Department of Applied Mathematics
Preliminary Examination in Numerical Analysis
August 2023

Instructions. You have three hours to complete this exam. Submit solutions to four (and no more) of the following six problems. All problems have equal value.

Please start each problem on a new page. You MUST prove your conclusions or show a counter-example for all problems unless otherwise noted.
Write your student ID number (not your name!) on your exam.

Problem 1: Root finding

(a) Consider the following iteration schemes of the form $x_{n+1} = f(x_n)$ each with a proposed fixed point α. Which of the following will converge (provided x_0 is sufficiently close to α)? If it does converge, give the order of convergence; for linear convergence, give the rate of linear convergence

(i) $x_{n+1} = -16 + 6x_n + \dfrac{12}{x_n}, \qquad \alpha = 2$

(ii) $x_{n+1} = \dfrac{2}{3}x_n + \dfrac{1}{x_n^2}, \qquad \alpha = 3^{1/3}$

(iii) $x_{n+1} = \dfrac{12}{1+x_n}, \qquad \alpha = 3$

(b) Consider an analytic function $f(x)$ such that the fixed point iteration

$$x_{n+1} = f(x_n)$$

for any initial value of $x_0 \neq 0$ eventually hops between $+1$ and -1. Describe the properties of $f(x)$ for $|x| = 1$, $|x| < 1$, and $|x| > 1$ that would make this limiting sequence possible. Sketch such a function.

Solution:

(a) If $f(\alpha) = \alpha$ then α is a fixed point. The iteration scheme converges/diverges near α according to $|f'(\alpha)| < 1$ or > 1.

(i) $f'(x) = 6 - \dfrac{12}{x^2}, \qquad f(2) = -16 + 6.2 + \dfrac{12}{2} = 2$ & $f'(2) = 6 - \dfrac{12}{4} = 3$ implies iterative scheme is divergent.

(ii) $f'(x) = \dfrac{2}{3} - \dfrac{2}{x^3}$, $f''(x) = \dfrac{6}{x^4} \qquad f(3^{1/3}) = \dfrac{2}{3}.3^{1/3} + 3^{-2/3} = 3^{1/3}$ & $f'(3^{1/3}) = \dfrac{2}{3} - \dfrac{2}{3} = 0$ implies iterative scheme is supercritically convergent. $f''(3^{1/3}) \neq 0$ convergence is quadratic (and not faster than that).

(iii) $f'(x) = -\dfrac{12}{(1+x)^2}, \qquad f(3) = \dfrac{12}{1+3} = 3$ & $f'(3) = -\dfrac{12}{4^2} = -\dfrac{3}{4}$ implies iterative scheme is linearly convergent with rate $3/4$.

(b)

- $|x| = 1$, the map should satisfy $f(1) = -1$ and $f(-1) = 1$.
- $|x| < 1$, the map should satisfy $|f(x)| > |x|$.
- $|x| > 1$, the map should satisfy $|f(x)| < |x|$.

Problem 2: Interpolation/Approximation

(a) Define what it means for a polynomial $p_N(x)$ to be a minimax approximation of degree N.

(b) Find the first degree Taylor polynomial approximating e^x in the interval $[-1, 1]$ centered at $a = 0$. Then find the maximum norm of the error in this approximation.

(c) Find the first degree polynomial least squares approximation of the function e^x that minimizes the error in the following norm:

$$\|f - g\| := \sqrt{\int_{-1}^{1} |f(x) - g(x)|^2 dx}$$

.

(d) Create the polynomial that interpolates e^x with the nodes $x_0 = -1$ and $x_1 = 1$.

(e) Which of the three polynomials that you created is the closest to the optimal approximating polynomial in the Minimax sense and why?

Solution:

(a) For a degree N polynomial, there are N values of x where the error is 0 and $N + 2$ values of x where the absolute value of the error is the maximum.

(b) The Taylor polynomial is $p_1(x) = 1 + x$. The maximum error is obtained at the right end point $\max_{x\in[-1,1]} |e^x - p_1(x)| = |e - 2|$

(c) There are two ways to obtain the L^2 approximation: (i) remembering that the Legendre polynomials are orthogonal on the interval $[-1, 1]$ and using them to create the approximation. (ii) Define the error and look for the minimum. Both ways give the same polynomial.

