Numerical methods - Exercises

Lectures:	Doc. Mgr. Jozef Kristek, PhD.	F1-207
Exercises:	Mgr. David Gregor	F1-204

Rounding

Let $\tilde{x}$ is approximation of x written in decimal representation

$$\tilde{x} = \pm\left[d_1 \cdot 10^{e} + d_2 \cdot 10^{e-1} + \cdots + d_k \cdot 10^{e+1-k} + \cdots\right], \quad d_1 \neq 0.$$

We say that k-th decimal digit d_k is significant if

$$\left|x - \tilde{x}\right| \leq 0,5 \cdot 10^{e+1-k} \tag{3}$$

i.e. if $\tilde{x}$ differs from x

at most of 5 units of order of subsequent digit.

If inequality (3) holds for $k \leq p$, but not for $k = p+1$,

we say, that $\tilde{x}$ has p significant digits

and is correctly rounded value of the number x

to the p significant digits.

Rounding

We say that k-th decimal place is significant if

$$\left| x - \tilde{x} \right| \leq 0,5 \cdot 10^{-k} \tag{4}$$

i.e. if $\tilde{x}$ differs from x

at most of 5 units of order of subsequent decimal place.

If inequality (4) hold for $k \leq p$ but not for $k = p + 1$,

we say that $\tilde{x}$ has p significant decimal places.

Rounding

x	374	-27,6473	100,0020
$\tilde{x}$	380	-27,5980	099,9973
	006	+00,0493	000,0047
significant digits	1	3	4
significant decimal places	-	1	2
	099,9973	-0,003728	$1,841.10^{-6}$
	100,0020	-0,004100	$2,500.10^{-6}$
	000,0047	+0,000372	$0,659.10^{-6}$
	5	1	0
	2	3	5
	0,9973	-10,0037	$1,82.10^{-2}$
	1,0084	-10,0042	$2,52.10^{-2}$
	0,0011	+00,0005	$0,70.10^{-2}$
	3	5	0
	2	3	1

Exercise 1.0

Determine the number of significant digits of finite decimal representation of Euler number, if

$$\tilde{x} = 2,718$$

Exercise 1.0

Determine the number of significant digits of finite decimal representation of Euler number, if

$$\tilde{x} = 2,718$$

Solution:

$$x = e = 2,718218\cdots = 10 \cdot 0,2718218...$$

$$\left| x - \tilde{x} \right| \le 10^{-3} \cdot 0,218\cdots \le 0,5 \cdot 10^{1-4}$$

Number $\tilde{x}$ is said to approximate Euler number to 4 significant digits.

Exercise 1.1:

Suppose that $x = 2,78493$ and $y = 2,78469$ are approximations of numbers ξ and η obtained by rounding these numbers to 5 decimal places.
Determine the absolute and relative error of $x-y$ difference.

Definition of errors

Let x is exact value of some number
and $\tilde{x}$ is its approximation

$$\Delta(x) = \tilde{x} - x$$

we call absolute error of approximation

Relative error

$$\frac{\Delta(x)}{x} = \frac{\tilde{x} - x}{x}$$

Definition of errors

Estimation of errors

Each non-negative number £ , for which holds

$$\left|\Delta\left(x\right)\right| \leq £$$

i.e.

$$\tilde{x} - £ \leq x \leq \tilde{x} + £$$

we call estimation of absolute error

Each non-negative number ¢ , for which

$$\left|\frac{\Delta\left(x\right)}{x}\right| \leq ¢$$

we call estimation of relative error

Usually we write

$$x = \tilde{x} \pm £ \qquad \tilde{x} = x\left(1 \pm ¢\right)$$

Error of basic arithmetic operations

Let $f(x, y) = x \pm y.$

Using eqs. (1) and (2) we obtain

absolute and relative error of addition and subtraction

$$\Delta(x \pm y) \le \Delta x + \Delta y \qquad \left|\frac{\Delta(x \pm y)}{x \pm y}\right| \le \left|\frac{x}{x \pm y}\right|\left|\frac{\Delta x}{x}\right| + \left|\frac{y}{x \pm y}\right|\left|\frac{\Delta y}{y}\right|$$

Relative error of addition or subtraction could be significantly larger then relative errors of each operand in case when $|x \pm y|$ is significantly smaller than $|x|$ or $|y|$.

