Numerical Analysis: Midterm

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

Urbain Vaes

October 20, 2024

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

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. In Julia, `eps(Float16)` returns the smallest strictly positive number that can be represented exactly in the `Float16` 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 $x \in \mathbf{R}$ and $y \in \mathbf{R}$ be two numbers that are exactly representable in the `Float64` format. Then $x \mathbin{\widehat{+}} y = x + y$: machine addition is exact in this case.

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

Question 2 (Interpolation and approximation, **10 marks**). Throughout this exercise, we use the notation $x_i^n = i/n$ and assume 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. We proved in class that, for all $n \geqslant 0$, there exists a unique polynomial $p_n \in \mathbf{P}(n)$ such that

$$\forall i \in \{0, \dots, n\}, \qquad p_n(x_i^n) = u(x_i^n). \tag{1}$$

Are the following assertions true or false?

1. If u is not the zero function, then the degree of p_n is exactly n.

2. If $u(x) = 2x + 1$, then $p_n = u$ for all $n \in \{1, 2, 3, \dots\}$.

3. Fix $u(x) = 1 + \sin(57\pi x)$. Then $p_3(x) = 1$.

4. Fix $u(x) = (2x - 1)^3$. Then $p_2(x) = 2x - 1$.

5. For all u that is smooth, it holds that

$$\max_{x \in [0,1]} \bigl| u(x) - p_n(x) \bigr| \xrightarrow[n \to \infty]{} 0.$$

6. Fix $u(x) = \cos(2x)$. Then

$$\max_{x \in [0,1]} \bigl| u(x) - p_n(x) \bigr| \xrightarrow[n \to \infty]{} 0.$$

7. Fix $u(x) = \sin(x)$. Then

$$\max_{x \in \mathbf{R}} \bigl| u(x) - p_n(x) \bigr| \xrightarrow[n \to \infty]{} 0.$$

8. Suppose that $p(x) \in \mathbf{P}(n)$ and let $q(x) = p(x+1) - p(x)$. Then $q \in \mathbf{P}(n-1)$.

9. Let $(f_0, f_1, f_2, \dots) = (1, 1, 2, \dots)$ denote the Fibonacci sequence. There exists a polynomial p such that

$$\forall n \in \mathbf{N}, \qquad f_n = p(n).$$

10. For any matrix $\mathsf{A} \in \mathbf{R}^{20 \times 10}$, the linear system

$$\mathsf{A}^T \mathsf{A} \boldsymbol{\alpha} = \mathsf{A}^T \boldsymbol{\alpha}$$

admits a unique solution.

1. There exists a unique polynomial $p \in \mathbf{P}(n+1)$ such that

$$\forall i \in \{0, \dots, n\}, \qquad p(x_i) = u(x_i). \tag{2}$$

2. Assume that $p \in \mathbf{P}(n)$ is such that (2) is satisfied. Then there is a constant $K \in \mathbf{R}$ independent of x such that

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

3. Assume that $p \in \mathbf{P}(n)$ is such that (2) is satisfied. Then p is of degree exactly n.

4. If $x_0, \dots, x_n$ are the roots of the Chebyshev polynomial of degree n, then

$$\sup_{x \in \mathbf{R}} \left| (x - x_0) \dots (x - x_n) \right| \leqslant \frac{\pi}{2^n}.$$

5. 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 $n \in \mathbf{N}$.)

6. Assume that $p \in \mathbf{P}(n)$ is such that (2) is satisfied. It holds that

$$\sup_{x \in \mathbf{R}} \left| u(x) - p(x) \right| \leqslant \pi^2 / n.$$

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

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

Specifically, we wish to find coefficients $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_m)^T$ 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

$$\mathsf{A}^T\mathsf{A}\boldsymbol{\alpha}_* = \mathsf{A}^T\boldsymbol{b}. \tag{3}$$

- (**1 mark**) Find a solution to (3) in terms of $u_0, \ldots, u_n$ and n when $m = 0$. Explain.

Solution.

