

Numerical Analysis: Final Exam

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

Urbain Vaes

12 December 2022

You are allowed to use a calculator, but not *Julia* or *Python*.

Academic integrity pledge

☐ I certify that I will not give or receive any unauthorized help on this exam, and that all work will be my own. (Tick ✓ or copy the sentence on your answer sheet).

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

1. Let $(\bullet)_3$ denote base 3 representation. It holds that

$$(222, 222)_3 + (1)_3 = (1, 000, 000)_3.$$

2. Let $(\bullet)_2$ denote base 2 representation. It holds that

$$3 \times (0.0101)_2 = (0.1111)_2.$$

3. The following equality holds

$$(0.\overline{011})_2 = \frac{3}{4}.$$

4. The number $x = (d_1 d_2 d_3)_3$ for $d_1, d_2, d_3 \in \{0, 1, 2\}$ is a multiple of 3 if and only if $d_3 = 0$.

5. In Julia, `Float64(0.375) == Float32(0.375)` evaluates to `true`.

6. The value of the machine epsilon is the same for the single precision (\mathbf{F}_{32}) and the double precision (\mathbf{F}_{64}) formats.

7. The spacing (in absolute value) between successive double-precision (`Float64`) floating point numbers is equal to the machine epsilon.

8. All the natural numbers can be represented exactly in the double precision floating point format \mathbf{F}_{64} .

9. Machine addition in the \mathbf{F}_{64} format is associative but not commutative.

10. In Julia `exp(eps()) == 1 + eps()` evaluates to `true`. (Remember that, by default, rounding is to the nearest representable number).

11. In Julia `sqrt(1 + eps()) == 1 + eps()` evaluates to `true`.

12. Let x and y be two numbers in \mathbf{F}_{64} . The result of the machine multiplication $x \hat{*} y$ is sometimes exact and sometimes not, depending on the values of x and y .

13. In Julia, let `f(x) = (x == x/100.0) ? x : f(x/100.0)`¹. Then `f(3.0)` returns `0.0`.

¹In Python, let `f = lambda x: x if x == x/100.0 else f(x/100.0)`

Question 2 (Interpolation, **10 marks**). Let $u: [-1, 1] \rightarrow \mathbf{R}$ be given by

$$u(x) = x^3.$$

Let $p: [-1, 1] \rightarrow \mathbf{R}$ denote the interpolating polynomial of u at nodes $x_0 < x_1 < x_2$, all contained in the interval $[-1, 1]$.

1. (**2 marks**) Let $e(x) := u(x) - p(x)$. Prove, without assuming any result shown in class, that the interpolation error satisfies

$$\forall x \in [0, 1], \quad e(x) = (x - x_0)(x - x_1)(x - x_2).$$

2. (**2 marks**) Using a method of your choice, calculate the interpolating polynomial p in the particular case where

$$x_0 = -1, \quad x_1 = 0, \quad x_2 = 1. \quad (1)$$

3. (**2 marks**) We denote the maximum absolute value of the error by

$$E := \max_{x \in [-1, 1]} |e(x)|. \quad (2)$$

Calculate the value of E in the particular case (1).

4. (**2 marks**) We denote by $T_3: [-1, 1] \rightarrow \mathbf{R}$ the Chebyshev polynomial given by

$$T_3(x) := \cos(3 \arccos(x)).$$

Show that

$$T_3(x) = 4x^3 - 3x$$

and calculate the roots z_0, z_1, z_2 of T_3 .

Hint: Note that $\cos(3\theta) = \Re(e^{i3\theta}) = \Re((e^{i\theta})^3)$, where $e^{i\theta} = \cos(\theta) + i \sin(\theta)$.

5. (**2 marks**) Find the expression of the error $e(x)$ and the maximum absolute error E given in (2) in the case where the interpolation nodes x_0, x_1, x_2 are given by z_0, z_1, z_2 .
6. ***(Bonus +2)** Show that the maximum absolute error (2), viewed as a function of the interpolation nodes x_0, x_1, x_2 , is minimized when $x_i = z_i$ for $i \in \{0, 1, 2\}$.