Option i: $p_1(x) = \frac{\int_{-1}^{1} e^x dx}{\int_{-1}^{1} 1 dx} + \frac{\int_{-1}^{1} xe^x dx}{\int_{-1}^{1} x^2 dx} x = \frac{e-e^{-1}}{2} + 3e^{-1}x$

Option ii: Define $E^2 = \int_{-1}^{1} |e^x - a - bx|^2 dx$. Take derivatives with respect to the parameters and set to 0; i.e. solve $\frac{\partial E^2}{\partial a} = 0$ and $\frac{\partial E^2}{\partial b} = 0$ for a and b. The polynomial as part a results.

(d) The interpolating polynomial through the points is $p_1(x) = e^{-1}\frac{x-1}{-2} + e^{1}\frac{x+1}{2}$.

(e) We expect the polynomial from part (d) to have the smallest maximum norm in the error because the interpolation nodes are Chebychev nodes which are nearly optimal interpolation nodes in the maximum norm sense.

Problem 3: Quadrature

Gaussian quadratures that are approximated as follows

$$\int_{-1}^{1} f(x)dx \sim \sum_{k=0}^{N} w_k f(x_k)$$

where the quadrature nodes include the endpoints (i.e. $x_0 = -1$ and $x_N = 1$) are called *Gauss-Legendre-Lobatto* quadratures.

(a) Show that if the interior nodes $x_1, \ldots, x_{N-1}$ in the quadrature are given by the roots of $P_N'(x)$ where $P_N(x)$ is the N^{th} degree Legendre polynomial, then the quadrature is exact for polynomials up to degree $2N-1$.
Hint: The following recurrence relation is true:

$$(x^2-1)P_N'(x) = xP_N(x) - P_{N-1}(x)$$

(b) Find the 4−point Gauss-Legendree-Lobatto quadrature (nodes and weights) for approximating the integral $\int_{-1}^{1} f(x)dx$.
It is enough to set up a closed formula which evaluates each of the weights independently.
Hint: The three term recursion for Legendre polynomials is given by

$$P_0(x) = 1,\ P_1(x) = x,\ (k)P_k(x) - (2k-1)xP_{k-1}(x) + (k-1)P_{k-2}(x) = 0$$

Solution:

(a) Let $s(x)$ denote a polynomial of degree $2N-1$. Then by polynomial long division

$$s(x) = q(x)(x^2-1)P_N'(x) + r(x)$$

where $q(x)$ is a polynomial of degree of at most $N-2$ and $r(x)$ is a polynomial of degree at most N.

Then

$$\int_{-1}^{1} s(x)dx = \int_{-1}^{1} q(x)(x^2-1)P_N'(x)dx + \int_{-1}^{1} r(x)dx.$$

First let's take a closer look at $\int_{-1}^{1} q(x)(x^2-1)P_N'(x)dx$.

By the hint, we know that

$$\int_{-1}^{1} q(x)(x^2-1)P_N'(x)dx = N\int_{-1}^{1} q(x)xP_N(x)dx - N\int_{-1}^{1} q(x)P_{N-1}(x)dx$$

The polynomial $xq(x)$ is of degree $N-1$ and thus is orthogonal to $P_N(x)$. Similarily, $q(x)$ is a polynomial of degree $N-2$ and is orthogonal to $P_{N-1}(x)$. Thus this integral is 0.

This means that

$$\int_{-1}^{1} s(x)dx = \int_{-1}^{1} r(x)dx$$

and that the quadrature should satisfy the following

$$\sum_{j=0}^{N} w_j s(x_j) = \sum_{j=0}^{N} w_j r(x_j)$$

and

$$\sum_{j=0}^{N} \left(q(x_j)(x_j^2 - 1)P_N'(x_j)\right) w_j$$

for all functions $s(x)$. The only way that this can be true for all polynomials of degree $2N-1$ is the quadrature nodes are ±1 and the roots of $P_N'(x)$.

(b) We need get two more nodes so $N = 3$. Using the provided recursion formula, we find that $P_3(x) = \frac{5}{3}x^3 - \frac{3}{2}$. Thus the two additional quadrature nodes are $x_1 = -\frac{\sqrt{5}}{5}$ and $x_2 = \frac{\sqrt{5}}{5}$. The weights can be found by the standard formula

$w_j = \int_{-1}^{1} L_j(x)dx$ where $L_j(x)$ is the Lagrange polynomial associated with the j^{th} interpolation node.

Problem 4: Linear algebra

In some computational settings a block LU decomposition is useful. In this problem you will build the block LU factorization of a matrix and determine the computational complexity of using such a technique for solving a linear system.

