Numerical methods

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

Contents

Introduction

real problem –> mathematical model

mathematical model - solved using computers

Numerical mathematics is a scientific discipline,
which develop and analyze methods
based on manipulations with numbers

Numerical problem

clear and unambiguous description
of functional relation
between the finite number
of input and output data

Algorithm of the numerical problem

clear and unambiguous specification
of finite sequence of operations.
The operations uniquely assign
n-tuple numbers of results
to the *m*-tuple numbers
from certain set of input data

Pre- a post-processing

Contents

Source and types of errors

Human errors

Errors of mathematical model

difference between solution
of mathematical (idealized) problem
and
solution of real problem

Example: Estimation of the surface of the Earth
using equation $S = 4 \pi r^2$

Errors of input data
due to inaccuracy of measurements

Errors of numerical method

comes from taking a numerical problem
instead of mathematical problem

The estimation of this error
is necessary part of solution of numerical problem

Errors of numerical method

comes from taking a numerical problem
instead of mathematical problem

The estimation of this error
is necessary part of solution of numerical problem

Example: Computation of the value of sin x for x=1 using the summation of the finite number of terms in Taylor expansion

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots$$

Errors of numerical method

comes from taking a numerical problem
instead of mathematical problem

The estimation of this error
is necessary part of solution of numerical problem

Example: Computation of the value of sin x for x=1 using the summation of the finite number of terms in Taylor expansion

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots$$

We know that summation of the first n terms of expansion gives the error in estimation at most

$$1/(2n+1)!$$

Round-off errors

Round-off errors can cumulate but also cancel

Example: The number π, result of 2/3

Round-off errors

Round-off errors can cumulate but also cancel

Example: The number π, result of 2/3

real problem – all kind of errors together

Contents

Definition of errors

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

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

we call **absolute error of approximation**

Definition of errors

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

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

we call **absolute error of approximation**

Relative error

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

Estimation of errors

Each non-negative number ε , for which

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

i.e.

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

we call **estimation of absolute error**

Estimation of errors

Each non-negative number ε, for which

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

i.e.

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

we call **estimation of absolute error**

Each non-negative number δ, for which

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

we call **estimation of relative error**

Estimation of errors

Each non-negative number ε, for which

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

i.e.

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

we call **estimation of absolute error**

Each non-negative number δ, for which

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

we call **estimation of relative error**

Usually we write

$$x = \tilde{x} \pm \varepsilon \qquad \tilde{x} = x\left(1 \pm \delta\right)$$

Definition of errors

Now evaluate an error

of value $f\left(x_1, x_2, \ldots, x_n\right)$ of function f,

if exact values x_i

will be changed by approximate values $\tilde{x}_i = x_i + \Delta x_i$.

Definition of errors

Now evaluate an error

of value $f\left(x_1, x_2, \ldots, x_n\right)$ of function f,

if exact values x_i

will be changed by approximate values $\tilde{x}_i = x_i + \Delta x_i$.

$$f\left(\tilde{x}\right) = f\left(\mathrm{x}\right) + \sum_{i=1}^{n} \Delta x_i \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{n} \Delta x_i \Delta x_j \frac{\partial^2 f\left(\mathrm{x}\right)}{\partial x_i \partial x_j} + \cdots .$$

Definition of errors

Now evaluate an error

of value $f\left(x_1, x_2, \ldots, x_n\right)$ of function f,

if exact values x_i

will be changed by approximate values $\tilde{x}_i = x_i + \Delta x_i$.

$$f\left(\tilde{x}\right) = f\left(\mathrm{x}\right) + \sum_{i=1}^{n} \Delta x_i \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{n} \Delta x_i \Delta x_j \frac{\partial^2 f\left(\mathrm{x}\right)}{\partial x_i \partial x_j} + \cdots .$$

Assuming that $\Delta x_i \Delta x_j$ are small

Definition of errors

Now evaluate an error
of value $f\left(x_1, x_2, \ldots, x_n\right)$ of function f,
if exact values x_i
will be changed by approximate values $\tilde{x}_i = x_i + \Delta x_i$.

