Quals Revision Notes Numerics Sequence

Tyler Chen

Contents

Summer 2011, Day 1, Problem 3 4
Winter 2011, Day 1, Problem 4 5
Winter 2012, Day 1, Problem 5 6
Summer 2013, Day 1, Problem 4 7
Summer 2013, Day 1, Problem 5 8
Winter 2013, Day 2, Problem 2 9
Practice 2010, Day 1, Problem 2 10
Summer 2013, Day 2, Problem 1 11
Summer 2014, Day 1, Problem 2 12
Summer 2014, Day 2, Problem 4 14
Winter 2014, Day 1, Problem 2 15
Winter 2014, Day 1, Problem 4 16
Winter 2017, Day 2, Problem 1 17
Practice 2010, Day 2, Problem 3 18
Winter 2015, Day 2, Problem 3 20
Summer 2017, Day 1, Problem 4 21
Summer 2011, Day 2, Question 5 22
Winter 2011, Day 2, Problem 5 24
Summer 2013, Day 3, Problem 1 27
Summer 2014, Day 1, Problem 5 29
Summer 2014, Day 3, Problem 3 30
Summer 2010, Day 1, Problem 4 31
Winter 2010, Day 1, Problem 3 33
Summer 2011, Day 2, Problem 6 35
Summer 2012, Day 2, Problem 3 37
Winter 2015, Day 1, Problem 3 40
Winter 2017, Day 1, Problem 2 41
Winter 2010, Day 2, Problem 1 42
Summer 2013, Day 1, Problem 2 43
Summer 2014, Day 2, Problem 1 44
Winter 2014, Day 2, Problem 1 46
Summer 2017, Day 1, Problem 5 48
Winter 2010, Day 1, Problem 1 49
Winter 2012, Day 2, Problem 1 50
Winter 2013, Day 2, Problem 1 51
Summer 2014, Day 2, Problem 3 52
Winter 2017, Day 2, Problem 4 54
Summer 2010, Day 2, Problem 1 55
Winter 2011, Day 1, Problem 2 57
Winter 2013 Day 1, Problem 3 58
Summer 2017, Day 1, Problem 2 60
Winter 2017, Day 1, Problem 1 61
Practice 2010, Day 2, Problem 5 62
Practice 2010, Day 3, Problem 1 63
Winter 2011, Day 3, Problem 1 64
Winter 2010, Day 1, Problem 5 65
Winter 2011, Day 2, Problem 4 66
Summer 2012, Day 1, Problem 3 68
Summer 2015, Day 2, Problem 1 69
Winter 2015, Day 1, Problem 2 70
Winter 2015, Day 2, Problem 1 72

Winter 2015, Day 2, Problem 2 73
Summer 2016, Day 1, Problem 2 74
Summer 2011, Day 3, Problem 2 76
Winter 2017, Day 3, Problem 1 77
Practice 2010, Day 1, Problem 3 79
Winter 2012, Day 2, Problem 3 80
Summer 2017, Day 2, Problem 4 81
Summer 2015, Day 2, Problem 3 82
Summer 2016, Day 1. Problem 1 84
Summer 2016, Day 2, Problem 1 85
Summer 2015, Day 3, Problem 1 86
Summer 2015, Day 1, Problem 2 87
Summer 2015, Day 2, Problem 4 88
Summer 2012, Day 3, Problem 1 90

Summer 2011, Day 1, Problem 3

Let $A \in \mathbb{R}^{n\times n}$.

- Define the singular value decomposition of $A \in \mathbb{R}^{n\times n}$.
- How can you compute the singular values and right and left singular vectors by computing eigenpairs of two operators related to A?
- How is the Frobenius norm of A related to its singular values?
- Let $b \in \mathbb{R}^n$. Consider the least-squares problem

  $$\min_{x\in\mathbb{R}^n} \|b - Ax\|_2$$

  (where $\|\cdot\|_2$ is the Euclidian norm). Give an expression for all the minimizer(s) x using the singular value decomposition of A. (Study all the possible cases.)
- Numerically, if all the singular values of A are positive, would you use all of them to solve the least-squares problems? Why?

Solution

- $A = U\Sigma V^*$ where U and V are unitary and Σ is diagonal.
- You could compute the eigenvalue eigenvector pairs of AA^* or A^*A. The singular values are the square roots of the nonzero eigenvalues of these matrices.
- $\|A\|_F = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$
- This least squares problem is equivalent to computing the pre-image of the orthogonal projection of b onto the span of A.

  In particular, suppose A has rank r and $A = \hat{U}\hat{\Sigma}\hat{V}^*$ is the rank-r SVD. Then the range of $\hat{U} \in \mathbb{R}^{n\times r}$ is equal to that of A.

  Thus, the orthogonal projection of b onto the span of A is given by,

  $$\hat{U}(\hat{U}^*\hat{U})^{-1}\hat{U}^*b = \hat{U}\hat{U}^*b$$

  Any point x such that $Ax = \hat{U}\hat{U}^*b$ solves the least squares problem. Since $\hat{U}^*\hat{U}$ is the identity, we seek all solutions to,

  $$\hat{V}^*x = \hat{\Sigma}^{-1}\hat{U}^*b$$

  If A is full rank then so is $\hat{V}$. This means the minimizer of the least squares problem is,

  $$x = \hat{V}\hat{\Sigma}^{-1}\hat{U}^*b$$

  If A is not full rank then there will be multiple solutions. One such solution can be obtained by using the psuedo inverse of $\hat{V}$, $\hat{V}(\hat{V}^*\hat{V})^{-1}\hat{V}^* = \hat{V}\hat{V}^*$. Since $\hat{V}^*\hat{V}$ is the identity, this gives the minimizer as above.

  Note that adding anything in the null space of A to x will not change the residual norm. Denote the last $n-r$ columns of V by $\tilde{V}$. Then the general solution set to the least squares problem is,

  $$\{x + r : x = \hat{V}\hat{\Sigma}^{-1}\hat{U}^*b, r = \tilde{V}y, y \in \mathbb{R}^{n-r}\}$$

- Numerically singular values which are zero in exact arithmetic might not be. Some threshold can be chosen and all singular values below this threshold discarded.

Winter 2011, Day 1, Problem 4

Let $A \in \mathbb{R}^{m\times n}$ $(m \leq n)$. Define the QR factorization of $A \in \mathbb{R}^{m\times n}$. Describe the algorithm you would recommend to compute the matrices Q and R?

Solution

A QR factorization of $A \in \mathbb{R}^{m\times n}$ with $m \leq n$ is a factorization $A = QR$ where $Q \in \mathbb{R}^{m\times m}$ is unitary and $R \in \mathbb{R}^{m\times n}$ is upper triangular.

I would use the (modified) Gram–Schmidt algorithm with an additional modification to ensure that if a column of A is linearly dependent with the previous columns of A that the algorithm will correctly project this vector onto the span of the previous vectors, and that Q will not be updated.

More specifically, using $\hat{Q}$ to denote the orthonormal basis available at each step,

1. take column of A
2. if column of A depends only on columns of $\hat{Q}$ compute projection onto the span of $\hat{Q}$ and add appropriate values to R. Do not update $\hat{Q}$
   if column of A is independent of columns of $\hat{Q}$ proceed with Gram-Schmidt algorithm, removing projections onto each of the columns of $\hat{Q}$ and normalizing the remainder which is added as a new column of $\hat{Q}$.
3. go to step 1

Note that this does not suggest how to determine if the current column of A lies in the span of $\hat{Q}$ when implementing such an algorithm on a computer. In general some threshold may need to be set for what it means to be “numerically linearly independent”.

