Course of Numerical Methods for Engineering
Lab 2

Luca Bonaventura, Tommaso Benacchio

MOX - Politecnico di Milano

PHYS-ENG, A.Y. 2020-21

21-22/09/2020

Topics of this session:

- **Floating point approximation of real numbers and error definitions**
- **Fixed point method**

Floating point approximation of real numbers

Let's run these two code snippets in MATLAB:

```
x = 0;
while (x  = 1)
x = x + 1/16
end
```

```
x = 0;
while (x  = 1)
x = x + 0.1
end
```

What is going on?

Computers use floating-point arithmetic, working with a finite set of numbers and finite precision. Given a real number x, its floating-point approximation is:

$$fl(x) = (-1)^s \cdot (0.a_1 a_2 \dots a_m) \cdot \beta^e$$

dove

- s **is the sign;**
- β **is the basis of the system;**
- e **is the exponent,** $e \in [e_{min}, e_{max}],\ e_{min} < 0,\ e_{max} > 0$**;**
- $0.a_1 a_2 \dots a_m$**, with** $0 \leq a_i \leq \beta - 1$**, is the mantissa.**

In MATLAB/Octave, floating point numbers are represented in double-precision format:

- **binary representation:** $\beta = 2$**;**
- **every number occupies a registry of 8-byte=64-bit, with**
  - **1 bit for the sign:** 0 **(**$+$**) or** 1 **(**$-$**);**
  - $m = 52$ **bit for the mantissa** $\rightarrow$ **limitation on computational precision ;**
  - **11 bits for the exponent** $e_{min} = -1022, e_{max} = 1023$ $\rightarrow$ **limitation on arithmetic extension;**

- **The minimum and maximum numbers (in absolute value) representable in double precision can be obtained in MATLAB via the commands:**

$$\texttt{realmin} = 2^{e_{min}} = 2.2251 \cdot 10^{-308}$$
$$\texttt{realmax} = (2 - 2^{-m}) \cdot 2^{e_{max}} = 1.7977 \cdot 10^{308}.$$

- **Any operation producing a value larger than `realmax` generates an overflow. The result takes the value `Inf`.**
- **Any operation producing a value smaller than `realmin` generates an underflow. Except for special cases, this is treated as 0.**
- **For any operation producing an indefinite result (e.g., `0/0`, `Inf/Inf`, `Inf-Inf`), the result takes the value `NaN` ("not a number").**

Floating-point arithmetic

The machine epsilon is the smallest number `eps` **such that** $1 + \texttt{eps} > 1$**, in a floating point system with basis** β **and** m **significant digits:**

$$\texttt{eps} = \frac{\beta^{1-m}}{2}$$

Note that `eps` **is very different from** `realmin`**.**

It makes no sense trying to achieve a relative error smaller than `eps` **in floating point computations.**

The value of machine epsilon can be obtained in MATLAB with the command `eps`**.**

Errors

The absolute error in the approximation of α with α_{approx} is defined as

$$\mathcal{E}_{abs} = |\alpha - \alpha_{approx}|.$$

The relative error in the approximation of α with α_{approx} is defined as

$$\mathcal{E}_{rel} = \frac{|\alpha - \alpha_{approx}|}{|\alpha|}.$$

Exercise 1

It is known that

$$1 = \lim_{h \to 0} \frac{\exp(h) - 1}{h}.$$

1. **Write a MATLAB script to compute** $\frac{\exp(h) - 1}{h}$ **for** $h = 10^{-k}, k = 0, \ldots, 20.$
2. **Compute for each case the absolute error of the result. Plot the error as a function of** k **on a semi-logarithmic plot using the MATLAB command** `semilogy`**.**
3. **Explain the error behaviour based on the theory of the floating point representation of real numbers.**

Exercise 2

Write a MATLAB script to compute $(1 + 10^{-k}) - 1$ **for** $k = 0, \dots, 20$. **Compute in each case the relative error of the result, using the formula**

$$\frac{|[(1 + 10^{-k}) - 1] - 10^{-k}|}{10^{-k}}.$$

Plot the error as a function of k. **Repeat the error computation using the formula**

$$\frac{|[1 + 10^{-k} - 10^{-k} - 1|}{10^{-k}}.$$

Repeat the computations for $(10^m + 10^{-k}) - 10^m$ **for the same values of** k **and** $m = 1, \dots, 10$. **Explain the error behaviour based on the theory of the floating point representation of real numbers.**

Fixed point method

Let $\phi \in \mathcal{C}^1[a, b]$ such that $\phi : [a, b] \to [a, b]$ and let $x^{(0)} \in [a, b]$. The **fixed point method** **for the solution of $x = \phi(x)$ is defined by the recursive sequence**

$$x^{(k+1)} = \phi(x^{(k)}) \quad k \geq 1$$

ϕ is also called iteration function.

Exercise 3

For an arbitrary real number $a > 0$, consider the fixed point problem

$$x = x(2 - ax),$$

whose solution is $1/a$.

1. **Determine the conditions on the initial guess x_0 such that the fixed point method is convergent.**
2. **Implement the method in two MATLAB scripts, one using a `for` cycle, one using a `while` cycle, interrupting the computation when the relative error is smaller that some quantity ϵ.**
3. **Run the script for different values of a, ϵ and different initial guesses.**

Exercise 4

For an arbitrary real number $a > 0$, consider the fixed point problem

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right),$$

whose solution is $\sqrt{a}$.

1. **Determine the conditions on the initial guess x_0 such that the fixed point method is convergent.**
2. **Implement the method in a MATLAB script using a `for` or a `while` cycle, interrupting the computation when the relative error is smaller that some quantity ϵ.**
3. **Run the script for different values of a, ϵ and different initial guesses.**

Exercise 5

Given a real number $a > 0$ determine the conditions on the initial guess x_0 such that the sequence

$$x_{n+1} = \frac{x_n}{2}(3 - a x_n^2) \quad n \geq 0$$

is convergent to $1/\sqrt{a}$. Write a MATLAB function that uses this fixed point algorithm to compute the square root of a at a given level of accuracy.