(a) Consider the $2n \times 2n$ block matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where each block is an $n \times n$ matrix. Derive the matrices $\hat{\mathbf{L}}_{21}$ and $\hat{\mathbf{A}}_{22}$ such that

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \hat{\mathbf{L}}_{21} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \hat{\mathbf{A}}_{22} \end{bmatrix}.$$

(b) What is the computational cost in the big O sense for constructing the factorization and solving a linear system $\mathbf{Ax} = \mathbf{b}$ with the precomputed factorization? The answer should be in terms of the number of blocks and the size of the blocks. Provide justification for your answer.

Note: you should assume that any inverse created while making the factorization is available for the solve stage.

(c) Building off your work in part (a), derive the formula for a $3n \times 3n$ block LU factorization of a matrix $\mathbf{A}$ where each of blocks is of size $n \times n$. This means that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{bmatrix}$$

where each bock is an $n \times n$ matrix. *Hint: Blocking will be helpful.*

(d) What is the computational cost in the big O sense for constructing the factorization and solving a linear system $\mathbf{Ax} = \mathbf{b}$ with the precomputed factorization in part (c)? The answer should be in terms of the number of blocks and the size of the blocks. Provide justification for your answer.

Note: you should assume that any inverse created while making the factorization is available for the solve stage.

Solution:

(a) $\hat{\mathbf{L}}_{21} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$ and $\hat{\mathbf{A}}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$.

(b) Constructing the LU factorization involves one $n \times n$ matrix inverse, two $n \times n$ matrix products and one matrix add. These have $O(n^3)$, $O(2n^3)$ and $O(n^2)$ cost respectively. Thus the total cost is $O(3n^3 + n^2)$.

The solve stage involves 2 steps: First, you must solve

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \hat{\mathbf{A}}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}$$

The cost of inverting $\hat{\mathbf{A}}_{22}$ is $O(n^3)$. This will give $\mathbf{v}_2$ which can be plugged into the first row equation. Then

$$\mathbf{v}_1 = \mathbf{A}_{11}^{-1}(\mathbf{b}_1 - \mathbf{A}_{12}\mathbf{v}_2)$$

The cost of applying $\mathbf{A}_{12}$ and $\mathbf{A}_{11}^{-1}$ is $O(2n^2)$. Since the matrix $\mathbf{A}_{11}^{-1}$ was computed in the building stage of the LU factorization, there is no additional cost associated with it.

The cost of solving $\mathbf{L}\mathbf{x} = \mathbf{v}$ is $O(n^2)$ since it only requires applying $\hat{\mathbf{L}}_{21}$ to a vector and adding two vectors.

(c) Following the hint we chose to block the matrix as follows.

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \tilde{\mathbf{A}}_{12} \\ \tilde{\mathbf{A}}_{21} & \tilde{\mathbf{A}}_{22} \end{bmatrix}$$

where $\tilde{\mathbf{A}}_{12} = [\mathbf{A}_{12}\mathbf{A}_{13}]$, $\tilde{\mathbf{A}}_{21} = \begin{bmatrix} \mathbf{A}_{21} \\ \mathbf{A}_{31} \end{bmatrix}$ and $\tilde{\mathbf{A}}_{22} = \begin{bmatrix} \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{A}_{32} & \mathbf{A}_{33} \end{bmatrix}$.

From part (a), we know that the following LU factorization holds.

$$\begin{bmatrix} \mathbf{A}_{11} & \tilde{\mathbf{A}}_{12} \\ \tilde{\mathbf{A}}_{21} & \tilde{\mathbf{A}}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 2n} \\ \hat{\mathbf{L}}_{21} & \mathbf{I}_{2n} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \tilde{\mathbf{A}}_{12} \\ \mathbf{0}_n & \hat{\mathbf{A}}_2 \end{bmatrix}$$

where $\hat{\mathbf{L}}_{21} = \tilde{\mathbf{A}}_{21}\mathbf{A}_{11}^{-1}$ and $\hat{\mathbf{A}}_2 = \tilde{\mathbf{A}}_{22} - \tilde{\mathbf{A}}_{21}\mathbf{A}_{11}^{-1}\tilde{\mathbf{A}}_{12}$.