Winter 2012, Day 1, Problem 5

Let B and C be real $m \times n$ matrices. Relate the singular values and vectors of $B + iC$ to those of

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix}$$

Solution

Suppose $(B + iC) = U\Sigma V^*$ is an SVD of $B + iC$. Write $U = X + iY$ and $V = W + iZ$. Then,

$$(BW - CZ) + i(CW + BZ) = (B + iC)(W + iZ) = (B + iC)V = U\Sigma = X\Sigma + iY\Sigma$$

These give the two equations,

$$BW - CZ = X\Sigma, \qquad CW + BZ = Y\Sigma$$

Equivalently,

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} W \\ Z \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} \Sigma$$

However, we could also write,

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} Z \\ -W \end{bmatrix} = \begin{bmatrix} Y \\ -X \end{bmatrix} \Sigma$$

Putting these together gives the singular value decomposition,

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} W & Z \\ Z & -W \end{bmatrix} = \begin{bmatrix} X & Y \\ Y & -X \end{bmatrix} \begin{bmatrix} \Sigma & \\ & \Sigma \end{bmatrix}$$

In standard form (up to ordering) and using the original SVD of $B + iC$,

$$\begin{bmatrix} B & -C \\ C & B \end{bmatrix} = \begin{bmatrix} \operatorname{Re}(U) & \operatorname{Im}(U) \\ \operatorname{Im}(U) & -\operatorname{Re}(U) \end{bmatrix} \begin{bmatrix} \Sigma & \\ & \Sigma \end{bmatrix} \begin{bmatrix} \operatorname{Re}(V) & \operatorname{Im}(V) \\ \operatorname{Im}(V) & -\operatorname{Re}(V) \end{bmatrix}^*$$

Summer 2013, Day 1, Problem 4

Call a vector $y \in \mathbb{R}^n$ a palindrome if it reads the same way forwards and back; i.e., if $y_i = y_{n+1-i}$ for $i = 1, 2, \dots, n$. Any vector $x \in \mathbb{R}^n$ can be mapped to a palindrome y by defining $y_i = \frac{1}{2}(x_i + x_{n+1-i})$. This mapping defines a matrix P.

(a) Write down P for $n = 4$ and $n = 5$.
(b) Determine all the eigenvalues and a basis for each eigenspace of P for general n. (Note: Consider both odd and even n.)
(c) Does P define an orthogonal projection? Justify your answer
(d) The SVD of P can be written as $P = \sum_{i=1}^{r} \sigma_i u_i v_i^T$ where r is the rank of P. Determine r, σ_i, u_i and v_i for general n.

Solution

(a)

$$P_4 = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \qquad P_5 = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(b) Suppose n is even. It is clear that for every $i = 1, 2, \dots, n$,

$$(x^i)_j = \begin{cases} 1 & j = i \text{ or } j = n+1-i \\ 0 & \text{otherwise} \end{cases}$$

are eigenvectors with eigenvalue 1.

Clearly the rank of the matrix is $n/2$. Since there are $n/2$ such vectors, and they are clearly linearly independent (in fact orthogonal) this forms an eigenbasis for the space corresponding to the eigenvalue 1.

The other eigenvalues must be zero and the eigenspace is the kernel of this matrix. One such basis would be of vectors for $j = 1, 2, \dots, n$ of the form,

$$(y^i)_j = \begin{cases} 1 & j = i \\ -1 & j = n+1-i \\ 0 & \text{otherwise} \end{cases}$$

Suppose n is odd. Then the rank is $(n+1)/2$, and there are $(n+1)/2$ vectors of the form of x above. These again correspond to the eigenvalue 1.

We must now remove the y^i above corresponding to $(n+1)/2$ as this is not even well defined. The rest of the vectors are unchanged.

(c) Yes. Since P does not change a palindrome then P is a projector. Clearly $P = P^*$. Together these imply P is an orthogonal projector.

(d) As stated above, $r = \lceil n/2 \rceil$. Clearly all nonzero singular values are equal to one. Since the eigenbasis found above is orthogonal, we need only make each vector unit length. Therefore, $u_i = v_i = x^i/\sqrt{2}$ except for when n is odd, in which case $u_{\lceil n/2 \rceil} = v_{\lceil n/2 \rceil} = x^{\lceil n/2 \rceil}$.

Summer 2013, Day 1, Problem 5

Let V be a real Hilbert space and A a bounded linear operator on V . Recall that

$$\|A\| = \sup_{u \in V \setminus \{0\}} \frac{\|Au\|}{\|u\|} = \sup_{\|u\|=1} \|Au\|$$

where $v \in V$, $\|v\| = \langle v, v\rangle^{1/2}$.

(a) Prove that $\|A\| = \sup_{\|u\|=\|v\|=1} \langle Au, v\rangle = \sup_{u,v\neq 0} \frac{\langle Au,v\rangle}{\|u\|\|v\|}$

(b) Suppose $V = \mathbb{R}^n$ with the usual Euclidean inner product. What property of the unit ball in $\mathbb{R}^n$ ($\{u \in \mathbb{R}^n : \|u\| = 1\}$ guarantees that the supremum in (a) is actually achieved by some vectors $u_*, v_* \in \mathbb{R}^n$ with $\|u_*\| = \|v_*\| = 1$? If the linear operator A is represented by an n by n matrix, what are the vectors u_* and v_* that achieve this supremum called? Are they necessarily unique? Explain why or why not.

Solution

(a) Trivially,

$$\sup_{\|u\|,\|v\|=1} \langle Au, v\rangle = \sup_{\|u\|,\|v\|=1} \frac{\langle Au, v\rangle}{\|u\|\,\|v\|} \le \sup_{\|u\|,\|v\|\neq 0} \frac{\langle Au, v\rangle}{\|u\|\,\|v\|}$$

By the Cauchy–Schwarz inequality,

$$\sup_{\|u\|,\|v\|\neq 0} \frac{\langle Au, v\rangle}{\|u\|\,\|v\|} \le \sup_{\|u\|,\|v\|\neq 0} \frac{|\langle Au, v\rangle|}{\|u\|\,\|v\|} \le \sup_{\|u\|,\|v\|\neq 0} \frac{\|Au\|\,\|v\|}{\|u\|\,\|v\|} = \sup_{\|u\|\neq 0} \frac{\|Au\|}{\|u\|} = \|A\|$$

Finally,

$$\|A\| = \sup_{\|u\|=1} \|Au\| = \sup_{\|u\|=1} \frac{\|Au\|^2}{\|Au\|} = \sup_{\|u\|=1} \left\langle Au, \frac{Au}{\|Au\|} \right\rangle \le \sup_{\|u\|,\|v\|=1} \langle Au, v\rangle$$

This proves the desired equalities. ☐

(b) Since the unit ball is compact and the function $\langle Au, v\rangle$ is continuous then the supremum is attained.

The vectors attaining the maximum would be the singular vectors corresponding to the largest singular value. In particular, if u_* and v_* are such that $\langle Au_*, v_*\rangle = \|A\|$, then this means $Au_* = \|A\|\, v_* = \sigma_{\max} v_*$, where $\sigma_{\max}$ is the largest singular value.

Thus, u_* is one of the right-singular vectors corresponding to the largest singular value, and v_* is the corresponding left-singular vector.

They are unique up to complex sign if the largest singular value is unique (SVD uniqueness).

Winter 2013, Day 2, Problem 2

Suppose that $A \in \mathbb{C}^{m\times m}$ has an SVD $A = U\Sigma V^*$. Find an eigenvalue decomposition of the form $B = X\Lambda X^{-1}$ for the $2m \times 2m$ matrix,

$$B = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}$$

where A^* is the conjugate transpose matrix, $A^* = \overline{A}^T$.

Check that the eigenvectors of B are mutually orthogonal as expected since this matrix is Hermetian.

Solution

Write,

$$X = \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix}$$

Then $BX = X\Lambda$ is,

$$\begin{bmatrix} A^*X_3 & A^*X_4 \\ AX_1 & AX_2 \end{bmatrix} = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}\begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \end{bmatrix}\begin{bmatrix} \Lambda_1 & \\ & \Lambda_2 \end{bmatrix} = \begin{bmatrix} X_1\Lambda_1 & X_2\Lambda_2 \\ X_3\Lambda_1 & X_4\Lambda_2 \end{bmatrix}$$

This gives the equations,

$$\begin{aligned} A^*X_3 &= X_1\Lambda_1, & A^*X_4 &= X_2\Lambda_2 \\ AX_1 &= X_3\Lambda_1, & AX_2 &= X_4\Lambda_2 \end{aligned}$$

Note that $A = U\Sigma V^*$ means $AV = U\Sigma$ and $A^*U^* = V\Sigma$. Then clearly the previous equation is satisfied when,

$$\begin{aligned} \Lambda_1 &= \Sigma, & X_1 &= V, & X_3 &= U \\ \Lambda_2 &= -\Sigma, & X_2 &= V, & X_4 &= -U \end{aligned}$$

Thus a full eigen-decomposition can be written,

$$\begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}\begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix}\begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}$$

Clearly,

$$X^*X = \begin{bmatrix} V & V \\ U & -U \end{bmatrix}^* \begin{bmatrix} V & V \\ U & -U \end{bmatrix} = \begin{bmatrix} V^*V + U^*U & V^*V - U^*U \\ V^*V - U^*U & V^*V + U^*U \end{bmatrix} = 2I$$

Therefore we see that the columns of X are mutually orthogonal, as expected.

