Department of Applied Mathematics

Preliminary Examination in Numerical Analysis

January 11, 2021, 9 am – 12 noon.

Submit solutions to four (and no more) of the following six problems. Show all your work, and justify all your answers. Start each problem on a new page, and write on one side only. No calculators allowed.
Do not write your name on your exam. Instead, write your student number on each page.

1. Root finding

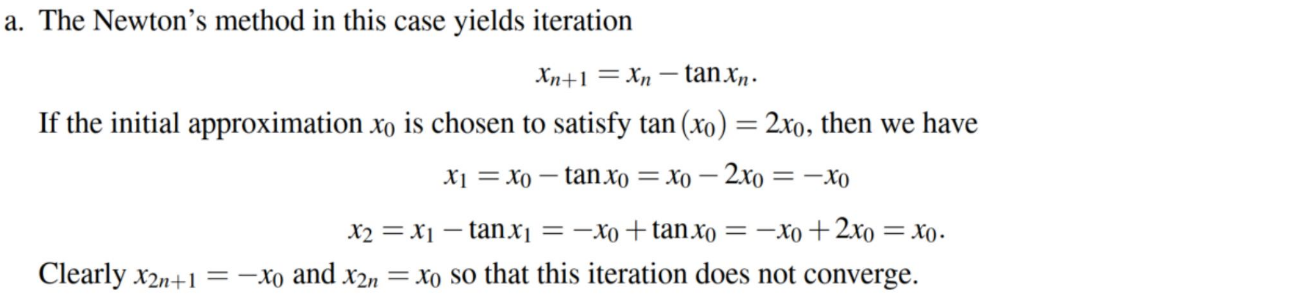
Consider Newton's method for solving the equation $\sin x = 0$ in the interval $(-\pi/2, \pi/2)$ starting with the initial approximation x_0 such that $\tan x_0 = 2x_0$ (nb. $x_0 \approx \pm 1.1656$).

a. What is the result of this iteration?

b. What is the result of the iteration if the initial approximation $\tilde{x}_0$ satisfies $|\tilde{x}_0| < |x_0|$?

c. What is the result of the iteration if the initial approximation $\tilde{x}_0$ satisfies $|\tilde{x}_0| > |x_0|$?

Solution:



a. The Newton's method in this case yields iteration

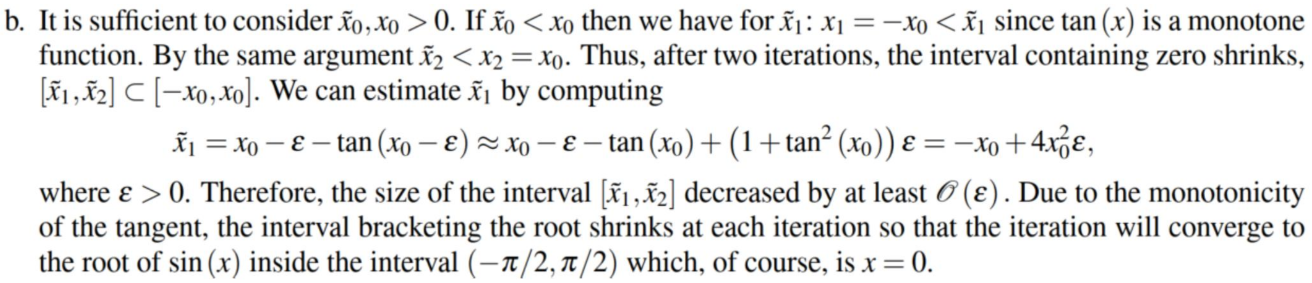
$$x_{n+1} = x_n - \tan x_n.$$

If the initial approximation x_0 is chosen to satisfy $\tan(x_0) = 2x_0$, then we have

