Lecture 2

Computational and Errors

Numerical methods are procedures that allow for efficient solution of a mathematically formulated problem in a finite number of steps to within an arbitrary precision. Computers are needed in most cases. A very important issue here is the errors caused in computations.
A numerical algorithm consists of a sequence of arithmetic and logical operations which produces an approximate solution to within any prescribed accuracy. There are often several different algorithms for the solution of any one problem. The particular algorithm chosen depends on the context from which the problem is taken. In economics, for example, it may be that only the general behavior of a variable is required, in which case a simple, low accuracy method which uses only a few calculations is appropriate. On the other hand, in precision engineering, it may be essential to use a complex, highly accurate method, regardless of the total amount of computational effort involved.
Once a numerical algorithm has been selected, a computer program is usually written for its implementation. The program is run to obtain numerical results, although this may not be the end of the story. The computed solution could indicate that the original mathematical model needs modifying with a corresponding change in both the numerical algorithm and the program.
Although the solution of 'real problems' by numerical techniques involves the use of a digital computer or calculator, Determination of the eigenvalues of large matrices, for ex-

ample, did not become a realistic proposition until computers became available because of the amount of computation involved. Nowadays any numerical technique can at least be demonstrated on a microcomputer, although there are some problems that can only be solved using the speed and storage capacity of much larger machines.

There exist four possible sources of error:

1. **Errors in the formulation of the problem** to be solved.
   (a) Errors in the mathematical model. For example, when simplifying assumptions are made in the derivation of the mathematical model of a physical system. (Simplifications).
   (b) Error in input data. (Measurements).
2. **Approximation errors**
   (a) Discretization error.
   (b) Convergence error in iterative methods.
   (c) Discretization/convergence errors may be estimated by an analysis of the method used.
3. **Roundoff errors**: This error is caused by the computer representation of numbers.
   (a) Roundoff errors arise everywhere in numerical computation because of the finite precision arithmetic.
   (b) Roundoff errors behave quite unorganized.
4. **Truncation error**: Whenever an expression is approximated by some type of a mathematical method. For

example, suppose we use the Maclaurin series representation of the sine function:

$$\sin\alpha = \sum_{n=odd}^{\infty} \frac{(-1)^{\frac{(n-1)}{2}}}{n!}\alpha^n = \alpha - \frac{1}{3!}\alpha^3 + \frac{1}{5!}\alpha^5 - \cdots + \frac{(-1)^{\frac{(m-1)}{2}}}{3!}\alpha^m + E_m$$

where E_m is the tail end of the expansion, neglected in the process, and known as the truncation error.

0.1 ERRORS AND STABILITY

The majority of numerical methods involve a large number of calculations which are best performed on a computer or calculator. Unfortunately, such machines are incapable of working to infinite precision and so small errors occur in nearly every arithmetic operation. Even an apparently simple number such as $2/3$ cannot be represented exactly on a computer. This number has a non-terminating decimal expansion

$$0.66666666666666\cdots$$

and if, for example, the machine uses ten-digit arithmetic, then it is stored as

$$0.666\,666\,666\,7$$

(In fact, computers use binary arithmetic. However, since the substance of the argument is the same in either case, we restrict our attention to decimal arithmetic for simplicity). **The difference between the exact and stored values is called the rounding error** which, for this example, is

$$-0.000\,000\,000\,033\,33...$$

Suppose that for a given real number α the digits after the decimal point are

$$d_1 d_2 \cdots d_n d_{n+1} \cdots$$

To round α to n decimal places (abbreviated to nD) we proceed as follows. If $d_{n+1} < 5$, then α is rounded down; all digits after the nth place are removed. If $d_{n+1} \geq 5$, then α is rounded up; d_n is increased by one and all digits after the nth place are removed. It should be clear that in either case the magnitude of the rounding error does not exceed 0.5×10^{-n}.
In most situations the introduction of rounding errors into the calculations does not significantly affect the final results. However, in certain cases it can lead to a serious loss of accuracy so that computed results are very different from those obtained using exact arithmetic. The term instability is used to describe this phenomenon.
There are two fundamental types of instability in numerical analysis - **inherent** and **induced**. The first of these is a fault of the problem, the second of the method of solution.

