 2,3, and 4 with Dynpao (all are direct copies of the screen).

For the wave equation, Figures 13,14,17,18,21 and 22 were obtained with PASCAL programmes, written (adapted) for these examples by van Beckum.

In section 4 some analytical calculations are needed to obtain the Hamiltonian (44) and, by taking its gradient, to find the equations (43). These calculations can be performed by hand or with an algebraic software package, such as DERIVE, MAPLE or MATHEMATICA. In order to avoid mistakes in the calculations on paper and in typing the commands, I used both methods. The results with the software packages can be used, by copying without retyping, in the Pascal programmes and in the final text of this report. These simple calculations give experience and confidence for dealing in the same way with more difficult problems.

A software package WAVEPACK is being developed at this moment as part of a cooperation between the mathematics institutes of the University of Twente and Institut Teknologi Bandung. WAVEPACK is aimed to be useful for both didactical and for research purposes. It visualises elementary concepts like superposition, dispersion, groupvelocity and Fourier mode interactions. At the same time it also allows to calculate the propagation of waves for KdV- and other equations. A particular facility enables the calculation of exact wave profiles of a soliton (or cnoidal wave) from an optimization problem. Also the distortion of waves over uneven bottoms can be calculated, as well as the effects of damping (viscosity) and selfexcitation. At this moment (April '92) a preliminary version is available. Interested readers can get a copy of the software, together with an extensive manual (including theory), from the author (at costprice, until further notice).

Acknowledgement.

Most of these ideas have been developed in discussions with F. van Beckum. Despite his great skill and experience in numerical calculations of much more difficult problems, he always showed great interest in my simple examples and questions. I am also gratefull for his direct help, in particular with the Pascal programmes.

References

- F. van Beckum & E. van Groesen: *Spatial discretizations of Hamiltonian Wave Equations*, University of Twente Memorandum, 1992, to be published.
- H.W. Broer, F. Dumortier, S.J. van Strien & F. Takens: *Structures in Dynamics, Finite dimensional deterministic studies*, Vol.2, North-Holland

Studies in Mathematical Physics, Amsterdam, 1991.

- R.L. Devaney: *An introduction to Chaotic Dynamical Systems*, Benjamin/Cummings, 1986.
- B. Ermentrout: *Phase Plane, The Dynamical Systems Tool*, Version 3.0, Brooks/Cole, California 1990.
- E. van Groesen, F. van Beckum, S. Redjeki & W. Djohan: WAVEPACK, *computer software for wave equations, including manual and theory*. Preliminary version, April 1992. Available upon request: E. van Groesen, Applied Mathematics, Univ. of Twente, P.O.Box 217, 7500 AE Enschede, The Netherlands.
- J. Guckenheimer & P. Holmes: *Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields*, Springer Verlag, Berlin, 1983.
- H. Koçak: *PHASER Differential and Difference Equations through Computer Experiments*, Springer Verlag 1986.
- H.G. Schuster: *Deterministic Chaos*, Physik Verlag, Weinheim, 1984.
- J.B. Whitham: *Linear and Nonlinear Waves*, Wiley, New York, 1974.