$$f\left(\tilde{x}\right) \approx f\left(\mathrm{x}\right) + \sum_{i=1}^{n} \Delta x_i \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i}$$

Assuming that $\Delta x_i \Delta x_j$ are small

Definition of errors

Now evaluate an error
of value $f\left(x_1, x_2, \ldots, x_n\right)$ of function f,
if exact values x_i
will be changed by approximate values $\tilde{x}_i = x_i + \Delta x_i$.

$$f\left(\tilde{x}\right) \approx f\left(\mathrm{x}\right) + \sum_{i=1}^{n} \Delta x_i \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i}$$

Assuming that $\Delta x_i \Delta x_j$ are small, we can define absolute error as

$$\left|\Delta f\left(\mathrm{x}\right)\right| := \left|f\left(\tilde{\mathrm{x}}\right) - f\left(\mathrm{x}\right)\right| \doteq \left|\sum_{i=1}^{n} \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i} \Delta x_i\right|$$

Definition of errors

Now evaluate an error
of value $f\left(x_1, x_2, \ldots, x_n\right)$ of function f,
if exact values x_i
will be changed by approximate values $\tilde{x}_i = x_i + \Delta x_i$.

$$f\left(\tilde{x}\right) \approx f\left(\mathrm{x}\right) + \sum_{i=1}^{n} \Delta x_i \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i}$$

Assuming that $\Delta x_i \Delta x_j$ are small, we can define absolute error as

$$\left|\Delta f\left(\mathrm{x}\right)\right| := \left|f\left(\tilde{\mathrm{x}}\right) - f\left(\mathrm{x}\right)\right| \doteq \left|\sum_{i=1}^{n} \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i} \Delta x_i\right| \lesssim \sum_{i=1}^{n} \left|\frac{\partial f\left(\mathrm{x}\right)}{\partial x_i}\right| \cdot \left|\Delta x_i\right| \qquad (1)$$

And for relative error

$$\left|\frac{\Delta f\left(\mathrm{x}\right)}{f\left(\mathrm{x}\right)}\right| \doteq \left|\sum_{i=1}^{n} \frac{x_i}{f\left(\mathrm{x}\right)} \frac{\partial f\left(\mathrm{x}\right)}{\partial x_i} \frac{\Delta x_i}{x_i}\right|$$

And for relative error

$$\left|\frac{\Delta f\left(\mathrm{x}\right)}{f\left(\mathrm{x}\right)}\right| \doteq \left|\sum_{i=1}^{n}\frac{x_i}{f\left(\mathrm{x}\right)}\frac{\partial f\left(\mathrm{x}\right)}{\partial x_i}\frac{\Delta x_i}{x_i}\right| \lesssim \sum_{i=1}^{n}\left|\frac{x_i}{f\left(\mathrm{x}\right)}\frac{\partial f\left(\mathrm{x}\right)}{\partial x_i}\right|.\left|\frac{\Delta x_i}{x_i}\right|. \qquad (2)$$

In practice the function values and values of derivatives are evaluated at $\tilde{\mathrm{x}}$.

Error of basic arithmetic operations

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

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**

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

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**

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

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 $|y|$.

Error of basic arithmetic operations

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

Then absolute and relative error of **multiplication**

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

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

Then absolute and relative error of **division**

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

Contents

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

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

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.

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.

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.

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

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380		
-27.6473	-27.598		
100.002	99.9973		
99.9973	100.002		
-0.003728	-0.0041		
$1.841*10^{-6}$	$2.5*10^{-6}$		

Rounding

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380	1	-
-27.6473	-27.598		
100.002	99.9973		
99.9973	100.002		
-0.003728	-0.0041		
$1.841*10^{-6}$	$2.5*10^{-6}$		

Rounding

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380	1	-
-27.6473	-27.598	3	1
100.002	99.9973		
99.9973	100.002		
-0.003728	-0.0041		
$1.841*10^{-6}$	$2.5*10^{-6}$		