Exercise 1.1:

Suppose that $x = 2,78493$ and $y = 2,78469$ are approximations of numbers ξ and η obtained by rounding these numbers to 5 decimal places. Determine the estimation of absolute and relative error of x-y difference.

Solution:

Estimations of absolute errors of x and y are

$\varepsilon(x) = \varepsilon(y) = 0,5.10^{-5}$.

Then $|\Delta(x-y)| <= \mathbf{10^{-5}} = \varepsilon(x-y)$.

Estimation of relative error of x is

$\delta(x) = [(0,5.10^{-5})/2,78493] = \mathbf{1,8 \cdot 10^{-6}}$

$\delta(y)$ similar to $\delta(x)$

Estimation of relative error

$[(\varepsilon(x-y))/(x-y)] = [(10^{-5})/0,00024] = \mathbf{4,2 \cdot 10^{-2}}$.

Exercise 1.2:

Suppose that $z=1,23456$ is approximation of numbers ζ obtained by rounding this number to 5 decimal places.

Determine the estimation of errors of $[z/(x-y)]$, where x and y are numbers from Exercise1.1

$x = 2,78493$ and $y = 2,78469$

Error of basic arithmetic operations

Let $f(x,y) = xy$.

Then the absolute and relative errors of multiplication are

$$\Delta(xy) \le |y|\Delta x + |x|\Delta y \qquad \left|\frac{\Delta(xy)}{xy}\right| \le \frac{\Delta x}{|x|} + \frac{\Delta y}{|y|}$$

Let $f(x,y) = x/y$.

Then the absolute and relative errors of division are

$$\Delta\left(\frac{x}{y}\right) \le \left|\frac{1}{y}\right|\Delta x + \left|\frac{x}{y^2}\right|\Delta y \qquad \left|\frac{\Delta(x/y)}{x/y}\right| \le \frac{\Delta x}{|x|} + \left|\frac{x}{y}\right|\frac{\Delta y}{|y|}$$

Exercise 1.2:

Suppose that z=1,23456 is approximation of numbers ζ obtained by rounding this number to 5 decimal places.

Determine the estimation of errors of [z/(x-y)], where x and y are numbers from Exercise1.1

Solution:

From Exercise 1.1 we already know the error of denominator.

We also know, that $\varepsilon(z)=0,5.10^{-5}$.

To obtain estimation of error, we just have to do substitution :

$$\Delta\left(\frac{z}{x-y}\right) \leq \left|\frac{1}{x-y}\right| \pounds\left(z\right) + \left|\frac{z}{\left(x-y\right)^2}\right| \pounds\left(x-y\right) =$$

$$= \frac{1}{0,00024}\frac{1}{2}10^{-5} + \frac{1,23456\cdot 10^{-5}}{0,00024^2} \cong 2,2\cdot 10^2$$

Whereas the input values x, y and z have error of order 10^{-5}, the result has error of order 10^{-2}!

One should avoid subtracting two nearly equal numbers!

Propagation of errors

notation 10.2324 represents: 10.2324 ± 0.00005
(all the digits of the number are significant)

Calculate (determine as precisely as possible):

$3.45 + 4.87 - 5.16$

3.55×2.73

$8.24 + 5.33$

$124.53 - 124.52$

4.27×3.13

$9.48 \times 0.513 - 6.72$

Calculate (determine as precisely as possible):

$3.45 + 4.87 - 5.16 = 3.16 \pm 0.015$ (3.145,3.175)

$3.55 \times 2.73 = 9.6915 \pm 0.0314$ (9.6601, 9.7229)

$8.24 + 5.33 = 13.57 \pm 0.01$ (13.56, 13.58)

$124.53 - 124.52 = 0.01 \pm 0.01$ (0, 0.02)

$4.27 \times 3.13 = 13.3651 \pm 0.037$ (13.3281, 13.4021)

$9.48 \times 0.513 - 6.72 = -1.85676 \pm 0.012305$
(-1.869065, -1.844455)

Exercise 1.3

Suppose that the number of digits kept in computer is *p.* *Assuming p=3, add* 1,24 and 0,0221.