$$x_1 = x_0 - \tan x_0 = x_0 - 2x_0 = -x_0$$

$$x_2 = x_1 - \tan x_1 = -x_0 + \tan x_0 = -x_0 + 2x_0 = x_0.$$

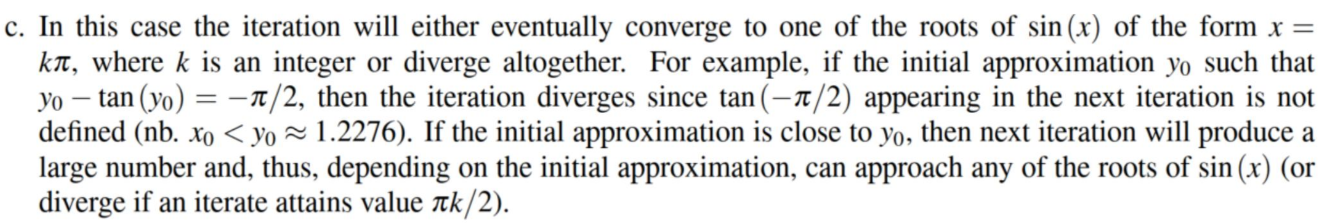
Clearly $x_{2n+1} = -x_0$ and $x_{2n} = x_0$ so that this iteration does not converge.



b. It is sufficient to consider $\tilde{x}_0, x_0 > 0$. If $\tilde{x}_0 < x_0$ then we have for $\tilde{x}_1$: $x_1 = -x_0 < \tilde{x}_1$ since $\tan(x)$ is a monotone function. By the same argument $\tilde{x}_2 < x_2 = x_0$. Thus, after two iterations, the interval containing zero shrinks, $[\tilde{x}_1, \tilde{x}_2] \subset [-x_0, x_0]$. We can estimate $\tilde{x}_1$ by computing

$$\tilde{x}_1 = x_0 - \varepsilon - \tan(x_0 - \varepsilon) \approx x_0 - \varepsilon - \tan(x_0) + \left(1 + \tan^2(x_0)\right)\varepsilon = -x_0 + 4x_0^2\varepsilon,$$

where $\varepsilon > 0$. Therefore, the size of the interval $[\tilde{x}_1, \tilde{x}_2]$ decreased by at least $\mathscr{O}(\varepsilon)$. Due to the monotonicity of the tangent, the interval bracketing the root shrinks at each iteration so that the iteration will converge to the root of $\sin(x)$ inside the interval $(-\pi/2, \pi/2)$ which, of course, is $x = 0$.



c. In this case the iteration will either eventually converge to one of the roots of $\sin(x)$ of the form $x = k\pi$, where k is an integer or diverge altogether. For example, if the initial approximation y_0 such that $y_0 - \tan(y_0) = -\pi/2$, then the iteration diverges since $\tan(-\pi/2)$ appearing in the next iteration is not defined (nb. $x_0 < y_0 \approx 1.2276$). If the initial approximation is close to y_0, then next iteration will produce a large number and, thus, depending on the initial approximation, can approach any of the roots of $\sin(x)$ (or diverge if an iterate attains value $\pi k/2$).

Yet another possibility is that the iterations will get stuck in a cycle of the type described in part (a).

2. Quadrature

a. What is the largest step size that makes the trapezoidal rule exact for trigonometric polynomials of the form

