Review Notes for the Numerical Analysis Qualifying Exam

Spring 2017 – Fall 2017

Written by

Howard Heaton

Purpose: This document is a compilation of notes generated to prepare for the Numerical Analysis Qualifying Exam at UCLA. I have documented the followings solutions that I have completed to speed up the review process. These are quite incomplete and certainly contain typos and errors. If the reader finds any mistakes, please feel free to email me at `heaton@math.ucla.edu` and I will post an updated set of notes on my webpage.

Contents

1 Introduction **1**

2 Initial Value Problems **2**

3 Iterative Techniques in Matrix Algebra **2**

4 Old Numerical Qual Solutions **6**

Fall 2002 6

Fall 2003 8

Winter 2004 11

Fall 2004 15

Winter 2005 16

Spring 2007 22

Fall 2008 26

Spring 2009 30

Spring 2010 34

Fall 2010 44

Spring 2011 49
Fall 2011 51
Spring 2012 53
Fall 2012 58
Spring 2013 65
Fall 2013 75
Spring 2014 84
Fall 2014 87
Spring 2015 90
Fall 2015 108
Spring 2016 124
Fall 2016 145
Spring 2017 167

References **185**

1 Introduction

Many solutions from old exams are contained herein. First we provide some background material that I found relevant for the solutions. Then the solutions themselves are presented. I attempted to make these self-contained, as if to model what I would hope to write on the actual exam. Consequently, some of the solutions may seem repetitive. Lastly, please note some of the older solutions here closely follow the notes of Alejandro Cantarero and Shyr-Shea.

2 Initial Value Problems

Definition: The **local truncation error** (LTE) of a scheme is the amount by which the exact solution fails to satisfy the difference equation associated with the approximation method. If $\{w_n\}_{n=0}^N$ gives the iterates of the scheme, then the LTE, denoted τ_n, is $\tau_n := y(t_n) - w_n$. △

Definition: The **region of absolute stability** of a numerical method is the set $S \subset \mathbb{C}$ of $h\lambda$ such that when the numerical method is applied to the model problem, the iterates satisfy $y_n \longrightarrow 0$ for all initial conditions. △

Definition: The **interval of absolute stability** is the intersection of the region of absolute stability with the real axis, i.e., $S \cap \mathbb{R}$. △

3 Iterative Techniques in Matrix Algebra

Definition: A matrix $T \in \mathbb{R}^{n\times n}$ is said to be **convergent** provided that

$$\lim_{n\to\infty} (A^n)_{ij} = 0 \quad \text{for } i,j = 1,\ldots,n. \tag{1}$$

△

Theorem: Let $A \in \mathbb{C}^{n\times n}$ with spectral radius $\rho(A)$. Then $\rho(A) < 1$ if and only if A is convergent. △

Proof:

Assume A is convergent and let v and λ be any eigenvector-eigenvalue pair for A. Then

$$0 = \left(\lim_{k\to\infty} A^k\right) v = \lim_{k\to\infty} \left(A^k v\right) = \lim_{k\to\infty} \lambda^k v = \left(\lim_{k\to\infty} \lambda^k\right) v. \tag{2}$$

Because eigenvectors are nonzero, the above implies $\lim_{k\to\infty} \lambda^k = 0$, and so $|\lambda| < 1$. Whence $\rho(A) < 1$.

Conversely, assume $\rho(A) < 1$ and write $A = PJP^{-1}$ where P is invertible and J is block diagonal, consisting of Jordan blocks. Let M be any Jordan block and note $M = \lambda I + N$ where N is the matrix of 1's along the super diagonal and λ is an eigenvalue of A. The binomial theorem asserts

$$M^k = (\lambda I + N)^k = \sum_{j=0}^{k} \binom{k}{j} \lambda^{k-j} N^j = \sum_{j=0}^{n} \binom{k}{j} \lambda^{k-j} N^j \tag{3}$$

where the second equality holds since N is nilpotent of order n. Moreover, for $j = 1, \ldots, n$,

$$0 \leq \lim_{k\to\infty} \lambda^k \binom{k}{j} \leq \frac{1}{j!} \lim_{k\to\infty} \lambda^k k^j = \frac{1}{j!} \cdot 0 = 0. \tag{4}$$

The first equality follows from the fact $|\lambda| < 1$. This shows

$$\lim_{k\to\infty} M^k = \lim_{k\to\infty} \sum_{j=0}^{n} \binom{k}{j} \lambda^{k-j} N^j = \sum_{j=0}^{n} \left(\lim_{k\to\infty} \binom{k}{j} \lambda^{k-j} \right) N^j = \sum_{j=0}^{n} 0 N^j = 0. \tag{5}$$

This shows each Jordan block goes to zero as $k \longrightarrow \infty$. Moreover, through induction we obtain $A^k = PJ^kP^{-1}$ for each $k \in \mathbb{Z}^+$, and so

$$\lim_{k\to\infty} A^k = P \left(\lim_{k\to\infty} J^k \right) P^{-1} = P0P^{-1} = 0, \tag{6}$$

as desired. □

Lemma: If T is a matrix with spectral radius $\rho(T) < 1$, then $(I - T)^{-1}$ exists and

$$(I - T)^{-1} = \sum_{j=0}^{\infty} T^j. \tag{7}$$

△

Proof:

Suppose λ and v for an eigenvalue/eigenvector pair for T. Then $Tx = \lambda x$ and $(I - T)x = (I - \lambda x) =$

$(1-\lambda)x$. This shows $(1-\lambda)$ is an eigenvalue of $(I-T)$ if and only if λ is an eigenvalue of T. By hypothesis, $|\lambda| \leq \rho(T) < 1$, and so 1 is not an eigenvalue of T, which implies $1-1=0$ is not an eigenvalue of $I-T$ and so $\det(I-T) \neq 0$. Whence $(I-T)^{-1}$ exists.

We now compute $(I-T)^{-1}$. Since $\rho(T) < 1$ and $\rho(T^n) = \rho(T)^n$, we see

$$\lim_{n\to\infty} \rho(T^n) = \lim_{n\to\infty} \rho(T)^n = 0. \tag{8}$$

And, $\rho(T) = \inf\{\|T\| \;\;|\;\; \|\cdot\| \text{ is a norm}\}$, which means $\lim\limits_{n\to\infty} \|T\|^n = 0$ for some norm $\|\cdot\|$. In the finite dimensional case, all norms are equivalent, and so we can say $\lim\limits_{n\to\infty} T^n = 0$. Now define $S_n = I + T + \cdots + T^n$. Then

$$\lim_{n\to\infty} (I-T)S_n = \lim_{n\to\infty} (I - T^{n+1}) = I, \tag{9}$$

and we conclude (7) holds, as desired. □

Lemma: Pick any $x^0 \in \mathbb{R}^n$. Then the sequence $\{x^k\}_{k=0}^{\infty}$ defined for $k \geq 1$ by

$$x^k = Tx^{k-1} + c \tag{10}$$

converges to the unique solution $x = Tx + c$ if and only if $\rho(T) < 1$. △

Proof:

Assume $\rho(T) < 1$. Then

$$x^k = Tx^{k-1} + c = T^k x^0 + \sum_{j=0}^{k-1} T^j c. \tag{11}$$

Because $\rho(T) < 1$, T is convergent and so $\lim\limits_{n \to \infty} T^n x^0 = 0$. Then our above lemma implies

$$\lim_{n \to \infty} x^n = \lim_{n \to \infty} T^k x^0 + \left(\sum_{j=0}^{\infty} \right) c = 0 + (I - T)^{-1} c = (I - T)^{-1} c. \tag{12}$$

This shows $x^n \longrightarrow \overline{x}$ where $\overline{x} = (I - T)^{-1} c$, which is equivalently written as $\overline{x} = T\overline{x} + c$. Uniqueness of $\overline{x}$ follows from the fact $(I - T)^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is a bijection.

Conversely, assume $\rho(T) < 1$. Then T is convergent and so $\lim\limits_{n \to \infty} T^n z = 0$ for each $z \in \mathbb{R}^n$. Pick any $x^0 \in \mathbb{R}^n$ and set $z = \overline{x} - x^0$ so that $x^0 = \overline{x} - z$. Then

$$\overline{x} - x^n = (T\overline{x} + c) - \left(Tx^{n-1} + c\right) = T\left(x - x^{n-1}\right) = T^n\left(x - x^0\right) = T^n z. \tag{13}$$

Thus, taking the limit as $n \longrightarrow \infty$, we obtain $\lim\limits_{n \to \infty} \overline{x} - x^n = \lim\limits_{n \to \infty} T^n z = 0$, which implies $x^n \longrightarrow \overline{x}$. This completes the proof. □

4 Old Numerical Qual Solutions

Fall 2002

F02.03: Let $A \in \mathbb{R}^{n\times n}$ be invertible and consider iterative methods of the form

$$Mx^{n+1} = Nx^n + b \tag{14}$$

where $A = M - N$.

a) Assuming M is non-singular, state a sufficient condition that ensures convergence of the iterates to the solution of $Ax = b$ for any starting vector x^0.

b) Describe matrices M and N for

   i) Jacobi's iteration

   ii) Gauss-Seidel iteration

c) If A is strictly diagonally dominant, prove that Jacobi's method converges.

Proof:

a) The spectral radius of $M^{-1}N$ is less than one, i.e., $\rho(M^{-1}N) < 1$.

b) i) For Jacobi's iteration $M = D$ and $N = D - A$ where D is the matrix containing the diagonal entries of A.

   ii) Write the matrix A as the sum of its lower triangular, diagonal, and upper triangular elements, i.e., write $A = D - L - U$ where L is strictly lower triangular, U is strictly upper triangular, and D is diagonal. Then Gauss-Seidel's iteration, $M = D - L$ and $N = U$

c) By the definition of strictly diagonally dominant,

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \quad \Longleftrightarrow \quad \sum_{j\neq i} \left|\frac{a_{ij}}{a_{ii}}\right| < 1 \quad \text{for } i = 1, \dots, n. \tag{15}$$

Now observe

$$\|M^{-1}N\|_\infty = \sup\left\{\|M^{-1}Nx \mid \|x\|_\infty = 1\right\} = \sup_{i=1,\dots,n} \left|M^{-1}Nx\right|_i = \sup_{i=1,\dots,n} \sum_{j\neq i} \left|\frac{a_{ij}}{a_{ii}}\right| < 1. \tag{16}$$

This shows $\|M^{-1}N\|_\infty < 1$. By definition, the spectral radius satisfies $\rho(M^{-1}N) \leq \|M^{-1}N\|_\infty$ and thus we conclude $\rho(M^{-1}N) < 1$. This shows our stated condition in a) is satisfied and we are done.

□

Fall 2003

F03.06: Consider the one-dimensional diffusion equations

$$\frac{\partial v}{\partial t} = \alpha \frac{\partial^2 v}{\partial x^2}, \quad \alpha > 0 \tag{17}$$

to be solved for $0 \le x \le 1$, $t > 0$, with periodic boundary conditions in x and the initial data $v(x,0) = v_0(x)$. Assume one uses the Dufort Frankel method

$$\frac{v_m^{n+1} - v_m^{n-1}}{2\Delta t} = \alpha \left(\frac{v_{m+1}^n - (v_m^{n+1} + v_m^{n-1}) + v_{m-1}^n}{\Delta x^2} \right) \tag{18}$$

as a means of computing approximate solutions to this equation.

a) Determine the truncation error associated with this approximation. Under what conditions does the scheme provide a consistent approximation to the diffusion equation? Would the condition required for consistency be difficult to satisfy in a set of computational experiments where Δx is repeatedly halved?

b) Surprisingly, this scheme is explicit and unconditionally stable. Show this, and explain why this does not violate the CFL condition.

Solution:

a) We use multiple Taylor expansions to derive the truncation error. Let $k := \Delta t$ and $h := \Delta x$. First observe

$$v_m^{n\pm1} = v_m^n \pm k(v_m^n)_t + \frac{k^2}{2}(v_m^n)_{tt} \pm \frac{k^3}{6}(v_m^n)_{ttt} + \frac{k^4}{24}(v_m^n)_{tttt} \pm \frac{k^5}{120}(v_m^n)_{ttttt} + \mathcal{O}(k^6). \tag{19}$$

This implies

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = (v_m^n)_t + \frac{k^2}{6}(v_m^n)_{ttt} + \mathcal{O}(k^4) \quad \text{and} \quad v_m^{n+1} + v_m^{n-1} = 2v_m^n + k^2(v_m^n)_{tt} + \mathcal{O}(k^4). \tag{20}$$

Using an analogous Taylor expansion for $v_{m\pm1}^n$ as in (19), we can combine the expansions to obtain

$$\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} = (v_m^n)_{xx} + \frac{h^2}{12}(v_m^n)_{xxxx} + \mathcal{O}(h^3). \tag{21}$$

Together (20) and (21) imply

$$\begin{aligned}\frac{v_{m+1}^n - (v_m^{n+1}+v_m^{n-1}) + v_{m-1}^n}{h^2} &= \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} - \frac{k^2}{h^2}(v_m^n)_{tt} + \mathcal{O}(k^4/h^2) \\ &= (v_m^n)_{xx} + \frac{h^2}{12}(v_m^n)_{xxxx} - \frac{k^2}{h^2}(v_m^n)_{tt} + \mathcal{O}(k^4/h^2 + h^3).\end{aligned} \tag{22}$$

Let $P := \partial_t - \alpha\partial_{xx}$ be the differential operator and $P_{k,h}$ be the analogous differential operator defined by the Dufort Frankel method. Then (19) and (22) together show the truncation error τ_m^n is

$$\begin{aligned}\tau_m^n := (P - P_{k,h})v_m^n &= \frac{h^2}{12}(v_m^n)_{xxxx} - \frac{k^2}{h^2}(v_m^n)_{tt} - \frac{k^2}{6}(v_m^n)_{ttt} + \mathcal{O}(k^4/h^2 + h^3 + k^4) \\ &= \mathcal{O}(h^2 + k^2 + (k/h)^2).\end{aligned} \tag{23}$$

We see $(P - P_{k,h})v_m^n \longrightarrow 0$ as $k, h \longrightarrow 0$ provided $k/h \longrightarrow 0$ as well. Thus the scheme is consistent provided $k/h \longrightarrow 0$ as $k, h \longrightarrow 0$.

Lastly, we claim the required condition for consistent would not be difficult to satisfy. For instance, if h is halved at each successive experiment, it would suffice to have k quartered. Let $\{k_n\}$ and $\{h_n\}$ give the time and space step sizes respectively. Then, with this quartering and halving, respectively, we see $k_n, h_n \longrightarrow 0$ as $n \longrightarrow \infty$ and

$$\lim_{n\to\infty} \frac{k_n}{h_n} = \lim_{n\to\infty} \frac{k_0/4^n}{h_0/2^n} = \frac{k_0}{h_0} \lim_{n\to\infty} \frac{1}{2^n} = 0, \tag{24}$$

as desired.

b) We first rewrite the scheme as

$$(1 + 2b\mu)v_m^{n+1} = 2b\mu(v_{m+1}^n + v_{m-1}^n) + (1 - 2b\mu)v_m^{n-1} \tag{25}$$

where $\mu = \Delta t/\Delta x^2$. Using Von Neumann analysis to determine the stability, we replace v_m^n with $g^n e^{im\theta}$. This gives

$$(1 + 2b\mu)g^2 - 4b\mu\cos\theta g - (1 - 2b\mu) = 0, \tag{26}$$

which implies the solutions g_+ and g_- are given by

$$g_\pm = \frac{2b\mu\cos\theta \pm \sqrt{1 - 4b^2\mu^2\sin^2\theta}}{1 + 2b\mu}. \tag{27}$$

If $1 - b^2\mu^2\sin^2\theta \geq 0$, then

$$|g_\pm| \leq \frac{2b\mu|\cos\theta| + \sqrt{1 - b^2\mu^2\sin^2\theta}}{1 + 2b\mu} \leq \frac{2b\mu + 1}{1 + 2b\mu} = 1. \tag{28}$$

Alternatively, $1 - 4b^2\mu^2\sin^2\theta < 0$ implies

$$|g_\pm|^2 = \frac{(2b\mu\cos\theta)^2 + 4b^2\mu^2\sin^2\theta - 1}{(1 + 2b\mu)^2} = \frac{4b^2\mu^2 - 1}{4b^2\mu^2 + 4b\mu + 1} < 1. \tag{29}$$

In any case, $|g_\pm| \leq 1$, and so we deduce the scheme is stable. Moreover, because we introduced no constraint on μ, the scheme is unconditionally stable.

The CFL condition is associated with a theorem stating there are no explicit unconditionally stable schemes for hyperbolic equations. Since one-dimensional diffusion equations are parabolic, we see the CFL condition is not directly applicable for this scheme.

□

Winter 2004

W04.01: Let $g(x)$ be continuously differentiable and consider the fixed point problem of finding x such that $x = g(x)$.

a) What conditions on $g(x)$ and α, $0 < \alpha \leq 1$, will guarantee convergence of the iteration

$$x_{n+1} = \alpha g(x_n) + (1-\alpha)x_n \tag{30}$$

to a solution $\overline{x}$ of the fixed point problem?

b) Prove that, under the conditions you derived in a), the solution $\overline{x}$ of the fixed point problem is unique.

Solution:

a) For each index n, Taylor's theorem asserts there is ξ_n along the line segment between x_n and $\overline{x}$ such that

$$\overline{x} = g(\overline{x}) = g(x_n) + g'(\xi_n)\cdot(\overline{x} - x_n) \qquad \Rightarrow \qquad g(x_n) = \overline{x} - g'(\xi_n)(\overline{x} - x_n). \tag{31}$$

Then, using the definition of our iteration, we discover

$$|x_{n+1} - x_n| = |\alpha g(x_n) + (1-\alpha)x_n - x_n| = \alpha\,|g(x_n) - x_n| = \alpha\,\big|1 - g'(\xi_n)\big|\,|\overline{x} - x_n| \tag{32}$$

Assume $\boxed{\|g'\|_\infty \in (0,2)}$. Then $\|1 - g'\|_\infty \in (0,1)$ and

$$|x_{n+1} - x_n| \leq \alpha\|1 - g'\|_\infty\,|\overline{x} - x_n| \leq \|1 - g'\|_\infty\,|\overline{x} - x_n|\,. \tag{33}$$

Through induction, we discover

$$|x_{n+1} - x_n| \leq \|1 - g'\|_\infty^n\,|\overline{x} - x_0|\,. \tag{34}$$

Since $\|1 - g'\|_\infty \in (0,1)$, taking the limit as $n \longrightarrow \infty$, the above shows $x_n \longrightarrow \overline{x}$, as desired.

b) Suppose $\overline{x}$ and α are fixed points of g. By our results in a), we deduce $x_n \longrightarrow \alpha$ and $x_n \longrightarrow \overline{x}$. Let $\varepsilon > 0$ be given. Then there is a positive index n such that $|x_n - \overline{x}| < \varepsilon$ and $|x_n - \alpha| < \varepsilon$. With the triangle inequality, this implies

$$|\alpha - \overline{x}| = |\alpha - x_n + x_n - \overline{x}| \leq |\alpha - x_n| + |x_n - \overline{x}| < \varepsilon + \varepsilon = 2\varepsilon. \tag{35}$$

We thus have $|\alpha - \overline{x}| < 2\varepsilon$. But, $\varepsilon > 0$ was arbitrarily chosen and so, taking the limit as $\varepsilon \longrightarrow 0$, we conclude $\alpha = \overline{x}$. Thus the fixed point of g is unique.

□

W04.05: Consider the hyperbolic equation

$$u_t + u_x + 2u_y = 0 \tag{36}$$

for $t > 0$ and $(x, y) \in [-1, 1] \times [-1, 1]$, and the initial data $u(x, y, 0) = \phi(x, y)$.

a) Boundary conditions on u are imposed to be zero on which sides of the square? Why?

b) Set up a finite difference approximation which converges to the correct solution. Justify your answer.

Solution:

a) Boundary conditions on the sides of the square corresponding to $x = -1$ and $y = -1$ are imposed because the waves move in the directions of the positive x and positive y axes.

b) We propose using the upwind scheme given by

$$\frac{u^{n+1}_{\ell,m} - u^n_{\ell,m}}{k} + \frac{u^n_{\ell,m} - u^n_{\ell-1,m}}{h_x} + 2\left(\frac{u^n_{\ell,m} - u^n_{\ell,m-1}}{h_y}\right) = 0. \tag{37}$$

The discrete operator P_{k,h_x,h_y} is defined by the left hand side of (37) and $R_{k,h_x,h_y} = f^n_{\ell,m}$. This gives the corresponding symbols

$$\begin{aligned} r_{k,h_x,h_y}(s,\xi,\eta) &= 1, \\ p(s,\xi,\eta) &= s + i\xi + 2i\eta, \\ p_{k,h_x,h_y}(s,\xi,\eta) &= \frac{1}{k}\left(e^{sk} - 1\right) + \frac{1}{h_x}\left(1 - e^{-i\xi h_x}\right) + \frac{2}{h_y}\left(1 - e^{-i\eta h_h}\right). \end{aligned} \tag{38}$$

Using the Taylor expansion for exponentials, we discover

$$\begin{aligned} p_{k,h_x,h_y} - r_{k,h_x,h_y}p &= \left([s + \mathcal{O}(k)] + \frac{1}{h_x}[i\xi h_x + \mathcal{O}(h_x^2)] + \frac{2}{h_y}[i\eta h_y + \mathcal{O}(h_y^2)]\right) - (s + i\xi + 2i\eta) \\ &= \mathcal{O}(k) + \mathcal{O}(h_x) + \mathcal{O}(h_y). \end{aligned} \tag{39}$$

This shows the method is first order. Hence $|p_{k,h_x,h_y} - r_{k,h_x,h_y}p| \longrightarrow 0$ as $k, h_x, h_y \longrightarrow 0$, i.e., the method is consistent. The Lax equivalence theorem states a consistent method converges if and only if it is stable. So, it suffices to show the method is stable. Rearranging (37), we see

$$\begin{aligned} |u_{\ell,m}^{n+1}| &= \left| u_{\ell,m}^n - \frac{k}{h_x}\left(u_{\ell,m}^n - u_{\ell-1,m}^n\right) - \frac{2k}{h_y}\left(u_{\ell,m}^n - u_{\ell,m-1}^n\right) \right| \\ &\le \left|1 - \frac{k}{h_x} - \frac{2k}{h_y}\right| |u_{\ell,m}^n| + \frac{k}{h_x}|u_{\ell-1,m}^n| + \frac{2k}{h_y}|u_{\ell,m-1}^n| \\ &\le \left(\left|1 - \frac{k}{h_x} - \frac{2k}{h_y}\right| + \frac{k}{h_x} + \frac{2k}{h_y} \right) \|u^n\|_\infty. \end{aligned} \tag{40}$$

The first inequality holds by the triangle inequality and the second by the definition of the sup norm. Since the right hand side has no dependence on ℓ and m, we may take the supremum over both sides to obtain

$$\|u^{n+1}\|_\infty \le \left(\left|1 - \frac{k}{h_x} - \frac{2k}{h_y}\right| + \frac{k}{h_x} + \frac{2k}{h_y} \right) \|u^n\|_\infty. \tag{41}$$

This shows $\|u^{n+1}\|_\infty \le \|u^{n+1}\|_\infty$ precisely when $k/h_x + 2k/h_y \le 1$. Thus, when this condition is met, we conclude the method is stable, and we are done.

□

Fall 2004

F04.01: Let $\overline{x}$ be a root of a continuous differentiable function $f : \mathbb{R} \to \mathbb{R}$. If x^* is an approximate root, then

a) Derive an expression that relates the magnitude of the residual at x^* to the magnitude of the error of the root x^*.

b) Give an example of a function where the magnitude of the residual at x^* over-estimates the error of the root x^*.

c) Give an example of a function where the magnitude of the residual at x^* under-estimates the error of the root x^*.

Solution:

a) By Taylor's theorem, there is ξ between $\overline{x}$ and x^* such that

$$f(x^*) = f(\overline{x}) + f(\xi)(x^* - \overline{x}) \qquad \Rightarrow \qquad |f(x^*) - f(\overline{x})| = f(\xi)\,|x^* - \overline{x}|\,. \tag{42}$$

The left hand side of the resulting equality is the residual and the right hand side is $f(\xi)$ multiplied by the error of the root x^*.

b) Consider $f(x) = 2x$. Then $f' = 2$ and so the residual is twice the error of the root x^*.

c) Consider the function $f(x) = x/2$. Then $f' = 1/2$ and so the residual is half the error of the root x^*.

□

Winter 2005

W05.01: Let $f(x) = \cos(x) - x$.

a) Prove $f(x)$ has exactly one root in the interval $[0, \pi/2]$.

b) Give a good estimate of the minimum number of bisection iterations requried to obtain an approximation that is within $10^{-6} \cdot \pi/2$ of this root when the initial interval used is $[0, \pi/2]$.

Solution:

a) First observe that $f(0) = \cos(0) - 0 = 1 - 0 = 1 > 0$ and $f(\pi/2) = cos(\pi/2) - \pi/2 = 0 - \pi/2 < 0$. Since f is continuous, it follows from the intermediate value theorem that there is a root of f in $[0, \pi/2]$. We claim $f' < 0$ in $[0, \pi/2]$. Indeed, for each $x \in [0, \pi/2]$, $\sin(x) \geq 0$ and so

$$f'(x) = -\sin(x) - 1 \leq 0 - 1 < 0 \tag{43}$$

Now, by way of contradiction, suppose f has distinct roots $x_1, x_2 \in [0, \pi/2]$. Then $f(x_1) = -f(x_2)$ and, because f is differentable, Rolle's theorem asserts there is c between x_1 and x_2 such that $f'(c) = 0$. However, this contradicts the fact $f' < 0$ in $[0, \pi/2]$. Thus f has exactly one root in $[0, \pi/2]$, denoted x^*.

b) The bisection method produces a sequence $\{x_n\}_{n=0}^{\infty}$ where

$$|x_n - x^*| \leq \frac{\pi/2 - 0}{2^{n+1}} = \frac{\pi}{2} \cdot 10^{-6} \tag{44}$$

for each index n. We make the approximation $2^{10} = 1024 \approx 1000 = 10^3$. This implies

$$\frac{\pi}{2} \cdot 10^{-6} \approx \frac{\pi}{2} \cdot \left(10^3\right)^{-2} \approx \frac{\pi}{2} \cdot \left(2^{10}\right)^{-2} = \frac{\pi}{2} \cdot 2^{-20}. \tag{45}$$

Whence we conclude by (44) and (45) that the approximate minimum number of iterations to obtain an estimate x^n within the desired error is roughly $\boxed{\text{19 iterations.}}$

□

W05.03: Let $A \in \mathbb{R}^{n\times n}$ be non-singular and consider iterative methods of the form

$$Mx^{n+1} = b + Nx^n \tag{46}$$

where $A = M - N$.

a) Assuming M is non-singular, state a sufficient condition that insures convergence of the iterates to the solution of $Ax = b$ for any starting vector x^0.

b) Describe the matrices M and N for i) Jacobi iteration and ii) Gauss-Siedel iteration.

c) If A is strictly diagonally dominant, prove Jacobi's method converges.

Solution:

a) A sufficient condition is that $\rho(M^{-1}N) < 1$ where ρ gives the spectral radius. Assume $Ax^* = b$ so that

$$(M - N)x^* = b \qquad \Rightarrow \qquad x^* = M^{-1}Nx^* + M^{-1}b. \tag{47}$$

Then, using the relation,

$$x^* - x^{n+1} = M^{-1}N(x^* - x^n) = \cdots = \left(M^{-1}N\right)^n (x^* - x^0). \tag{48}$$

Thus

$$\|x^* - x^{n+1}\| = \|\left(M^{-1}N\right)^n (x^* - x^0)\| \leq \|M^{-1}N\|^n \|x^* - x^*\|. \tag{49}$$

By hypothesis, $\|M^{-1}N\| < 1$ and so, taking the limit as $n \longrightarrow \infty$, the right hand side of (49) goes to zero, thereby establishing the convergence of $\{x^n\}_{n=0}^{\infty}$ to x^*.

b) Let $A = A_L + D + A_R$ where A_L and A_R contain the strictly lower and upper triangular elements of A, respectively, and D contains the diagonal elements of A.

   i) For the Jacobi iteration we take $M = D$ and $N = -A_L - A_R$.

   ii) For the Gauss-Siedel iteration we take $M = A_L + D$ and $N = -A_R$.

c) From a), it suffices to show that if A is strictly diagonally dominant, then $\rho(D^{-1}(A_L + A_R)) < 1$. Let

(λ, v) be an eigenvalue/eigenvector pair for $D^{-1}(A_L + A_R)$. Then $v \neq 0$ and

$$|(\lambda v)_i| = \left|\left(D^{-1}(A_L + A_R)v\right)_i\right| = \left|\frac{1}{a_{ii}}\sum_{i \neq j} a_{ij} v_j\right| \leq \frac{1}{|a_{ii}|}\sum_{i \neq j} |a_{ij}| \|v\|_\infty < \|v\|_\infty, \tag{50}$$

where the final inequality holds since A is strictly diagonally dominant. Since this holds for each i, we deduce

$$|\lambda| \|v\|_\infty = \|\lambda v\|_\infty < \|v\|_\infty \quad \Rightarrow \quad |\lambda| < 1. \tag{51}$$

Since λ was an arbitrary eigenvalue, we conclude $\rho(D^{-1}(A_L + A_R)) < 1$, as desired.

□

W05.07: Consider the boundary value problem

$$\begin{aligned} -\Delta u + u &= f(x,y), \quad (x,y) \in \Omega = [0,1] \times [0,1], \\ u &= 0, \quad (x,y) \in \partial\Omega, \ x = 0,1, \\ u_y &= 0, \quad (u,xy) \in \partial\Omega, \ y = 0,1. \end{aligned} \tag{52}$$

a) Give a weak variational formulation of this problem.

b) Analyze the existence and uniqueness of the solution to this problem. Justify your answer and assume $f \in L^2(\Omega)$.

c) Formulate a finite element approximation of the elliptic problem using piecewise-linear elements. Discuss the form and properties of the stiffness matrix and the existence and uniqueness of the solution of the linear system thus obtained. Justify your answers.

Solution:

a) Set $\Gamma_1 := \{(x,y) \in \partial\Omega \ : \ x \in \{0,1\}\}$ and $\Gamma_2 := \{(x,y) \in \partial\Omega \ : \ y \in \{0,1\}\}$. Then define the Hilbert space $H := \{u \in H^1(\Omega) \ : \ u = 0 \text{ on } \Gamma_1\}$. Letting $v \in H$ be a test function, we discover, for a solution u,

$$\int_\Omega fv = \int_\Omega -\Delta uv + uv. \tag{53}$$

Integrating by parts, we obtain

$$\underbrace{\int_\Omega fv}_{\ell(v)} = \int_\Omega \nabla u \cdot \nabla v + uv - \cancelto{0}{\int_{\Gamma_1} v(n \cdot \nabla u)} - \cancelto{0}{\int_{\Gamma_2} v(n \cdot \nabla u)} = \underbrace{\int_\Omega \nabla u \cdot \nabla v + uv}_{a(u,v)}, \tag{54}$$

where the integral over Γ_1 equals zero since $v = 0$ on Γ_1 and the integral over Γ_2 also equals zero since $\nabla u = 0$ on Γ_2. Define a and ℓ to be the underbraced quantities, noting a is bilinear and ℓ is linear. Then the weak variational formulation is

$$\text{Find } u \in H \text{ such that } a(u,v) = \ell(v) \ \ \forall \ v \in H. \tag{55}$$

b) We claim a unique solution exists to this problem. We show this by verifying the assumptions of the

Lax-Milgram theorem are satisfied. We must show a is bounded and coercive and ℓ is bounded. First note the triangle and Hölder's inequality imply

$$|a(u,v)| \leq \int_\Omega |\nabla u \cdot \nabla v + \int_\Omega |uv| \leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq 2\|u\|_H \|v\|_H, \tag{56}$$

where $\|u\|_H = \|u\|_{H^1(\Omega)} := \left(\int_\Omega |\nabla u|^2 + u^2\right)^{1/2}$. This shows a is bounded. And,

$$a(u,u) = \int_\Omega |\nabla u|^2 + u^2 = \|u\|_H^2 \tag{57}$$

shows a is coercive. Lastly,

$$|\ell(v)| = \left| \int_\Omega fv \right| \leq \int_\Omega |fv| = \|fv\|_{L^1(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_H. \tag{58}$$

Whence ℓ is bounded and the assumptions of the Lax-Milgram theorem are satisfied.

c) Let $\mathcal{T}_h$ be a triangulation of the domain Ω. Define $H_h := \{v \in C(\Omega) \ : \ v|_K \text{ is linear } \forall\ K \in \mathcal{T}_h\}$. Let n_i for $i = 1, \ldots, M$ be the nodes corresponding to $\mathcal{T}_h$. Let $\{\phi_i\}_{i=1}^M$ be the basis for H_h such that $\phi_i(n_j) = \delta_{ij}$ with δ_{ij} the Kronecker δ. Consider a solution $u = \sum_{i=1}^M \xi_i \phi_i$ where $\xi \in \mathbb{R}^n$ is constant. Our finite element problem becomes

$$\text{Find } u_h \in H_h \text{ such that } a(u_h, v) = \ell(v) \ \ \forall\ v \in H_h. \tag{59}$$

The bilinearity of a implies

$$\ell(\phi_j) = a(u_h, \phi_j) = a\left(\sum_{i=1}^M \xi_i \phi_i, \phi_j\right) = \sum_{i=1}^M \xi_i a(\phi_i, \phi_j) \quad j = 1, \ldots, M. \tag{60}$$

This can be written as the matrix equation $A\xi = b$ where $A_{ij} := a(\phi_i, \phi_j)$ and $b_i := \ell(\phi_i)$. Note A is symmetric and

$$\xi \cdot A\xi = \sum_{i,j=1}^M \xi_i A_{i,j} \xi_j = a\left(\sum_{i=1}^M \xi_i \phi_i,\ \sum_{j=1}^M \xi_j \phi_j\right) = a(u_h, u_h) > 0, \tag{61}$$

where the final inequality holds since a is coercive. Because A is symmetric and positive definite, the system $A\xi = b$ has a unique solution. Moreover, A is sparse since $A_{i,j} = a(\phi_i, \phi_j) = 0$ whenever there does not exists $K \in \mathcal{T}_h$ such that $n_i, n_j \in K$.

$\square$

Spring 2007

S07.03: Consider the system of ODEs

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}_t = \begin{pmatrix} -4 & 1 \\ 1 & -4 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{62}$$

If y and z are any two solutions of this equation with distinct initial data, then $\|y - z\| \longrightarrow 0$ as $t \longrightarrow \infty$. Let y^n and z^n be approximate solutions obtained with Euler's method starting with distinct initial data at time t_0. What time step is required to ensure $\|y^n - z^n\| \longrightarrow 0$ as $n \longrightarrow \infty$? Justify your result.

Solution:

Let A be the matrix so that $y_t = Ay$. Then the Forward Euler (FE) method is defined so that

$$y^{n+1} = y^n + h(y^n)' = y^n + hAy^n = (I + hA)y^n. \tag{63}$$

This implies

$$\|y^{n+1} - z^{n+1}\| = \|(I + hA)(y^n - z^n)\| \leq \|I + hA\|\|y^n - z^n\| \leq \cdots \leq \|I + hA\|^n\|y^0 - z^0\|. \tag{64}$$

where all norms are the two norm. If $\|I + hA\| < 1$, then $\|I + hA\|^n \longrightarrow 0$ and we obtain $\|y^{n+1} - z^{n+1}\| \longrightarrow 0$, as desired. All that remains is to identify h such that $\|I + hA\| < 1$. Since $I + hA$ is symmetric, its two norm is the absolute value of its maximum eigenvalue. So, we compute the characteristics polynomial

$$\begin{aligned} \chi(\lambda) &= \det(\lambda I - (I + hA)) \\ &= \begin{vmatrix} \lambda - (1 - 4h) & -h \\ -h & \lambda - (1 - 4h) \end{vmatrix} \\ &= [\lambda - (1 - 4h)]^2 + h^2 \\ &= \lambda^2 - 2(1 - 4h)\lambda + (1 - 4h)^2 - h^2. \end{aligned} \tag{65}$$

Using the quadratic formula, we discover the eigenvalues are

$$\lambda = \frac{2(1 - 4h) \pm \sqrt{4(1 - 4h)^2 - 4[(1 - 4h)^2 - h^2]}}{2} = 1 - 4h \pm h. \tag{66}$$

We need $|1 - 4h \pm h| < 1$. For both eigenvalues, we have$1 - 3h, 1 - 5h < 1$ since $h > 0$. And, $1 - 3h > 1 - 5h > -1$ implies $h < 2/5$. Thus to ensure the desired convergence we need $\boxed{h \in (0, 2/5).}$ □

S07.05: Consider the convection-diffusion equation

$$u_t + au_x = bu_{xx}, \tag{67}$$

with $a \neq 0$, $b > 0$ as constants, to be solved for $t > 0$, $0 \leq x \leq 1$, with $u(x,t)$ periodic in x and $u(x,0)$ given.

a) Construct an explicit second order scheme of the form

$$u_i^{n+1} + u_i^n + c_2 u_{i+2}^n + c_1 u_{i+1}^n + c_0 u_i^n + c_{-1} u_{i-1}^n + c_{-2} u_{i-2}^n \tag{68}$$

by using the Lax-Wendroff procedure, i.e., the procedure where one uses the Taylor series expansion

$$u(x, t+k) = u(x,t) + ku_t + \frac{k^2}{2} u_{tt} + \mathcal{O}(k^3) \tag{69}$$

and replaces the t derivatives by x derivatives using the equation.

b) Derive stability conditions involving k and h. Justify your statements.

Solution:

a) First observe

$$u_t = bu_{xx} - au_x \qquad \Rightarrow \qquad u_t = (bu_{xx} - au_x)_t = b(u_t)_{xx} - a(u_t)_x = b^2 u_{xxxx} - 2ab u_{xxx} + a^2 u_{xx}. \tag{70}$$

Through substitution into our Taylor series, we discover

$$\begin{aligned} u_i^{n+1} &= u_i^n + k(u_i^n)_t + \frac{k^2}{2}(u_i^n)_{tt} + \mathcal{O}(k^3) \\ &= u_i^n + k\left(b(u_i^n)_{xx} - a(u_i^n)_x\right) + \frac{k^2}{2}\left(b^2(u_i^n)_{xxxx} - 2ab(u_i^n)_{xxx} + a^2(u_i^n)_{xx}\right) + \mathcal{O}(k^3) \\ &= u_i^n - ka(u_i^n)_x + \left(kb + \frac{k^2a^2}{2}\right)(u_i^n)_{xx} - k^2ab(u_i^n)_{xxx} + \frac{k^2b^2}{2}(u_i^n)_{xxxx} + \mathcal{O}(k^3). \end{aligned} \tag{71}$$

We then use second order central difference approximations for each of these derivatives to obtain the desired scheme. That is, we set

$$\begin{aligned}
u_i^{n+1} &= u_i^n - ka\left(\frac{u_{i+1}^n - u_{i-1}^n}{2h}\right) + \left(kb + \frac{k^2a^2}{2}\right)\left(\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}\right) \\
&\quad - k^2ab\left(\frac{u_{i+2}^n - 2u_{i+1}^n + 2u_{i-1}^n - u_{i-2}^n}{2h^3}\right) + \frac{k^2b^2}{2}\left(\frac{u_{i+2}^n - 4u_{i+1}^n + 6u_i^n - 4u_{i-1}^n + u_{i-2}^n}{h^4}\right) \\
&= u_i^n + \left(-\frac{k^2ab}{2h^3} + \frac{k^2b^2}{2h^4}\right)u_{i+2}^n + \left(-\frac{ka}{2h} + \frac{2kb + k^2a^2}{2h^2} + \frac{k^2ab}{h^3} - \frac{2k^2b^2}{h^4}\right)u_{i+1}^n \\
&\quad + \left(\frac{-2kb - k^2a^2}{h^2} + \frac{3k^2b^2}{h^4}\right)u_i^n + \left(\frac{ka}{2h} + \frac{2kb + k^2a^2}{2h^2} - \frac{k^2ab}{h^3} - \frac{2k^2b^2}{h^4}\right)u_{i-1}^n \\
&\quad + \left(\frac{k^2ab}{2h^3} + \frac{k^2b^2}{2h^4}\right)u_{i-2}^n.
\end{aligned} \tag{72}$$

The form of the second equality enables us to directly write off $c_2, c_1, c_0, c_{-1}, c_{-2}$, which we do not do here for sake of space.

b) We proceed using Von Neumann analysis, substituting $g^n e^{im\theta}$ for u_m^n. Doing so and then canceling common terms, we obtain

$$g = 1 + c_2e^{2i\theta} + c_1e^{i\theta} + c_0 + c_{-1}e^{-i\theta} + c_{-2}e^{-2i\theta}. \tag{73}$$

To obtain stability, we need our amplification factor g to satisfy $|g| \leq 1$. Going further and plugging in $c_2, \ldots, c_{-2}$ to find the stability conditions involving k and h explicitly would take longer than I have time for...

□

Fall 2008

F08.04: Consider the ordinary differential equation $\mathrm{d}y/\mathrm{d}t = f(t,y)$.

a) Give a derivation of the Taylor series method that is of global second order accuracy.

b) What is the interval of absolute stability for this method? Justify your answer.