- Notice that

$$\mathsf{A}\boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 + \alpha_1 x_0 + \cdots + \alpha_m x_0^m \\ \vdots \\ \alpha_0 + \alpha_1 x_n + \cdots + \alpha_m x_n^m \end{pmatrix} = \begin{pmatrix} \widehat{u}(x_0) \\ \vdots \\ \widehat{u}(x_n) \end{pmatrix}.$$

Therefore

$$\frac{1}{2}\sum_{i=1}^{n} \left|\widehat{u}(x_i) - u_i\right|^2 = \frac{1}{2}\sum_{i=1}^{n} \left|(\mathsf{A}\boldsymbol{\alpha} - \boldsymbol{b})_i\right|^2 = \frac{1}{2}(\mathsf{A}\boldsymbol{\alpha} - \boldsymbol{b})^T(\mathsf{A}\boldsymbol{\alpha} - \boldsymbol{b})$$

- A necessary condition is that $\nabla J(\boldsymbol{\alpha}_*) = 0$. We calculate that

$$\frac{\partial}{\partial_{x_i}}\left(\boldsymbol{b}^T\boldsymbol{x}\right) = \frac{\partial}{\partial_{x_i}}\left(\sum_{j=1}^{n} b_j x_j\right) = \sum_{j=1}^{n} b_j \delta_{ij} = b_i.$$

Similarly, for any matrix $\mathsf{M} \in \mathbf{R}^{n\times n}$, it holds that

$$\frac{\partial}{\partial_{x_i}}\left(\boldsymbol{x}^T\mathsf{M}\boldsymbol{x}\right) = \frac{\partial}{\partial_{x_i}}\left(\sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk} x_j x_k\right) = \sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk} \frac{\partial}{\partial_{x_i}}(x_j x_k).$$

Applying the formula for the derivative of a product, we obtain

$$\begin{aligned} \frac{\partial}{\partial_{x_i}}\left(\boldsymbol{x}^T\mathsf{M}\boldsymbol{x}\right) &= \sum_{j=1}^{n}\sum_{k=1}^{n} m_{jk}\delta_{ij}x_k + m_{jk}x_j\delta_{ik} \\ &= \sum_{k=1}^{n} m_{ik}x_k + \sum_{j=1}^{n} m_{ji}x_j = (\mathsf{M}\boldsymbol{x} + \mathsf{M}^T\boldsymbol{x})_i. \end{aligned}$$

Employing these formulae, we calculate that (representing the gradient with a column vector)

$$\nabla_{\boldsymbol{\alpha}} \left(\boldsymbol{b}^T \boldsymbol{\alpha} \right) = \boldsymbol{b}, \qquad \nabla_{\boldsymbol{\alpha}} \left(\boldsymbol{\alpha}^T \mathsf{A}^T \mathsf{A} \boldsymbol{\alpha} \right) = 2\mathsf{A}^T \mathsf{A} \boldsymbol{\alpha}.$$

It is then simple to conclude.

- In this case $\mathsf{A}^T\mathsf{A} = n+1$ and α_* is a scalar. The solution is given by

$$\alpha_* = \frac{u_0 + \cdots + u_n}{n+1},$$

which is the average of the values $u_0, \dots, u_{n+1}$.

△

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

$$I(u) := \int_{-1}^{1} u(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i\, u(x_i) =: \widehat{I}_n(u), \tag{4}$$

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.

Solution. A necessary and sufficient condition in order for (4) to be satisfied for any polynomial $p \in \mathbf{P}(5)$ is that

$$\int_{-1}^{1} x^d\,\mathrm{d}x = \sum_{i=1}^{n} w_i x_i^d, \qquad \text{for all } d \in \{0,1,2,3,4,5\}.$$

This leads to the following system of equations

$$\begin{cases} 2 = w_1 + w_2 + w_3, \\ 0 = w_1 x_1 + w_2 x_2 + w_3 x_3, \\ \dfrac{2}{3} = w_1 x_1^2 + w_2 x_2^2 + w_3 x_3^2, \\ 0 = w_1 x_1^3 + w_2 x_2^3 + w_3 x_3^3, \\ \dfrac{2}{5} = w_1 x_1^4 + w_2 x_2^4 + w_3 x_3^4, \\ 0 = w_1 x_1^5 + w_2 x_2^5 + w_3 x_3^5. \end{cases}$$

Given the symmetry of the problem, it is reasonable to look for a solution of the form

$$(x_1, x_2, x_3, w_1, w_2, w_3) = (-x, 0, x, w_1, w_2, w_1),$$

where only 3 unknown parameters remain. For such a set of parameters, the second, fourth and sixth equations are satisfied, and the other three equations give

$$\begin{cases} 2 = 2w_1 + w_2, \\ \dfrac{2}{3} = 2w_1 x^2, \\ \dfrac{2}{5} = 2w_1 x^4. \end{cases}$$

Dividing the third equation by the second, we obtain $x^2 = 3/5$ and so $x = \pm\sqrt{\frac{3}{5}}$ (both values lead to the same integration rule in the end). It is then simple to deduce that $w_1 = \frac{5}{9}$ and $w_2 = \frac{8}{9}$. We have thus derived the formula

$$\int_{-1}^{1} u(x) \approx \frac{5}{9} u\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} u(0) + \frac{5}{9} u\left(\sqrt{\frac{3}{5}}\right).$$

△

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?

