Numerical Analysis: Final Exam

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

Urbain Vaes

December 12, 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 $x \mathbin{\widehat{*}} y$ 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`.

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

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

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

Let $p\colon [-1,1] \to \mathbf{R}$ 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

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

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

$$x_0 = -1, \qquad x_1 = 0, \qquad x_2 = 1. \tag{1}$$

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

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

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

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

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

Show that

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

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

Hint: Note that $\cos(3\theta) = \Re\left(\mathrm{e}^{\mathrm{i}3\theta}\right) = \Re\bigl(\left(\mathrm{e}^{\mathrm{i}\theta}\right)^3\bigr)$, where $\mathrm{e}^{i\theta} = \cos(\theta) + \mathrm{i}\sin(\theta)$.

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

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

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

$$I := \int_0^1 u(x)\,\mathrm{d}x.$$

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

$$I \approx w_1 u(0) + w_2 u(1). \tag{3}$$

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

$$I \approx \frac{1}{2n}\big(u(x_0) + 2u(x_1) + 2u(x_2) + \cdots + 2u(x_{n-2}) + 2u(x_{n-1}) + u(x_n)\big) =: \widehat{I}_n. \tag{4}$$

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

$$\big|I - \widehat{I}_n\big| \leqslant \frac{C_2}{12n^2}, \qquad C_2 := \sup_{\xi \in [0,1]} \big|u''(\xi)\big|. \tag{5}$$

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,

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

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

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

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

$$\int_0^1 u(x)\,\mathrm{d}x \approx w_1 u(x_1) + w_2 u(x_2). \tag{6}$$

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 $\mathsf{A} \in \mathbf{R}^{n\times n}$ is a *symmetric positive definite* matrix and that $\boldsymbol{b} \in \mathbf{R}^n$. We wish to solve the linear system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b}. \tag{7}$$

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

$$\mathsf{M}\boldsymbol{x}_{k+1} = \mathsf{N}\boldsymbol{x}_k + \boldsymbol{b}. \tag{8}$$

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 (7). Prove that

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

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

$$\|\boldsymbol{x}\|_{\mathsf{A}} := \sqrt{\boldsymbol{x}^T\mathsf{A}\boldsymbol{x}}, \tag{9}$$

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

$$\forall k \in \mathbf{N}, \qquad \|\boldsymbol{e}_k\|_{\mathsf{A}} \leqslant L^k\|\boldsymbol{e}_0\|_{\mathsf{A}}, \qquad L := \|\mathsf{M}^{-1}\mathsf{N}\|_{\mathsf{A}}. \tag{10}$$

3. (**1 mark**) Is the condition $\|\mathsf{M}^{-1}\mathsf{N}\|_{\mathsf{A}} < 1$ sufficient to ensure convergence for all $\boldsymbol{x}_0$?
4. *(**3 marks**) Show that

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

   Hint: Eliminate N from both sides of the equation by rewriting $\mathsf{N} = \mathsf{M} - \mathsf{A}$. Then substitute the expression of $\boldsymbol{y}$ 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 $\mathsf{M} = \mathsf{L} + \mathsf{D}$ contains just the lower triangular and diagonal parts of A, it holds that

$$\mathsf{M}^T + \mathsf{N} = \mathsf{D}. \tag{12}$$

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

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

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

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

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 $\omega \in \left(\frac{1}{3}, 1\right)$, 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

$$f(x) = 0, \tag{14}$$

for an appropriate function $f\colon \mathbf{R} \to \mathbf{R}$.

