Numerical Analysis Preliminary Exam

9 AM to 12 Noon, January 5, 2022

Instructions. You have three hours to complete this exam. Submit solutions to four (and no more) of the following six problems. 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 on your exam; **do not write your name on your exam**.

Problem 1: Rootfinding

Consider a 2D fixed point iteration of the form

$$\begin{aligned} x_{k+1} &= f(x_k, y_k) \\ y_{k+1} &= g(x_k, y_k) \end{aligned}$$

Assume that the vector-valued function $\vec{h}(x,y) = (f(x,y), g(x,y))^T$ is continuously-differentiable, and the infinity norm of the Jacobian matrix is less than 1 at a unique fixed point (x_∞, y_∞).

Now consider the "nonlinear Gauss-Seidel" version of the iteration:

$$\begin{aligned} x_{k+1} &= f(x_k, y_k) \\ y_{k+1} &= g(x_{k+1}, y_k) \end{aligned}$$

Prove that the "nonlinear Gauss-Seidel" version is convergent, to the same fixed point, for initial conditions sufficiently close to the fixed point.

Solution: First check that it has the same fixed point: Plug $x_\infty = f(x_\infty, y_\infty), y_\infty = g(x_\infty, y_\infty)$ into the new iteration to obtain

$$\begin{aligned} f(x_\infty, y_\infty) &= x_\infty \\ g(f(x_\infty, y_\infty), y_\infty) = g(x_\infty, y_\infty) &= y_\infty \end{aligned}$$

Next, find the Jacobian of the new iteration function:

$$\mathbf{J}_h = \begin{bmatrix} \partial_1 f(x,y) & \partial_2 f(x,y) \\ (\partial_1 g(f(x,y),y))\partial_1 f(x,y) & (\partial_1 g(f(x,y),y))\partial_2 f(x,y) + \partial_2 g(f(x,y),y) \end{bmatrix}$$

The infinity norm of this Jacobian is the maximum absolute row sum. The first row has exactly the same absolute row sum as the Jacobian of the original iteration, so we know that

$$|\partial_1 f(x_\infty, y_\infty)| + |\partial_2 f(x_\infty, y_\infty)| < 1$$

The absolute row sum for the second row is

$$\begin{aligned} &|(\partial_1 g(f(x,y),y))\partial_1 f(x,y)| + |(\partial_1 g(f(x,y),y))\partial_2 f(x,y) + \partial_2 g(f(x,y),y)| \\ &\le |\partial_1 g(f(x,y),y)| \left(|\partial_1 f(x,y)| + |\partial_2 f(x,y)|\right) + |\partial_2 g(f(x,y),y)| \\ &\le |\partial_1 g(f(x,y),y)| + |\partial_2 g(f(x,y),y)| \end{aligned}$$

Evaluating at the fixed point:

$$|\partial_1 g(f(x_\infty, y_\infty), y_\infty)| + |\partial_2 g(f(x_\infty, y_\infty), y_\infty)| = |\partial_1 g(x_\infty, y_\infty)| + |\partial_2 g(x_\infty, y_\infty)| < 1$$

We've thus proven that the Jacobian of the new iteration function has infinity norm less than 1 at the fixed point. Since the new iteration function is continuously-differentiable, there must be a neighborhood of the fixed point such that iterations initialized in this neighborhood will converge.

Problem 2: Interpolation & Approximation

Let the ordinary Legendre polynomial of degree k be denoted $P_k(x)$ for $k \geq 0$. The *associated* Legendre polynomials are

$$P_k^m(x) = (-1)^m (1-x^2)^{m/2} \frac{\mathrm{d}^m}{\mathrm{d}x^m} P_k(x),\ m > 0, k \geq m.$$

(Note that despite the name, for odd m they are not actually polynomials.)

(a) Consider the *interpolation* problem of finding coefficients a_k such that

$$\sum_{k=1}^{N} a_k P_k^1(x_i) = y_i, \quad i = 1, \ldots, N.$$

Prove that this linear system of equations for the unknown coefficients a_k is nonsingular whenever the set of interpolation points x_i does not include ±1, and does not include duplicates.

(b) Consider the *approximation* problem of finding coefficients a_k to minimize the squared approximation error

$$\left\| f(x) - \sum_{k=1}^{N} a_k P_k^1(x) \right\|_2^2$$

where the L^2 norm is over $x \in [-1, 1]$. Write down a linear system for the unknown coefficients a_k and explain why it is nonsingular. You should give an explicit integral expression for the entries of the coefficient matrix and right hand side, but the expression does not need to be simplified.