Definition 2. *A problem is said to be **inherently unstable** (or **ill - conditioned**) if small changes in the data of the problem cause large changes in its solution.*

This concept is important for two reasons. Firstly, the data may be given as a set of readings from an analogue device such as a thermometer or voltmeter and as such cannot be measured exactly. If the problem is ill-conditioned then any numerical results, irrespective of the method used to obtain them, will be highly inaccurate and may be worthless. The second reason is that even if the data is exact it will not necessarily be stored exactly on a computer. Consequently, the problem which the computer is attempting to solve may differ slightly from the one originally posed. This does not usually matter, but if the problem is ill-conditioned then the computed results may differ wildly from those expected.

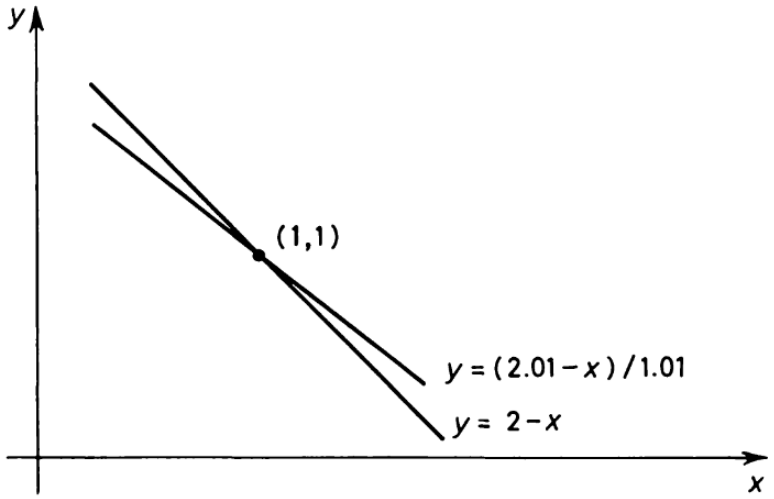

Figure 1: sketche of example 0.3

Example 0.3. *Consider the simultaneous linear equations*

$$x + y = 2$$
$$x + 1.01y = 2.01$$

which have solution $x = y = 1$*. If the number* 2.01 *is changed to* 2.02*, the corresponding solution is* $x = 0$*,* $y = 2$*. We see that a* 0.5% *change in the data produces a* 100% *change in the solution. It is instructive to give a geometrical interpretation of this result. The solution of the system is the point of intersection of the two lines* $y = 2 - x$ *and* $y = (2.01 - x)/1.01$*. These lines are sketched in figure 1. It is clear that the point of intersection is sensitive to small movements in either of these lines since they are nearly parallel. In fact, if the coefficient of* y *in the second equation is* 1.00*, the two lines are exactly parallel and the system has no solution. This is fairly typical of ill-conditioned problems. They are often close to 'critical' problems which either possess infinitely many solutions or no solution whatsoever.*

Example 0.4. *Consider the initial value problem*

$$y'' - 10y' - 11y = 0; \quad y(0) = 1, \quad y'(0) = -1$$

defined on $x \geq 0$*. The corresponding auxiliary equation has roots* -1 *and* 11*, so the general solution of the differential equation is*

$$y = Ae^{-x} + Be^{11x}$$

for arbitrary constants A *and* B*. The particular solution which satisfies the given initial conditions is*

$$y = e^{-x}$$

Now suppose that the initial conditions are replaced by

$$y(0) = 1 + \delta, \qquad y'(0) = -1 + \epsilon$$

for some small numbers δ *and* ϵ*. The particular solution satisfying these conditions is*

$$y = \left(1 + \frac{11\delta}{12} - \frac{\epsilon}{12}\right) e^{-x} + \left(\frac{\delta}{12} + \frac{\epsilon}{12}\right) e^{11x}$$

and the change in the solution is therefore

$$\left(\frac{11\delta}{12} - \frac{\epsilon}{12}\right) e^{-x} + \left(\frac{\delta}{12} + \frac{\epsilon}{12}\right) e^{11x}$$

The term $\dfrac{(\delta + \epsilon)e^{11x}}{12}$ *is large compared with* e^{-x} *for* $x > 0$*, indicating that this problem is ill-conditioned.*
To inherent stability depends on the size of the solution to the original problem as well as on the size of any changes in the data. Under these circumstances, one would say that the problem is ill-conditioned.

We now consider a different type of instability which is a consequence of the method of solution rather than the problem itself.