Practice 2010, Day 1, Problem 2

Define the inner product of two functions $f(x)$ and $g(x)$ defined on the interval $[0,1]$ by

$$\langle f, g\rangle = \int_0^1 f(x)g(x)\mathrm{d}x$$

We say f and g are orthogonal if $\langle f, g\rangle = 0$.

Let $\mathcal{P}$ be the space of cubic polynomials $p(x)$ satisfying $p(1) = p'(1) = 0$. This is a two-dimensional linear function space. Determine an orthogonal basis for this space with respect to the inner product above. Note: Orthogonal is enough, it need not be orthonormal.

Solution

Given that we are told the space is two dimensional, we can start with any two linearly independent polynomials from $\mathcal{P}$ and use Gram-Schmidt to orthogonalize these polynomials. The resulting polynomial will be in $\mathcal{P}$ (since the derivative is linear, but also since we are given that $\mathcal{P}$ is a linear function space).

We know polynomials with $p(1) = 0$ and $p'(1) = 0$ are divisible by $(x-1)^2$. Let,

$$p(x) = (x-1)^2 = x^2 - 2x + 1, \qquad q(x) = xp(x) = x^3 - 2x^2 + x$$

These are both in the space.

$$\begin{aligned}
\langle p, p\rangle &= \int_0^1 p(x)p(x)\mathrm{d}x = \int_0^1 (x-1)^4\mathrm{d}x = \frac{1}{5}(x-1)^5 + c = \frac{1}{5} \\
\langle p, q\rangle &= \int_0^1 p(x)q(x)\mathrm{d}x = \int_0^1 x(x-1)^4 = \frac{x^2}{30}(15 - 40x + 45x^2 - 24x^3 + 5x^4) + c = \frac{1}{30}
\end{aligned}$$

Let,

$$r = q - \frac{\langle p, q\rangle}{\langle p, p\rangle}p = q(x) - 6p(x) = 1 - 8x + 13x^2 - 6x^3$$

This is clearly orthogonal (by linear algebra rules) and also by verifying the inner product in Mathematica.

Summer 2013, Day 2, Problem 1

On Day 1, you showed that if V is a real Hilbert space and A is a bounded linear operator on V then,

$$\|A\| = \sup_{\|u\|=\|v\|=1} \langle Au, v\rangle = \sup_{u,v\neq 0} \frac{\langle Au, v\rangle}{\|u\|\,\|v\|}$$

where for $v \in V$, $\|v\| = \langle v, v\rangle^{1/2}$. Now assume additionally that A is a compact, self-adjoint, and bounded linear operator on V such that,

$$\langle Au, u\rangle > 0, \qquad \forall u \in V \setminus \{0\}$$

(a) Prove that $\|A\| = \sup_{\|u\|=1}\langle Au, u\rangle = \sup_{u\neq 0} \frac{\langle Au,u\rangle}{\|u\|^2}$.

(b) Without assuming that V is finite dimensional (as you did on Day 1), prove that there exists a nonzero vector $v \in V$ such that $Av = \|A\|\, v$; i.e., $\|A\|$ is the largest eigenvalue of A.

Solution

(a) Trivially,

$$\sup_{\|u\|=1} \langle Au, u\rangle = \sup_{\|u\|=1} \frac{\langle Au, u\rangle}{\|u\|^2} \leq \sup_{\|u\|\neq 0} \frac{\langle Au, u\rangle}{\|u\|^2}$$

By the Cauchy–Schwarz inequality,

$$\sup_{\|u\|\neq 0} \frac{\langle Au, u\rangle}{\|u\|^2} \leq \sup_{\|u\|\neq 0} \frac{|\langle Au, u\rangle|}{\|u\|^2} \leq \sup_{\|u\|\neq 0} \frac{\|Au\|\,\|u\|}{\|u\|^2} = \sup_{\|u\|\neq 0} \frac{\|Au\|}{\|u\|} = \|A\|$$

Now observe,

$$\begin{aligned}\langle A(u+v), (u+v)\rangle &= \langle Au, u\rangle + \langle Au, v\rangle + \langle Av, u\rangle + \langle Av, v\rangle \\ \langle A(u-v), (u-v)\rangle &= \langle Au, u\rangle - \langle Au, v\rangle - \langle Av, u\rangle + \langle Av, v\rangle\end{aligned}$$

Since V is a real Hilbert space, $\langle Au, v\rangle = \langle u, A^* v\rangle = \langle u, Av\rangle = \langle Av, u\rangle$. Thus,

$$\langle Au, v\rangle = \left(\langle A(u+v), (u+v)\rangle - \langle A(u-v), (u-v)\rangle\right)/4$$

Let $\alpha = \sup_{\|x\|=1} \langle Ax, x\rangle$. Applying the triangle inequality, definition of α, and parallelogram rule,

$$\begin{aligned}|\langle Au, v\rangle| &= |\langle A(u+v), (u+v)\rangle - \langle A(u-v), (u-v)\rangle|/4 \\ &\leq (|\langle A(u+v), (u+v)\rangle| + |\langle A(u-v), (u-v)\rangle|)/4 \\ &\leq \alpha(\langle (u+v), (u+v)\rangle + \langle (u-v), (u-v)\rangle)/4 \\ &= \alpha(2\langle u, u\rangle + 2\langle v, v\rangle)/4\end{aligned}$$

Thus,

$$\|A\| = \sup_{\|u\|,\|v\|=1} = \alpha(2\|u\| + 2\|v\|)/4 = \alpha = \sup_{\|u\|=1} \langle Au, u\rangle$$

The result is then proved. ☐

(b) Since the unit ball is compact and the function we are maximizing is continuous, we know the supremum is attained. Let v be such that $\|A\| = \langle Av, v\rangle$.

Note: IDK HOW TO DO THIS?????

Summer 2014, Day 1, Problem 2

Let A and B be $n \times n$ Hermitian matrices. Further, assume that A is non-negative definite, while B is positive definite. Define,

$$\lambda = \sup_{v \neq 0} \frac{v^* A v}{v^* B v}$$

Show that λ is the largest generalized eigenvalue of the pair (A, B): λ is the largest scalar satisfying $Ax = \lambda Bx$, for some nonzero vector x.

Solution

Recall that for any matrix M,

$$\nabla \left[x^* M x \right] = (M + M^*) x$$

Therefore, for M Hermetian, $\frac{\mathrm{d}}{\mathrm{d}x} \left[x^* M x \right] = 2x^* M$.

Observe, by the quotient rule,

$$\nabla \left[\frac{x^* A x}{x^* B x} \right] = \frac{(\nabla x^* A x)\,(x^* B x) - (x^* A x)(\nabla x^* B x)}{(x^* B x)^2} = \frac{(2Ax)(x^* B x) - (x^* A x)(2Bx)}{(x^* B x)^2}$$

Setting the gradient equal to zero we find,

$$(x^* B x)(Ax) = (x^* A x)(Bx)$$

We now dive by $x^* B x$ and obtain,

$$Ax = \frac{x^* A x}{x^* B x} Bx = \lambda Bx$$

We now show that the supremum is actually attained. It is clear that $x^* A x / x^* B x$ is invariant under $x \mapsto cx$ for $c > 0$. Thus taking the supremum over $x \neq 0$ is the same as taking it over $\|x\| = 1$. Our function is continuous and $\{x : \|x\| = 1\}$ is a compact. Therefore the supremum is attained somewhere in this set. In particular, this means there is a point x such that $x^* A x / x^* B x = \lambda$.

