Numerical Analysis: Practice Midterm (30 marks)

Urbain Vaes

March 31, 2022

Question 1 (8 marks)**.** True or false?

1. Let $(\bullet)_2$ denote binary representation. It holds that $(0.1111)_2 + (0.0001)_2 = 1$.

2. It holds that $(1000)_2 \times (0.001)_2 = 1$.

3. It holds that
$$(0.\overline{1})_3 = \frac{1}{2}.$$

4. In base 16, all the natural numbers from 1 to 200 can be represented using 2 digits.

5. In Julia, `Float64(.1) == Float32(.1)` evaluates to `true`.

6. The spacing (in absolute value) between successive double-precision (`Float64`) floating point numbers is constant.

7. It holds that $(0.\overline{10101})_2 = (1.2345)_{10}$.

8. Machine addition $\widehat{+}$ is an associative operation. More precisely, given any three double-precision floating point numbers x, y and z, the following equality holds:
$$(x \mathbin{\widehat{+}} y) \mathbin{\widehat{+}} z = x \mathbin{\widehat{+}} (y \mathbin{\widehat{+}} z).$$

9. The machine epsilon is the smallest strictly positive number that can be represented in a floating point format.

10. Let ε denote the machine epsilon for the double-precision format. Let also $\widehat{+}$ and $\widehat{/}$ denote respectively the machine addition and the machine division operators for the double-precision format. It holds that $1 \mathbin{\widehat{+}} (\varepsilon \mathbin{\widehat{/}} 64) = 1$ and that $\varepsilon \mathbin{\widehat{/}} 64 \neq 0$.

11. Assume that $x \in \mathbf{R}$ belongs to the double-precision floating point format (that is, assume that $x \in \mathbf{F}_{64}$). Then $-x \in \mathbf{F}_{64}$.

A correct (resp. incorrect) answer leads to +1 mark (resp. -1 mark).

Question 2 (8 marks)**.** Assume that $\mathsf{A} \in \mathbf{R}^{n\times n}$ is an invertible matrix and that $\boldsymbol{b} \in \mathbf{R}^n$ and $\boldsymbol{\beta} \in \mathbf{R}^n$ are two nonzero vectors in $\mathbf{R}^n$. We denote by $\boldsymbol{x}$ and $\boldsymbol{\xi}$ the solutions to the linear equations $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ and $\mathsf{A}\boldsymbol{\xi} = \boldsymbol{\beta}$, respectively. Show that

$$\frac{\|\boldsymbol{x} - \boldsymbol{\xi}\|}{\|\boldsymbol{x}\|} \leq \|\mathsf{A}\|\|\mathsf{A}^{-1}\|\frac{\|\boldsymbol{b} - \boldsymbol{\beta}\|}{\|\boldsymbol{b}\|}.$$

Here $\|\bullet\|$ denotes both the Euclidean vector norm and the induced matrix norm.
Bonus question (1 mark): Let $\kappa := \|\mathsf{A}\|\|\mathsf{A}^{-1}\|$. Prove that $\kappa \geq 1$.

Question 3 (8 marks)**.** Let $\mathsf{A} \in \mathbf{R}^{n\times n}$ be a symmetric positive definite matrix and let $\boldsymbol{b} \in \mathbf{R}^n$. The steepest descent algorithm for solving $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ is given hereafter:

Pick $\varepsilon > 0$ and initial $\boldsymbol{x}$
$\boldsymbol{r} \leftarrow \mathsf{A}\boldsymbol{x} - \boldsymbol{b}$
while $\|\boldsymbol{r}\| \geq \varepsilon\|\boldsymbol{b}\|$ **do**
  $\omega \leftarrow \boldsymbol{r}^T\boldsymbol{r}/\boldsymbol{r}^T\mathsf{A}\boldsymbol{r}$
  $\boldsymbol{x} \leftarrow \boldsymbol{x} - \omega\boldsymbol{r}$
  $\boldsymbol{r} \leftarrow \mathsf{A}\boldsymbol{x} - \boldsymbol{b}$
end while

- Why is this method called the *steepest descent* algorithm? (1 mark)
- How many floating point operations does an iteration of this algorithm require? (5 marks)
- Are the following statements true of false? (2 marks)
  1. There exists a unique solution $\boldsymbol{x}_*$ to the linear system $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$.
  2. The iterates converge to $\boldsymbol{x}_*$ in at most n iterations.
  3. We consider the following modification of the algorithm:

     Pick $\varepsilon > 0$, $\omega > 0$ and initial $\boldsymbol{x}$
     $\boldsymbol{r} \leftarrow \mathsf{A}\boldsymbol{x} - \boldsymbol{b}$
     while $\|\boldsymbol{r}\| \geq \varepsilon\|\boldsymbol{b}\|$ **do**
       $\boldsymbol{x} \leftarrow \boldsymbol{x} - \omega\boldsymbol{r}$
       $\boldsymbol{r} \leftarrow \mathsf{A}\boldsymbol{x} - \boldsymbol{b}$
     end while

     If ω is sufficiently small, then this algorithm converges.
  4. Here we no longer assume that A is positive definite. Instead, we consider that

     $$\mathsf{A} = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}.$$

     In this case, the steepest descent algorithm is convergent for any initial $\boldsymbol{x}$.

Question 4 (6 marks)**.** We proved in class the quadratic convergence of the Newton–Raphson method for a smooth function with a simple root. The aim of this exercise is to study the convergence of the method in the case of a function with a double root. To this end, we consider the simple one-dimensional equation

$$f(x) := (x-1)^2 = 0. \tag{1}$$

1. Write down one iteration of the Newton–Raphson method for (1) in the form:

$$x_{k+1} = F(x_k).$$

2. Let $e_k = x_k - x_*$, where x_* is the exact solution to (1). Find a recurrence relation for the error and, assuming that the initial guess is $x_0 = 2$, write down an explicit expression for e_k.

3. What is the order of convergence of the method in this case?

4. **Bonus question** (1 mark): Repeat the previous exercises for the equation $(x-1)^3 = 0$. What is the order of convergence in this case, and what is the rate of convergence?