

Numerical Analysis PhD Qualifying Exam
University of Vermont, Spring 2010

1. This question concerns number representation and errors. Normalized floating point numbers can be represented by $\pm 1.b_1b_2b_3 \dots b_N \times 2^{\pm p}$ where b_i is either 0 or 1, N is the number of bits in the mantissa, p is an M -bit binary exponent, and two additional bits are used to store the signs. Assume we are using a machine for which $N = 23$ and $M = 7$. Note that you may not need all of the information above to solve the problem.

(a): Let $x = 2^{16} + 2^{-8} + 2^{-9} + 2^{-10}$ and let x^* be the machine number closest to x on the machine above. What is $|x - x^*|$?

(b): The **Theorem on Loss of Precision** states that if x and y are positive normalized floating point binary machine numbers such that $x > y$ and

$$2^{-q} \leq 1 - \frac{y}{x} \leq 2^{-p}$$

then at most q and at least p significant binary bits are lost in the subtraction $x - y$. The theorem is useful in estimating the likelihood of catastrophic cancellation, which is common when performing modular arithmetic.

Use the theorem to show that if $x > \pi \cdot 2^{25}$ is a number represented *exactly* on the machine described above, then $z \equiv x \pmod{2\pi} = x - 2k\pi$ can be computed with *no* significant digits. The value of z is required to compute $\cos x$, for example. *Hint:* Use $y = 2k\pi$ and solve for p .

Solution:

(a):

$$x = 2^{16}(2^0 + 2^{-24} + 2^{-25} + 2^{-26})$$

$$(1.00000 \ 00000 \ 00000 \ 00000 \ 00011 \ 1)_2 \times 2^{16}$$

Writing x_- for the closest machine number below x and x_+ for the closest machine number above x , we have

$$x_- = (1.00000 \ 00000 \ 00000 \ 00000 \ 000)_2 \times 2^{16} = 2^{16}$$

$$x_+ = (1.00000 \ 00000 \ 00000 \ 00000 \ 001)_2 \times 2^{16} = 2^{16} + 2^{-7}$$

Now

$$x - x_- = 2^{16}(2^{-24} + 2^{-25} + 2^{-26}) = 2^{-8} + 2^{-9} + 2^{-10} = 7 \times 2^{-10}$$

$$x_+ - x = 2^{-7} - 2^{-8} - 2^{-9} - 2^{-10} = 1 \times 2^{-10}$$

So $|x_+ - x| < |x - x_-|$. Therefore $x^* = x_+$ and $|x - x^*| = 2^{-10}$.

(b): Letting $z \equiv x \pmod{2\pi}$, we have $0 < z = x - 2k\pi < 2\pi$. Using the Theorem and letting $y = 2k\pi$, we have

$$1 - \frac{2k\pi}{x} \equiv \frac{x - 2k\pi}{x} < \frac{2\pi}{x} < \frac{2\pi}{\pi \cdot 2^{25}} = 2^{-24}$$

Therefore, at least 24 bits are lost in the modular arithmetic.

□

2. This question concerns root finding. To avoid computing the derivative at each step in Newton's method, it has been proposed to replace $f'(x_n)$ by $f'(x_0)$. Define the error at step n to be $e_n = x_n - r$ where the function f has a single root at the point r , i.e. $f(r) = f(x_n - e_n) = 0$. Derive the rate of convergence for this method by finding the relationship between e_{n+1} and e_n . *Hint:* You will need a Taylor remainder at one point.

Solution:

This modification of Newton's method is of the form

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}$$

Plugging in what we know about the error term, we find

$$e_{n+1} = x_{n+1} - r = e_n - \frac{f(x_n)}{f'(x_0)} = \frac{e_n f'(x_0) - f(x_n)}{f'(x_0)}$$

Now we have $f(r) = f(x_n - e_n) = 0$, where Taylor series tell us $f(x_n - e_n) = f(x_n) - e_n f'(c)$ for some $c \in [x_n - e_n, x_n]$. So

$$e_{n+1} = \frac{e_n f'(x_0) - e_n f'(c)}{f'(x_0)} = \left(1 - \frac{f'(c)}{f'(x_0)}\right) e_n$$

Which implies that the method exhibits linear convergence.

