МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2005
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2005
Proceedings of the Thirty Fourth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 6–9, 2005

13 MYTHS IN NUMERICAL ANALYSIS*

Mihail M. Konstantinov, Petko H. Petkov, Zvezdalina S. Gancheva

In this paper we consider a number of myths, which are popular among users of modern computational systems and may be very misleading. We also propose a useful heuristic rule.

1. Introduction. The invention of the digital computer in the middle of XX century led to a significant change of the viewpoint on numerical methods and algorithms. Moreover, it became clear that many classical computational schemes are not suitable for implementation in finite arithmetics and, in particular, in floating-point machine arithmetic (MA).

In order to organize a reliable computational procedure, one has to take into account the main three factors determining the accuracy of the computed solution: 1) the properties of the MA and, in particular, the rounding unit, 2) the properties of the computational problem and, in particular, its sensitivity, and 3) the properties of the numerical algorithm and, in particular, its numerical stability.

Unfortunately, the accounting of these factors is not a common practice, especially among the users and developers of mathematical software, who are not numerical analysts. Moreover, many myths are wide spread among this category of users. These myths deserve a special consideration. Analyses of such misconceptions have been presented in [1,2].

2. 13 myths and one heuristic rule. There are sustainable myths[1], or misconceptions, in computational practice which are popular even among experienced users of modern computer software. Below we describe some ot them and consider a number of instructive examples.

Myth 1 *Large errors in the computed solution are due to the total effect of a great number of small round-off errors done at each arithmetic operation performed in MA.*

This is rarely true or is only partially true. Moreover, when performing a large number of arithmetic operations there is a tendency of their mutual compensation. If there is a large error in the computed result then most probably this is due to the performance of a small number of critical operations (or even of a single such operation), when a

*2000 Mathematics Subject Classification: 65-01
[1]This is not a definable concept.

numerical disaster has occured. Such an operation can be the catastrophic cancellation of true digits during a subtraction of close numbers.

For example, consider the computation of e by the formula $e \approx e_n := (1 + 1/n)^n$ for n sufficiently large. The justification of this (obviuously bad) way to compute e is that $e_n \to e$ for $n \to \infty$. In MA with rounding unit $\varepsilon \approx 10^{-16}$ we obtain the good approximation e_8 with $|e - e_8|/e = 10^{-8}$ and the the catastrophe $e_{17} = 1$ with $|e - e_{17}|/e = 0.6321 \approx 1 - 1/e$. The reason is not in the computing of the high power with exponent 10^{16} but in the machine summation $1+10^{-17}$ which produces the result 1.

There is also a very "brave" myth which reveals simple ignorance.

Myth 2 *Rounding errors are not important because they are small* (*a variant: because they are completely compensated*).

The believers in this myth at least have heard about rounding errors. Because some users have not. The latter ones are maybe the happier part of users. But their happiness cannot last long.

The next myth seems almost like a true statement [2].

Myth 3 *A short computation free of cancellation, overflow and underflow, should be accurate.*

Consider the following algorithm, which transforms $x \geq 0$ into $y \in \mathbb{R}$ by the recursive formulae $x_{k+1} = x_k^{1/2}$ for $k = 1, \ldots, n$ and $x_{k+1} = x_k^2$ for $k = n+1, \ldots, 2n$, $y := x_{2n}$. The theoretical result is $y = x$. In practice, even for moderate values of n in the range of 50 to 100, at all available computer platforms and for all computing environments (we write this in November, 2004), the computed result will be $\widehat{y} = 0$ if $0 \leq x < 0$ and $\widehat{y} = 1$ if $x \geq 1$. Thus we can achieve an arbitrarily large relative error in the result.

The next myth deserves a special attention.

Myth 4 *Subtraction of close numbers in MA is dangerous because the relative error of subtraction is large.*

Something is true here: the relative error of subtracion is large (and even unbounded for arbitrary close numbers). But subtraction of close numbers is usually done *exactly* in MA and hence the subtraction itself does not introduce any error. What happens then? It happens that if the close numbers are approximate (which is the typical case in computer computations) then the left-most significant digits are cancelled and the possible inaccuracies in the right-most digits become important. So the useful information is lost even when the subtraction itself is exact.

Of course, if the above close numbers were exact, then the computed result would also be exact. Moreover, in many situations catastrophic cancellation may be harmless. For instance, the machine operation $a + (b - c)$ is OK when $1 \gg |b - c|/|a|$, $a \neq 0$.

So the following statement is also a myth.

Myth 5 *Cancellation in the subtraction of near numbers is always very dangerous.*

Consider now the following myths.

Myth 6 *Increasing the precision of the arithmetics (i.e., decreasing the rounding unit) always increases the accuracy of the computed result.*

Sometimes this is not true, as it is shown below. The sum

$$10^{65.5} + 1005 - 10^{77} + 999 - 10^{65.5} + 10^{77} = 2004$$

will be computed as 0 on most computers in single, double and extended precision.

It is true, however, that *decreasing the rounding unit decreases the known bounds on the error* of the computations, since in many of them the rounding unit is a multiplier.