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:

$$\forall x \in [x_* - \delta, x_* + \delta], \qquad |F(x) - F(x_*)| \leqslant L|x - x_*|. \tag{15}$$

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 $\mathsf{A} \in \mathbf{R}^{n\times n}$ is symmetric and nonsingular, and that all the eigenvalues of 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 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 $\boldsymbol{v}_n$. We let

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

   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 $(\widehat{\boldsymbol{v}}, \widehat{\lambda})$, what is the smallest perturbation E that we need to apply to A in order to guarantee that $(\widehat{\boldsymbol{v}}, \widehat{\lambda})$ is an exact eigenpair, i.e. that

$$(\mathsf{A} + \mathsf{E})\widehat{\boldsymbol{v}} = \widehat{\lambda}\widehat{\boldsymbol{v}}\,?$$

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

$$\min_{\mathsf{E}\in\mathcal{E}} \|\mathsf{E}\| = \|\boldsymbol{r}\|, \qquad \boldsymbol{r} := \mathsf{A}\widehat{\boldsymbol{v}} - \widehat{\lambda}\widehat{\boldsymbol{v}}. \tag{16}$$

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

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

$$\forall \mathsf{E} \in \mathcal{E}, \qquad \|\mathsf{E}\| \geqslant \|\boldsymbol{r}\|.$$

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

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 $f\colon [a,b] \to \mathbf{R}$ is a function in $C^2([a,b])$ and let $\widehat{f}$ denote the interpolation of f at two distinct interpolation nodes y_1, y_2. Then there exists $\xi\colon [a,b] \to [a,b]$ such that*

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

Proposition 2. *Assume that $F\colon (0,\infty) \to (0,\infty)$ 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.

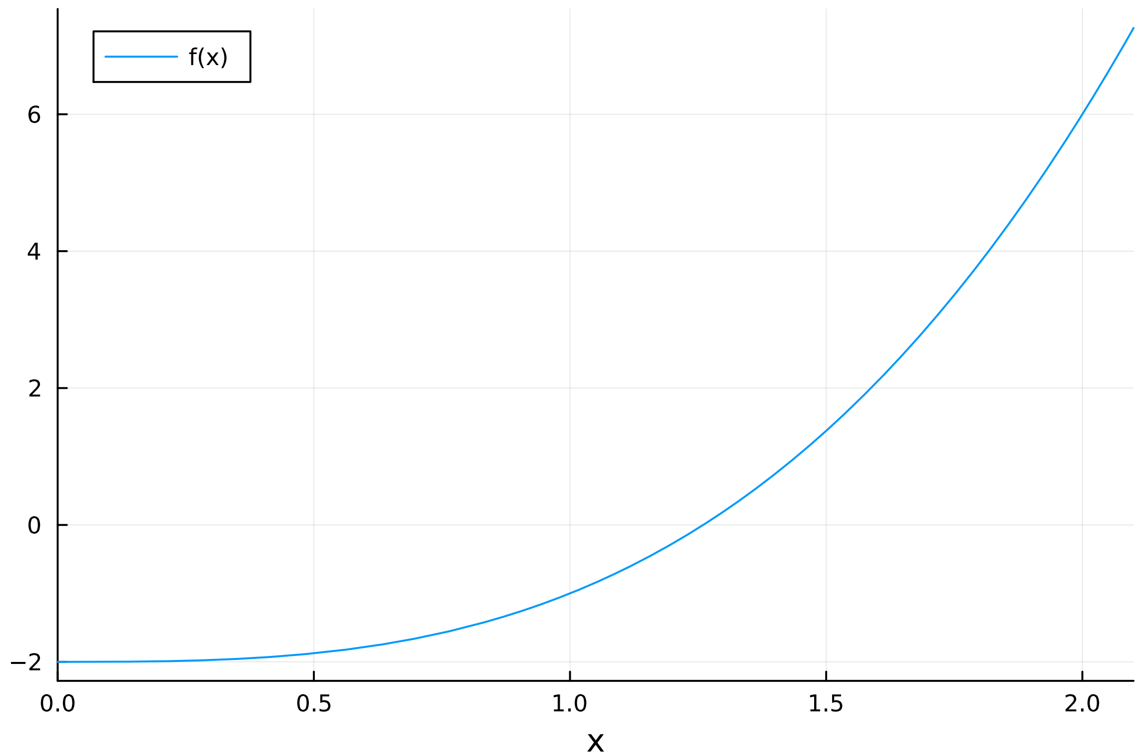

Figure 1: You can use this figure to illustrate the Newton–Raphson method.