Solution:

a) We claim the desired Taylor series is given by the sequence $\{w_i\}$ where

$$w_{n+1} = w_n + hf(t_n, w_n) + \frac{h^2}{2}(f_t + f_y f)(t_n, w_n). \tag{74}$$

We verify this as follows. Let $y_n := y(t_n)$. Then we Taylor expand about y_n to discover

$$y_{n+1} = y_n + hy_n' + \frac{h^2}{2}y_n'' + \frac{h^3}{6}y_n''' + \mathcal{O}(h^4). \tag{75}$$

Note $y_n' = f(t_n, y_n)$ and $y_n'' = \mathrm{d}f(t_n, y_n)/\mathrm{d}t = (f_t + f_y f)(t_n, y_n)$. Whence, if the current information is exact, then the iterative step w_{n+1} introduces the local truncation error τ_{n+1} given by

$$\tau_{n+1} := y_{n+1} - w_{n+1} = \frac{h^3}{6}y_n''' + \mathcal{O}(h^4). \tag{76}$$

Thus $\tau_n/h = \mathcal{O}(h^2)$ and we conclude the method is $\mathcal{O}(h^2)$, as desired.

b) To compute the region of stability, we consider the test equation $f(t,y) = \lambda y$ with $\lambda \in \mathbb{C}$. We must identify h such that $w_{n+1} \longrightarrow 0$ for all initial iterates w_0. First note that for the test equation

$$w_{n+1} = w_n + \lambda h w_n + \frac{h^2\lambda^2}{2}w_n = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right)w_n. \tag{77}$$

We will be done if we identify $h\lambda$ such that $|w_{n+1}| = \alpha|w_n|$ for some $\alpha \in (0,1)$. Thus we obtain

$$1 > \left|1 + h\lambda + \frac{(h\lambda)^2}{2}\right| = \frac{1}{2} + \frac{1}{2}(1 + h\lambda)^2 \quad \Rightarrow \quad |1 + h\lambda| < 1. \tag{78}$$

Thus the region of absolute stability is the unit circle in the complex plane centered at $(-1,0)$, i.e., the set $\{h\lambda \; : \; |1 + h\lambda| < 1\}$. This implies the interval of absolute stability is $(-2,0)$. □

F08.05: Consider the trapezoidal (TM) method and backward Euler method (BE)for the ordinary differential equation $\mathrm{d}y/\mathrm{d}t = f(y)$,

$$\text{(TM)} \quad y^{k+1} = y^k + \frac{h}{2}\left(f(y^k) + f(y^{k+1})\right) \quad \text{and} \quad \text{(BE)} \quad y^{k+1} = y^k + hf(y^{k+1}). \tag{79}$$

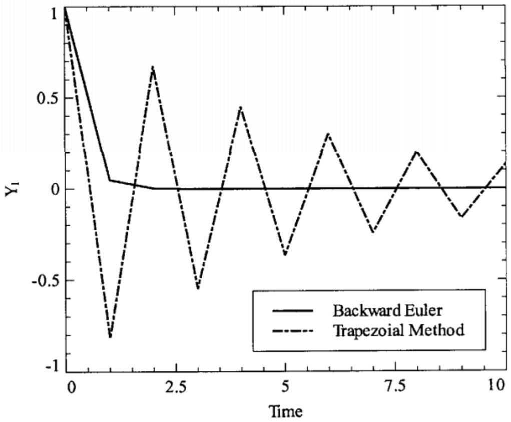
a) Show that, for each of these methods, the interval $(-\infty, 0)$ is contained with its interval of absolute stability.

b) If these methods are applied to the following system of ordinary differential equations

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \begin{pmatrix} -11 & -9 \\ -9 & -11 \end{pmatrix}, \quad y(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{80}$$

with $h = 1.0$, then, as shown in the plot below, the value of the first component of the numerical solution with the TM exhibits undesirable oscillations while the first component of the numerical solution obtained with BE doesn't. Explain the presence of the oscillations in the first solution obtained with the TM and the absence of oscillations obtained with BE. Justify your explanation.

c) For what value of the time step will the solution obtained with the TM cease to given an oscillatory solution? Justify your result.



Solution:

a) To compute the region of absolute stability, we assume $f(y) = \lambda y$ for $\lambda \in \mathbb{C}$. We must identify h

such that $y^k \longrightarrow 0$ for all initial iterates y^0. This will be accomplished if we identify $h\lambda$ such that $|y^{k+1}| = \alpha|y^k|$ for some $\alpha \in (0,1)$. Applying the TM with the test equation, we find

$$y^{k+1} = y^k + \frac{h\lambda}{2}\left(y^k + y^{k+1}\right) \qquad \Rightarrow \qquad y^{k+1} = \frac{1+h\lambda/2}{1-h\lambda/2}y^k. \tag{81}$$

This implies the absolute stability region is the set

$$\left\{h\lambda \in \mathbb{C} \ : \ \left|\frac{1+h\lambda/2}{1-h\lambda/2}\right| < 1\right\}. \tag{82}$$

If $\text{Re}(h\lambda) < 0$, then $|1+h\lambda/2| < |1-h\lambda/2|$. This implies the region of absolute stability contains the left half plane of the complex plane, of which $(-\infty, 0)$ is a subset.

We now consider BE with the test equation. Here we see

$$y^{k+1} = y^k + h\lambda y^{k+1} \qquad \Rightarrow \qquad y^{k+1} = \frac{1}{1-h\lambda}y^k, \tag{83}$$

and so the absolute stability region for BE is

$$\left\{h\lambda \ : \ \left|\frac{1}{1-h\lambda}\right| < 1\right\} = \{h\lambda \ : \ \text{Re}(h\lambda) < 0\}. \tag{84}$$

Thus the BE also contains the left half of the complex plane, of which $(-\infty, 0)$ is a subset. This completes the solution.

b) We first diagonalize our matrix, which will be denoted by A. Observe that

$$\det(A - \lambda I) = (-11-\lambda)^2 - 81 = \lambda^2 + 22\lambda + 40 = (\lambda+20)(\lambda+2), \tag{85}$$

and so the eigenvalues of A are $\lambda_1 = -20$ and $\lambda_2 = -2$. These correspond to the eigenvectors $(1,1) \in \mathbb{R}^2$ and $(1,-1) \in \mathbb{R}^2$, respectively. This implies

$$A = \underbrace{\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}}_{P} \underbrace{\begin{pmatrix} -20 & 0 \\ 0 & -2 \end{pmatrix}}_{D} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{-1} = PDP^{-1} \tag{86}$$

where P and D denote the underbraced matrices. Set $w = P^{-1}y$. Then we obtain the decoupled differential equation $w' = Dw$ since

$$w' = P^{-1}y' = P^{-1}Ay = P^{-1}PDP^{-1}y = IDw = Dw. \tag{87}$$

So, with the TM, we obtain

$$w_1^{k+1} = \frac{1+h\lambda_1/2}{1-h\lambda_1/2}w_1^k = -\frac{9}{11}w_1^k \quad \text{and} \quad w_2^{k+1} = \frac{1+h\lambda_1/2}{1-h\lambda_1/2}w_2^k = 0w_2^k = 0. \tag{88}$$

This shows w_1^k flips sign with each successive time step and that $w_2^k = 0$. Using the definition of w, we discover

$$y_1^k = (Pw^k)_1 = 1w_1^k - 1w_2^k = w_1^k - 0 = w_1^k. \tag{89}$$

The results in (88) and (89) and the fact $y_1^0 = 1 \neq 0$ together imply $y_1^{k+1} = -9y_1^k/11 \neq 0$. This shows y_1^k will oscillate for the given choice of h, as illustrated in the plot.

Now we investigate the BE approach. There

$$w_1^{k+1} = \frac{1}{1-h\lambda_1}w_1^k = \frac{1}{21}w_1^k \quad \text{and} \quad w_2^{k+1} = \frac{1}{1-h\lambda_2}w_2^k = \frac{1}{3}w_2^k. \tag{90}$$

Because y_1^k is a linear combination of w_1^k and w_2^k and, as shown in (90), neither w_1^k nor w_2^k exhibit oscillatory behavior, the linear combination for y_1^k will not exhibit oscillatory behavior either. That is, the BE method does not exhibit any oscillatory behavior.

c) In order for the TM to cease to given an oscillatory solution, we need the sign of w_1^k and w_2^k to not flip between iterations. Returning to (88), this will be accomplished if

$$\frac{1+h\lambda_1/2}{1-h\lambda_1/2} = \frac{1-10h}{1+10h} > 0 \quad \text{and} \quad \frac{1+h\lambda_2/2}{1-h\lambda_2/2} = \frac{1-h}{1+h} > 0. \tag{91}$$

Since $h > 0$, this will be satisfied if $1-10h > 0$ and $1-h > 0$. This is equivalent to asserting $h < 1/10$ and $h < 1$. So, the TM will not exhibit oscillatory behavior when $\boxed{h < 1/10.}$

□

Spring 2009

S09.01: Give a derivation of the nodes and weights of a Gaussian quadrature formula $\sum_{i=1}^{n} c_i f(x_i)$ with $n = 2$ of highest order that approximates the integral $\int_{-1}^{1} f(x)\ \mathrm{d}x$. What is the order of the approximation?

Solution:

We have four degrees of freedom from the variables $c_1, c_2, x_1, x_2 \in \mathbb{R}$. This implies we can make the formula exact for $f(x) = x^k$ for $k = 0, 1, 2, 3$. For such a formula to be exact in these cases, Taylor's theorem implies the error must then be proportional to $f^{(4)}$, i.e., the approximation will be fourth order. All that remains is to identify these nodes and weights. First note that must each be nonzero since if any variable equals zero, then we only have two degrees of freedom. And, $x_1 \neq x_2$ since then the formula can equivalently be rewritten using only two variables. Integrating, we obtain

$$2 = \int_{-1}^{1} x^0\ \mathrm{d}x = c_1 + c_2, \tag{92}$$

$$0 = \int_{-1}^{1} x^0\ \mathrm{d}x = c_1x_1 + c_2x_2, \tag{93}$$

$$\frac{2}{3} = \int_{-1}^{1} x^0\ \mathrm{d}x = c_1x_1^2 + c_2x_2^2, \tag{94}$$

$$0 = \int_{-1}^{1} x^0\ \mathrm{d}x = c_1x_1^3 + c_2x^3. \tag{95}$$

Equation (93) implies $c_1x_2 = -c_2x_2$ and plugging this into (94) gives

$$0 = (-c_2x_2)x_1^2 + c_2x_2^3 = c_2x_2(x_2 - x_1)(x_2 + x_1). \tag{96}$$

Since $c_2, x_2 \neq 0$ and $x_2 \neq x_1$, we conclude $x_1 = -x_2$. Then (94) implies, using (92),

$$\frac{2}{3} = c_1x_1^2 + c_2(-x_1)^2 = (c_1 + c_2)x_1^2 = 2x_1^2 \quad \Rightarrow \quad x_1 = \pm\frac{1}{\sqrt{3}}. \tag{97}$$

And (93) shows $0 = c_1x_1 + c_2(-x_1) = x_1(c_1 - c_2)$, which implies $c_1 = c_2$. With (92), we see $c_1 = c_2 = 1$. Thus the quadrature formula is

$$\boxed{f(1/\sqrt{3}) + f(-1/\sqrt{3}),} \tag{98}$$

and we are done. □

S09.05: The corrected Heun's method is given by

$$\begin{aligned} y^p &= y^k + dt f(y^k), \\ y^C &= y^k + \frac{dt}{2}\left(f(y^k) + f(y^p)\right), \\ y^{k+1} &= y^k + \frac{dt}{2}\left(f(y^k) + f(y^C)\right), \end{aligned} \tag{99}$$

while the standard Heun's method consists of using only the first two of the above equations and setting $y^{k+1} = y^C$. Consider applying the corrected Heun's method to the model problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y, \quad y(0) = y_0. \tag{100}$$

a) Based on the difference equation that results, derive the leading term of the local truncation error associated with the corrected Heun's method applied to this model problem. Justify your answer.

b) How much smaller do you expect the error to be when using the corrected Huen's method instead of the standard Heun's method? Explain.

c) Recall that the interval of absolute stability for the standard Heun's method is $[-2, 0]$. Does the corrected Heun's method have a larger region of absolute stability? Justify your answer.

Solution:

a) We first expand the corrected Heun's method for the model problem. Letting $h := dt$, this gives

$$\begin{aligned} y^{k+1} &= y^k + \frac{h\lambda}{2}\left(y^k + y^C\right) \\ &= y^k + \frac{h\lambda}{2}\left(y^k + \left[y^k + \frac{k\lambda}{2}(y^k + y^p)\right]\right) \\ &= y^k + \frac{h\lambda}{2}\left(y^k + \left[y^k + \frac{h\lambda}{2}\left(y^k + y^k + h\lambda y^k\right)\right]\right) \\ &= \underbrace{\left(1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{4}\right)}_{\alpha_h} y^k. \end{aligned} \tag{101}$$

Let α_h be the underbraced expression. Taylor expanding about t^k, we obtain

$$\begin{aligned} y(t^{k+1}) &= y(t^k) + k\lambda y(t^k) + \frac{(h\lambda)^2}{2} y(t^k) + \frac{(h\lambda)^3}{6} y(t^k) + \mathcal{O}(h^4) \\ &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{6}\right) y(t^k) + \mathcal{O}(h^4). \end{aligned} \tag{102}$$

Taking $y^k = y(t^k)$, this implies the local truncation error τ^{k+1} is

$$\begin{aligned} \tau^{k+1} &:= y(t^{k+1}) - y^{k+1} \\ &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{6}\right) y(t^k) + \mathcal{O}(h^4) - \left(1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{4}\right) y(t^k) \\ &= -\frac{(h\lambda)^3}{12} y(t^k) + \mathcal{O}(h^4). \end{aligned} \tag{103}$$

The first term on the right hand side of the final line gives the desired leading term of the local truncation error.

b) Let $\{z^k\}$ be the sequence of iterates generated by the standard Heun's method. Then

$$z^{k+1} = z^k + \frac{h\lambda}{2}\left(z^k + z^p\right) = z^k + \frac{h\lambda}{2}\left(z^k + z^k + h\lambda z^k\right) = \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) z^k. \tag{104}$$

Taking $z^k = y(t^k)$, we see the local truncation error $\tilde{\tau}^{k+1}$ for this method is

$$\begin{aligned} \tilde{\tau}^{k+1} := y(t^{k+1}) - z^{k+1} &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{6}\right) y(t^k) + \mathcal{O}(h^4) - \left(1 + h\lambda + \frac{(h\lambda)^2}{2}\right) y(t^k) \\ &= \frac{(h\lambda)^3}{6} y(t^k) + \mathcal{O}(h^4). \end{aligned} \tag{105}$$

Assuming h is sufficiently small to say the error is well approximated by the leading terms, we see the ratio of truncation errors is

$$\frac{\tau^{k+1}}{\tilde{\tau}^{k+1}} \approx \frac{-(h\lambda)^3 y(t^k)/12}{(h\lambda)^3 y(t^k)/6} = -\frac{1}{2}. \tag{106}$$

This shows we expect the error using the corrected Heun's method to be half the magnitude of that obtained with the standard Heun's method.

c) No, the corrected Heun's method does not have a larger interval of absolute stability. Here we take the definition of absolute stability to be that for each $T > 0$ there is $C_T > 0$ such that $|y^k| \leq C_T|y^0|$ for $0 \leq t^k = kh \leq T$. Since $y^k = \alpha_h y^{k-1} = (\alpha_h)^k y^0$, the method is absolute stable if and only if

$|\alpha_h| \leq 1$. In other words, we need

$$-1 \leq \alpha_h = 1 + h\lambda + \frac{(h\lambda)^2}{2} + \frac{(h\lambda)^3}{4} \leq 1 \quad \Leftrightarrow \quad -2 \leq h\lambda\left(1 + \frac{h\lambda}{2} + \frac{(h\lambda)^2}{4}\right) \leq 0. \tag{107}$$

For the interval of absolute stability, we take $\lambda \in \mathbb{R}$. Then the method is unstable if $h\lambda > 0$ since this would imply $\alpha_h > 1$. Now let $\beta > 0$ and take $h\lambda = -(2+\beta)$. Then

$$\begin{aligned} h\lambda\left(1 + \frac{h\lambda}{2} + \frac{(h\lambda)^2}{4}\right) &= -(2+\beta)\left(1 - \frac{2+\beta}{2} + \frac{4+2\beta+\beta^2}{4}\right) \\ &= -(2+\beta)\left(1 + \frac{\beta^2}{4}\right) \\ &< -(2+\beta) \\ &< -2. \end{aligned} \tag{108}$$

Whence the interval of absolute stability, denoted R, for the corrected Heun's method excludes $(-\infty, 2)$. From this we conclude $R \subseteq [-2, 0]$ and so the interval of absolute stability for the corrected Heun's method is not larger than the standard Heun's method.

□

Spring 2010

S10.01: Let $S(x)$ be a cubic spline with knots $t_0, t_1, t_2, \ldots, t_n$. If it is determined that $S(x)$ is linear over $[t_1, t_2]$ and $[t_3, t_4]$, what can be said about $S(x)$ over $[t_2, t_3]$?

Solution:

We claim S is also linear over $[t_2, t_3]$. First define $p : \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} p(x) = &\frac{S''(t_3)(x-t_2)^3}{6(t_3-t_2)} + \frac{S''(t_2)(t_3-x)^3}{6(t_3-t_2)} \\ &+ \left[\frac{S(t_3)}{t_3-t_2} - \frac{S''(t_3)(t_3-t_2)}{6}\right](x-t_2) + \left[\frac{S(t_2)}{t_3-t_2} - \frac{S''(t_2)(t_3-t_2)}{6}\right](t_3-x). \end{aligned} \tag{109}$$

We claim $p = S$ in $[t_2, t_3]$. Since $\deg(p) = \deg(S) = 3$, we will be done if we show p and S match at four distinct constraints. Observe

$$\begin{aligned} p(t_2) &= 0 + \frac{S''(t_2)(t_3-t_2)^2}{6} + \cancelto{0}{\left[\frac{S(t_3)}{t_3-t_2} - \frac{S''(t_3)(t_3-t_2)}{6}\right]} 0 + \left[\frac{S(t_2)}{t_3-t_2} - \frac{S''(t_2)(t_3-t_2)}{6}\right](t_3-t_2) \\ &= S(t_2). \end{aligned} \tag{110}$$

In similar fashion, we see $p(t_3) = S(t_3)$. Moreover,

$$p''(x) = S''(t_3)\frac{x-t_2}{t_3-t_2} + S''(t_2), \frac{x-t_3}{t_2-t_3} \tag{111}$$

which is the Lagrange interpolating polynomial for S'' between $S''(t_2)$ and $S''(t_3)$, i.e., $p''(t_i) = S''(t_i)$ for $i = 2, 3$. This shows four degrees of freedom for which p matches S, and so we conclude $p = S$ in $[t_2, t_3]$.

We use the fact $p = S$ to show S is linear over $[t_2, t_3]$. Because S is linear over $[t_1, t_2]$ and $[t_3, t_4]$, we have $S''(t_2) = S''(t_3) = 0$. This implies, for $x \in [t_2, t_3]$,

$$S(x) = p(x) = 0 + 0 + \left[\frac{S(t_3)}{t_3-t_2} - 0\right](x-t_2) + \left[\frac{S(t_2)}{t_3-t_2} - 0\right](t_3-x) = S(t_3)\frac{x-t_2}{t_3-t_2} + S(t_2)\frac{x-t_3}{t_2-t_3}, \tag{112}$$

i.e., $\deg(S) = 1$. Thus S is linear over $[t_2, t_3]$.

□

S10.02: Consider the iteration

$$x_{n+1} = x_n - \left(\frac{x_n - x_0}{f(x_n) - f(x_0)}\right) f(x_n) \tag{113}$$

for finding the roots of a two times continuously differentiable function $f(x)$. Assuming the method converges to a simple root x^*, what is the rate of convergence? Justify your answer.

Solution:

The iteration may be rewritten as

$$x_{n+1} = \frac{[x_n f(x_n) - x_n f(x_0)] - [x_n f(x_n) - x_0 f(x_n)]}{f(x_n) - f(x_0)} = \frac{x_0 f(x_n) - x_n f(x_0)}{f(x_n) - f(x_0)}. \tag{114}$$

Thus

$$x_{n+1} - x^* = \frac{x_0 f(x_n) - x_n f(x_0)}{f(x_n) - f(x_0)} - x^* = \frac{(x_0 - x^*) f(x_n) - (x_n - x^*) f(x_0)}{f(x_n) - f(x_0)}. \tag{115}$$

Taylor's theorem asserts there is ξ_n between x_n and x^* such that

$$0 = f(x^*) = f(x_n) + f'(\xi_n)(x^* - x_n) \quad \Rightarrow \quad f(x_n) = f'(\xi_n)(x_n - x^*). \tag{116}$$

This implies

$$x_{n+1} - x_n = \frac{(x_0 - x^*) f'(\xi_n) - f(x_0)}{f(x_n) - f(x_0)}(x_n - x^*). \tag{117}$$

Evaluating the limit as $n \longrightarrow \infty$, $\xi_n \longrightarrow x^*$ and

$$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|} = \lim_{n\to\infty} \left| \frac{(x_0 - x^*) f'(\xi_n) - f(x_0)}{f(x_n) - f(x_0)} \right| = \left| \frac{(x_0 - x^*) f'(x^*) - f(x_0)}{0 - f(x_0)} \right|. \tag{118}$$

Taylor expanding once more, we know there is η between x^* and x_0 such that

$$f(x_0) = \cancelto{0}{f(x^*)} + f'(x^*)(x_0 - x^*) + \frac{f''(\eta)}{2}(x_0 - x^*)^2 \quad \Rightarrow \quad (x_0 - x^*) f'(x^*) - f(x_0) = -\frac{f''(\eta)}{2}(x_0 - x^*)^2. \tag{119}$$

Therefore

$$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|} = \left| \frac{f''(\eta)}{f(x_0)} \right| \frac{(x_0 - x^*)^2}{2}. \tag{120}$$

Note the right hand side is dependent only upon x_0 and x^*. Since we know $x_n \longrightarrow x^*$, this shows the rate of convergence is linear. □

S10.03: Let $P_{0,1,\ldots,n} := P_{x_0,x_1,\ldots,x_n}$ be the interpolating Lagrange polynomial of degree at most n through the points $x_0, x_1, \ldots, x_n$ and the values $f(x_0), \ldots, f(x_n)$, such that $P_{0,1,\ldots,n}(x_i) = f(x_i)$.

a) Let $i, j \in \{0, 1, \ldots, n\}$ be two distinct integers. Express $P_{0,1,\ldots,n}$ in terms of $P_{0,\ldots,i-1,i+1,\ldots,n}$ and $P_{0,\ldots,j-1,j+1,\ldots,n}$.

b) Suppose $x_j = j$ for $j = 0, 1, 2, 3$ and it is known that $P_{0,1}(x) = x + 1$, $P_{1,2}(x) = 3x - 1$, and $P_{1,2,3}(1.5) = 4$. Find $P_{0,1,2,3}(1.5)$.

Solution:

a) For simplicity, set $P := P_{0,1,\ldots,n}$, $P_{\neq i} := P_{0,\ldots,i-1,i+1,\ldots,n}$ and $P_{\neq j} := P_{0,\ldots,j-1,j+1,\ldots,n}$. We claim $P(x) = p(x)$ where

$$p(x) := P_{\neq i}(x)\frac{x - x_i}{x_j - x_i} + P_{\neq j}(x)\frac{x - x_j}{x_i - x_j}. \tag{121}$$

Indeed, observe

$$\begin{aligned} p(x_i) &= P_{\neq i}(x_i)0 + P_{\neq j}(x_i)1 = P_{\neq j}(x_i) = f(x_i), \\ p(x_j) &= P_{\neq i}(x_j)1 + P_{\neq j}(x_j)0 = P_{\neq i}(x_j) = f(x_j), \\ p(x_k) &= P_{\neq i}(x_k)\frac{x_k - x_i}{x_j - x_i} + P_{\neq j}(x_j)\frac{x_k - x_j}{x_i - x_j} \\ &= f(x_k)\left(\frac{(x_k - x_i) - (x_k - x_j)}{x_j - x_i}\right) = f(x_k) \quad \text{for } k \in \{0, 1, \ldots, n\} - \{i, j\}. \end{aligned} \tag{122}$$

This shows p interpolates the points $x_0, x_1, \ldots, x_n$ and the values $f(x_0), \ldots, f(x_n)$, such that $p(x_i) = f(x_i)$. Moreover, $\deg(P_{\neq i}), \deg(P_{\neq j}) \leq (n-1)$ and so $\deg(xP_{\neq i}), \deg(xP_{\neq j}) \leq n$, which implies $\deg(p) \leq n$. Since the polynomial of degree $\leq n$ interpolating $n + 1$ points is unique, we conclude $p = P$, as desired.

b) Observe $f(0) = P_{0,1}(0) = 1$, $f(1) = P_{0,1}(1) = 2$, and $f(2) = P_{1,2}(2) = 5$. Then the Lagrange interpolating polynomial $P_{0,1,2}$ is given by

$$\begin{aligned} P_{0,1,2}(x) &= f(0)\frac{(x-1)(x-2)}{(0-1)(0-2)} + f(1)\frac{(x-0)(x-2)}{(1-0)(1-2)} + f(2)\frac{(x-0)(x-1)}{(2-0)(2-1)} \\ &= \frac{(x-1)(x-2)}{2} - 2x(x-2) + \frac{5x(x-1)}{2}, \end{aligned} \tag{123}$$

which implies

$$P_{0,1,2}(3/2) = \frac{(1/2)(-1/2)}{2} - 2(3/2)(-1/2) + \frac{5(3/2)(1/2)}{2} = -\frac{1}{8} + \frac{12}{8} + \frac{15}{8} = \frac{13}{4}. \tag{124}$$

Thus, using the result in a),

$$\begin{aligned} P_{0,1,2,3}(3/2) &= P_{0,1,2}(3/2)\frac{3/2-3}{0-3} + P_{1,2,3}(3/2)\frac{3/2-0}{3-0} \\ &= \frac{1}{2}\left[P_{0,1,2}(3/2) + P_{1,2,3}(3/2)\right] \\ &= \frac{1}{2}\left[\frac{13}{4} + 4\right] \\ &= \boxed{\frac{29}{8}.} \end{aligned} \tag{125}$$

□

S10.04: The trapezoidal rule applied to $\int_0^2 f(x)\ \mathrm{d}x$ gives the value of 4, and Simpson's rule gives the value 2. What is $f(1)$?

Solution:

The trapezoidal rule applied to this problem asserts

$$\int_0^2 f(x)\ \mathrm{d}x \approx \frac{2}{2}\left[f(0)+f(2)\right] = f(0)+f(2) = 4. \tag{126}$$

Simpson's rule applied here asserts

$$\int_0^2 f(x)\ \mathrm{d}x \approx \frac{1}{3}\left[f(0)+4f(1)+f(2)\right] = 2. \tag{127}$$

This implies

$$f(1) = \frac{1}{4}\left[3\cdot 2 - \left[f(0)+f(2)\right]\right] = \frac{1}{4}\left[6-4\right] = \frac{1}{2}. \tag{128}$$

□

S10.05: Consider the numerical method $y^{n+1} = y^n + kAy^{n+1}$ used to create approximate solutions of the linear system of equations

$$\frac{\mathrm{d}y}{\mathrm{d}t} = Ay, \quad y(t_0) = y_0 \quad \text{for } t \in [t_0, T]. \tag{129}$$

a) Derive a bound for the local truncation error in the $\|\cdot\|_2$ norm of the form $C(T)k^p$ where the constant $C(T)$ is explicitly expressed in terms of $\sup_{t\in[t_0,T]} \|y(t)\|_2$ and powers of $\|A\|_2$ and holds for $t \in [t_0, T]$.

b) Assume A is symmetric and negative definite. If $e^n = y^n - y(t^n)$ is the error at the n-th step and $C(T)k^p$ the bounded derived in a), show that

$$\|e^n\|_2 \leq [T - t_0]C(T)k^{p-1} \quad \text{for all } n \text{ with } t_0 \leq nk \leq T, \tag{130}$$

assuming $e^0 = 0$.

Note: The defining equation for the local truncation error assume for this problem is not based on the numerical method by k.

Solution:

a) Taylor's theorem asserts there is ξ^{n+1} between $y(t^{n+1})$ and $y(t^n)$ such that

$$y(t^n) = y(t^{n+1}) - k\frac{\mathrm{d}y}{\mathrm{d}t}(t^{n+1}) + \frac{k^2}{2}\frac{\mathrm{d}^2y}{\mathrm{d}t}(\xi^{n+1}) = y(t^{n+1}) - kAy(t^{n+1}) + \frac{k^2}{2}\left\langle y(\xi^{n+1}), Ay(\xi^{n+1})\right\rangle. \tag{131}$$

This implies the local truncation error τ^{n+1} is given by

$$\begin{aligned}
\tau^{n+1} &:= \|y(t^{n+1}) - \left[y(t^n) + kAy(t^{n+1})\right]\|_2 \\
&= \frac{k^2}{2}\|\left\langle y(\xi^{n+1}), Ay(\xi^{n+1})\right\rangle\|_2 \\
&\leq \frac{k^2}{2}\|y(\xi^{n+1})\|_2\|Ay(\xi^{n+1})\|_2 \\
&\leq \frac{k^2}{2}\|A\|_2\|y(\xi^{n+1})\|_2^2 \\
&\leq k^2 \cdot \underbrace{\frac{1}{2}\|A\|_2 \sup_{t\in[t_0,T]} \|y(t)\|_2^2}_{C(T)}.
\end{aligned} \tag{132}$$

The second line follows from (131), the third from the triangle inequality, and the fourth from the definition of $\|A\|_2$. Since the final line of (132) is independent of n, we deduce the truncation error at each step is bounded by $C(T)k^2$.

b) Let (λ, v) be an eigenvalue/eigenvector pair for A. By hypothesis, A is invertible and $\lambda < 0$. This implies $(I - kA)v = (1 - k\lambda)v$ with $1 - k\lambda > 1$. So, $(I - kA)$ is symmetric and positive definite with eigenvalues greater than unity. This implies $(I - kA)^{-1}$ has eigenvalues contained in $(0, 1)$, and so $\rho((I - kA)^{-1}) < 1$. Whence $\|(I - kA)^{-1}\| < 1$. Then observe

$$\begin{aligned} e^n := y^n - y(t^n) &= (I - kA)\left(y^{n+1} - y(t^{n+1})\right) - \frac{k^2}{2}\left\langle y(\xi^{n+1}), Ay(\xi^{n+1})\right\rangle \\ &= (I - kA)e^{n+1} - \frac{k^2}{2}\left\langle y(\xi^{n+1}), Ay(\xi^{n+1})\right\rangle, \end{aligned} \tag{133}$$

which implies

$$\begin{aligned} \|e^{n+1}\| &= \left\|(I - kA)^{-1}\left(e^n - \langle y(\xi^{n+1}), Ay(\xi^{n+1})\rangle\right)\right\| \\ &\le \left\|(I - kA)^{-1}\right\| \left\|\left(e^n - \langle y(\xi^{n+1}), Ay(\xi^{n+1})\rangle\right)\right\| \\ &\le \left\|e^n - \langle y(\xi^{n+1}), Ay(\xi^{n+1})\rangle\right\| \\ &\le \|e^n\| + \left\|\langle y(\xi^{n+1}), Ay(\xi^{n+1})\rangle\right\| \\ &\le \|e^n\| + C(T)k^2. \end{aligned} \tag{134}$$

Thus we conclude, for all n with $t_0 \le t_n \le T$ and $t_n := t_0 + nk$,

$$\|e^n\| \le \cdots \le \|e^0\| + \sum_{i=1}^{n} C(T)k^2 \le 0 + (nk)C(T)k = [t_n - t_0]C(T)k \le [T - t_0]C(T)k, \tag{135}$$

a desired.

□

We assume the problem statement meant to say $t_0 \le t_n \le T$ where $t_n := t_0 + nk$.

S10.08: Consider the problem

$$\begin{aligned} -\operatorname{div}(a(x)\nabla u) + b(x)u &= f(x), \quad && x \in \Omega, \\ u &= 0, && x \in \partial\Omega_1, \\ \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} + u &= 2, && x \in \partial\Omega_2, \end{aligned} \tag{136}$$

where

$$\begin{aligned} \Omega &= \{x \ : \ x_1, x_2 > 0, \ x_1 + x_2 < 1\}, \\ \partial\Omega_1 &= \{x \ : \ x_1 = 0, 0 \le x_2 \le 1\} \cup \{x \ : \ x_2 = 0, 0 \le x_1 \le 1\}, \\ \partial\Omega_2 &= \{x \ : \ x_1, x_2 > 0, x_1 + x_2 = 1\}, \end{aligned} \tag{137}$$

and $0 < \alpha \le a(x) \le A$, $0 < \beta \le b(x) \le B$, with $a(x)$ and $b(x)$ smooth functions and $f \in L^2(\Omega)$.

a) Find the weak variational formulation and show that the problem is well-posed, by verifying the assumptions of the Lax-Milgram lemma and by analyzing the appropriate bilinear and linear forms.

b) Develop and describe the piecewise linear Galerkin finite element approximation of the problem and a set of basis functions such that the corresponding linear system is sparse. Show that this linear system has a unique solution.

Solution:

a) Assume u is a solution to the problem (136). Define the Hilbert space $H := \{v \in H^1(\Omega) \ : \ v = 0 \text{ on } \partial\Omega_1\}$. Then, for each $v \in H$,

$$\begin{aligned} \int_\Omega fv &= \int_\Omega -\operatorname{div}(aDu)v + buv \\ &= \int_\Omega (aDu)\cdot Dv + buv - \int_{\partial\Omega} (n \cdot aDu)v \\ &= \int_\Omega (aDu)\cdot Dv + buv - \int_{\partial\Omega_2} \frac{av}{\sqrt{2}}\left(\frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2}\right) \\ &= \int_\Omega (aDu)\cdot Dv + buv - \int_{\partial\Omega} \frac{av}{\sqrt{2}}(2-u). \end{aligned} \tag{138}$$

The second equality holds by integration by parts; the third holds since u has zero trace on $\partial\Omega_1$ and $n = \frac{1}{\sqrt{2}}(1,1)$ on $\partial\Omega_2$; the final equality holds by substituting from (136). Then the weak variational form of the problem is given by

$$\text{Find } u \in H \text{ such that } \sigma(u,v) = \ell(v) \ \ \forall v \in H, \tag{139}$$

where

$$\sigma(u,v) := \int_\Omega aDu\cdot Dv + buv + \int_{\partial\Omega_2}\frac{auv}{\sqrt{2}} \quad \text{and} \quad \ell(v) := \int_\Omega fv + \int_{\partial\Omega_2}\sqrt{2}av. \tag{140}$$

We now verify the assumptions of the Lax-Milgram lemma. We must show the bilinear form σ is coercive and bounded and the linear form ℓ is bounded. First note σ is coercive since

$$\begin{aligned}\sigma(u,u) &= \int_\Omega a|Du|^2 + bu^2 + \int_{\partial\Omega_2}\frac{au^2}{\sqrt{2}}\\ &\geq \int_\Omega a|Du|^2 + bu^2\\ &\geq \min\{\alpha,\beta\}\int_\Omega |Du|^2 + u^2\\ &= \min\{\alpha,\beta\}\|u\|_H^2.\end{aligned} \tag{141}$$

The form ℓ is bounded because

$$\begin{aligned}|\ell(v)| &\leq \|fv\|_{L^1(\Omega)} + \|\sqrt{2}av\|_{L^1(\partial\Omega_2)}\\ &\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \sqrt{2}\|a\|_{L^2(\partial\Omega_2)}\|v\|_{L^2(\partial\Omega_2)}\\ &\leq \|f\|_{L^2(\Omega)}\|v\|_H + \sqrt{2}\|a\|_{L^2(\partial\Omega_2)}\|v\|_{L^2(\partial\Omega)}\\ &\leq \|f\|_{L^2(\Omega)}\|v\|_H + \sqrt{2}\|a\|_{L^2(\partial\Omega_2)}C\|v\|_H\\ &= \left(\|f\|_{L^2(\Omega)} + \sqrt{2}C\|a\|_{L^2(\partial\Omega_2)}\right)\|v\|_H.\end{aligned} \tag{142}$$

Note $a \in L^2(\partial\Omega_2)$ since a is smooth and $\partial\Omega_2$ is bounded. Lastly, we show σ is bounded. Observe

$$\begin{aligned}|\sigma(u,v)| &\leq A\|Du\cdot Dv\|_{L^1(\Omega)} + B\|uv\|_{L^1(\Omega)} + \frac{A}{\sqrt{2}}\|uv\|_{L^1(\partial\Omega_2)}\\ &\leq A\|Du\|_{L^2(\Omega)}\|Dv\|_{L^2(\Omega)} + B\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \frac{A}{\sqrt{2}}\|u\|_{L^2(\partial\Omega_2)}\|v\|_{L^2(\partial\Omega_2)}.\end{aligned} \tag{143}$$

Together with the facts $\|Du\|_{L^2(\Omega)} \leq \|u\|_H$ and there is $C > 0$ such that

$$\|u\|_{L^2(\partial\Omega_2)} \leq \|u\|_{L^2(\partial\Omega)} \leq C\|u\|_H, \tag{144}$$

this implies

$$|\sigma(u,v)| \leq \left(A + B + \frac{AC^2}{\sqrt{2}}\right)\|u\|_H\|v\|_H, \tag{145}$$

and we are done.

b) Let $\mathcal{T}_h$ be a triangulation of the domain Ω. Define $H_h := \{v \in H \ : \ v|_K \text{ is linear } \forall\ K \in \mathcal{T}_h\}$. Let n_i for $i = 1, \dots, M$ be the nodes corresponding to $\mathcal{T}_h$. Let $\{\phi_i\}_{i=1}^M$ be the basis for H_h such that $\phi_i(n_j) = \delta_{ij}$ with δ_{ij} the Kronecker δ. Consider a solution $u = \sum_{i=1}^M \xi_i \phi_i$ where $\xi \in \mathbb{R}^n$ is constant. Our finite element problem becomes

$$\text{Find } u_h \in H_h \text{ such that } a(u_h, v) = \ell(v) \ \ \forall\ v \in H_h. \tag{146}$$

The bilinearity of σ implies

$$\ell(\phi_j) = \sigma(u_h, \phi_j) = \sigma\left(\sum_{i=1}^M \xi_i \phi_i, \phi_j\right) = \sum_{i=1}^M \xi_i \sigma(\phi_i, \phi_j) \quad j = 1, \dots, M. \tag{147}$$

This can be written as the matrix equation $A\xi = b$ where $A_{ij} := \sigma(\phi_i, \phi_j)$ and $b_i := \ell(\phi_i)$. Note A is symmetric and

$$\xi \cdot A\xi = \sum_{i,j=1}^M \xi_i A_{i,j} \xi_j = \sigma\left(\sum_{i=1}^M \xi_i \phi_i, \ \sum_{j=1}^M \xi_j \phi_j\right) = \sigma(u_h, u_h) > 0, \tag{148}$$

where the final inequality holds since σ is coercive. Because A is symmetric and positive definite, the system $A\xi = b$ has a unique solution. Moreover, A is sparse since $A_{i,j} = a(\phi_i, \phi_j) = 0$ whenever there does not exists $K \in \mathcal{T}_h$ such that $n_i, n_j \in K$.

□

Fall 2010

F10.01: Simpson's rule with its error term for numerical integration is given by

$$\int_{x_0}^{x_2} f(x)\ \mathrm{d}x = \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right] - \frac{h^5}{90} f^{(4)}(\xi), \tag{149}$$

where $f \in C^4[x_0, x_2]$ and $x_1 - x_0 = x_2 - x_1 = h > 0$. Assume $f \in C^4[a, b]$, n is even, $h = (b - a)/n$, and $x_j = a + jh$ for $j = 0, 1, \ldots, n$. Show there is $\mu \in (a, b)$ for which the composite Simpson's rule for n subintervals can be written with its error term as

$$\int_a^b f(x)\ \mathrm{d}x = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{b-a}{180} h^4 f^{(4)}(\mu). \tag{150}$$

Solution:

Define $S_i = \int_{x_i}^{x_i+2} f(x)\ \mathrm{d}x$ for $i = 0, \ldots, n-1$. Then

$$\begin{aligned}\int_a^b f(x)\ \mathrm{d}x &= \sum_{i=0}^{(n/2)-1} S_{2i} \\ &= \sum_{i=0}^{(n/2)-1} \left(\frac{h}{3}\left[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right] - \frac{h^5}{90} f^{(4)}(\xi_{2i})\right) \\ &= \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j-1}) + f(b)\right] - \frac{h^5}{90}\sum_{i=0}^{(n/2)-1} f^{(4)}(\xi_{2i}).\end{aligned} \tag{151}$$

Here $\xi_i \in (x_i, x_{i+2})$ for each i. Define

$$\alpha := \frac{2}{n}\sum_{i=0}^{(n/2)-1} f^{(4)}(\xi_{2i}) \tag{152}$$

and observe α is the average of all the $f^{(4)}(\xi_{2i})$'s. This implies

$$\min_{0\le i\le n/2-1} f^{(4)}(\xi_{2i}) \le \alpha \le \max_{0\le i\le n/2-1} f^{(4)}(\xi_{2i}). \tag{153}$$

Since $f^{(4)}$ is continuous, the intermediate value theorem asserts there is $\mu \in (a, b)$ such that $f^{(4)}(\mu) = \alpha$. Whence we conclude

$$\frac{h^5}{90}\sum_{i=0}^{(n/2)-1} f^{(4)}(\xi_{2i}) = \frac{h^5}{90}\frac{n}{2} f^{(4)}(\mu) = \frac{nh}{180} h^4 f^{(4)}(\mu) = \frac{b-a}{180} h^4 f^{(4)}(\mu), \tag{154}$$

as desired. □

F10.02: Let $g : [a, b] \to [a, b]$ be a continuously differentiable function. Assume there is a constant $0 < k < 1$ such that $|g'(x)| \leq k$ for all $x \in (a, b)$. Let $p \in [a, b]$ be a unique fixed point of g. For any $p_0 \in [a, b]$, define the sequence $p_n = g(p_{n-1})$ for $n \geq 1$.

a) Show the sequence $\{p_n\}$ converges to p.

b) If $g'(p) \neq 0$, show that $\{p_n\}$ converges only linearly to p.

Solution:

a) We claim g is Lipschitz continuous with Lipschitz constant k. This implies

$$|p_{n+1} - p_n| = |g(p_n) - g(p_{n-1})| \leq k|p_n - p_{n-1}| \leq \cdots \leq k^n |p_1 - p_0|. \tag{155}$$

Thus, for $m > n$,

$$\begin{aligned} |p_m - p_n| &\leq |p_m - p_{m-1}| + \cdots + |p_{n+1} - p_n| \\ &\leq k^{m-1}|p_1 - p_0| + \cdots + k^n|p_1 - p_0| \\ &= k^n|p_1 - p_0| \sum_{j=0}^{(m-1)-n} k^j \\ &\leq k^n|p_1 - p_0| \sum_{j=0}^{\infty} k^j \\ &= k^n \frac{|p_1 - p_0|}{1 - k}. \end{aligned} \tag{156}$$

Since $\lim_{n\to\infty} k^n = 0$, the right hand side of (156) goes to zero as $n \longrightarrow \infty$, which shows the sequence $\{p_n\}$ is Cauchy. Because $\mathbb{R}$ is complete, this means $\{p_n\}$ converges to some limit $x \in \mathbb{R}$. Thus, by the continuity of g,

$$x = \lim_{n\to\infty} p_n = \lim_{n\to\infty} p_{n+1} = \lim_{n\to\infty} g(p_n) = g\left(\lim_{n\to\infty} p_n\right) = g(x). \tag{157}$$

But, p is the unique fixed point of g and so $x = p$. This shows $p_n \longrightarrow p$, as desired.