Myth 7 *Rounding errors are always harmful.*

Not true again. Sometimes rounding errors can (and do!) help in certain computational procedures. For example, the QR algorithm cannot start (theoretically) for certain matrices but due to rounding errors it actually starts.

Myth 8 *The final result cannot be more accurate than the intermediate results, i.e. errors do not cancel.*

A counterexample to this myth is given in [2].

Many myths are connected with the solution of linear and nonlinear equations $f(x) = 0$, where x and $f(x)$ are vectors of the same size and the function f is continuous.

Let $\widehat{x}$ be the solution computed in MA. Then the quantity $\|f(\widehat{x})\|$ is the *residual* corresponding to $\widehat{x}$. Since the residual is scale dependent, it is preferable to work with some scaled quantity, e.g. $r(\widehat{x}) := \|f(\widehat{x})\|/\|f\|$, where $\|f\|$ is the supremum of $\|f(\xi)\|$ when ξ varies over a compact containing the solution x.

The continuity of the function r and the fact that $r(x) = 0$ for the exact solution x, are the basis of many myths, some of them considered below.

Myth 9 *The accuracy of the computed solution $\widehat{x}$ can be checked by the size of the residual $r(\widehat{x})$ – the smaller the residual, the better the approximation.*

There is a close variant to Myth 9.

Myth 10 *Of two approximate solutions the better one corresponds to the smaller residual.*

Myth 9 and 10 are equvalent and they are both untrue. That these myths fail for nonlinear equations is almost obvious as the next example shows.

The scalar equation

$$f(x) := x^3 - 23.001x^2 + 143.022x - 121.021 = 0.$$

has a single real root $x = 1$. Let $\widehat{x}_1 = 0.99$ and $\widehat{x}_2 = 11.00$ be two approximations to the solution. By the way, only the first may be considered as an approximation. Computing the residuals $r(\widehat{x}_k) = |f(\widehat{x}_k)|$ we have $r(\widehat{x}_1) = 1.0022$ and $r(\widehat{x}_2) = 0.1$. Thus the bad approximation $\widehat{x}_2$ with a relative error of 1000 percent has a 10 times less residual than the good approximation $\widehat{x}_1$ with relative error of 1 percent!

But maybe Myths 9 and 10 fail only for nonlinear equations? Unfortunately, not. They are false for linear vector equations as well!

Consider the linear algebraic equation $Ax = b$, where $A = \begin{bmatrix} 0.2161 & 0.1441 \\ 1.2969 & 0.8648 \end{bmatrix}$, $b = [0.1440, 0.8642]\mathrm{T}$. The approximate solution $\widehat{x} = [0.9911, 0.4870]\mathrm{T}$ has small residual $r(\widehat{x}) = \|Ax - b\| = 0.1414 \times 10^{-7}$ and according to Myth 9 should be close to the exact one. But the exact solution is $x = [2, -2]\mathrm{T}$ and there is no true digit in $\widehat{x}$. This should be expected since the relative error here is $\|\widehat{x} - x\|/\|x\| = 0.643$. This example (due to W. Kahan) is remarkable because *all* approximate solutions, whose first three decimal digits coincide with these of x, have larger residuals than $r(\widehat{x})$!

This phenomenon for linear equations is explained in [1] and can be observed even for $n = 2$. Briefly, it is possible in equations with ill-conditioned matrix A, for which the condition number $\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|$ is large. However, this observation is often true for nonlinear equations as well.

Consider the system $Ax = b$ with $A = \begin{bmatrix} \varepsilon^3 & 1 \\ 0 & 1 \end{bmatrix}$ and $b = [1, 1]\mathrm{T}$ which has a solution $x = [0, 1]\mathrm{T}$, where $\varepsilon > 0$ is a small parameter. Let $y = [0, 1 + \varepsilon]\mathrm{T}$ and $z = [1/\varepsilon, 1]\mathrm{T}$ be two approximate solutions. The relative error of y is $e_y = \varepsilon \ll 1$, while this of z is $e_z = 1/\varepsilon \gg 1$. At the same time the residual for y is $\varepsilon\sqrt{2}$, while the residual for z is ε^2. We see that for $\varepsilon \to 0$ the bad solution z has a relative error $1/\varepsilon$ tending to ∞ but its residual ε^2 is arbitrarily smaller than the residual $\varepsilon\sqrt{2}$ of the good solution y with relative error ε. So the check by residuals is completely misleading even for linear systems with two equations. At this point we advise the reader to explain geometrically the observed phenomenon.

In the general case, for a nonsingular A, $Ax = b$ and arbitrary $\widehat{x} \neq x$, we have (in the 2-norm)

$$\frac{1}{\|A^{-1}\|} \leq \frac{\|A\widehat{x} - b\|}{\|\widehat{x} - x\|} \leq \|A\|$$

and these inequalities are reachable. Denote by $\widehat{x}_1$ and $\widehat{x}_2$ the vectors, for which

$$\frac{1}{\|A^{-1}\|} = \frac{\|A\widehat{x}_1 - b\|}{\|\widehat{x}_1 - x\|}, \quad \frac{\|A\widehat{x}_2 - b\|}{\|\widehat{x}_2 - x\|} = \|A\|.$$

Then we have

$$\frac{e_1}{e_2} = \mathrm{cond}(A)\frac{r_1}{r_2},$$

where $r_k := \|A\widehat{x}_k - b\|$. This clearly explains why the bettter approximation can have larger residual.