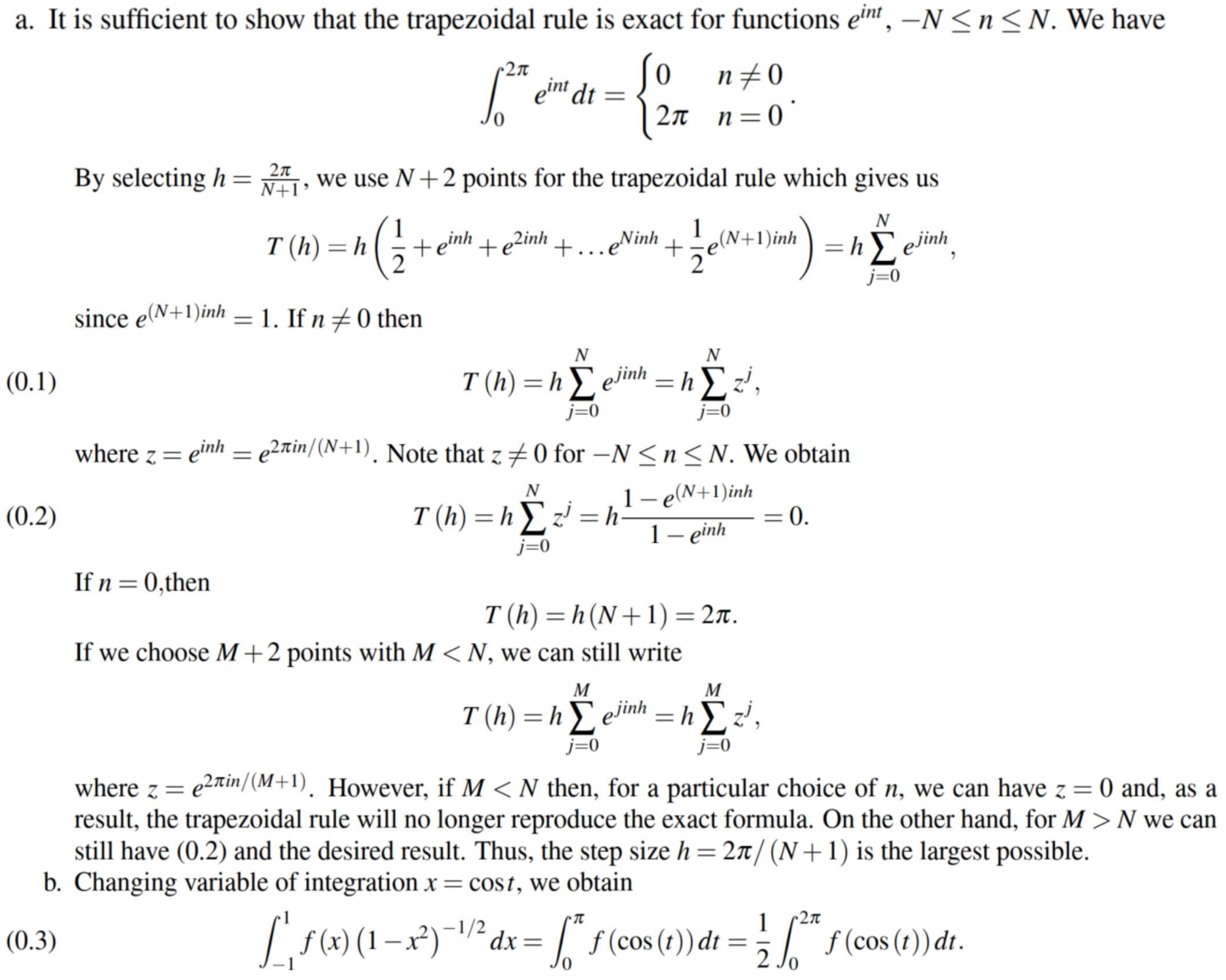
$$\sum_{n=-N}^{N} c_n e^{int}, \quad t \in [0, 2\pi).$$

b. Show that the formula

$$\int_{-1}^{1} f(x)(1-x^2)^{-1/2}\,dx = \frac{\pi}{N}\sum_{n=1}^{N} f\left(\cos\left(\frac{2n-1}{2N}\pi\right)\right)$$

is exact for all polynomials *f* of degree $2N-1$.

Solution:



a. It is sufficient to show that the trapezoidal rule is exact for functions e^{int}, $-N \le n \le N$. We have

$$\int_0^{2\pi} e^{int}dt = \begin{cases} 0 & n \neq 0 \\ 2\pi & n = 0 \end{cases}.$$

By selecting $h = \frac{2\pi}{N+1}$, we use $N+2$ points for the trapezoidal rule which gives us

$$T(h) = h\left(\frac{1}{2} + e^{inh} + e^{2inh} + \ldots e^{Ninh} + \frac{1}{2}e^{(N+1)inh}\right) = h\sum_{j=0}^{N} e^{jinh},$$

since $e^{(N+1)inh} = 1$. If $n \neq 0$ then

$$T(h) = h\sum_{j=0}^{N} e^{jinh} = h\sum_{j=0}^{N} z^j, \tag{0.1}$$

where $z = e^{inh} = e^{2\pi in/(N+1)}$. Note that $z \neq 0$ for $-N \le n \le N$. We obtain

$$T(h) = h\sum_{j=0}^{N} z^j = h\frac{1-e^{(N+1)inh}}{1-e^{inh}} = 0. \tag{0.2}$$

If $n = 0$,then

$$T(h) = h(N+1) = 2\pi.$$

If we choose $M+2$ points with $M < N$, we can still write

$$T(h) = h\sum_{j=0}^{M} e^{jinh} = h\sum_{j=0}^{M} z^j,$$

where $z = e^{2\pi in/(M+1)}$. However, if $M < N$ then, for a particular choice of n, we can have $z = 0$ and, as a result, the trapezoidal rule will no longer reproduce the exact formula. On the other hand, for $M > N$ we can still have (0.2) and the desired result. Thus, the step size $h = 2\pi/(N+1)$ is the largest possible.