(c) Let $\mathbf{M}$ be the coefficient matrix from (b). Prove that $\mathbf{M}_{k,j} = 0$ when $k + j$ is odd.

Solution: **(a)** The linear system for the unknown coefficients is

$$\begin{aligned}
\sqrt{1-x_1^2} P_1'(x_1) a_1 + \ldots + \sqrt{1-x_1^2} P_N'(x_1) a_N &= y_1 \\
\vdots &= \vdots \\
\sqrt{1-x_N^2} P_1'(x_N) a_1 + \ldots + \sqrt{1-x_N^2} P_N'(x_N) a_N &= y_N
\end{aligned}$$

In matrix form

$$\begin{bmatrix} \sqrt{1-x_1^2} P_1'(x_1) & \cdots & \sqrt{1-x_1^2} P_N'(x_1) \\ \vdots & & \vdots \\ \sqrt{1-x_N^2} P_1'(x_N) & \cdots & \sqrt{1-x_N^2} P_N'(x_N) \end{bmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

If any of the x_i are ±1 then an entire row of the coefficient matrix is 0, which makes it singular. If none of the x_i are ±1 then we can rescale the rows to arrive at the following system

$$\begin{aligned}
P_1'(x_1) a_1 + \ldots + P_N'(x_1) a_N &= \frac{y_1}{\sqrt{1-x_1^2}} \\
\vdots &= \vdots \\
P_1'(x_N) a_1 + \ldots + P_N'(x_N) a_N &= \frac{y_N}{\sqrt{1-x_N^2}}
\end{aligned}$$

This is simply a polynomial interpolation problem where we are trying to find the interpolating polynomial of degree at most $N-1$, but using the unusual basis $\{P_1', \dots, P_N'\}$. We can establish that these functions are a basis for the space of polynomials of degree at most $N-1$ by simply noting that P_i' has degree exactly $i-1$. Since we know that the polynomial interpolation problem always has a unique solution (regardless of the choice of basis) whenever the set of interpolation points does not include duplicates, we conclude that a solution of the linear system must always exist. Since it is a square system and a solution exists for any right hand side, we conclude that the coefficient matrix is nonsingular.

(b) First note that we can expand the formula for the squared approximation error as

$$\left\| f(x) - \sum_{k=1}^{N} a_k P_k^1(x) \right\|_2^2 = \left\langle f(x) - \sum_{k=1}^{N} a_k P_k^1(x), f(x) - \sum_{j=1}^{N} a_j P_j^1(x) \right\rangle$$

$$= \langle f, f\rangle - 2\sum_{k=1}^{N} a_k \langle f, P_k^1\rangle + \sum_{k=1}^{N}\sum_{j=1}^{N} a_k a_j \langle P_k^1, P_j^1\rangle$$

Computing the gradient with respect to the unknown coefficients and setting it to zero yields the linear system

$$\sum_{j=1}^{N} \langle P_k^1, P_j^1\rangle a_j = \langle f, P_k^1\rangle, \quad k = 1, \dots, N$$

The coefficient matrix is a Gram matrix (sometimes called a mass matrix); i.e. its entries are the inner products of a set of vectors/functions in a space. Gram matrices are positive definite (and thus nonsingular) whenever the set of vectors/functions is linearly independent. Linear independence is clear from the fact that P_k^1 is a polynomial of degree exactly $k-1$ multiplied by $\sqrt{1-x^2}$.

(c) By definition

$$\mathbf{M}_{kj} = \langle P_k^1, P_j^1\rangle = \int_{-1}^{1} \left(\sqrt{1-x^2}P_k'(x)\right)\left(\sqrt{1-x^2}P_j'(x)\right) \mathrm{d}x = \int_{-1}^{1} (1-x^2)P_k'(x)P_j'(x)\mathrm{d}x.$$

The Legendre polynomials P_k have odd parity when k is odd, and even parity when k is even. Thus, when $k+j$ is odd the integrand is odd, and the integral equals zero.