Rounding

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380	1	-
-27.6473	-27.598	3	1
100.002	99.9973	4	2
99.9973	100.002		
-0.003728	-0.0041		
$1.841*10^{-6}$	$2.5*10^{-6}$		

Rounding

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380	1	-
-27.6473	-27.598	3	1
100.002	99.9973	4	2
99.9973	100.002	5	2
-0.003728	-0.0041		
$1.841*10^{-6}$	$2.5*10^{-6}$		

Rounding

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380	1	-
-27.6473	-27.598	3	1
100.002	99.9973	4	2
99.9973	100.002	5	2
-0.003728	-0.0041	1	3
$1.841*10^{-6}$	$2.5*10^{-6}$		

Rounding

Several examples

x	$\tilde{x}$	# of significant digits	# of significant decimal places
374	380	1	-
-27.6473	-27.598	3	1
100.002	99.9973	4	2
99.9973	100.002	5	2
-0.003728	-0.0041	1	3
$1.841*10^{-6}$	$2.5*10^{-6}$	0	5

Subtracting of two close numbers leads to decrement of significant digits

Subtracting of two close numbers leads to decrement of significant digits

Example:

$$x = 4.998949 \cdot 10^{1}, \quad \tilde{x} = 4.999 \cdot 10^{1}, \quad |\Delta x| = 5.10 \cdot 10^{-4}, \quad \left|\frac{\Delta x}{x}\right| \doteq 1.020 \cdot 10^{-5},$$

$$y = 5.001848 \cdot 10^{1}, \quad \tilde{y} = 5.002 \cdot 10^{1}, \quad |\Delta y| = 1.52 \cdot 10^{-4}, \quad \left|\frac{\Delta y}{y}\right| \doteq 3.039 \cdot 10^{-5}$$

Subtracting of two close numbers leads to decrement of significant digits

Example:

$$x = 4.998949 \cdot 10^{1}, \quad \tilde{x} = 4.999 \cdot 10^{1}, \quad |\Delta x| = 5.10 \cdot 10^{-4}, \quad \left|\frac{\Delta x}{x}\right| \doteq 1.020 \cdot 10^{-5},$$

$$y = 5.001848 \cdot 10^{1}, \quad \tilde{y} = 5.002 \cdot 10^{1}, \quad |\Delta y| = 1.52 \cdot 10^{-4}, \quad \left|\frac{\Delta y}{y}\right| \doteq 3.039 \cdot 10^{-5}$$

then for subtractions $z = y - x, \quad \tilde{z} = \tilde{y} - \tilde{x}$ we get

$$z = 2.899 \cdot 10^{-2}, \quad \tilde{z} = 3 \cdot 10^{-2}, \quad |\Delta z| = 1.01 \cdot 10^{-3}, \quad \left|\frac{\Delta z}{z}\right| \doteq 3.484 \cdot 10^{-2}$$

Subtracting of two close numbers leads to decrement of significant digits

Example:

$$x = 4.998949 \cdot 10^{1}, \quad \tilde{x} = 4.999 \cdot 10^{1}, \quad |\Delta x| = 5.10 \cdot 10^{-4}, \quad \left|\frac{\Delta x}{x}\right| \doteq 1.020 \cdot 10^{-5},$$

$$y = 5.001848 \cdot 10^{1}, \quad \tilde{y} = 5.002 \cdot 10^{1}, \quad |\Delta y| = 1.52 \cdot 10^{-4}, \quad \left|\frac{\Delta y}{y}\right| \doteq 3.039 \cdot 10^{-5}$$

then for subtractions $z = y - x, \quad \tilde{z} = \tilde{y} - \tilde{x}$ we get

$$z = 2.899 \cdot 10^{-2}, \quad \tilde{z} = 3 \cdot 10^{-2}, \quad |\Delta z| = 1.01 \cdot 10^{-3}, \quad \left|\frac{\Delta z}{z}\right| \doteq 3.484 \cdot 10^{-2}$$

therefore $\tilde{z}$ has one significant digit
while $\tilde{x}$ and $\tilde{y}$ have four significant digits.