Let μ be a generalized eigenvalue satisfying $Ax = \mu Bx$. This implies $x^* A x = \mu x^* B x$ so,

$$\mu = \frac{x^* A x}{x^* B x} \leq \sup_{v \neq 0} \frac{v^* A v}{v^* B v} = \lambda$$

This proves that λ is the largest generalized eigenvalue of (A, B). ☐

Solution (Alternate)

Since B is positive-definite it admits a Cholesky decomposition $B = R^* R$. Note that B and therefore R are invertible. In particular this means $v \neq 0$ if and only if $Rv \neq 0$. Thus,

$$\lambda = \sup_{v \neq 0} \frac{v^* A v}{v^* B v} = \sup_{v \neq 0} \frac{v^* R^* R^{-*} A R^{-1} R v}{v^* R^* R v} = \sup_{y \neq 0} \frac{y^* (R^{-*} A R^{-1}) y}{y^* y} = \left\| R^{-*} A R^{-1} \right\|$$

Since A is Hermeitian then so is $R^{-*} A R^{-1}$. This means $\left\| R^{-*} A R^{-1} \right\|$ is the largest eigenvalue of $R^{-*} A R^{-1}$ implying that there exists y such that,

$$R^{-*} A R^{-1} y = \lambda y$$

Therefore, setting $x = R^{-1}y$,

$$Ax = AR^{-1}y = \lambda R^* y = \lambda R^* Rx = \lambda Bx$$

This proves that λ is a generalized eigenvalue of (A, B).

Let μ be a generalized eigenvalue satisfying $Ax = \mu Bx$. This implies $x^*Ax = \mu x^*Bx$ so,

$$\mu = \frac{x^*Ax}{x^*Bx} \leq \sup_{v \neq 0} \frac{v^*Av}{v^*Bv}$$

The result is then proved. □

Summer 2014, Day 2, Problem 4

Let A and B be $n \times n$ Hermitian non-negative definite matrices. Define,

$$\lambda = \sup_{v \notin \mathcal{N}(B)} \frac{v^* A v}{v^* B v}$$

where $\mathcal{N}(B)$ denotes the null space of B. What are the necessary and sufficient conditions so that λ is finite? **Note.** This problem setting differs from Problem 2 on Day 1, by the fact that B is not positive definite here. As a consequence, the sup above is taken over a different space.

Solution

We require $\mathcal{N}(A) \supseteq \mathcal{N}(B)$. Clearly this is a necessary condition. We prove that it is sufficient.

Let $\{u_j, \beta_j\}_{j=1}^n$ be an orthonomal eigen-decomposition of B with $\beta_1 \geq \beta_1 \geq \cdots \beta_n \geq 0$. Let r denote the rank of A and let $u \perp \mathcal{N}(B)$. Let β be the smallest positive eigenvalue of B. Then,

$$u^* B u = (a_1 u_1 + \cdots a_r u_r)^* B (a_1 u_1 + \cdots a_r u_r) = \sum_{j=1}^{r} a_j^2 \beta_j \leq \sum_{j=1}^{r} a_j^2 \beta = \beta u^* u$$

Let $\{v_j, \alpha_j\}_{j=1}^n$ be an orthonomal eigen-decomposition of A with $\alpha = \alpha_1 \geq \alpha_1 \geq \cdots \alpha_n \geq 0$. Let $v = a_1 v_1 + \cdots a_n v_n$ be any vector. Then,

$$v^* A v = (a_1 v_1 + \cdots a_n v_n)^* A (a_1 v_1 + \cdots a_n v_n) = \sum_{j=1}^{n} a_j^2 \alpha_j \leq \sum_{j=1}^{n} a_j^2 \alpha = \alpha v^* v$$

Write $v = u + w$ for $u \perp \mathcal{N}(B)$ and $w \in \mathcal{N}(B)$. Then, since B is Hermetian,

$$v^* B v = (u + w)^* B (u + w) = u^* B u + u^* B w + w^* B u + w^* B w = u^* B u \geq \beta u^* u$$

Likewise, since A is Hermetian and by hypothesis $w \in \mathcal{N}(A)$,

$$v^* A v = (u + w)^* A (u + w) = u^* A u + u^* A w + w^* A u + w^* A w = u^* A u \leq \alpha u^* u$$

Therefore, for each $v \notin \mathcal{N}(B)$,

$$\frac{v^* A v}{v^* b v} \leq \frac{\alpha u^* u}{\beta u^* u} = \frac{\alpha}{\beta} < \infty$$

This proves that $\lambda < \alpha/\beta < \infty$. $\square$

Winter 2014, Day 1, Problem 2

Suppose we have an $m \times n$ matrix W that has n orthonormal columns and $m > n$. Write Matlab statements to find an $m \times (m - n)$ matrix V that has $m - n$ orthonormal columns so that the matrix $Q = \begin{bmatrix} W & V \end{bmatrix}$ is unitary.

Solution

In full precision arithmetic we could append random vectors to the columns of W until it is $m \times m$. Since random matrices are almost always non-singular we expect this new matrix to be full rank and so we can run QR on the new matrix.

Numerically, this may work depending on what it means for the columns to be "orthogonal", and how much we care that the submatrix of Q is exactly that of the input. This is also computationally wasteful, but the simplest to implement.

Alternatively, we could pick up Modified Gram–Schmidt from the $(n+1)$-th column and continue from there.

```
Q = np.zeros((10,10))
Q[:,:n] = W

for i in range(n,m):
    Q[:,i] = np.random.rand(m) # pick random column
    # you could check the new column is linearly independent to the span of Q if
        you want
    for j in range(i):
        Q[:,i] -= np.dot(Q[:,i],Q[:,j])*Q[:,j]
    Q[:,i] /= np.linalg.norm(Q[:,i])
```

Winter 2014, Day 1, Problem 4

Let S be a symmetric matrix. The following is an incorrect proof that S is non-negative definite! Find the flaw in the proof.

Pf. Let $S = U\Sigma U^*$ be the SVD of S where U is unitary and Σ is a diagonal matrix of nonzero real elements ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. Then,

$$x^*Sx = x^*U\Sigma U^*x = y^*\Sigma y$$

where $y = U^*x$. Also, since U is unitary $x^*x = y^*y$, and we have,

$$\frac{x^*Sx}{x^*x} = \frac{y^*\Sigma y}{y^*y}$$

for all nonzero x (and hence nonzero y). Note now that $\sigma_n \leq \frac{y^*\Sigma y}{y^*y} \leq \sigma_1$ and since both σ_n and σ_1 are non-negative, it follows that S is non-negative definite.

Solution

We cannot assume the SVD of S has the form, $S = U\Sigma U^*$. For instance, take $S = [-1]$. Then if $u\sigma u$ is the SVD, $u\sigma u = \sigma u^2 > 0$, a contradiction.

Winter 2017, Day 2, Problem 1

Let A be an n by n real symmetric matrix with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$. Let x be any real n-vector with $\|x\|_2 = 1$. Show that the Rayleigh quotient

$$r(x) = x^T Ax$$

satisfies $\lambda_1 \leq r(x) \leq \lambda_n$. Show that, by varying x while keeping $\|x\|_2 = 1$ the Rayleigh quotient can take on every value in the interval $[\lambda_1, \lambda_n]$.

Solution

Since A is real and symmetric the eigenvalues are real and A can be unitarily diagonalized as $A = U\Lambda U^T$. Thus, with $y = U^T x$,

$$x^T Ax = x^T UAU^T x = y^T \Lambda y = \sum_{j=1}^{n} \lambda_j y_j^2$$

If $x^T x = 1$, then $y^T y = x^T UU^T x = x^T x = 1$. Therefore,

$$\sum_{j=1}^{n} y_j^2 = 1$$

Since $y_j^2 \geq 0$, this means that $x^T Ax$ is just the convex combination of the eigenvalues so it must be bounded by the largest and smallest.