Problem 3: Quadrature
The quadrature formula

$$\int_0^h f(x)\mathrm{d}x \approx \frac{3h}{4} f\left(\frac{h}{3}\right) + \frac{h}{4} f(h)$$

integrates polynomials of degree ≤ 2 exactly. Derive an error bound of the form Ch^4 for this quadrature rule, where C is a constant independent of h, assuming that $f \in C^3[0,h]$. Hint: Use the Peano kernel theorem.
Solution: Letting $Q[f]$ denote the quadrature approximation, we start from

$$R[f] = Q[f] - \int_0^h f(x)\mathrm{d}x = \int_0^h f^{(3)}(t)K(t)\mathrm{d}t$$

where

$$K(t) = \frac{1}{2}R[(x-t)_+^2] = \frac{1}{2}\left[\frac{3h}{4}\left(\frac{h}{3}-t\right)_+^2 + \frac{h}{4}(h-t)_+^2 - \int_0^h (x-t)_+^2\mathrm{d}x\right].$$

$$K(t) = \frac{1}{2}R[(x-t)_+^2] = \frac{1}{2}\left[\frac{3h}{4}\left(\frac{h}{3}-t\right)_+^2 + \frac{h}{4}(h-t)^2 - \int_t^h (x-t)^2\mathrm{d}x\right].$$

$$K(t) = \frac{1}{2}R[(x-t)_+^2] = \frac{1}{2}\left[\frac{3h}{4}\left(\frac{h}{3}-t\right)_+^2 + \frac{h}{4}(h-t)^2 - \frac{(h-t)^3}{3}\right].$$

$$K(t) = \frac{1}{2}R[(x-t)_+^2] = \frac{h^3}{2}\left[\frac{3}{4}\left(\frac{1}{3}-\frac{t}{h}\right)_+^2 + \frac{1}{4}\left(1-\frac{t}{h}\right)^2 - \frac{1}{3}\left(1-\frac{t}{h}\right)^3\right].$$

The kernel is non-negative over $[0,h]$, so we could use the integral mean value theorem to arrive at

$$R[f] = \frac{f^{(3)}(\xi)}{2}\int_0^h K(t)\mathrm{d}t = \frac{f^{(3)}(\xi)}{216}h^4$$

for some $\xi \in [0,h]$. Proving that the kernel is non-negative on the interval is extra work in the context of this exam, so a simpler alternative is to bound the error

$$|R[f]| \leq \|f^{(3)}\|_\infty \int_0^h |K(t)|\mathrm{d}t \leq h^3\frac{\|f^{(3)}\|_\infty}{2}\int_0^h \left[\frac{1}{4}\left(\frac{1}{3}-\frac{t}{h}\right)_+^2 + \frac{3}{4}\left(1-\frac{t}{h}\right)^2 + \frac{1}{3}\left(1-\frac{t}{h}\right)^3\right]\mathrm{d}t$$

$$= \frac{19\|f^{(3)}\|_\infty}{216}h^4.$$

To obtain a bound of the form Ch^4 without evaluating the integrals, it suffices to note that the integrals can all be written as $h\times$(a constant) via the substitution $u = t/h$.

Problem 4: Numerical Linear Algebra
For all parts of this problem, assume all matrices and vectors are real.

(a) Write down the steepest descent method for solving $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is symmetric and positive definite.

(b) Explain how the formulas from (a) can break down if $\mathbf{A}$ is symmetric, but only non-negative definite (also called positive semi-definite).

(c) Suppose that $\mathbf{A}$ is symmetric but only non-negative definite. Show that if $\mathbf{b}$ is in the range of $\mathbf{A}$, then the steepest-descent method will still converge to a solution. You may assume that the steepest-descent method converges whenever $\mathbf{A}$ is symmetric and positive definite.

Solution: **(a)** The steepest-descent iteration is

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{r}_k, \quad \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \quad \alpha_k = \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{r}_k^T \mathbf{A} \mathbf{r}_k}$$

(b) If $\mathbf{A}$ has a nontrivial null space and $\mathbf{r}_k$ lies in that null space then the denominator in the definition of α_k could be 0.

(c) Since $\mathbf{A}$ is symmetric, the range (column space) equals the co-range (row space), and the kernel (null space) equals the co-kernel (left hand null space). Any vector can be uniquely decomposed into components in the range and kernel, and these components can be obtained by orthogonal projections onto the associated subspaces. Let $\boldsymbol{x}_k = \boldsymbol{w}_k + \boldsymbol{z}_k$ where $\boldsymbol{w}_k$ is in the range of $\mathbf{A}$ and $\boldsymbol{z}_k$ is in the kernel of $\mathbf{A}$. Since $\boldsymbol{b}$ is in the range, the residual $\boldsymbol{r}_k$ is always in the range as well. Projecting the iteration onto the range yields

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \alpha_k \boldsymbol{r}_k.$$

Projecting onto the kernel yields

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k.$$

Note that since $\boldsymbol{r}_k$ is in the range of $\mathbf{A}$, we have that $\boldsymbol{r}_k^T \mathbf{A} \boldsymbol{r}_k$ is only zero if $\boldsymbol{r}_k = \mathbf{0}$.

