Department of Applied Mathematics
Preliminary Examination in Numerical Analysis
January, 2020

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 number (not your name!) on your exam.

1. **Root Finding.** Consider the equation $e^x = \sin x$.

   (a) Show that there is a solution $x^* \in (-\frac{5}{4}\pi, -\pi)$.

   (b) Consider the following iterative methods (i) $x_{k+1} = \ln(\sin x_k)$ and (ii) $x_{k+1} = \arcsin(e^{x_k})$. What can you say about the local convergence of each of the methods for x^* as in (a) and their convergence order? If you use a theorem give its precise statement.

   (c) For x^* as in (a) give a method that is quadratically convergent. Justify why the method is quadratically convergent.

   Solution:

   (a) For $x < 0$. $1/(1-x) > e^x$ and for all x $e^x > 0$. Since

   $$\sin(-5\pi/4) = 1/\sqrt{2} > 1/(1+5\pi/4) > e^{-\frac{5\pi}{4}},$$

   and $\sin(-\pi) = 0$ and since the functions are continuous the intermediate value theorem guarantees the existence of a root.

   (b) Writing the iterations as $x_{k+1} = g(x_k)$ we have that for the first iteration $|g'(x)| = |\frac{\cos(x)}{\sin(x)}| > 1$ for $x \in (-\frac{5}{4}\pi, -\pi)$ so there is no convergence to the root inside that interval. For the second iteration we have that

   $$|g'(x)| = |\frac{e^x}{\sqrt{1-e^{2x}}}| < 1,$$

   on the interval under consideration. However as arcsin is only defined for $x \in [-\pi/2, \pi/2]$ the iteration cannot find this root. Note that there is no root to the equation inside $[-\pi/2, \pi/2]$.

(c) Use Newton's method on $f(x) = e^x - \sin(x)$. Use arguments similar to those in (a) to argue that $f'(x) \neq 0$ for $x \in [-\pi/2, \pi/2]$.

2. **Linear Alegbra.**

(a) Let $\mathbf{A}$ be a real $n \times n$ matrix with distinct eigenvalues such that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n| \geq 0$$

with corresponding eigenvectors $\{\mathbf{v}_j\}_{j=1}^n$.
The power iteration is given by $\mathbf{z}_m = \sigma_m \frac{\mathbf{A}^m \mathbf{z}_0}{\|\mathbf{A}^m \mathbf{z}_0\|_\infty}$ where $\sigma_m = \pm 1$. Prove that the power iteration converges to $(\pm 1)\frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_\infty}$.

(b) Consider the Eudoxos iteration with initial guess $x_0 = y_0 = 1$ given by

$$\begin{aligned} x_{n+1} &= x_n + y_n \\ y_{n+1} &= x_{n+1} + x_n. \end{aligned}$$

Rewrite the iteration as a linear system iteration. In other words, rewrite the iteration as $\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} x_n \\ y_n \end{bmatrix}$.

(c) Using the result from the power iteration, prove that the ratio $\frac{y_n}{x_n}$ converges to $\sqrt{2}$.

Solution:

(a) The power iteration has m^{th} iterate

$$\begin{aligned} \mathbf{A}^m \mathbf{z}_0 &= \sum_{j=1}^n \alpha_j \mathbf{A}^m \mathbf{v}_j \\ &= \sum_{j=1}^n \alpha_j \lambda_j^m \mathbf{v}_j \\ &= \lambda_1^m \left[\alpha_1 \mathbf{v}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^m \mathbf{v}_j \right] \end{aligned}$$

Note that as $m \to \infty$, $\left(\frac{\lambda_j}{\lambda_1}\right)^m \to 0$ for $2 \leq j \leq n$.
Then $\|\mathbf{A}^m \mathbf{z}_0\|_\infty \sim |\lambda_1|^m |\alpha_1| \|\mathbf{v}_1\|_\infty$. So

$$\begin{aligned} \lim_{m\to\infty} \frac{\mathbf{A}^m \mathbf{z}_0}{\|\mathbf{A}^m \mathbf{z}_0\|_\infty} &= \lim_{m\to\infty} \left(\frac{\lambda_1}{|\lambda_1|} \right)^m \frac{\alpha_1 \mathbf{v}_1}{|\alpha_1| \|\mathbf{v}_1\|_\infty} \\ &= \pm 1 \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_\infty} \end{aligned}$$

(b) We need to rewrite y_{n+1} so that it does not involve x_{n+1}. We do this by simply plugging in the definition of x_{n+1} to find $y_{n+1} = 2x_n + y_n$.
Then the linear system iteration is

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix}.$$

(c) We know that the power iteration converges to the eigenvector corresponding to the largest (in magnitude) eigenvalue. For our system, the largest eigenvalue is $1+\sqrt{2}$ with corresponding eigenvector $\mathbf{v}_1 = \pm \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}$. Thus $\frac{y_n}{x_n} \to \sqrt{2}$.

