Numerical Analysis: Final exam

(**50 marks**, only the 5 best questions count)

Urbain Vaes

12 May 2022

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

1. Let $(\bullet)_3$ denote base 3 representation. It holds that
$$(120)_3 + (111)_3 = (1001)_3.$$

2. Let $(\bullet)_2$ denote binary representation. It holds that
$$(1000)_2 \times (0.1\overline{01})_2 = (101.\overline{01})_2.$$

3. In Julia, `Float64(.25) == Float32(.25)` evaluates to `true`.

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

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

6. Let $\mathbf{F}_{64} \subset \mathbf{R}$ denote the set of double-precision floating point numbers. If $x \in \mathbf{F}_{64}$, then x admits a finite decimal representation.

7. Let x be a real number. If $x \in \mathbf{F}_{64}$, then $2x \in \mathbf{F}_{64}$.

8. The following equality holds
$$(0.\overline{101})_2 = \frac{7}{3}.$$

9. In Julia, `256.0 + 2.0*eps(Float64) == 256.0` evaluates to `true`.

10. The set $\mathbf{F}_{64}$ of double-precision floating point numbers contains twice as many real numbers as the set $\mathbf{F}_{32}$ of single-precision floating point numbers.

11. Let x and y be two numbers in $\mathbf{F}_{64}$. The result of the machine addition $x \mathbin{\widehat{+}} y$ is sometimes exact and sometimes not, depending on the values of x and y.

Solution. The correct answers are the following:

1. True. The equality can be checked by converting the numbers to base 10 and then adding them, or by performing a long addition in base 3 directly.

2. True. Multiplication by $(1000)_2$ shifts the binary expansion 3 positions to the left.

3. True, because $0.25 = (0.01)_2$ in binary, which belongs to $\mathbf{F}_{32} \cap \mathbf{F}_{64}$.

4. False. This is why they are called *floating point* numbers.

5. False. The machine epsilon is related to the *relative* accuracy.

6. True, because all the powers of 2 admit a decimal representation with finitely many digits. Here we employ the word "admit" because the decimal expansion is not unique; for example, $(0.1)_2 = (0.5)_10 = (0.4\overline{9})_10$.

7. False. If the statement were true, then there would be an infinite amount of floating point numbers.

8. False. The left-hand side is < 1, and the right-hand side is > 1.

9. True. The next floating point number after 256 is $256(1 + \varepsilon)$.

10. False. It would take just one additional bit to store twice as many numbers.

11. True. It depends on whether $x + y$ belongs to $\mathbf{F}_{64}$ or not.

Question 2 (Iterative method for linear systems, **10 marks**). Assume that $\mathsf{A} \in \mathbf{R}^{n\times n}$ is a nonsingular matrix and that $\boldsymbol{b} \in \mathbf{R}^n$. We wish to solve the linear system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

using an iterative method where each iteration is of the form

$$\mathsf{M}\boldsymbol{x}_{k+1} = \mathsf{N}\boldsymbol{x}_k + \boldsymbol{b}. \tag{2}$$

Here $\mathsf{A} = \mathsf{M} - \mathsf{N}$ is a splitting of A such that M is nonsingular, and $\boldsymbol{x}_k \in \mathbf{R}^n$ denotes the k-th iterate of the numerical scheme.

1. (**3 marks**) Let $\boldsymbol{e}_k := \boldsymbol{x}_k - \boldsymbol{x}_*$, where $\boldsymbol{x}_*$ is the exact solution to (1). Prove that

$$\boldsymbol{e}_{k+1} = \mathsf{M}^{-1}\mathsf{N}\boldsymbol{e}_k.$$

2. (**3 marks**) Let $L = \|\mathsf{M}^{-1}\mathsf{N}\|_\infty$. Prove that

$$\forall k \in \mathbf{N}, \qquad \|\boldsymbol{e}_k\|_\infty \leq L^k \|\boldsymbol{e}_0\|_\infty.$$

3. (**1 mark**) Is the condition $\|\mathsf{M}^{-1}\mathsf{N}\|_\infty < 1$ necessary for convergence when $\boldsymbol{x}_0 \neq \boldsymbol{x}_*$?