All that remains is to verify the initial claim. Let $\alpha, \beta \in [a, b]$. Then Taylor's theorem asserts there is ξ between α and β such that

$$g(\alpha) = g(\beta) + g'(\xi)(\alpha - \beta) \quad \Rightarrow \quad |g(\alpha) - g(\beta)| = |g'(\xi)||\alpha - \beta| \leq k|\alpha - \beta|, \tag{158}$$

and we are done.

b) For each n, Taylor's theorem asserts there is ξ_n between p_n and p such that

$$p_{n+1} = g(p_n) = \cancelto{p}{g(p_n)} + g'(\xi_n)(p_n - p) \qquad \Rightarrow \qquad \frac{|p_{n+1} - p|}{|p_n - p|} = |g'(\xi_n)|. \tag{159}$$

Since $p_n \longrightarrow p$, $\xi_n \longrightarrow p$ and we deduce

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = \lim_{n\to\infty} |g'(\xi_n)| = |g'(p)| \neq 0. \tag{160}$$

Thus $p_n \longrightarrow p$ linearly.

□

F10.03: Let x be the solution of $Ax = b$ and $\tilde{x}$ be the solution of $A\tilde{x} = \tilde{b}$ where A is an $n \times n$ matrix.

a) Define $\kappa_2(A)$, the condition number of A using the 2-norm.

b) Give a derivation of the error bound

$$\frac{\|x - \tilde{x}\|_2}{\|x\|_2} \le \kappa_2(A) \frac{\|b - \tilde{b}\|_2}{\|b\|_2}. \tag{161}$$

Solution:

a) The condition number $\kappa_2(A)$ of A is defined by

$$\kappa_2(A) := \|A\|_2 \|A^{-1}\|_2, \tag{162}$$

where for each $n \times n$ matrix M we define

$$\|M\|_2 := \sup\{\|Ax\|_2 \ : \ \|x\|_2 = 1\}. \tag{163}$$

b) First observe that

$$\|x - \tilde{x}\|_2 = \|(A^{-1}A)(x - \tilde{x})\|_2 = \|A^{-1}(b - \tilde{b})\|_2 \le \|A^{-1}\|_2 \|b - \tilde{b}\|_2, \tag{164}$$

and

$$\|b\|_2 = \|Ax\|_2 \le \|A\|_2 \|x\|_2 \quad \Rightarrow \quad \frac{1}{\|x\|_2} \le \frac{\|A\|_2}{\|b\|_2}. \tag{165}$$

Combining these relations, we deduce

$$\frac{\|x - \tilde{x}\|_2}{\|x\|_2} \le \frac{\|A\|_2 \|x - \tilde{x}\|_2}{\|b\|_2} \le \frac{\|A\|_2 \|A^{-1}\|_2 \|b - \tilde{b}\|_2}{\|b\|_2} = \kappa_2(A) \frac{\|b - \tilde{b}\|_2}{\|b\|_2}, \tag{166}$$

as desired.

□

Spring 2011

S11.07: Consider the equation

$$u_t + uu_x = \varepsilon u_{xx} \tag{167}$$

with $\varepsilon > 0$, to be solved for $t \geq 0$, $u(x,0) = \Phi(x)$, and $0 \leq x \leq 1$ with periodic boundary conditions $u(x+1,t) = u(x,t)$.

a) Construct a finite difference scheme which converges with a rate independent of ε.

b) Justify your statement.

Solution:

a) Define the function $f(u) := u^2/2$ and note this allows the PDE to be rewritten as $u_t + f(u)_x = \varepsilon u_{xx}$. Then we propose the scheme

$$\frac{u_i^{n+1} - \left[(1-k)u_i^n + \frac{k}{2}(u_{i+1}^n + u_{i-1}^n)\right]}{k} + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2h} - \frac{\varepsilon}{h^2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) = 0. \tag{168}$$

b) If $\Phi(x) = 0$, then $u^n = 0$ for all $n \geq 0$ and we are done. So, in what follows, assume $\|u^0\|_\infty \neq 0$. Our scheme may be rewritten as

$$u_i^{n+1} = u_i^n - \lambda\left[F(u_i^n, u_{i+1}^n) - F(u_{i-1}^n, u_i^n)\right] \tag{169}$$

where the numerical flux function F is defined by

$$F(u,v) := \left(\frac{h}{2} + \frac{\varepsilon}{h}\right)(u - v) + \frac{f(u) + f(v)}{2}. \tag{170}$$

Equation (169) shows the scheme is in conservation form. Below we establish the consistency and stability of the scheme.

We first verify consistency.

Now we verify stability. Assume $k \in (0,1]$, $\lambda \leq h/2\varepsilon$, and

$$h + \frac{2\varepsilon}{h} \geq \|u^0\|_\infty^2. \tag{171}$$

Also observe the scheme may be rewritten as

$$\begin{aligned} u_i^{n+1} &= \left(1 - k - \frac{2\lambda\varepsilon}{h}\right) u_i^n \\ &+ \left[\frac{k}{2} - \frac{\lambda}{4}\left(u_{i+1}^n + u_{i-1}^n\right) + \frac{\varepsilon\lambda}{h}\right] u_{i+1}^n \\ &+ \left[\frac{k}{2} + \frac{\lambda}{4}\left(u_{i+1}^n + u_{i-1}^n\right) + \frac{\varepsilon\lambda}{h}\right] u_{i-1}^n \end{aligned} \tag{172}$$

We claim our choice of k and h causes all the coefficients to be nonnegative. THE FOLLOWING IS INCORRECT. Indeed, observe

$$1 - k - \frac{2\lambda\varepsilon}{h} \geq 0 \quad \Leftrightarrow \quad 1 - \frac{2\lambda\varepsilon}{h} \geq k \geq 0 \quad \Leftrightarrow \quad \frac{1}{2} \geq \frac{\lambda\varepsilon}{h}. \tag{173}$$

And

$$\frac{k}{2} \pm \frac{\lambda}{4}\left(u_{i+1}^n + u_{i-1}^n\right) + \frac{\varepsilon\lambda}{h} \geq 0 \quad \Leftrightarrow \quad \frac{k}{2} + \frac{\varepsilon\lambda}{h} \geq \frac{\lambda}{2}\max\{|u_{i+1}^n|, |u_{i-1}^n|\}. \tag{174}$$

Equivalently,

$$\left(h + \frac{2\varepsilon}{h}\right) \geq \max\{|u_{i+1}^n|, |u_{i-1}^n|\}. \tag{175}$$

Thus we conclude $\|u^n\|_\infty \leq \|u^0\|_\infty$ for all $n \in \mathbb{Z}^+$ and the method is stable.

The Lax-Wendroff theorem states that if a conservative scheme $\{u_i^n\}$ converges as $k, h, \longrightarrow 0$ a.e. to a function $u(x,t)$, then u is a weak solution of the conservation law. And, the above shows the Lax-Friedrichs method on a sequence of grids converges to the "vanishing viscosity" solution as $\varepsilon \longrightarrow 0$, which may thus be used to define the physically relevant weak solution to the conservation law $u_t + f(u)_x = 0$.

□

Fall 2011

F11.08: Give a variational formulation of the problem

$$\begin{aligned} u'''' &= f \quad \text{for} \quad x \in (0,1) \\ u(0) = u''(0) &= u'(1) = u'''(1) = 0, \end{aligned} \tag{176}$$

and show the assumptions of the Lax-Milgram Lemma are satisfied (assume that $f \in L^2(0,1)$). Which boundary conditions are essential and which are natural? Develop and describe a finite element approximation of the problem using piecewise-cubic functions and a uniform partition; describe the basis functions, the degrees of freedom of the finite-dimensional space and the corresponding linear system. Show that the linear system is sparse and has a unique solution.

Solution:

If u is a solution to our ODE, then for each $v \in W := \{v \in H^2(\Omega) \ | \ v(0) = v'(1) = 0\}$ we obtain

$$(f,v) = (u''',v) = -(u'',v') + [u'''v]_0^1 = (u'',v'') - [u''v']_0^1 + [u'''v]_0^1 = (u'',v'') \tag{177}$$

where $(\cdot,\cdot)$ is the inner product for $L^2(0,1)$. Define $\ell : H^2(0,1) \to \mathbb{R}$ and $B : H^2(0,1) \times H^2(0,1) \to \mathbb{R}$ by $\ell(v) := (f,v)$ and $B(u,v) := (u'',v'')$, respectively. Then the weak variational form of the ODE is

$$\text{Find } u \in W \text{ such that } B(u,v) = \ell(v) \text{ for all } v \in W. \tag{178}$$

We now show the assumptions of the Lax-Milgram theorem are satisfied. Of course, ℓ and B are linear in each of there arguments since the integral is linear. Now observe

$$B(u,v) = (u'',v'') \leq \|u''v''\|_{L^1(0,1)} \leq \|u''\|_{L^2(0,1)}\|v''\|_{L^2(0,1)} \leq \|u\|_{H^2(0,1)}\|v\|_{H^2(0,1)}, \tag{179}$$

and so B is bounded. Similarly,

$$\ell(v) = (f,v) \leq \|fv\|_{L^1(0,1)} \leq \|f\|_{L^2(0,1)}\|v\|_{L^2(0,1)} \leq \|f\|_{L^2(0,1)}\|v\|_{H^2(0,1)}, \tag{180}$$

which verifies ℓ is bounded. The last assumption of the Lax-Milgram Lemma to be verified is that B is coercive.

□

Spring 2012

S12.01: Let A be a $m \times n$ matrix and $b \in \mathbb{R}^m$ with $m > n$.

1. Assume $\text{rank}(A) = n$. Derive the equations that determine the solution to
$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2.$$
2. Outline a procedure for obtaining the solution to these equations that avoids problems due to ill-conditioning that may occur if one uses Gaussian elimination on the equations directly.

Solution:

a) First note that, since norms are convex, the expression has a minimizer over $\mathbb{R}^n$ and no maximizers. Moreover,
$$0 \leq \|Ax - b\|_2^2 = (Ax - b)^T(Ax - b) = x^TA^TAx - 2b^TAx + b^Tb. \tag{181}$$
Thus, at the minimizer x^*, we obtain
$$0 = \frac{\mathrm{d}}{\mathrm{d}x}\left[x^TA^TAx - 2b^TAx + b^Tb\right]_{x=x^*} = 2A^TAx^* - 2b^TA \quad\Rightarrow\quad A^TAx^* = b^TA. \tag{182}$$
Thus a minimizer x^* satisfies $A^TAx^* = b^TA$. We claim A^TA is invertible and so
$\boxed{x^* = (A^TA)^{-1}b^TA.}$

All that remains is to show A^TA is invertible. Note this is equivalent to showing $\ker(A^TA) = \{0\}$. Since A is linear, $A^TA0 = 0$ and so $\{0\} \subseteq \ker(A^TA)$. Conversely, if $A^TAx = 0$, then
$$0 = x^T0 = x^TA^TAx = \|Ax\|^2 \quad\Rightarrow\quad Ax = 0. \tag{183}$$
Since $\text{rank}(A) = n$, we deduce $x = 0$. This shows $\ker(A^TA) \subset \{0\}$ and completes the proof.

b) Idk... See Shea's notes???

□

S12.02: Consider the polynomial $p(x)$ defined by

$$p(x) := 1 + x + x(x-1) + \frac{1}{6}x(x-1)(x-2) + \frac{1}{24}x(x-1)(x-2)(x-3) + q(x). \tag{184}$$

a) Determine a fifth degree polynomial $q(x)$ so that $p(x)$ interpolates the data

$$\{(-1,-3),(0,1),(1,2),(2,5),(3,11),(4,22)\}. \tag{185}$$

b) Is $q(x)$ unique?

Solution:

a) Using the given data and pluging values into $p(x)$, we discover

$$\begin{aligned} -3 &= p(-1) = 2 + q(-1), \\ 1 &= p(0) = 1 + q(0), \\ 2 &= p(1) = 2 + q(1), \\ 5 &= p(2) = 5 + q(2), \\ 11 &= p(3) = 11 + q(3), \\ 22 &= p(4) = 22 + q(4). \end{aligned} \tag{186}$$

This implies

$$0 = q(0) = q(1) = q(2) = q(3) = q(4) \quad \text{and} \quad q(-1) = -5. \tag{187}$$

Then the Lagrange interpolating polynomial for q is given by

$$q(x) = -5 \cdot \frac{(x-0)(x-1)(x-2)(x-3)(x-4)}{(-1-0)(-1-1)(-1-2)(-1-3)(-1-4)} = \frac{x(x-1)(x-2)(x-3)(x-4)}{24}. \tag{188}$$

b) Yes, the polynomial q is unique. It is a fifth degree polynomial that interpolates six points.

□

S12.03: Determine constants $c_1, c_2, x_1, x_2 \in \mathbb{R}$ such that the integration formula

$$\int_{-1}^{1} f(x)\ \mathrm{d}x \approx c_1 f(x_1) + c_2 f(x_2) \tag{189}$$

has degree of precision 3.

Solution:

We must find $c_1, c_2, x_1, x_2 \in \mathbb{R}$ such that

$$c_1 x_1^k + c_2 x_2^k = \int_{-1}^{1} x^k\ \mathrm{d}x \quad \text{for } k = 0, 1, 2, 3, \tag{190}$$

i.e., so the approximation has degree of precision 3. This follows from the fact we have 4 unknowns and each value k gives one equation. Note this is only obtainable if each variable is nonzero. Evaluating the integral at each k, we obtain

$$2 = c_1 + c_2, \tag{191}$$

$$0 = c_1 x_1 + c_2 x_2, \tag{192}$$

$$\frac{2}{3} = c_1 x_1^2 + c_2 x_2^2, \tag{193}$$

$$0 = c_1 x_1^3 + c_2 x_2^3. \tag{194}$$

From (192), we deduce $c_1 x_1 = -c_2 x_2$. Plugging this into (193), we discover

$$\frac{2}{3} = c_1 x_1^2 + c_2 x_2^2 = c_1 x_1 (x_1 - x_2) \quad \Rightarrow \quad x_1 - x_2 = \frac{2}{3 c_1 x_1}. \tag{195}$$

Together with (194) this reveals

$$0 = c_1 x_1^3 + c_2 x_2^3 = c_1 x_1 (x_1 - x_2)(x_1 + x_2) = \frac{2}{3}(x_1 + x_2) \quad \Rightarrow \quad x_1 = -x_2. \tag{196}$$

This implies $c_1 x_1 = -c_2 x_2 = c_2 x_1$ and so $c_1 = c_2$. But, (191) then shows $c_1 = c_2 = 1$. Returning to (193), we see

$$\frac{2}{3} = x_1^2 + x_2^2 = x_1^2 + (-x_1)^2 \quad \Rightarrow \quad x_1 = \pm\frac{1}{\sqrt{3}}. \tag{197}$$

Thus we conclude $\boxed{c_1 = c_2 = 1,\ x_1 = 1/\sqrt{3}, \text{ and } x_2 = -1/\sqrt{3}.}$ □

S12.04: Consider the matrix A defined by

$$\begin{pmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}. \tag{198}$$

a) Derive the Jacobi iteration for solving the linear system $Ax = b$ for some $b \in \mathbb{R}^3$.

b) Is this iteration convergent? Justify your answer.

Solution:

a) Write $A = M - N$ where M contains the diagonal entries of A and $-N$ contains all the nondiagonal entries of A. Then the Jacobi iteration is defined by

$$Mx_{k+1} = Nx_k + b. \tag{199}$$

Not sure on the meaning... Since $\det(M) = 4^3 \neq 0$, M is invertible. This implies

$$x_{k+1} = M^{-1}\left(Nx_k + b\right) = \frac{1}{4^3} I \left(Nx_k + b\right) = \frac{1}{4^3}\left(\begin{pmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} x_k + b\right). \tag{200}$$

b) Finish

□

Fall 2012

F12.01: Assume $f : \mathbb{R} \to \mathbb{R}$ is a smooth function. Let $\varepsilon > 0$ be given and consider the three data values $(0, f(0))$, $(\varepsilon, f(\varepsilon))$, and $(1, f(1))$. Let $p(x)$ be the polynomial that arises as the limit of the polynomial interpolant of the data as $\varepsilon \longrightarrow 0$.

a) What is the degree of $p(x)$?

b) What data (if any) does $p(x)$ interpolate?

c) What data (if any) does $p'(x)$ interpolate?

Solution:

Infinitely many polynomials interpolate the three data points. We shall presume "the polynomial interpolant" means the unique polynomial interpolant of order 2, i.e., the Lagrange interpolating polynomial.

a) We derive an expression for p by taking the limit as $\varepsilon \longrightarrow 0$, i.e.,

$$\begin{aligned} p(x) &= \lim_{\varepsilon\to 0} f(0)\frac{(x-\varepsilon)(x-1)}{(0-\varepsilon)(0-1)} + f(\varepsilon)\frac{(x-0)(x-1)}{(\varepsilon-0)(\varepsilon-1)} + f(1)\frac{(x-0)(x-\varepsilon)}{(1-0)(1-\varepsilon)} \\ &= \lim_{\varepsilon\to 0} f(0)\frac{x(x-1)}{\varepsilon} + f(0)(1-x) + f(\varepsilon)\frac{x(x-1)}{\varepsilon(\varepsilon-1)} + f(1)\frac{x(x-\varepsilon)}{1-\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{x(x-1)}{\varepsilon}\left(f(0) + \frac{f(\varepsilon)}{\varepsilon-1}\right) + f(0)(1-x) + f(1)x^2 \\ &= \lim_{\varepsilon\to 0} \frac{x(x-1)}{\varepsilon}\left(f(0) + \frac{f(0)+\varepsilon f'(0)+\mathcal{O}(\varepsilon^2)}{\varepsilon-1}\right) + f(0)(1-x) + f(1)x^2 \\ &= \lim_{\varepsilon\to 0} x(x-1)\left(\frac{f(0)+f'(0)+\mathcal{O}(\varepsilon)}{\varepsilon-1}\right) + f(0)(1-x) + f(1)x^2 \\ &= x(1-x)\left[f(0)+f'(0)\right] + f(0)(1-x) + f(1)x^2 \\ &= x(1-x)f'(0) + f(0)(1-x^2) + f(1)x^2. \end{aligned} \tag{201}$$

This shows $\boxed{\deg(p) = 2.}$

b) We claim p interpolates $\boxed{(0, f(0)) \text{ and } (1, f(1)).}$ Indeed, using (201), we see

$$\begin{aligned} p(0) &= 0(1-0)f'(0) + f(0)(1-0^2) + f(1)0^2 = f(0), \\ p(1) &= 1(1-1)f'(0) + f(0)(1-1^2) + f(1)1^2 = f(1). \end{aligned} \tag{202}$$

c) We claim p' interpolates $\boxed{(0, f'(0)).}$ To see this, note

$$p'(x) = (1 - 2x)f'(0) - 2xf(0) + 2xf(1), \tag{203}$$

which implies

$$p'(0) = (1 - 0)f'(0) - 2(0)f(0) + 2(0)f(1) = f'(0). \tag{204}$$

□

F12.03: Find a bound for the number of iterations of the bisection method needed to achieve an approximation with accuracy 10^{-3} to the solution of $x^3+x-4=0$ lying in the interval $[1,4]$. Justify your answer.

Solution:

Set $b_0 = 4$, $a_0 = 1$, $x^0 = (4+1)/2 = 5/2$, and $\varepsilon = 10^{-3}$. Then the bisection method for finding a root x^* of $f(x) = x^3 + x - 4$ is given by the algorithm

```
$k \leftarrow 0$
while $b_k - a_k \geq \varepsilon$ and $x^k \neq x^*$
  $x^k \leftarrow \frac{a_k + b_k}{2}$
  if $\text{sgn} f(b_k) \text{sgn} f(x_k) \leq 0$
    $a_{k+1} \leftarrow x^k$
    $b_{k+1} \leftarrow b_k$
  else
    $a_{k+1} \leftarrow a_k$
    $b_{k+1} \leftarrow x^k$
  $k \leftarrow k+1$
end while loop
```

By construction, $b_{k+1} - a_{k+1} = (b_k - a_k)/2 = (b_0 - a_0)/2^k$, which converges to zero as $k \longrightarrow \infty$. This implies the algorithm will terminate in a finite number of steps. And, because a root x^* of f is contained in $[a_k, b_k]$ for each index k and $x^k = (a_k + b_k)/2$, we discover

$$|x^k - x^*| \leq \frac{b_k - a_k}{2} = \frac{b_0 - a_0}{2^{k+1}} = 3 \cdot 2^{-(k+1)}. \tag{205}$$

Because $3 \cdot 2^{-(10+1)} = 3 \cdot 2^{-11} = 3/2048 > 10^{-3} = \varepsilon$ and $3 \cdot 2^{-(11+1)} = 3 \cdot 2^{-12} = 3/4096 < 10^{-3} = \varepsilon$, we see the number of iterations needed to approximate a root x^* of f within $\varepsilon = 10^{-3}$ is no more than $\boxed{\text{12 iterations.}}$ □

F12.04: For a single panel, the midpoint rule

$$\int_{x_{-1}}^{x_1} f(x)\ \mathrm{d}x = 2hf(x_0) + \frac{h^3}{3} f''(\xi), \tag{206}$$

where $x_1 - x_0 = x_0 - x_{-1} = h$ and $\xi \in (x_{-1}, x_1)$, is third order accurate. What is the order of the composite midpoint rule? Justify your answer.

Solution:

We claim the composite midpoint rule is second order accurate. Let $a, b \in \mathbb{R}$ with $a < b$ and set $x_j = a + (j+1)h$ for $j = -1, 0, \dots, n+1$ where $h = (b-a)/(n+2)$. Then $x_{-1} = a$, $x_{n+1} = b$, and

$$\int_{x_{-1}}^{x_{n+1}} f(x)\ \mathrm{d}x = \sum_{j=0}^{n/2} \int_{x_{j-1}}^{x_{j+1}} f(x)\ \mathrm{d}x = \sum_{j=0}^{n/2} 2hf(x_{2j}) + \frac{h^3}{3} f''(\xi_j) = 2h \sum_{j=0}^{n/2} f(x_{2j}) + \frac{h^3}{3} \sum_{j=0}^{n/2} f''(\xi_j), \tag{207}$$

where $\xi_j \in (x_{j-1}, x_{j+1})$ for each j. Now observe that

$$\frac{h^3}{3} \sum_{j=0}^{n/2} f''(\xi_j) = \frac{h^3}{3} \cdot \frac{n+2}{2} \cdot \underbrace{\left[\frac{1}{n/2+1} \cdot \sum_{j=0}^{n/2} f''(\xi_j) \right]}_{\overline{f''}} \tag{208}$$

and the underbraced term, denoted $\overline{f''}$, gives the average value of the $f''(\xi_j)$'s. This implies $\overline{f''} \in [\min_j f''(\xi_j),\ \max_j f''(\xi_j)]$ so that $\overline{f''}$ is contained in the range of f on $[a, b]$. By the intermediate value theorem and smoothness of f, we deduce there is $\mu \in [a, b]$ such that $f''(\mu) = \overline{f''}$. Whence, using the definition of h, the error for the composite midpoint rule may be expressed as

$$\frac{h^3}{3} \sum_{j=0}^{n/2} f''(\xi_j) = \frac{h^3}{3} \cdot \frac{n+2}{2} \cdot f''(\mu) = \frac{(b-a)h^2}{6} \cdot f''(\mu). \tag{209}$$

This shows the composite midpoint rule is second order accurate, as desired. □

F12.05: Consider the numerical method

$$y^* = y_{n-1} + \frac{2h}{3} f(y_{n-1}), \quad y_n = y_{n-1} + \frac{h}{4} f(y_{n-1}) + \frac{3h}{4} f(y^*) \tag{210}$$

to obtain approximations to

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y), \quad y(t_0) = y_0. \tag{211}$$

a) Assuming $f : \mathbb{R} \to \mathbb{R}$ is smooth, give the leading term of the expansion of the local truncation error for this method.

b) Derive the relation between $|e_n| = |y(t_n) - y_n|$ and $|e_{n-1}| = |y(t_{n-1}) - y_{n-1}|$ and the local truncation error. You may assume f has global Lipschitz constant L.

c) Give a derivation of an error bound that uses the results from a) and b) to obtain an error bound for this method over a time interval $[0, T]$.

Solution:

a) Set $\alpha := \frac{2h}{3} f(y_{n-1})$. Then Taylor expanding about $f(y^*)$ gives

$$f(y^*) = f(y_{n-1} + \alpha) = f(y_{n-1}) + \alpha f'(y_{n-1}) + \frac{\alpha^2}{2} f''(y_{n-1}) + \mathcal{O}(\alpha^3). \tag{212}$$

Using the formula for y_n and substituting for α, (212) implies

$$\begin{aligned} y_n &= y_{n-1} + \frac{h}{4} f(y_{n-1}) + \frac{3h}{4} f(y^*) \\ &= y_{n-1} + h f(y_{n-1}) + \frac{h^2}{2} f(y_{n-1}) f'(y_{n-1}) + \frac{h^3}{3} f(y_{n-1})^2 f''(y_{n-1}) + \mathcal{O}(h^4). \end{aligned} \tag{213}$$

Taylor expanding about $y(t_n)$, we also find

$$\begin{aligned} y(t_n) &= y(t_{n-1}) + h y'(t_{n-1}) + \frac{h^2}{2} y''(t_{n-1}) + \frac{h^3}{6} y'''(t_{n-1}) + \mathcal{O}(h^4) \\ &= y(t_{n-1}) + h y'(t_{n-1}) + \frac{h^2}{2} f'(y(t_{n-1})) f(y(t_{n-1})) \\ &+ \frac{h^3}{6} \left[f''(y(t_n)) f(y(t_n))^2 + f'(y(t_n))^2 f(y(t_n)) \right] + \mathcal{O}(h^4) \end{aligned} \tag{214}$$

Assume $y_{n-1} = y(t_{n-1})$. Then (213) and (214) together imply the local truncation error τ_n is

$$\tau_n := y_n - y(t_n) = \frac{h^3}{6}\left[f''(y_n)f(y_n)^2 - f'(y_n)^2 f(y_n)\right] + \mathcal{O}(h^4). \tag{215}$$

The right hand side give the leading term in the expansion of τ_n.

b)

c)

□

F12.06: a) Consider the initial value problem

$$\begin{cases} u_t &= u_x + v_x \\ v_t &= v_x \end{cases} \tag{216}$$

to be solve for $x \in [0,1]$, $t \geq 0$ with the initial and boundary conditions

$$u(x,0) = \phi(x), \quad u(1,t) = u(0,t), \quad v(x,0) = \psi(x), \quad v(1,t) = v(0,t) \tag{217}$$

where ϕ and ψ are smooth and periodic functions.

i) Can you write a stable, convergent finite difference scheme for this problem? Explain your answer.

ii) Give an example if one exists. Explain your answer.

b) Consider the related system

$$\begin{cases} u_t &= u_x + v_x \\ v_t &= \dfrac{1}{1000} u_x + v_x \end{cases} \tag{218}$$

with the same initial and boundary conditions.

i) Can you write a stable, convergent finite difference scheme for this problem? Explain your answer.

ii) Give an example if one exists. Explain your answer.

Solution:

a) i) Yes

ii)

b) i) No. We get eigenvalues $1 \pm \sqrt{\varepsilon}$ where $\varepsilon = 1/1000$.

ii)

□

Spring 2013

S13.01: Consider the linear system $Ax = b$ with $x, b \in \mathbb{R}^n$ and nonsingular $A = M - N \in \mathbb{R}^{n\times n}$.

a) If M is nonsingular and $(M^{-1}N)^k \longrightarrow 0$ as $k \longrightarrow \infty$, show that the iterates $\{x_k\}$ defined by

$$Mx_{k+1} = Nx_k + b \tag{219}$$

converge to $x = A^{-1}b$ for any starting vector x_0.

b) Find a splitting $A = M - N$ for the matrix

$$A = \begin{pmatrix} 10 & -1 \\ -1 & 10 \end{pmatrix} \tag{220}$$

so that the iteration in a) is convergent. Justify your answer.

Solution:

a) Set $T := M^{-1}N$ and $c := M^{-1}b$. Then the iteration for each k is defined by

$$x_k = Tx_{k-1} + c = T^2x_{k-1} + (T + I)\,c = \cdots = T^kx_0 + \left(\sum_{j=0}^{k-1} T^k\right) c. \tag{221}$$

Using our hypothesis that $\lim\limits_{k\to\infty} T^k = 0$, we also deduce

$$\lim_{k\to\infty} (I - T) \left(\sum_{j=0}^{k-1} T^k\right) = \lim_{k\to\infty} I - T^k = I - 0 = I \quad \Rightarrow \quad (I - T)^{-1} = \sum_{j=0}^{\infty} T^k. \tag{222}$$

Combining the above two results, we evaluate our limit to find

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} T^kx_0 + \left(\sum_{j=0}^{k-1} T^k\right) c = \lim_{k\to\infty} T^kx_0 + \lim_{k\to\infty} \left(\sum_{j=0}^{k-1} T^k\right) c = 0x_0 + (I - T)^{-1}c. \tag{223}$$

Back substituting for T and c gives our desired relation. Namely,

$$\lim_{k\to\infty} x_k = (I-T)^{-1}c = [M^{-1}(M-N)]^{-1}\left(M^{-1}b\right) = (M-N)^{-1}MM^{-1}b = A^{-1}Ib = A^{-1}b, \quad (224)$$

and we are done.

b) Take

$$M = \begin{pmatrix} 10 & 0 \\ -1 & 10 \end{pmatrix} \quad \text{and} \quad N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \quad (225)$$

Then

$$M^{-1}N = \begin{pmatrix} 10 & 0 \\ -1 & 10 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1/10 & 0 \\ 1/100 & 1/10 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1/10 \\ 0 & 1/100 \end{pmatrix}. \quad (226)$$

This implies

$$\det(M^{-1}N - \lambda I) = \begin{vmatrix} -\lambda & 1/10 \\ 0 & 1/100-\lambda \end{vmatrix} = \lambda(\lambda - 1/100), \quad (227)$$

and so $\rho(M^{-1}N) = 1/100 < 1$. Because $M^{-1}N$ is convergent if and only if $\rho(M^{-1}N) < 1$, we conclude $\lim_{k\to\infty} M^{-1}N = 0$. Then the result in a) establishes the convergence of the sequence $\{x_k\}$, as desired.

□

S13.02: Let $g \in C([a,b])$ with $g(x) \in [a,b]$ for all $x \in [a,b]$. Prove the following:

a) g has at least one fixed point in the interval $[a,b]$.

b) If there is a value $\gamma \in (0,1)$ such that

$$|g(x) - g(y)| \leq \gamma |x - y| \tag{228}$$

for all $x, y \in [a,b]$, then the fixed point p is unique and the iteration $x_{n+1} = g(x_n)$ converges to p for any initial guess $x_0 \in [a,b]$.

Proof:

a) If $g(a) = a$ or $g(b) = b$, then we are done. Now suppose this is not the case and set $f(x) = g(x) - x$. Then $f(a) = g(a) - a > 0$ and $f(b) = g(b) - b < 0$. Because g is continuous, so also is f. Then the intermediate value theorem asserts there is $p \in (a,b)$ such that $0 = f(p) = g(p) - p$, which is equivalent to asserting p is a fixed point of g. This completes the proof.

b) First observe that, for each nonnegative integer n,

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq \gamma |x_n - x_{n-1}| \leq \cdots \leq \gamma^n |x_1 - x_0|. \tag{229}$$

Now suppose $m, n \geq 0$ with $m > n$. Then

$$\begin{aligned} |x_m - x_n| &= |x_m - x_{m-1} + x_{m-1} + \cdots + x_n| \\ &\leq |x_m - x_{m-1}| + \cdots + |x_{n+1} - x_n| \\ &\leq \gamma^n |x_1 - x_0| \cdot \sum_{j=0}^{m-n} \gamma^j \\ &\leq \gamma^n |x_1 - x_0| \cdot \sum_{j=0}^{\infty} \gamma^j \\ &= \gamma^n \cdot \frac{x_1 - x_0}{1 - \gamma}. \end{aligned} \tag{230}$$

Because $\gamma^n \cdot \dfrac{x_1 - x_0}{1 - \gamma} \longrightarrow 0$ as $n \longrightarrow \infty$, we see $\{x_n\}$ is Cauchy. Because $\mathbb{R}$ is complete, this implies there is $\overline{x} \in \mathbb{R}$ such that $x_n \longrightarrow \overline{x}$. By the continuity of g, we discover

$$\overline{x} = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) = g\left(\lim_{n\to\infty} x_n\right) = g(\overline{x}), \tag{231}$$

i.e., $\overline{x}$ is a fixed point of g.

All that remains is to verify $\overline{x}$ is unique. If there were $\overline{y} \neq \overline{x}$ such that $g(\overline{y}) = \overline{y}$, then this would imply

$$|\overline{x} - \overline{y}| = |g(\overline{x}) - g(\overline{y})| \leq \gamma|\overline{x} - \overline{y}| \quad \Rightarrow \quad 1 = \frac{|\overline{x} - \overline{y}|}{|\overline{x} - \overline{y}|} \leq \gamma, \tag{232}$$

contradicting the fact $\gamma \in (0, 1)$. Whence $\overline{x}$ must be unique.

□

S13.03: Let $u : \mathbb{R}^2 \to \mathbb{R}$ be a smooth function.

a) For $(x, y) \in [0, \delta x] \times [0, \delta y]$ derive the bilinear interpolation formula for $u(x, y)$ that uses the function values $u(0, 0)$, $u(\delta x, 0)$, $u(0, \delta y)$, and $u(\delta x, \delta y)$ (e.g., the formula that results when you linearly interpolate in one direction followed by linear interpolation in the other direction).

b) Derive the leading term of error expansion for the error in the interpolated value when using the formula in a).

Solution:

a) Define the linear functions q_0 and $q_{\delta y}$ by

$$q_0(x) := u(0,0)\frac{x-\delta x}{0-\delta x} + u(\delta x, 0)\frac{x-0}{\delta x - 0} \quad \text{and} \quad q_{\delta y} := u(0,\delta y)\frac{x-\delta x}{0-\delta x} + u(\delta x, \delta y)\frac{x-0}{\delta x-0}. \tag{233}$$

Then q_0 and $q_{\delta y}$ are linear interpolations along $y = 0$ and $y = \delta y$, respectively. We linearly interpolate along the direction of the y axis between points $(x, 0)$ and $(x, \delta y)$ in $[0, \delta x] \times [0, \delta y]$ by defining the bilinear interpolation p such that

$$p(x,y) := q_0(x)\frac{y-\delta y}{0-\delta y} + q_{\delta y}(x)\frac{y-0}{\delta y-0}. \tag{234}$$

This gives the desired bilinear interpolation.

b) Let $(x, y) \in [0, \delta x] \times [0, \delta y]$ be given and set

$$(x^*, y^*) := \text{argmin}\left\{\|(x^*,y^*)-(x,y)\| \ : \ (x^*,y^*) \in \{(0,0),(\delta x,0),(0,\delta y),(\delta x,\delta y)\}\right\}. \tag{235}$$

This implies $\|(x, y) - (x^*, y^*)\| \leq \frac{1}{2}\sqrt{(\delta x)^2 + (\delta y)^2}$. Now define the error $e(x, y) := u(x, y) - p(x, y)$. Then through Taylor expansion we discover

$$0 = e(x^*,y^*) = e(x,y) + \langle (x^*,y^*)-(x,y), \ De(x,y)\rangle + (\text{higher order terms}). \tag{236}$$

This shows, taking only the leading term of the error expansion,

$$\begin{aligned}|u(x,y)-p(x,y)| = |e(x,y)| &\approx |\langle (x^*,y^*)-(x,y),\ De(x,y)\rangle| \\ &\leq \|(x^*,y^*)-(x,y)\|\|De(x,y)\| \\ &\leq \frac{1}{2}\sqrt{(\delta x)^2+(\delta y)^2}\sqrt{(u_x(x,y)-p_x(x,y))^2+(u_y(x,y)-p_y(x,y))^2}.\end{aligned} \tag{237}$$

Note u_x and u_y are unknown; however, p_x and p_y can be computed directly from the definition of p above. The term on the right side of the $\approx$ in (237) gives the leading error term and the final inequality gives a more explicit bound for this term.

□

S13.04: Let Δ_h be the following three point difference operator that approximates $\mathrm{d}^2u/\mathrm{d}x^2$ using a mesh spacing h, i.e.,

$$\Delta_h u := \frac{u(x+h)-2u(x)+u(x-h)}{h^2}. \tag{238}$$

a) Derive the combination of Δ_h and Δ_{2h} that yields a fourth order approximation to $\mathrm{d}^2u/\mathrm{d}x^2$.

b) Give a derivation of the leading term of the local truncation error for the difference approximation you obtained in a).

Solution:

a) First we Taylor expand about x to discover

$$u(x\pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) \pm \frac{h^3}{3!}u^{(3)}(x) + \frac{h^4}{4!}u^{(4)}(x) \pm \frac{h^5}{5!}u^{(5)}(x) + \frac{h^6}{6!}u^{(6)}(x) + \mathcal{O}(h^7). \tag{239}$$

Adding the expansions for $u(x+h)$ and $u(x-h)$, subtracting $2u(x)$, and then dividing by h^2, we obtain

$$\Delta_h u = 0 + u''(x) + 0 + \frac{h^2}{4!}u^{(4)}(x) + \frac{h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5). \tag{240}$$

This further implies

$$\Delta_{2h} u = u''(x) + \frac{4h^2}{4!}u^{(4)}(x) + \frac{16h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5). \tag{241}$$

We must form a combination of $\Delta_h u$ and $\Delta_{2h}u$ that cancels the $\mathcal{O}(h^2)$ terms. Thus, we write

$$\begin{aligned}\frac{4\Delta_h u - \Delta_{2h}u}{3} &= \frac{1}{3}\left[4\left(u''(x) + \frac{h^2}{4!}u^{(4)}(x) + \frac{h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5)\right)\right.\\ &\qquad \left. - \left(u''(x) + \frac{4h^2}{4!}u^{(4)}(x) + \frac{16h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5)\right)\right]\\ &= u''(x) - \frac{4h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5).\end{aligned} \tag{242}$$

This shows we should form the combination $(4\Delta_h u - \Delta_{2h}u)/3$ to obtain a fourth order approximation to $\mathrm{d}^2u/\mathrm{d}x^2$.

b) Referring to (242), we see the local truncation error $\tau(x)$ for the difference approximation is given by

$$\tau(x) = u''(x) - \frac{4\Delta_h u - \Delta_{2h}u}{3} = u''(x) - \left[u''(x) - \frac{4h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5)\right] = \frac{4h^4}{6!}u^{(6)}(x) + \mathcal{O}(h^5). \tag{243}$$

Thus the leading term of $\tau(x)$ is $\boxed{4h^4u^{(6)}(x)/6!.}$

□

S13.05: Consider the following general 2-stage explicit Runge-Kutta method for advancing the solution of $\mathrm{d}y/\mathrm{d}t = F(y)$ with time step k,

$$y^* = y^n + \alpha k F(y^n), \quad y^{n+1} = y^n + \beta k F(y^n) + \gamma k F(y^*). \tag{244}$$

a) Derive conditions on the coefficients α, β, and γ that ensure the method converges to at least *first* order.

b) Assuming the coefficients of the method are selected so that it is first order, derive the expression that determines the interval of absolute stability for the method.

c) Show that there is at least one set of value $\alpha, \beta, \gamma > 0$ so that the resulting method is first order accurate and has an interval of absolute stability that is larger than $[-2, 0]$ (the latter being the interval of absolute stability for all second order methods of the given form).

Solution:

a) Henceforth assume $y^n := y(t^n)$. We first Taylor expand our Runge-Kutta method about y^n to find

$$\begin{aligned} y^n + \beta k F(y^n) + \gamma k F(y^*) &= y^n + \beta k F(y^n) + \gamma k F\left(y^n + \alpha k F(y^n)\right) \\ &= y^n + \beta k F(y^n) + \gamma k \left[F(y^n) + \alpha k F'(y^n) F(y^n) + \mathcal{O}(k^2)\right] \\ &= y^n + k F(y^n)\left[\beta + \gamma\right] + \gamma \alpha k^2 F'(y^n) F(y^n) + \mathcal{O}(k^3). \end{aligned} \tag{245}$$

This implies the local truncation error τ^{n+1} for the method is

$$\begin{aligned} \tau^{n+1} &:= y^{n+1} - y^n + \beta k F(y^n) + \gamma k F(y^*) \\ &= \left(y^n + k F(y^n) + \frac{k^2}{2} F'(y^n) F(y^n) + \mathcal{O}(k^3)\right) \\ &\quad - \left(y^n + k F(y^n)\left[\beta + \gamma\right] + \gamma \alpha k^2 F'(y^n) F(y^n) + \mathcal{O}(k^3)\right) \\ &= k F(y^n)\left[1 - \beta - \gamma\right] + k^2 F'(y^n) F(y^n) \left[\frac{1}{2} - \gamma \alpha\right] + \mathcal{O}(k^3). \end{aligned} \tag{246}$$

In order to ensure the convergence is at least first order, we need $\tau^{n+1} = \mathcal{O}(k^2)$, which is obtained precisely when $\boxed{1 = \beta + \gamma.}$

b) Let w^{n+1} be the $(n+1)$-st step of the method and w^n be the n-th step. Then for the model equation

$F(y) = \lambda y$ for some $\lambda \in \mathbb{C}$ we discover

$$w^{n+1} = w^n + \beta k\lambda y^n + \gamma k\lambda \left(y^n + \alpha k\lambda y^n\right) = \left(1 + (\beta+\gamma)k\lambda + \alpha\gamma(k\lambda)^2\right) w^n = \left(1 + k\lambda + \alpha\gamma(k\lambda)^2\right) w^n \tag{247}$$

where the final equality holds by assuming $\beta + \gamma = 1$. Recall the region of absolute stability is the set of all $k\lambda$ such that $w^n \longrightarrow 0$ for all initial conditions w^0. Since w^{n+1} is a scalar multiple of w^n, this is obtained precisely when

$$\left|1 + k\lambda + \alpha\gamma(k\lambda)^2\right| < 1. \tag{248}$$

The interval of absolute stability is the intersection of the real axis with the region of absolute stability. So, taking $\lambda \in \mathbb{R}$, the expression determining the interval of absolute stability for the method is

$$-2 < k\lambda + \alpha\gamma(k\lambda)^2 < 0. \tag{249}$$

c) Take $\alpha = 1/10$, $\beta = \gamma = 1/2$. Then $\beta + \gamma = 1$ and $\alpha\gamma = 5$, which implies the interval of stability is

$$-2 < k\lambda\left(1 + \frac{1}{20}(k\lambda)\right) < 0. \tag{250}$$

In order to satisfy the right hand inequality, we need $-20 < k\lambda < 0$. When this holds, $k\lambda/20 > -1$, and so

$$k\lambda\left(1 + \frac{1}{20}(k\lambda)\right) > k\lambda\,(1-1) = 0 > -2. \tag{251}$$

This shows the interval of absolute stability is $(-20, 0)$, which is larger than $(-2, 0)$. As noted in the problem statement, if $\alpha = 1$ so that $\alpha\gamma = 1/2$, then we would obtain a second order method (c.f. (246)) and have interval of absolute stability $(-2, 0)$. Note: There are differing definitions of absolute stability, which explains the potential discrepancy between my soft brackets and the hard brackets used in the problem statement.