□

3. **(a):** Compute a singular value decomposition $A = USV^\top = \sum_{i=1}^2 s_i \vec{u}_i \vec{v}_i^\top$ of the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 3/2 \end{bmatrix}$$

It is advised that you keep your entries as fractions to avoid nasty numbers.

(b): What is the best rank-1 approximation of the matrix A ?

(c): What is the condition number of the matrix A and what does this say about the number of significant digits d in the solution to $A\vec{x} = \vec{b}$? *Note* that a good explanation of your reasoning is more important than an exact answer for d .

Solution:

(a):

$$\begin{aligned}
A = \begin{bmatrix} 0 & 1 \\ 1 & 3/2 \end{bmatrix} &= USV^T = \begin{bmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \left(\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1/2 \end{bmatrix} \right) \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{bmatrix} \\
&= 2 \begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -\frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{2}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{8}{5} \end{bmatrix} + \begin{bmatrix} -\frac{2}{5} & \frac{1}{5} \\ \frac{1}{5} & -\frac{1}{10} \end{bmatrix}
\end{aligned}$$

(b): The original matrix A is separated above into a larger contribution plus a smaller contribution, due to the magnitude of the singular values. The best rank-one approximation of A is given by the first rank-one matrix above. The second matrix provides small corrections.

(c): The condition number of A is the ratio of its largest singular value to its smallest, namely $s_1/s_2 = 4$. Since $4 \approx O(10^1)$, we know that if \vec{b} has d significant digits, then the solution vector \vec{x} will have $d - 1$ significant digits.

□

4. (a): Determine the coefficients of an implicit, one-step, ODE method of the form

$$x(t+h) = ax(t) + bx'(t) + cx'(t+h)$$

so that it is exact for polynomials of as high a degree as possible. Begin by letting LHS = $x(h)$ and RHS = $ax(0) + bx'(0) + cx'(h)$ and fill in the missing entries in the table below. The first row and column have been filled in for you.

$x(t)$	$x'(t)$	LHS	RHS
1	0	1	a
t			
t^2			
...

(b): Once you have obtained the coefficients a, b, c in part (a), use Taylor Series to find the order of the local truncation error term.

Solution:

(a): Setting LHS = RHS, we find $a = 1, b = c = h/2$, but only up to degree 2 polynomials.

(b): By Taylor series we know

$x(t)$	$x'(t)$	LHS	RHS
1	0	1	a
t	1	h	$b + c$
t^2	$2t$	h^2	$2hc$
t^3	$3t^2$	h^3	$3h^2c$

$$\text{LHS} = x + hx' + \frac{h^2}{2}x'' + \frac{h^3}{3!}x''' + \dots$$

Then

$$\begin{aligned} \text{RHS} &= x + \frac{h}{2}x' + \frac{h}{2}(x' + hx'' + \frac{h^2}{2}x''' + \dots) \\ &= x + hx' + \frac{h^2}{2}x'' + \frac{h^3}{4}x''' + \dots \end{aligned}$$

Since $\text{LHS} \neq \text{RHS}$ after the third term, we know the error term is $O(h^3)$.

□

5. The Dahlquist method

$$Y_{n+1} - 2Y_n + Y_{n-1} = \frac{h^2}{4}(f_{n+1} + 2f_n + f_{n-1}), \quad \text{where } f_n \equiv f(x_n, Y_n), \text{ etc.} \quad (1)$$

can be used to solve the initial-value problem

$$y'' = f(x, y), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0. \quad (2)$$

(a) Show that method (1) has the global error of order 2 when applied to (2).

(b) Show that this method is stable for any value of $h\omega$ when applied to the oscillator equation

$$y'' = -\omega^2 y, \quad \omega > 0. \quad (3)$$

Solution:

$$\begin{aligned}
 (a) \quad Y_{n+1} &= 2y_n - y_{n-1} + \frac{h^2}{4}(f_{n+1} + 2f_n + f_{n-1}) \\
 &= y_n - (-hy_n' + \frac{h^2}{2}y_n'' - \frac{h^3}{6}y_n''' + \frac{h^4}{24}y_n'''' + \dots \\
 &\quad + \frac{h^2}{4}(\cancel{f_n} + \cancel{hf_n'} + \frac{h^2}{2}\cancel{f_n''} + \dots + 2f_n + \\
 &\quad + \cancel{f_n} - \cancel{hf_n'} + \frac{h^2}{2}\cancel{f_n''} + \dots)) \\
 \text{use } y_n'' = f_n &\quad \downarrow \\
 &= y_n + hy_n' + (-\frac{h^2}{2} + h^2)y_n'' + \frac{h^3}{6}y_n''' \\
 &\quad + (-\frac{h^4}{24} + \frac{h^4}{4})y_n'''' + \dots \\
 &= y_n + hy_n' + \frac{h^2}{2}y_n'' + \frac{h^3}{6}y_n''' + \frac{5h^4}{24}y_n'''' + \dots
 \end{aligned}$$