3. **Numerical quadrature.** Consider the task of numerically approximating

$$I(f) = \int_a^b f(x)dx,$$

where $f \in C^\infty[a,b]$.

(a) Derive the trapezoidal rule and corresponding error for approximating $I(f)$.
Useful information: $\int_a^b (x-a)(x-b)dx = -\frac{1}{6}(b-a)^3$.

(b) Find the formula for the composite trapezoidal rule using uniform intervals of size $h = \frac{b-a}{n}$ where $n+1$ is the number of quadrature points. i.e. the quadrature points are $x_j = a + jh$ for $j = 0, \ldots, n$.

(c) Derive the error for the composite trapezoidal rule.

Solution:

(a) The trapezoidal rule is based on integrating the linear interpolation with interpolation points $x_0 = a$ and $x_1 = b$. Using this information, we use Taylor's Theorem with a modified Lagrange remainder term to rewrite $f(x)$ as

$$f(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} + \frac{f''(\eta_x)}{2}(x-a)(x-b)$$

for some $\eta_x \in [a,b]$.
Integrating the linear approximation we find that the trapezoidal rule is given by

$$I_1(f) = \int_a^b \left[f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} \right] = \frac{f(a)+f(b)}{2}(b-a).$$

The error in approximating the integral using the trapezoidal rule has an upper bound given by

$$\begin{aligned} E_1(f) = I(f) - I_1(f) &= -\int_a^b \frac{f''(\eta_x)}{2}(x-a)(x-b)dx \\ &= -f''(\eta)\frac{(b-a)^3}{12} \end{aligned}$$

for some $\eta \in [a,b]$ by the mean value theorem. (See for example Chapter 1 Thm 1.3 of Atkinson Numerical Analysis text.) The trapezoidal rule is given by $I_1(f) = \frac{f(a)+f(b)}{2}(b-a)$ and the error term is $E_1(f) = -f''(\eta)\frac{(b-a)^3}{12}$.

(b) $I_n(f) = h\left[\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right]$

(c)

$$E_n(f) \le I(f) - I_n(f) = \frac{-h^3}{12}\sum_{j=1}^{n} f''(\eta_j)$$
$$= \frac{-h^3 n}{12}\left(\frac{1}{n}\sum_{j=1}^{n} f''(\eta_j)\right)$$
$$= \frac{-h^3 n(b-a)}{12(b-a)}\left(\frac{1}{n}\sum_{j=1}^{n} f''(\eta_j)\right)$$

We know that

$$\min_{x\in[a,b]} f''(x) \le 1/n\sum_{j=1}^{n} f''(\eta_j) \le \max_{x\in[a,b]} f''(x).$$

Since $f''(x)$ is continuous in $[a,b]$, there exists an $\eta \in [a,b]$ such that

$$1/n\sum_{j=1}^{n} f''(\eta_j) = f''(\eta).$$

Thus $E_n(f) = \frac{-h^2(b-a)}{12} f''(\eta)$ for some $\eta \in [a,b]$.

4. **Interpolation/Approximation.** Consider the Hermite problem of constructing a polynomial $p(x)$ of degree $\le$ 3 such that

$$p(x_1) = y(x_1) \quad p'(x_1) = y'(x_1) \quad p(x_2) = y(x_2) \quad p'(x_2) = y'(x_2).$$

(a) Derive a Lagrange type formula for $p(x)$. (*Hint:* For the basis functions satisfying $l(x_2) = l'(x_2) = 0$, use $l(x) = (x-x_2)^2 g(x)$, where $g(x)$ is a polynomial of degree ≤ 1. Find $g(x)$.)

(b) Derive an error formula.

(c) Prove that the interpolation is unique.