b. Changing variable of integration $x = \cos t$, we obtain

$$\int_{-1}^{1} f(x)\left(1-x^2\right)^{-1/2}dx = \int_0^{\pi} f(\cos(t))\,dt = \frac{1}{2}\int_0^{2\pi} f(\cos(t))\,dt. \tag{0.3}$$

The last identity follows by replacing $t' = 2\pi - t$ in

$$\int_{\pi}^{2\pi} f(\cos(t))\,dt = -\int_{\pi}^{0} f(\cos(t'))\,dt' = \int_{0}^{\pi} f(\cos(t))\,dt.$$

We need to show that the formula is correct for all polynomials f of degree up to and including $2N-1$. One possible proof uses Part 1 of this problem. Since $f(\cos(t))$ is a trigonometric polynomial of degree $2N-1$, we need to use $2N+1$ points for the trapezoidal rule,

$$2\pi\frac{2n-1}{4N},\ n = 0,\dots 2N,$$

with the step size

$$h = \frac{\pi}{N}.$$

We obtain

$$T(h) = \frac{1}{2}h\sum_{n=1}^{2N} f\left(\cos\left(\pi\frac{2n-1}{2N}\right)\right) = h\sum_{n=1}^{N} f\left(\cos\left(\pi\frac{2n-1}{2N}\right)\right),$$

where we used

$$\begin{aligned}\sum_{n=N+1}^{2N} f\left(\cos\left(\pi\frac{2n-1}{2N}\right)\right) &= \sum_{n=N+1}^{2N} f\left(\cos\left(2\pi-\pi\frac{2n-1}{2N}\right)\right)\\ &= \sum_{n=1}^{N} f\left(\cos\left(\pi\frac{4N-2(n+N)+1}{2N}\right)\right)\\ &= \sum_{n=1}^{N} f\left(\cos\left(\pi\frac{2N-2n+1}{2N}\right)\right)\\ &= \sum_{n=1}^{N} f\left(\cos\left(\pi\frac{2n-1}{2N}\right)\right)\end{aligned}$$

and where the last identity is obtained by summing in reverse order, $n = N, N-1, \dots, 1$.
In an alternative proof, it is sufficient to show that the formula is correct for the Chebyshev polynomials of degree up to $2N-1$. From (0.3) we have

$$\begin{aligned}\int_{-1}^{1} T_k(x)\left(1-x^2\right)^{-1/2}dx &= \frac{1}{2}\int_{0}^{2\pi} T_k(\cos(t))\,dt\\ &= \frac{1}{2}\int_{0}^{2\pi}\cos(kt)\,dt\\ &= \begin{cases}0 & k\neq 0\\ \pi & k = 0\end{cases}\end{aligned}$$

Next, we need to compute

$$\begin{aligned}\frac{\pi}{N}\sum_{n=1}^{N} T_k\left(\cos\left(\frac{2n-1}{2N}\pi\right)\right) &= \frac{\pi}{N}\sum_{n=1}^{N}\cos\left(\frac{2n-1}{2N}k\pi\right)\\ &= \frac{\pi}{N}\mathscr{Re}\left[\sum_{n=1}^{N}\exp\left(i\frac{2n-1}{2N}k\pi\right)\right]\\ &= \frac{\pi}{N}\mathscr{Re}\left[\exp\left(i\frac{k\pi}{2N}\right)\sum_{n=0}^{N-1}\exp\left(i\frac{n}{N}k\pi\right)\right]\end{aligned}$$

If $k\neq 0$ and k is even, then we have

$$\frac{\pi}{N}\sum_{n=1}^{N} T_k\left(\cos\left(\frac{2n-1}{2N}\pi\right)\right) = \frac{\pi}{N}\mathscr{Re}\left[\exp\left(i\frac{k\pi}{2N}\right)\frac{1-\exp(ik\pi)}{1-\exp\left(i\frac{k\pi}{N}\right)}\right] = 0.$$

If k is odd, then we have a sum of values of an odd polynomial evaluated on a grid symmetric with respect to zero so that

$$\frac{\pi}{N}\sum_{n=1}^{N} T_k\left(\cos\left(\frac{2n-1}{2N}\pi\right)\right) = 0.$$