Hint: Reason by contradiction and notice that

$$|T_3(y)| = 1 \quad \text{for } y \in \left\{-1, -\frac{1}{2}, \frac{1}{2}, 1\right\}.$$

Question 3 (Numerical integration, **10 marks**). Let $u: [0, 1] \rightarrow \mathbf{R}$ be a function we wish to integrate and

$$I := \int_0^1 u(x) \, dx.$$

1. (**3 marks**) Consider the following integration rule:

$$I \approx w_1 u(0) + w_2 u(1). \quad (3)$$

Find the weights $w_1, w_2 \in \mathbf{R}$ so that this integration rule has the highest possible degree of precision. What is the degree of precision of the rule constructed?

2. (**3 marks**) Let $x_i = i/n$ for $i = 0, \dots, n$. The composite trapezoidal rule is given by

$$I \approx \frac{1}{2n} (u(x_0) + 2u(x_1) + 2u(x_2) + \dots + 2u(x_{n-2}) + 2u(x_{n-1}) + u(x_n)) =: \hat{I}_n. \quad (4)$$

Explain how this rule can be obtained by applying a generalization of the integration rule (3) in each interval $[x_i, x_{i+1}]$.

3. (**3 marks**) Assume that $u \in C^2([0, 1])$. Show that, for all $n \in \mathbf{N}_{>0}$,

$$|I - \hat{I}_n| \leq \frac{C_2}{12n^2}, \quad C_2 := \sup_{\xi \in [0, 1]} |u''(\xi)|. \quad (5)$$

You may use **Proposition 1** at the end of this document for the interpolation error.

4. (**1 mark**) In this part of the question, we assume that u is a quadratic polynomial. It is possible to show that, in this case,

$$I - \hat{I}_n = -\frac{u''(0)}{12n^2}.$$

Explain how, given two approximations \hat{I}_n and \hat{I}_{2n} obtained with (4), a better approximation of the integral I can be obtained by a linear combination of the form

$$\alpha_1 \hat{I}_n + \alpha_2 \hat{I}_{2n}.$$

5. ***(Bonus +2)** Instead of (3), consider a more general integration rule of the form

$$\int_0^1 u(x) \, dx \approx w_1 u(x_1) + w_2 u(x_2). \quad (6)$$

Find the weights $w_1, w_2 \in \mathbf{R}$ and the nodes $x_1, x_2 \in [0, 1]$ so that this integration rule has the highest possible degree of precision. What is the degree of precision obtained?

Question 4 (Iterative method for linear systems, **10 marks**). Assume that $\mathbf{A} \in \mathbf{R}^{n \times n}$ is a *symmetric positive definite* matrix and that $\mathbf{b} \in \mathbf{R}^n$. We wish to solve the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \quad (7)$$

To this end we consider an iterative method where each iteration is of the form

$$\mathbf{M}\mathbf{x}_{k+1} = \mathbf{N}\mathbf{x}_k + \mathbf{b}. \quad (8)$$

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 (7). Prove that

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

2. (**2 marks**) We denote by $\|\bullet\|_{\mathbf{A}}$ the vector norm

$$\|\mathbf{x}\|_{\mathbf{A}} := \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}}, \quad (9)$$

and we use the same notation for the induced matrix norm. Prove that

$$\forall k \in \mathbf{N}, \quad \|\mathbf{e}_k\|_{\mathbf{A}} \leq L^k \|\mathbf{e}_0\|_{\mathbf{A}}, \quad L := \|\mathbf{M}^{-1}\mathbf{N}\|_{\mathbf{A}}. \quad (10)$$

3. (**1 mark**) Is the condition $\|\mathbf{M}^{-1}\mathbf{N}\|_{\mathbf{A}} < 1$ sufficient to ensure convergence for all \mathbf{x}_0 ?

4. ***(3 marks)** Show that

$$\|\mathbf{M}^{-1}\mathbf{N}\mathbf{x}\|_{\mathbf{A}}^2 = \|\mathbf{x}\|_{\mathbf{A}}^2 - \mathbf{y}^T (\mathbf{M}^T + \mathbf{N}) \mathbf{y}, \quad \mathbf{y} := \mathbf{M}^{-1}\mathbf{A}\mathbf{x}. \quad (11)$$

Hint: Eliminate \mathbf{N} from both sides of the equation by rewriting $\mathbf{N} = \mathbf{M} - \mathbf{A}$. Then substitute the expression of \mathbf{y} and expand both sides. Remember that a scalar quantity transposed is equal to itself.