This only shows that the iteration will not break down, which is the easy part. To see that it converges, note that the range is an invariant subspace of $\mathbf{A}$. Choose any basis for the range of $\mathbf{A}$ and then let the vector $\hat{\boldsymbol{w}}_k$ contain the coordinates of $\boldsymbol{w}_k$ in this coordinate system. Let $\hat{\mathbf{A}}$ be the representation of $\mathbf{A}$ acting on this subspace using this basis. Then $\hat{\mathbf{A}}$ is symmetric and positive definite, and $\hat{\boldsymbol{w}}_k$ obeys the steepest-descent iteration on this subspace. Since $\hat{\mathbf{A}}$ is symmetric and positive definite, the sequence $\hat{\boldsymbol{w}}_k$ converges, which implies that $\boldsymbol{w}_k$ also converges.

Problem 5: Ordinary Differential Equations
Consider a system of two ODEs of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x,y), \quad \frac{\mathrm{d}y}{\mathrm{d}t} = g(x,y).$$

Suppose that it is more computationally expensive to evaluate g than to evaluate f.

(a) Prove that the multi-rate explicit Euler method defined by

$$\begin{aligned} x_{k+1/2} &= x_k + \frac{\Delta}{2} f(x_k, y_k) \\ x_{k+1} &= x_{k+1/2} + \frac{\Delta}{2} f(x_{k+1/2}, y_k) \\ y_{k+1} &= y_k + \Delta g(x_k, y_k) \end{aligned}$$

is locally second order, where Δ is the size of the time step. (Only consider the order at integer time subscripts.)

(b) Consider applying the method from (a) to the following linear problem:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -x + y, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = -y.$$

Under what conditions on the time step $\Delta > 0$ will the discrete solution remain stable, i.e. satisfying $\lim_{k\to\infty} x_k = \lim_{k\to\infty} y_k = 0$ for any initial condition?

Solution: (a) The proof for the y variable is straightforward:

$$y(t_k + \Delta) = y(t_k) + \Delta y'(t_k) + \mathcal{O}(\Delta^2) = y_k + \Delta g(x_k, y_k) + \mathcal{O}(\Delta^2) = y_{k+1} + \mathcal{O}(\Delta^2).$$

For the x variable we proceed in several steps. First use Taylor series to get an expression for the error at $t_k + \Delta/2$:

$$x\left(t_k + \frac{\Delta}{2}\right) = x(t_k) + \frac{\Delta}{2} x'(t_k) + \mathcal{O}(\Delta^2) = x_k + \frac{\Delta}{2} f(x_k, y_k) + \mathcal{O}(\Delta^2) = x_{k+1/2} + \mathcal{O}(\Delta^2)$$

Next Taylor expand about $t_k + \Delta/2$

$$x(t_k + \Delta) = x\left(t_k + \frac{\Delta}{2}\right) + \frac{\Delta}{2} x'\left(t_k + \frac{\Delta}{2}\right) + \mathcal{O}(\Delta^2)$$

Next plug in the approximation $x_{k+1/2}$ and the definition of x'

$$x(t_k + \Delta) = x_{k+1/2} + \frac{\Delta}{2} f\left(x\left(t_k + \frac{\Delta}{2}\right), y\left(t_k + \frac{\Delta}{2}\right)\right) + \mathcal{O}(\Delta^2)$$

Note the simple Taylor approximation

$$y\left(t_k + \frac{\Delta}{2}\right) = y(t_k) + \mathcal{O}(\Delta) = y_k + \mathcal{O}(\Delta)$$

Plug this in, as well as the approximation $x_{k+1/2}$:

$$x(t_k + \Delta) = x_{k+1/2} + \frac{\Delta}{2} f\left(x_{k+1/2} + \mathcal{O}(\Delta^2), y_k + \mathcal{O}(\Delta)\right) + \mathcal{O}(\Delta^2)$$

Now Taylor expand f

$$f\left(x_{k+1/2}+\mathcal{O}(\Delta^2), y_k+\mathcal{O}(\Delta)\right) = f(x_{k+1/2}, y_k) + (\partial_x f)\,\mathcal{O}(\Delta^2) + (\partial_y f)\,\mathcal{O}(\Delta) + \mathcal{O}(\Delta^2)$$

Plugging this back in above yields

$$x(t_k+\Delta) = x_{k+1/2} + \frac{\Delta}{2} f\left(x_{k+1/2}, y_k\right) + \mathcal{O}(\Delta^2) = x_{k+1} + \mathcal{O}(\Delta^2)$$

This proves that the scheme is locally second order.