If $k = 0$,then we have

$$\frac{\pi}{N}\sum_{n=1}^{N} T_0\left(\cos\left(\frac{2n-1}{2N}\pi\right)\right) = \frac{\pi}{N}\sum_{n=1}^{N} 1 = \pi.$$

3. Linear Algebra

a. Define what is meant by a matric being *Hermitian*, and show that such a matrix has only real eigenvalues.

b. A matrix A is called *circulant* if its elements $a_{i,j}$ are all the same whenever $(i-j) \bmod N$ is the same. In other words, each row is the same as the row above shifted periodically one step to the right. Show that such a matrix can be diagonalized by similarity transforming it using the DFT (Discrete Fourier Transform) matrix, as given by

$$U = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)^2} \end{bmatrix},$$

where ω is the N^{th} root of unity.

c. The matrix $A = \begin{bmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{bmatrix}$ is both Hermitian and circulant.

Determine all its eigenvalues.

Solution:

a. A is Hermitian if $A = A^*$, where the star denotes transpose and conjugate.

From $A\underline{x} = \lambda\underline{x}$ follows $\underline{x}^* A^* = \overline{\lambda}\underline{x}^*$ and then $\underline{x}^* A\underline{x} = \lambda\underline{x}^*\underline{x}$, $\underline{x}^* A^*\underline{x} = \overline{\lambda}\underline{x}^*\underline{x}$. Subtracting the last two relations (noting that $\underline{x}^*\underline{x} > 0$) gives $\lambda = \overline{\lambda}$, i.e. λ is real.

b. In forming $U^* A U$, first consider

$$AU=\frac{1}{\sqrt{N}}\begin{bmatrix} a_0 & a_1 & \cdots & \cdots & a_{N-1}\\ a_{N-1} & a_0 & a_1 & \cdots & a_{N-2}\\ \ddots & \ddots & \ddots & \ddots & \ddots\\ \ddots & \ddots & \ddots & \ddots & \ddots\\ \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1\\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1}\\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)^2} \end{bmatrix}=$$

$$=\frac{1}{\sqrt{N}}\begin{bmatrix} \Sigma a_k & \Sigma a_k\omega^k & \Sigma a_k\omega^{2k} & \cdots & \Sigma a_k\omega^{(N-1)k}\\ \Sigma a_k & \omega\Sigma a_k\omega^k & \omega^2\Sigma a_k\omega^{2k} & \cdots & \omega^{N-1}\Sigma a_k\omega^{(N-1)k}\\ \Sigma a_k & \omega^2\Sigma a_k\omega^k & \omega^4\Sigma a_k\omega^{2k} & \cdots & \omega^{2(N-1)}\Sigma a_k\omega^{(N-1)k}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \qquad \text{(S1)}$$

(having utilized that $\omega=e^{2\pi i/N}$ satisfies $\omega^N=1$), and then

$$U^*AU=\frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1\\ 1 & \omega^{-1} & \omega^{-2} & \cdots & \omega^{-(N-1)}\\ 1 & \omega^{-2} & \omega^{-4} & \cdots & \omega^{-2(N-1)}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & \omega^{-(N-1)} & \omega^{-2(N-1)} & \cdots & \omega^{-(N-1)^2} \end{bmatrix}AU=$$

$$=\frac{1}{N}\begin{bmatrix} N(\Sigma a_k) & (\Sigma\omega^k)(\Sigma a_k\omega^k) & (\Sigma\omega^{2k})(\Sigma a_k\omega^{2k}) & \cdots & (\Sigma\omega^{(N-1)k})(\Sigma a_k\omega^{(N-1)k})\\ (\Sigma\omega^{-k})(\Sigma a_k) & N(\Sigma a_k\omega^k) & (\Sigma\omega^k)(\Sigma a_k\omega^{2k}) & \cdots & \cdots\\ (\Sigma\omega^{-2k})(\Sigma a_k) & (\Sigma\omega^{-k})(\Sigma a_k\omega^k) & N(\Sigma a_k\omega^{2k}) & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}.$$

All off-diagonal entries vanish (due to their first factor), and we read off *A*’s eigenvalues in the diagonal.