Explicitly, since $y_j^2 \geq 0$,

$$\lambda_1 = \sum_{j=1}^{n} y_j^2 = \sum_{j=1}^{n} \lambda_1 y_j^2 \leq \sum_{j=1}^{n} \lambda_j y_j^2 \leq \sum_{j=1}^{n} \lambda_n y_j^2 = \lambda_n \sum_{j=1}^{n} y_j^2 = \lambda_n$$

Note that $v = cv_1 + \sqrt{1-c^2}v_n$ satisfies,

$$v^T v = c^2 v_1^* v_1 + c\sqrt{1-c^2}(v_1^* v_n + v_n^* v_1) + (1-c^2)v_n^* v_n$$

Therefore, if we cake v_1 and v_n as the unit eigenvectors corresponding to λ_1 and λ_n we have,

$$v^* v = c^2 v_1^* v_1 + 0 + 0 + (1-c^2)v_n^* v_n = c^2 + (1-c)^2 = 1$$

Moreover, with $c = 1$, $v^* Av = \lambda_1$ and when $c = 0$, $v^* Av = \lambda_n$. Clearly $v^T Av$ is a continuous function of c so by the intermediate value theorem the Rayleigh quotient can take on every value in the interval $[\lambda_1, \lambda_n]$. $\square$

Practice 2010, Day 2, Problem 3

Consider the boundary value problem,

$$-u_{xx} + 2u = f, \qquad 0 \le x \le 1, \qquad u(0) = 0, u(1) = 0$$

Assume that $f \in C^\infty([0,1])$.

(a) On a uniform grid with spacing $h = 1/n$, show that the following set of difference equations has local truncation error $\mathcal{O}(h^2)$:

$$\frac{2u_i - u_{i+1} - u_{i-1}}{h^2} + 2u_i = f(x_i), \qquad i = 1, \dots, n-1$$

Here u_i is then approximate solution at node $x_i = ih$, and the local truncation error $\tau = (\tau_1, \dots, \tau_{n-1})^T$ is is defined as the amount by which the true solution u fails to satisfy the difference equation at each node; i.e.

$$\tau_i = \frac{2u(x_i) - u(x_{i+1}) - u(x_{i-1})}{h^2} + 2u(x_i) - f(x_i), \qquad i = 1, \dots, n-1$$

(b) Use Gerschgorins Theorem to determine upper and lower bounds on the eigenvalues of the coefficient matrix for this set of difference equations.

(c) Show that the L_2-norm of the global error (the difference between the true solution and the approximate solution at the nodes) is of the same order as the local truncation error; i.e., $\mathcal{O}(h^2)$.

Solution

(a) Since we have used the second order approximation,

$$u_{xx} = \frac{2u_i - u_{i+1} - u_{i-1}}{h^2} + \mathcal{O}(h^2)$$

to the second derivative, the total error is $\mathcal{O}(h^2)$.

(b) The matrix for this finite difference method is tridiagonal. We have entries of $2 + 2/h^2$ on the main diagonal and $-1/h^2$ of the sub and super diagonals.

Gershgorin's theorem states that the eigenvalues are contained in the union of the $(n-1)$-disks centered at the diagonal entries, with radius equal to the sum of the modulii of the non-diagonal entires in the corresponding row. That is, all eigenvalues are contained in,

$$\bigcup_{i=1}^{n-1} \left\{ z : |a_{ii} - z| \le \sum_{j \ne i} |a_{ij}| \right\}$$

Since a_{ii} is the same for all i, we are concerned only with the largest radius of a disk. This clearly happens for the disks corresponding to $i = 2, \dots, n-2$. Each of these disks have radius $|-1/h^2| + |-1/h^2| = 2/h^2$.

Since the finite difference method's coefficient matrix is Hermetian all eigenvalues are real. Therefore, all eigenvalues are contained in,

$$\left\{ z : |2 + 2/h^2 - z| \le 2/h^2 \right\} \cap \mathbb{R} = [2, 2 + 4/h^2]$$

(c) Denoting the computed solution by $\hat{U}$, and the exact solution evaluated on the mesh by U we have error given by $E = U - \hat{U}$ and local truncation error given by $\tau = AU - F$. Thus,

$$\tau = AU - F = AU - A\hat{U} = A(U - \hat{U}) = AE$$

Therefore,

$$E = A^{-1}\tau$$

This gives the inequality,

$$\|E\|_{L_2} = \left\|A^{-1}\tau\right\|_{L_2} \leq \left\|A^{-1}\right\|_{L_2} \|\tau\|_{L_2}$$

We know $\|\tau\|_{L_2} = \mathcal{O}(h^2)$ since each entry of τ is $\mathcal{O}(h^2)$. It remains to show that $\left\|A^{-1}\right\|_{L_2} < C < \infty$ for some $C > 0$ independent of the number of mesh points.
Since A is Hermetian and all eignevalues are positive then the eignevalues are the singular values. Thus, $\sigma_{\min}(A) \in [2, 2 + 4/h^2]$. Therefore,

$$\left\|A^{-1}\right\|_2 = 1/\sigma_{\min}(A) \leq 1/2$$

Thus,

$$\left\|A^{-1}\right\|_{L_2} = \left\|A^{-1}\right\|_2 < \left\|A^{-1}\right\|_2 \leq 1/2$$

This proves that the L_2 norm of the global error is $\mathcal{O}(h^2)$.

Winter 2015, Day 2, Problem 3

Consider the two-point boundary value problem,

$$u_x - \epsilon u u_{xx} = 0$$

for $0 \leq x \leq 1$ with $u(0) = 2$ and $u(1) = 1$.

(a) What do you expect the solution to look like for small ϵ? In particular, is there a boundary layer, and where is it?

(b) Suggest a finite difference method to solve this equation on a uniform grid with grid spacing $\Delta x = 1/N$ for some integer N. Be sure to discuss boundary conditions.

(c) Determine the local truncation error and (local) order of accuracy for the method you proposed

(d) Explain how you would solve the resulting system of equations that result from your method. You do **not** need to implement it, but explain what is required.

(e) Suppose you wanted to use this method to obtain a decent approximation to the solution with $\epsilon = 10^{-6}$, e.g. a couple digits of accuracy at all points. Roughly how large must N be taken? Justify your answer.

Solution

(a) If ϵ is small the approximation looks like $u_x = 0$ so we expect the solution to be mostly constant. Since u is positive, we need u_x and u_{xx} to have the same sign. This means there must be a boundary layer on the right side.

(b) Second order centered approximations to u_x and u_{xx} could be used. The boundary conditions would be satisfied by inserting the values at the boundary into the equation.

(c) If we used second order for both derivatives the local order of accuracy is $\mathcal{O}(h^2)$. The LTE depends on the exact method chosen.

(d) This is a non-linear system of equations $F(x) = 0$, so an iterative method will probably have to be used. Newton's method is a common choice for this.

(e) We assume the boundary layer has width $\mathcal{O}(\epsilon)$. We would like a couple mesh points in this region. Therefore we would like a mesh spacing of something like $10^{-6}/n$, where n is the number of points in the boundary layer. Using something like a spacing of 10^{-7} gives $N = 10^7$ points.

Summer 2017, Day 1, Problem 4

Consider the two-point boundary-value problem,

$$u''(x) = f(x), \qquad 0 \le x \le 1, \qquad u(0) = u(1) = 0$$

Assume f is as smooth as you like.

(a) Show that

$$u(x) = \int_0^1 G(x, \xi) f(\xi) \mathrm{d}\xi$$

where

$$G(x, \xi) = \begin{cases} \xi(x-1) & 0 \le \xi \le x \le 1 \\ x(\xi - 1) & 0 \le x \le \xi \le 1 \end{cases}$$

(b) Replacing $f(x)$ by $f(x) + \Delta f(x)$, where $|\Delta f(x)| \le \epsilon$ for all x changes the solution $u(x)$ to $u(x) + \Delta u(x)$. Prove that

$$|\Delta u(x)| \le \frac{\epsilon}{2} x(1-x), \qquad 0 \le x \le 1$$

Solution

(a) We know that $(\partial^2/\partial x^2)G(x;\xi) = \delta(x-\xi)$. Thus, since the integral is with respect to ξ,

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[\int_0^1 G(x;\xi) f(\xi) \mathrm{d}\xi\right] = \int_0^1 \frac{\partial^2}{\partial x^2}\left[G(x;\xi)f(\xi)\right] \mathrm{d}\xi = \int_0^1 \delta(x-\xi) f(\xi) \mathrm{d}\xi = f(x)$$