5. (**1 mark**) Show that, for the Gauss–Seidel method, i.e. when $\mathbf{M} = \mathbf{L} + \mathbf{D}$ contains just the lower triangular and diagonal parts of \mathbf{A} , it holds that

$$\mathbf{M}^T + \mathbf{N} = \mathbf{D}. \quad (12)$$

6. (**Bonus +2**) Deduce from (11) and (12) that, for the Gauss–Seidel method,

$$\|\mathbf{M}^{-1}\mathbf{N}\|_{\mathbf{A}} < 1.$$

Question 5 (Nonlinear equations, **10 marks**). We consider the following iterative method for calculating $\sqrt[3]{2}$:

$$x_{k+1} = F(x_k) := \omega x_k + (1 - \omega) \frac{2}{x_k^2}, \quad (13)$$

with $\omega \in (0, 1)$ a fixed parameter.

1. **(1 mark)** Show that $x_* := \sqrt[3]{2}$ is a fixed point of the iteration (13).
2. **(2 marks)** Write down in pseudocode a computer program based on the iteration (13) for calculating $\sqrt[3]{2}$. Use an appropriate stopping criterion that does not require to know the value of $\sqrt[3]{2}$.
3. **(2 marks)** Prove that if $\omega \in (\frac{1}{3}, 1)$, then x_* is locally exponentially stable. You may take for granted **Proposition 2** at the end of this document.
4. **(1 mark)** Do you expect faster convergence of (13) with $\omega = \frac{1}{2}$ or with $\omega = \frac{2}{3}$?
5. **(2 marks)** Show that, in the particular case where $\omega = \frac{2}{3}$, the iterative scheme (13) coincides with the Newton–Raphson method applied to the nonlinear equation

$$f(x) = 0, \quad (14)$$

for an appropriate function $f: \mathbf{R} \rightarrow \mathbf{R}$.

6. **(2 marks)** Illustrate graphically a few iterations of the Newton–Raphson method for solving (14) when starting from $x_0 = 2$. You may either create your own figure or write on **Figure 1** at the end of this document.
7. ***(Bonus +2)** Prove **Proposition 2** in the appendix. More precisely, show that the assumptions of the proposition imply that there is $\delta > 0$ and $L < 1$ such that the following local Lipschitz condition is satisfied:

$$\forall x \in [x_* - \delta, x_* + \delta], \quad |F(x) - F(x_*)| \leq L|x - x_*|. \quad (15)$$

For completeness, one should then show that (15) is sufficient to guarantee local exponential stability, but this is taken for granted here; you do not need to prove this.

Question 6 (Iterative methods for eigenvalue problems, **10 marks**). Let $\|\bullet\|$ denote both the Euclidean norm on vectors and the induced matrix norm. Assume that $\mathbf{A} \in \mathbf{R}^{n \times n}$ is symmetric and nonsingular, and that all the eigenvalues of \mathbf{A} have different moduli:

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|.$$

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. (**2 marks**) Suppose that we have calculated the smallest eigenvalue in modulus λ_n , as well as the associated normalized eigenvector \mathbf{v}_n . We let

$$\mathbf{B} := \mathbf{A}^{-1} - \frac{1}{\lambda_n} \mathbf{v}_n \mathbf{v}_n^T.$$

If we apply the power iteration to this matrix, what convergence can we expect? Justify your answer.

3. ***(3 marks)** The aim of this part is to provide an answer to the following question: given an approximate eigenpair $(\hat{\mathbf{v}}, \hat{\lambda})$, what is the smallest perturbation \mathbf{E} that we need to apply to \mathbf{A} in order to guarantee that $(\hat{\mathbf{v}}, \hat{\lambda})$ is an exact eigenpair, i.e. that

$$(\mathbf{A} + \mathbf{E})\hat{\mathbf{v}} = \hat{\lambda}\hat{\mathbf{v}}?$$

Assume that $\hat{\mathbf{v}}$ is normalized and let $\mathcal{E} = \{\mathbf{E} \in \mathbf{C}^{n \times n} : (\mathbf{A} + \mathbf{E})\hat{\mathbf{v}} = \hat{\lambda}\hat{\mathbf{v}}\}$. Prove that

$$\min_{\mathbf{E} \in \mathcal{E}} \|\mathbf{E}\| = \|\mathbf{r}\|, \quad \mathbf{r} := \mathbf{A}\hat{\mathbf{v}} - \hat{\lambda}\hat{\mathbf{v}}. \quad (16)$$

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

- Show first that $\mathbf{E} \in \mathcal{E}$ if and only if $\mathbf{E}\hat{\mathbf{v}} = -\mathbf{r}$.
- Deduce from the previous item that

$$\forall \mathbf{E} \in \mathcal{E}, \quad \|\mathbf{E}\| \geq \|\mathbf{r}\|.$$

- Find a rank one matrix $\mathbf{E}_* \in \mathcal{E}$ such that $\|\mathbf{E}_*\| = \|\mathbf{r}\|$, and then conclude. Recall that any rank 1 matrix can be written in the form $\mathbf{E}_* = \mathbf{u}\mathbf{w}^*$, with norm $\|\mathbf{u}\|\|\mathbf{w}\|$.