Solution. Clearly $L_0 = 1$. Then $L_1 = x + a_1$ and the requirement that $\langle L_1, L_0 \rangle = 0$ implies that $a_1 = 0$. We then use the ansatz $L_2 = x^2 + b_2 x + a_2$ for L_2. The requirement that $\langle L_2, L_1 \rangle$ leads to $b_2 = 0$, and then

$$\langle L_2, L_0 \rangle = \frac{2}{3} + 2a_2,$$

and so $L_2(x) = x^2 - \frac{1}{3}$. Finally, for L_3, we use the ansatz $L_3 = x^3 + c_3 x^2 + b_3 x + a_3$. We calculate

$$\begin{aligned}
\langle L_3, 1 \rangle &= \frac{2}{3} c_3 + 2a_3, \\
\langle L_3, x \rangle &= \frac{2}{5} + \frac{2}{3} b_3, \\
\langle L_3, x^2 \rangle &= \frac{2}{5} c_3 + \frac{2}{3} a_3.
\end{aligned}$$

The second equation gives $b_3 = -\frac{3}{5}$, and the other two equations lead to $c_3 = a_3 = 0$. We conclude that $L_3(x) = x^3 - \frac{3}{5} x$. The roots of L_3 are given by $\left\{-\sqrt{\frac{3}{5}}, 0, \sqrt{\frac{3}{5}}\right\}$, and they coincide with the nodes of the Gauss–Legendre quadrature with 3 nodes. △

3. Assume that $x_1, \dots, x_n$ and $w_1, \dots, w_n$ are such that (4) 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 ℓ_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.

Solution. Since (4) holds true for all $u \in \mathbf{P}(2n-1)$, it holds true in particular for the function $u = \ell_i \in \mathbf{P}(2n-1)$, which implies that

$$\int_{-1}^{1} \ell_i(x) \, \mathrm{d}x = \sum_{i=1}^{n} w_j \ell_i(x_j) = w_i.$$

Similarly, since (4) holds true also for $u \in \ell_i^2 \in \mathbf{P}(2n-1)$, we deduce that

$$\int_{-1}^{1} \bigl(\ell_i(x)\bigr)^2 \, \mathrm{d}x = \sum_{i=1}^{n} w_j \bigl(\ell_i(x_j)\bigr)^2 = w_i.$$

Since the left-hand side is positive, we deduce that $w_i > 0$. △

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

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

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

- First show that
$$I(u) - \widehat{I}_n(u) = \int_{-1}^{1} u(x) - p(x) \, \mathrm{d}x, \tag{5}$$
for *any* polynomial $p \in \mathbf{P}(2n-1)$ such that
$$\forall i \in \{1, \dots, n\}, \qquad p(x_i) = u(x_i). \tag{6}$$

- Notice that equation (5) 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
$$\forall x \in \mathbf{R}, \qquad u(x) - p(x) = \frac{u^{(2n)}\bigl(\xi(x)\bigr)}{(2n)!}(x - x_1)^2 \dots (x - x_n)^2.$$

Solution. Assume that $p \in \mathbf{P}(2n-1)$ is such that (6) is satisfied. Then by (4) we deduce that

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

Consequently, we obtain that

$$I(u) - \widehat{I}_n(u) = \int_{-1}^{1} u(x)\,\mathrm{d}x - \int_{-1}^{1} p(x)\,\mathrm{d}x = \int_{-1}^{1} u(x) - p(x)\,\mathrm{d}x.$$

This equation holds true in particular with p being the Hermite interpolation of u at the nodes $x_1, \dots, x_n$. Then, using the formula for the interpolation error, we obtain

$$u(x) - u(x) = \frac{u^{(2n)}\big(\xi(x)\big)}{(2n)!}(x-x_1)^2 \dots (x-x_n)^2 = \frac{u^{(2n)}\big(\xi(x)\big)}{(2n)!}\big(L_n(x)\big)^2.$$

Indeed, as shown in class, L_n is a polynomial of degree n with single roots at $x_1, \dots, x_n$. Now we conclude by noting that

$$\big|I(u) - \widehat{I}_n(u)\big| = \left|\int_{-1}^{1} u(x) - p(x)\,\mathrm{d}x\right| \leqslant \int_{-1}^{1} |u(x) - p(x)|\,\mathrm{d}x \leqslant \int_{-1}^{1} \frac{C_{2n}}{(2n)!}\big(L_n(x)\big)^2\,\mathrm{d}x,$$

which concludes the exercise. △