□

Fall 2013

F13.01: Consider a piecewise linear interpolant $L(x)$ to $\sin(kx)$ with $k \in \mathbb{Z}$, $x \in [0, 2\pi]$, based upon $N + 1$ equispaced points x_j where $x_j = 2\pi j/N$.

a) Give a derivation of an estimate of the smallest integer N, denoted N^*, depending on k, such that

$$\max_{x \in [0,2\pi]} |\sin(kx) - L(x)| < 0.01. \tag{252}$$

b) Does the number of points *per wavelength* required to insure a bound depend on the value of k?

Solution:

a) Define $g(x) := \sin(kx) - L(x)$. Then For $j = 0, 1, \ldots, N$, define

$$z_j := \operatorname*{argmax}_{x \in [x_j, x_{j+1}]} |g(x)|. \tag{253}$$

Since g is smooth and $g(x_j) = 0$ for $j = 0, 1, \ldots, N$, we know $g'(z_j) = 0$. Now define x_{k_j} to be the closest of the two points x_j and x_{j+1} to z_j. This ensures

$$|x_{k_j} - z_j| \leq \frac{x_{j+1} - x_j}{2} = \frac{\pi}{N}. \tag{254}$$

Then Taylor's theorem asserts there is ξ between z_j and x_{k_j} such that

$$0 = g(x_{k_j}) = g(z_j) + \cancelto{0}{g'(z_j)} \cdot (x_{k_j} - z_j) + \frac{g''(\xi)}{2}(x_{k_j} - z_j)^2 \quad \Rightarrow \quad g(z_j) = -\frac{g''(\xi)}{2}(x_{k_j} - z_j)^2. \tag{255}$$

Note $g''(\xi) = -k^2 \sin(kx) - L''(x) = -k^2 \sin(kx)$ since L is linear. Then

$$|g(z_j)| = \left|\frac{g''(\xi)}{2}\right| \left|x_{k_j} - z_j\right|^2 \leq \left|\frac{g''(\xi)}{2}\right| \left(\frac{\pi}{N}\right)^2 = \left|\frac{-k^2 \sin(k\xi)}{2}\right| \left(\frac{\pi}{N}\right)^2 \leq \frac{1}{2}\left(\frac{k\pi}{N}\right)^2. \tag{256}$$

To obtain the desired error bound, we see the right hand side of (256) to be less than the error bound, i.e.,

$$\frac{1}{2}\left(\frac{k\pi}{N}\right)^2 < \frac{1}{100} \quad \Rightarrow \quad \frac{1}{\sqrt{2}}\frac{k\pi}{N} < \frac{1}{10} \quad \Rightarrow \quad N > \frac{10k\pi}{\sqrt{2}}. \tag{257}$$

With this, we estimate N^* to be the smallest integer greater than $10k\pi/\sqrt{2}$.

b) No, the number of points *per wavelength* needed for our bound does *not* depend on k. Our expression giving N^* is linear in k, and so dividing $10k\pi/\sqrt{2}$ by k to obtain the points *per wavelength* gives $10\pi/\sqrt{2}$, which is independent of k.

□

F13.02: Consider the $(m_1 + m_2) \times (m_1 + m_2)$ block matrix

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}, \tag{258}$$

where $A_{1,1} \in \mathbb{R}^{m_1 \times m_1}$, $A_{1,2} \in \mathbb{R}^{m_1 \times m_2}$, $A_{2,1} \in \mathbb{R}^{m_2 \times m_1}$, and $A_{2,2} \in \mathbb{R}^{m_2 \times m_2}$.

a) Derive an expression for a lower block triangular matrix L and an upper block triangular matrix U in terms of the block components of A such that $LA = U$.

b) Consider the system of equations

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}. \tag{259}$$

Derive an expression for the $m_2 \times m_2$ matrix S and the vector $\tilde{f}$ in terms of the block components of A and components of f such that

$$Sx_2 = \tilde{f} \tag{260}$$

is the set of equations that determines the x_2 component fo the solution of the original system. *The equations you derive should not include x_1.*

Solution:

In this solution, we assume $A_{1,1}$ and $A_{2,2}$ are invertible.

a) Let $L_{i,j}$ and $U_{i,j}$ denote the corresponding blocks in L and U, respectively. Then we must find L and U satisfying

$$\begin{bmatrix} L_{1,1} & 0 \\ L_{2,1} & L_{2,2} \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} = \begin{bmatrix} U_{1,1} & U_{1,2} \\ 0 & U_{2,2} \end{bmatrix}. \tag{261}$$

This gives rise to 4 matrix equations, but there are 6 unknowns. So take $L_{1,1} = I_{m_1}$ and $L_{2,2} = I_{m_2}$. Then note we have

$$0 = L_{2,1}A_{1,1} + L_{2,2}A_{2,1} = L_{2,1}A_{1,1} + A_{2,1} \quad \Rightarrow \quad L_{2,1} = -A_{2,1}A_{1,1}^{-1}. \tag{262}$$

We now have each block entry of L. Then by direct computation we obtain

$$\underbrace{\begin{bmatrix} I_{m_1} & 0 \\ -A_{2,1}A_{1,1}^{-1} & I_{m_2} \end{bmatrix}}_{L}\underbrace{\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & -A_{2,1}A_{1,1}^{-1}A_{1,2} + A_{2,2} \end{bmatrix}}_{U}, \tag{263}$$

and we are done.

b) From the linear system we obtain the equations

$$f_1 = A_{11}x_1 + A_{12}x_2, \tag{264}$$

$$f_2 = A_{21}x_1 + A_{22}x_2. \tag{265}$$

From (264), we write $x_1 = A_{11}^{-1}(f_1 - A_{12}x_2)$. Then substituting this for x_1 in (265) yields

$$f_2 = A_{21}A_{11}^{-1}(f_1 - A_{12}x_2) + A_{22}x_2 = A_{21}A_{11}^{-1}f_1 + \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)x_2. \tag{266}$$

Rearranging, we obtain

$$\underbrace{\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)}_{S}x_2 = \underbrace{\left(f_2 - A_{21}A_{11}^{-1}f_1\right)}_{\tilde{f}}, \tag{267}$$

and so S and $\tilde{f}$ are given by the respective underbraced quantities.

□

Can the above problem be done without assuming invertibility???

F13.03: Find the approximation of the integral $\int_{-1}^{1} f(x)\ \mathrm{d}x$ using a Gaussian quadrature formula $c_1 f(x_1) + c_2 f(x_2)$. Give the degree of precision of the approximation.

Solution:

We must find $c_1, c_2, x_1, x_2 \in \mathbb{R}$ such that

$$c_1 x_1^k + c_2 x_2^k = \int_{-1}^{1} x^k\ \mathrm{d}x \quad \text{for } k = 0, 1, 2, 3, \tag{268}$$

i.e., so the approximation has degree of precision 3. This follows from the fact we have 4 unknowns and each value k gives one equation. Note this is only obtainable if each variable is nonzero. Evaluating the integral at each k, we obtain

$$2 = c_1 + c_2, \tag{269}$$

$$0 = c_1 x_1 + c_2 x_2, \tag{270}$$

$$\frac{2}{3} = c_1 x_1^2 + c_2 x_2^2, \tag{271}$$

$$0 = c_1 x_1^3 + c_2 x_2^3. \tag{272}$$

From (270), we deduce $c_1 x_1 = -c_2 x_2$. Plugging this into (271), we discover

$$\frac{2}{3} = c_1 x_1^2 + c_2 x_2^2 = c_1 x_1 (x_1 - x_2) \quad \Rightarrow \quad x_1 - x_2 = \frac{2}{3 c_1 x_1}. \tag{273}$$

Together with (272) this reveals

$$0 = c_1 x_1^3 + c_2 x_2^3 = c_1 x_1 (x_1 - x_2)(x_1 + x_2) = \frac{2}{3}(x_1 + x_2) \quad \Rightarrow \quad x_1 = -x_2. \tag{274}$$

This implies $c_1 x_1 = -c_2 x_2 = c_2 x_1$ and so $c_1 = c_2$. But, (269) then shows $c_1 = c_2 = 1$. Returning to (271), we see

$$\frac{2}{3} = x_1^2 + x_2^2 = x_1^2 + (-x_1)^2 \quad \Rightarrow \quad x_1 = \pm\frac{1}{\sqrt{3}}. \tag{275}$$

Thus we conclude $\boxed{c_1 = c_2 = 1,\ x_1 = 1/\sqrt{3}, \text{ and } x_2 = -1/\sqrt{3}.}$ The degree of precision is 3. □

F13.04: Let $f(0)$, $f(h)$, $f(2h)$ be the value of a real valued function at $x = 0$, $x = h$, and $x = 2h$.

a) Derive the coefficients c_0, c_1, and c_2 so that

$$Df_h(x) = c_0 f(0) = c_1 f(h) + c_2 f(2h) \tag{276}$$

is as accurate an approximation to $f'(0)$ as possible.

b) Derive the leading term of a truncation error estimate for the formula derived in a).

Solution:

a) We first Taylor expand f about $x = 0$ to discover

$$f(h) = f(0) + f'(0)h + \frac{f''(0)}{2}h^2 + \frac{f'''(0)}{6}h^3 + \mathcal{O}(h^4), \tag{277}$$

$$f(2h) = f(0) + f'(0)h + \frac{f''(0)}{2}4h^2 + \frac{f'''}{6}8h^3 + \mathcal{O}(h^4). \tag{278}$$

Then we multiply (277) by 4, subtract (278), and then divide by 2 to obtain

$$2f(h) - \frac{f(2h)}{2} = \frac{3f(0)}{2} + f'(0) + 0 - \frac{f'''(0)}{3}h^3 + \mathcal{O}(h^4). \tag{279}$$

Note this linear combination makes the $\mathcal{O}(h^2)$ terms cancel. Then through rearranging term we see

$$f'(0) = -\frac{3}{2}f(0) + 2f(h) - \frac{1}{2}f(2h) + \frac{f'''(0)}{3}h^3 + \mathcal{O}(h^4). \tag{280}$$

This shows $\boxed{c_0 = 3/2, c_1 = 2, c_2 = -1/2.}$

b) As seen in (280), the leading term of truncation error for the formula in a) is given by

$$\boxed{\frac{f'''(0)}{3}h^3.} \tag{281}$$

□

F13.05: Consider the following ODE method for creating approximate solutions of $\mathrm{d}y/\mathrm{d}t = F(y)$ with time step k,

$$y_n = \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + k\frac{2}{3}F(y_n). \tag{282}$$

a) Derive an expression for the local truncation error.

b) Show that this method satisfies the root condition.

c) Is this a convergent method? If it is, what is the global order of accuracy of this method?

d) Derive the conditions that determine the region of absolute stability of this method.

Solution:

a) Assume $y_n = y(t_n)$ for each n and, instead of the notation given, let $\tilde{y}_n := 4\tilde{y}_{n-1}/3 - \tilde{y}_{n-2}/3 + 2kF(\tilde{y}_n)/3$ denote the ODE method. Through Taylor expansion about y_n, we discover

$$y_{n-1} = y_n - ky_n' - \frac{k^2}{2}y_n'' - \frac{k^3}{6}y_n''' + \mathcal{O}(k^4), \tag{283}$$

$$y_{n-2} = y_n - 2ky_n' - -\frac{4k^3}{3}y_n''' + \mathcal{O}(k^4). \tag{284}$$

This implies

$$\begin{aligned} &\frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + \frac{2k}{3}y_n' \\ &= y_n + \underbrace{\left(-\frac{4k}{3}y_n' + \frac{2k}{3}y_n' + \frac{2}{3}y_n'\right)}_{=0} + \underbrace{\left(-\frac{2k^2}{3}y_n'' + \frac{2k^2}{3}y_n''\right)}_{=0} + \left(-\frac{2k^3}{9}y_n''' + \frac{4k^3}{9}y_n'''\right) + \mathcal{O}(k^4) \\ &= y_n + \frac{2k^3}{9}y_n''' + \mathcal{O}(k^4). \end{aligned} \tag{285}$$

Thus, if $\tilde{y}_{n-1} = y_{n-1}$ and $\tilde{y}_{n-2} = y_{n-2}$, then we see the local truncation error τ_n is given by

$$\tau_n := y(t_n) - y_n = -\frac{2k^3}{9}y_n''' + \mathcal{O}(k^4). \tag{286}$$

b) The characteristic polynomial $\chi(\lambda)$ for this method is given by

$$\chi(\lambda) := \lambda^2 - \frac{4}{3}\lambda - \left(-\frac{1}{3}\right) = \lambda^2 - \frac{4}{3}\lambda + \frac{1}{3}. \tag{287}$$

This implies the roots are given by

$$\lambda = \frac{4/3 \pm \sqrt{16/9 - 4\cdot 1 \cdot 1/3}}{2\cdot 1} = \frac{4/3 \pm 2/3}{2} = \frac{2\pm 1}{3}. \tag{288}$$

That is, $\lambda = 1/3$ and $\lambda = 1$ are roots of the characteristic polynomial. Since each root of the characteristic polynomial $\chi(\lambda)$ either has norm less than unity or has norm unity and is simple, we see the root condition is satisfied.

c) Observe that

$$\lim_{k\to 0}\left|\frac{\tau_n(k)}{k}\right| = \lim_{k\to 0}\left|-\frac{2k^2}{3}y_n''' + \mathcal{O}(k^3)\right| = 0. \tag{289}$$

Since this holds for each n, the method is consistent. Since the method also satisfies the root condition, it is stable. And, a multi-step method of this form is consistent and stable if and only if it is convergent. Whence we conclude the method is convergent. And, since $\tau_n(k) = \mathcal{O}(k^3)$, the order of accuracy of this method is $\mathcal{O}(k^2)$.

d) Applying the method to the test equation $y' = \lambda y$, we obtain

$$\left(1 - k\lambda\frac{2}{3}\right)\tilde{y}_n - \frac{4}{3}\tilde{y}_{n-1} - \left(-\frac{1}{3}\right)\tilde{y}_{n-2} = 0. \tag{290}$$

The characteristic polynomial $Q(z;k\lambda)$ associated with the test equation for this method is

$$Q(z;k\lambda) := \left(1 - k\lambda\frac{2}{3}\right)z^2 - \left(\frac{4}{3}\right)z - \left(-\frac{1}{3}\right). \tag{291}$$

The region of absolute stability is given by the set of all $k\lambda$ such that the norm of every root of $Q(z;k\lambda)$ is less than unity. The zeros of $Q(z;k\lambda)$ are

$$z = \frac{4/3 \pm \sqrt{16/9 - 4(1-2k\lambda/3)(1/3)}}{2(1-2k\lambda/3)} = \frac{4/3 \pm \sqrt{(4/9)(1+2k\lambda)}}{2(1-2k\lambda/3)} = \frac{1 \pm \frac{1}{2}\sqrt{1+2k\lambda}}{3-2k\lambda}. \tag{292}$$

Thus the region of absolute stability is defined by the set

$$\boxed{\left\{k\lambda \in \mathbb{C} \ : \ \left|\frac{1\pm\frac{1}{2}\sqrt{1+2k\lambda}}{3-2k\lambda}\right| < 1\right\}.} \tag{293}$$

□

F13.08: Consider the problem in two dimensions

$$\begin{cases} -\Delta u + u = f, & \text{in } T, \\ u = g_1, & \text{on } T_1, \\ u = g_2, & \text{on } T_2, \\ \dfrac{\partial u}{\partial n} = h, & \text{on } T_3, \end{cases} \tag{294}$$

where

$$\begin{aligned} T &= \{(x,y) \ : \ x > 0, \ y > 0, \ x + y < 1\}, \\ T_1 &= \{(x,y) \ : \ 0 < x < 1, \ y = 0\}, \\ T_2 &= \{(x,y) \ : \ x = 0, \ 0 < y < 1\}, \\ T_3 &= \{(x,y) \ : \ x > 0, \ y > 0, \ x + y = 1\}. \end{aligned} \tag{295}$$

a) Find the weak variational formulation of the problem and verify the assumptions of the Lax-Milgram Lemma by analyzing the appropriate bilinear and linear forms. Impose the weakest necessary assumptions on the functions f, g_1, g_2, and h.

b) Develop and describe the piecewise linear Galerkin finite element approximation of the problem and a set of basis functions such that the corresponding linear system is sparse. Show that this linear system has a unique solution. Give a convergence estimate and quote the appropriate theorems for convergence.

Solution:

a)

b)

□

Spring 2014

S14.04: Consider the implicit Euler's method (or the backwards Euler's method)

$$y_{i+1} = y_i + hf\left(x_{i+1}, y_{i+1}\right) \tag{296}$$

for the ODE $y' = f(x, y)$ with $y(0)$ the initial condition. Derive the region of absolute stability for the method. Give an ODE for which $\partial f/\partial y > 0$. Does the backwards Euler always give the qualitatively correct solution? Explain.

Solution:

To derive the region of absolute stability, we use the model equation. That is, we assume $f(x, y) = \lambda y$ for some $\lambda \in \mathbb{C}$. Then the method becomes

$$y_{i+1} = y_i + h\lambda y_{i+1} \qquad \Rightarrow \qquad y_{i+1} = \frac{1}{1-h\lambda} y_i. \tag{297}$$

Recall the region of absolute stability is defined to be the set of all $h\lambda$ such that $y_i \longrightarrow 0$ for all initial conditions $y(0)$. Since y_{i+1} is a scalar multiple of y_i, this is accomplished if $|y_{i+1}| = \alpha|y_i|$ for some $\alpha \in (0, 1)$, i.e.,

$$\left|\frac{1}{1-h\lambda}\right| < 1. \qquad \Rightarrow \qquad \frac{1}{|1-h\lambda|} < 1 \qquad \Rightarrow \qquad 1 < |1 - h\lambda|. \tag{298}$$

This means the region of absolute stability is the entire complex plane except for the unit circle centered at $(1, 0)$. In mathematical terms, the region of absolute stability for the backwards Euler method is the set $\{h\lambda \in \mathbb{C} \ : \ 1 < |1 - h\lambda|\}$.

An example of an ODE for which $\partial f/\partial y > 0$ is $y' = f(x, y) = y$ since here $\partial f/\partial y = 1 > 0$. And, no, the backwards Euler method does not always give the qualitatively correct solution. We demonstrate this with the provided example ODE. The solution to this ODE is well-known to be $y = y(0)e^x$, which grows like an exponential. However, the iteration for this method becomes

$$y_{i+1} = \frac{1}{1-h} y_i. \tag{299}$$

Using our above result, the region of absolute stability is where $h > 2$. However, this implies $1/(1-h) < 0$, and so the sign of y_{i+1} is the opposite of the sign of y_i. Thus there will be oscillatory behavior exhibited for h in the stability region, which is not qualitatively correct. And, when $h > 2$, $|y_{i+1}| < |y_i|$ so that the iterates are not growing in time. This also shows the iteration is not qualitatively correct and completes our solution. □

S14.05: Let $f(y)$ and $g(y)$ be smooth real valued functions of y and consider the differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y) + g(y), \quad y(0) = y_0. \tag{300}$$

a) Derive the leading term of the local truncation error for the method

$$y^* = y^n + kf(y^n), \quad y^{n+1} = y^* + kg(y^{n+1}). \tag{301}$$

b) Assume one can evaluate the derivatives $\mathrm{d}f/\mathrm{d}y$ and $\mathrm{d}g/\mathrm{d}y$. Determine additional terms, which may incorporate these derivatives, that can be added to the method in a) and result in a higher order method. Justify your results.

Solution:

a) We henceforth take $y^n := y(t^n)$ and let $f^n := f(y^n)$ and $g^n = g(y^n)$. We also let Φ define the subsequent iterate of the method when given appropriate inputs. Taylor expansion about y^n gives

$$\begin{aligned}\Phi(y^n, f^n, g^{n+1}) &:= y^n + k\left(f^n + g^{n+1}\right)\\ &= y^n + k\left(f^n + g^n + k(g^n)'(y^n)' + \mathcal{O}(k^2)\right)\\ &= y^n k(y^n)' + k^2(g^n)'(y^n)' + \mathcal{O}(k^3).\end{aligned} \tag{302}$$

This implies the local truncation error τ^{n+1} is given by

$$\begin{aligned}
\tau^{n+1} &:= y^{n+1} - \Phi(y^n, f^n, g^{n+1}) \\
&= \left(y^n + k(y^n)' + \frac{k^2}{2}(y^n)'' + \mathcal{O}(k^3)\right) - \left(y^n k(y^n)' + k^2(g^n)'(y^n)' + \mathcal{O}(k^3)\right) \\
&= k^2\left(\frac{(y^n)''}{2} - (g^n)'(y^n)'\right) + \mathcal{O}(k^3) \\
&= k^2\left(\frac{(f^n+g^n)'(f^n+g^n)}{2} - (g^n)'(f^n+g^n)\right) + \mathcal{O}(k^3) \\
&= \frac{k^2}{2}\left(f^n - g^n\right)'\left(f^n+g^n\right) + \mathcal{O}(k^3).
\end{aligned} \tag{303}$$

The first term on the final line gives the leading term of τ^{n+1}.

b) Assuming we can evaluate $(f^n)'$ and $(g^n)'$, then we can incorporate the leading term of τ^{n+1} derive in (303) into our method Φ to obtain a new method $\tilde{\Phi}$ with local truncation error $\tilde{\tau}^{n+1} = \mathcal{O}(k^3)$. Indeed, define a new method $\tilde{\Phi}$ by

$$\tilde{\Phi}(y^n, f^n, g^{n+1}, (f^n)', (g^n)') := y^n + k\left(f^n + g^{n+1}\right) + \frac{k^2}{2}\left(f^n - g^n\right)'\left(f^n + g^n\right). \tag{304}$$

□

Fall 2014

F14.05: Let $f(y)$ be a smooth real valued function of y and consider the differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y), \quad y(0) = y_0. \tag{305}$$

Consider a numerical method that consists of using the Backward Euler (BE) method with a time step Δt and a fixed number, M, of fixed-point iterations ot create approximate solutions of the equations that must be solved to advance the solution one time step. Specifically, to determine y_{n+1} from y_n one uses

i) $\tilde{y}^0 = y_n$

ii) $\tilde{y}^k = y_n + \Delta t f(\tilde{y}^{k-1}) \quad$ for $k = 1, 2, \ldots, M$.

iii) $y_{n+1} = \tilde{y}^M$.

a) Derive the interval of absolute stability for the method when $M = 2$.

b) How does the interval of absolute stability determined in a) compare with that of the BE method if one assumes that the implicit equations are solved exactly?

c) How does the interval of absolute stability determined in a) compare with the of the Forward Euler (FE) method?

d) For general f, derive constraints on the time step in terms of f and it's derivatives that will ensure that the fixed point iteration in i)) will converge as $M \longrightarrow \infty$.

e) Briefly comment on the advisability or inadvisability of using the BE method in combination with fixed point iteration to create approximate solutions of stiff ODE's.

Solution:

a) When $M = 2$, we obtain

$$y_{n+1} = \tilde{y}^2 = y_n + \Delta t f(\tilde{y}^1) = y_n + \Delta t f\left(y_n + \Delta t f(y_n)\right). \tag{306}$$

To derive the region of absolute stability, we assume $f(y) = \lambda y$ for some $\lambda \in \mathbb{C}$. This yields

$$y_{n+1} = y_n + \Delta t\lambda(y_n + \Delta t\lambda y_n) = \left(1 + \Delta t\lambda + (\Delta t\lambda)^2\right) y_n. \tag{307}$$

Recall the region of absolute stability is defined to be the set of $\Delta t\lambda$ such that $y_n \longrightarrow 0$ for all initial conditions y_0. Since y_{n+1} is here a scalar multiple of y_n, this will be accomplished if there is $\alpha \in (0,1)$ such that $|y_{n+1}| = \alpha |y_n|$, i.e.,

$$\left|1 + \Delta t\lambda + (\Delta t\lambda)^2\right| < 1. \tag{308}$$

The interval of absolute stability is defined to be the intersection of the region of absolute stability with the real axis. So, assuming $\Delta t\lambda \in \mathbb{R}$, we obtain the constraint

$$-1 < 1 + \Delta t\lambda + (\Delta t\lambda)^2 < 1 \quad \Rightarrow \quad -2 < \Delta t(\Delta t + 1) < 0. \tag{309}$$

If $\Delta t + 1 < 0$, then $\Delta t < -1 < 0$, which implies $\Delta t(\Delta t + 1) > 0$. So, in order for the right hand constraint to be satisfied, we need $\Delta t + 1 > 0$ and $\Delta t < 0$, which implies $-1 < \Delta t < 0$. This implies

$$\Delta t(\Delta t + 1) = (\Delta t)^2 + \Delta t > 0 + (-1) = -1, \tag{310}$$

and so both conditions in (309) are satisfied. Thus the interval of absolute stability for the method when $M = 2$ is $(-1, 0)$.

b) For the BE method, we have

$$y_{n+1} = y_n + \Delta t f(y_{n+1})y_n + \Delta t\lambda y_{n+1} \quad \Rightarrow \quad y_{n+1} = \frac{1}{1 - \Delta t\lambda} y_n, \tag{311}$$

where the second equality follows from assuming $f(y) = \lambda y$. Following in similar fashion as in a), we see deduce the region of absolute stability is when $1/|1 - \Delta t\lambda| < 1$, equivalently expressed as $|1 - \Delta t\lambda| > 1$. This is precisely all points outside the unit circle centered at $(1, 0)$. So, the interval of absolute stability for the BE method is $(-\infty, 0) \cup (2, \infty)$. This is larger than the interval of absolute stability $(-1, 0)$ obtained in a), and contains $(-1, 0)$ as a proper subset.

c) For the FE method, we have

$$y_{n+1} = y_n + \Delta t f(y_n)y_n + \Delta t\lambda y_n \quad \Rightarrow \quad y_{n+1} = 1 + \Delta t\lambda y_n, \tag{312}$$

where the second equality follows from assuming $f(y) = \lambda y$. Following in similar fashion as in a), we see deduce the region of absolute stability is when $|1 + \Delta t\lambda| < 1$. This is precisely all points inside the unit circle centered at $(-1,0)$. So, the interval of absolute stability for the FE method is $(-2,0)$. This is larger than the interval of absolute stability $(-1,0)$ obtained in a), and contains $(-1,0)$ as a proper subset.

d) We claim that sufficient constraints to ensure convergence of the fixed point iteration are that $|f'|$ is bounded by some $M > 0$ and $\Delta t < 1/M$. Set $F(z) := y_n + \Delta t f(z)$. We seek to show the iteration $z^{n+1} := F(z^n)$ will converge to a fixed point of F. Since $\mathbb{R}$ is complete, the Banach fixed point theorem asserts this will be true for arbitrary z^0 provided that F is a contraction.

We now show that, with the listed constraints, F is a contraction. We must show there is $L \in (0,1)$ such that

$$|F(z) - F(y)| \leq L|z - y| \quad \forall\ y, z \in \mathbb{R}. \tag{313}$$

Indeed, observe that

$$\begin{aligned} F(z) - F(y) &= (y_n + \Delta t f(z)) - (y_n + \Delta t f(y)) \\ &= \Delta t \left[f(z) - f(y)\right] \\ &= \Delta t \left[f(z) - \big(f(z) + f'(\xi)(y - z)\big)\right] \\ &= \Delta t f'(\xi)(y - z). \end{aligned} \tag{314}$$

The first equality holds by the definition of F, the second by cancellation, and the third by Taylor's thereom where ξ is between y and z. Then we see

$$|F(z) - F(y)| = |\Delta t f'(\xi)||y - z| \leq \Delta t M|z - y|, \tag{315}$$

and so, if $\Delta t < M$, then F is a contraction.

e) In order to combined the BE method with fixed point iteration, we presume it is meant that we assume $y_{n+1} = y_n + \Delta t f(y^*)$ where y^* is an approximation to y^{n+1} obtained by performing the fixed point iteration. Since the BE method has nearly the entire real axis as its interval of absolute stability, it would be advised to use this combination over common methods (e.g., the FE method, which has quite a limited interval of absolute stability). Moreover, the fixed point iteration is an advisable choice for computing y^* since there are nice error bound formulas for this iteration.

□

Spring 2015

S15.01: Let $A \in \mathbb{R}^{n\times n}$ be non-singular with real eigenvalues and, given $x_0, \alpha \in \mathbb{R}$, consider the iteration

$$x_{n+1} = x_n + \alpha(b - Ax_n) \quad \text{for } k \geq 0. \tag{316}$$

a) Assume A has positive and negative eigenvalues. Show that, for any nonzero α, there exists at least one initial vector x_0 such that this iteration diverges.

b) Assume A has only positive eigenvalues. Derive conditions on α under which the iteration will converge for any x_0. Show how to choose α so that the spectral radius of $I - \alpha A$ will be the smallest.

Solution:

a) Set $T := I - \alpha A$ so that the iteration may be expresses as

$$x_{n+1} = Tx_n + \alpha b \quad \text{for } n \geq 0. \tag{317}$$

Now let v be any eigenvector of A such that its eigenvalue λ satisfies $-\alpha\lambda > 0$. This implies

$$Tv = (I - \alpha A)v = (1 - \alpha\lambda)v, \tag{318}$$

and so $\rho(T) \geq (1-\alpha\lambda) > 1$. We claim that if $\{x_n\}$ converges for arbitrary x_0, then $\rho(T) < 1$. Because $\rho(T) > 1$, we conclude there is x_0 such that $\{x_n\}$ does not converge.

All that remains to verify is the claim. Assume $x_n \longrightarrow x$ for arbitrary x_0. Then pick any $z \in \mathbb{R}$ and set $x_0 := x - z$. This implies

$$x - x_n = (Tx + \alpha b) - (Tx_{n-1} + \alpha b) = T\left(x - x_{n-1}\right) = \cdots = T^n(x - x_0) = T^n z. \tag{319}$$

Taking the limit as $n \longrightarrow \infty$, we deduce

$$0 = \lim_{n\to\infty} x - x_n = \lim_{n\to\infty} T^n z. \tag{320}$$

This shows T is convergent, which is locally equivalent to asserting $\rho(T) < 1$, and we are done.

b) We claim if $\rho(T) < 1$, then the iteration will converge for arbitrary x_0. So, it suffices to find α such that $\rho(T) < 1$. Let Λ be the set of eigenvalues of A. For each $\lambda \in \Lambda$, $1 - \alpha\lambda$ is an eigenvalue of T, and so

$$\rho(T) = \max_{\lambda \in \Lambda} |1 - \alpha\lambda|. \tag{321}$$

We claim picking $\alpha \in (0, 2/\lambda_{max})$ will guarantee $\rho(T) < 1$ where $\lambda_{max} := \max \Lambda$. Indeed, for each $\lambda \in \Lambda$, this implies

$$1 > 1 - \alpha\lambda > 1 - \frac{2\lambda}{\lambda_{max}} \geq 1 - 2 = -1. \tag{322}$$

We now verify our claim. Assume $\rho(T) < 1$. Then

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \left[T^n x_0 + \left(\sum_{j=0}^{n-1} T^j \right) \alpha b \right] = \left[\lim_{n\to\infty} T^n x_0 \right] + \left(\sum_{j=0}^{\infty} T^j \right) \alpha b. \tag{323}$$

Because $\rho(T) < 1$, T is convergent and

$$\lim_{n\to\infty} (I - T) \sum_{j=0}^{n-1} T^j = \lim_{n\to\infty} I - T^n = I - \lim_{n\to\infty} T^n = I - 0 = I. \tag{324}$$

This shows $\sum_{j=0}^{\infty} T^j = (I - T)^{-1}$. Therefore

$$A \lim_{n\to\infty} x_n = A \left[0 x_0 + (I - T)^{-1} \alpha b \right] = A \left[0 + (\alpha A)^{-1} \alpha b \right] = b, \tag{325}$$

thereby proving $\{x_n\}$ converges to a solution of $Ax = b$.

Now we pick α to minimize $\rho(T)$. First note $|1 - \alpha\lambda| \leq \max\{|1 - \alpha\lambda_{max}|, |1 - \alpha\lambda_{min}|\}$, and optimal α satisfies $1 - \alpha\lambda_{max} \leq 0$ and $1 - \alpha\lambda_{min} \geq 0$. Let δ be the difference between the magnitudes in these two extreme cases so that

$$\delta + \alpha\lambda_{max} - 1 = 1 - \alpha\lambda_{min} \quad \Rightarrow \quad \alpha = \frac{2 + \delta}{\lambda_{min} + \lambda_{max}}. \tag{326}$$

Then

$$|1 - \alpha\lambda_{max}| = \alpha\lambda_{max} - 1 = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} + \frac{\delta}{\lambda_{max} + \lambda_{min}} \tag{327}$$

and

$$|1-\alpha\lambda_{min}| = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}} - \frac{\delta}{\lambda_{max}+\lambda_{min}}. \tag{328}$$

This shows

$$|1-\alpha\lambda| \leq \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}} + \frac{|\delta|}{\lambda_{max}+\lambda_{min}}, \tag{329}$$

with equality at $\lambda = \lambda_{min}$ or $\lambda = \lambda_{max}$. Thus the minimum is obtained when $\delta = 0$, thereby implying

$$\boxed{\alpha = \frac{2}{\lambda_{min}+\lambda_{max}}.} \tag{330}$$

□

S15.02: Let x_0, x_1, x_2, x_3 be four equispaced points that are a distance h apart.

a) Derive the coefficients α_1 and α_2 so that the "open" integration formula

$$\int_{x_0}^{x_3} f(s)\ \mathrm{d}s \approx \alpha_1 f(x_1)h + \alpha_2 f(x_2)h \tag{331}$$

is as high order as possible

b) Given an alternate "open" integration formula for approximating the integral over $[x_0, x_3]$ that has the same order as a) but requires fewer function evaluations.

Solution:

a) Define the map $F : \mathbb{R} \to \mathbb{R}$ by $F(x) := \displaystyle\int_{x_0}^{x} f(s)\ \mathrm{d}s$. This gives the following Taylor expansions:

$$F(x_3) = F(x_2) + F'(x_2)\cdot h + F''(x_2)\cdot \frac{h^2}{2} + F'''(x_2)\cdot\frac{h^3}{6} + \mathcal{O}(h^4), \tag{332}$$

$$F(x_0) = F(x_2) - F'(x_2)\cdot 2h + F''(x_2)\cdot 2h^2 - F'''(x_2)\cdot\frac{8h^3}{6} + \mathcal{O}(h^4), \tag{333}$$

$$F(x_3) = F(x_1) + F'(x_1)\cdot 2h + F''(x_1)\cdot 2h^2 + F'''(x_1)\cdot\frac{8h^3}{6} + \mathcal{O}(h^4), \tag{334}$$

$$F(x_0) = F(x_1) - F'(x_1)\cdot h + F''(x_1)\cdot \frac{h^2}{2} - F'''(x_1)\cdot\frac{h^3}{6} + \mathcal{O}(h^4). \tag{335}$$

Note that $F(x_0) = 0$. Then subtracting (333) from (332) and subtracting (335) from (334), we obtain

$$F(x_3) = 3h\cdot f(x_2) - \frac{3h^2}{2}\cdot f'(x_2) + \frac{3h^3}{2}f''(x_2) + \mathcal{O}(h^4), \tag{336}$$

$$F(x_3) = 3h\cdot f(x_1) + \frac{3h^2}{2}\cdot f'(x_1) + \frac{3h^3}{2}f''(x_1) + \mathcal{O}(h^4), \tag{337}$$

where $F'(x) = f(x)$. We seek to form a linear combination of $f(x_1)$ and $f(x_2)$ that makes the approximation as high order as possible. Since we only have two terms from which to make the linear combination, the best we can do is cancel the $\mathcal{O}(h^2)$ terms. We do this as follows.

(continued on next page)

Using the Taylor expansion

$$f'(x_2) = f'(x_1) + h \cdot f''(x_1) + \mathcal{O}(h^2), \tag{338}$$

we rewrite (336) as

$$F(x_3) = 3h \cdot f(x_2) - \frac{3h^2}{2} \cdot f'(x_1) + \frac{3h^3}{2}\left[f''(x_2) - f''(x_1)\right] + \mathcal{O}(h^4). \tag{339}$$

Thence adding (337) and (339) and dividing by 2 yields

$$F(x_3) = \frac{3h}{2}\left[f(x_1) + f(x_2)\right] + \frac{3h^3}{4}\left[f''(x_2) - f''(x_1)\right] + \mathcal{O}(h^4), \tag{340}$$

which is a $\mathcal{O}(h^3)$ approximation of the desired integral. Thus, the coefficients are $\boxed{\alpha_1 = \alpha_2 = 3/2.}$

b) Define $\bar{x} = (x_1 + x_2)/2$. Then we obtain the Taylor expansions

$$F(x_3) = F(\bar{x}) + \frac{3h}{2} \cdot F'(\bar{x}) + \frac{9h^2}{8} \cdot F''(\bar{x}) + \frac{9h^3}{16} \cdot F'''(\bar{x}) + \mathcal{O}(h^4), \tag{341}$$

$$F(x_0) = F(\bar{x}) - \frac{3h}{2} \cdot F'(\bar{x}) + \frac{9h^2}{8} \cdot F''(\bar{x}) - \frac{9h^3}{16} \cdot F'''(\bar{x}) + \mathcal{O}(h^4). \tag{342}$$

Again noting $F(x_0) = 0$, we subtract (342) from (341) to discover

$$F(x_3) = \frac{3h}{2} \cdot F'(\bar{x}) + \frac{9h^3}{8} \cdot F'''(\bar{x}) + \mathcal{O}(h^4). \tag{343}$$

This implies our alternative "open" integral formula is given by

$$\int_{x_0}^{x_3} f(s)\ \mathrm{d}s = \frac{3h}{2} \cdot f(\bar{x}) + \mathcal{O}(h^3), \tag{344}$$

which requires a single function evaluation and is $\mathcal{O}(h^3)$, as desired.

□

S15.03: Let $f(0)$, $f(-h)$, $f(-2h)$ be the values of a real valued function at $x = 0$, $x = -h$, and $x = -2h$.

a) Derive the coefficients c_0, c_1, c_2 so that

$$Df_h(x) = c_0 f(0) + c_1 f(-h) + c_2 f(-2h) \tag{345}$$

is as accurate an approximation to $f'(0)$ as possible.

b) Derive the leading term of a truncation error estimate for the formula derived in a).

Solution:

a) Through Taylor expansion about $x = 0$ we discover

$$f(-h) = f(0) - hf'(0) + \frac{h^2}{2} f''(0) - \frac{h^3}{6} f'''(0) + \mathcal{O}(h^4), \tag{346}$$

$$f(-2h) = f(0) - 2hf'(0) + 2h^2 f''(0) - \frac{4h^3}{6} f'''(0) + \mathcal{O}(h^4). \tag{347}$$

This implies

$$\begin{aligned} Df_h(0) &= c_0 f(0) + c_1 \left(f(0) - hf'(0) + \frac{h^2}{2} f''(0) - \frac{h^3}{6} f'''(0) + \mathcal{O}(h^4) \right) \\ &\quad + c_2 \left(f(0) - 2hf'(0) + 2h^2 f''(0) - \frac{4h^3}{6} f'''(0) + \mathcal{O}(h^4) \right) \\ &= (c_0 + c_1 + c_2) f(0) + (-c_1 - 2c_2) hf'(0) + \left(\frac{c_1}{2} + 2c_2 \right) h^2 f''(0) + \left(-\frac{c_1}{6} - \frac{4c_2}{3} \right) h^3 f'''(0) + \mathcal{O}(h^4). \end{aligned} \tag{348}$$

Since we have three terms, c_0, c_1, c_2, we can simultaneously solve three linear equations, one for each of the first three terms in the expansion above. This gives rise to a linear system to be solved for (c_0, c_1, c_2). Writing this in augmented form yields

$$\left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & -1 & -2 & 1/h \\ 0 & 1/2 & 2 & 0 \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & -1/h \\ 0 & 0 & 1 & 1/2h \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 1 & 0 & -1/2h \\ 0 & 1 & 0 & -2/h \\ 0 & 0 & 1 & 1/2h \end{array} \right) \sim \left(\begin{array}{ccc|c} 1 & 0 & 0 & 3/2h \\ 0 & 1 & 0 & -2/h \\ 0 & 0 & 1 & 1/2h \end{array} \right). \tag{349}$$

Whence we conclude $c_0 = 3/2h$, $c_1 = -2/h$, and $c_2 = 1/2h$.

b) From a), we see

$$\begin{aligned} Df_h(0) &= f'(0) + \left(-\frac{c_1}{6} - \frac{4c_2}{3}\right) h^3 f'''(0) + \mathcal{O}(h^3) \\ &= f'(0) + \left(-\frac{1}{6}\left(-\frac{2}{h}\right) - \frac{4}{3}\frac{1}{2h}\right) h^3 f'''(0) + \mathcal{O}(h^3) \\ &= f'(0) - \frac{h^2}{3} f'''(0) + \mathcal{O}(h^3), \end{aligned} \tag{350}$$

where we note the final term is $\mathcal{O}(h^3)$ rather than $\mathcal{O}(h^4)$ because our coefficients c_1 and c_2 are equal to constants divided by h. Thus the truncation error $\tau_h(0)$ is

$$\tau_h(0) = f'(0) - Df_h(0) = \frac{h^2}{3} f'''(0) + \mathcal{O}(h^3), \tag{351}$$

which reveals the leading term of our truncation error to be $h^2 f'''(0)/3$.

□

S15.04: Consider the function $g(x) = 2^{-x}$ on the interval $[1/3, 1]$.

a) Show g has a unique fixed point p on this interval.

b) Estimate the number of iterations necessary to achieve an accuracy of 10^{-4} when applying the fixed point iteration for approximating p. Does the method converge? Justify your answers.

