# Numerical Analysis: Midterm

(**30 marks**, only the 3 best questions count)

Urbain Vaes

October 24, 2022

**Question 1** (Floating point arithmetic, 10 marks)**.** True or false? (+1/0/-1)

**1.** Let $(\bullet)_2$ denote binary representation. It holds that $(0.1011)_2 + (0.0101)_2 = 1$.

**2.** Let $(\bullet)_3$ denote base 3 representation. 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.** 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 $x \mathbin{\widehat{+}} y = y \mathbin{\widehat{+}} x$.

**6.** Let $\mathbf{F}_{32}$ and $\mathbf{F}_{64}$ denote respectively the sets of single and double precision floating point numbers. It holds that $\mathbf{F}_{32} \subset \mathbf{F}_{64}$.

**7.** The machine epsilon of a floating point format is the smallest strictly positive number that can be represented exactly in the format.

**8.** 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 $x + 1 \in \mathbf{F}_{64}$.

**9.** Let $a_i \in \{0, 1\}$ for $i \in \{1, 2, 3\}$. If $(a_1 a_2 a_3)_2$ is a multiple of 3, then $(a_1 a_2 a_3)_4$ is a multiple of 6. Here $(\bullet)_4$ denotes base 4 representation.

**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 $n$.

**12.** It holds that $(0.\overline{2200})_3 = (0.9)_{10}$.

**13.** Let $p \in \mathbf{N}$. The set $\bigl\{(b_0.b_1 b_2 \dots b_{p-1})_2 \colon b_i \in \{0, 1\}\bigr\}$ contains $2^p$ distinct real numbers.

**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. The notation $\mathbf{P}(n)$ denotes the set of polynomials of degree less than or equal to $n$.

**1.** (**4 marks**) Are the following statements true or false? (+1/0/-1)

- There exists a unique polynomial $p \in \mathbf{P}(n)$ 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 of degree exactly $n$.
- If $x_0, \ldots, x_n$ are the roots of the Chebyshev polynomial of degree $n$, then
$$\sup_{x \in \mathbf{R}} \Big| (x - x_0) \ldots (x - x_n) \Big| \leqslant \frac{\pi}{2^n}.$$
- The function $S\colon \mathbf{N} \to \mathbf{R}$ given by
$$S(n) = \sum_{i=1}^{n} \left( i + i^2 + i^3 + i^4 \right)$$
is a polynomial of degree 5. (More precisely, there exists a polynomial of degree 5, say $q$, such that $S(n) = q(n)$ for all $n \in \mathbf{N}$.)

**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 coefficients $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_m)^T$ such that the error
$$J(\boldsymbol{\alpha}) := \frac{1}{2} \sum_{i=0}^{n} \left| u_i - \widehat{u}(x_i) \right|^2$$

is minimized. Throughout this exercise, we use the notations

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

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

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

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

$$\mathsf{A}^T\mathsf{A}\boldsymbol{\alpha}_* = \mathsf{A}^T\boldsymbol{b}. \tag{2}$$

- (**1 mark**) Find a solution to (2) in terms of $u_0, \dots, u_n$ and $n$ when $m = 0$. Explain.

**Question 3** (Numerical integration, 10 marks)**.** The Gauss–Legendre quadrature formula with $n$ nodes is an approximate integration formula of the form

$$I(u) := \int_{-1}^{1} u(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i\, u(x_i) =: \widehat{I}_n(u), \tag{3}$$

which is exact when $u$ is a polynomial of degree less than or equal to $2n-1$. (Note that the nodes are here numbered starting from 1.)

**1.** (**5 marks**) Find the nodes and weights of the Gauss–Legendre rule with $n = 3$ nodes.

**2.** (**2 marks**) Let $\{L_0, L_1, \dots\}$ denote orthogonal polynomials for the inner product

$$\langle f, g \rangle := \int_{-1}^{1} f(x) g(x)\,\mathrm{d}x$$

which, in addition, satisfy the following two conditions:

- For all $i \in \mathbf{N}$, the polynomial $L_i$ is of degree $i$.
- The leading coefficient of $L_i$, which multiplies $x^i$, is equal to 1.

Calculate $L_0$, $L_1$, $L_2$ and $L_3$. What is the connection between $L_3$ and the rule found in the first item?

**3.** Assume that $x_1, \dots, x_n$ and $w_1, \dots, w_n$ are such that (3) is satisfied for all $u \in \mathbf{P}(2n-1)$.

- (**2 marks**) Show that the weights are given by

$$\forall i \in \{1, \dots, n\}, \qquad w_i = \int_{-1}^{1} \ell_i(x)\,\mathrm{d}x,$$

where $\ell_i$ is the Lagrange polynomial

$$\ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

- (**1 marks**) Show that the weights are all positive: $w_i > 0$ for all $i$.

**4.** (**Bonus +2**) Prove the following error estimate: if $u$ is a smooth function, then

$$\left|I(u) - \widehat{I}_n(u)\right| \leqslant \frac{C_{2n}}{(2n)!} \int_{-1}^{1} \big(L_n(x)\big)^2 \, \mathrm{d}x, \qquad C_{2n} := \sup_{\xi \in [-1,1]} \left|u^{(2n)}(\xi)\right|.$$

**Hint**: You may find it useful to proceed as follows:

- First show that
$$I(u) - \widehat{I}_n(u) = \int_{-1}^{1} u(x) - p(x) \, \mathrm{d}x, \tag{4}$$
for *any* polynomial $p \in \mathbf{P}(2n-1)$ such that
$$\forall i \in \{1, \dots, n\}, \qquad p(x_i) = u(x_i).$$

- Notice that equation (4) is true in particular when $p$ is the Hermite interpolation of $u$ at the nodes $x_1, \dots, x_n$. Finally, conclude by using the formula for the interpolation error proved in class: if $p$ is the Hermite interpolant of $u$ at the nodes $x_1, \dots, x_n$, then
$$\forall x \in \mathbf{R}, \qquad u(x) - p(x) = \frac{u^{(2n)}\big(\xi(x)\big)}{(2n)!} (x - x_1)^2 \dots (x - x_n)^2.$$

**Question 4** (Vector and matrix norms, 10 marks)**.** The 1-norm and the $\infty$-norm of a vector $\boldsymbol{x} \in \mathbf{R}^n$ are defined as follows:

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

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

1. (**10 marks**) $\|\mathsf{A}\|_1$ is given by the maximum absolute column sum:

$$\|\mathsf{A}\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|. \tag{5}$$

2. (**Bonus +2**) $\|\mathsf{A}\|_\infty$ 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 (5), 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 (5) 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 (5) by showing that, for any $\boldsymbol{x} \in \mathbf{R}^n$ with $\|\boldsymbol{x}\|_1 = 1$,

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