Example:

$$x = 1.3262 \pm 5 \cdot 10^{-5}, \quad y = -6.5347 \pm 5 \cdot 10^{-5}, \quad z = 13.235 \pm 5 \cdot 10^{-4}$$

Find the approximation of

$$f = xy / z,$$

absolute and relative errors
and the number of significant digits of result.

Contents

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

$$\begin{aligned} \beta &\quad \text{base} && \left(\beta \geq 2\right) \\ p &\quad \text{precision} && \left(p \geq 1\right) \\ \left[e_{\min}, e_{\max}\right] &\quad \text{exponent range} && \left(e_{\min} < 0 < e_{\max}\right) \end{aligned}$$

Representation of numbers

$$\begin{aligned} \beta &\quad \text{base} && (\beta \geq 2) \\ p &\quad \text{precision} && (p \geq 1) \\ [e_{\min}, e_{\max}] &\quad \text{exponent range} && (e_{\min} < 0 < e_{\max}) \end{aligned}$$

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

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

$$\begin{aligned} \beta &\quad \text{base} && (\beta \geq 2) \\ p &\quad \text{precision} && (p \geq 1) \\ [e_{\min}, e_{\max}] &\quad \text{exponent range} && (e_{\min} < 0 < e_{\max}) \end{aligned}$$

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

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

m is **normalized mantissa (**or **significand)**,

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

p is the number of digits of mantissa and

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

β base $(\beta \geq 2)$

p precision $(p \geq 1)$

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

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

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

m is **normalized mantissa (**or **significand)**,

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

p is the number of digits of mantissa and

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

Normalized mantissa means that for $x \neq 0$ is $d_1 \geq 1$.

Representation of numbers

$\beta = 2$	binary numeral system
$\beta = 16$	hexadecimal system
$\beta = 8$	octal system
$\beta = 10$	decimal system

Representation of numbers

$\beta = 2$	binary numeral system
$\beta = 16$	hexadecimal system
$\beta = 8$	octal system
$\beta = 10$	decimal system

Set $\mathcal{F}$ of floating point numbers is a finite set,

count of number is

Representation of numbers

$\beta = 2$ binary numeral system

$\beta = 16$ hexadecimal system

$\beta = 8$ octal system

$\beta = 10$ decimal system

Set $\mathcal{F}$ of floating point numbers is a finite set,

count of number is

2

two signs,

Representation of numbers

$\beta = 2$ binary numeral system

$\beta = 16$ hexadecimal system

$\beta = 8$ octal system

$\beta = 10$ decimal system

Set $\mathcal{F}$ of floating point numbers is a finite set,

count of number is

$$2(\beta - 1)$$

two signs,

$\beta - 1$ options for the first digit of mantissa,

Representation of numbers

$\beta = 2$ binary numeral system

$\beta = 16$ hexadecimal system

$\beta = 8$ octal system

$\beta = 10$ decimal system

Set $\mathcal{F}$ of floating point numbers is a finite set,

count of number is

$$2(\beta - 1)\beta^{p-1}$$

two signs,

$\beta - 1$ options for the first digit of mantissa,

β options for other $p - 1$ digits of mantissa,

$\beta = 2$	binary numeral system
$\beta = 16$	hexadecimal system
$\beta = 8$	octal system
$\beta = 10$	decimal system

Set $\mathcal{F}$ of floating point numbers is a finite set,

count of number is

$$2\left(\beta - 1\right)\beta^{p-1}\left(e_{\max} - e_{\min} + 1\right)$$

two signs,
$\beta - 1$ options for the first digit of mantissa,
β options for other $p - 1$ digits of mantissa,
$e_{\max} - e_{\min} + 1$ possible values of exponent

$\beta = 2$ binary numeral system

$\beta = 16$ hexadecimal system

$\beta = 8$ octal system

$\beta = 10$ decimal system

Set $\mathcal{F}$ of floating point numbers is a finite set,

count of number is

$$2\left(\beta - 1\right)\beta^{p-1}\left(e_{\max} - e_{\min} + 1\right) + 1$$

two signs,

$\beta - 1$ options for the first digit of mantissa,

β options for other $p - 1$ digits of mantissa,

$e_{\max} - e_{\min} + 1$ possible values of exponent

and one zero

The smallest positive number in $\mathcal{F}$ is *UFL* (UnderFlow Level)

$$UFL = \beta^{e_{\min}},$$

which has the first digit of mantissa equal to one, others zero and exponent the smallest one.