4. (**Bonus +2**) Suppose that we have calculated λ_n and λ_{n-1} together with the associated normalized eigenvectors. Propose a method for calculating the third smallest eigenvalue in modulus, i.e. λ_{n-2} .

Auxiliary results

Proposition 1. Assume that $f: [a, b] \rightarrow \mathbf{R}$ is a function in $C^2([a, b])$ and let \hat{f} denote the interpolation of f at two distinct interpolation nodes y_1, y_2 . Then there exists $\xi: [a, b] \rightarrow [a, b]$ such that

$$\forall y \in [a, b], \quad f(y) - \hat{f}(y) = \frac{f''(\xi(y))}{2}(y - y_1)(y - y_2).$$

Proposition 2. Assume that $F: (0, \infty) \rightarrow (0, \infty)$ is continuously differentiable, and suppose that $x_* \in (0, \infty)$ is a fixed point of the iteration $x_{k+1} = F(x_k)$. If

$$|F'(x_*)| < 1,$$

then the fixed point x_* is locally exponentially stable.

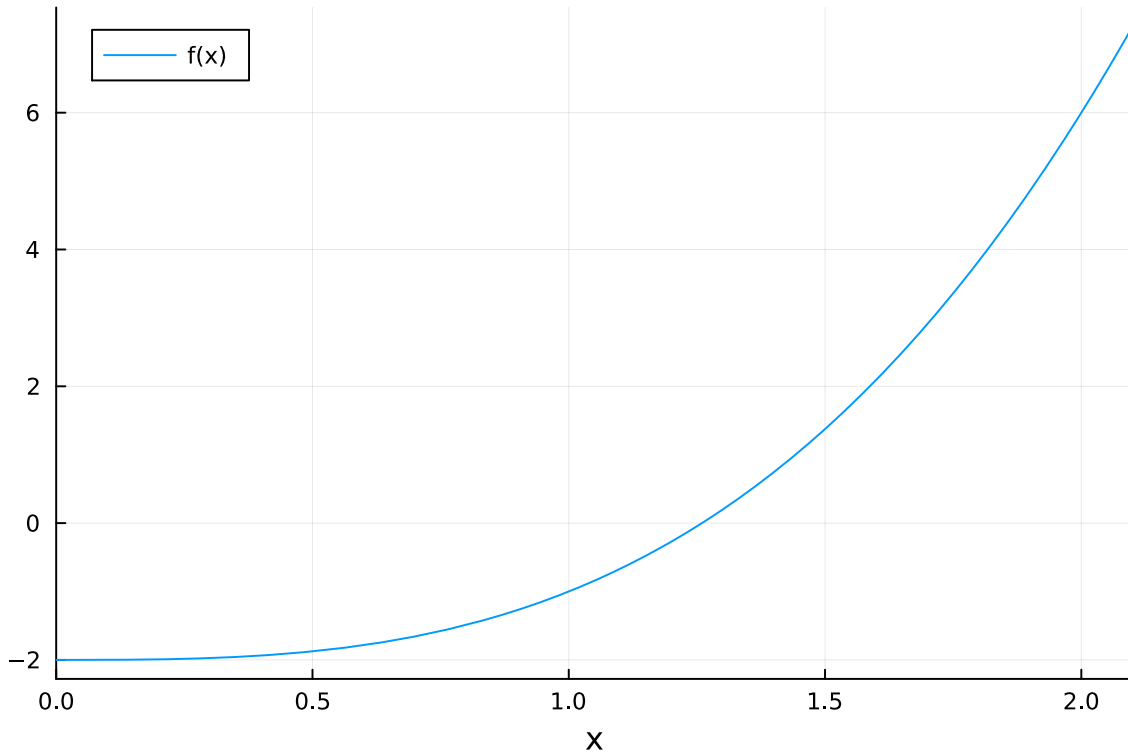


Figure 1: You can use this figure to illustrate the Newton–Raphson method.