4. (**3 marks**) Assume that A is strictly row diagonally dominant, in the sense that

$$\forall i \in \{1, \dots, n\}, \qquad |a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Show that, in this case, the inequality $\|\mathsf{M}^{-1}\mathsf{N}\|_\infty < 1$ holds for the Jacobi method, i.e. when M contains just the diagonal of A. You may take for granted the following expression for the ∞-norm of a matrix $\mathsf{X} \in \mathbf{R}^{n\times n}$:

$$\|\mathsf{X}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |x_{ij}|.$$

5. (**Bonus +1**) Write down a few iterations of the Jacobi method when

$$\mathsf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{b} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Is the method convergent?

Solution. **1.** We have

$$\begin{cases} \mathsf{M}\boldsymbol{x}_{k+1} = \mathsf{N}\boldsymbol{x}_k + \boldsymbol{b} \\ \quad \mathsf{M}\boldsymbol{x}_* = \mathsf{N}\boldsymbol{x}_* + \boldsymbol{b}. \end{cases}$$

The second equation holds because $\boldsymbol{x}_*$ is a solution to (1). Subtracting the second equation from the first, and multiplying both sides by M^{-1}, we obtain the required result.

2. By induction we have

$$\boldsymbol{e}_k = (\mathsf{M}^{-1}\mathsf{N})^k \boldsymbol{e}_0.$$

By definition of the $\|\bullet\|_\infty$ operator norm, we deduce that

$$\|\boldsymbol{e}_k\|_\infty \leq \|(\mathsf{M}^{-1}\mathsf{N})^k\|_\infty \|\boldsymbol{e}_0\|_\infty.$$

Since the norm $\|\bullet\|_\infty$ is submultiplicative, we conclude that

$$\|\boldsymbol{e}_k\|_\infty \leq \|\mathsf{M}^{-1}\mathsf{N}\|_\infty^k \|\boldsymbol{e}_0\|_\infty = L^k \|\boldsymbol{e}_0\|_\infty.$$

3. No. The condition is sufficient, because $\rho(\mathsf{M}^{-1}\mathsf{N}) \leq \|\mathsf{M}^{-1}\mathsf{N}\|_\infty$, but not necessary. See the bonus question for an example where convergence occurs but $\|\mathsf{M}^{-1}\mathsf{N}\|_\infty > 1$.

4. We have that

$$(\mathsf{M}^{-1}\mathsf{N})_{ij} = \begin{cases} 0 & \text{if } i = j \\ \frac{a_{ij}}{a_{ii}} & \text{if } i \neq j. \end{cases}.$$

By strict diagonal dominance, we deduce

$$\forall i \in \{1, \dots, n\}, \qquad \sum_{j=1}^{n} \left|(\mathsf{M}^{-1}\mathsf{N})_{ij}\right| = \sum_{j=1, j\neq i}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \frac{1}{|a_{ii}|} \sum_{j=1, j\neq i}^{n} |a_{ij}| < 1.$$

Therefore, we conclude that

$$\|\mathsf{M}^{-1}\mathsf{N}\|_\infty = \max_{1\leq i \leq n} \sum_{j=1}^{n} \left|(\mathsf{M}^{-1}\mathsf{N})_{ij}\right| < 1.$$

5. In this case

$$\mathsf{M}^{-1}\mathsf{N} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix},$$

which is a nilpotent matrix and so $\boldsymbol{e}_2 = (\mathsf{M}^{-1}\mathsf{N})^2 \boldsymbol{e}_0 = \boldsymbol{0}$; the method converges in two iterations.

Question 3 (Nonlinear equations, **10 marks**)**.** Assume that $\boldsymbol{x}_* \in \mathbf{R}^n$ is a solution to the equation

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x},$$

where $\boldsymbol{F} \colon \mathbf{R}^n \to \mathbf{R}^n$ is a smooth nonlinear function. We consider the following fixed-point iterative method for approximating $\boldsymbol{x}_*$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}(\boldsymbol{x}_k). \tag{3}$$

1. (**8 marks**) Assume in this part that $\boldsymbol{F}$ satisfies the local Lipschitz condition

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}_*)\| \leq L\|\boldsymbol{x} - \boldsymbol{x}_*\|, \tag{4}$$

with $0 \leq L < 1$ and $\delta > 0$. Here $B_\delta(\boldsymbol{x}_*)$ denotes the open ball of radius δ centered at $\boldsymbol{x}_*$. Show that the following statements hold:

- (**2 marks**) There is no fixed point of $\boldsymbol{F}$ in $B_\delta(\boldsymbol{x}_*)$ other than $\boldsymbol{x}_*$.
- (**2 marks**) If $\boldsymbol{x}_0 \in B_\delta(\boldsymbol{x}_*)$, then all the iterates $(\boldsymbol{x}_k)_{k \in \mathbf{N}}$ belong to $B_\delta(\boldsymbol{x}_*)$.
- (**3 marks**) If $\boldsymbol{x}_0 \in B_\delta(\boldsymbol{x}_*)$, then the sequence $(\boldsymbol{x}_k)_{k \in \mathbf{N}}$ converges to $\boldsymbol{x}_*$ and

$$\forall k \in \mathbf{N}, \qquad \|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leq L^k \|\boldsymbol{x}_0 - \boldsymbol{x}_*\|.$$

2. (**3 marks**) Explain with an example how the iterative scheme (3) can be employed for solving a nonlinear equation of the form

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}.$$

3. (**Bonus +1**) Let $\mathsf{J}_F \colon \mathbf{R}^n \to \mathbf{R}^{n \times n}$ denote the Jacobian matrix of $\boldsymbol{F}$. Show that if

$$\forall \boldsymbol{x} \in B_\delta(\boldsymbol{x}_*), \qquad \|\mathsf{J}_F(\boldsymbol{x})\| \leq L,$$

then the local Lipschitz condition (4) is satisfied.

Solution.

1.

- Assume by contradiction that there was another fixed point $\boldsymbol{y}_*$. Then, using the Lipschitz continuity, it would hold

$$\|\boldsymbol{y}_* - \boldsymbol{x}_*\| = \|\boldsymbol{F}(\boldsymbol{y}_*) - \boldsymbol{F}(\boldsymbol{x}_*)\| \leq L\|\boldsymbol{y}_* - \boldsymbol{x}_*\|,$$

  which is a contradiction because $L < 1$.

- The first iterate $\boldsymbol{x}_0$ is in $B_\delta(\boldsymbol{x}_*)$ by assumption. Reasoning by induction we assume that all the iterates up to $\boldsymbol{x}_k$ belong to $B_\delta(\boldsymbol{x}_*)$. Then, since $\boldsymbol{F}(\boldsymbol{x}_*) = \boldsymbol{x}_*$ by definition of $\boldsymbol{x}_*$, we have

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_*\| = \|\boldsymbol{F}(\boldsymbol{x}_k) - \boldsymbol{F}(\boldsymbol{x}_*)\| \leq L\|\boldsymbol{x}_k - \boldsymbol{x}_*\| < L\delta < \delta,$$

  implying that $\boldsymbol{x}_{k+1}$ is also in $B_\delta(\boldsymbol{x}_*)$. Note that we used the induction hypothesis twice: in the first inequality, because we need to know that $\boldsymbol{x}_k \in B_\delta(\boldsymbol{x}_*)$ in order to apply the local Lipschitz continuity (4), and then in the second inequality for the bound $\|\boldsymbol{x}_k - \boldsymbol{x}_*\| < \delta$.

- In the previous item, we showed that

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_*\| \leq L\|\boldsymbol{x}_k - \boldsymbol{x}_*\|.$$

  Iterating this inequality, we deduce that

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_*\| \leq L\|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leq \ldots \leq L^{k+1}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|.$$

2. A possible approach is to use the Newton–Raphson method.

Question 4 (Error estimate for eigenvalue problem, **10 marks**). Let $\|\bullet\|$ denote the Euclidean norm, and assume that $\mathsf{A} \in \mathbf{R}^{n \times n}$ is symmetric and nonsingular.

1. (**5 marks**) Describe with words and pseudocode a simple numerical method for calculating the eigenvalue of A of smallest modulus, as well as the corresponding eigenvector.