Representation of numbers

Real numbers in computers are represented in the floating point format.

Basic idea is similar to the semilogarithmic notation (i.e. $2.457*10^5$)

System of normalized floating point numbers $\mathcal{F}$ is characterized by 4 integer numbers:

Representation of numbers

$$\text{base} \quad \left(\quad \geq 2 \right)$$

$$p \quad \text{precision} \left(p \geq 1 \right)$$

$$\left[e_{\min}, e_{\max} \right] \quad \text{exponent range} \left(e_{\min} < 0 < e_{\max} \right)$$

Each number $x \in \mathcal{F}$ has form of

$$x = \pm m \cdot \quad ^{e}, \quad \text{where } m = d_1 + \frac{d_2}{\quad} + \frac{d_3}{\quad ^2} + \cdots + \frac{d_p}{\quad ^{p-1}}$$

m is normalized mantissa (or significand),

$d_i \in \{0, 1, ..., \quad -1\}, \ i = 1, 2, ..., p$ are digits of mantissa,

p is the number of digits of mantissa and

$e \in \left\langle e_{\min}, e_{\max} \right\rangle$ is integer exponent.

Exercise 1.3

Suppose that the number of digits kept in computer is *p. Assuming p=3, add* 1,24 and 0,0221.

Exercise 1.3

Suppose that the number of digits kept in computer is *p. Assuming p=3, add* 1,24 and 0,0221.

Solution:

At first comparison of exponents with potential denormalization takes place.

$$0{,}124 \cdot 10^{1} + 0{,}221 \cdot 10^{-1} = \left(0{,}124\middle|0 + 0{,}002\middle|21\right) \cdot 10^{1} \doteq 0{,}126 \cdot 10^{1}$$

It should be noted, that due to roundoff errors, the associative and commutative laws of algebra do not necessarily hold for floating-point numbers.

IEEE Standard

IEEE Standard for Floating-Point Arithmeic
The result of arithmetic operation in computer
is exactly the same as
if the operation had been
computed exactly and then rounded

The term underflow is a condition in a computer program where the result of a calculation is a number of smaller absolute value than the computer can actually store in memory.

The term overflow is a condition in a computer program where the result of a calculation is a number of greater absolute value than the computer can actually store in memory.

```
program test_overflow
implicit none
real(4):: x, y
x = 3.1E38
write(*,*) 'x = ', x
y = x + 1.3E38
write(*,*) 'y = ', y  ! +Infinity
end program
```

IEEE Standard

Consequences of floating-point arithmetics:

1. addition of small nonzero might have no effect

   $5.18 \times 10^2 + 4.37 \times 10^{-1} = 5.18 \times 10^2 + 0.00437 \times 10^2 = 5.18437 \times 10^2$ =(rounding)= 5.18×10^2

   machine epsilon: smallest positive machine number ε such that 1 + ε = 1

```
program test_epsilon
implicit none
real(4):: x = 1.0, y

y = 5.96046412227E-008
write(*,'(A4,F16.14)') 'x = ', x
write(*,'(A4,F16.14)') 'y = ', y

write(*,'(A8,F16.14)') 'x + y = ', x+y ! 1.00000000000000

y = 1.19209282445e-007
write(*,'(A4,F16.14)') 'x = ', x
write(*,'(A4,F16.14)') 'y = ', y

write(*,'(A8,F16.14)') 'x + y = ', x+y ! 1.00000011920929

end program
```

IEEE Standard

Consequences of floating-point arithmetics:

2. inverse property of multiplication might not exist

   a x 1/a ≠ 1   :   3.000 x 0.333 = 0.999

rounding eliminate error in representation:

```
program test_inverse
implicit none
real(4):: x = 3, y = 0

y = 1.0/x
write(*,'(A4,F16.14)') 'x = ', x        ! 3.00000000000000
write(*,'(A4,F16.14)') 'y = ', y        ! 0.33333334326744

write(*,'(A8,F16.14)') 'x * y = ', x*y ! 1.00000000000000

end program
```

similar situation might happend in operation of addition:

```
program test_rounding
implicit none
real(4):: x, y
x = 1.7                                  ! 1.70000004768372
write(*,'(A4,F16.14)') 'x = ', x
y = 2.3                                  ! 2.29999995232628
write(*,'(A4,F16.14)') 'y = ', y
y = x + y
write(*,'(A4,F16.14)') 'v = ', y        ! 4.0000000000000

write(*,*)

x = 1.7                                  ! 1.70000004768372
write(*,'(A4,F16.14)') 'x = ', x
y = 0.3                                  ! 0.30000001192093
write(*,'(A4,F16.14)') 'y = ', y
y = x + y
write(*,'(A4,F16.14)') 'v = ', y        ! 2.0000000000000
end program
```

Consequences of floating-point arithmetics:

3. associative law might not hold

$(a + b) + c \neq a + (b + c)$
$a = 6.31 \times 10^{1}$, $b = 4.24 \times 10^{0}$, $c = 2.47 \times 10^{-1}$

$(6.31 \times 10^{1} + 0.424 \times 10^{1}) + 2.47 \times 10^{-1} =$
$6.73 \times 10^{1} + 2.47 \times 10^{-1} = 6.73 \times 10^{1} + 0.0247 \times 10^{1} = 6.75 \times 10^{1}$

$6.31 \times 10^{1} + (4.24 \times 10^{0} + 0.247 \times 10^{0}) =$
$6.31 \times 10^{1} + 4.49 \times 10^{0} = 6.31 \times 10^{1} + 0.449 \times 10^{1} = 6.76 \times 10^{1}$

4. loss of significant digits

Conditionality of numerical problems and numerical stability of algorithms

Exercises:

1. Roots of quadratic equation $x^2 - 2bx + c = 0$

(standard approach can produce error,

while substracting two nearly equal numbers.

It‘s better to use Vieta’s formulas)

```
program test_kvadr

implicit none
real(4):: a = 1.0, b = -400.005, c = 2.0
real(4):: x, y, D

D = b*b - 4*a*c
x = (- b + sqrt(d))/2/a
y = (- b - sqrt(d))/2/a
write(*,*) 'x = ', x, ' y = ', y
! x =    400.0000      y =   5.0024572E-03

y = c / x
write(*,*) 'x = ', x, ' y = ', y
! x =    400.0000      y =   4.9999999E-03
end program
```

2. Computation of integral

(recurrence relation from $n = 0$ to some $\mathrm{n} > 0$ it's not stable,

more accurate is to start from some big $\mathrm{n} > 0$)

$$E_n = \int_0^1 x^n e^{x-1} dx \qquad n = 1, 2, \ldots \qquad E_n = 1 - nE_{n-1} \qquad E_0 = 1 - 1/e$$

```
program test_rekurencia

implicit none
integer :: n
real(4):: E

E = 1.0 - 1.0/exp(1.0)
do n = 1, 12
    E = 1.0 - n*E
    write(*,'(A2,I2,A4,F12.8)') 'E(',n,') = ', E
enddo

write(*,*)

E = 0.0
do n = 20, 13, -1
    E = (1.0 - E)/n
    write(*,'(A2,I2,A4,F12.8)') 'E(',n-1,') = ', E
enddo

end program
```

Exercise 1.4

Suppose that the number of digits kept in computer is *p*. Calculate partial sum $\sum_{n=0}^{\infty} 0,9^n$,assuming *p=3.*

Exercise 1.4

Suppose that the number of digits kept in computer

is *p.* Calculate partial sum $\sum_{n=0}^{\infty} 0,9^n$,assuming *p=3.*

Solution:

Since sum corresponds to geometric series with common ratio q=0,9, we can calculate value of the summation as s = 1/(1-0.9) = 10. Partial sums will be computed using two different approaches:

$$s_k = \left(\left(1+0,9\right)+0,9^2\right)+\cdots+0,9^k - \text{forward summation,}$$

$$r_k = \left(\left(0,9^k+0,9^{k-1}\right)+\cdots+1\right) - \text{backward summation.}$$

Results for different k are written down into table.

k	s_k	r_k	$s - s_k$	$s - r_k$
50	9.98	9.97	0.02	0.03
100	9.98	10.0	0.02	0
150	9.98	10.0	0.02	0

Exercise 1.5

Determine the condition number for value of polynomial

$$p(x) = x^2 + x - 1150$$

in point $x = 33$. Let‘s

$$x \approx \tilde{x} = \frac{100}{3}.$$

Conditionality of numerical problems and numerical stability of algorithms

We say that the correct problem is well-conditioned, if
small change in input data
will cause small change of result.

Condition number is defined as

$$C_p = \frac{|\text{relative error of output}|}{|\text{relative error of input}|}$$

If $C_p \approx 1$, the problem is well-conditioned.

For large C_p (>100) the problem is ill-conditioned.

Exercise 1.5

Determine the condition number for value of polynomial

$$p(x) = x^2 + x - 1150$$

in point $x = 33$. Let‘s $x \approx \tilde{x} = \dfrac{100}{3}.$

Solution:

$$p(33) = -28,\ p\left(\frac{100}{3}\right) = -\frac{50}{9},\ c_p = \frac{\left|\frac{-28-\frac{50}{9}}{-28}\right|}{\left|\frac{33-\frac{50}{9}}{33}\right|} = \frac{\frac{22,4}{28}}{\frac{1/3}{33}} \doteq 79,$$

Problem is ill-conditioned.

Banach fixed-point theorem: Let (X, d) be a non-empty complete metric space with a contraction mapping $g: X \to X$. Then g admits a unique fixed-point x^* in X. Furthermore, x^* can be found as follows: start with an arbitrary element x_0 in X and define a sequence $\{x_n\}$ by $g(x_{n-1}) = x_n$, then $x_n \to x^*$.

What it is good for? Suppose we want to solve $f(x) = 0$.

Let‘s rewrite the $f(x) = 0$ as $\underbrace{\frac{f(x)}{h(x)} + x}_{g(x)} = x$

$h(x_p) \neq 0$

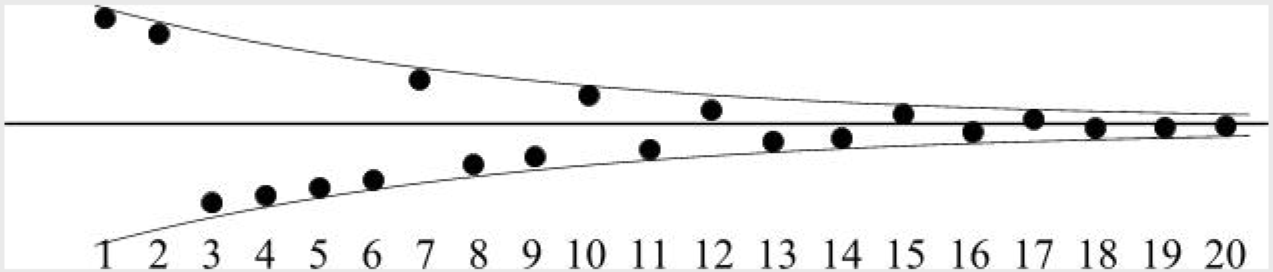
We‘ll get fixed-point problem for $g(x)$, while th solution of $g(x_p) = x_p$ is root of $f(x_p) = 0$.