The largest positive number in $\mathcal{F}$ is *OFL* (OverFlow Level)

$$OFL = \left(\beta - \beta^{1-p}\right)\beta^{U},$$

which has all digits of mantissa equal to $\beta - 1$ and exponent is the largest one.

The gaps between adjacent numbers
scale
with the size of the numbers

The gaps between adjacent numbers
scale
with the size of the numbers

Relative resolution is given by **machine epsilon**

$$\varepsilon_{\text{machine}} = 0.5\,\beta^{1\text{-}p}$$

For all x, there exists a floating point x'
such that $|x'-x| \leq \varepsilon_{\text{machine}}|x|$

The gaps between adjacent numbers
scale
with the size of the numbers

Relative resolution is given by **machine epsilon**

$$\varepsilon_{\text{machine}} = 0.5\,\beta^{1\text{-}p}$$

For all x, there exists a floating point x'
such that $|x'-x| \leq \varepsilon_{\text{machine}}|x|$

Example: $\beta = 2, \quad p = 3, \quad e_{\min} = -1, \quad e_{\max} = 2$

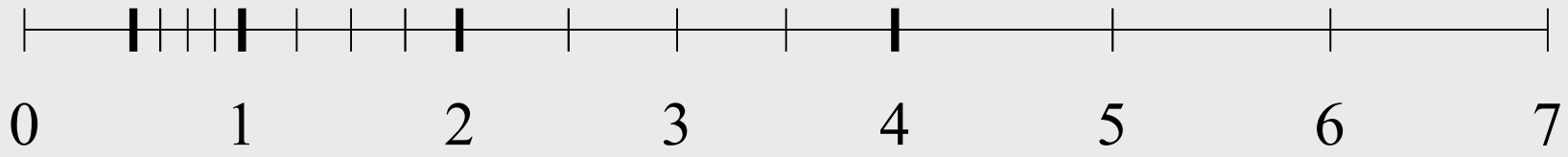
The gaps between adjacent numbers
scale
with the size of the numbers

Relative resolution is given by **machine epsilon**

$$\varepsilon_{\text{machine}} = 0.5\,\beta^{1\text{-}p}$$

For all x, there exists a floating point x'
such that $|x'-x| \leq \varepsilon_{\text{machine}}|x|$

Example: $\beta = 2, \quad p = 3, \quad e_{\min} = -1, \quad e_{\max} = 2$



Representation of numbers

- $\pm\infty$ is returned when an operation overflows
- $x/\pm\infty = 0$ for any number x,
- $x/0 = \pm\infty$ for any nonzero number x
- Operations with infinity are defined as limits, e.g.

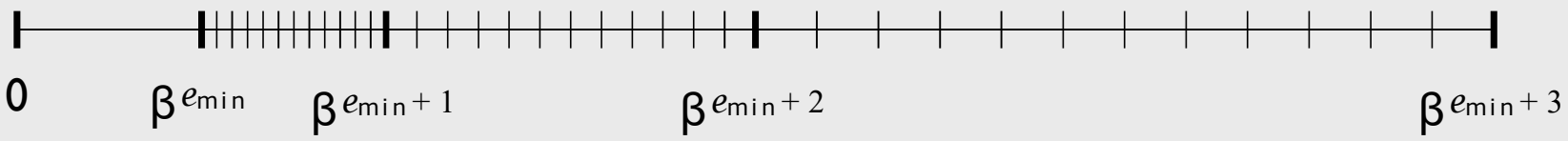
$$4 - \infty = \lim_{x \to \infty} 4 - x = -\infty$$

- NaN (Not a Number) is returned when the an operation has no well-defined finite or infinite result