Solution:

(a) Our goal is write $p(x) = y_1 l_1(x) + y_2 l_2(x) + y_1' l_3(x) + y_2' l_4(x)$ We need to create 4 basis functions. First we will create $l_1(x)$ which satisfies $l_1(x_1) = 1$, $l_1'(x_1) = 0$, $l_1'(x_2) = 0$ and $l_1(x_2) = 0$.
Utilizing the hint, we write $l_1(x) = (x-x_2)^2(ax+b)$ where a and b are constants to be determined. This choice of representing l_1 already satisfies the last two conditions.

So we need to chose a and b so that l_1 satisfies the first two conditions. After some algebra you find $a = \frac{2}{(x_1-x_2)^3}$ and $b = \frac{-(x_1+x_2)}{(x_1-x_2)^2}$. Thus $l_1(x) = \frac{(x-x_2)^2}{(x_1-x_2)^3}(2x-(x_1+x_2))$. Through the same process, we know $l_2(x) = \frac{(x-x_1)^2}{(x_2-x_1)^3}(2x-(x_1+x_2))$. Now we create $l_3(x)$ which satisifies $l_3(x_1) = 0$, $l_3'(x_1) = 1$, $l_3(x_2) = 0$, and $l_3'(x_2) = 0$. As for the previous basis function, we set $l_3(x) = (x-x_2)^2(ax+b)$. The constants a and b which make $l_3(x)$ satisfy all the conditions are $a = \frac{1}{(x_1-x_2)^2}$ and $b = \frac{-x_1}{(x_1-x_2)^2}$. Thus $l_3(x) = \frac{(x-x_2)^2}{(x_1-x_2)^2}(x-x_1)$. Through the same process, $l_4(x) = \frac{(x-x_1)^2}{(x_2-x_1)^2}(x-x_2)$.

(b) The Taylor remainder theorem gives the error formula

$$E = \frac{f^{(4)}(\eta)}{4!}(x-x_1)^2(x-x_2)^2.$$

for some $\eta \in (x_1, x_2)$.

(c) Suppose there are two distinct polynomials $p(x)$ and $q(x)$ of degree ≤ 3 that satisfy the four conditions. Let $w(x) = p(x) - q(x)$. Then $w(x)$ is also a polynomial of degree ≤ 3 and we know $w(x_1) = w(x_2) = w'(x_1) = w'(x_2) = 0$. This means that x_1 and x_2 are double roots of $w(x)$. The only way a polynomial of degree ≤ 3 can have more than 3 roots is if it is the zero function; i.e. $w(x) = 0$. Thus $p(x) = q(x)$ which contracts the assumption that there are distinct interpolation polynomials.

5. **ODEs** Consider the one-step method applied to IVP $y' = f(t, y)$,

$$y_{n+1} = y_n + \alpha h f(t_n, y_n) + \beta h f(t_n + \gamma h, y_n + \gamma h f(t_n, y_n))$$

where α, β, γ are real parameters.

(a) Prove that the method is consistent if and only if $\alpha+\beta = 1$, and the order of the method cannot exceed 2.

(b) Suppose that a second-order method of the above form is applied to the initial value problem $y' = -\lambda y, y(0) = 1$, where λ is a positive real number. Show that the sequence $(y_n)_{n\geq 0}$ is bounded if and only if $h \leq \frac{2}{\lambda}$. Show further that for such h,

$$|y(t_n) - y_n| \leq \frac{1}{6}\lambda^3 h^2 t_n, \quad n \geq 0.$$

Solution (a)
Expand

$$\begin{aligned} y_{n+1} &= y_n + \alpha hf + \beta h[f + \gamma h f_t + \gamma h f f_y + \frac{1}{2}\left(\gamma^2 h^2 f_{tt} + 2\gamma^2 h^2 f f_{ty} + \gamma^2 h^2 f^2 f_{yy} + \ldots\right) \\ &= y_n + (\alpha+\beta)hf + h^2\beta\gamma(f_t + f f_y) + h^3\frac{\beta\gamma^2}{2}[f_{tt} + 2f f_{ty} + f^2 f_{yy}] + \ldots. \end{aligned}$$

The exact solution satisfies

$$y(t_{n+1}) = y_n + hf + \frac{h^2}{2}[f_t + f f_y] + \frac{h^3}{6}(f_{tt} + 2f_{ty}f + f_y f_t + f_{yy}f^2 + f_y^2 f).$$

Consistency requires $\alpha + \beta = 1$ and second order accuracy $2\beta\gamma = 1$. Third order cannot be achieved since the higher order terms do not match.