Solution:

a) We first show g has a fixed point in $[1/3, 1]$. Define $f(x) = g(x) - x$ and note f is continuous. Then

$$f(1/3) = \frac{1}{\sqrt[3]{2}} - \frac{1}{3} > 0 \quad \text{and} \quad f(1) = \frac{1}{2} - 1 < 0. \tag{352}$$

Note the first inequality follows form the fact $\left(\sqrt[3]{2}\right)^3 = 2 < 27 = 3^3$. Since f is continuous, the intermediate value theorem thus implies there is $p \in (1/3, 1)$ such that $f(p) = 0$, which is equivalent to saying $g(p) = p$. Thus g has a fixed point $p \in [1/3, 1]$. Moreover, this fixed point is unique since g is strictly decreasing on $[1/3, 1]$ Indeed,

$$g'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[e^{-\ln(2)x}\right] = - \nLeftarrow (2) e^{-\ln(2)x} < 0 \quad \forall\, x \in \mathbb{R}. \tag{353}$$

b) For fixed point iteration, we define the sequence $\{x_n\}$ by $x_{n+1} = g(x_n)$ for $n \geq 0$ with x_0 an arbitrary point in $[1/3, 1]$. Taylor's theorem asserts there is ξ_n between x_n and p such that

$$p = g(p) = g(x_n) + g'(\xi_n)(p - x_n) = x_{n+1} + g'(\xi_n)(p - x_n). \tag{354}$$

And, $\ln(2) \leq 0.7$ and $2^{-\xi_n} \leq 2^{-1/3} < 2^0 = 1$. Thus

$$|x_{n+1} - p| = |g'(\xi_n)||x_n - p| \leq 0.7|x_n - p| \leq \cdots \leq 0.7^n |x_0 - p|. \tag{355}$$

Since $\lim\limits_{n\to\infty} 0.7^n = 0$, we deduce $x_n \longrightarrow p$, i.e., the method converges. We know $|x_0 - p| \leq 1 - 1/3 = 2/3$. So, taking

$$0.7^n \cdot \frac{2}{3} = 10^{-4} \quad \Rightarrow \quad n\ln(0.7) = \ln\left(\frac{3}{2}\cdot 10^{-4}\right) \quad \Rightarrow \quad n = \frac{\ln\left(\frac{3}{20000}\right)}{\ln(0.7)}. \tag{356}$$

Thus, we conclude $|x_n - p| < 10^{-4}$ when $n \geq \dfrac{\ln\left(\frac{3}{20000}\right)}{\ln(0.7)}$. □

S15.05: Let A be an $p \times p$ symmetric matrix and consider the initial value problem (IVP)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = Ay, \quad y(t_0) = y_0, \tag{357}$$

for $t \in [0, T]$.

a) Derive the implicit Taylor series method to advance an approximate solution of the IVP from t_n to $t_{n+1} = t_n + k$, based upon the approximation of $y(t_n)$ given by

$$y(t_n) \approx y(t_{n+1}) - k\frac{\mathrm{d}y}{\mathrm{d}t}(t_{n+1}) + \frac{k^2}{2}\frac{\mathrm{d}^2y}{\mathrm{d}t}(t_{n+1}). \tag{358}$$

b) What is the order of the local truncation error for the method derived in a)?

c) Derive an error bound for the solution to the IVP obtained with the method in a). Specifically, derive an error bound for $\|e_n\|_2 := \|y(t_N) - y_N\|_2$ in terms of the time step $k = T/N$.

d) What conditions (if any) on the time step k are needed for the error bound derivation to be valid?

Solution:

a) For notational clarity, let $\{z_n\}$ be the sequence of iterates generated by the implicit Taylor method. Then

$$z_n = z_{n+1} - k\frac{z_{n+1}}{t} + \frac{k^2}{2}\frac{\mathrm{d}^2 z_{n+1}}{\mathrm{d}t^2} = z_{n+1} - kAz_{n+1} + \frac{k^2}{2}A^2 z_{n+1} = \left(I - kA + \frac{(kA)^2}{2}\right) z_{n+1}. \tag{359}$$

b) Define $y_n := y(t_n)$. Taylor's theorem asserts there is $\xi_n \in (t_n, t_{n+1})$ such that

$$y_n = y_{n+1} - k\frac{\mathrm{d}y_n}{\mathrm{d}t} + \frac{k^2}{2}\frac{\mathrm{d}^2 y_{n+1}}{\mathrm{d}t^2} - \frac{k^3}{6}\frac{\mathrm{d}^3 y(\xi_n)}{\mathrm{d}t^3} = \underbrace{\left(I - kA + \frac{(kA)^2}{2}\right)}_{P_k} y_{n+1} - \frac{(kA)^3}{6}y(\xi_n). \tag{360}$$

Let P be the underbraced matrix so we may write $z_n = P_k z_{n+1}$. Then the local truncation error τ_n is

$$\tau_n := \frac{1}{k}\left[y_n - P_k y_{n+1}\right] = -\frac{k^2 A^3}{6}y(\xi_n) = \mathcal{O}(k^2). \tag{361}$$

c) First observe f is Lipschitz with constant $L = \|A\|_2$ since, for each $y, z \in \mathbb{R}^p$,

$$\|f(y) - f(z)\|_2 = \|A(y-z)\|_2 \leq \|A\|_2 \|y - z\|_2. \tag{362}$$

Moreover, the solution to the IVP is analytically given by $y(t) = \exp(At)y_0$. Thus we define $M > 0$ such that

$$M := \sup_{t\in[0,T]} \|y''(t)\|_2 = \sup_{t\in[0,T]} \|A^2 \exp(At)y_0\|_2 < \infty. \tag{363}$$

Then Taylor's theorem asserts there is $\xi_n \in (t_n, t_{n+1})$ and $\eta_n \in \mathbb{R}^p$ along the line segment between y_{n+1} and z_{n+1} such that

$$y_n = y_{n+1} - kf(y_{n+1}) + \frac{k^2}{2}y''(\xi_n) = y_{n+1} - k\left[f(z_{n+1}) + f'(\eta_n)(y_{n+1} - z_{n+1})\right] + \frac{k^2}{2}y''(\xi_n). \tag{364}$$

□

S15.06: Consider the initial value problem $u_{tt} = u_{xx} - u$ to be solved for $x \in (0,1)$, $t > 0$ with $u(x,0) = \phi(x)$, $u_t(x,0) = \psi(x)$, where ϕ and ψ are smooth.

a) For which constants a, b, c, d do the boundary conditions

$$au_x + bu_t = 0 \ \text{ at } x = 0 \quad \text{and} \quad cu_x + du_t = 0 \ \text{ at } x = 1 \tag{365}$$

lead to a well posed problem?

b) Write a convergent finite difference scheme for these well posed problems.

c) Justify your answers.

Solution:

This problem is an exact duplicate of Fall 2015 Problem 7 on page 118. □

S15.07: Let $\varepsilon > 0$ and consider the initial value problem

$$u_t = -\frac{\partial}{\partial x}\left(\frac{u^3}{3}\right) + \varepsilon u_{xx} \tag{366}$$

to be solved for $0 \leq x \leq 1$ with initial data $u(x,0) = \phi(x)$, smooth and periodic boundary conditions $u(x+1,t) = u(x,t)$.

a) Write a finite difference scheme that converges uniformly as $\varepsilon \longrightarrow 0$ for all $t > 0$.

b) Justify your answers.

Solution:

a) Define $f : \mathbb{R} \to \mathbb{R}$ by $f(u) := u^3/3$ so that $u_t + f(u)_x = \varepsilon u_{xx}$. Then we propose using the scheme

$$\frac{u_i^{n+1} - [ku_i^n + \frac{(1-k)}{2}(u_{i+1}^n + u_{i-1}^n)]}{k} + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2h} - \frac{\varepsilon}{h^2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) = 0, \tag{367}$$

taking $k \in (0,1)$. Note this scheme becomes the Lax-Friedrichs scheme as $k \longrightarrow 0$.

b) **Theorem:**[1] Suppose $\{u_i^n\}$ with step sizes k and h is generated by a numerical method in conservation form with a Lipschitz continuous numerical flux, consistent with some scalar law. If the method is TV-stable, then the method will converge to a weak solution of the conservation law as $k, h \longrightarrow 0$.

We show the conditions in the theorem are met. In the limit as $\varepsilon \longrightarrow 0$, the scalar conservation law at hand is $u_t + (f(u))_x = 0$. In order to apply the theorem, we must therefore assume $\lim\limits_{k,h\to 0} \varepsilon = 0$. In particular, we keep $\varepsilon = h^2/2$. (Step 1) We first show the scheme may be expressed in conservation form. (Step 2) We show the numerical is consistent with the actual flux up to a Lipschitz constant. (Step 3) We lastly verify the scheme TV-stable.

Step 1: Our scheme may be rewritten in conservation form as

$$u_i^{n+1} = u_i^n - \lambda\left[F(u_i^n, u_{i+1}^n) - F(u_{i-1}^n, u_i^n)\right], \tag{368}$$

where the numerical flux function F is defined by

$$F(u,v) := \underbrace{\left(\frac{1-k}{2\lambda} + \frac{\varepsilon}{h^2}\right)}_{\beta}(u-v) + \frac{f(u)+f(v)}{2}. \tag{369}$$

[1] See Theorem 15.2 on page 164 of Leveque.

Define β to be the underbraced quantity. Keeping λ fixed, our assumptions on ε ensure β remains bounded as $k, h \longrightarrow 0$.

Step 2: To show the numerical flux F is Lipschitz continuous, we must show there is $L > 0$ such that

$$|F(u,v) - f(w)| \leq L \cdot \max\{|u-w|, |v-w|\} \tag{370}$$

for all u, v with $|u-w|$ and $|v-w|$ sufficiently small. Recall we assume the grid functions are bounded. This implies the support of f is closed and bounded and, thus, compact. Since f is continuously differentiable on a compact set, it is Lipschitz continuous. Then

$$\begin{aligned} |F(u,v) - f(w)| &= \left| \beta(u-v) + \frac{1}{2}\left(f(u) + f(v)\right) - f(w) \right| \\ &= \left| \beta\left[(u-w) + (w-v)\right] + \frac{1}{2}\left[(f(u) - f(w)) + (f(v) - f(w))\right] \right| \\ &\leq (2\beta + \operatorname{lip}(f)) \cdot \max\{|u-w|, |v-w|\}, \end{aligned} \tag{371}$$

and we see (370) holds by taking $K = (2\beta + \operatorname{lip}(f))$.

Step 3: In this step we show the method is TV-stable. We first show it is ℓ_1 contracting. Let $\{u_i^n\}$ and $\{v_i^n\}$ be two collections of grid functions and define $e_i^n = u_i^n - v_i^n$. Also note our scheme may be written in the form

$$u_i^{n+1} = (k - 2\varepsilon\mu)\, u_i^n + \frac{(1-k)}{2}\left(u_{i+1}^n + u_{i-1}^n\right) + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2h} + \varepsilon\mu\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right). \tag{372}$$

Note the first term is cancels since $k - 2\varepsilon\mu = k - 2(h^2/2)(k/h^2) = 0$. This implies

$$\begin{aligned} e_i^{n+1} &= \frac{1-k}{2}\left(e_{i-1}^n + e_{i+1}^n\right) + \frac{\lambda}{2}\left[\left(f(u_{i+1}^n) - f(v_{i+1}^n)\right) - \left(f(u_{i-1}^n) - f(v_{i-1}^n)\right)\right] + \varepsilon\mu\left[e_{i+1}^n - 2e_i^n + e_{i-1}^n\right] \\ &= \frac{1-k}{2}\left(e_{i-1}^n + e_{i+1}^n\right) + \frac{\lambda}{2}\left[f'(\theta_{i+1}^n)e_{i+1}^n - f'(\theta_{i-1}^n)e_{i-1}^n\right] + \varepsilon\mu\left[e_{i+1}^n - 2e_i^n + e_{i-1}^n\right] \\ &= \left(\frac{(1-k)}{2} + \frac{\lambda}{2}f'(\theta_{i+1}) + \varepsilon\mu\right)e_{i+1}^n + \left(\frac{(1-k)}{2} - \frac{\lambda}{2}f'(\theta_{i-1}) + \varepsilon\mu\right)e_{i-1}^n. \end{aligned} \tag{373}$$

Here $\theta_{i\pm1}$ exists, by Taylor's theorem, between $u_{i\pm1}^n$ and $v_{i\pm1}^n$, respectively. We assume λ is fixed so the CFL conditions holds, i.e.,

$$\left|\lambda f'(u)\right| \leq 1 \quad \forall \ \min_j(u_j^n, v_j^n) \leq u \leq \max_j(u_j^n, v_j^n). \tag{374}$$

In particular, $|\lambda f'(\theta_{i\pm 1})| \leq 1$, and so

$$\frac{1-k}{2} \pm \frac{1}{2}\left(\lambda f'(\theta_{j\pm 1})\right) + \varepsilon\mu \geq \frac{1-k}{2} - \frac{1}{2} + \varepsilon\mu = -\frac{k}{2} + \frac{h^2}{2}\cdot\frac{k}{h^2} = 0. \tag{375}$$

This shows the coefficients of e^n_{i+1} and e^n_{i-1} are nonnegative. Whence

$$\begin{aligned}
\|e^{n+1}\|_1 &= h\sum_i |e^{n+1}_i| \\
&\leq \left(\frac{1-k}{2} + \frac{\lambda}{2}f'(\theta_{i+1}) + \varepsilon\mu\right) h\sum_i |e^n_{i+1}| + \left(\frac{1-k}{2} - \frac{\lambda}{2}f'(\theta_{i-1}) + \varepsilon\mu\right) h\sum_i |e^n_{i-1}| \\
&= \left(\frac{1-k}{2} + \frac{\lambda}{2}f'(\theta_i) + \varepsilon\mu\right) h\sum_i |e^n_i| + \left(\frac{1-k}{2} - \frac{\lambda}{2}f'(\theta_i) + \varepsilon\mu\right) h\sum_i |e^n_i| \\
&= (1-k+2\varepsilon\mu)\, h\sum_i |e^n_i| \\
&= h\sum_i |e^n_i| \\
&= \|e^n\|_1.
\end{aligned} \tag{376}$$

Note $\sum_i |e^n_i| = \sum_i |e^n_{i+j}|$ for all j since the grid functions are periodic, which is how the third line follows from the second. This shows the method is ℓ_1 contracting. Choosing $v^n_i = u^n_{i+1}$, (376) implies the method is TV diminishing (TVD) since

$$\text{TV}(u^{n+1}) = \frac{1}{h}\|u^{n+1}_{i+1} - u^{n+1}_i\|_1 \leq \frac{1}{h}\|u^n_{i+1} - u^n_i\|_1 = \text{TV}(u^n). \tag{377}$$

Then, by induction, we conclude $\text{TV}(u^n) \leq \text{TV}(u^0)$ for all $n \in \mathbb{Z}^+$ and the scheme is TV-stable.

By the above, we conclude the sequence converges to a weak solution of the conservation law. And, the above shows the method on a sequence of grids with $\varepsilon \longrightarrow 0$ converges to the "vanishing viscosity" solution that might be used to define the physically relevant weak solution to the conservation law.

□

S15.08: Consider the problem in two dimensions[2]

$$\begin{cases} -\Delta u + u = f, & \text{in } T, \\ u = g_1, & \text{on } T_1, \\ u = g_2, & \text{on } T_2, \\ \dfrac{\partial u}{\partial n} = h, & \text{on } T_3, \end{cases} \tag{378}$$

where

$$\begin{aligned} T &= \{(x,y) \ : \ x > 0,\ y > 0,\ x + y < 1\}, \\ T_1 &= \{(x,y) \ : \ 0 < x < 1,\ y = 0\}, \\ T_2 &= \{(x,y) \ : \ x = 0,\ 0 < y < 1\}, \\ T_3 &= \{(x,y) \ : \ x > 0,\ y > 0,\ x + y = 1\}. \end{aligned} \tag{379}$$

a) Find the weak variational formulation of the problem and verify the assumptions of the Lax-Milgram Lemma by analyzing the appropriate bilinear and linear forms. Impose the weakest necessary assumptions on the functions f, g_1, g_2, and h.

b) Develop and describe the piecewise linear Galerkin finite element approximation of the problem and a set of basis functions such that the corresponding linear system is sparse. Show that this linear system has a unique solution. Give a convergence estimate and quote the appropriate theorems for convergence.

[2] This is the exact same problem as F13.08, except for the last sentence.

Solution:

a) First assume $f \in L^2(T)$ and $h \in L^2(T_3)$. We also assume g_1 and g_2 are sufficiently smooth so that there are liftings $u_{g_1}, u_{g_2} \in H^2(T)$ such that $u_{g_1} = g_1$ on T_1 and $u_{g_2} = g_2$ on T_2. Then define $\phi := u - u_{g_1} - u_{g_2}$ so that $u = \phi + u_{g_1} + u_{g_2}$ and ϕ is in the Hilbert space $H := \{v \in H^2(T) \ : \ v|_{T_1 \cup T_2} = 0\}$. Then

$$\begin{aligned} -\Delta\phi + \phi &= \tilde{f}, \ \text{ in } T, \\ \phi &= 0, \ \text{ on } T_1, \\ \phi &= 0, \ \text{ on } T_2, \\ \frac{\partial \phi}{\partial n} &= \tilde{h}, \ \text{ on } T_3, \end{aligned} \tag{380}$$

where

$$\tilde{f} := f + \Delta(u_{g_1} + u_{g_2}) - (u_{g_1} + u_{g_2}) \quad \text{and} \quad \tilde{h} := h - \frac{\partial}{\partial n}(u_{g_1} + u_{g_2}). \tag{381}$$

Note $\tilde{f} \in L^2(T)$ by our choices of f, u_{g_1}, and u_{g_2}. In particular, $\Delta(u_{g_1} + u_{g_2}) \in L^2(T)$. Similarly, $\tilde{h} \in L^2(T_3)$. For each test function $v \in H$,

$$\int_T \tilde{f} v = \int_T -\Delta\phi v + \phi v = \int_T D\phi \cdot Dv + \phi v - \int_{T_3} \frac{\partial \phi}{\partial n} v = \int_T D\phi \cdot Dv + \phi v - \int_{T_3} \tilde{h} v. \tag{382}$$

Thus, defining the bilinear form b and linear form ℓ by

$$b(\phi, v) := \int_T D\phi \cdot Dv + \phi v \quad \text{and} \quad \ell(v) := \int_T \tilde{f} v + \int_{T_3} \tilde{h} v, \tag{383}$$

we obtain the weak variational problem

$$\text{Find } u = u_{g_1} + u_{g_2} + \phi \in H^2(T) \text{ such that } \phi \in H \text{ and } b(\phi, v) = \ell(v) \text{ for all } v \in H. \tag{384}$$

We now verify the assumptions of the Lax-Milgram theorem. The bilinear form b is symmetric by commutativity of scalar multiplication. We must show b is coercive and bounded and ℓ is bounded. We see b is coercive since

$$b(v, v) = \int_T |Dv|^2 + v^2 = \|v\|_H^2. \tag{385}$$

And, b is bounded because

$$\begin{aligned} |b(\phi, v)| &\leq \|D\phi \cdot Dv\|_{L^1(T)} + \|\phi v\|_{L^1(T)} \\ &\leq \|D\phi\|_{L^2(T)}\|Dv\|_{L^2(T)} + \|\phi\|_{L^2(T)}\|v\|_{L^2(T)} \\ &\leq 2\|\phi\|_H\|v\|_H. \end{aligned} \tag{386}$$

The first inequality follows from the triangle inequality; the second holds by applying Hölder's inequality; the final inequality follows from making use of the fact $\|v\|_H^2 = \|Dv\|_{L^2(T)}^2 + \|v\|_{L^2(T)}^2$. Next observe

$$|\ell(v)| \leq \|\tilde{f}v\|_{L^1(T)} + \|\tilde{h}v\|_{L^1(T_3)} \leq \|\tilde{f}\|_{L^2(T)}\|v\|_{L^2(T)} + \|\tilde{h}\|_{L^2(T_3)}\|v\|_{L^2(T_3)}. \tag{387}$$

Also note there is $C > 0$ such that

$$\|v\|_{L^2(T_3)} \leq \|v\|_{L^2(\partial T)} \leq C\|v\|_{L^2(T)}. \tag{388}$$

Together (387) and (388) imply

$$|\ell(v)| \leq C\|\tilde{h}\|_{L^2(T_3)}\|v\|_{L^2(T)} \leq C\|\tilde{h}\|_{L^2(T_3)}\|v\|_H. \tag{389}$$

Whence ℓ is bounded, and we are done.

b) For the finite element approximation, let $\mathcal{T}_h$ be a triangulation of Ω where h denotes the fineness of the triangulation mesh and the nodes are denoted by $\{N_j\}$. Let

$$H_h = \{v \in H \mid v|_K \in P_1(K) \ \forall\ K \in \mathcal{T}_h\}. \tag{390}$$

The approximate variational formulation then becomes to find $u = u_{g_1} + u_{g_2} + \phi_h \in H^2(T)$ such that $\phi_h \in H_h$ and $b(\phi_h, v) = \ell(v)$ for all $v \in H_h$. By linearity, if $\{\gamma_i\}$ is a basis for H_h, this is equivalent to finding $\phi_h \in H_h$ such that $a(\phi_h, \gamma_i) = \ell(\gamma_i)$ for all γ_i. We take γ_i such that $\gamma_i(N_j) = \delta_{ij}$. Now we can also express $\phi_h = \sum_j \xi_j \phi_j$, thus obtaining the linear system

$$\sum_j \xi_j b(\gamma_i, \gamma_j) = \ell(\gamma_i) \qquad \Rightarrow \qquad A\xi = b, \tag{391}$$

where the entries of the stiffness matrix are $A_{ij} := b(\gamma_i, \gamma_j)$ and the entries of the load vector are

$b_i := \ell(\gamma_i)$. If the enumeration of the N_j's is done efficiently, A will be sparse (since $b(\gamma_i, \gamma_j) = 0$ if $|i - j|$ is too large) and banded, allowing for efficient solving of the system. Moreover, A is positive definite (since b is an inner product). Whence is A invertible and the system has a unique solution.

If ϕ is the solution to the weak variational formulation and ϕ_h is the solution to the approximate variational formulation, then we have the bound $\|\phi - \phi_h\|_b \leq \|\phi - v\|_b$ for any $v \in H_h$ where $\|\cdot\|_b$ is the norm $\|\cdot\|_H$ induced by the inner product $b(\cdot, \cdot)$. In particular, we can take the linear interpolant $\pi_h\phi \in V_h$ of ϕ, and we know $\|\phi - \pi_h\phi\|_b \leq C_w h^2$ for some constant C_w, dependent on ϕ and independent of h. From these two inequalities, we obtain the convergence rate estimate $\|\phi - \phi_h\|_b \leq C_w h^2$.

□

Fall 2015

F15.01: Let $A \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix with a non-trivial null space, i.e., $\ker(A) \neq \{0\}$. Consider the problem of finding x such that $Ax = b$.

a) State the condition that guarantees a solution x will exist.

b) Give a derivation of the condition on the matrix A that insures the iterative method

$$x_{k+1} = x_k + (b - Ax_k) \quad \text{for } k \geq 0 \tag{392}$$

will converge to a solution x when such a solution exists.

Solution:

a) We could say $b = 0$ or $b \in \text{range}(A)$....Not sure what they want here with "the" condition.

b) Ummmm.... We could rearrange this to be $x_k = Tx_{k-1} + b$ where $T = (I - A)$. If $\rho(T) < 1$, then we have convergence. However, that is equivalent to saying A is positive definite with eigenvalues strictly between zero and one. And we are given that $\ker(A) \neq \{0\}$.....

□

F15.02: Let $f, g : \mathbb{R} \to \mathbb{R}$ be smooth functions and consider the problem of finding x such that

$$f(x) + g(x) = b, \tag{393}$$

for some $b \in \mathbb{R}$.

a) Assume one is given an approximation solution value x_{n-1}. Derive the formula for the next approximation x_n that is obtained by using one step of Newton's method with the starting iterate x_{n-1} applied to the problem of finding x such that

$$f(x) + g(x_{n-1}) = b. \tag{394}$$

b) Assume f is a linear function. Under what conditions on f and g would you be able to prove the iteration defined in part a) will converge? Explain your answer.

Solution:

a) For each nonnegative integer n, define the function $F_n : \mathbb{R} \to \mathbb{R}$ by $F_n(x) := f(x) + g(x_n) - b$. Our desired iteration is then found by using one step of Newton's method to approximate F_n, and then the successive iterate by repeating the process with F_{n+1}, and so on. In mathematical terms, we write

$$x_n = x_{n-1} - \frac{F_{n-1}(x_{n-1})}{F'_{n-1}(x_{n-1})} = x_{n-1} - \frac{f(x_{n-1}) + g(x_{n-1}) - b}{f'(x_{n-1})}. \tag{395}$$

b) If f is linear, then there exists $a_1, a_2 \in \mathbb{R}$ such that $f(x) = a_1 x + a_2$. We claim the iteration defined in a) will converge provided g' is bounded and $L := \|g'\|_\infty / a_1 < 1$. We verify this as follows. First observe that, with our choice of f, the iteration in a) becomes

$$x_n = x_{n-1} - \frac{a_1 x_{n-1} + a_2 + g(x_{n-1}) - b}{a_1} = \frac{b - a_2 - g(x_{n-1})}{a_1}. \tag{396}$$

This implies

$$x_{n+1} - x_n = \left(\frac{b - a_2 - g(x_{n-1})}{a_1} \right) - \left(\frac{b - a_2 - g(x_{n-1})}{a_1} \right) = \frac{g(x_{n-1}) - g(x_n)}{a_1}. \tag{397}$$

Recall Taylor's theorem asserts that, for each n, there is ξ_n between x_n and x_{n-1} such that

$$g(x_n) = g(x_{n-1}) + g'(\xi_n)(x_n - x_{n-1}) \quad \Rightarrow \quad g(x_n) - g(x_{n-1}) = g'(\xi_n)(x_n - x_{n-1}). \tag{398}$$

Substituting the result from (398) into (397) and setting $\gamma := \|g'\|_\infty / a_1$, we discover

$$|x_{n+1} - x_n| = \left|\frac{g'(\xi_n)}{a_1}\right| |x_n - x_{n-1}| \le \gamma |x_n - x_{n-1}| \le \cdots \le \gamma^n |x_1 - x_0|. \tag{399}$$

Whence for $m > n$

$$\begin{aligned} |x_m - x_n| &\le |x_m - x_{m-1}| + \cdots |x_{n+1} - x_n| \\ &\le \gamma^n |x_1 - x_0| \cdot \sum_{j=0}^{m-n} \gamma^j \\ &\le \gamma^n \cdot \frac{x_1 - x_0}{1 - \gamma}. \end{aligned} \tag{400}$$

In the limit as $n \longrightarrow \infty$, $\gamma^n \cdot \frac{|x_1 - x_0|}{1-\gamma} \longrightarrow 0$. This shows $\{x_n\}$ is Cauchy. Because $\mathbb{R}$ is complete, we conclude $\{x_n\}$ converges.

□

F15.03: Recall an $n \times n$ matrix A is strictly diagonally dominant if

$$|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}| \quad \text{for } i = 1, 2, \ldots, n. \tag{401}$$

a) Show the linear system $Ax = b$ has a unique solution when A is strictly diagonally dominant.

b) Discuss the convergence of the Jacobi method for solving the linear system $Ax = b$ when A is such a matrix. Justify your answer.

Solution:

a) We claim that if A is strictly diagonally dominant, then A is invertible. Thus A is invertible. We then claim the unique solution to $Ax = b$ is $A^{-1}b$. Indeed, this is a solution since $AA^{-1}b = Ib = b$. To verify uniqueness, suppose z is any solution to $Ax = b$. Then

$$z - A^{-1}b = I(z - A^{-1}b) = A^{-1}A(z - A^{-1}b) = A^{-1}(Az - b) = A^{-1}0 = 0. \tag{402}$$

This shows $z = A^{-1}b$ and verifies the solution is unique.

All that remains is to verify our initial claim. We do this by proving its contrapositive form. Suppose A is singular. Then there is nonzero y such that $Ay = 0$. Set $j := \underset{i=1,\ldots,n}{\text{argmax}} |x_i|$. Then, in particular, we have $0 = \sum_{i=1}^{n} a_{ji}x_i$. This implies

$$|a_{jj}||x_j| = \left| - \sum_{i=1, i \neq j}^{n} a_{ji}x_i \right| \leq \sum_{i=1, i \neq j}^{n} |a_{ji}||x_i| \leq |x_j| \sum_{i=1, i \neq j}^{n} |a_{ji}|. \tag{403}$$

The final inequality follows from our choice of j. Dividing both sides of (403) by $|x_j|$, we obtain

$$|a_{jj}| \leq \sum_{i=1, i \neq j}^{n} |a_{ji}|, \tag{404}$$

which shows A is not strictly diagonally dominant. This completes the proof.

b) We claim the Jacobi method converges when A is strictly diagonally dominant. Let D be the matrix containing the diagonal entries of A. Given an initial iterate x_0, the Jacobi iteration is then defined

$$Dx_{n+1} = (D - A)x_n + b. \tag{405}$$

Since A is diagonally dominant, each of its diagonal entries are nonzeoro and so D is invertible. Letting $T := D^{-1}(D - A)$, this implies

$$x_{n+1} = Tx_n + D^{-1}b. \tag{406}$$

A well-known result asserts this iteration converges if and only if the spectral radius of T is less than unity, i.e., $\rho(T)) < 1$. All that remains is to verify $\rho(T) < 1$. First note

$$\sum_{j=1}^{n} |t_{ij}| = \sum_{j=1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 \quad \text{for } i = 1, 2, \ldots, n. \tag{407}$$

The equality above follows from dividing each side of the inequality in (398) by a_{ii}. Using this fact with the definition of the sup norm, we deduce

$$\|T\|_\infty := \sup_{\|x\|_\infty = 1} \|Tx\|_\infty = \sup_{\|x\|_\infty = 1} \sup_i \left| \sum_{j=1}^{n} t_{ij} x_j \right| \leq \sup_{\|x\|_\infty = 1} \sup_i \sum_{j=1}^{n} |t_{ij}||x_j| = \sup_i \sum_{j=1}^{n} |t_{ij}| < 1. \tag{408}$$

Because $\rho(M) \leq \|M\|$ for each natural norm $\| \cdot \|$ of a matrix M, the above shows $\rho(T) < 1$, and we are done.

□

F15.04: Let $f(x) = \ln(x+1)$, $x_0 = 0$, $x_1 = 0.6$, and $x_2 = 0.9$.

a) Construct an interpolating polynomial of degree at most two to approximate f using the three points (you can use $f(0.6) = 0.47$ and $f(0.9) = 0.6$).

b) Find an error bound for the approximation.

Solution:

a) The Lagrange interpolating polynomial for this data is given by

$$p(x) := f(x_0)L_{2,0}(x) + f(x_1)L_{2,1}(x) + f(x_2)L_{2,2}(x) \tag{409}$$

where

$$L_{2,i}(x) := \prod_{j=0, j\neq i}^{2} \frac{(x-x_j)}{(x_i - x_j)}. \tag{410}$$

More explicitly, we can write

$$\begin{aligned} p(x) &= 0L_{2,0}(x) + 0.47L_{2,1}(x) + 0.6L_{2,2}(x) \\ &= 0 + 0.47 \cdot \frac{(x-0)(x-0.9)}{(0.6-0)(0.6-0.9)} + 0.6 \cdot \frac{(x-0)(x-0.6)}{(0.9-0)(0.9-0.6)} \\ &= \boxed{\frac{0.47}{-0.18} \cdot x(x-0.9) + \frac{0.6}{0.27} \cdot x(x-0.6).} \end{aligned} \tag{411}$$

b) The remainder theorem for Lagrange polynomials asserts that for each $x \in [x_0, x_2]$ there is $\xi_x \in (x_0, x_2)$ such that

$$f(x) = p(x) + \frac{f^{(3)}(\xi_x)}{3!} \cdot (x-x_0)(x-x_1)(x-x_2). \tag{412}$$

This implies that for each $x \in [0, 0.9]$, the error $e(x)$ satisfies

$$\begin{aligned} |e(x)| &\leq \left| \frac{f'''(\xi_x)}{6} \cdot x(x-0.6)(x-0.9) \right| \\ &= \left| \frac{x(x^2 - 1.5x + 0.54)}{6 \cdot 2(\xi_x + 1)^3} \right| \\ &\leq \left| \frac{x(x^2 - 1.5x + 0.54)}{6} \right| \\ &\leq \left| \frac{x^2 - 1.5x + 0.54}{6} \right|. \end{aligned} \tag{413}$$

The first inequality holds since $\xi_x \geq 0$ and the second by the fact $x \in [0, 1]$. Now note $x^2 - 1.5x + 0.54$ is concave up and evaluates to 0.54 when $x = 0$ and 0.04 when $x = 1$. So, we deduce $x^2 - 1.5x + 0.54 \leq 0.54$ for $x \in [0, 0.9]$. Whence we obtain

$$|e(x)| \leq \left| \frac{0.54}{6} \right| \leq \frac{0.6}{6} = 0.1 \quad \text{for } x \in [0, 0.9]. \tag{414}$$

□

F15.05: Assume $f : \mathbb{R} \to \mathbb{R}$ is a smooth function and consider the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y), \quad y(0) = y_0. \tag{415}$$

a) Using expansions about t_n, derive the leading term of the local truncation error for the following method used to create approximate solutions to (415),

$$y_n = y_{n-1} + \frac{3h}{2} f_{n-1} - \frac{h}{2} f_{n-2}. \tag{416}$$

b) Using a) and the fact that the local truncation error estimate for the second order BDF method has the form

$$y_n = \frac{4}{3} y_{n-1} - \frac{1}{3} y_{n-2} + \frac{2h}{3} f_n - \frac{2h^3}{9} \frac{\mathrm{d}^3 y}{\mathrm{d}t^3}(t_n) + \mathcal{O}(h^4), \tag{417}$$

derive the leading term of the local truncation error for the method

$$\begin{aligned} y^* &= y_{n-1} + \frac{3h}{2} f_{n-1} - \frac{h}{2} f_{n-2}, \\ y_n &= \frac{4}{3} y_{n-1} - \frac{1}{3} y_{n-2} + \frac{2h}{3} f(y^*). \end{aligned} \tag{418}$$

c) Derive the polynomial whose roots determine the region of absolute stability for the method in b).

Solution:

a) Set $\Phi(y_{n-1} + f_{n-1}, f_{n-2}) := y_{n-1} + \frac{3h}{2} f_{n-1} - \frac{h}{2} f_{n-2}$. Then

$$\begin{aligned} \Phi(y_{n-1} + f_{n-1}, f_{n-2}) &= y_{n-1} + \frac{3h}{2}\left[f_n - h f_n' f_n + \frac{h^2}{2}\left(f_n'' f_n + f_n'^2 \right) + \mathcal{O}(h^3) \right] \\ &\quad - \frac{h}{2}\left[f_n - 2h f_n' f_n + 2h^2 \left(f_n'' f_n + f_n'^2 \right) + \mathcal{O}(h^3) \right] \\ &= y_{n-1} + h f_n - \frac{h^2}{2} f_n' f_n - \frac{h^3}{4}\left(f_n'' f_n + f_n'^2 \right) + \mathcal{O}(h^4) \end{aligned} \tag{419}$$

This implies the local truncation error τ_n is given by

$$\begin{aligned}\tau_n &= y_n - \Phi(y_{n-1} + f_{n-1}, f_{n-2}) \\ &= \left(y_{n-1} + hf_n - \frac{h^2}{2}f_n'f_n + \frac{h^3}{6}\left(f_n''f_n + f_n'^2\right) + \mathcal{O}(h^4)\right) \\ &- \left(y_{n-1} + hf_n - \frac{h^2}{2}f_n'f_n - \frac{h^3}{4}\left(f_n''f_n + f_n'^2\right) + \mathcal{O}(h^4)\right) \\ &= \frac{5h^3}{12}\left(f_n''f_n + f_n'^2\right) + \mathcal{O}(h^4),\end{aligned} \tag{420}$$

which gives the leading term of τ_n.

b) Here observe that

$$f(y^*) = f\left(y_n - \frac{5h^3}{12}y_n''' + \mathcal{O}(h^4)\right) = f(y_n) - \frac{5h^3}{12}y_n'''y_n' + \mathcal{O}(h^4). \tag{421}$$

This implies the local truncation error $\tilde{\tau}_n$ of the method is

$$\begin{aligned}\tilde{\tau}_n &= y_n - \left(\frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + \frac{2h}{3}f(y_n) - \frac{2h}{3}\frac{5h^3}{12}\left(f_n''f_n' + f_n'^2\right)f_n + \mathcal{O}(h^5)\right) \\ &= -\frac{2h^3}{9}(f_n''f_n + f_n'^2) + \frac{5h^4}{18}\left(f_n''f_n' + f_n'^2\right)f_n + \mathcal{O}(h^5) \\ &= -\frac{2h^3}{9}(f_n''f_n + f_n'^2) + \mathcal{O}(h^4),\end{aligned} \tag{422}$$

i.e., the leading term of the local truncation error for the method is the same as the leading term of local truncation error in (417).

c) For determining the region of absolute stability, we presume $f(y) = \lambda y$ for some $\lambda \in \mathbb{C}$. Then the method becomes

$$\begin{aligned}y_n &= \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + \frac{2h}{3}\lambda\left(y_{n-1} + \frac{3h}{2}\lambda y_{n-1} - \frac{h}{2}\lambda y_{n-2}\right) \\ &= \left(\frac{4}{3} + \frac{2h\lambda}{3} + (h\lambda)^2\right)y_{n-1} - \frac{1}{3}\left(1 + (h\lambda)^2\right)y_{n-2}.\end{aligned} \tag{423}$$

From this, we derive the polynomial $p(r; h\lambda)$ whose roots determine the region of absolute stability for the method. Indeed, substituting for each y_i, we obtain

$$\boxed{p(r; h\lambda) := r^2 - \left(\frac{4}{3} + \frac{2h\lambda}{3} + (h\lambda)^2\right)r + \frac{1}{3}\left(1 + (h\lambda)^2\right).} \tag{424}$$

□

F15.06: Consider the initial value problem

$$u_t = -\left(\frac{u^3}{3}\right)_x + \varepsilon u_{xx} \tag{425}$$

for $\varepsilon > 0$ to be solved for $0 \leq x \leq 1$ with initial data $u(x, 0) = \phi(x)$ and smooth and periodic boundary conditions $u(x + 1, t) = u(x, t)$.

a) Write a finite difference scheme that converges uniformly in ε as $\varepsilon \longrightarrow 0$ for all $t > 0$.

b) Justify your answers.

Solution:

This was already done in S15.07. See page 101. □

F15.07: Consider the initial value problem $u_{tt} = u_{xx} - u$ to be solved for $x \in (0,1)$, $t > 0$ with $u(x,0) = \phi(x)$, $u_t(x,0) = \psi(x)$, where ϕ and ψ are smooth.

a) For which constants a, b, c, d do the boundary conditions

$$au_x + bu_t = 0 \text{ at } x = 0 \quad \text{and} \quad cu_x + du_t = 0 \text{ at } x = 1 \tag{426}$$

lead to a well posed problem?

b) Write a convergent finite difference scheme for these well posed problems.

c) Justify your answers.

Solution:

a) I have no clue how to do this.......

b) First set $v = (u,\ (\partial_t - \partial_x)u) \in \mathbb{R}^2$. Then $(v_1)_t - (v_1)_x = u_t - u_x = -v_2$ and $(v_2)_t + (v_2)_x = u_{tt} - u_{xx} = -u = -v_1$. This shows our PDE can be written as the system

$$v_t + Av_x - Bv = 0 \tag{427}$$

where

$$A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{428}$$

We propose the Lax-Friedrichs scheme

$$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} + A\left(\frac{v_{m+1}^n - v_{m-1}^n}{2h}\right) - Bv_m^n = 0. \tag{429}$$

c) By the Lax equivalence theorem, it suffices to show the presented scheme is consistent and stable. We first verify consistency. We Taylor expand component-wise to find

$$v_{m\pm1}^n = v_m^n \pm h(v_m^n)_x + \frac{h^2}{2}(v_m^n)_{xx} \pm \frac{h^3}{6}(v_m^n)_{xxx} + \mathcal{O}(h^4). \tag{430}$$

This implies

$$\frac{v_{m+1}^n - v_{m-1}^n}{2h} = (v_m^n)_x + \frac{h^2}{6}(v_m^n)_{xxx} + \mathcal{O}(h^3) \tag{431}$$

and

$$\frac{v_{m+1}^n + v_{m-1}^n}{2} = v_m^n + \frac{h^2}{2}(v_m^n)_{xx} + \mathcal{O}(h^4). \tag{432}$$

In similar fashion, we discover

$$\frac{v_m^{n+1} - v_m}{k} = (v_m^n)_t + \frac{k}{2}(v_m^n)_{tt} + \mathcal{O}(k^2). \tag{433}$$

Let $P_{k,h}$ be the difference operator corresponding to the left side of (429). Combining (431), (432), and (433), we see

$$P_{k,h}v_m^n = (v_m^n)_t + \mathcal{O}(k) + \mathcal{O}(h^2k^{-1}) + A(v_m^n)_x + \mathcal{O}(h^2) - Bv_m^n. \tag{434}$$

Set P to be the differential operator for the PDE in (427). Then

$$P_{k,h}v_m^n - Pv_m^n = \mathcal{O}(k + h^2k^{-1} + h^2). \tag{435}$$

We therefore conclude the scheme is consistent provided $h^2k^{-1} \longrightarrow 0$ as $k, h \longrightarrow 0$.

We now verify the stability of the scheme using Von Neumann analysis. Let G denote the amplification factor. Then our method is stable if and only if for each $T > 0$ there is $C_T > 0$ such that

$$\|G^n\| \leq C_T \quad \text{for } 0 \leq nk \leq T. \tag{436}$$

Since all norms in finite dimensional space are equivalent, we can proceed using the two norm. We replace each v_m^n in the scheme by $G^n e^{im\theta}$ to obtain

$$B = \frac{G - \frac{e^{i\theta}+e^{-i\theta}}{2}I}{k} + A\left(\frac{e^{i\theta} - e^{-i\theta}}{2h}\right) = \frac{G - \cos(\theta)I}{k} + \frac{i\sin(\theta)}{h}A, \tag{437}$$

which implies

$$G = \underbrace{I\cos\theta - \lambda Ai\sin\theta}_{M} + kB = M + kB \tag{438}$$

where $\lambda := k/h$ and M is the underbraced quantity. Now recall $\|G\|_2 = \sqrt{\lambda_{max}(G^*G)}$. Noting M is diagonal, $B = -B^*$, and $B^*B = I$, we find

$$G^*G = (M^* + kB^*)(M + kB) = M^*M + k^2B^*B = M^*M + k^2I = (\cos^2\theta + \lambda^2\sin^2\theta + k^2)I. \quad (439)$$

This shows the single eigenvalue of G^*G is $(\cos^2\theta + \lambda^2\sin^2\theta + k^2)$. Assuming $\lambda = k/h \leq 1$, we deduce

$$\|G\|_2^2 = \cos^2\theta + \lambda^2\sin^2\theta + k^2 \leq 1 + k^2 \leq 1 + 2k + k^2 = (1+k)^2 \quad \Rightarrow \quad \|G\|_2 \leq 1 + k. \quad (440)$$

Thus

$$\|G^n\|_2 \leq \|G\|_2^n \leq (1+k)^n \leq e^{nk} \leq e^T \quad \text{for } 0 \leq nk \leq T. \quad (441)$$

Whence (436) holds with $C_T = e^T$, thereby showing the method is stable when $\lambda := k/h \leq 1$. This completes the proof. □

F15.08: Consider the problem

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega_1, \\ \dfrac{\partial u}{\partial n} + u = x & \text{on } \partial\Omega_2, \end{cases} \tag{442}$$

where

$$\begin{aligned} \Omega &= \{(x,y) \;:\; x^2+y^2<1\}, \\ \partial\Omega_1 &= \{x,y \;:\; x^2+y^2=1,\ x \le 0\}, \\ \partial\Omega_2 &= \{x,y \;:\; x^2+y^2=1,\ x > 0\}, \end{aligned} \tag{443}$$

and $f \in L^2(\Omega)$.

a) Determine an appropriate weak variational formulation.

b) Is the obtained bilinear form symmetric? If yes, give an equivalent minimization formulation.

c) Verify conditions on the corresponding linear and bilinear forms needed for existence and uniqueness of the solution to the weak variational formulation.

d) Assume that the boundary $\partial\Omega$ is approximated by a symmetric polygonal curve. Describe a finite element approximation of the problem using P_1 elements and a set of basis functions. Prove the necessary properties of the obtained linear system and discuss its structure. Give a rate of convergence.