(b) We have,

$$\begin{aligned} \int_0^1 G(x;\xi)(f(\xi) + \Delta f(\xi)) \mathrm{d}\xi &= \int_0^1 G(x;\xi) f(\xi) \mathrm{d}\xi + \int_0^1 G(x;\xi) \Delta f(\xi) \mathrm{d}\xi \\ &\le u(x) + \int_0^1 G(x;\xi) |\Delta f(\xi)| \mathrm{d}\xi \\ &\le u(x) + \epsilon \int_0^1 G(x;\xi) \mathrm{d}\xi \\ &= u(x) + \epsilon \left[\int_0^x \xi(x-1) \mathrm{d}\xi + \int_x^1 x(\xi - 1) \mathrm{d}\xi\right] \\ &= u(x) + \epsilon \left[x^2(x-1)/2 + -x/2 + x^2 - x^3/3\right] \\ &= u(x) + \epsilon x(1-x) \end{aligned}$$

Summer 2011, Day 2, Question 5

Consider the boundary value problem,

$$\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + k^2u = 0, \qquad 0 < x < 1, \qquad u(0) = 1, \qquad \frac{\mathrm{d}u}{\mathrm{d}x}(1) = iku(1)$$

- Write a second order finite difference scheme for this boundary value problem. (Denoting h the mesh size, the system of algebraic equations, $A_hu_h = f_h$, should involve a symmetric matrix)
- Prove the existence and uniqueness of a discrete solution.
- Investigate numerically whether the discrete solution u_h converges or not when $k = 75$. Write a simple Matlab script that solves the discrete problem. Do log-log plots for $\left\|(A_h)^{-1}\right\|_\infty$, $\|A_h\|_\infty$, and the maximum nodal error as a function of h. Use this information to discuss the stability of the method, the convergence, and the convergence rate.
- Assume that you use the Jacobi iterative method to solve the linear system. Under which condition does the Jacobi method converge? State the general convergence condition. Study numerically the convergence by doing a log-log plot of h vs. the quantity in the general convergence condition. Discuss when or whether this condition is satisfied. Your plot should exhibit a vertical asymptote. Explain why.

Solution

- Let $0 = x_0 < x_1 < \ldots < x_{m-1} < x_m = 1$ be equally spaced mesh points.

  On the interior will use the difference scheme,

  $$\frac{u_{i+1} + u_{i-1} - 2u_i}{h^2} + k^2u_i = 0$$

  We will satisfy the left boundary condition by setting $u_0 = 0$. In order to obtain a second order approximation of the solution on the right boundary we will introduce an extra point x_{m+1} and add the equations,

  $$\frac{u_{m+1} + u_{m-1} - 2u_m}{h^2} + k^2u_m = 0, \qquad \frac{u_{m+1} - u_{m-1}}{2h} = iku_m$$

  We now eliminate x_{m+1}. The second equation gives $x_{m+1} = 2ikhu_m + u_{m-1}$. Thus,

  $$0 = \frac{(2ikhu_m + u_{m-1}) + u_{m-1} - 2u_m}{h^2} + k^2u_m = \frac{2(ikh-1)u_m + 2u_{m-1}}{h^2} + k^2u_m$$

  Therefore,

  $$\frac{u_{m-1}}{h^2} + \left(\frac{ikh-1}{h^2} + \frac{k^2}{2}\right)u_m = 0$$

  In matrix form we have,

  $$\begin{bmatrix} k^2-2/h^2 & 1/h^2 & & & \\ 1/h^2 & k^2-2/h^2 & 1/h^2 & & \\ & \ddots & \ddots & \ddots & \\ & & 1/h^2 & k^2-2/h^2 & 1/h^2 \\ & & & 1/h^2 & ik/h+k^2/2-1/h^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{m-1} \\ u_m \end{bmatrix} = \begin{bmatrix} -1/h^2 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

- Will this be full rank??

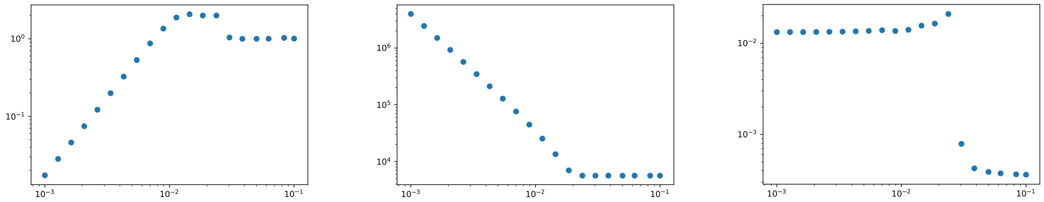
  If $k^2 \approx 2/h^2$ it will become highly ill conditioned for large h. In this case the vector $x = [1, 1, -1, -1, 1, 1, \dots]$ maps to a point very near the origin. However the vector $y = [1, 1, 1, \dots]$ maps to somewhere not that close to the origin. Therefore the ratio of the singular values of A (condition number) must be large.

- The solution to this differential equation is,

  $$u(x) = \cos(kx) + i\sin(kx)$$

  Since $\left\|A^{-1}\right\|_\infty$ is bounded for sufficiently small h the method is stable.

  We see that the error does begin to decrease like $\mathcal{O}(h^2)$ as expected. Therefore the method is convergent (as expected since it is consistent with $\mathcal{O}(h^2)$ LTE).



(a) infinity norm of error vs. h (b) $\|A\|_\infty$ vs. h (c) $\left\|A^{-1}\right\|_\infty$ vs. h

Figure 1

- Jacobi iteration will converge if and only if $\rho(I - M^{-1}A) < 1$, where $M = \operatorname{diag}(A)$.

  We see that the Jacobi iteration does not converge for values of h above a certain threshold.

  The vertical asymtote happens around $m = 53 \approx k/\sqrt{2}$ where the diagonal entries of most of the matrix are zero.

  Note: WHY

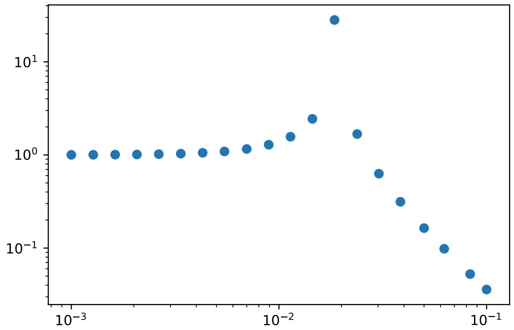


Figure 2: spectral radius of $I - M^{-1}A$ vs. h

Winter 2011, Day 2, Problem 5

Consider the Helmholtz problem,

$$u_{xx} + u_{yy} + k^2 u = f(x,y) = (k^2 - 5\pi^2)\sin(\pi x)\sin(2\pi y)$$

with $u(x,y) = 0$ on the boundary of the unit square, $(x,y) \in [0,1]^2$.

(a) Solve the Helmholtz problem using the 5-point Laplacian (second order finite difference) and the backslash as linear system solver. Verify your code works for $k = 5$, $k = 10$, and $k = 60$ by giving log-log plots for the maximum nodal error as a function of h.

(b) Suppose we change the solver to Jacobi (say take 200 iterations of Jacobis method). Program this in your code using the matrix version of Jacobi. Derive the spectral radius of Jacobis iteration matrix in terms of h and k. Recall the eigenvalues λ_{pq} of the 5-point Laplacian are,

$$\frac{2}{h^2}(\cos(p\pi h) + \cos(q\pi h) - 2)$$

where $h = 1/(m+1)$, $p = 1, 2, \ldots, m$, $q = 1, 2, \ldots, m$.

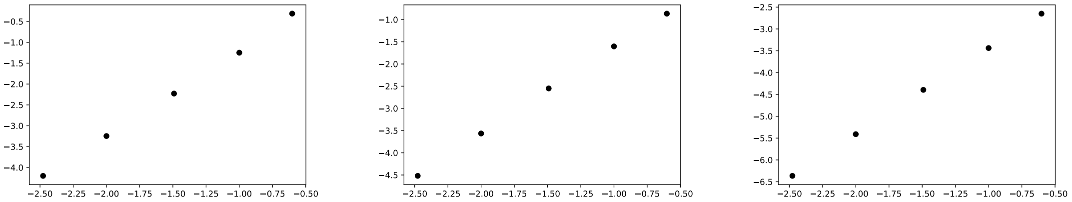
(c) If we fix $h = 1/21$, for what values of k will Jacobi's method converge? Verify this in your code by trying $k = 5$, $k = 10$, and $k = 60$, and any other k values, say $k = 0$ for example, you deem appropriate.