Now we will apply the block LU factorization to $\hat{\mathbf{A}}_{22}$. First begin by blocking the matrix into its original form.

$$\hat{\mathbf{A}}_2 = \begin{bmatrix} \hat{\mathbf{A}}_{22} & \hat{\mathbf{A}}_{23} \\ \hat{\mathbf{A}}_{32} & \hat{\mathbf{A}}_{33} \end{bmatrix}$$

Then again by part (a), the LU factorization of $\hat{\mathbf{A}}_2$ is

$$\hat{\mathbf{A}}_2 = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_n \\ \hat{\mathbf{L}}_{32} & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_{22} & \tilde{\mathbf{A}}_{23} \\ \mathbf{0}_n & \hat{\mathbf{A}}_3 \end{bmatrix}$$

where $\hat{\mathbf{L}}_{32} = \tilde{\mathbf{A}}_{23}\tilde{\mathbf{A}}_{22}^{-1}$ and $\hat{\mathbf{A}}_3 = \hat{\mathbf{A}}_{33} - \tilde{\mathbf{A}}_{32}\tilde{\mathbf{A}}_{22}^{-1}\tilde{\mathbf{A}}_{23}$.

Thus the full LU factorization of $\mathbf{A}$ has the following factors.

$$\mathbf{L} = \left[\begin{array}{ccc} \mathbf{I}_n & \mathbf{0}_n & \mathbf{0} \\ \hat{\mathbf{L}}_{21} & \mathbf{I}_n & \mathbf{0}_n \\ & \hat{\mathbf{L}}_{32} & \mathbf{I}_n \end{array}\right]$$

$$\mathbf{U} = \left[\begin{array}{ccc} \mathbf{A}_{11} & \multicolumn{2}{c}{\tilde{\mathbf{A}}_{21}} \\ \mathbf{0}_n & \tilde{\mathbf{A}}_{22} & \tilde{\mathbf{A}}_{23} \\ \mathbf{0}_n & \mathbf{0}_n & \hat{\mathbf{A}}_3 \end{array}\right]$$

(d) To count the cost of constructing the LU factorization, we count the cost of each of the steps. The cost for zeroing out the first block column is $O(5n^3 + 4n^2)$. (One $O(n^3)$ inversion plus a matrix-multiply involving an $n \times n$ matrix with an $n \times 2n$ matrix that matrix is then left multiplied by a $2n \times n$ matrix. Finally there is matrix addition of $2n \times 2n$ matrices.) The cost of processing the lower right block of the matrix is the same as in part (b) of this problem $O(3n^3 + n^2)$.

Again the solve is two steps: (i) solve $\mathbf{Uv} = \mathbf{b}$ and solve $\mathbf{Lx} = \mathbf{v}$.

The first system that is solve is

$$\begin{bmatrix} \mathbf{A}_{11} & \tilde{\mathbf{A}}_{21} & \\ \mathbf{0}_n & \tilde{\mathbf{A}}_{22} & \tilde{\mathbf{A}}_{23} \\ \mathbf{0}_n & \mathbf{0}_n & \hat{\mathbf{A}}_3 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix}$$

The cost of solving for $\mathbf{v}_3$ is $O(n^3)$. The cost of solving for $\mathbf{v}_2$ is $O(2n^2)$ since it requires to matrix vector multiplies of $n \times n$ matrices. The cost of solving for $\mathbf{v}_1$ is $O(3n^2)$ since it requires 3 matvecs of $n \times n$ matrices. Thus the cost of solving for $\mathbf{v}$ is $O(n^3 + (b + (b-1))n^2)$ where b is the number of blocks in the matrix.

The cost of solving $\mathbf{Lx} = \mathbf{v}$ is $O(3n^2)$. The cost of geting $\mathbf{x}_1$ is free. The cost of solving for $\mathbf{x}_2$ is 1 matvec plus a vector add thus $O(n^2)$. The cost of solving for $\mathbf{x}_3$ is two matvecs plus adding 3 vectors thus $O(2n^2)$. So the total cost in terms of the number of blocks is $O(bn^2)$.

Note that these solves are asympotically less than a ful

Problem 5: Numerical ODEs

Consider the following implicit three-step method

$$y_{n+3} - y_n = h\left[\mu f(t_{n+3}, y_{n+3}) + \frac{9}{8} f(t_{n+2}, y_{n+2}) + \frac{9}{8} f(t_{n+1}, y_{n+1}) + \frac{3}{8} f(t_n, y_n)\right]$$

with undetermined coefficient μ to be designed to numerically solve

$$y' = f(t, y), \quad y(t_0) = y_0$$

i) Determine the value of μ that makes this scheme consistent.

ii) Determine the order of the consistent scheme by looking at the truncation error.

iii) Is the consistent scheme convergent?