Note that you can also start from

$$x_{k+1} = x_k + \frac{\Delta}{2} f(x_k, y_k) + \frac{\Delta}{2} f\left(x_k + \frac{\Delta}{2} f(x_k, y_k), y_k\right).$$

If you then Taylor expand $x(t_k+\Delta)$ around $x(t_k)=x_k$ and Taylor expand the second f around x_k then you arrive at the same result.

(b) Applying the method to the specified linear problem yields the following linear iteration:

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{bmatrix} \left(1-\frac{\Delta}{2}\right)^2 & \Delta\left(1-\frac{\Delta}{4}\right) \\ 0 & 1-\Delta \end{bmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix}.$$

The iteration is stable when the eigenvalues are within the unit circle. Since it is a triangular matrix, the eigenvalues are obvious

$$\left(1-\frac{\Delta}{2}\right)^2, \text{ and } 1-\Delta.$$

Both eigenvalues lie within the unit circle when

$$0 < \Delta < 2.$$

This problem is inherently a system and cannot be solved using the usual scalar methods for stability analysis of numerical methods for ODEs.

Problem 6: Partial Differential Equations
Consider following discretization

$$u_j^{n+1} = u_j^n + \frac{\kappa \Delta t}{\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n)$$

of the heat equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}$$

on a periodic 1D domain. Assume that the spatial grid is equispaced with size Δx.

(a) Show that the discretization is first-order accurate in time and second-order accurate in space.

(b) Show that if the time step is chosen to be $\Delta t = \Delta x^2/(6\kappa)$, then the discretization becomes second-order accurate in time.

Solution: **(a)** Re-arrange the discretization as follows

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \kappa \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

First Taylor expand in time

$$u(x, t + \Delta t) = u(x, t) + \Delta t \partial_t u + \frac{\Delta t^2}{2} \partial_t^2 u + \mathcal{O}(\Delta t^3)$$

Re-arrange to

$$\frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} = \partial_t u + \frac{\Delta t}{2} \partial_t^2 u + \mathcal{O}(\Delta t^2).$$

Now Taylor expand in space

$$u(x + \Delta x, t) = u(x, t) + \Delta x \partial_x u + \frac{\Delta x^2}{2} \partial_x^2 u + \frac{\Delta x^3}{6} \partial_x^3 u + \frac{\Delta x^4}{4!} \partial_x^4 u + \frac{\Delta x^5}{5!} \partial_x^5 u + \mathcal{O}(\Delta x^6)$$

$$u(x - \Delta x, t) = u(x, t) - \Delta x \partial_x u + \frac{\Delta x^2}{2} \partial_x^2 u - \frac{\Delta x^3}{6} \partial_x^3 u + \frac{\Delta x^4}{4!} \partial_x^4 u - \frac{\Delta x^5}{5!} \partial_x^5 u + \mathcal{O}(\Delta x^6)$$

Soving for the centered difference yields

$$\frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{\Delta x^2} = \partial_x^2 u + \frac{\Delta x^2}{12} \partial_x^4 u + \mathcal{O}(\Delta x^4)$$

Plugging these expansions into the discretization yields

$$\partial_t u + \frac{\Delta t}{2} \partial_t^2 u + \mathcal{O}(\Delta t^2) = \kappa \partial_x^2 u + \frac{\kappa \Delta x^2}{12} \partial_x^4 u + \mathcal{O}(\Delta x^4).$$

The leading order terms are $\mathcal{O}(\Delta t)$ and $\mathcal{O}(\Delta x^2)$, so the discretization satisfies the PDE to first order in time and second order in space.

(b) Note that the PDE implies

$$\partial_t^2 u = \partial_t(\kappa \partial_x^2 u) = \kappa^2 \partial_x^4 u.$$

Plugging this in to the Taylor expansion from (a) yields

$$\partial_t u + \frac{\kappa^2 \Delta t}{2} \partial_x^4 u + \mathcal{O}(\Delta t^2) = \kappa \partial_x^2 u + \frac{\kappa \Delta x^2}{12} \partial_x^4 u + \mathcal{O}(\Delta x^4).$$

If the time step is set to $\Delta t = \Delta x^2/(6\kappa)$ then the leading-order truncation error terms in time and space cancel, leaving $\mathcal{O}(\Delta t^2)$. It is crucial to note that the next term in space is $\mathcal{O}(\Delta x^4)$, not $\mathcal{O}(\Delta x^3)$. If it had been the latter, then since $\Delta x^3 \sim \Delta t^{3/2}$ the discretization would only have had sesquilinear accuracy in time.