A faster solution is obtained by noting that, if *U* diagonalizes *A*, then the columns of *U* are the eigenvectors of *A*. We thus look at column *n* of *AU* (as calculated above, $n=0,1,2,\ldots,N-1$):

$$\frac{1}{\sqrt{N}}\begin{bmatrix} \Sigma a_k\omega^{nk}\\ \omega^n\Sigma a_k\omega^{nk}\\ \omega^{2n}\Sigma a_k\omega^{nk}\\ \vdots\\ \omega^{(N-1)n}\Sigma a_k\omega^{nk} \end{bmatrix}=\frac{1}{\sqrt{N}}(\Sigma a_k\omega^{nk})\begin{bmatrix} 1\\ \omega^n\\ \omega^{2n}\\ \vdots\\ \omega^{(N-1)n} \end{bmatrix}.$$

This agrees with the n^{th} column of *U*, multiplied with $(\Sigma a_k\omega^{nk})$, which shows $(\Sigma a_k\omega^{nk})$ to be the n^{th} eigenvalue, and it also confirms that *U* indeed diagonalizes *A*,

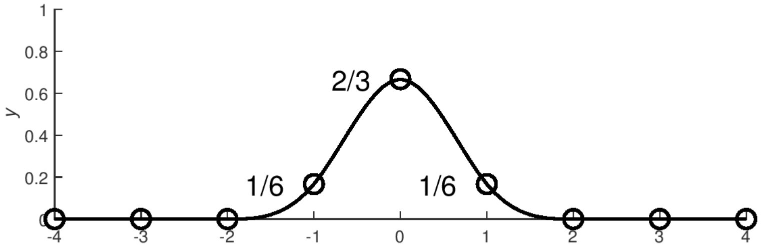
c. By the result just above, the eigenvalues of the given matrix A become:

$$\begin{array}{lll}
1-2+1 & =0 & (=-2+2\cos 0) \\
\omega^{-1}-2+\omega^{1} & =e^{-i\omega}-2+e^{i\omega} & =-2+2\cos\left(1\cdot\dfrac{2\pi}{N}\right) \\
\omega^{-2}-2+\omega^{2} & \cdots & =-2+2\cos\left(2\cdot\dfrac{2\pi}{N}\right) \\
\omega^{-3}-2+\omega^{3} & \cdots & =-2+2\cos\left(3\cdot\dfrac{2\pi}{N}\right) \\
\quad\vdots & & \\
\omega^{-(N-1)}-2+\omega^{(N-1)} & \cdots & =-2+2\cos\left((N-1)\cdot\dfrac{2\pi}{N}\right)
\end{array}$$

As expected, these are all real.

4. Interpolation / Approximation

The figure below illustrates a cubic *B*-spline on a unit-spaced grid, uniquely defined when using its standard normalization.



a. Tell what the defining property is of a *B-spline* (as opposed to any other spline). Also tell what is the customary normalization of any *B*-spline.

b. Verify that the *B*-spline in the figure can be written explicitly as

$$B(x)=\frac{1}{12}\left(1\cdot|x+2|^{3}-4\cdot|x+1|^{3}+6\cdot|x|^{3}-4\cdot|x-1|^{3}+1\cdot|x-2|^{3}\right). \qquad (1)$$

c. Translates of *B*-splines form an excellent set of basis functions for representing a general spline. Consider the nodes $x_i=i, i=0,1,2,\ldots,N$ with matching function values y_i, and let $B_i(x)$ denote the *B*-spline centered at $x=i,\ i=-1,0,1,2,\ldots,N+1.$ Write out the linear system that needs to be solved for obtaining the *B*-spline coefficients for the *natural* cubic spline that obeys this data.

Hint: For the *B*-spline (as given in (1)), $B''(x)$ takes the values {1,-2, 1} at $x=\{-1,0,1\}$.

Solution:

a. A *B*-spline is the spline that, in the narrowest way possible, transitions non-trivially from identically zero back to identically zero. The customary normalization is that its integral becomes one.

b. The given expression is clearly a cubic within each subinterval and, at each node, it is discontinuous only in its third derivative. Given the uniqueness mentioned in the problem statement, there are just two further items we need to verify:

i. The formula (1) for *B*(*x*) evaluates to zero when $|x| > 2$:

Say $x > 2$ (equivalent for $x < -2$). We can then change all magnitude signs to parentheses, and the expression becomes a single cubic:

$$12B(x) = \left(1\cdot(x+2)^3 - 4\cdot(x+1)^3 + 6\cdot(x)^3 - 4\cdot(x-2)^3 + 1\cdot(x-2)^3\right) \quad = \qquad (2)$$

$$\begin{array}{rl} 1\cdot & (x^3+6x^2+12x+8) \\ -4\cdot & (x^3+3x^2+3x+1) \\ 6\cdot & (x^3) \\ -4\cdot & (x^3-3x^2+3x-1) \\ 1\cdot & (x^3-6x^2+12x-8) \quad = 0 \end{array}$$

since the coefficient for every power of *x* has vanished.