Solution:

a) Define the Hilbert space $H := \{v \in H^1(\Omega) \;:\; v|_{\partial\Omega_1} = 0\}$ with the norm $\|\cdot\|_H := \|\cdot\|_{H^1(\Omega)}$ and let u be a solution to the PDE. Then, for each test function $v \in H$,

$$\begin{aligned} \int_\Omega fv &= \int_\Omega -\Delta uv + uv \\ &= \int_\Omega Du\cdot Dv + uv - \int_{\partial\Omega_1} \frac{\partial u}{\partial n} v - \int_{\partial\Omega_2} \frac{\partial u}{\partial n} v \\ &= \int_\Omega Du\cdot Dv + uv - 0 - \int_{\partial\Omega_2} (x-u)v. \end{aligned} \tag{444}$$

Define the bilinear form b and the linear form ℓ by

$$b(u,v) := \int_\Omega Du\cdot Dv + uv + \int_{\partial\Omega_2} uv \quad \text{and} \quad \ell(v) := \int_\Omega fv + \int_{\partial\Omega_2} xv. \tag{445}$$

Then the weak variational formulation of the PDE problem is

$$\text{Find } u \in H \text{ such that } b(u,v) = \ell(v) \ \ \forall\ v \in H. \tag{446}$$

b) Yes, the bilinear form b is symmetric. Define the linear form F by

$$F(v) := \frac{1}{2} b(v,v) - \ell(v). \tag{447}$$

Then the equivalent minimization formulation is

$$\text{Find } u \in H \text{ such that } F(u) = \min_{v\in H} F(v). \tag{448}$$

c) We proceed by verifying the conditions of the Lax-Milgram theorem hold. We must show b is bounded and coercive and ℓ is bounded. First observe b is coercive since

$$b(v,v) = \int_\Omega |Dv|^2 + v^2 + \int_{\partial\Omega_2} v^2 \geq \int_\Omega |Dv|^2 + v^2 = \|v\|_H^2. \tag{449}$$

Also,

$$\begin{aligned} |b(u,v)| &\leq \|Du\cdot Dv\|_{L^1(\Omega)} + \|uv\|_{L^1(\Omega)} + \|uv\|_{\partial\Omega_2} \\ &\leq \|Du\|_{L^2(\Omega)}\|Dv\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|u\|_{L^2(\partial\Omega_2)}\|v\|_{L^2(\partial\Omega_2)}. \end{aligned} \tag{450}$$

The first inequality is a direct application of the triangle inequality; the second inequality follows from Hölder's inequality. Now there is $C > 0$, dependent only on Ω, such that

$$\|v\|_{L^2(\partial\Omega_2)} \leq \|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{L^2(\Omega)}. \tag{451}$$

Together (450) and (451) and the fact $\|v\|_H^2 = \|v\|_{L^2(\Omega)}^2 + \|Dv\|_{L^2(\Omega)}^2$ imply

$$|b(u,v)| \leq \|Du\|_{L^2(\Omega)}\|Dv\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + C^2\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \left(1 + C^2\right)\|u\|_H\|v\|_H, \tag{452}$$

i.e., b is bounded. We now verify ℓ is bounded. Note $\sup_{x\in\partial\Omega_2} = 1$ and so

$$\begin{aligned} |\ell(v)| &\leq \|fv\|_{L^1(\Omega)} + \|xv\|_{L^1(\partial\Omega_2)} \\ &\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|x\|_{L^2(\partial\Omega_2)}\|v\|_{L^2(\partial\Omega_2)} \\ &\leq \|f\|_{L^2(\Omega)}\|v\|_H + |\partial\Omega_2|C\|v\|_H \\ &= \left(\|f\|_{L^2(\Omega)} + \pi C\right)\|v\|_H. \end{aligned} \tag{453}$$

This completes the proof.

d) For the finite element approximation, let $\mathcal{T}_h$ be a triangulation of Ω where h denotes the fineness of the triangulation mesh and the nodes are denoted by $\{N_j\}$. Let

$$H_h = \{v \in H \mid v|_K \in P_1(K) \ \forall\ K \in \mathcal{T}_h\}. \tag{454}$$

The approximate variational formulation then becomes to find $\phi_h \in H_h$ such that $b(\phi_h, v) = \ell(v)$ for all $v \in H_h$. By linearity, if $\{\gamma_i\}$ is a basis for H_h, this is equivalent to finding $\phi_h \in H_h$ such that $a(\phi_h, \gamma_i) = \ell(\gamma_i)$ for all γ_i. We take γ_i such that $\gamma_i(N_j) = \delta_{ij}$. Now we can also express $\phi_h = \sum_j \xi_j\phi_j$, thus obtaining the linear system

$$\sum_j \xi_j b(\gamma_i, \gamma_j) = \ell(\gamma_i) \qquad \Rightarrow \qquad A\xi = b, \tag{455}$$

where the entries of the stiffness matrix are $A_{ij} := b(\gamma_i, \gamma_j)$ and the entries of the load vector are $b_i := \ell(\gamma_i)$. If the enumeration of the N_j's is done efficiently, A will be sparse (since $b(\gamma_i, \gamma_j) = 0$ if $|i - j|$ is too large) and banded, allowing for efficient solving of the system. Moreover, A is positive definite (since b is an inner product). Whence is A invertible and the system has a unique solution.

If ϕ is the solution to the weak variational formulation and ϕ_h is the solution to the approximate variational formulation, then we have the bound $\|\phi - \phi_h\|_b \leq \|\phi - v\|_b$ for any $v \in H_h$ where $\|\cdot\|_b$ is the norm induced by the inner product $b(\cdot, \cdot)$. In particular, we can take the linear interpolant $\pi_h\phi \in V_h$ of ϕ, and we know $\|\phi - \pi_h\phi\|_b \leq C_w h^2$ for some constant C_w, dependent on ϕ and independent of h. From these two inequalities, we obtain the convergence rate estimate $\|\phi - \phi_h\|_b \leq C_w h^2$.

□

Spring 2016

S16.01: Determine the number of iterations necessary to find a root of $x^3 + 4x^2 - 10 = 0$ with accuracy 1×10^{-3} using the bisection method with a starting interval $[1, 2]$. Justify your answer.

Solution:

Observe $1^3 + 4 \cdot 1^2 - 10 = -5 < 0$ and $2^3 + 4 \cdot 2^2 - 10 = 14 > 0$. Since $x^3 + 4x^2 - 10$ is continuous, the intermediate value theorem asserts there is $\overline{x} \in (1, 2)$ such that $\overline{x}$ is a root of $x^3 + 4x^2 - 10$. Thus the bisection method applied here produces a sequence $\{x_n\}$ converging to a root of $x^3 + 4x^2 - 10$. Set $a_0 = 1$ and $b_0 = 2$. Then the bisection method defines $x_n := (a_n + b_n)/2$ for each $n \geq 0$ with

$$a_{n+1} := \begin{cases} a_n & \text{if } f(a_n)f(x_n) \leq 0, \\ x_n & \text{otherwise,} \end{cases} \quad \text{and} \quad b_{n+1} := \begin{cases} x_n & \text{if } f(a_n)f(x_n) \leq 0, \\ b_n & \text{otherwise.} \end{cases} \tag{456}$$

By construction, $b_{k+1} - a_{k+1} = (b_k - a_k)/2 = (b_0 - a_0)/2^k$, which converges to zero as $k \longrightarrow \infty$. And, because a $x_n \longrightarrow x^*$ with $(x^*)^3 + 4(x^*)^2 - 10 = 0$, we discover

$$|x^k - x^*| \leq \frac{b_k - a_k}{2} = \frac{b_0 - a_0}{2^{k+1}} = 2^{-(k+1)}. \tag{457}$$

Since $2^{-(8+1)} = 2^{-9} = 1/512 > 10^{-3}$ and $2^{-(9+1)} = 2^{-10} = 1/1024 < 10^{-3}$, we conclude the number of iterations needed to guarantee x_n approximates a root x^* within 10^{-3} is $\boxed{\text{9 iterations.}}$ □

S16.02: The second column of the table below gives the errors associated witht he derivative of $f(x) := e^x$ at $x = 0$ using a centered difference approximation $(f(x+h) - f(x-h))/2h$. The third column gives the errors associated with the approximation of the integral $\int_0^1 e^x$ obtained using the composite trapezoidal rule with panel width h.

h	Error in derivative	Error in integral
10^{-1}	-1.6675e-03	-1.4316e-03
10^{-2}	-1.6667e-05	-1.4318e-05
10^{-3}	-1.6667e-07	-1.4319e-07
10^{-4}	-1.6669e-09	-1.4319e-09
10^{-5}	-1.2102e-11	-1.4296e-11
10^{-6}	2.6756e-11	-1.4033e-13
10^{-7}	5.2636e-10	5.7509-14
10^{-8}	6.0775e-09	4.0767e-13
10^{-9}	-2.7229e-08	9.4156e-14

a) For h ranging from 10^{-1} to 10^{-4} why do the errors associated with the derivative and the integral decrease by a factor of $1/100$ with each refinement?

b) Why do the errors grow when $h < 10^{-5}$ for the approximation of the derivative?

c) Why don't the errors grow as much for the approximation of the integral when h is very small?

d) Why do the errors for the integral never get any smaller than $\mathcal{O}(10^{-14})$?

Solution:

a) First we analyze the derivative. Let $x_{\pm h} = x \pm h$. Then

$$f(x_{\pm h}) = f(x) \pm hf'(x) + \frac{h^2}{2}f''(x) \pm \frac{h^3}{6}f'''(x) + \mathcal{O}(h^4). \tag{458}$$

This implies

$$\frac{f(x_{+h}) - f(x_{-h})}{2h} = f'(x) + \frac{h^2}{6}f'''(x) + \mathcal{O}(h^3). \tag{459}$$

So, the leading term of the local truncation error τ_h is $e(h) = \dfrac{h^2}{6}f'''(x)$. As $h \longrightarrow 0$, $e(h)$ dominates the truncation error τ_h and so we deduce

$$\frac{\tau_{h/10}}{\tau_h} \approx \frac{e(h/10)}{e(h)} = \frac{(h/100)^2 f'''(x)/6}{h^2 f'''(x)/6} = \frac{1}{100}, \tag{460}$$

as desired.

Now we consider the integral approximation. For the composite trapezoidal rule, we write

$$\int_a^b f(x) = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] \underbrace{-\frac{b-a}{12}h^2 f''(\mu_h)}_{e(h)} \tag{461}$$

where $x_j = a + jh$ and $\mu_h \in (a,b)$ and $e(h)$ is the error term. Here $a = 0$ and $b = 1$. Then

$$\frac{e(h/10)}{e(h)} = \frac{(h/10)^2 f''(\mu_{h/10})}{h^2 f''(\mu_h)} = \frac{1}{100}\cdot e^{\mu_{h/10}-\mu_h} \tag{462}$$

Note $\mu_{h/10} - \mu_h \in (-1,1)$ and so, on average, we approximately say $e^{\mu_{h/10}-\mu_h} \approx 1$. So, we again find the factor of roughly 1/100.

b) The errors grow when $h < 10^{-5}$ for the derivative approximation because this approximation is unstable with respect to roundoff error. Suppose the precision of computations is accurate to within some max roundoff error ε. Then assume $f(x-h)$ and $f(x+h)$ are approximated $f(x-h) + \tau_1(h)$ and $f(x+h) + \tau_2(h)$, respectively, for some $\tau_1(h), \tau_2(h) \in [-\varepsilon, \varepsilon]$. Then the accumulated roundoff error $e_r(h)$ in the derivative approximation is

$$e_r(h) = \frac{\tau_1(h) - \tau_2(h)}{2h}. \tag{463}$$

The denominator in $e_r(h)$ approaches zero as $h \longrightarrow 0$. However, the numerator does not converge as $h \longrightarrow 0$. Whence $e_r(h)$ begins to "blow up" as h gets small. Being more precise, floating point computations are accurate to some machine epsilon typically on the order of 10^{-15}. So, taking $\varepsilon = 10^{-15}$ and $h = 10^{-m}$ for some $m \in \mathbb{Z}^+$, we find

$$\max_h |e_r(h)| = \frac{2\varepsilon}{2h} = \frac{\varepsilon}{h} = \frac{10^{-15}}{10^{-m}} = 10^{-15+m}, \tag{464}$$

where the max occurs when $\tau_1(h) = \tau_2(h) = \pm\varepsilon$. For $m \geq 5$, the truncation error is roughly

$$e_t(h) \approx h^2/6 \approx 10^{-2m}/6 < 2 \times 10^{-11}. \tag{465}$$

By (587) and (556), we conclude for $m > 5$ the roundoff error may begin to exceed the truncation error, and from (587) we also deduce the roundoff error increases by roughly order of magnitude for each successive refinement. This explains the error behavior seen in the derivative approximations.

c) Assume $f(x_i)$ is approximated by $\tilde{f}(x_i)$ with

$$\tilde{f}(x_i) = f(x_i) + e_i \quad \text{for } i = 0, 1, \ldots, n \tag{466}$$

where $e_i \in [-\varepsilon, \varepsilon]$ denotes the roundoff error associated with the approximation $\tilde{f}(x_i)$. Then the accumulated roundoff error $e_r(h)$ in the composite trapezoidal rule is

$$e_r(h) = \frac{h}{2}\left[e_0 + 2\sum_{j=1}^{n-1} e_j + e_n\right], \tag{467}$$

and we obtain the bound

$$|e_r(h)| \leq \frac{h}{2}\left[\varepsilon + 2\sum_{j=1}^{n-1}\varepsilon + \varepsilon\right] = \frac{h\varepsilon}{2}\left[1 + 2(n-1) + 1\right] = nh\varepsilon = (b-a)\varepsilon. \tag{468}$$

Taking the supremum over both sides, we discover

$$\sup_{h\geq 0} |e_r(h)| \leq (b-a)\varepsilon. \tag{469}$$

This shows the composite trapezoidal rule is stable with respect to roundoff error, i.e., $e_r(h)$ remains bounded as $h \longrightarrow 0$. Hence the errors in the integral approximation do not grow as much as those in the derivative approximation when h becomes very small.

d) The errors in the integral never become smaller than $\mathcal{O}(10^{-14})$ because machine epsilon is $\varepsilon = 10^{-15}$ and so roundoff errors limit the accuracy of the integral approximation.

□

S16.03: The midpoint rule for numerical integration, with error term, is given by

$$\int_{x_{-1}}^{x_1} f(x) = 2hf(x_0) + \frac{h^3}{3} f''(\xi) \tag{470}$$

where $f \in C^2[x_{-1}, x_1]$, $\xi \in (x_{-1}, x_1)$, and $x_0 - x_{-1} = x_1 - x_0 = h > 0$. Assume the interval $[0, 2]$ is decomposed into $2n$ sub-intervals of length $h = 1/n$. Determine the value of n required to approximate $\int_0^2 e^{2x}\sin(3x)$ with an accuracy 1×10^{-4} using the composite midpoint rule. Justify your answer.

Solution:

Let $x_i = 0 + h \cdot i$. Then repeatedly applying the result in (470) gives

$$\int_0^2 f(x) = \sum_{i=0}^{2n-1} \int_{x_i}^{x_{i+1}} f(x) = \sum_{i=0}^{2n-1} hf(x_{i+1/2}) + \frac{(h/2)^3}{3} \cdot f''(\xi_i) \tag{471}$$

where $\xi_i \in (x_i, x_{i+1})$ for $i = 0, \dots, 2n-1$. The error term is therefore given by

$$e(h) = \frac{h^3}{24} \sum_{i=0}^{2n-1} f''(\xi_i) = \frac{h^3}{24} \cdot 2n \cdot \underbrace{\frac{1}{2n} \sum_{i=0}^{2n-1} f''(\xi_i)}_{\overline{f''}} = \frac{h^3}{24} \cdot 2n\overline{f''} \tag{472}$$

where $\overline{f''}$ is the average value of sum terms. This gives

$$\min_i f''(\xi_i) \leq \overline{f''} \leq \max_i f''(\xi_i). \tag{473}$$

Combining (473) with the fact $\xi_i \in (0, 2)$ for $i = 0, \dots, 2n-1$, we deduce from the intermediate value theorem that there exists $\xi \in (0, 2)$ such that $\overline{f''} = f''(\xi)$.

$$e(h) = \frac{h^3}{24} \cdot 2nf''(\xi) = \frac{h^2}{12} \cdot f''(\xi). \tag{474}$$

Now let $f(x) := e^{2x}\sin(3x) \in C^2[0,2]$ and observe

$$\begin{aligned} f''(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left[2e^{2x}\sin(3x) + 3e^{2x}\cos(3x)\right] \\ &= e^{2x}\left[4\sin(3x) + 12\cos(3x) - 9\sin(3x)\right] \\ &= e^{2x}\left[12\cos(3x) - 5\sin(3x)\right]. \end{aligned} \tag{475}$$

Thus we obtain the bound

$$|f''(\xi)| \leq e^{2\cdot 2}\left[12 + 5\right] = 17e^4. \tag{476}$$

This implies that, to achieve an accuracy of 10^{-4}, it suffices to have

$$\frac{1}{12n^2}\cdot 17e^4 = \frac{h^2}{12}\cdot 17e^4 < 10^{-4} \quad \Rightarrow \quad n^2 > 10^4 e^4 \cdot \frac{17}{12} \quad \Rightarrow \quad \boxed{n > 100e^2\sqrt{\frac{17}{12}}.} \tag{477}$$

□

S16.04: Let $u(x)$ and $a(x)$ be smooth functions. Determine the order of accuracy of

$$\frac{(a_{i+1}+a_i)(u_{i+1}-u_i)-(a_i+a_{i-1})(u_i-u_{i-1})}{2h^2} \tag{478}$$

as an approximation to

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[a(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right]_{x=x_i} \tag{479}$$

where h is the mesh width, $x_i = ih$, $a_i = a(x_i)$, and $u_i = u(x_i)$.

Solution:

First observe

$$a_{i+1/2\pm1/2} = a_{i+1/2} \pm \frac{h}{2}a'_{i+1/2} + \frac{h^2}{8}a''_{i+1/2} + \mathcal{O}(h^3) \tag{480}$$

Adding the two Taylor series and then dividing by two gives

$$\frac{a_{i+1}+a_i}{2} = a_{i+1/2} + \frac{h^2}{8}a''_{i+1/2} + \mathcal{O}(h^3). \tag{481}$$

Next observe

$$u_{i+1/2\pm1/2} = u_{i+1/2} \pm \frac{h}{2}\cdot u'_{i+1/2} + \frac{h^2}{8}u''_{i+1/2} \pm \frac{h^3}{48}u'''_{i+1/2} + \mathcal{O}(h^4). \tag{482}$$

Thus

$$\frac{u_{i+1}-u_i}{h} = u'_{i+1/2} + \frac{h^3}{24}u'''_{i+1/2} + \mathcal{O}(h^4). \tag{483}$$

This shows

$$\begin{aligned}\left(\frac{a_{i+1}+a_i}{2}\right)\left(\frac{u_{i+1}-u_i}{h}\right) &= \left(a_{i+1/2} + \frac{h^2}{8}a''_{i+1/2} + \mathcal{O}(h^3)\right)\left(u'_{i+1/2} + \frac{h^3}{24}u'''_{i+1/2} + \mathcal{O}(h^4)\right)\\ &= (au')_{i+1/2} + \frac{h^2}{8}(a''u')_{i+1/2} + \mathcal{O}(h^3).\end{aligned} \tag{484}$$

With this, we see our finite difference formula is equal to

$$\begin{aligned}&\frac{1}{h}\left[\left((au')_{i+1/2} + \frac{h^2}{8}(a''u')_{i+1/2} + \mathcal{O}(h^3)\right) - \left((au')_{i-1/2} + \frac{h^2}{8}(a''u')_{i-1/2} + \mathcal{O}(h^3)\right)\right]\\ &= \frac{(au')_{i+1/2} - (au')_{i-1/2}}{h} + \frac{h}{8}\left((a''u')_{i+1/2} - (a''u')_{i-1/2}\right) + \mathcal{O}(h^2).\end{aligned} \tag{485}$$

Finally, observe that

$$(au')_{i\pm1/2} = (au')_i + \frac{h}{2}\cdot(au')'_i + \frac{h^2}{8}(au')''_i \pm \frac{h^3}{48}(au')'''_i + \mathcal{O}(h^4). \tag{486}$$

Subtracting these Taylor series from each other and then dividing by h, we find

$$\frac{(au')_{i+1/2} - (au')_{i-1/2}}{h} = (au')'_i + \frac{h^2}{24}(au')'''_i + \mathcal{O}(h^3). \tag{487}$$

Combining (485) and (487), our finite difference formula becomes

$$\begin{aligned}&\left[(au')'_i + \frac{h^2}{24}(au')'''_i + \mathcal{O}(h^3)\right] + \left[\frac{h}{8}\left((a''u')_{i+1/2} - (a''u')_{i-1/2}\right) + \mathcal{O}(h^2)\right]\\ &= (au')'_i + \frac{h^2}{24}\left((au')'''_i + \frac{(a''u')'}{8}\right) + \mathcal{O}(h^2).\end{aligned} \tag{488}$$

Thus we conclude the finite difference is an $\mathcal{O}(h^2)$ approximation of $(au')'_i$. □

S16.05: Consider the initial value problem (IVP)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\alpha y + f(y), \quad y(0) = y_0, \tag{489}$$

where $\alpha > 0$ and $|\partial f/\partial y| < \beta$.

a) What value of λ associated with the model problem $\mathrm{d}y/\mathrm{d}t = \lambda y$ should be used when estimating an acceptable time step for the numerical solution of the IVP?

b) Assume that Forward Euler (FE) is used to create approximation solutions to the IVP. Give an estimate of the largest time step that should be used when seeking a qualitatively accurate solution.

c) If $y(t)$ is a solution of the IVP, let $z(t)$ be defined by the relation $y(t) = e^{-\alpha t}z(t)$. Derive the initial value problem $z(t)$ satisfies.

d) Assume that FE is used to create approximation solutions to the IVP for $z(t)$. Give an estimate of the largest time step that should be used when seeking a qualitatively accurate solution.

e) For what values of α and β would it be advisable to solve the IVP for $z(t)$ rather than $y(t)$? Explain.

Solution:

a) We claim we can use any λ satisfying the inequality $|\lambda + \alpha| < \beta$. Indeed, observe

$$\lambda y = \frac{\mathrm{d}y}{\mathrm{d}t} = -\alpha y + f(y) \quad \Leftrightarrow \quad (\lambda + \alpha)y = f(y) \quad \Rightarrow \quad |\lambda + \alpha| = \left|\frac{\partial f}{\partial y}\right| < \beta. \tag{490}$$

b) Let $\{w_n\}$ be the sequence produced by the FE method. To analyze the numerical stability, we use the model problem $y' = \lambda y$. In this case,

$$w_{n+1} = w_n + h(w_n)' = w_n + h\lambda w_n = (1 + h\lambda)w_n. \tag{491}$$

The solution is absolutely stable provided there is $C > 0$ such that $|w_n| \leq C|w_0|$ for all $n \geq 0$. Since w_{n+1} is a scalar multiple of w_n, we deduce the scheme is absolutely stable provided $|1 + h\lambda| \leq 1$. Taking $\lambda \in \mathbb{R}$ with $\lambda \neq 0$, this implies $\lambda < 0$ and

$$-1 \leq 1 + h\lambda \leq 1 \quad \Leftrightarrow \quad 0 \leq h \leq -\frac{2}{\lambda}. \tag{492}$$

However, from a),

$$-\beta < \lambda + \alpha < \beta \quad \Leftrightarrow \quad -(\beta+\alpha) < \lambda < \beta - \alpha \quad \Rightarrow \quad -\frac{1}{\lambda} > \frac{1}{\alpha+\beta}. \tag{493}$$

Thus, choosing $\boxed{h < 2/(\alpha+\beta)}$ causes the needed relation (492) to hold, i.e.,

$$h < \frac{2}{\alpha+\beta} < -\frac{2}{\lambda}. \tag{494}$$

c) Through direct differentiation, we find

$$z'(t) = \left[y'(t) + \alpha y(t)\right] e^{\alpha t} = \left[-\alpha y(t) + f(y(t)) + \alpha y(t)\right] e^{\alpha t} = f(y(t)) e^{\alpha t} = f\left(z(t)e^{-\alpha t}\right) e^{\alpha t}. \tag{495}$$

Thus z satisfies

$$\frac{\mathrm{d}z}{\mathrm{d}t} = f\left(z(t)e^{-\alpha t}\right) e^{\alpha t}, \quad z(0) = y_0. \tag{496}$$

d) We assume $f(0) = 0$ so that

$$|z'| = \left|f\left(ze^{-\alpha t}\right)\right| = \left|\int_0^{ze^{-\alpha t}} f'(s)\ \mathrm{d}s\right| \leq \beta\left|ze^{-\alpha t}\right| \leq \beta|z|. \tag{497}$$

So, here we need $|\lambda| \leq \beta$ when estimating an acceptable time step for the numerical solution of the IVP for z. And, as in b), the region of absolute stability for the FE method is $h\lambda$ satisfying $|1 + h\lambda| \leq 1$. So, here choosing $\boxed{h < 2/\beta}$ gives the desired relation, i.e.,

$$0 \leq h < \frac{2}{\beta} \leq -\frac{2}{\lambda}. \tag{498}$$

e) When $\beta \ll \alpha$, it may be advisable to solve the IVP for $z(t)$ rather than the one for $y(t)$ since in this case our estimate for the largest acceptable time step for a qualitatively accurate solution approximation of $z(t)$ is $2/\beta$, which is much greater the acceptable time step $2/(\alpha+\beta)$ for the IVP for $y(t)$, i.e., $2/\beta \gg 2/(\alpha+\beta)$.

□

S16.06: Consider the system of partial differential equations

$$u_t + Au_x = 0 \tag{499}$$

with

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \tag{500}$$

to be solved for $0 \leq x \leq 1$ and $t \geq 0$ with smooth initial data $u(x, 0) = u_0(x)$ and boundary conditions at $x = 0$ and $x = 1$. The entries of A, denoted a_{ij}, are constant real values.

a) What conditions on A and what boundary conditions are need for well-posedness of this problem?

b) Give a stable, convergent numerical approximation to this initial value problem. (Justify your statements.)

Solution:

a) For the system to be hyperbolic the matrix A must be diagonalizable with real eigenvalues. Let d_1 and d_2 be the eigenvalues of A. If $d_i > 0$, then a wave will propagate in the positive x direction and a boundary condition $g_i(t)$ must be prescribed at $x = 0$. If instead $d_i < 0$, then a boundary condition $g_i(t)$ must be prescribed at $x = 1$.

b) We propose the Crank-Nicolson scheme

$$\frac{v_i^{n+1} - v_i^n}{k} + A\left(\frac{v_{i+1}^{n+1} - v_{i-1}^{n+1} + v_{i+1}^n - v_{i-1}^n}{4h}\right) = 0. \tag{501}$$

The Lax-Equivalence theorem asserts a consistent scheme is stable if and only if it converges. So, we show the scheme is consistent (Step 1). Then we show the scheme is unconditionally stable (Step 2).

Step 1: The Crank-Nicolson scheme stencil is centered about $v_i^{n+1/2}$. So, Taylor expanding about this point, we find

$$v_i^{n+1/2\pm 1/2} = v_i^{n+1/2} \pm \frac{k}{2}(v_i^{n+1/2})_t + \frac{k^2}{8}(v_i^{n+1/2})_{tt} \pm \frac{k^3}{48}(v_i^{n+1/2})_{ttt} + \mathcal{O}(k^4). \tag{502}$$

This implies

$$\frac{v_i^{n+1} - v_i^n}{k} = (v_i^{n+1/2})_t + \frac{k^2}{24}(v_i^{n+1/2})_{ttt} + \mathcal{O}(k^3) \tag{503}$$

and

$$\frac{v_i^{n+1} + v_i^n}{2} = v_i^{n+1/2} + \frac{k^2}{4}(v_i^{n+1/2})_{tt} + \mathcal{O}(k^3). \tag{504}$$

In analogous fashion to (503), we obtain

$$\frac{v_{i+1}^{n+1/2} - v_{i-1}^{n+1/2}}{2h} = (v_i^{n+1/2})_x + \frac{h^2}{6}(v_i^{n+1/2})_{xxx} + \mathcal{O}(h^2). \tag{505}$$

Combining (504) and (505) yields

$$\frac{v_{i+1}^{n+1} - v_{i-1}^{n+1} + v_{i+1}^n - v_{i-1}^n}{4h} = \frac{v_{i+1}^{n+1/2} - v_{i-1}^{n+1/2}}{2h} + \mathcal{O}(k^2/h) = (v_i^{n+1/2})_x + \mathcal{O}(h^2 + k^2/h). \tag{506}$$

Let $P_{k,h}$ be the difference operator associated with the scheme defined by (501) and $P := \partial_t + A\partial_x$. Then (503) and (506) together imply

$$(P - P_{k,h})v_i^{n+1/2} = \mathcal{O}(k^2 + h^2 + k^2/h). \tag{507}$$

Thus the scheme is consistent provided $k^2/h \longrightarrow 0$ as $k, h \longrightarrow 0$.

Step 2: We proceed by using Von Neumann analysis. The condition for stability is that for each $T > 0$, there is a constant C_T such that for $0 \leq nk \leq T$,

$$\|G^n\| \leq C_T, \tag{508}$$

where G is the amplification matrix and $\|\cdot\|$ is any matrix norm (since all finite dimensional norms are equivalent). Substituting v_i^n with $G^n e^{im\theta}$ in the scheme gives

$$\frac{G - I}{k} = \frac{A}{4h}(G + I)\left(e^{i\theta} - e^{-i\theta}\right) \quad \Leftrightarrow \quad G = I + \frac{\lambda i \sin\theta}{2} A(I + G). \tag{509}$$

Rearranging, we see

$$G = \left(I + \frac{\lambda i \sin\theta}{2} A\right)^{-1} \left(I + \frac{\lambda i \sin\theta}{2} A\right) = P\left(I + \frac{\lambda i \sin\theta}{2} D\right)^{-1} \left(I + \frac{\lambda i \sin\theta}{2} D\right) P^{-1} \tag{510}$$

where $A = PDP^{-1}$ with $D = \text{diag}(d_1, d_2)$ diagonal and P invertible. Then

$$G = P \underbrace{\begin{pmatrix} \frac{1+i(d_1\lambda\sin\theta)/2}{1-i(d_1\lambda\sin\theta)/2} & 0 \\ 0 & \frac{1+i(d_2\lambda\sin\theta)/2}{1-i(d_2\lambda\sin\theta)/2} \end{pmatrix}}_{M} P^{-1} = PMP^{-1}, \tag{511}$$

where M is the underbraced matrix. Note the diagonal entries of M each have norm one since they are the ratio of complex conjugates. Thus $\|M\|_2 = 1$. And, it follows from induction that $G^n = PM^nP^{-1}$. Hence

$$\|G^n\|_2 = \|PM^nP^{-1}\|_2 \leq \|P\|_2\|M^n\|_2\|P^{-1}\|_2 \leq \|P\|_2\|M\|_2^n\|P^{-1}\|_2 = \|P\|_2\|P^{-1}\|_2. \tag{512}$$

This shows (508) holds with $C_T = \|P\|_2\|P^{-1}\|_2$, and we are done.

□

S16.07: Consider the initial value problem

$$u_t = -u^2 u_x + \varepsilon u_{xx} \tag{513}$$

for $\varepsilon > 0$, to be solved for $0 \leq x \leq 1$ and $t > 0$ with smooth initial data $u(x,0) = u_0(x)$ and periodic boundary conditions $u(x+1,t) = u(x,t)$.

a) Construct a second order accurate convergent method.

b) Construct a method which remains convergent as $\varepsilon \longrightarrow 0$.

Solution:

a) See b).

b) We shall define a single method that answers both parts at once. Also, we presume the problem statement means an order $(1,2)$ accurate method. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(u) := u^3/3$ so that $u_t + f(u)_x = \varepsilon u_{xx}$. Then we propose using the scheme

$$\frac{u_i^{n+1} - [2\mu\varepsilon u_i^n + \frac{1-2\mu\varepsilon}{2}(u_{i+1}^n + u_{i-1}^n)]}{k} + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2h} - \frac{\varepsilon}{h^2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) = 0, \tag{514}$$

taking $\mu := k/h^2$. Note this scheme becomes the Lax-Frierichs scheme when $\varepsilon = 0$.

This problem is in two parts. When $\varepsilon > 0$, the PDE is parabolic with a smooth solution and so we may use the usual notion of consistency and stability to obtain convergence. Note in this case μ is held fixed. When $\varepsilon = 0$, we hold λ fixed and verify Lipschitz continuity of the numerical flux function and the method is TV-stable, which is sufficient to conclude convergence as $k, h \longrightarrow 0$. Thus, the method is convergent as $k, h \longrightarrow 0$ for arbitrary $\varepsilon > 0$ and also for $\varepsilon = 0$. (Step 1) We first show the method is ℓ_1 contracting. (Step 2) Then we show convergence for $\varepsilon > 0$ and (Step 3) for $\varepsilon = 0$.

Step 1: We now show the method is ℓ_1 contracting. Let $\{u_i^n\}$ and $\{v_i^n\}$ be two collections of grid functions and define $e_i^n = u_i^n - v_i^n$. Also note our scheme may be written in the form

$$\begin{aligned} u_i^{n+1} &= 2\mu\varepsilon u_i^n + \frac{1-2\mu\varepsilon}{2}\left(u_{i+1}^n + u_{i-1}^n\right) + \frac{\lambda}{2}\left(f(u_{i+1}^n) - f(u_{i-1}^n)\right) + \varepsilon\mu\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) \\ &= \frac{1}{2}(u_{i+1}^n + u_{i-1}^n) + \frac{\lambda}{2}\left(f(u_{i+1}^n) - f(u_{i-1}^n)\right), \end{aligned} \tag{515}$$

where $\lambda := k/h$. This implies

$$\begin{aligned} e_i^{n+1} &= \frac{1}{2}\left(e_{i-1}^n + e_{i+1}^n\right) + \frac{\lambda}{2}\left[\left(f(u_{i+1}^n) - f(v_{i+1}^n)\right) - \left(f(u_{i-1}^n) - f(v_{i-1}^n)\right)\right] \\ &= \frac{1}{2}\left(e_{i-1}^n + e_{i+1}^n\right) + \frac{\lambda}{2}\left[f'(\theta_{i+1}^n)e_{i+1}^n - f'(\theta_{i-1}^n)e_{i-1}^n\right] \\ &= \left(\frac{1}{2} + \frac{\lambda}{2}f'(\theta_{i+1})\right)e_{i+1}^n + \left(\frac{1}{2} - \frac{\lambda}{2}f'(\theta_{i-1})\right)e_{i-1}^n. \end{aligned} \tag{516}$$

Here $\theta_{i\pm1}$ exists, by Taylor's theorem, between $u_{i\pm1}^n$ and $v_{i\pm1}^n$, respectively. We assume k and h are chosen so the CFL conditions holds, i.e.,

$$\left|\lambda f'(u)\right| \leq 1 \quad \forall \ \min_j(u_j^n, v_j^n) \leq u \leq \max_j(u_j^n, v_j^n). \tag{517}$$

(Note this can be done even when μ is fixed by choosing h sufficiently small since $\lambda = h\mu$.) In particular, we have $|\lambda f'(\theta_{i\pm1})| \leq 1$, and so

$$\frac{1}{2} \pm \frac{1}{2}\left(\lambda f'(\theta_{j\pm1})\right) \geq \frac{1}{2} - \frac{1}{2} = 0. \tag{518}$$

This shows the coefficients of e_{i+1}^n and e_{i-1}^n are nonnegative. Whence

$$\begin{aligned} \|e^{n+1}\|_1 &= h\sum_i |e_i^{n+1}| \\ &\leq \left(\frac{1}{2} + \frac{\lambda}{2}f'(\theta_{i+1})\right) h\sum_i |e_{i+1}^n| + \left(\frac{1}{2} - \frac{\lambda}{2}f'(\theta_{i-1})\right) h\sum_i |e_{i-1}^n| \\ &= \left(\frac{1}{2} + \frac{\lambda}{2}f'(\theta_i)\right) h\sum_i |e_i^n| + \left(\frac{1}{2} - \frac{\lambda}{2}f'(\theta_i)\right) h\sum_i |e_i^n| \\ &= h\sum_i |e_i^n| \\ &= \|e^n\|_1. \end{aligned} \tag{519}$$

Note $\sum_i |e_i^n| = \sum_i |e_{i+j}^n|$ for all j since the grid functions are periodic, which is how the third line follows from the second. This shows the method is ℓ_1 contracting.

Step 2: Assume $\varepsilon > 0$. The method is second order in space since we use the standard centered difference formulas for each spatial derivative. For the time derivative, note we use the standard forward time with the second term being an $\mathcal{O}(k^2)$ approximation of u_i^n by means of a linear combination. So, we conclude the method is accurate of order $(1,2)$ and, thus, consistent. Now, taking $e_i^n = u_i^n - 0$, the method being ℓ_1 contracting implies

$$\|u^n\|_1 \leq \|u^{n-1}\|_1 \leq \cdots \leq \|u^0\|_1. \tag{520}$$

Thus the method is stable and we conclude it converges when $\varepsilon > 0$.

Step 3: Now we assume $\varepsilon = 0$ and recall the following theorem:

Theorem:[3] Suppose $\{u_i^n\}$ with step sizes k and h is generated by a numerical method in conservation form with a Lipschitz continuous numerical flux, consistent with some scalar law. If the method is TV-stable, then the method will converge to a weak solution of the conservation law as $k, h \longrightarrow 0$.

The scalar law at hand is $u_t + f(u)_x = 0$. Observe our scheme may be rewritten in conservation form as

$$u_i^{n+1} = u_i^n - \lambda\left[F(u_i^n, u_{i+1}^n) - F(u_{i-1}^n, u_i^n)\right], \tag{521}$$

where the numerical flux function F is defined by

$$F(u,v) := \left(\frac{1}{2\lambda}\right)(u-v) + \frac{f(u)+f(v)}{2}. \tag{522}$$

Note, keeping λ fixed, $1/2\lambda$ remains constant as $k, h \longrightarrow 0$.

To show the numerical flux F is Lipschitz continuous, we must show there is $L > 0$ such that

$$|F(u,v) - f(w)| \leq L \cdot \max\{|u-w|, |v-w|\} \tag{523}$$

for all u, v with $|u-w|$ and $|v-w|$ sufficiently small. Recall we assume the grid functions are bounded. This implies the support of f is closed and bounded and, thus, compact. Since f is continuously

[3] See Theorem 15.2 on page 164 of Leveque.

differentiable on a compact set, it is Lipschitz continuous. Then

$$\begin{aligned}|F(u,v) - f(w)| &= \left|\frac{1}{2\lambda}(u-v) + \frac{1}{2}\left(f(u)+f(v)\right) - f(w)\right| \\ &= \left|\frac{1}{2\lambda}\left[(u-w)+(w-v)\right] + \frac{1}{2}\left[(f(u)-f(w))+(f(v)-f(w))\right]\right| \\ &\leq \left(\frac{1}{\lambda} + \operatorname{lip}(f)\right)\cdot \max\{|u-w|,|v-w|\},\end{aligned} \tag{524}$$

and we see (523) holds by taking $K = (1/\lambda + \operatorname{lip}(f))$.

We now verify TV-stability. Choosing $v_i^n = u_{i+1}^n$, (519) implies the method is TVD since

$$\operatorname{TV}(u^{n+1}) = \frac{1}{h}\|u_{i+1}^{n+1} - u_i^{n+1}\|_1 \leq \frac{1}{h}\|u_{i+1}^n - u_i^n\|_1 = \operatorname{TV}(u^n). \tag{525}$$

Then, by induction, we conclude $\operatorname{TV}(u^n) \leq \operatorname{TV}(u^0)$ for all $n \in \mathbb{Z}^+$ and the scheme is TV-stable. By the above, we conclude the sequence converges to a weak solution of the conservation law. And, the above shows the method converges to the "vanishing viscosity" solution that might be used to define the physically relevant weak solution to the conservation law.

□

S16.08: The following elliptic problem is approximated by the finite element method,

$$\begin{aligned} -\nabla \cdot (a(x)\nabla u(x)) = f(x) &\quad \text{in } \Omega, \\ u = u_0 &\quad \text{on } \Gamma_1, \\ \frac{\partial u}{\partial x_1} + u = 0 &\quad \text{on } \Gamma_2, \\ \frac{\partial u}{\partial x_2} = 0 &\quad \text{on } \Gamma_3, \end{aligned} \tag{526}$$

where

$$\begin{aligned} \Omega &= \{(x_1, x_2) \ : \ x_1, x_2 \in (0,1)\}, \\ \Gamma_1 &= \{(x_1, x_2) \ : \ x_1 = 0, \ x_2 \in [0,1]\}, \\ \Gamma_2 &= \{(x_1, x_2) \ : \ x_1 = 1, \ x_2 \in [0,1]\}, \\ \Gamma_3 &= \{(x_1, x_2) \ : \ x_1 \in (0,1), \ x_2 = 0, 1\}, \end{aligned} \tag{527}$$

and

$$0 < A \leq a(x) \leq B \ \text{ a.e. in } \Omega, \ f \in L^2(\Omega), \tag{528}$$

and $u_0|_{\Gamma_1}$ is the trace of a function $u_0 \in H^1(\Omega)$.

a) Determine an appropriate weak variational formulation of the problem.

b) Prove conditions on the corresponding linear and bilinear forms which are needed for the existence and uniqueness of the solution.

c) Set up a finite element approximation using P_1 elements and a set of basis functions such that the associated linear system is sparse and of band structure. Discuss the properties of the linear system thus obtained and give the rate of convergence of the approximation.

