Numerical Analysis: Practice Midterm (30 marks)

Urbain Vaes

October 22, 2022

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

1. Let $(\bullet)_2$ denote binary representation. It holds that $(0.1011)_2 + (0.0101)_2 = 1$.
2. It holds that $(1000)_3 \times (0.002)_3 = 2$.
3. A natural number with binary representation $(b_4 b_3 b_2 b_1 b_0)_2$ is even if and only if $b_0 = 0$.
4. In Julia, `Float64(.4) == Float32(.4)` evaluates to `true`.
5. Let $(\bullet)_3$ denote base 3 representation. It holds that $(0.\overline{2200})_3 = (0.9)_{10}$.
6. Machine addition $\widehat{+}$ is a commutative operation. More precisely, given any two double-precision floating point numbers $x \in \mathbf{F}_{64}$ and $y \in \mathbf{F}_{64}$, it holds that $x \mathbin{\widehat{+}} y = y \mathbin{\widehat{+}} x$.
7. Let $\mathbf{F}_{32}$ and $\mathbf{F}_{64}$ denote respectively the sets of single and double precision floating point numbers. It holds that $\mathbf{F}_{32} \subset \mathbf{F}_{64}$.
8. The machine epsilon of a floating point format is the smallest strictly positive number that *(i)* is a power of 2 and *(ii)* can be represented exactly in the format.
9. Let $\mathbf{F}_{64}$ denote the set of double precision floating point numbers. For any $x \in \mathbf{R}$ such that $x \in \mathbf{F}_{64}$, it holds that $x + 1 \in \mathbf{F}_{64}$.
10. Let $f \colon \mathbf{R} \to \mathbf{R}$ denote the function that maps $x \in \mathbf{R}$ to the number of double precision floating point numbers contained in the interval $[x-1, x+1]$. Then f is a decreasing function of x.
11. Let $n \in \mathbf{N}$. The number of bits in the binary representation of n is less than or equal to 4 times the number of digits in the decimal representation of n.

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

Question 2 (Interpolation and approximation, 10 marks)**.** Throughout this exercise, we assume that $x_0 < \ldots < x_n$ are distinct values and that $u\colon \mathbf{R} \to \mathbf{R}$ is a smooth function.

1. (**3 marks**) Are the following statements true or false?

   - There exists a unique polynomial p of degree less than or equal n such that

   $$\forall i \in \{0, \ldots, n\}, \qquad p(x_i) = u(x_i). \tag{1}$$

   - Assume that $p \in \mathbf{P}(n)$ is such that (1) is satisfied. Then there is a constant $K \in \mathbf{R}$ independent of x such that

   $$\forall x \in \mathbf{R}, \qquad u(x) - p(x) = K(x - x_0) \ldots (x - x_n).$$

   - Assume that $p \in \mathbf{P}(n)$ is such that (1) is satisfied. Then p is necessarily of degree n.

2. For $i \in \{0, \ldots, n\}$, let $u_i = u(x_i)$, and let $m \leqslant n$ be a given natural number. We wish to fit the data $(x_0, u_0), \ldots, (x_n, u_n)$ with a function $\widehat{u}\colon \mathbf{R} \to \mathbf{R}$ of the form

   $$\widehat{u}(x) = \alpha_0 + \alpha_1 x + \ldots + \alpha_m x^m.$$

   Specifically, we wish to find the coefficients $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_m)^T$ such that the error

   $$J(\boldsymbol{\alpha}) := \frac{1}{2} \sum_{i=0}^{n} |u_i - \widehat{u}(x_i)|^2$$

   is minimized. Throughout this exercise, we use the notations

   $$\mathsf{A} \begin{pmatrix} 1 & x_0 & \ldots & x_0^m \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^m \end{pmatrix}, \qquad \boldsymbol{b} := \begin{pmatrix} u_0 \\ \vdots \\ u_n \end{pmatrix}$$

   - (**3 marks**) Show that $J(\boldsymbol{\alpha})$ may be rewritten as

   $$J(\boldsymbol{\alpha}) = \frac{1}{2} (\mathsf{A}\boldsymbol{\alpha} - \boldsymbol{b})^T (\mathsf{A}\boldsymbol{\alpha} - \boldsymbol{b}).$$

   - (**4 marks**) Prove that if $\boldsymbol{\alpha}_* \in \mathbf{R}^{m+1}$ is a minimizer of J, then

   $$\mathsf{A}^T \mathsf{A} \boldsymbol{\alpha}_* = \mathsf{A}^T \boldsymbol{b}.$$

   - (**1 mark**) Show that the matrix $\mathsf{A}^T\mathsf{A}$ is positive definite. You can take for granted that the columns of A are linearly independent.

Question 3 (Numerical integration).

Question 4 (Vector and matrix norms, 6 marks)**.** The 1-norm and the ∞-norm of a vector $\boldsymbol{x} \in \mathbf{R}^n$ are defined as follows:

$$\|\boldsymbol{x}\|_1 = |x_1| + \cdots + |x_n| \qquad \text{and} \qquad \|\boldsymbol{x}\|_\infty = \max\Big\{|x_1, \ldots, |x_n|\Big\}.$$

These norms both induce a matrix norm through the formula

$$\|\mathsf{A}\|_p := \sup\Big\{\|\mathsf{A}\boldsymbol{x}\|_p : \|\boldsymbol{x}\|_p = 1\Big\}.$$

Prove that, for $\mathsf{A} \in \mathbf{R}^{n\times n}$,

- (**6 marks**) $\|\mathsf{A}\|_1$ is given by the maximum absolute column sum:

$$\|\mathsf{A}\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|. \tag{2}$$

- (**1 mark**) $\|\mathsf{A}\|_\infty$ is given by the maximum absolute row sum:

$$\|\mathsf{A}\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|.$$

Hint: In order to prove (2), you may find it useful to proceed as follows:

- Introduce j_* as the index of the column with maximum absolute sum:

$$j_* = \operatorname*{arg\,max}_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|.$$

- Prove the direction $\geqslant$ in (2) by finding a vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_1 = 1$ such that

$$\|\mathsf{A}\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |a_{ij_*}|.$$

- Prove the direction $\leqslant$ in (2) by showing that, for a general $\boldsymbol{x} \in \mathbf{R}^n$ with $\|\boldsymbol{x}\|_1 = 1$,

$$\|\mathsf{A}\boldsymbol{x}\| \leqslant \sum_{i=1}^{n} |a_{ij_*}|.$$