More 'elegantly' we can alternatively note that the right hand side of (2) can be seen as applying the finite difference (FD) approximation $[1, -4, 6, -4, 1]/h^4$ with $h = 1$ to the cubic function x^3. This FD formula is the second order approximation to the 4th derivative (the second derivative approximation $[1, -2, 1]/h^2$ applied twice). Applied to a cubic, the result has to become zero.

Note: If one defines $\lfloor x \rfloor_+ = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } x > 0 \end{cases}$, a related way to represent the cubic *B*-spline is $B(x) = \frac{1}{6}\left(1\cdot\lfloor x+2 \rfloor_+^{\,3} - 4\cdot\lfloor x+1 \rfloor_+^{\,3} + 6\cdot\lfloor x \rfloor_+^{\,3} - 4\cdot\lfloor x-1 \rfloor_+^{\,3} + 1\cdot\lfloor x-2 \rfloor_+^{\,3}\right)$. While this formula looks similar to (1), the scaling factor differs.

ii. It is correctly normalized:

The value of *B*(0) is given (in the figure) as 2/3. That matches substituting zero into (1).

c. Let the spline be $s(x) = \sum_{i=-1}^{N+1} \alpha_i B_i(x)$. We next recall that a *natural spline* is characterized by second derivative zero at both end points. Enforcing this (using the hint) together with it producing the correct value at each node point gives the system

$$\frac{1}{6}\begin{bmatrix} 1 & -2 & 1 & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & 1 & -2 & 1 \end{bmatrix}\begin{bmatrix}\alpha_{-1}\\ \alpha_0\\ \alpha_1\\ \vdots\\ \alpha_N\\ \alpha_{N+1}\end{bmatrix}=\begin{bmatrix}0\\ y_0\\ y_1\\ \vdots\\ y_N\\ 0\end{bmatrix}\begin{array}{l}\leftarrow \text{Left BC}\\ \leftarrow \text{Match}\\ \leftarrow \text{values at}\\ \leftarrow \text{nodes}\\ \vdots\\ \leftarrow \text{Right BC}\end{array}$$

5. Numerical ODEs

a. Define what is meant by the stability domain of an ODE solver.

b. Determine the stability domains for the Forward Euler (FE) and Backward Euler (BE) methods (first order Adams-Bashforth and Adams-Moulton methods, respectively).

c. Suppose one uses FE as a predictor and BE as a corrector. What is the order of accuracy of the resulting method? Either derive it, or quote a specific, more general theorem.

d. Give an equation for the stability domain of the FE – BE predictor corrector method. Determine what (if any) intervals along the real and imaginary axes fall within this domain.

Solution:

a. When applying the solver to the ODE $y'(t)=\lambda y(t)$, using a time step *k*, the stability domain is the domain in the complex $\xi=\lambda k$ plane in which the solution is not growing.

b. FE applied to $y'=\lambda y$ gives $y(t+k)=y(t)+k\lambda y(t)=(1+\xi)y(t)$, which is a linear recursion relation with characteristic equation $r=1+\xi$, i.e., the stability domain is the circle $|1+\xi|\le 1$. BE applied to $y'=\lambda y$ similarly gives $y(t+k)=y(t)+k\lambda y(t+k)$ and $r=\dfrac{1}{1-\xi}$ with the (unbounded) domain $|1-\xi|\ge 1$.

c. If the predictor has accuracy *p* and the corrector accuracy *q*, the resulting order of accuracy is min(*p*+1,*q*). Since here, $p=q=1$, the combination is only first order accurate.

d. The combined scheme, applied to $y'(t)=\lambda y(t)$ can be written as $y(t+k)=y(t)+\xi(y(t)+\xi y(t))$, with characteristic equation $r=1+\xi+\xi^2$. The stability domain is given by $|1+\xi+\xi^2|\le 1$.

For ξ real, we can write the characteristic equation as $r=\left(\xi+\frac{1}{2}\right)^2+\frac{3}{4}$, which satisfies $|r|\le 1$ for $-1\le\xi\le 0$.

