Basic notions: numerical methods, modeling, types of errors, correctness, stability

By Snezhana G. Gocheva-Ilieva, Plovdiv University,
snow@pu.acad.bg

Throughout its history from its creation up to now the mathematics has dealt with practical problems. In the ancient times it served to calculate lengths, surfaces and volumes, in the Renaissance – for discovering the qualitative laws in physics, astronomy and other sciences. Nowadays particularly intensively develops the calculation aspect of mathematics in parallel with its high degree of abstraction. A separate branch of it called numerical analysis and numerical methods have formed.

Subject of numerical methods is the development, studying and implementation of numerical algorithms for solving the wide number of basic mathematical and practical problems. Numerical methods are developed in almost all aspects of the mathematics itself – mathematic analysis, algebra, geometry, and differential equations etc. that enable solving specific problems.

The majority of the existing methods are orientated to implementation of computers and obtaining the end results in numerical form. It is incorrect to consider that solving of particular problem can be attained directly through implementation of one or few numerical methods. Quite the contrary, the way from setting the problem to attaining the end results of the calculations suitable for implementation in practice usually is rather long. First numerically and statistically defined number of observations, measurements and experiments has to be systemized and designed. They form the initial data of the problem.

The data are obtained with certain acceptable degree of accuracy regardless the accuracy of instruments or the quantity of. That is called irremovable error or initial data error. For example the length may be measured with accuracy ± 5 mm, the weight with accuracy ± 0.001gr etc.

The second step towards solving of the problem is the most difficult and has a decisive role. It is about composing adequate mathematical model, which describes precisely enough the investigated process, phenomenon or complex of phenomena. The mathematical model may be system of algebraic equations, system of partial equations or another mathematical problem. For example an arbitrary chemical equation that represents the process of reaction between specific substances and the obtaining of another substance is described mathematically with a system of ordinary differential equations at some initial conditions.

Different mathematical models may be constructed for one problem and every one to approximate more or better to the investigated problem. The error that is made at that stage is called model error.

Only when there are initial data and the mathematical model is selected comes the time to employ some numerical method. On the other side every numerical method creates error of approximation called error of the method. We have to mention that unlike the initial data and the model errors the error of the numerical method in principle can not be estimated beforehand. That makes possible to set the accuracy that the numerical method will guarantee for the obtained results.

The last stage of the overall solving of practical problem is the calculation. Manually done or by the aid of computer this inevitably impose rounding of the intermediate and the final results, an error as a result of transition from one counting system to another etc. Therefore we do not obtain exact solution either but some approximation of it. The error made on this stage is called error of calculation and it can be estimated though more difficultly.

In this manner the following four types of errors emerge through the process of solving certain practical problem:

- initial data error, referred as irremovable error;
- error of the model;
- error of the numerical method;
- error of calculation.

The overall error is called total error.

Approximation of numbers. Absolute and relative error

Finally the implementation of numerical methods is reduced to performing definite number of arithmetic operations with numbers. That is why naturally arises the issue of the approximation of numbers and evaluation of the error of the approximation.

Let us for example the number $\bar{x} = \frac{2}{3}$ is given. That is exact rational number. However manually or by machine it is more convenient to work with decimal fractions. Then $\bar{x} = 0.666666...$

Apparently we can perform operations or even to write exactly one infinite continued fraction. That's why it is necessary to introduce approximations of $\bar{x}$. One approximation is for example the number

$$\tilde{x}_1 = 0.66 .$$

Another approximations are

$$\tilde{x}_2 = 0.6667$$

and

$$\tilde{x}_3 = 0.666432 .$$

Every one of these numbers may be replaced by the other, but of course with certain accuracy. How as a matter of fact we should evaluate the closeness or the error of the approximation? Most often we use the notions of absolute and relative error.

Let $\bar{x}$ is exact number, and $\tilde{x}$ - its approximation. The <u>absolute error</u> of $\tilde{x}$ to $\bar{x}$ is defined as $\alpha(\tilde{x})$ and is expressed with the formula:

$$\alpha(\tilde{x}) \geq |\bar{x} - \tilde{x}| . \qquad (1)$$

The simplest way is to calculate the exact value of $\alpha(\tilde{x})$. For example for $\bar{x} = \frac{2}{3}$ and $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$ above we will have correspondingly:

$$\alpha(\tilde{x}_1) = |\bar{x} - \tilde{x}_1| = 0.006666...$$
$$\alpha(\tilde{x}_2) = |\bar{x} - \tilde{x}_2| = 0.000333...$$
$$\alpha(\tilde{x}_3) = |\bar{x} - \tilde{x}_3| = 0.00023466...$$

Apparently the smallest absolute error has x_3. But in all three cases the error is also indefinite decimal fraction which is inconvenient. For that reason as an absolute error we understand not the exact difference but slightly inflated value which is definite number. Let take for example:

$$\alpha(\tilde{x}_2) = 0.0005 \geq |\bar{x} - \tilde{x}_2|$$

or $\alpha(\tilde{x}_2) = 0.001 \geq |\bar{x} - \tilde{x}_2|$

If we solve the first inequality in respect of $\bar{x}$ we'll receive

$$\tilde{x}_2 - 0.0005 \leq \bar{x} \leq \tilde{x}_2 + 0.0005 ,$$

i.e. exactly the number within the interval [0.6662, 0.6672].

As the exact number $\bar{x}$ is unknown and certain proximate $\tilde{x}$ and absolute error $\alpha(\tilde{x})$ are known in fact this means that with tolerance $\pm\alpha(\tilde{x})$ for approximation of $\bar{x}$ may be applied every number of the interval

$$[\tilde{x} - \alpha(\tilde{x}),\ \tilde{x} + \alpha(\tilde{x})].$$

Another criterion for evaluation of the closeness between the numbers is the relative error. It is expressed with $\Delta(\tilde{x})$ and is calculated by the inequality:

$$\Delta(\tilde{x}) \geq \frac{\alpha(\tilde{x})}{|\tilde{x}|} \tag{2}$$

or

$$\Delta(\tilde{x}) \geq \frac{|\bar{x}-\tilde{x}|}{|\tilde{x}|} \ . \tag{3}$$

Usually $\Delta(\tilde{x})$ is expressed in percents.

For example for the relative error of $\tilde{x}_1$ to $\bar{x} = \frac{2}{3}$ we find:

$$\frac{|\tilde{x}_1 - \bar{x}|}{|\tilde{x}_1|} = \frac{0.0066...}{0.66} = 0.001010... \leq 0.002\,, \quad \Delta(\tilde{x}_1) = 0.2\%\,.$$

Stability

An enormous amount of calculations is performed for solving great number of problems. It’s not unlikely errors to pile up and to grow infinitely throughout this process. In this case we say that the calculation process is unstable, i.e. small errors in the initial data lead to big errors in the result. This may be due to instability of the mathematic problem itself, to instability of the numerical method or to instability of calculations. Typical example of instability is the numerical differentiation, summation of series etc.

Correctness

While solving real practical problems often is not paid sufficient attention to the correct formulation and classification. We select and apply some numerical method without thoroughly checking the conditions that it works in. Then it may turn out that as the formulation of the model and the numerical method are inappropriate. That’s why we will introduce the concept correct problem.

Certain problem is called correct or set correctly when the following conditions are met:

1) Exists a solution of the problem;

2) There is only one solution in particular domain;

3) The solution is stable.