The first 4 terms are the expansion terms of the exact solution y_{n+1} ; the 5th term is the local truncation error:

$$LTE = O(h^4).$$

Since it is applied to the 2nd-order ODE (2), the global ~~truncation~~ error is $O(h^{4-2}) = O(h^2)$,
 \Rightarrow the method is 2nd order. ✓

(b) let $y_n = p^n$. Then (1) applied to (3) yields:

$$p^2 - 2p + 1 = \frac{h^2}{4} \cdot (-\omega^2) \cdot (p^2 + 2p + 1)$$

$$\Rightarrow p^2(1 + \frac{h^2}{4}\omega^2) - 2p(1 - \frac{h^2}{4}\omega^2) + (1 - \frac{h^2}{4}\omega^2) = 0$$

$$\Rightarrow p^2 - 2z \cdot p + 1 = 0, \quad z = \frac{1 - \frac{h^2}{4}\omega^2}{1 + \frac{h^2}{4}\omega^2},$$

where $|z| < 1$ for any real ω^2 .

$$p_{1,2} = z \pm \sqrt{z^2 - 1} = z \pm i \underbrace{\sqrt{1 - z^2}}_{\text{real since } |z| < 1}$$

$$\Rightarrow |p_{1,2}| = z^2 + (1 - z^2) = 1 \text{ for any } z$$

and hence for any $(h\omega) \in \mathbb{R}$.

Thus the Dahlquist method for (3) is stable. \square

6. (a) Propose a 2nd-order accurate discretization of the equation

$$(p(x) u_x)_x = q(x)u + r(x). \quad (1)$$

(b) Use this discretization to set up a linear system for the boundary-value problem given by Eq. (1) and by the boundary conditions

$$u_x(0) = \alpha, \quad u(1) = \beta, \quad (2)$$

where α, β are some given constants. Use $h = 1/3$ and write out each equation in the linear system in question. Make sure to use the 2nd-order accurate approximation for the Neumann boundary conditions.

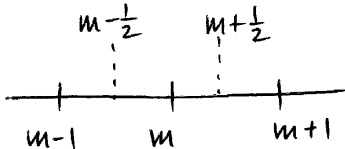
Solution:

Solution:

(a)

$$\underbrace{\frac{p_{m+\frac{1}{2}}(u_{m+1}-u_m)}{h} - \frac{p_{m-\frac{1}{2}}(u_m-u_{m-1})}{h}}_{\text{second order about } m} = q_m u_m + r_m$$

2nd-order about $m+\frac{1}{2}$ 2nd-order about $m-\frac{1}{2}$



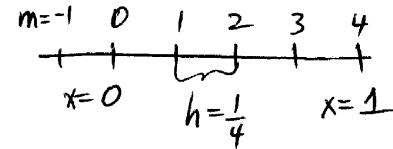
Simplifying:

$$p_{m+\frac{1}{2}} u_{m+1} - (p_{m+\frac{1}{2}} + p_{m-\frac{1}{2}} + h^2 q_m) u_m + p_{m-\frac{1}{2}} u_{m-1} = r_m. \quad (3)$$

(b) 2nd-order approximation to Neumann at 0:

$$\frac{u_1 - u_{-1}}{2h} = \alpha \Rightarrow u_1 - u_{-1} = 2\alpha h. \quad (4)$$

~~Specify~~ Specify (3) & (4) to



(4): $u_{-1} - u_1 = -2h\alpha$

(3), $m=0$: $p_{\frac{1}{2}} u_{-1} - (p_{\frac{1}{2}} + p_{\frac{3}{2}} + h^2 q_0) u_0 + p_{\frac{3}{2}} u_1 = r_0$