For ξ purely imaginary, we write ξ as $\xi=i\alpha$ with α real. and obtain $r=1+i\alpha-\alpha^2$, $|r|^2=(1-\alpha^2)^2+\alpha^2=\alpha^4-\alpha^2+1$, and $|r|^2-1=\alpha^2(\alpha+1)(\alpha-1)$. This is less than or equal to zero for $-1\le\alpha\le 1$.

6. <u>Numerical PDEs</u>

Consider the Crank-Nicolson method

$$u_j^{(n+1)} - u_j^{(n)} = \frac{1}{2}\frac{k}{h^2}\left(\left(u_{j-1}^{(n+1)} - 2u_j^{(n+1)} + u_{j+1}^{(n+1)}\right) + \left(u_{j-1}^{(n)} - 2u_j^{(n)} + u_{j+1}^{(n)}\right)\right)$$

for the heat equation

$$\begin{cases} u_t = u_{xx}, t \ge 0, x \in [0, 2\pi] \\ u(x,0) = u_0(x) \\ u(x+2\pi, t) = u(x,t), t \ge 0 \end{cases} .$$

a. Show that the scheme is unconditionally stable.

b. Show that it is second order accurate in both space and time.

<u>Solution:</u>

a. There are two primary approaches for determining stability:

(i) Consider the scheme as a Method-of-Lines solution, and refer to ODE stability domains:

Discretizing first in space only, we get the ODE system

$$\frac{du_j}{dt} = \frac{1}{h^2}(u_{j-1} - 2u_j + u_{j+1}),\ j = 0,1,\ldots,N-1, u_N = u_0, u_{-1} = u_{N-1} \ ,$$

or

$$\frac{d\underline{u}}{dt} = A\underline{u} \ .$$

The matrix *A* of this system is symmetric and negative semi-definite which can be observed by using the Gershgorin theorem which provides a spectral bound $-4/h^2 \le \lambda \le 0$. The time stepping scheme is the trapezoidal rule (Adams-Moulton of second order) which is *A*-stable (the entire left half-plane falls within the stability domain). Whatever the values are for the (positive) time and space steps, all the eigenvalues λ fall within the stability domain, i.e., the scheme is unconditionally stable.

(ii) Apply von Neumann analysis.

Substitute $u(x,t) = \xi^{t/k} e^{i\omega x}$ into the discrete approximation. This gives, after a brief simplification

$$\xi - 1 = \frac{1}{2}\frac{k}{h^2}(\xi + 1)\left(e^{-i\omega x} - 2 + e^{i\omega x}\right).$$

With $e^{-i\omega x} - 2 + e^{i\omega x} = -4\left(\frac{e^{i\omega h/2} - e^{-i\omega h/2}}{2i}\right)^2 = -4\left(\sin\frac{\omega h}{2}\right)^2$, we obtain

$$\xi = \frac{1-2\frac{k}{h^2}\left(\sin\frac{\omega h}{2}\right)^2}{1+2\frac{k}{h^2}\left(\sin\frac{\omega h}{2}\right)^2}$$, and therefore $|\xi| \le 1$ for all ω and (positive) *k* and *h*.

b. We give again two solutions. Note that it is insufficient to just argue that the centered approximations used in the Crank-Nicholson scheme each by itself is second order accurate, since it also matters how the terms are combined.

(i) Refer to ODE theory

Given that the scheme is of Method-of-Lines (MOL) type, then the overall order of accuracy will indeed be that of the space scheme and of the time stepping method. These are in this case well known to both be second order accurate.

(ii) Direct series expansion

Given the symmetries of the scheme, it is natural to expand around $\left\{x, y+\frac{k}{2}\right\}$ and call this point {0,0}. We then get

$$\frac{1}{k}\left(u(0,\frac{k}{2})-u(0,-\frac{k}{2})\right)=u_t(0,0)+O(k^2)\ ,$$

$$\frac{1}{h^2}\left(u(-h,\frac{k}{2})-2u(0,\frac{k}{2})+u(h,\frac{k}{2})\right)=u_{xx}(0,0)+\frac{k}{2}u_{xxt}(0,0)+O(h^2)+O(k^2)\ .$$

Averaging the last equation with its counterpart for $k \to -k$ cancels its second term, and the Crank-Nicolson combination shows a total error of $O(h^2)+O(k^2)$.