Solution (b)
Plug in the right hand side to find

$$y_{n+1} = y_n - \alpha h\lambda y_n + \beta h(-\lambda(y_n - \lambda\gamma h y_n)) = (1 - (\alpha+\beta)h\lambda + \beta\gamma h^2\lambda^2)y_n.$$

We know that for second order we require $\alpha + \beta = 1$ and $2\beta\gamma = 1$ so that in order for the sequence to be bounded

$$|1 - z + \frac{z^2}{2}| \leq 1, \quad z = h\lambda.$$

Thus

$$-1 \leq 1 - z + \frac{z^2}{2} \leq 1.$$

or

$$\frac{z^2}{2} - z \leq 0 \Rightarrow z \geq 0, \quad z \leq 2.$$

That is $h \leq \frac{2}{\lambda}$.

To find the error estimate note that $y(t) = e^{-\lambda t}$ so that

$$y(t_n) - y_n = e^{-\lambda t_n} - (1 - h\lambda + \frac{h^2\lambda^2}{2})^n = (e^{-\lambda h})^n - (1 - h\lambda + \frac{h^2\lambda^2}{2})^n.$$

Recall that

$$x^n - y^n = (x - y)\sum_{i=1}^{n} x^{n-i}y^{i-1},$$

so that

$$\begin{aligned} |y(t_n) - y_n| &= \left| \left(e^{-\lambda h} - (1 - h\lambda + \frac{h^2\lambda^2}{2})\right) \sum_{i=1}^{n} \underbrace{(e^{-\lambda h})^{n-1}(1 - h\lambda + \frac{h^2\lambda^2}{2})^{i-1}}_{\leq 1} \right| \\ &\leq \frac{h^3\lambda^3}{6} n = \frac{h^2\lambda^3}{6} t_n. \end{aligned}$$

6. **PDEs** Consider the approximation $v_j \approx u(x_j, t)$ on the grid $x_j = jh, h = 2\pi/(N+1)$, $j = 1, \ldots, N+1$, to the solution of the advection equation $u_t + u_x = 0$ on the 2π-periodic domain $0 \leq x \leq 2\pi$ and with initial data $u(x, 0) = \exp(-(x - \pi)^2)$.

   Define difference operators acting on a grid function v_i as

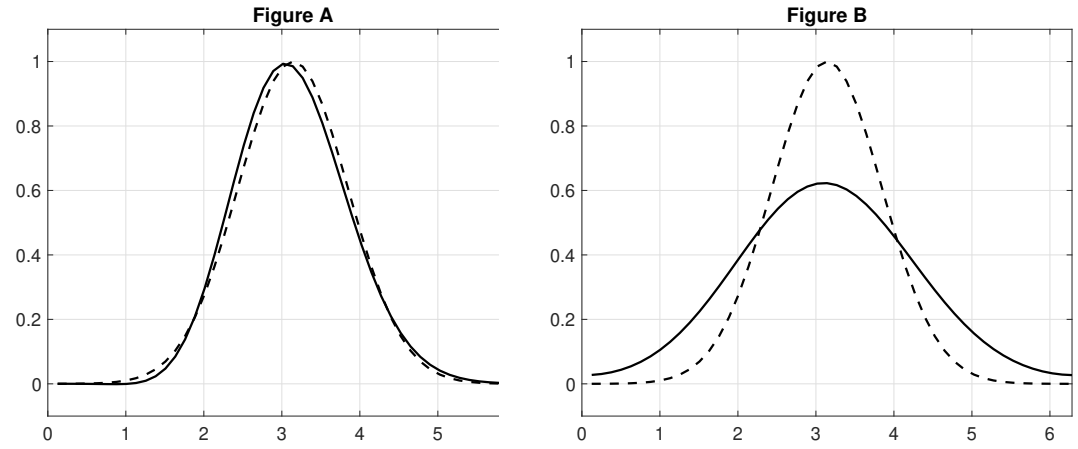
$$D_- v_j = \frac{v_j - v_{j-1}}{h}, \quad D_+ v_j = \frac{v_{j+1} - v_j}{h}, \quad D_0 v_j = \frac{v_{j+1} - v_{j-1}}{2h}.$$

   Further define the inner product $(v, w)_h = \sum_{i=1}^{n} h v_i w_i$ and the norm $\|v\|_h^2 = (v, v)_h$.