Metric space

A metric space is an ordered pair (X, d) where X is a set and d is a metric on X, such that for any $x, y, z \in X$, the following holds:

1. $d(x,y) \geq 0$ 2. $d(x,y) = 0 \Leftrightarrow x = y$ 3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

Convergence: If there is some distance e such that no matter how far you go out in the sequence, you can find all subsequent elements which are closer to the limit than e



Cauchy sequence <- term in functional analysis

Contraction mapping: images of two elements are closer then originals

$$\forall x, y \in M \quad d\left(F(x), F(y)\right) \leq \sim d(x, y); \quad \sim \in \left\langle 0,1 \right)$$

Banach fixed-point theorem – states, that exist only one $\P = \lim_{n \to \infty} x_n$,

$x_{k+1} = F(x_k), \; k = 0,1,\ldots$ if $F(x)$ is contraction

$$\sim d(\P, x_{n-1}) \geq d\left(F(\P), F(x_{n-1})\right) = d(\P, x_n)$$

$$\sim [d(\P, x_n) + d(x_n, x_{n-1})] \geq \sim d(\P, x_{n-1}) \geq d(\P, x_n)$$

$$\sim d(x_n, x_{n-1}) \geq (1 - \sim) d(\P, x_n)$$

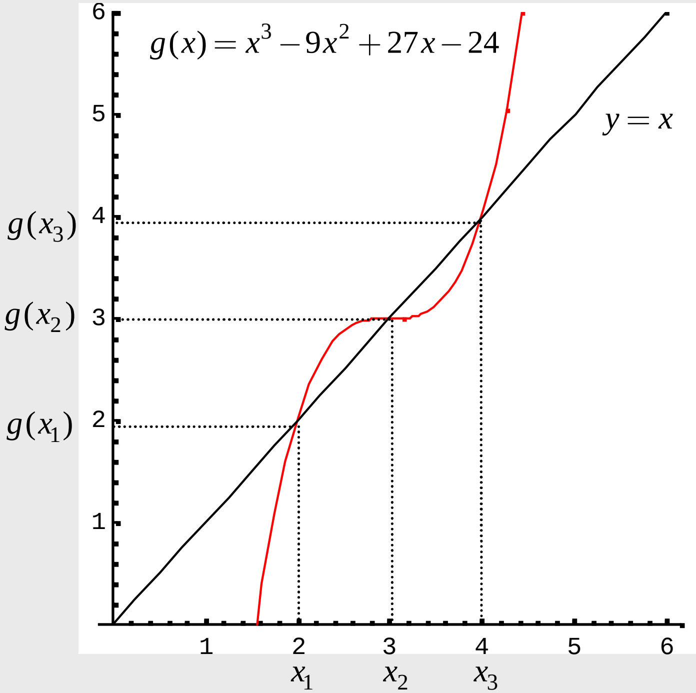
$$\frac{\sim}{1-\sim} d(x_n, x_{n-1}) \geq d(\P, x_n)$$

$$\sim d(x_{n-1}, x_{n-2}) \geq d(x_n, x_{n-1})$$

$$\frac{\sim^n}{1-\sim} d(x_o, x_1) \geq d(\P, x_n)$$

sequence is uniformly approaching limit

Finding roots of nonlinear equations



$$f(x)=0$$

$$g(x)=x$$

$$f(x)=x^3-9x^2+26x-24$$

$$g(x)=f(x)+x$$

$$g(x)=x^3-9x^2+27x-24$$

$$g(x_1)=x_1$$

$$g(x_2)=x_2$$

$$g(x_3)=x_3$$

fixed points

fixed-point problem can be solved(finding roots of previous problem),

constructing contraction mapping

we can construct sequence (cauchy sequence) that converges

to fixed point (= converges to the root of previous problem)

g must be contraction mapping: $d\left(g(x_1), g(x_2)\right) \leq \sim d\left(x_1, x_2\right)$

$$d\left(y_1, y_2\right) \leq \sim d\left(x_1, x_2\right)$$

$$\frac{d\left(y_1, y_2\right)}{d\left(x_1, x_2\right)} \leq \sim < 1$$

so the derivative of $g(x)$ must be from interval $-1 < g'(x) < 1$

this guarantees

that $g(x)$ is a contraction mapping

and therefore convergences to

fixed point