

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

$$(1000)_2 \times (0.1\overline{01})_2 = (101.0\overline{1})_2.$$

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

$$(0.1\overline{01})_2 = \frac{7}{3}.$$

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 $x \hat{+} y$ is sometimes exact and sometimes not, depending on the values of x and y .

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

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1}$$

using an iterative method where each iteration is of the form

$$\mathbf{M}\mathbf{x}_{k+1} = \mathbf{N}\mathbf{x}_k + \mathbf{b}.$$

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

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

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

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

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

3. (1 marks) Is the condition $\|\mathbf{M}^{-1}\mathbf{N}\|_\infty < 1$ necessary for convergence when $\mathbf{x}_0 \neq \mathbf{x}_*$?
4. (3 marks) Assume that \mathbf{A} is strictly row diagonally dominant, in the sense that

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

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

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

5. (**Bonus +2**) Write down a few iterations of the Jacobi method when

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Is the method convergent?

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

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

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

$$\mathbf{x}_{k+1} = \mathbf{F}(\mathbf{x}_k). \quad (2)$$

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

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{x}_*)\| \leq L\|\mathbf{x} - \mathbf{x}_*\|, \quad (3)$$

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

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

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

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

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

3. (Bonus +2) Let $\mathbf{J}_F: \mathbf{R}^n \rightarrow \mathbf{R}^{n \times n}$ denote the Jacobian matrix of \mathbf{F} . Show that if

$$\forall \mathbf{x} \in B_\delta(\mathbf{x}_*), \quad \|\mathbf{J}_F(\mathbf{x})\| \leq L,$$

then the local Lipschitz condition (3) is satisfied.

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

1. (5 marks) Describe with words and pseudocode a simple numerical method for calculating the eigenvalue of \mathbf{A} of smallest modulus, as well as the corresponding eigenvector.
2. (**Bonus +1**) Let $\mathbf{M} \in \mathbf{R}^{n \times n}$ denote a nonsingular symmetric matrix. Prove that

$$\forall \mathbf{x} \in \mathbf{R}^n, \quad \|\mathbf{M}\mathbf{x}\| \geq \|\mathbf{M}^{-1}\|^{-1} \|\mathbf{x}\|. \quad (4)$$

Let $\lambda_{\min}(\mathbf{M})$ denote the eigenvalue of \mathbf{M} of smallest modulus. Deduce from (4) that

$$\forall \mathbf{x} \in \mathbf{R}^n, \quad \|\mathbf{M}\mathbf{x}\| \geq |\lambda_{\min}(\mathbf{M})| \|\mathbf{x}\|. \quad (5)$$

3. (5 marks) Assume that $\hat{\lambda} \in \mathbf{R}$ and $\hat{\mathbf{v}} \in \mathbf{R}^n$ are such that

$$\|\mathbf{A}\hat{\mathbf{v}} - \hat{\lambda}\hat{\mathbf{v}}\| = \varepsilon > 0, \quad \|\hat{\mathbf{v}}\| = 1. \quad (6)$$

Using (5), prove that there exists an eigenvalue λ of \mathbf{A} such that

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

4. (**Bonus +1**) Show that, in the more general case where $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ is diagonalizable but not necessarily Hermitian, equation (6) implies the existence of an eigenvalue λ of \mathbf{A} with

$$|\hat{\lambda} - \lambda| \leq \|\mathbf{V}\| \|\mathbf{V}^{-1}\| \varepsilon.$$

Hint: Introduce $\mathbf{r} = \mathbf{A}\hat{\mathbf{v}} - \hat{\lambda}\hat{\mathbf{v}}$ and rewrite

$$\|\hat{\mathbf{v}}\| = \|(\mathbf{A} - \hat{\lambda}\mathbf{I})^{-1}\mathbf{r}\| = \|\mathbf{V}(\mathbf{D} - \hat{\lambda}\mathbf{I})^{-1}\mathbf{V}^{-1}\mathbf{r}\|.$$

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

$$\begin{aligned} u: [0, 2\pi] &\rightarrow \mathbf{R}; \\ x &\mapsto \cos(x). \end{aligned}$$

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

$$x_i = \frac{2\pi i}{n}, \quad i = 0, \dots, n.$$

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

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

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 $n \rightarrow \infty$?

(Bonus +2) 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

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

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

$$\int_0^\infty u(x) e^{-x} \, dx \approx \sum_{i=1}^n w_i u(x_i),$$

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 +2**) 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.