   (a) Two of the three semi-discretizations

$$(1): \frac{dv_j}{dt} + D_- v_j = 0, \quad (2): \frac{dv_j}{dt} + D_+ v_j = 0, \quad (3): \frac{dv_j}{dt} + D_0 v_j = 0,$$

   are stable and produce the results in the figures below when evolved one period in time using the classic fourth order Runge-Kutta method. Which method is not stable? For the remaining two methods, what method goes with what figure? Clearly motivate your answer.



Solid lines represents the numerical solution and dashed lines the exact solution.

(b) Let D denote one of the difference operators above. Then if we discretize in time using the trapezoidal rule we have (the superscript now denotes the time index)

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} + D\left(\frac{v_j^{n+1} + v_j^n}{2}\right) = 0.$$

Show that with this timestepping the spatial discretization corresponding to "Figure A" satisfies $\|v^{n+1}\|_h^2 = \|v^n\|_h^2$ while the discretiztion corresponding to "Figure B" satisfies $\|v^{n+1}\|_h^2 \leq \|v^n\|_h^2$. Hint: First find α_+ and / or α_- such that $D_\pm v_j = D_0 v_j + \alpha_\pm D_+ D_- v_j$.

Solution (a):
The continuous problem can be treated by Fourier series. Assume that the expansion of the initial data is

$$u(x,0) = \sum_{k=-\infty}^{\infty} \hat{u}_k e^{ikx},$$

Then each mode solves the ordinary differential equation

$$\frac{d\hat{u}_k}{dt} + \lambda_k \hat{u}_k = 0,$$

with $\lambda_k = ik$ being purely imaginary indicating that there is no decay but only translation of the initial data.
Now, as the discrete problem is periodic the complex exponential basis e^{ikjh} diagonalizes the semi discrete problems. The three operators D_-, D_+ and D_0 thus satisfy

$$hD_- e^{ikx} = (1 - e^{ikh})e^{ikx}, \quad hD_+ e^{ikx} = (e^{ikh} - 1)e^{ikx}, \quad hD_0 e^{ikx} = i\sin(kh)e^{ikx},$$

and the discrete modes satisfy

$$\frac{d\hat{u}_k}{dt} + \frac{(1 - e^{ikh})}{h}\hat{u}_k = 0, \quad \frac{d\hat{u}_k}{dt} + \frac{(e^{ikh} - 1)}{h}\hat{u}_k = 0, \quad \frac{d\hat{u}_k}{dt} + \frac{i\sin(kh)}{h}\hat{u}_k = 0.$$

Since the "λ's" will lie in the right half plane, the left half plane and on the imaginary axis, respectively the first and the last (denoted by (1) and (3)) schemes will be stable. The first scheme will be dissipative and produces the small amplitude solution in Figure B. The last (centered) scheme has errors that are purely dispersive and belongs to Figure A.

Solution (b):
Multiply by $v_i^{n+1} + v_i^n$ and sum to find

$$\|v^{n+1}\|_h^2 - \|v^n\|_h^2 + \frac{\Delta t}{2}(v^{n+1} + v^n, D(v^{n+1} + v^n))_h = 0.$$

First note that for any periodic grid functions r, s we have $(r, D_0 s) = -(D_0 r, s)$ (just write out the expressions term by term and use the boundary conditions) so that

$$(v^{n+1} + v^n, D_0(v^{n+1} + v^n))_h = 0,$$

and the first part follows.

Second, as indicated by the hint, we have the identity

$$D_- v_j = D_0 v_j - \frac{h}{2} D_+ D_- v_j.$$

The second part then follows by noting that $(r, D_+ s) = -(D_- r, s)$ so that for scheme (1) we have

$$\|v^{n+1}\|_h^2 - \|v^n\|_h^2 + \frac{\Delta t h}{4}(D_-(v^{n+1} + v^n), D_-(v^{n+1} + v^n))_h = 0.$$

The

$$\|v^{n+1}\|_h^2 = \|v^n\|_h^2 - \frac{\Delta t h}{4}\|D_-(v^{n+1} + v^n)\|_h^2.$$