(Solution on next page.)

Solution:

a) By hypothesis, $u_0 \in H^1(\Omega)$. Define $\phi := u - u_0$ so that $u = u_0 + \phi$ and ϕ is in the Hilbert space $H := \{v \in H^1(\Omega) \ : \ v|_{\Gamma_1} = 0\}$ equipped with the norm $\|\cdot\|_H := \|\cdot\|_{H^1(\Omega)}$. Then ϕ satisfies

$$\begin{aligned} -\nabla\cdot(a\nabla\phi) &= \tilde{f} \ \ \text{in } \Omega, \\ \phi &= 0 \ \ \text{on } \Gamma_1, \\ \frac{\partial\phi}{\partial x_1} + \phi &= g \ \ \text{on } \Gamma_2, \\ \frac{\partial\phi}{\partial x_2} &= h \ \ \text{on } \Gamma_3, \end{aligned} \tag{529}$$

where

$$\tilde{f} := f + \nabla\cdot(a\nabla u_0), \quad g := -\frac{\partial u_0}{\partial x_1} - u_0, \quad h := -\frac{\partial u_0}{\partial x_2}. \tag{530}$$

Note that $\tilde{f} \in L^2(\Omega)$, $g \in L^2(\Gamma_2)$, and $h \in L^2(\Gamma_3)$ since $u_0 \in H^1(\Omega)$ and $f \in L^2(\Omega)$. Then, for each test function $v \in H$, integrating by parts yields

$$\begin{aligned} \int_\Omega \tilde{f}v &= -\int_\Omega \nabla\cdot(a\nabla\phi)v \\ &= \int_\Omega a\nabla\phi\cdot\nabla v - \int_{\partial\Omega}(n\cdot a\nabla\phi)v \\ &= \int_\Omega a\nabla\phi\cdot\nabla v - \int_{\Gamma_2}(n\cdot a\nabla\phi)v - \int_{\Gamma_3,x_2=0}(n\cdot a\nabla\phi)v - \int_{\Gamma_3,x_2=1}(n\cdot a\nabla\phi)v \\ &= \int_\Omega a\nabla\phi\cdot\nabla v - \int_{\Gamma_2} a(g-\phi)v + \int_{\Gamma_3,x_2=0} ahv - \int_{\Gamma_3,x_2=1} ahv. \end{aligned} \tag{531}$$

Define the bilinear form b and linear form ℓ by

$$b(\phi,v) := \int_\Omega a\nabla\phi\cdot\nabla v + \int_{\Gamma_2} a\phi v \quad \text{and} \quad \ell(v) := \int_\Omega \tilde{f}v - \int_{\Gamma_3,x_2=0} ahv + \int_{\Gamma_3,x_2=1} ahv + \int_{\Gamma_2} agv. \tag{532}$$

Then the weak variational form of the problem is

$$\text{Find } \phi \in H \text{ such that } b(\phi,v) = \ell(v) \ \ \forall\, v \in H. \tag{533}$$

b) We prove there exists a unique solution to the problem by verifying the assumptions of the Lax-Milgram theorem are satisfied. Namely, we show b is bounded and coercive and ℓ is bounded. (n.b. symmetry of b follows directly from its definition and commutativity of scalar multiplication.) We see ℓ is bounded since

$$\begin{aligned} |\ell(v)| &\leq \|\tilde{f}v\|_{L^1(\Omega)} + B\|hv\|_{L^1(\Gamma_3)} + B\|gv\|_{L^1(\Gamma_2} \\ &\leq \|\tilde{f}\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + B\|h\|_{L^2(\Gamma_3)}\|v\|_{L^2(\Gamma_3)} + B\|g\|_{L^2(\Gamma_2)}\|v\|_{L^2(\Gamma_2)}. \end{aligned} \tag{534}$$

The first inequality is a direct application of the triangle inequality and the second follows from Hölder's inequality. There are $C_2, C_3 > 0$ such that

$$\|v\|_{L^2(\Gamma_i)} \leq \|v\|_{L^2(\partial\Omega)} \leq C_i\|v\|_H \quad \text{for } i = 1, 2. \tag{535}$$

Combining (534) and (535) with the fact $\|u\|_H^2 = \|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2$, we deduce

$$|\ell(v)| \leq \left(\|\tilde{f}\|_{L^2(\Omega)} + BC_3\|h\|_{L^2(\Gamma_3)} + BC_2\|g\|_{L^2(\Gamma_2)}\right)\|v\|_H, \tag{536}$$

as desired. We now establish the boundedness and coercivity of b. Indeed,

$$\begin{aligned} |b(\phi, v)| &\leq B\left(\|\nabla\phi \cdot \nabla v\|_{L^1(\Omega)} + \|\phi v\|_{L^1(\Gamma_1)}\right) \\ &\leq B\left(\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + C_2^2\|\phi\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}\right) \\ &\leq B\left(1 + C_2^2\right)\|\phi\|_H\|v\|_H. \end{aligned} \tag{537}$$

The first inequality holds by applying the triangle inequality. The second follows Hölder's inequality and using (535). Next observe

$$b(u, u) = \int_\Omega a|\nabla u|^2 + \int_{\Gamma_2} au^2 \geq A\left(\int_\Omega |\nabla u|^2 + \int_{\Gamma_2} u^2\right) \geq A\|\nabla u\|_{L^2(\Omega)}^2. \tag{538}$$

Because $u|_{\Gamma_1} = 0$ and $|\Gamma_1| > 0$, there is $C^* > 0$ such that $\|\nabla u\|_{L^2(\Omega)} \geq C^*\|u\|_H$. Thus we obtain the inequality

$$b(u, u) \geq AC^*\|u\|_H^2, \tag{539}$$

thereby establishing the coercivity of b. This also shows b is an inner product and completes the proof.

c) For the finite element approximation, let $\mathcal{T}_h$ be a triangulation of Ω where h denotes the fineness of the triangulation mesh and the nodes are denoted by $\{N_j\}$. Let

$$H_h = \{v \in H \mid v|_K \in P_1(K) \; \forall \; K \in \mathcal{T}_h\}\,. \tag{540}$$

The approximate variational formulation then becomes to find $\phi_h \in H_h$ such that $b(\phi_h, v) = \ell(v)$ for all $v \in H_h$. By linearity, if $\{\gamma_i\}$ is a basis for H_h, this is equivalent to finding $\phi_h \in H_h$ such that $a(\phi_h, \gamma_i) = \ell(\gamma_i)$ for all γ_i. We take γ_i such that $\gamma_i(N_j) = \delta_{ij}$. Now we can also express $\phi_h = \sum_j \xi_j \phi_j$, thus obtaining the linear system

$$\sum_j \xi_j b(\gamma_i, \gamma_j) = \ell(\gamma_i) \qquad \Rightarrow \qquad A\xi = b, \tag{541}$$

where the entries of the stiffness matrix are $A_{ij} := b(\gamma_i, \gamma_j)$ and the entries of the load vector are $b_i := \ell(\gamma_i)$. If the enumeration of the N_j's is done efficiently, A will be sparse (since $b(\gamma_i, \gamma_j) = 0$ if $|i - j|$ is too large) and banded, allowing for efficient solving of the system. Moreover, A is positive definite (since b is an inner product). Whence is A invertible and the system has a unique solution.

If ϕ is the solution to the weak variational formulation and ϕ_h is the solution to the approximate variational formulation, then we have the bound $\|\phi - \phi_h\|_b \leq \|\phi - v\|_b$ for any $v \in H_h$ where $\|\cdot\|_b$ is the norm induced by the inner product $b(\cdot, \cdot)$. In particular, we can take the linear interpolant $\pi_h\phi \in V_h$ of ϕ, and we know $\|\phi - \pi_h\phi\|_b \leq C_w h^2$ for some constant C_w, dependent on ϕ and independent of h. From these two inequalities, we obtain the convergence rate estimate $\|\phi - \phi_h\|_b \leq C_w h^2$.

□

Fall 2016

F16.01: Let $A \in \mathbb{R}^{n\times n}$ be positive-definite and symmetric. A set of vectors $\{p_i\}_{i=1}^n$ are A-orthogonal if they are orthogonal with respect to the A-inner product, specifically, $\langle Ap_i, p_j\rangle = \delta_{i,j}$ where $\delta_{i,j}$ is the Kronecker δ. Consider the problem of finding a solution to $Ax = b$.

a) Derive expressions for c_i for $i = 1, \dots, n$ such that $x = \sum_{i=1}^n c_i p_i$ is a solution to $Ax = b$.

b) Given a set of n linearly independent vectors $\{q_i\}$, present formulas for a method that can be used to construct a set of A-orthogonal vectors $\{p_i\}$ from the set of vectors $\{q_i\}$.

Solution:

a) We assume we are given the A-orthogonal set of vectors $\{p_i\}$. Then observe that

$$\langle p_j, b\rangle = \langle p_j, Ax\rangle = \langle p_j, A\left(\sum_{i=1}^n c_i p_i\right)\rangle = \sum_{i=1}^n c_i \langle p_j, Ap_i\rangle = \sum_{i=1}^n c_i \delta_{i,j} = c_j \quad \text{for } j = 1, \dots, n. \tag{542}$$

This shows $\boxed{c_i = \langle p_i, b\rangle \text{ for } i = 1, \dots, n.}$

b) We construct the set $\{p_i\}$ is analogous fashion to the Gram-Schmidt procedure for constructing an orthonormal set of vectors. The difference is that here we use the A-inner product rather than the standard inner product on $\mathbb{R}^n$. That is, we write

$$\boxed{p_i = \frac{q_i - \sum_{k=1}^{i-1} \langle p_k, Aq_i\rangle p_k}{\sqrt{\langle q_i, Aq_i\rangle - \sum_{k=1}^{i-1} \langle p_k, Aq_i\rangle^2}} \quad \text{for } i = 1, \dots, n.} \tag{543}$$

We could do this in two steps more intuitively. First‘define the set of vectors $\{v_i\}_{i=1}^n$ by

$$v_i := q_i - \sum_{k=1}^{i-1} \langle p_k, Aq_i\rangle \quad \text{for } i = 1, \dots, n. \tag{544}$$

This definition makes it so that $\langle Av_i, v_j\rangle = 0$ whenever $i \neq j$. By appropriate scaling we can use each v_i to write an alternate formula for each p_i. Namely,

$$p_i = \frac{v_i}{\sqrt{\langle Av_i, v_i\rangle}} \quad \text{for } i = 1, \dots, n. \tag{545}$$

□

F16.02: Consider using Gauss-Seidel to compute the solution to the following system of equations

$$\begin{pmatrix} -1 & 0 & 0 \\ 0 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{546}$$

a) Give the vector b and the matrix G that arise when the Gauss-Seidel method is expressed in the form $x^{k+1} = Gx^k + b$.

b) Does the Gauss-Seidel iteration converge for this system of equations? Justify your answer.

Solution:

a) Write $A = A_L + D + A_R$ where A_L is the strictly lower triangular portion of A, D consists of A's diagonal elements, and A_R is the strictly upper triangular portion. Then $y = (A_L + D + A_R)x$ implies

$$(A_L + D)x = -A_R x + y \qquad \Rightarrow \qquad x = -(A_L + D)^{-1} A_R x + (A_L + D)^{-1} y. \tag{547}$$

This gives rise to the Gauss-Seidel iteration

$$x^{k+1} = \underbrace{-(A_L + D)^{-1} A_R}_{G} x^k + \underbrace{(A_L + D)^{-1} y}_{b}. \tag{548}$$

Observe we may perform elementary row operations on $(A_L + D \mid I)$ to obtain

$$\left(\begin{array}{ccc|ccc} -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 & 1 & 0 \\ 0 & 1 & -2 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1/2 & 0 \\ 0 & 1 & -2 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1/2 & 0 \\ 0 & 0 & 1 & 0 & -1/4 & -1/2 \end{array}\right). \tag{549}$$

This implies

$$G = -\begin{pmatrix} -1 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 0 & 1/4 \end{pmatrix} \tag{550}$$

and

$$b = (A_L + D)^{-1} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & -1/4 & -1/2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -1/2 \\ -3/4 \end{pmatrix}. \tag{551}$$

b) Recall the iteration converges if and only if $\rho(G) < 1$. Observe the characteristic polynomial for G is

$$\chi(\lambda) := \det(\lambda I - G) = \lambda^2(\lambda - 1/4), \tag{552}$$

which implies the eigenvalues of G are 0 and $1/4$. Whence $\rho(G) = 1/4 < 1$, and so the iteration does converge to a solution to the system of equations. □

F16.03: A local truncation error estimate for the forward-difference approximation

$$f'(x_0) \approx \frac{1}{h}\left[f(x_0+h) - f(x_0)\right] \tag{553}$$

can be expressed as

$$f'(x_0) = \frac{1}{h}\left[f(x_0+h) - f(x_0)\right] - \frac{h}{2}f''(x_0) - \frac{h^2}{6}f'''(x_0) + \mathcal{O}(h^3). \tag{554}$$

Use Richardson extrapolation to derive an $\mathcal{O}(h^3)$ approximation for $f'(x_0)$.

Solution:

Define $N(h)$ to be the forward-difference approximation, i.e.,

$$N(h) := \frac{1}{h}\left[f(x_0+h) - f(x_0)\right]. \tag{555}$$

With Richardson extrapolation, we form a linear combination of $N(h)$, $N(h/2)$, and $N(2h)$ such that their combination yields a $\mathcal{O}(h^3)$ approximation of $f'(x_0)$, e.g., of the form

$$M(h) = aN(h) + bN(2h) + cN(4h) \tag{556}$$

for some $a, b, c \in \mathbb{R}$. We will be done if we cancel the $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ terms as shown in (554) for the truncation error in $N(h)$, $N(2h)$, and $N(4h)$, and if we additionally have $a + b + c = 1$ so the result is scaled appropriately. This gives rise to the linear system

$$\begin{bmatrix} 1 & 2 & 4 \\ 1^2 & 2^2 & 4^2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{557}$$

Writing this as an augmented matrix and row reducing we find

$$\left[\begin{array}{ccc|c} 1 & 2 & 4 & 0 \\ 0 & 2 & 12 & 0 \\ 0 & 1 & 3 & -1 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & 4 & 0 \\ 0 & 1 & 6 & 0 \\ 0 & 0 & 3 & 1 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 2 & 0 & -4/3 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 1/3 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 0 & 8/3 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 1/3 \end{array}\right], \tag{558}$$

which implies $(a, b, c) = (8/3, -2, 1/3)$. Whence we conclude

$$M(h) = \frac{8}{3} \cdot N(h) - 2 \cdot N(2h) + \frac{1}{3} \cdot N(4h) \tag{559}$$

gives an $\mathcal{O}(h^3)$ approximation of $f'(x_0)$. □

F16.04: Consider a function $f \in C^2[a, b]$.

a) Derive Newton's method for approximating a zero $p \in [a, b]$ of this function, starting with an initial approximation $p_0 \in [a, b]$. Can you state conditions under which Newton's method will converge at least quadratically?

b) Consider $f(x) = e^x - x - 1$. Do we expect any difficulty when applying Newton's method to approximate the zero $p = 0$ of f? If yes, how could this be avoided? How can we improve the rate of convergence? Observe that the first nine iterations obtained using Newton's method starting from $p_0 = 1$ are

$$\begin{aligned} &p_1 = 0.58198, \quad p_2 = 0.31906, \quad p_3 = 0.16800, \quad p_4 = 0.08635, \\ &p_5 = 0.04380, \quad p_6 = 0.02206, \quad p_7 = 0.01107, \quad p_8 = 0.005545. \end{aligned} \tag{560}$$

Solution:

a) Let $x \in [a, b]$. Then Taylor's remainder theorem asserts there is ξ between x and p such that

$$0 = f(p) = f(x) + f'(x) \cdot (p - x) + \frac{f''(\xi)}{2}(p - x)^2. \tag{561}$$

Assuming $|p - x|$ is sufficiently small, we obtain $|p - x|^2 \ll |p - x|$ and make the approximation

$$0 = f(p) \approx f(x) + f'(x) \cdot (p - x). \tag{562}$$

Rearranging to solve for p gives

$$p \approx x - \frac{f(x)}{f'(x)}. \tag{563}$$

So, suppose we have an initial iterate x_0 that is sufficiently close to p. Then (563) implies we can obtain a better approximation by writing

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}. \tag{564}$$

This sets the stage for Newton's method, which we define, in general, by the iteration

$$p_{n+1} = p_n - \frac{f(p_{n-1})}{f'(p_{n-1})} \quad \text{for } n \in \mathbb{Z}^+. \tag{565}$$

We claim Newton's method converges at least quadratically when $f'(p) \neq 0$. Using the above results, observe we have the equality

$$|p - p_{n+1}| = \left|\frac{f''(\xi_n)}{2f'(p_n)}\right| |p - p_n|^2 . \tag{566}$$

Since ξ_n is between p_n and p and $p_n \longrightarrow p$, we deduce $\xi_n \longrightarrow p$ as well. Whence

$$\lim_{n\to\infty} \frac{|p - p_{n+1}|}{|p - p_n|^2} = \lim_{n\to\infty} \left|\frac{f''(\xi_n)}{2f'(p_n)}\right| = \frac{f''(p)}{2f'(p)}, \tag{567}$$

i.e., we obtain the desired quadratic convergence.

b) Yes, we expect difficulty when applying Newton's method here since 0 is root of f with multiplicity 2, i.e., $f(0) = 0$, $f'(0) = 0$, $f''(0) = 1$. So, as $p_n \longrightarrow 0$, the fractional term in the Newton iteration approaches a division of two small numbers, which eventually becomes extremely prone to introduce notable computational errors when using floating point arithmetic. Moreover, the convergence of Newton's method is only linear when 0 is a root of multiplicity 2.

One method of handling the problem of multiple roots of f is to define

$$\mu(x) = \frac{f(x)}{f'(x)}. \tag{568}$$

If p is a root of f of multiplicity m, then we can write $f(x) = (x - p)^m q(x)$ for some function $q(x)$ with $q(p) \neq 0$. This implies

$$\mu(x) = \frac{(x-p)^m q(x)}{m(x-p)^{m-1}q(x) + (x-p)q'(x)} = (x - p) \cdot \frac{q(x)}{mq(x) + (x-p)q'(x)}, \tag{569}$$

and note

$$\frac{q(p)}{mq(p) + (p-p)q'(p)} = \frac{1}{m} \neq 0. \tag{570}$$

This shows 0 is a simple zero of μ, i.e., $\mu'(0) \neq 0$. Our result in a) shows Newton's method can then be applied to μ to find 0 with quadratic convergence, yielding the iteration

$$\begin{aligned} p_{n+1} &= p_n - \frac{\mu(p_n)}{\mu'(p_n)} \\ &= p_n - \frac{f(p_n)/f'(p_n)}{[f'(p_n)^2 - f(p_n)f''(p_n)]/f'(p_n)^2} \\ &= p_n - \frac{f(p_n)f'(p_n)}{f'(p_n)^2 - f(p_n)f''(p_n)}. \end{aligned} \tag{571}$$

This completes our solution.

□

F16.05: Assume $f : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable function with a global Lipschitz constant K. Consider using Backward Euler (BE) to construct approximate solution of

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y), \quad y(0) = y_0 \quad \text{for } t \in [0, T], \tag{572}$$

using a uniform time step of size of k.

a) Derive the leading term of the local truncation error for the BE method applied to (572).

b) Derive an error bound for the approximate solution obtained with BE assuming that the implicit equations are solved exactly.

c) Derive an error bound for the approximation solution obtained with BE assuming that at each step the residual associated with the solution of the implicit equations is less than a value $\varepsilon > 0$.

d) How should ε be chosen as we converge to the solution by letting $k \longrightarrow 0$? Explain this choice.

Solution:

a) Let $\Phi(y_n)$ be the approximation of y_{n+1} obtained by applying the BE iteration to the exact solution, i.e., set $\Phi(y_n) := y_n + kf(y_{n+1})$. Then through Taylor expanding about y_n we obtain the local truncation error, which here is taken to be to be the difference between $\Phi(y_n)$ and y_{n+1},

$$\begin{aligned}
\tau_{n+1} &:= y_{n+1} - \Phi(y_n) \\
&= \left[y_n + kf(y_n) + \frac{k^2}{2} f'(y_n) f(y_n) + \mathcal{O}(k^3) \right] - [y_n + kf(y_{n+1})] \\
&= \left[y_n + kf(y_n) + \frac{k^2}{2} f'(y_n) + \mathcal{O}(k^3) \right] - \left[y_n + kf(y_n) + k^2 f'(y_n) f(y_n) + \mathcal{O}(k^3) \right] \\
&= -\frac{k^2}{2} f'(y_n) f(y_n) + \mathcal{O}(k^3).
\end{aligned} \tag{573}$$

The final line in (573) gives the leading term in the local truncation error.

b) Let $\{z_n\}_{n=1}^N$ be the iterates for the BE method. Then Taylor's theorem asserts there is ξ_n and η_n between y_n and y_{n+1} such that

$$y_n = y_{n+1} - kf(y_{n+1}) + \frac{k^2}{2} y''(\xi_n) = y_{n+1} - k \left[f(z_{n+1}) + f'(\eta_n)(y_{n+1} - z_{n+1}) \right] + \frac{k^2}{2} y''(\xi_n). \tag{574}$$

Noting $z_{n+1} := z_n + kf(z_{n+1})$, we rearrange to obtain

$$\begin{aligned} y_n - z_n &= \left(y_{n+1} - k\left[f(z_{n+1}) + f'(\eta_n)(z_{n+1} - y_{n+1}) \right] + \frac{k^2}{2} y''(\xi_n) \right) - z_{n+1} - kf(z_{n+1}) \\ &= \left(1 - kf'(\eta_n)\right)(y_{n+1} - z_{n+1}) + \frac{k^2}{2} y''(\xi_n). \end{aligned} \tag{575}$$

Letting $k < 1/K$, the error $e_n := y_n - z_n$ satisfies

$$(1 - kK)|e_{n+1}| \leq |e_n| + \frac{k^2}{2}|y''(\xi_n)| \quad \Rightarrow \quad |e_{n+1}| \leq \frac{1}{1-kK}\left(|e_n| + \frac{Mk^2}{2} \right) \tag{576}$$

where $M := \sup_{x\in[0,T]} |y''(x)| < \infty$ since f is Lipschitz. Since $e_0 = y_0 - z_0 = 0$, we obtain

$$|e_{n+1}| \leq \frac{Mk^2}{2} \sum_{i=1}^{n} \frac{1}{(1-kK)^i} = \frac{Mk^2}{2} \sum_{i=1}^{n} \left(1 + \frac{kK}{1-kK} \right)^i. \tag{577}$$

Evaluating the geometric sum gives

$$\sum_{i=1}^{n} \left(1 + \frac{kK}{1+kK} \right)^i = \frac{1 - (1 + kK/(1-kK))^{n+1}}{1 - (1 + kK/(1-kK))} = \frac{1-kK}{kK} \left[\left(1 + \frac{kK}{1-kK} \right)^{n+1} - 1 \right]. \tag{578}$$

This implies

$$|e_{n+1}| \leq \frac{Mh(1-kK)}{2K} \left[\left(1 + \frac{hK}{1-kK} \right)^{n+1} - 1 \right] \leq \frac{Mh(1-kK)}{2K} \left[\exp\left(\frac{t_{n+1}K}{1-kK} \right) - 1 \right]. \tag{579}$$

c) Assume $z_{n+1} = z_n + kf(z_{n+1}) + \varepsilon_{n+1}$ with $|\varepsilon_{n+1}| < \varepsilon$. Then the right hand side of (575) must include a term ε_{n+1}, i.e.,

$$y_n - z_n = \left(1 - kf'(\eta_n)\right)(y_{n+1} - z_{n+1}) + \frac{k^2}{2} y''(\xi_n) + \varepsilon_{n+1}. \tag{580}$$

This implies

$$|e_{n+1}| \leq \frac{1}{1-kK} \left(|e_n| + \frac{Mk^2}{2} + \varepsilon_{n+1} \right) \leq \frac{1}{1-kK} \left(|e_n| + \frac{Mk^2}{2} + \varepsilon \right), \tag{581}$$

and so the error bound in (579) becomes

$$|e_{n+1}| \leq \left(\frac{Mk^2}{2} + \varepsilon \right) \frac{(1-kK)}{kK} \left[\exp\left(\frac{t_{n+1}K}{1-kK} \right) - 1 \right]. \tag{582}$$

d) From (582), we see the residual has added a term in the error bound proportional to ε/k. This implies we need $\varepsilon = o(k)$ as $k \longrightarrow 0$ to obtain the desired limit $|e_{n+1}| \longrightarrow 0$.

□

F16.06: Consider the initial value problem

$$u_t = u_{xx} + 2bu_{xy} + c^2 u_{xx} \tag{583}$$

to be solved for $t > 0$, $x, y \in [0, 2\pi]$, with $u(x, y, t)$ periodic in x and y with period 2π.

a) For what real values of b and c is the initial value problem with smooth, periodic initial data $u(x, y, 0) = u_0(x, y)$ well posed?

b) Write a stable, convergent finite difference equation for this problem. Justify your answers.

Solution:

a) We first rewrite our PDE as

$$\left(\partial_t - \partial_{xx} - 2b\partial_{xy} - c^2\partial_{xx}\right) u = 0 \tag{584}$$

so that we see the corresponding symbol $p(s, w_x, w_y)$ is

$$p(s, w_x, w_y) = s - (iw_x)^2 - 2b(iw_x)(iw_y) - c^2(iw_x)^2 = s + (1 + c^2)w_x^2 + 2bw_xw_y. \tag{585}$$

This implies the root $\overline{s}$ of the symbol is

$$\overline{s} = \underbrace{-(1 + c^2)w_x}_{\leq 0} + 2bw_xw_y. \tag{586}$$

The first term is nonnegative and the second term can take on any real value when $b \neq 0$. Recall the necessary and sufficient condition for the problem to be well posed is that there is $M \in \mathbb{R}$ such that

$$\text{Re}(\overline{s}) \leq M \tag{587}$$

for all values of w_x and w_y. This holds precisely when $b = 0$ and c is any real number, and in this case (587) holds with $M = 0$.

b) When $b = 0$, the initial value problem reduces to the one dimensional heat equation $u_t = (1 + c^2)u_{xx}$.

Thus propose using the forward-time central-space scheme, i.e.,

$$\frac{v_{m,\ell}^{n+1} - v_{m,\ell}^{n}}{k} = b\left(\frac{v_{m+1,\ell}^{n} - 2v_{m,\ell}^{n} + v_{m-1,\ell}^{n}}{h^2}\right). \tag{588}$$

By the Lax equivalence theorem, to verify convergence, it suffices to show the scheme is consistent and stable. Using a Taylor expansion, we see

$$v_{m,\ell}^{n+1} = v_{m,\ell}^{n} + k(v_{m,\ell}^{n})_t + \frac{k^2}{2}(v_{m,\ell}^{n})_{tt} + \mathcal{O}(k^3) \quad \Rightarrow \quad \frac{v_{m,\ell}^{n+1} - v_{m,\ell}^{n}}{k} = (v_{m,\ell}^{n})_t + \frac{k}{2}(v_{m,\ell}^{n})_{tt} + \mathcal{O}(k^2). \tag{589}$$

Then observe

$$v_{m\pm1,\ell}^{n} = v_{m,\ell}^{n} + h(v_{m,\ell}^{n})_x + \frac{h^2}{2}(v_{m,\ell}^{n})_{xx} \pm \frac{h^3}{6}(v_{m,\ell}^{n})_{xxx} + \frac{h^4}{24}(v_{m,\ell}^{n})_{xxxx} + \mathcal{O}(h^5). \tag{590}$$

Adding the expansions for $v_{m+1,\ell}^{n}$ and $v_{m-1,\ell}^{n}$, subtracting $2v_{m,\ell}^{n}$, and then dividing by h^2, we deduce

$$\frac{v_{m+1,\ell}^{n} - 2v_{m,\ell}^{n} + v_{m-1,\ell}^{n}}{h^2} = (v_{m,\ell}^{n})_{xx} + \frac{h^2}{12}(v_{m,\ell}^{n})_{xxxx} + \mathcal{O}(h^3). \tag{591}$$

Combining (589) and (591), we deduce the discrete differential operator $P_{k,h}$ satisfies

$$P_{k,h}v_{m,\ell}^{n} = (v_{m,\ell}^{n})_t + \mathcal{O}(k) + (v_{m,\ell}^{n})_{xx} + \mathcal{O}(h^2). \tag{592}$$

This implies $(P - P_{k,h})u = \mathcal{O}(k + h^2) \longrightarrow 0$ as $k, h \longrightarrow 0$. So, the method is consistent.

All that remains is to verify stability. We proceed using Von Neumann analysis, replacing each v_m^n in the scheme by $g^n e^{im\theta}$. Doing this, we obtain

$$\frac{g-1}{k} = b\left(\frac{e^{i\theta} - 2 + e^{-i\theta}}{h^2}\right), \tag{593}$$

which implies

$$g = 1 + \frac{kb}{h^2}\left(e^{i\theta/2} - e^{-i\theta/2}\right)^2 = 1 + \frac{kb}{h^2}\left(2i\sin(\theta/2)\right)^2 = 1 - \underbrace{\frac{4kb}{h^2}\sin^2(\theta/2)}_{\geq 0}. \tag{594}$$

This shows $g \leq 1$. To obtain stability, we need $|g| \leq 1$, i.e.,

$$-1 \leq g = 1 - \frac{4kb}{h^2}\sin^2(\theta/2) \quad \Rightarrow \quad \frac{kb}{h^2}\sin^2(\theta/2) \leq \frac{1}{2}. \tag{595}$$

Whence we obtain stability when $kb/h^2 \leq 1/2$. Thus the scheme is stable and converges when this condition is met.

□

F16.07: Consider the initial value problem

$$u_t + u^4 u_x = \varepsilon u_{xx} \tag{596}$$

to be solved for $0 \le x \le 2\pi$ and $t > 0$ with $\varepsilon > 0$, u periodic in x with period 2π, and smooth periodic initial data $u(x,0) = u_0(x)$. Write a stable convergent difference approximation that remains convergent even as $\varepsilon \longrightarrow 0$. Justify your answers.

Solution:

Define $f : \mathbb{R} \to \mathbb{R}$ by $f(u) := u^5/5$ so that $u_t + f(u)_x = \varepsilon u_{xx}$. Then we propose using the scheme

$$\frac{u_i^{n+1} - [k u_i^n + \frac{(1-k)}{2}(u_{i+1}^n + u_{i-1}^n)]}{k} + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2h} - \frac{\varepsilon}{h^2}\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right) = 0, \tag{597}$$

taking $k \in (0,1)$. Note this scheme becomes the Lax-Friedrichs scheme as $k \longrightarrow 0$.

Theorem:[4] Suppose $\{u_i^n\}$ with step sizes k and h is generated by a numerical method in conservation form with a Lipschitz continuous numerical flux, consistent with some scalar law. If the method is TV-stable, then the method will converge to a weak solution of the conservation law as $k, h \longrightarrow 0$.

We show the conditions in the theorem are met. In the limit as $\varepsilon \longrightarrow 0$, the scalar conservation law at hand is $u_t + (f(u))_x = 0$. In order to apply the theorem, we must therefore assume $\lim\limits_{k,h\to 0} \varepsilon = 0$. In particular, we keep $\varepsilon = h^2/2$. (Step 1) We first show the scheme may be expressed in conservation form. (Step 2) We show the numerical is consistent with the actual flux up to a Lipschitz constant. (Step 3) We lastly verify the scheme TV-stable.

Step 1: Our scheme may be rewritten in conservation form as

$$u_i^{n+1} = u_i^n - \lambda\left[F(u_i^n, u_{i+1}^n) - F(u_{i-1}^n, u_i^n)\right], \tag{598}$$

where the numerical flux function F is defined by

$$F(u,v) := \underbrace{\left(\frac{1-k}{2\lambda} + \frac{\varepsilon}{h^2}\right)}_{\beta}(u - v) + \frac{f(u)+f(v)}{2}. \tag{599}$$

[4] See Theorem 15.2 on page 164 of Leveque.

Define β to be the underbraced quantity. Keeping λ fixed, our assumptions on ε ensure β remains bounded as $k, h \longrightarrow 0$.

Step 2: To show the numerical flux F is Lipschitz continuous, we must show there is $L > 0$ such that

$$|F(u,v) - f(w)| \leq L \cdot \max\{|u-w|, |v-w|\} \tag{600}$$

for all u, v with $|u-w|$ and $|v-w|$ sufficiently small. Recall we assume the grid functions are bounded. This implies the support of f is closed and bounded and, thus, compact. Since f is continuously differentiable on a compact set, it is Lipschitz continuous. Then

$$\begin{aligned} |F(u,v) - f(w)| &= \left|\beta(u-v) + \frac{1}{2}\left(f(u) + f(v)\right) - f(w)\right| \\ &= \left|\beta\left[(u-w) + (w-v)\right] + \frac{1}{2}\left[(f(u) - f(w)) + (f(v) - f(w))\right]\right| \\ &\leq (2\beta + \text{lip}(f)) \cdot \max\{|u-w|, |v-w|\}, \end{aligned} \tag{601}$$

and we see (523) holds by taking $K = (2\beta + \text{lip}(f))$.

Step 3: In this step we show the method is TV-stable. We first show it is ℓ_1 contracting. Let $\{u_i^n\}$ and $\{v_i^n\}$ be two collections of grid functions and define $e_i^n = u_i^n - v_i^n$. Also note our scheme may be written in the form

$$u_i^{n+1} = (k - 2\varepsilon\mu)\, u_i^n + \frac{(1-k)}{2}\left(u_{i+1}^n + u_{i-1}^n\right) + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2h} + \varepsilon\mu\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right). \tag{602}$$

Note the first term is cancels since $k - 2\varepsilon\mu = k - 2(h^2/2)(k/h^2) = 0$. This implies

$$\begin{aligned} e_i^{n+1} &= \frac{1-k}{2}\left(e_{i-1}^n + e_{i+1}^n\right) + \frac{\lambda}{2}\left[\left(f(u_{i+1}^n) - f(v_{i+1}^n)\right) - \left(f(u_{i-1}^n) - f(v_{i-1}^n)\right)\right] + \varepsilon\mu\left[e_{i+1}^n - 2e_i^n + e_{i-1}^n\right] \\ &= \frac{1-k}{2}\left(e_{i-1}^n + e_{i+1}^n\right) + \frac{\lambda}{2}\left[f'(\theta_{i+1}^n)e_{i+1}^n - f'(\theta_{i-1}^n)e_{i-1}^n\right] + \varepsilon\mu\left[e_{i+1}^n - 2e_i^n + e_{i-1}^n\right] \\ &= \left(\frac{(1-k)}{2} + \frac{\lambda}{2}f'(\theta_{i+1}) + \varepsilon\mu\right)e_{i+1}^n + \left(\frac{(1-k)}{2} - \frac{\lambda}{2}f'(\theta_{i-1}) + \varepsilon\mu\right)e_{i-1}^n. \end{aligned} \tag{603}$$

Here $\theta_{i\pm1}$ exists, by Taylor's theorem, between $u_{i\pm1}^n$ and $v_{i\pm1}^n$, respectively. We assume λ is fixed so the CFL conditions holds, i.e.,

$$\left|\lambda f'(u)\right| \leq 1 \quad \forall \ \min_j(u_j^n, v_j^n) \leq u \leq \max_j(u_j^n, v_j^n). \tag{604}$$

In particular, $|\lambda f'(\theta_{i\pm 1})| \leq 1$, and so

$$\frac{1-k}{2} \pm \frac{1}{2}\left(\lambda f'(\theta_{j\pm 1})\right) + \varepsilon\mu \geq \frac{1-k}{2} - \frac{1}{2} + \varepsilon\mu = -\frac{k}{2} + \frac{h^2}{2}\cdot\frac{k}{h^2} = 0. \tag{605}$$

This shows the coefficients of e^n_{i+1} and e^n_{i-1} are nonnegative. Whence

$$\begin{aligned}\|e^{n+1}\|_1 &= h\sum_i |e^{n+1}_i| \\ &\leq \left(\frac{1-k}{2} + \frac{\lambda}{2}f'(\theta_{i+1}) + \varepsilon\mu\right) h\sum_i |e^n_{i+1}| + \left(\frac{1-k}{2} - \frac{\lambda}{2}f'(\theta_{i-1}) + \varepsilon\mu\right) h\sum_i |e^n_{i-1}| \\ &= \left(\frac{1-k}{2} + \frac{\lambda}{2}f'(\theta_i) + \varepsilon\mu\right) h\sum_i |e^n_i| + \left(\frac{1-k}{2} - \frac{\lambda}{2}f'(\theta_i) + \varepsilon\mu\right) h\sum_i |e^n_i| \\ &= (1-k+2\varepsilon\mu)\, h\sum_i |e^n_i| \\ &= h\sum_i |e^n_i| \\ &= \|e^n\|_1.\end{aligned} \tag{606}$$

Note $\sum_i |e^n_i| = \sum_i |e^n_{i+j}|$ for all j since the grid functions are periodic, which is how the third line follows from the second. This shows the method is ℓ_1 contracting. Choosing $v^n_i = u^n_{i+1}$, (606) implies the method is TV diminishing (TVD) since

$$\text{TV}(u^{n+1}) = \frac{1}{h}\|u^{n+1}_{i+1} - u^{n+1}_i\|_1 \leq \frac{1}{h}\|u^n_{i+1} - u^n_i\|_1 = \text{TV}(u^n). \tag{607}$$

Then, by induction, we conclude $\text{TV}(u^n) \leq \text{TV}(u^0)$ for all $n \in \mathbb{Z}^+$ and the scheme is TV-stable.

By the above, we conclude the sequence converges to a weak solution of the conservation law. And, the above shows the method on a sequence of grids with $\varepsilon \longrightarrow 0$ converges to the "vanishing viscosity" solution that might be used to define the physically relevant weak solution to the conservation law. □

F16.08: Consider the biharmonic problem in a two-dimensional domain Ω with sufficiently smooth boundary,

$$\begin{cases} \Delta\Delta u = f & \text{in } \Omega, \\ u = \dfrac{\partial u}{\partial n} = 0 & \text{on } \Gamma = \partial\Omega, \end{cases} \tag{608}$$

where $\partial/\partial n$ denotes differentiation in the outward normal direction to the boundary Γ.

a) Formally show using a Green's formula that, for any $u \in H^2(\Omega)$ satisfying the above boundary conditions, we have

$$\int_\Omega |\Delta u|^2 \, \mathrm{d}x\mathrm{d}y = \int_\Omega (u_{xx})^2 + (u_{yy})^2 + (u_{xy})^2 + (u_{yx})^2 \, \mathrm{d}x\mathrm{d}y. \tag{609}$$

b) Derive a weak variational formulation of the biharmonic problem and show that this has a unique solution u in an appropriate space of functions that you will specify. Assume $f \in L^2(\Omega)$. Justify your answers.

c) Briefly describe a finite element approximation of the problem using P_5 elements and a set of basis functions such that the corresponding linear system is parse. Show that this linear system has a unique solution.

d) Assume convexity and sufficient regularity of the domain Ω. State a standard error estimate for the approximation.