Solution:

i) $\mu = 3/8$ by consistency condition

ii) order 4

iii) yes

Problem 6: Numerical PDEs
Consider the initial value problem for one-dimensional wave propagation

$$\partial_{tt} u = c^2 \partial_{xx} u, \quad t \geq 0, \quad u(x,0) = f(x), \quad u_t(x,0) = g(x).$$

(a) An explicit time-stepping numerical method using central differences to discretize space and time derivatives gives

$$U(x,t+k) - 2U(x,t) + U(x,t-k) = \alpha^2 \left[U(x+h,t) - 2U(x,t) + U(x-h,t) \right]$$

where $\alpha = c(h/k)$, $k = \Delta t$ and $h = \Delta x$. Assuming $U(x,t) = \zeta^{t/k} e^{i\omega x}$, a Von Newmann analysis performed on this disctetization gives the amplification equations

$$\zeta^2 - 2\beta\zeta + 1 = 0 \quad \text{where} \quad \beta = 1 - 2\alpha^2 \sin^2\left(\frac{\omega h}{2}\right).$$

Show that the scheme is conditionally stable and establish explicitly the stability condition.

(b) Consider the following implicit time-stepping numerical method using central differences to discretize space and time derivatives

$$\begin{aligned} U(x,t+k) - 2U(x,t) + U(x,t-k) &= \frac{\alpha^2}{2} \left[U(x+h,t-k) - 2U(x,t-k) + U(x-h,t-k) \right] & (1) \\ &+ \frac{\alpha^2}{2} \left[U(x+h,t+k) - 2U(x,t+k) + U(x-h,t+k) \right] & (2) \end{aligned}$$

(i) **Find** the amplification equation. ($\beta = 1 + 2\alpha^2 \sin^2\left(\frac{\omega h}{2}\right)$ is a useful definition).

(ii) **Determine** whether the scheme is conditionally or absolutely stable.

Solution:

(a) Assume

$$U(x,t) = \zeta^{t/k} e^{\mathrm{i}\omega x}$$

To get

$$\zeta - 2 + \frac{1}{\zeta} = \alpha^2 \left(e^{i\omega h} - 2 + e^{-i\omega h} \right) \implies \begin{aligned} \zeta^2 - 2\zeta + 1 &= 2\alpha^2 \zeta \left(\cos(\omega h) - 1 \right) \\ &= -4\alpha^2 \zeta \sin^2(\omega h/2) \end{aligned}$$

Thus

$$\zeta^2 - 2(1 - 2\alpha^2 \sin^2(\omega h/2))\zeta + 1 \equiv \zeta^2 - 2\beta\zeta + 1 = 0$$

as stated in the question. Thus

$$\zeta_\pm = \beta \pm \sqrt{\beta^2 - 1}$$

if $|\beta| > 1$ then magnitude of one of the roots $|\zeta_\pm|$ is > 1 implying instability. Thus we must choose $|\beta| < 1$ implying

$$\zeta_\pm = \beta \pm i\sqrt{1 - \beta^2} \quad s.t. \quad |\zeta| = 1.$$

Thus the scheme is conditionally stable. The stability condition is

$$-1 \le \beta \le 1, \quad \beta = 1 - 2\alpha^2 \sin^2\left(\frac{\omega h}{2}\right) \qquad \Longrightarrow \alpha \le 1$$

(b) Assume

$$U(x,t) = \zeta^{t/k} e^{\mathrm{i}\omega x}$$

To get

$$\begin{aligned} \zeta - 2 + \frac{1}{\zeta} = \frac{\alpha^2}{2}\left(e^{i\omega h} - 2 + e^{-i\omega h}\right)\left(\frac{1}{\zeta} + \zeta\right) \quad \Longrightarrow \quad \zeta^2 - 2\zeta + 1 &= \alpha^2\left(\cos(\omega h) - 1\right)\left(\zeta^2 + 1\right) \\ &= -2\alpha^2 \sin^2(\omega h/2)\left(\zeta^2 + 1\right) \end{aligned}$$

Thus

$$(1 + 2\alpha^2 \sin^2(\omega h/2))\zeta^2 - 2\zeta + (1 + 2\alpha^2 \sin^2(\omega h/2)) \equiv \beta\zeta^2 - 2\zeta + \beta = 0$$

where $\beta = (1 + 2\alpha^2 \sin^2(\omega h/2))$. Thus we find $\beta \ge 1, \ \forall \omega$. Thus

$$\zeta_{\pm} = \frac{1 \pm i\sqrt{\beta^2 - 1}}{\beta}, \quad |\zeta| = 1.$$

No restriction on α, so the scheme is absolutely stable.