2. (**1 mark**) Let $\mathsf{M} \in \mathbf{R}^{n \times n}$ denote a nonsingular symmetric matrix. Prove that

$$\forall \boldsymbol{x} \in \mathbf{R}^n, \qquad \|\mathsf{M}\boldsymbol{x}\| \geq \|\mathsf{M}^{-1}\|^{-1}\|\boldsymbol{x}\|. \tag{5}$$

   Let $\lambda_{\min}(\mathsf{M})$ denote the eigenvalue of M of smallest modulus. Deduce from (5) that

$$\forall \boldsymbol{x} \in \mathbf{R}^n, \qquad \|\mathsf{M}\boldsymbol{x}\| \geq |\lambda_{\min}(\mathsf{M})|\|\boldsymbol{x}\|. \tag{6}$$

3. (**4 marks**) Assume that $\widehat{\lambda} \in \mathbf{R}$ and $\widehat{\boldsymbol{v}} \in \mathbf{R}^n$ are such that

$$\|\mathsf{A}\widehat{\boldsymbol{v}} - \widehat{\lambda}\widehat{\boldsymbol{v}}\| = \varepsilon > 0, \qquad \|\widehat{\boldsymbol{v}}\| = 1. \tag{7}$$

   Using (6), prove that there exists an eigenvalue λ of A such that

$$|\lambda - \widehat{\lambda}| \leq \varepsilon.$$

4. (**Bonus +1**) Show that, in the more general case where $\mathsf{A} = \mathsf{V}\mathsf{D}\mathsf{V}^{-1}$ is diagonalizable but not necessarily Hermitian, equation (7) implies the existence of an eigenvalue λ of A with

$$|\widehat{\lambda} - \lambda| \leq \|\mathsf{V}\|\|\mathsf{V}^{-1}\|\varepsilon.$$

   Hint: Introduce $\boldsymbol{r} = \mathsf{A}\widehat{\boldsymbol{v}} - \widehat{\lambda}\widehat{\boldsymbol{v}}$ and rewrite

$$\|\widehat{\boldsymbol{v}}\| = \|(\mathsf{A} - \widehat{\lambda}\mathsf{I})^{-1}\boldsymbol{r}\| = \|\mathsf{V}(\mathsf{D} - \widehat{\lambda}\mathsf{I})^{-1}\mathsf{V}^{-1}\boldsymbol{r}\|.$$

Question 5 (Interpolation error, **10 marks**). Let u denote the function

$$u\colon [0, 2\pi] \to \mathbf{R};$$
$$x \mapsto \cos(x).$$

Let $p_n\colon [0, 2\pi] \to \mathbf{R}$ denote the interpolating polynomial of u through at the nodes

$$x_i = \frac{2\pi i}{n}, \qquad i = 0, \ldots, n.$$

1. (**3 marks**) Using a method of your choice, calculate p_n for $n = 2$.

2. (**6 marks**) Let $n \in \mathbf{N}_{>0}$ and $e_n(x) := u(x) - p_n(x)$. Prove that

$$\forall x \in [0, 2\pi], \qquad |e_n(x)| \leq \frac{|\omega(x)|}{(n+1)!},$$

   where we introduced

$$\omega_n(x) := \prod_{i=0}^{n} (x - x_i).$$

   Hint: You may find it useful to introduce the function

$$g(t) = e_n(t)\omega_n(x) - e_n(x)\omega_n(t).$$

3. (**1 mark**) Does the maximum absolute error

$$E_n := \sup_{x \in [0, 2\pi]} |e_n(x)|$$

   tend to zero in the limit as $n \to \infty$?

(**Bonus +1**) Using the Gregory–Newton formula, find a closed expression for the sum

$$S(n) = \sum_{k=1}^{n} k^2.$$

Question 6 (Numerical integration, **10 marks**). The third exercise below is independent of the first two.

1. (**5 marks**) Construct an integration rule of the form

$$\int_{-1}^{1} u(x)\,\mathrm{d}x \approx w_1 u\left(-\frac{1}{2}\right) + w_2 u(0) + w_3 u\left(\frac{1}{2}\right)$$

with a degree of precision equal to at least 2.

2. (**1 mark**) What is the degree of precision of the rule constructed?

3. (**4 marks**) The Gauss–Laguerre quadrature rule with n nodes is an approximation of the form

$$\int_{0}^{\infty} u(x)\,\mathrm{e}^{-x}\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i u(x_i),$$

such that the rule is exact when u is a polynomial of degree less than or equal to $2n-1$. Find the Gauss–Laguerre rule with one node ($n = 1$).

4. (**Bonus +1**) Find the Gauss–Laguerre quadrature rule with two nodes ($n = 2$). You may find it useful to first calculate the Laguerre polynomial of degree 2.