  Examples: $\infty - \infty,\ \infty / \infty,\ 0/0,$ NaN $\odot\ x$

Denormalized Numbers

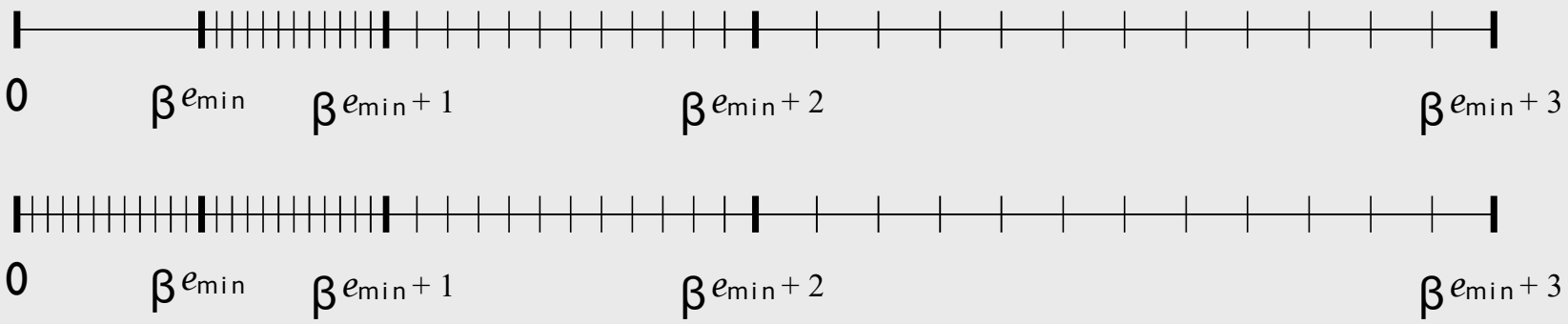
- With normalized significand there is a “gap” between 0 and $\beta^{e_{min}}$



Denormalized Numbers

- With normalized significand there is a “gap” between 0 and $\beta^{e_{min}}$
- Solution: Allow non-normalized significand when the exponent is e_{min}
- This *gradual underflow* garantees that

$$x = y \Leftrightarrow x - y = 0$$



IEEE Single Precision

- 1 sign bit, 8 exponent bits, 23 significand bits:

0	00000000	00000000000000000000000
S	E	M

- Represented number:

$$(-1)^S \times 1.M \times 2^{E-127}$$

- Special cases:

	$E = 0$	$0 < E < 255$	$E = 255$
$M = 0$	± 0	Powers of 2	$\pm\infty$
$M \neq 0$	Denormalized	Ordinary numbers	NaN

IEEE Single Precision, Examples

S	E	M	Quantity
0	11111111	00000100000000000000000	NaN
1	11111111	00100010001001010101010	NaN
0	11111111	00000000000000000000000	∞
0	10000001	10100000000000000000000	$+1 \cdot 2^{129-127} \cdot 1.101 = 6.5$
0	10000000	00000000000000000000000	$+1 \cdot 2^{128-127} \cdot 1.0 = 2$
0	00000001	00000000000000000000000	$+1 \cdot 2^{1-127} \cdot 1.0 = 2^{-126}$
0	00000000	10000000000000000000000	$+1 \cdot 2^{-126} \cdot 0.1 = 2^{-127}$
0	00000000	00000000000000000000001	$+1 \cdot 2^{-126} \cdot 2^{-23} = 2^{-149}$
0	00000000	00000000000000000000000	0
1	00000000	00000000000000000000000	-0
1	10000001	10100000000000000000000	$-1 \cdot 2^{129-127} \cdot 1.101 = -6.5$
1	11111111	00000000000000000000000	$-\infty$

IEEE Floating Point Data Types

	Single precision	Double precision
Significand size (p)	24 bits	53 bits
Exponent size	8 bits	11
Total size	32 bits	64 bits
$e_{\max}$	+127	+1023
$e_{\min}$	-126	-1022
Smallest normalized	$2^{-126} \approx 10^{-38}$	$2^{-1022} \approx 10^{-308}$
Largest normalized	$2^{127} \approx 10^{38}$	$2^{1023} \approx 10^{308}$
$\epsilon_{\text{machine}}$	$2^{-24} \approx 6 \cdots 10^{-8}$	$2^{-53} \approx 10^{-16}$