Solution

(a) Note that $u(x) = \sin(\pi x)\sin(2\pi y)$ is the solution to this equation for any k.

```
for k in [5,10,60]:
    def f(x,y):
        return (k**2-5*np.pi**2)*np.sin(np.pi*x)*np.sin(2*np.pi*y)

    def u_true(x,y):
        return np.sin(np.pi*x)*np.sin(2*np.pi*y)

    mesh_sizes = np.array([3,9,30,99,299]);
    max_error = np.zeros(len(mesh_sizes));

    for j,m in enumerate(mesh_sizes): # number of interior mesh points in a
        given direction
        h = 1/(m+1)


        # construct A
        T = sp.sparse.diags([-4*np.ones(m),np.ones(m-1),np.ones(m-1)
            ],[0,1,-1])

        A = sp.sparse.kron(sp.sparse.eye(m),T) + sp.sparse.kron(sp.sparse.
            diags([np.ones(m-1),np.ones(m-1)],[-1,1]),sp.sparse.eye(m))

        A /= h**2
        A += k**2 * sp.sparse.eye(m*m)

        # construct right hand side F
        xy = np.linspace(0,1,m+2)[1:-1] # get position of interior points

        F = np.reshape([[f(x,y) for x in xy] for y in xy],-1)
```

```
        # solve system
        U = sp.sparse.linalg.spsolve(A,F)

        U_true = np.reshape([[u_true(x,y) for x in xy] for y in xy],-1)

        max_error[j] = np.max(np.abs(U-U_true))


    plt.figure()
    plt.scatter(np.log10(1/(mesh_sizes+1)),np.log10(max_error),color='k')
    plt.savefig('w2011d2p5_'+str(k)+'.pdf')
```



(a) $k = 5$ (b) $k = 10$ (c) $k = 60$

Figure 3: max nodal error vs. mesh size

(b) Recall the Jacobi method is of the form,

$$x_k = x_{k-1} + M^{-1}(b - Ax_{k-1}) = b + (I - M^{-1}A)x_{k-1}$$

Thus,

$$e_k = (I - M^{-1}A)e_{k-1} = (I - M^{-1}A)e_0$$

We implement Jacobi iteration is Numpy as,

```
def jacobi(A,b,x,max_iter):
    M = sp.sparse.diags(A.diagonal())
    for n in range(max_iter):
        r = b-A@x
        x += sp.sparse.linalg.spsolve(M,r)
    return x
```

Note that the eigenvalues of A are,

$$k^2 + \frac{2}{h^2}(\cos(p\pi h) + \cos(q\pi h) - 2)$$

Now note that $M = \operatorname{diag}(A) = (k^2 - 4/h^2)I$ so the eigenvalues of $I - M^{-1}A = I - (k^2 - 4/h^2)^{-1}A$ are,

$$1 - \left(k^2 - \frac{4}{h^2}\right)^{-1}\left(k^2 + \frac{2}{h^2}(\cos(p\pi h) + \cos(q\pi h) - 2)\right)$$

We can rewrite this as,

$$\lambda(h,k,p,q) = 1 - \frac{h^2k^2 + 2(\cos(p\pi h) + \cos(q\pi h) - 2)}{h^2k^2 - 4}$$

It is clear that fixed h and k, the above expression is identical when $p = q = 1$ or $p = q = m$.

Note that this is a linear function of $\cos(p\pi h) + \cos(q\pi h)$. Now take p away from one of these points. The expression decreases. Thus,

$$\rho(I - M^{-1}A) = \left|1 - \frac{h^2k^2 + 2(\cos(\pi h) + \cos(\pi h) - 2)}{h^2k^2 - 4}\right|$$

(c) Jacobi iteration will converge if and only if $\rho(I - M^{-1}A) < 1$. We require,

$$-1 < 1 - \frac{h^2k^2 + 2(\cos(m\pi h) + \cos(m\pi h) - 2)}{h^2k^2 - 4} < 1$$

We solve $|1 - (h^2k^2 - 4)^{-1}(h^2k^2 + 2(\cos(m\pi h) + \cos(m\pi h) - 2))| = 1$ with Mathematica and obtain solutions,

$$k = \pm 42\sqrt{1 + \cos(\pi/21)}, \pm 42\sqrt{1 - \cos(\pi/21)}$$

Using the plot (and assuming $k > 0$) it is clear that for any

$$k \notin \left(42\sqrt{1 - \cos(\pi/21)}, 42\sqrt{1 + \cos(\pi/21)}\right) \approx (4.43874, 59.2309)$$

the method will converge. This is observed in our tests.

Summer 2013, Day 3, Problem 1

Consider the PDE,

$$4u_{xx} + 12u_{xy} + 9u_{yy} = 0$$

on the rectangle $0 \leq x \leq 2$, $0 \leq y \leq 3$ with boundary conditions,

$$u(x,0) = 0, \qquad u(x,3) = 3, \qquad u(0,y) = 2, \qquad u(2,y) = 1$$

1. Develop a finite difference method for this problem on a uniform grid with $\Delta x = \Delta y = h$, and use it to solve the problem. Discuss the order of accuracy and convergence of your method.
2. Find an analytical solution to the problem.
3. Suppose that the coefficient of the cross-term u_{xy} in the PDE were $12 - \epsilon$ instead of 12, where ϵ is a small positive number. Call the solution to this modified problem $u_\epsilon(x,y)$. Let $M(\epsilon)$ be the max-norm of $|u_\epsilon(xmy) - u(x,y)|$ over the rectangle. Can you estimate how $M(\epsilon)$ scales with ϵ?

Solution

1. Let $m = 2/h - 1$ and $n = 3/h - 1$ be the number of interior points in the x and y directions.

   For $i \in (1, m+1) \cap \mathbb{Z}$,

   $$u_{xx} \approx \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2}$$

   For $j \in (1, n+1) \cap \mathbb{Z}$,

   $$u_{yy} \approx \frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{h^2}$$

   For $(i,j) \in (1, m+1) \times (1, n+1) \cap \mathbb{Z}^2$,

   $$u_{xy} \approx \frac{\mathrm{d}}{\mathrm{d}x} \frac{u_{j+1}(x) - u_{j-1}(x)}{2h} = \frac{u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1}}{4h^2}$$

   For i, j with $1 < i < m+1$ and $1 < j < m+1$,

   $$4\frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2} + 12\frac{u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1}}{4h^2} + 9\frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{h^2} = 0$$

   This gives the stencil,

   $$\begin{matrix} -3 & 9 & 3 \\ 4 & -26 & 4 \\ 3 & 9 & -3 \end{matrix}$$

   Using the standard ordering we have matrix equation,

   $$A = \begin{bmatrix} T & T_1 & & \\ T_2 & T & T_1 & \\ & \ddots & \ddots & \ddots \end{bmatrix}$$

where,

$$T = \begin{bmatrix} -26 & 4 & & \\ 4 & -26 & 4 & \\ & \ddots & \ddots & \ddots \end{bmatrix}, \qquad T_1 = \begin{bmatrix} 9 & 3 & & \\ -3 & 9 & 3 & \\ & \ddots & \ddots & \ddots \end{bmatrix}, \qquad T_2 = T_1^T$$

We deal with the boundary conditions by modifying the right hand side.

We implement this in Python,

```
m = 30
n = 50


T = sp.sparse.diags([-26*np.ones(m),4*np.ones(m-1),4*np.ones(m-1)],[0,1,-1])

T1 = sp.sparse.diags([9*np.ones(m),3*np.ones(m-1),-3*np.ones(m-1)],[0,1,-1])

T2 = T1.T

A = sp.sparse.kron(sp.sparse.eye(n,k=0),T)
A += sp.sparse.kron(sp.sparse.eye(n,k=1),T1)
A += sp.sparse.kron(sp.sparse.eye(n,k=-1),T2)

F = np.zeros(m*n)

for j in range(n):
    for i in range(m):
        k = j*m+i
        print((i,j),k)
        if i == 0: #left boundary
            F[k] += 4*2
        if i == m-1: #right boundary
            F[k] += 4*1
        if j == 0: #bottom boundary
            F[k] += 9*0
        if j == n-1: #top boundary
            F[k] += 9*3

U = sp.sparse.linalg.spsolve(A,-F)

u = np.reshape(U,(n,m))
plt.pcolor(u)
plt.axis(’image’)
```

2. **Note:** no idea, not relevant
3. **Note:** no idea, not relevant

Summer 2014, Day 1, Problem 5

The Matlab code poisson.m provided to you solves the problem,

$$u_{xx} + u_{yy} = -5\pi^2 \sin(\pi x)\sin(2\pi y)$$

with $u(x, y) = 0$ on the boundary of the unit square $0 \leq x \leq 1$, $0 \leq y \leq 1$.