$m=1$: $p_{\frac{3}{2}} u_0 - (p_{\frac{3}{2}} + p_{\frac{5}{2}} + h^2 q_1) u_1 + p_{\frac{5}{2}} u_2 = r_1$

$m=2$: $p_{\frac{5}{2}} u_1 - (p_{\frac{5}{2}} + p_{\frac{7}{2}} + h^2 q_2) u_2 + p_{\frac{7}{2}} u_3 = r_2$

$m=3$ $p_{\frac{7}{2}} u_2 - (p_{\frac{7}{2}} + p_{\frac{9}{2}} + h^2 q_3) u_3 + p_{\frac{9}{2}} u_4 = r_3$

~~Specify~~

///

□

7. Consider a unidirectional wave equation

$$u_t = c u_x, \quad -\infty < x < \infty \quad (1)$$

where $c = \text{const.}$

(a) Use the von Neumann analysis to determine under what condition on the ratio

$$\mu = \frac{c\kappa}{h}$$

the scheme

$$\frac{U_m^{n+1} - U_m^n}{\kappa} = c \frac{U_{m+1}^n - U_m^n}{h}, \quad (2)$$

approximating (1), is stable.

(b) Similarly, show that the scheme

$$\frac{U_m^{n+1} - U_m^n}{\kappa} = c \frac{U_{m+1}^n - U_{m-1}^n}{2h} \quad (3)$$

is unstable for any μ .

(c) Note that for a Fourier harmonic $u = \exp[i\beta x]$, the right-hand sides of (2) and (3) equal λu for some λ . (Of course, this λ is different for (2) and (3).) Use this fact to interpret your results in parts (a) and (b) in light of the stability of a certain numerical method for ODEs.

Solution:

Solution:

(a) Substituting $u_m^n = \rho^n e^{i\beta h m}$ into (2) obtain:

$$\rho - 1 = \mu(e^{i\beta h} - 1).$$

$$\text{Then } |\rho|^2 \leq 1 \Rightarrow$$

$$1 + 2\mu(\cos\beta h - 1) + \mu^2(\cos^2\beta h - 2\cos\beta h + 1) + \mu^2 m^2 \beta^2 h \leq 1$$

$$\Rightarrow 1 + 2\mu z - 2\mu^2 z \leq 1$$

$$\text{where } z = \cos\beta h - 1.$$

Continuing, we have (upon cancelling by $\mu > 0$):

$$z - \mu z \leq 0 \Rightarrow$$

$$1 - \mu \geq 0 \quad (\text{since } z \leq 0)$$

$$\Rightarrow \boxed{\mu \leq 1}.$$

Thus, scheme (2) is stable when

$$\boxed{\frac{C K}{h} \leq 1}.$$

(b) Similarly,

$$\rho - 1 = \mu \left(\frac{e^{i\beta h} - e^{-i\beta h}}{2} \right) \Rightarrow$$

$$\rho = 1 + i\mu \cdot \sin \beta h \Rightarrow$$

$$|\rho|^2 = 1 + \mu^2 \sin^2 \beta h \geq 0 \text{ for all } \beta h \neq 0, \pi.$$

Thus, scheme (3) is unstable for all μ .

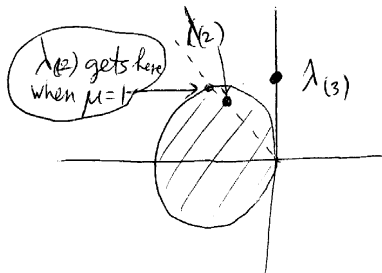
(c) For (2), $\lambda = \frac{c}{h}(e^{i\beta h} - 1) = \frac{c}{h}[(\cos \beta h - 1) + i \sin \beta h]$

$$\lambda_{(2)} = c \cdot \left[-\frac{(1 - \cos \beta h)}{h} + i \frac{\sin \beta h}{h} \right].$$

For (3),

$$\lambda_{(3)} = c \cdot i \cdot \frac{\sin \beta h}{h}.$$

We can plot these values in the complex plane along with the stability region for the simple Euler method (which is represented by the l.h.s. of (2) and (3)):



Note that $\lambda_{(2)}$ can be inside the stability region only for $\mu \leq 1$, but $\lambda_{(3)}$ cannot be in it at all (for any μ).

□