Floating Point Arithmetic

- Define $\mathrm{fl}(x)$ as the closest floating point approximation to x
- By the definition of $\epsilon_{\text{machine}}$, we have for the relative error:

  For all $x \in \mathbb{R}$, there exists ϵ with $|\epsilon| \leq \epsilon_{\text{machine}}$
  such that $\mathrm{fl}(x) = x(1+\epsilon)$

- The result of an operation $\circledast$ using floating point numbers is $\mathrm{fl}(a \circledast b)$
- If $\mathrm{fl}(a \circledast b)$ is the nearest floating point number to $a \circledast b$, the arithmetic *rounds correctly* (IEEE does), which leads to the following property:

  For all floating point x,y, there exists ϵ with $|\epsilon| \leq \epsilon_{\text{machine}}$ such that $x \circledast y = (x * y)(1+\epsilon)$

Contents

Conditionality of numerical problems and numerical stability of algorithms

We have to investigate the effects
of small changes in input data and rounding
on the result.

Conditionality of numerical problems and numerical stability of algorithms

We have to investigate the effects
of small changes in input data and rounding
on the result.

Mathematical problem can be treated as mapping $y = f(x)$,
which each data from x the set D of input data
assigns the result y from the set R of output data.

Conditionality of numerical problems and numerical stability of algorithms

We have to investigate the effects
of small changes in input data and rounding
on the result.

Mathematical problem can be treated as mapping $y = f(x)$,
which each data from x the set D of input data
assigns the result y from the set R of output data.

We say that the mathematical problem

$$y = f(x), \quad x \in D, \; y \in R,$$

is **correct,** when

1. For each input $x \in D$ exists only one result $y \in R$,
2. this result continuously depends on input data,
   i.e. if $x \to a$, then $f(x) \to f(a)$.

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

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

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

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

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

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

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

We say that the **algorithm is well-conditioned**,
if it is less sensitive to errors in input data.

If the effect of round-off errors on result is small then
we say about **numerically stable algorithm.**

Well-conditioned and numerically stable algorithm
is called **stable**.

Example:

Estimate the condition number of the problem:

evaluate the functional value of (differentiable) function

$$y = f\left(x\right)$$

show on example function

$$f\left(x\right) = \tan x$$

Examples:

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

2. *Evaluation of integral* $E_n = \int_0^1 x^n e^{x-1} dx \qquad n = 1, 2, \ldots$

$$E_n = \int_0^1 x^n e^{x-1}\,\mathrm{d}x \qquad n = 1, 2, \ldots$$

$$\int_0^1 x^n e^{x-1}\,\mathrm{d}x = [x^n e^{x-1}]_0^1 - \int_0^1 n x^{n-1} e^{x-1}\,\mathrm{d}x\,,$$

$$E_n = 1 - nE_{n-1}\,.$$

$$E_1 = 1/e\,, \qquad E_n = 1 - nE_{n-1}\,, \qquad n = 2, 3, \ldots$$

$$E_{n-1} = \frac{1 - E_n}{n},$$

$$E_{n-1} = \frac{1 - E_n}{n},$$

$$E_n = \int_0^1 x^n e^{x-1}\,\mathrm{d}x \leq \int_0^1 x^n\,\mathrm{d}x = \frac{1}{n+1}$$

$$E_n \to 0 \quad \text{for} \quad n \to \infty$$

$$E_{n-1} = \frac{1 - E_n}{n}\,,$$

$$E_n = \int_0^1 x^n e^{x-1}\,\mathrm{d}x \leq \int_0^1 x^n\,\mathrm{d}x = \frac{1}{n+1}$$

$$E_n \to 0 \quad \text{for} \quad n \to \infty$$

$$E_N = 0\,, \qquad E_{n-1} = \frac{1 - E_n}{n}\,, \qquad n = N, N-1, \ldots\,,$$