Solution:

a) Note the right hand side of (609) is equal to $\|D^2 u\|^2_{L^2(\Omega)}$. We verify the desired relation by integrating by parts twice via Green's formula and noting the boundary terms cancel each time. Indeed,

$$\int_\Omega |\Delta u|^2 = \int_\Omega \left(\sum_i u_{x_i x_i}\right)\left(\sum_j u_{x_j x_j}\right) = \sum_{i,j} \int_\Omega u_{x_i x_i} u_{x_j x_j} = \sum_{i,j} - \int_\Omega u_{x_i x_i x_j} u_{x_j} + \int_\Gamma u_{x_i x_i} u_{x_j} n^j, \tag{610}$$

and our boundary conditions yield

$$\sum_{i,j} \int_\Gamma u_{x_i x_i} u_{x_j} n^j = \sum_i \int_\Gamma u_{x_i x_i} \frac{\partial u}{\partial n} = \sum_i \int_\Gamma u_{x_i x_i} 0 = 0. \tag{611}$$

This implies

$$\int_\Omega |\Delta u|^2 = -\sum_{i,j}\int_\Omega u_{x_ix_ix_j}u_{x_j} = \sum_{i,j}\int_\Omega u_{x_ix_j}^2 - \int_\Gamma u_{x_ix_j}u_{x_j}n^i. \tag{612}$$

Integrating again, we see

$$\sum_{i,j}\int_\Gamma u_{x_ix_j}u_{x_j}n^i = \sum_{i,j}\int_\Gamma \partial_{x_i}\left(\frac{u_j^2}{2}\right)n^i = \int_\Gamma \frac{\partial}{\partial n}\left(\frac{|Du|^2}{2}\right) = \int_\Gamma |Du|\frac{\partial |Du|}{\partial n} = \int_\Gamma n\cdot Du = 0. \tag{613}$$

The final equality holds since $\partial u/\partial n := n\cdot Du$. Thus

$$\int_\Omega |\Delta u|^2 = \sum_{i,j}\int_\Omega u_{x_ix_j}^2 - \int_\Gamma u_{x_ix_j}u_{x_j}n^i = \sum_{i,j}\int_\Omega u_{x_ix_j}^2 = \|D^2u\|_{L^2(\Omega)}^2, \tag{614}$$

as desired.

b) Set $H := H_0^2(\Omega)$ and let $v\in H$. Then $v=0$ and $Dv=0$ on Γ, and for a solution u

$$\begin{aligned}\int_\Omega fv &= \int_\Omega \Delta\Delta uv\\ &= -\int_\Omega D(\Delta u)\cdot Dv + \cancelto{0}{\int_\Gamma \frac{\partial(\Delta u)}{\partial n}v}\\ &= -\int_\Omega D(\Delta u)\cdot Dv\\ &= \int_\Omega \Delta u\Delta v - \cancelto{0}{\int_\Gamma \Delta u\frac{\partial v}{\partial n}}\\ &= \int_\Omega \Delta u\Delta v.\end{aligned} \tag{615}$$

The first equality holds since u is a solution, the second through integration by parts, the third since v has zero trace, the fourth by again using integration by parts, and the final equality holds since Dv has zero trace. Then the weak variational formulation of the problem is

$$\text{Find } u\in H_0^2(\Omega) \text{ such that } B(u,v)=\ell(v)\ \ \forall\ v\in H_0^2(\Omega), \tag{616}$$

where

$$B(u,v) := \int_\Omega \Delta u\Delta v \quad\text{and}\quad \ell(v) := \int_\Omega fv. \tag{617}$$

We now show this has a unique solution by verifying each of the conditions for the Lax-Milgram theorem hold. We must show B is coercive and bounded and that ℓ is bounded. Indeed,

$$|B(u,v)| = \|\Delta u \Delta v\|_{L^1(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}\|\Delta v\|_{L^2(\Omega)} = \|D^2 u\|_{L^2(\Omega)}\|D^2 v\|_{L^2(\Omega)} \leq \|u\|_H \|v\|_H. \tag{618}$$

The first inequality is Hölder's inequality. The following equality was shown in a), and the final inequality holds by the definition of $\|\cdot\|_H$. Hence B is bounded. Poincaré's inequality implies there is $C > 0$, depending on Ω, such that

$$\|v\|_{L^2(\Omega)} \leq C\|Dv\|_{L^2(\Omega)} \quad \forall\ v \in H^1_0(\Omega). \tag{619}$$

This implies $\|u\|_{L^2(\Omega)} \leq C\|Du\|_{L^2(\Omega)}$ and, because $u_{x_i} \in H^1_0(\Omega)$ for each i,

$$\|Du\|^2_{L^2(\Omega)} = \sum_i \|u_{x_i}\|^2_{L^2(\Omega)} \leq \sum_i C\|Du_{x_i}\|^2_{L^2(\Omega)} = C\sum_{i,j}\|u_{x_i x_j}\|^2_{L^2(\Omega)} = C\|D^2 u\|^2_{L^2(\Omega)}. \tag{620}$$

Therefore $\|u\|^2_{L^2(\Omega)} \leq C^2\|D^2 u\|^2_{L^2(\Omega)}$ and

$$\begin{aligned} B(u,u) = \int_\Omega |\Delta u|^2 &= \|D^2 u\|^2_{L^2(\Omega)} \\ &\geq \frac{1}{C^2+C+1}\left(\|u\|^2_{L^2(\Omega)} + \|Du\|^2_{L^2(\Omega)} + \|D^2 u\|^2_{L^2(\Omega)}\right) \\ &= \frac{1}{C^2+C+1}\|u\|^2_H. \end{aligned} \tag{621}$$

This shows B is coercive. Lastly, ℓ is bounded since $f \in L^2(\Omega)$ and

$$|\ell(v)| \leq \|fv\|_{L^1(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_H. \tag{622}$$

c) Let $\mathcal{T}_h$ be a triangulation of the domain Ω with node $\{N_i\}$. Define H_h to be the Hilbert space

$$H_h := \{v \in H\ :\ v|_K \in P_5(K)\ \forall\ K \in \mathcal{T}_h\}. \tag{623}$$

Since $\dim(P_5(K)) = 21$, it follows that each $v \in H_h$ is uniquely determined by

$$D^\alpha v(a^i) \quad \text{for } i = 1,2,3,\ |\alpha| \leq 2, \quad \text{and} \quad \frac{\partial v}{\partial n}(a^{ij}) \quad \text{for } i,j = 1,2,3, \text{ with } i < j, \tag{624}$$

where K has vertices a^i for $i = 1, 2, 3$ and a^{ij} is the midpoint on the side $a^i a^j$ for $i, j = 1, 2, 3$ with $i < j$. Let $\{\phi_j\}_{j=1}^P$ be a basis for H_h using these constraints so the ϕ_j each have minimal compact support. Then the corresponding weak variational problem becomes

$$\text{Find } u_h \in H_h \text{ such that } B(u_h, v) = \ell(v) \ \forall \ v \in H_h. \tag{625}$$

By the bilinearity of B, we see this is equivalent to the problem

$$\text{Find } u_h \in H_h \text{ such that } B(u_h, \phi_j) = \ell(\phi_j) \text{ for } j = 1, \dots, P. \tag{626}$$

Now consider a solution $u_h = \sum_{i=1}^M \xi_i \phi_i$ where $\xi \in \mathbb{R}^P$ is constant. Then the bilinearity of B implies

$$\ell(\phi_j) = B(u_h, \phi_j) = B\left(\sum_{i=1}^P \xi_i \phi_i, \phi_j\right) = \sum_{i=1}^M \xi_i B(\phi_i, \phi_j) \quad j = 1, \dots, P. \tag{627}$$

This can be written as the matrix equation $A\xi = b$ where $A_{ij} := B(\phi_i, \phi_j)$ and $b_i := \ell(\phi_i)$. This shows (626) is equivalent to finding ξ such that $A\xi = b$. Note A is symmetric and, for each $\alpha \in \mathbb{R}^P$ and $v = \sum_j \alpha_j \phi_j \in H_h$,

$$\alpha \cdot A\alpha = \sum_{i,j=1}^M \alpha_i A_{i,j} \alpha_j = B\left(\sum_{i=1}^M \alpha_i \phi_i, \ \sum_{j=1}^M \alpha_j \phi_j\right) = B(v, v) > 0, \tag{628}$$

where the final inequality holds since B is coercive. Because A is symmetric and positive definite, the system $A\xi = b$ has a unique solution. Moreover, A is sparse since the basis functions have small compact support, i.e., because $B(\phi_i, \phi_j) = 0$ for all i and j except the few i and j for which the compact support of ϕ_i and ϕ_j overlap.

d) We now derive an error estimate for the problem. Let h_K be the diameter of $K \in \mathcal{T}_h$ and ρ_K be the diameter of the largest circle inscribable in K. If $h = \max_{K \in \mathcal{T}_h} h_K$ and there is $\beta > 0$ such that

$$\frac{\rho_K}{h_K} \geq \beta \ \ \forall \ K \in \mathcal{T}_h, \tag{629}$$

then we can apply the following error estimate. Note this assumption above is that the triangles do not become arbitrarily thin. Let πu be the P_5 polynomial interpolation of u. Then πu agrees with u up to fifth order and so there is C_u, dependent on u and independent of h, such that $\|u - \pi u\|_H \leq C_u h^6$. However, because u_h is a solution to the discrete variational problem, we know $\|u - u_h\|_H \leq \|u - v\|_H$ for all $v \in H_h$. Since $\pi u \in H_h$, we thus obtain the error estimate $\|u - u_h\|_H \leq C_u h^6$ and are done.

$\square$

Spring 2017

S17.01:

a) Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function with an isolated minimum at a point b and both $f(b)$ and $f''(b)$ are $\mathcal{O}(1)$. Let x_n be an approximation to the point b and assume one has the bound $|f(x_n) - f(b)| < \varepsilon$. Derive the leading term for an estimate of $|x_n - b|$ in terms of ε.

b) Suppose that, due to numerical errors in evaluating the function f, we make the assumption $|f(x) - f(b)| > \beta$ where $\beta = 10^{-14}$ for all numerically representable values x. How close an we expect to be able to find an approximation x_n to the value b, the point where f obtains its minimum? Explain.

Solution:

a) The notation used in this question is messed up for the big-oh definition.... I don't know what it means... I will proceed with my guess of an interpretation. For each x, Taylor's theorem asserts there is ξ_x between x and b such that

$$f(x) = f(b) + f'(b)(x-b) + \frac{f''(b)}{2}(x-b)^2 + \frac{f'''(\xi_x)}{6}(x-b)^3, \tag{630}$$

which implies

$$|f(x) - f(b)| = \left| \frac{f''(b)}{2}(x-b)^2 + \frac{f'''(\xi_x)}{6}(x-b)^3 \right|. \tag{631}$$

To derive a leading term for an estimate of $|x_n - b|$, assume the second term on the right hand side is small in comparison to the first term on the right hand side. This gives the estimate

$$|x_n - b| \approx \sqrt{\frac{2|f(x_n) - f(b)|}{f''(b)}} < \sqrt{\frac{2\varepsilon}{f''(b)}}. \tag{632}$$

b) This part follows in similar fashion to above, but instead swapping the direction of the inequality. Namely, we obtain

$$|x_n - b| \approx \sqrt{\frac{2|f(x_n) - f(b)|}{f''(b)}} > \sqrt{\frac{2\beta}{f''(b)}}. \tag{633}$$

□

S17.02: Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function and $\overline{x}$ be a fixed point of f. Consider determining the fixed point $\overline{x}$ by evolving the differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x) - x \tag{634}$$

to steady state.

a) Assume the Forward Euler (FE) method is used with time step size k to approximate the solution to the differential equation (634). Derive a relation between $e_{n+1} = |x^{n+1} - \overline{x}|$ and $e_n = |x^n - \overline{x}|$ where x^n is the approximation to $x(nk)$.

b) Assume $f'(x) < \kappa < 0$ for all x. Determine the restriction that must be imposed to ensure $x^n \longrightarrow \overline{x}$.

Solution:

a) Observe

$$\begin{aligned} x^{n+1} - \overline{x} &= \left[x^n + k(x^n)'\right] - \overline{x} \\ &= \left[x^n + k(f(x^n) - x^n)\right] - \overline{x} \\ &= \left[x^n + k\left(f(\overline{x}) + f'(\xi_n)(x^n - \overline{x}) - x^n\right)\right] - \overline{x} \\ &= \left[x^n + k\left(\overline{x} + f'(\xi_n)(x^n - \overline{x}) - x^n\right)\right] - \overline{x} \\ &= \left[1 + k\left(f'(\xi_n) - 1\right)\right](x^n - \overline{x}). \end{aligned} \tag{635}$$

The first equality holds since we are using the FE method. The second holds by substituting in the differential equation. The third holds by Taylor's theorem where ξ_n is between $\overline{x}$ and x^n. The fourth equality holds since $\overline{x} = f(\overline{x})$. The final equality is obtained by collecting common terms. This implies the error satisfies

$$e_{n+1} = \left|1 + k(f'(\xi_n) - 1)\right| e_n. \tag{636}$$

b) Since e_{n+1} is a scalar multiple of e_n for each step n, it suffices to find k such that $|1+k(f'(\xi_n)-1)| \leq \alpha$ for every time step n where $\alpha \in (0,1)$. This way

$$0 \leq e_{n+1} \leq \alpha e_n \leq \cdots \leq \alpha^n e_0 \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty. \tag{637}$$

First note

$$1 + k(f'(\xi_n) - 1) < 1 + k(\kappa - 1) = 1 - k(|\kappa| + 1) < 1 \tag{638}$$

for all $k > 0$. Now set $m := \sup_{x\in\mathbb{R}} |f'(x)|$ and assume $m < \infty$. Then

$$1 - k(|f'(\xi_n) + 1) \geq 1 - k(m+1) > -1 \qquad \Rightarrow \qquad k < \frac{2}{m+1}. \tag{639}$$

Thus we need $k < 2/(m+1)$.

□

S17.03: Consider the boundary-value problem

$$\begin{cases} y''(x) = q(x)y(x) + r(x) \quad \text{in } [a,b], \\ y(a) = \alpha, \quad y(b) = \beta. \end{cases} \tag{640}$$

a) Using standard centered-difference approximations for $y''(x_i)$, derive the system of equations whose solution can be used to approximate the solution to this problem.

b) What is the expected order of accuracy of your approximation? Explain.

Solution:

a) Pick $n \in \mathbb{Z}^+$ and set $x_i = a + ih$ where $h = (b-a)/(n+1)$. Put $y_i = y(x_i)$, $q_i = q(x_i)$, and $r(x_i) = r_i$ for $i = 0, \dots, n+1$. We are given $y_0 = \alpha$ and $y_{n+1} = \beta$. We must solve for $y_1, \dots, y_n$. Taylor expanding about x_i, we find

$$y_{i\pm 1} = y_i \pm h(y_i)' + \frac{h^2}{2}(y_i)'' \pm \frac{h^3}{6}(y_i)''' + \frac{h^4}{24}(y_i)'''' + \mathcal{O}(h^5). \tag{641}$$

Adding the expansions for y_{i+1} and y_{i-1}, subtracting $2y_i$, and then dividing by h^2, we deduce

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = (y_i)'' + \frac{h^2}{12}(y_i)'''' + \mathcal{O}(h^3). \tag{642}$$

This implies

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = q_i y_i + r_i + \mathcal{O}(h^2) \quad \text{for } i = 1, \dots, n, \tag{643}$$

which gives rise to the linear system

$$\left(\frac{1}{h^2} \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} - \begin{bmatrix} q_1 & & & & \\ & q_2 & & & \\ & & \ddots & & \\ & & & q_{n-1} & \\ & & & & q_n \end{bmatrix} \right) \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} r_1 - \alpha/h^2 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n - \beta/h^2 \end{bmatrix}. \tag{644}$$

b) As shown in (643), our difference approximation gives an $\mathcal{O}(h^2)$ approximation to the actual boundary-value problem. In other words, this shows the expected order of accuracy is $\mathcal{O}(1/n^2)$.

□

S17.04: For the midpoint rule I_m and the trapezoidal rule I_T asymptotic expansions for the error are given by

$$\begin{aligned}\int_0^h f(s)\ \mathrm{d}s - I_m &= \frac{1}{24}f''(0)h^3 + \frac{1}{48}f'''(0)h^4 + \frac{11}{1920}f^{(4)}(0)h^5 + \frac{13}{11520}f^{(5)}(0)h^6 ...\\ \int_0^h f(s)\ \mathrm{d}s - I_T &= -\frac{1}{12}f''(0)h^3 - \frac{1}{24}f'''(0)h^4 - \frac{1}{80}f^{(4)}(0)h^5 - \frac{1}{360}f^{(5)}(0)h^6 ...\end{aligned} \tag{645}$$

a) Derive the coefficients of the approximation

$$\int_0^h f(s)\ \mathrm{d}s = a_0 f(0) + a_1 f\left(\frac{h}{2}\right) + a_2 f(h) \tag{646}$$

that results in an approximation of greater accuracy than either the midpoint or the trapezoidal rule.

b) What is the order of the resulting approximation?

Solution:

a) In order to obtain a greater accuracy, we must cancel the $\mathcal{O}(h^3)$ terms found in the error expansions for I_m and I_T. This can be done by forming a linear combination of I_m and I_T. Indeed, adding $2I_m + I_T$ would cancel the $\mathcal{O}(h^3)$ terms, and then dividing by 3 would normalize the result appropriately. This gives rise to the method

$$\frac{2I_m + I_T}{3} = \frac{2}{3}\left[hf(h/2)\right] + \frac{1}{3}\left[\frac{h}{2}\left(f(0) + f(h)\right)\right] = \underbrace{\frac{h}{6}}_{a_0} f(0) + \underbrace{\frac{2h}{3}}_{a_1} f(h/2) + \underbrace{\frac{h}{6}}_{a_2} f(h). \tag{647}$$

The terms a_0, a_1, a_2 are given by the underbraced quantities.

b) Applying the method from above, we deduce

$$\begin{aligned}\int_0^h f(s)\ \mathrm{d}s - \frac{2I_m + I_T}{3} &= \left(\frac{2}{24} - \frac{1}{12}\right) f''(0)h^3 + \left(\frac{2}{48} - \frac{1}{24}\right) f'''(0)h^4 + \left(\frac{22}{1920} - \frac{1}{80}\right) f^{(4)}(0)h^5 + \mathcal{O}(h^6)\\ &= \underbrace{\left(\frac{11}{860} - \frac{1}{80}\right)}_{\neq 0} f^{(4)}(0)h^5 + \mathcal{O}(h^6).\end{aligned} \tag{648}$$

This implies the resulting approximation is $\mathcal{O}(h^5)$.

□

S17.05: Consider the two-step method

$$w_{n+1} = \frac{1}{2}\left(w_n + w_{n-1}\right) + \frac{h}{4}\left[4w'_{n+1} - w'_n + 3w'_{n-1}\right] \tag{649}$$

for the initial value problem $y' = f(t,y)$, $y(0) = y_0$ with $y'_n := f(t_n, y_n)$.

a) Derive the leading order of the truncation error for this method.

b) Is this a convergent method, and if so, what is the global order of convergence? Explain and give a derivation of the conditions that support your conclusion.

c) To use this method you need two starting values, w_0 and w_1. Give a procedure for determining the required starting value w_1 so that the global order of convergence is the same as that determined in b). Explain.

d) Derive the conditions that determine the region of absolute stability for this method.

Solution:

a) Using a Taylor expansion for y_{n-1} centered about y_n, we deduce

$$\frac{1}{2}\left(y_n + y_{n-1}\right) = \frac{1}{2}\left(2y_n - hy'_n + \frac{h^2}{2}y''_n - \frac{h^3}{6}y'''_n + \mathcal{O}(h^4)\right) = y_n - \frac{h}{2}y'_n + \frac{h^2}{4}y''_n - \frac{h^3}{12}y'''_n + \mathcal{O}(h^4). \tag{650}$$

In similar fashion, Taylor expanding about y'_n gives

$$\begin{aligned} 4y'_{n+1} - y'_n + 3y'_{n-1} &= 4\left(y'_n + hy''_n + \frac{h^2}{2}y'''_n + \mathcal{O}(h^3)\right) - y'_n + 3\left(y'_n - hy''_n + \frac{h^2}{2}y'''_n + \mathcal{O}(h^3)\right) \\ &= 6y'_n + hy''_n + \frac{7h^2}{2}y'''_n + \mathcal{O}(h^3). \end{aligned} \tag{651}$$

This implies

$$\frac{h}{4}\left[4y'_{n+1} - y'_n + 3y'_{n-1}\right] = \frac{3h}{2}y'_n + \frac{h^2}{4}y''_n + \frac{7h^3}{8}y'''_n + \mathcal{O}(h^4). \tag{652}$$

Adding our results in (650) and (652), we discover

$$\frac{1}{2}\left(y_n + y_{n-1}\right) + \frac{h}{4}\left[4y'_{n+1} - y'_n + 3y'_{n-1}\right] = y_n + hy'_n + \frac{h^2}{2}y''_n + \frac{19h^3}{24}y'''_n + \mathcal{O}(h^4). \tag{653}$$

If w_n, w_{n-1}, w'_{n+1}, w'_n, and w'_{n-1} agree with the exact solution, then we obtain the local truncation error τ_{n+1}, given by

$$\begin{aligned}\tau_{n+1} &:= y_{n+1} - w_{n+1} \\ &= \left[\left(y_n + hy_n' + \frac{h^2}{2}y_n'' + \frac{19h^3}{24}y_n''' + \mathcal{O}(h^4)\right) - \left(y_n + hy_n' + \frac{h^2}{2}y_n'' + \frac{h^3}{6}y_n''' + \mathcal{O}(h^4)\right)\right] \\ &= \frac{5h^3}{8}y_n''' + \mathcal{O}(h^4).\end{aligned} \tag{654}$$

b) We claim the method is consistent. Indeed,

$$\lim_{h\to 0}\left|\frac{\tau_n(h)}{h}\right| = \lim_{h\to 0}\left|\frac{5h^2}{6}y_{n-1}''' + \mathcal{O}(h^3)\right| = 0 \quad \text{for each } n, \tag{655}$$

which implies the method is consistent. The truncation error $\tau_n(h) = \mathcal{O}(h^3)$ implies the method is $\mathcal{O}(h^2)$. Recall that a consistent method is convergent if and only if it is stable. So, we will be done if we show the method is stable. We can write our method in the form

$$w_{n+1} = \sum_{j=0}^{1}\alpha_j w_{n-j} + h\sum_{j=-1}^{1}\beta_j w_{n-j}' \tag{656}$$

where $\alpha_0 = \alpha_1 = 1/2$, $\beta_{-1} = 1$, $\beta_0 = -1/4$, and $\beta_1 = 3/4$. Then the characteristic polynomial $\chi(r)$ for this method is

$$\chi(\lambda) = \lambda^2 - \sum_{j=0}^{1}\alpha_j\lambda^{1-j} = \lambda^2 - \frac{\lambda}{2} - \frac{1}{2}. \tag{657}$$

The roots of $\chi(\lambda)$ are given by

$$\lambda = \frac{1/2 \pm \sqrt{1/2^2 - 4\cdot 1\cdot(-1/2)}}{2} = \frac{1\pm 3}{4}, \tag{658}$$

i.e., the roots are $-1/2$ and 1. Since $|-1/2| < 1$ and 1 is a simple root of $\chi(\lambda)$, the difference method satisfies the root condition. The linear multistep method is stable if and only if it satisfies the root condition. Therefore it is stable and we have convergence.

c) We must use a method to compute w_1 that has local truncation error with order $\mathcal{O}(h^3)$. Taylor expanding about y_0, we find

$$y_1 = y_0 + hy_0' + \frac{h^2}{2} y_0'' + \frac{h^3}{6} y_0''' + \mathcal{O}(h^4). \tag{659}$$

So, we can set

$$w_1 = y_0 + hy_0' + \frac{h^2}{2} y_0'' \tag{660}$$

so that

$$y_1 - w_1 = \frac{h^3}{6} y_0''' + \mathcal{O}(h^4) \tag{661}$$

and the order of the approximation of w_1 agrees with the local truncation error of the linear multistep method.

d) To determine the region of absolute stability, we assume $f(t,y) = \lambda y$ for some $\lambda \in \mathbb{C}$. The region of absolute stability is then defined to be the set of all $h\lambda$ such that $w_n \longrightarrow 0$ for all initial conditions w_0. We first substitute for f in our method to find

$$w_{n+1} = \frac{1}{2}(w_n + w_{n-1}) + h\lambda w_{n+1} - \frac{h\lambda}{4} w_n + \frac{3h\lambda}{4} w_{n-1}, \tag{662}$$

which implies

$$(1 - h\lambda)\, w_{n+1} + \left(-\frac{1}{2} + \frac{h\lambda}{4}\right) w_n + \left(-\frac{1}{2} - \frac{3h\lambda}{4}\right) w_{n-1} = 0. \tag{663}$$

Then the characteristic polynomial $p(r; h\lambda)$ associated with this homogeneous difference equation is

$$p(r; h\lambda) := (1 - h\lambda)\, r^2 + \left(-\frac{1}{2} + \frac{h\lambda}{4}\right) r + \left(-\frac{1}{2} - \frac{3h\lambda}{4}\right). \tag{664}$$

The region of absolute stability is therefore

$$\{h\lambda \in \mathbb{C} \;:\; |\beta| < 1 \text{ for all } \beta \text{ such that } p(\beta; h\lambda) = 0\}. \tag{665}$$

□

S17.06: Consider the initial value problem

$$\begin{cases} u_t = 0 \\ v_t = u_x \end{cases} \tag{666}$$

to be solve for $x \in [0,1]$, $t \geq 0$ with initial and boundary conditions

$$u(x,0) = \phi(x), \quad u(1,t) = u(0,t), \quad v(x,0) = \psi(x), \quad v(1,t) = v(0,t). \tag{667}$$

a) i) Can you write a stable, convergent finite difference scheme for this problem?

ii) Explain your answer and give an example of such a scheme if one exists.

b) Consider the related system

$$\begin{cases} u_t = v_x/1000 \\ v_t = u_x \end{cases} \tag{668}$$

with the same initial and boundary conditions.

i) Can you write a stable, convergent finite difference scheme for this problem?

ii) Explain your answer and give an example of such a scheme if one exists.

Solution:

a) i) No, we can not write a stable, convergent finite difference scheme for this problem.

ii) No stable convergent finite difference scheme can be written for this problem because it is not well posed. We verify this as follows. First note the PDE can be written as

$$p_t = \begin{pmatrix} u \\ v \end{pmatrix}_t = \underbrace{\begin{pmatrix} 0 & \varepsilon \\ 1 & 0 \end{pmatrix}}_{M_\varepsilon} \begin{pmatrix} u \\ v \end{pmatrix}_x = M_\varepsilon p_x \tag{669}$$

where here $\varepsilon = 0$, $p = (u, v)$ and M_ε is the underbraced matrix. Taking the Fourier transform, we discover

$$\hat{p}_t = M_\varepsilon \hat{p}_x = iwM\hat{p} \quad \Rightarrow \quad \hat{p} = \exp(iwtM_\varepsilon)\hat{p}_0. \tag{670}$$

Then recall the necessary and sufficient condition for the system to be well posed is that for each $t \geq 0$ there is $C_t > 0$ such that

$$\| \exp(iwtM_\varepsilon)\| \leq C_t \quad \forall\ w \in \mathbb{R}. \tag{671}$$

We evaluate $\exp(iwtM)$ and show its norm is unbounded as $|w| \longrightarrow \infty$. This implies no C_t exists such that (671) holds for $t > 0$. Set

$$S := \exp(iwtM_0) = \sum_{j=0}^{\infty} \frac{(iwtM)^j}{j!} = I + iwtM = \begin{pmatrix} 1 & 0 \\ iwt & 1 \end{pmatrix}, \tag{672}$$

noting $M^j = 0$ for $j > 1$. Recall $\|S\|_2 = \sqrt{\lambda_{max}(S^*S)}$. So, we compute

$$S^*S = \begin{pmatrix} 1 & -iwt \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ iwt & 1 \end{pmatrix} = \begin{pmatrix} 1+(wt)^2 & -iwt \\ iwt & 1 \end{pmatrix}. \tag{673}$$

The corresponding characteristic polynomial is

$$\chi(\lambda) = (\lambda - (1+(wt)^2))(\lambda - 1) - (wt)^2 = \lambda^2 - (2+(wt)^2)\lambda + 1. \tag{674}$$

This implies the eigenvalues of S^*S are

$$\lambda = \frac{2+(wt)^2 \pm \sqrt{(2+wt)^2 - 4}}{2} = \frac{2+(wt)^2 \pm (wt)\sqrt{2+(wt)^2}}{2}. \tag{675}$$

Thus

$$\lim_{|w|\to\infty} \|\exp(iwtM_0)\|_2 = \lim_{|w|\to\infty} \|S\|_2 = \lim_{|w|\to\infty} \sqrt{\lambda_{max}(S^*S)} > \lim_{|w|\to\infty} \sqrt{1 + \frac{(wt)^2}{2}} = \infty, \tag{676}$$

and we are done.

b) i) Yes, it is possible to write such a scheme for this problem.

ii) We claim this problem is well posed. To verify this, we find C_t such that (671) holds for all $t \geq 0$.

Here $\varepsilon = 1/1000 > 0$. Then observe the characteristic polynomial for M_ε is

$$\chi(\lambda) = \lambda^2 - \varepsilon. \tag{677}$$

This implies the eigenvalues of M_ε are $\pm\sqrt{\varepsilon}$, and so M_ε is diagonalizable. Indeed, we obtain the diagonalization

$$M = \underbrace{\left(\begin{array}{cc} -\sqrt{\varepsilon} & \sqrt{\varepsilon} \\ 1 & 1 \end{array}\right)}_{P} \underbrace{\left(\begin{array}{cc} -\sqrt{\varepsilon} & 0 \\ 0 & \sqrt{\varepsilon} \end{array}\right)}_{D} \left(\begin{array}{cc} -\sqrt{\varepsilon} & \sqrt{\varepsilon} \\ 1 & 1 \end{array}\right)^{-1} = PDP^{-1} \tag{678}$$

where P and D are the underbraced matrices. It follows from induction that $M^j = (PDP^{-1})^j = PD^jP^{-1}$. Then

$$\begin{aligned} \exp(iwtM_\varepsilon) &= \lim_{n\to\infty} \sum_{j=0}^{n} \frac{(iwtM_\varepsilon)^j}{j!} \\ &= \lim_{n\to\infty} P \sum_{j=0}^{n} \frac{(iwtD)^j}{j!} P^{-1} \\ &= P \sum_{j=0}^{\infty} \frac{(iwtD)^j}{j!} P^{-1} \\ &= P \exp(iwtD) P^{-1} \\ &= P \left(\begin{array}{cc} e^{-iwt\sqrt{\varepsilon}} & 0 \\ 0 & e^{iwt\sqrt{\varepsilon}} \end{array}\right) P^{-1}. \end{aligned} \tag{679}$$

The two norm of the diagonal matrix on the final line is unity since $|e^{ix}| = 1$ for all $x \in \mathbb{R}$. This implies

$$\|\exp(iwtM_\varepsilon)\|_2 \leq \|P\|_2 \|P^{-1}\|_2 \quad \forall\ w \in \mathbb{R}. \tag{680}$$

This shows (671) holds with $C_t := \|P\|_2\|P^{-1}\|_2$.

For a stable convergent scheme for this problem, consider the Crank-Nicolson scheme

$$\frac{p_i^{n+1} - p_i^n}{k} = M_\varepsilon \frac{(p_{i+1}^{n+1} + p_{i+1}^n) - (p_{i-1}^{n+1} + p_{i-1}^n)}{4h}, \tag{681}$$

with centered differences about $v_i^{n+1/2}$ to be order $(2,2)$ and is unconditionally stable. □

S17.07: Consider the differential equation

$$u_t = au_{xx} + 2bu_{xy} + cu_{yy} \tag{682}$$

with constants a, b, c, to be solve for $t > 0$, $x, y \in [0,1]$ with $u(x,y,0) = \phi(x,y)$ smooth and boundary conditions $u(0,s,t) = u(1,s,t) = u(s,0,t) = u(s,1,t) = 0$ for $s \in [0,1]$.

a) For what values of a, b and c would you expect good behavior of the solution?

b) Write a convergent difference approximation to this problem.

c) Justify your answers.

Solution:

a) The problem is well posed when $a, c \geq 0$ and $b^2 \leq ac$.

b) Use the Peaceman-Rachford ADI type method to generate the sequence $\{v^n\}_{n=0}^{\infty}$ defined by

$$\left(I - \frac{k}{2}a\delta_x^2\right)\left(I - \frac{k}{2}c\delta_y^2\right)v_{m,\ell}^{n+1} = \left(I + \frac{k}{2}a\delta_x^2 + kb\delta_x\delta_y\right)\left(I + \frac{k}{2}c\delta_y^2 + kb\delta_x\delta_y\right)v_{m,\ell}^n. \tag{683}$$

where δ_x is the centered difference operator and δ_x^2 is the central second difference operator, and similarly for y.

c) Taking the Fourier transform, we discover

$$\hat{u}_t = -\left(aw_1^2 + 2bw_1w_2 + cw_2^2\right)\hat{u} \quad \Rightarrow \quad \hat{u} = \exp\left(-\left(aw_1^2 + 2bw_1w_2 + cw_2^2\right)t\right)\hat{\phi}. \tag{684}$$

We say the problem is well posed if and only if for each $t \geq 0$ there is C_t such that

$$|\hat{u}(w_1, w_2, t)| \leq C_t|\hat{\phi}(w_1, w_2)| \quad \forall\ (w_1, w_2) \in \mathbb{R}^2. \tag{685}$$

Note an exponential increases without bound as its argument increases to infinity. This implies we must identify constraints on a, b, c such that

$$aw_1^2 + 2bw_1w_2 + cw_2^2 \geq 0. \tag{686}$$

Since $w_1^2, w_2^2 \geq 0$, we immediately deduce that $a, c \geq 0$. Then observe

$$0 \leq aw_1^2 + cw_2^2 = \left(\sqrt{a}w_1 \pm \sqrt{c}w_2\right)^2 \mp 2\sqrt{ac}w_1w_2. \tag{687}$$

Thus, (686) holds provided $|2bw_1w_2| \leq |2\sqrt{ac}w_1w_2|$, which implies $|b| \leq \sqrt{ac}$ and so $b^2 \leq ac$. We conclude the solution is well posed if $\boxed{a, c \geq 0 \text{ and } b^2 \leq ac.}$

We now show the chosen scheme is convergent. To do this, it suffices to show it is consistent and stable. (Step 1) We first derive the scheme to verify its consistency. (Step 2) Then we verify stability.

Step 1: Define the operators $A_1u = au_{xx}$, $A_2u = cu_{yy}$, and $A_3u = bu_{xy}$. Then our PDE becomes

$$u_t = (A_1 + A_2 + 2A_3)u. \tag{688}$$

We use the Crank-Nicolson approach to expand about $(t_{n+1/2}, x_m, y_\ell)$. Taylor expanding reveals

$$u_{m,\ell}^{n+1/2\pm1/2} = u_{m,\ell}^{n+1/2} \pm \frac{k}{2}(u_{m,\ell}^{n+1/2})_t + \frac{k^2}{8}(u_{m,\ell}^{n+1/2})_{tt} + \mathcal{O}(k^3). \tag{689}$$

Combined with our PDE, this implies

$$\frac{u_{m,\ell}^{n+1} - u_{m,\ell}^n}{k} = (A_1 + A_2 + 2A_3)\left(\frac{u_{m,\ell}^{n+1} + u_{m,\ell}^n}{2}\right) + \mathcal{O}(k^2) \tag{690}$$

since the left hand side is an $\mathcal{O}(k^2)$ approximation of $(u_{m,\ell}^{n+1/2})_t$ and the right hand side contains an $\mathcal{O}(k^2)$ approximation of $u_{m,\ell}^{n+1/2}$. We may rewrite the scheme as

$$\left(I - \frac{k}{2}A_1 - \frac{k}{2}A_2\right)u_{m,\ell}^{n+1} = \left(I + \frac{k}{2}A_1 + \frac{k}{2}A_2 + 2kA_3\right)u_{m,\ell}^n + \mathcal{O}(k^2), \tag{691}$$

noting $u_{m,\ell}^n$ is a $\mathcal{O}(k)$ approximation of $u_{m,\ell}^{n+1}$. Adding $k^2A_1A_2/4$ to each side enables us to complete the square and write

$$\begin{aligned}\left(I-\frac{k}{2}A_1\right)\left(I-\frac{k}{2}A_2\right)u_{m,\ell}^{n+1} &= \left[\left(I+\frac{k}{2}A_1\right)\left(I+\frac{k}{2}A_2\right)+2kA_3\right]u_{m,\ell}^n + \underbrace{\frac{k^2}{4}A_1A_2(u_{m,\ell}^{n+1}-u_{m,\ell}^n)}_{=\mathcal{O}(k^3)}+\mathcal{O}(k^2)\\ &= \left[\left(I+\frac{k}{2}A_1\right)\left(I+\frac{k}{2}A_2\right)+2kA_3\right]u_{m,\ell}^n+\mathcal{O}(k^2)\end{aligned} \tag{692}$$

But,

$$\left(I+\frac{k}{2}A_1+kA_3\right)\left(I+\frac{k}{2}A_2+kA_3\right)=\left(I+\frac{k}{2}A_1\right)\left(I+\frac{k}{2}A_2\right)+2kA_3+\mathcal{O}(k^2), \tag{693}$$

which implies

$$\left(I-\frac{k}{2}A_1\right)\left(I-\frac{k}{2}A_2\right)u_{m,\ell}^{n+1}=\left(I+\frac{k}{2}A_1+kA_3\right)\left(I+\frac{k}{2}A_2+kA_3\right)u_{m,\ell}^n+\mathcal{O}(k^2). \tag{694}$$

This shows the scheme is $\mathcal{O}(k)$ in time. Choosing a second order approximation for each space derivative via centered differences gives $\mathcal{O}(h_x^2+h_y^2)$. In particular, we have

$$A_1u_{m,\ell}^n=a\delta_x^2u_{m,\ell}^n+\mathcal{O}(h_x^2),\quad A_2u_{m,\ell}^n=c\delta_y^2u_{m,\ell}^n+\mathcal{O}(h_y^2),\quad A_3u_{m,\ell}^n=b\delta_x\delta_yu_{m,\ell}^n+\mathcal{O}(h_x^2h_y^2). \tag{695}$$

Whence we conclude the scheme is $\mathcal{O}(k+h_x^2+h_y^2)$ and, thus, it is consistent.

Step 2: This part looks difficult.

□

S17.08: Consider the elliptic boundary value problem[5]

$$\begin{aligned} -\Delta u + u &= f(x,y) && (x,y) \in \Omega, \\ u &= 1 && (x,y) \in \partial\Omega_1, \\ \frac{\partial u}{\partial n} + u &= x && (x,y) \in \partial\Omega_2, \end{aligned}$$

where

$$\begin{aligned} \Omega &= \left\{(x,y) \mid x^2+y^2 < 1\right\}, \\ \partial\Omega_1 &= \left\{(x,y) \mid x^2+y^2 = 1,\ x \le 0\right\}, \\ \partial\Omega_2 &= \left\{(x,y) \mid x^2+y^2 = 1,\ x > 0\right\}, \end{aligned}$$

and n denotes the exterior unit normal to $\partial\Omega$.

1. Derivate a weak variational formulation.
2. Assuming the appropriate condition on the function f, analyze the assumptions of the Lax-Milgram theorem that ensure existence and uniqueness of a weak solution.
3. Set up a piecewise-linear Galerkin finite element approximation for this problem. Show that the obtained system has a unique solution. Give a convergence estimate and quote the appropriate theorems for convergence.

Solution:

a) We set $w := u - 1$ (so $u = w + 1$) and reformulate the problem in terms of w to obtain homogeneous boundary conditions:

$$\begin{aligned} -\Delta w + w &= g && \text{in } \Omega, \\ w &= 0 && \text{on } \partial\Omega_1, \\ \frac{\partial w}{\partial \nu} + w &= h && \text{on } \partial\Omega_2, \end{aligned}$$

where $g := f - 1$ and $h := x - 1$. Let $V = \left\{v \in H^1(\Omega) \mid v|_{\partial\Omega_1} \equiv 0\right\}$ equipped with the norm $\|\cdot\|_{H^1(\Omega)}$.

[5] Credit for the solution of this problem is due to Jeffrey Hellrung who solved it for W06.07.

We determine a weak variational formulation by multiplying the differential equation by $v \in V$, applying integration by parts, and noting that $v|_{\partial\Omega_1} \equiv 0$:

$$\begin{aligned}
&(-\Delta w + w)v = fv \\
&\Rightarrow \quad \int_\Omega (-\Delta w + w)v = \int_\Omega fv \\
&\Rightarrow \quad -\int_{\partial\Omega} v\frac{\partial w}{\partial \nu} + \int_\Omega \nabla w \cdot \nabla v + \int_\Omega wv = \int_\Omega fv \\
&\Rightarrow \quad -\int_{\partial\Omega_2} v(h-w) + \int_\Omega (\nabla w \cdot \nabla v + wv) = \int_\Omega fv \\
&\Rightarrow \quad \int_\Omega (\nabla w \cdot \nabla v + wv) + \int_{\partial\Omega_2} wv = \int_\Omega fv + \int_{\partial\Omega_2} hv.
\end{aligned}$$

Let

$$\begin{aligned}
a(w,v) &= \int_\Omega (\nabla w \cdot \nabla v + wv) + \int_{\partial\Omega_2} wv \\
Lv &= \int_\Omega fv + \int_{\partial\Omega_2} hv
\end{aligned}$$

such that the weak variational formulation is to find $w \in V$ such that

$$a(w,v) = Lv \text{ for all } v \in V.$$

b) The Lax-Milgram Lemma provides sufficient conditions the bilinear form a and the linear form L must satisfy for existence and uniqueness of w:

- *a is symmetric.* Clearly $a(v_1, v_2) = a(v_2, v_1)$ for $v_1, v_2 \in V$.
- *a is continuous.* For $v_1, v_2 \in V$, by the Cauchy-Schwarz Inequality,

$$\begin{aligned}
|a(v_1,v_2)| &= \left| \int_\Omega (\nabla v_1 \cdot \nabla v_2 + v_1 v_2) + \int_{\partial\Omega_2} v_1 v_2 \right| \\
&\leq \|v_1\|_{H^1(\Omega)} \|v_2\|_{H^1(\Omega)} + \|v_1\|_{L^2(\partial\Omega_2)} \|v_2\|_{L^2(\partial\Omega_2)}.
\end{aligned}$$

But

$$\|v_i\|_{L^2(\partial\Omega_2)} \leq C \|v_i\|_{H^1(\Omega)}$$

for some $C > 0$, so, in fact,

$$|a(v_1, v_2) \leq (1 + C)\|v_1\|_{H^1(\Omega)}\|v_2\|_{H^1(\Omega)},$$

and we conclude that a is continuous.

- *a is V-elliptic.* For $v \in V$,

$$\begin{aligned} a(v,v) &= \int_\Omega \left(|\nabla v|^2 + v^2\right) + \int_{\partial\Omega_2} v^2 \\ &\geq \int_\Omega \left(|\nabla v|^2 + v^2\right) \\ &= \|v\|^2_{H^1(\Omega)}, \end{aligned}$$

  and so a is indeed V-elliptic.

- *L is continuous.* For $v \in V$, by the Cauchy-Schwarz Inequality,

$$\begin{aligned} |Lv| &= \left|\int_\Omega fv + \int_{\partial\Omega_2} hv\right| \\ &\leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|h\|_{L^2(\partial\Omega_2)}\|v\|_{L^2(\partial\Omega_2)} \\ &\leq \left(\|f\|_{L^2(\Omega)} + C\|h\|_{L^2(\partial\Omega_2)}\right)\|v\|_{H^1(\Omega)}, \end{aligned}$$

  hence L is continuous.

Therefore, we have existence and uniqueness of the solution w.

c) For the finite element approximation, we suppose some triangulation $\{K\}_h$, where h denotes the fineness of the triangulation mesh, with nodes denoted by $\{N_j\}$. Let

$$V_h = \{v \in V \mid v|_K \in P_1(K) \text{ for each } K\}.$$

The approximate variational formulation then becomes to find $w_h \in V_h$ such that $a(w_h, v) = Lv$ for all $v \in V_h$. By linearity, if $\{\phi_i\}$ is a basis for V_h, this is equivalent to finding $w_h \in V_h$ such that $a(w_h, \phi_i) = L\phi_i$ for all ϕ_i. We take ϕ_i such that $\phi_i(N_j) = \delta_{ij}$. Now we can also express $w_h = \sum_j \xi_j\phi_j$, thus obtaining the linear system

$$\sum_j \xi_j a(\phi_j, \phi_i) = L\phi_i \;\Rightarrow\; A\xi = b,$$

where the entries of the stiffness matrix are $A_{ij} = a(\phi_j, \phi_i)$ and the entries of the load vector are $b_i = L\phi_i$. If the enumeration of the N_j's is done efficiently, A will be sparse (since $a(\phi_j, \phi_i) = 0$ if $|i - j|$ is too large) and banded, allowing for efficient solving of the system. Further, A is positive definite (since a is an inner product), hence is nonsingular, so the system has a unique solution.

If w is the solution to the weak variational formulation and w_h is the solution to the approximate variational formulation, then we have the bound $\|w - w_h\|_a \leq \|w - v\|_a$ for any $v \in V_h$, where $\| \cdot \|_a$ is the norm induced by the inner product $a(\cdot, \cdot)$. In particular, we can take the linear interpolant $\pi_h w \in V_h$ of w, and we know that $\|w - \pi_h w\|_a \leq C_w h^2$ for some constant C_w (dependent on w but independent of h), from which we obtain the convergence rate estimate $\|w - w_h\|_a \leq C_w h^2$.

□

References

[1] Richard L Burden and J Douglas Faires. Numerical analysis, 2011.

[2] Patrick M. Fitzpatrick. *Advanced Calculus*, volume II. American Mathematical Society, 2nd edition, 2009.

[3] Erwin Kreyszig. *Introductory Functional Analysis.* John Wiley & Sons (Asia), 1989.

[4] Walter Rudin. *Principles of Mathematial Analysis.* McGraw Hill Education (India), 3rd edition, 2013.