So there is a bit of irony in Myths 9 and 10: the check of the accuracy of the computed solution by the size of the residual can be successfull if the equation is well-conditioned. But then the computed solution is probably good enough, so there is nothing to check. But if the equation is ill-conditioned and there is a danger of large errors in $\widehat{x}$, then the check based on the residual can be misleading.

The bad thing is that 80 percent of experienced computer users beleive in such myths. And many books on numerical analysis do not worn them.

There are some very sophisticated myths that may even be usefull sometimes. But not always – otherwise they would not be myths.

Myth 11 *A reliable way to check the accuracy of $\widehat{x}$ is to repeat the computations with double, or some other extended precision.*

Myth 11 even has a procedural variant.

Myth 12 *If, after a repeated computation with extended precision, the first several digits in the approximate solutions computed in both ways coincide, then these digits are true.*

Why 11 and 12 are myths is explained in detail in [1] (see also the example after Myth 6). It is true that if we work with a very small rounding unit ε we can achieve an arbitrary number of true digits in the computed result. And it is also true that to achieve this goal we shall need an arbitrary large computing time.

Experienced computer users know that if small changes in the data lead to large changes in the result (i.e. if the computational problem is very sensitive) then the computed solution may be contaminated with large errors. Sometimes this correct observation is reversed assuming that *if, for a given set of small perturbations, the result is slightly changed, then the accuracy of the solution is satisfatory.* Thus we come to the next myth. And the reason is in the words "a given". If there was "any" instead, everything would be OK. But it is impossible to make an experiment including all possible initial data. At least because of the lack of time.

Myth 13 *If the check of the result by a repeated computation with slightly perturbed data gives a slightly perturbed result, this guarantees that the computational procedure is reliable and the solution is computed with a good accuracy.*

A counterexample to this myth is given in [1]. The reason is that in very sensitive problems there are "insensitive directions" (in nonlinear problems they are manifolds) along which large changes of data cause a small change in the result. At the same time a small perturbation of the data along other directions can change the result dramatically.

Consider the system $Ax = b$ with $A = \begin{bmatrix} a+1 & a \\ a & a-1 \end{bmatrix}$, where $a > 0$ is large. The matrix A is nonsingular since $\det(A) = -1$. Let $b = [2a, 2a]\mathrm{T}$. Then $x = [2a, -2a]\mathrm{T}$. So if we change a to $a + \delta$, where δ is arbitrary, then the relative change in both the data b and the result x will be $|\delta|/|a|$. The system looks very well conditioned with a relative condition number of order 1. But if we take $b = [2a + 1, 2a - 1]\mathrm{T}$ the solution becomes $x = [1, 1]\mathrm{T}$. So a relative change of order $1/(2a)$ in the data caused a relative change of order $2a$ in the result – an amplification of order $4a^2$. This indicates an ill-conditioning of the problem. The things become clear after computing the condition number of A, namely $\mathrm{cond}_2(A) = 4a^2 + 2 + O(a^{-2})$, $a \to \infty$.

Now we can ask whether it is possible to save something of Myths 11–13? The answer is yes. Or almost yes. And the next statement is not a myth but a useful euristics.

A heuristics *If a check of the result by several sets of randomly chosen small perturbations in the initial data and by several MA with different rounding units shows small perturbation in the computed result, then with a high degree of reliability we can expect that the computed solution is close to the exact one.*

Of course, we can find a counterexample to this proposition as well – that is why it is heuristics and not a theorem. But the reader will hardly find such an example in his or her computational practice.

REFERENCES

[1] N. Vulchanov, M. Konstantinov. *Modern Mathematical Methods for Computer Calculations. Part I: Foundations of Numerical Computations. Numerical Differentiation and Integration*. Studies in Math. Sci., Bulgarian Inst. Anal. Res., vol. 1, Sofia, 1996, ISBN 954-8949-01-6.
[2] N. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 2002.

Mihail Mihaylov Konstantinov
University of Architecture, Civil
Engineering and Geodesy
1046 Sofia, Bulgaria
e-mail: mmk_fte@uacg.bg

Petko Hristov Petkov
Department of Automatics
Technical University–Sofia
1756 Sofia, Bulgaria
e-mail: php@tu-sofia.bg

Zvezdalina Staikova Gancheva
Municipal Child Center
6 Panayot Volov Str.
9700 Shumen, Bulgaria
e-mail: zvezda_g@yahoo.com

13 МИТА В ЧИСЛЕНИЯ АНАЛИЗ

Михаил М. Константинов, Петко Х. Петков, Звездалина С. Ганчева

Анализирани са 13 популярни мита в числения анализ, които понякога са доста подвеждащи. Предложена е и една полезна евристика.