The mesh spacing h is taken the same in both coordinate directions and $h = 1/(m+1)$ where there are m interior unknowns in each row and each column. This code uses sparse storage to create a matrix problem $A^h u^h = F^h$ where the interior unknowns are by rows, bottom to top, and within each row, from left to right using the natural rowwise ordering. The code is set up to solve this problem using $m = 20$. It uses Matlabs backslash command to solve the linear system. The true solution to this PDE is known and used in the code to set the correct boundary conditions. The norm $\left\| u^h u_{pde} \right\|_\infty$ measuring the max error in the discrete solution relative to the PDE at the nodal points is printed at the end.

Recall, the eigenvalues λ_{pq} of A^h are $(2/h^2)(\cos(p\pi h) + \cos(q\pi h) - 2)$ where $p = 1, 2, \ldots, m$, $q = 1, 2, \ldots, m$.

1. Modify the poisson.m code to solve the Helmholtz problem,

   $$u_{xx} + u_{yy} + k^2 u = (k^2 - 5\pi^2)\sin(\pi x)\sin(2\pi y)$$

   with $u(x, y) = 0$ on the boundary of the unit square $0 \leq x \leq 1$, $0 \leq x \leq 1$. You will need to know the exact solution to this new PDE to set the boundary conditions using the codes approach.
2. Solve the new problem using your modified code which still uses backslash as the linear system solver. Verify it works for $k = 3$, $k = 10$, and $k = 60$. Do you think the backslash command did any pivoting for any of these problems?

Solution

1. This was done in Winter 2011, Day 2, Problem 5
2. Probably need pivoting if $k > \sqrt{2}/\pi$ as the matrix will not longer be SPD in this case.

   Note: Check this later

Summer 2014, Day 3, Problem 3

A heavily used method for solving $Ax = b$ where A is sparse, nonsingular, and nonsymmetric is the Krylov-space method called GMRES. Starting with any initial guess x_1, the method at step k chooses $x_{k+1} = x_k + Qy$ where Q is an $n \times k$ matrix with orthonormal columns and y is a $k \times 1$ vector.

1. Derive the method by showing how Q and y are chosen.
2. Show how the Arnoldi process is used in the method.
3. Discuss the arithmetic complexity and convergence properties. How does the truncated GMRES compare to the untruncated algorithm in both complexity and convergence properties?
4. Program as much of the algorithm as you can by either using pseudo code or actual Matlab code.

Solution

1. The GMRES algorithm minimizes the 2-norm of the residual over successive Krylov spaces. In particular the columns of Q_k form an orthonormal basis for the space $\mathcal{K}_k = \operatorname{span}\{r_0, Ar_0, \ldots, A^k r_0\}$ (where we have used the k to make the dependence of Q on k explicit) and at each step we pick $y \in \mathbb{R}^k$ minimizing,

   $$\|r_{k+1}\| = \|b - Ax_{k+1}\| = \|b - Ax_k - AQ_k y\| = \|r_0 - AQ_k y\|$$

   Once we have found this y, the iterate x_{k+1} is obtained by $x_{k+1} =$.

   Note that the y we have described is not quite the same as the y listed in the question statement. In particular, since $x_k \in \mathcal{K}_k$ we can write it as $Q_{k-1}z$, where the last entry of z is zero. Therefore the y in the solution varies from the $y$in the problem by this amount.

   Note: Indexing might be off by one

2. The Arnoldi algorithm produces an orthonormal basis for $\mathcal{K}_k = \operatorname{span}\{r_0, Ar_0, \ldots A^k r_0\}$. When used in GMRES, this is used to construct Q. In particular, the Arnoldi algorithm start with some $q_1 = r_0 / \|r_0\|$ and orthogonalizes Aq_1 against this vector to produce q_2. This process is repeated, and can be written

   $$AQ_k = Q_{k+1}H_{k+1,1}$$

   for an upper Hessenberg matrix H,

   Thus, instead of explicitly minimizing $\|r_{k+1}\|$ we can instead solve the equivalent system,

   $$\min_{y \in \mathbb{R}^k} \|r_0 - Q_k H_{k+1,k} y\| = \min_{y \in \mathbb{R}^k} \|\beta \xi_1 - H_{k+1,k} y\|$$

   where we have used the fact that $r_0 = \|r_0\| q_1 = \|r_0\| Q_k \xi_1$ and defined $\beta = \|r_0\|$.

3. Unlike conjugate gradient for HPD matrices, GMRES requires that all of Q be stored.

   Similarly to CG, in exact arithmetic GMRES will converge in at most n iterations. However, due to floating point error, it may be the case that more than n iterations are required.

   One variant of GMRES does not save all of Q, but ocassional resets itself. This saves storage space and reduces the number of operations needed, however it will obviously converge less quickly on some problems.

4.
   - start with any initial guess x_0 and compute $r_0 = b - Ax_0$. Set $q_0 = r_0 / \|r_0\|$.
   - for $k = 1, 2, \ldots$:
     - Run one more step of arnoldi to obtain Q_{k+1} and $H_{k+1,k}$ satisfying, $AQ_k = H_{k+1,k}Q_{k+1}$
     - compute $x = r_0 + Q_k y_k$ where y_k minimizes $\|\beta \xi_1 - H_{k+1,k} y\|$

   Note that a QR decomposition of $H_{k+1,k}$ can be used at each step and can be saved and updated rather than recomputed.

Summer 2010, Day 1, Problem 4

Consider the two-step Adams-Bashforth method to solve the scalar equation $y' = f(t,y)$:

$$y_{n+2} = y_{n+1} + h\left[\frac{3}{2}f(t_{n+1}, y_{n+1}) - \frac{1}{2}f(t_n, y_n)\right]$$

Show that this method is convergent, find its order, and sketch its region of absolute stability. In particular, determine where this region intersects the real and imaginary axes.

Solution

This is a linear multistep method of the form,

$$\sum_{j=0}^{r} \alpha_j U^{n+1} = h \sum_{j=0}^{r} \beta_j f(U^{n+j}, t_{n+j})$$

where $r = 2$, $\alpha_0 = 0$, $\alpha_1 = -1$, $\alpha_2 = 1$, $\beta_0 = -1/2$, $\beta_1 = 3/2$, and $\beta_2 = 0$.

The local truncation error is given by,

$$\tau_n = \frac{y(t_{n+2}) - y(t_{n+1})}{h} - \frac{3}{2}y'(t_{n+1}) + \frac{1}{2}y'(t_n)$$

We expand,

$$y(t_{n+2}) = y(t_{n+1}) + hy'(t_{n+1}) + \frac{h^2}{2}y''(t_{n+1}) + \frac{h^3}{3!}y'''(t_{n+1}) + \mathcal{O}(h^4)$$

Thus,

$$\frac{y(t_{n+2}) - y(t_{n+1})}{h} = y'(t_{n+1}) + \frac{h}{2}y''(t_{n+1}) + \frac{h^2}{3!}y'''(t_{n+1}) + \mathcal{O}(h^3)$$

Now expand,

$$y'(t_n) = y'(t_{n+1}) - hy''(t_{n+1}) + \frac{h^2}{2}y'''(t_{n+1}) + \mathcal{O}(h^3)$$

Finally we see that,

$$\begin{aligned}
\tau_n &= y'(t_{n+1}) + \frac{h}{2}y''(t_{n+1}) + \frac{h^2}{3!}y'''(t_{n+1}) + \mathcal{O}(h^3) - \frac{3}{2}y'(t_{n+1}) \\
&\quad + \frac{1}{2}\left[y'(t_{n+1}) - hy''(t_{n+1}) + \frac{h^2}{2!}y'''(t_{n+1}) + \mathcal{O}(h^3)\right] \\
&= \frac{5}{12}y'''(t_{n+1})h^2 + \mathcal{O}(h^3)
\end{aligned}$$

Therefore the method is consistent and has order h^2.

The region of absolute stability is the region,

$$\{z : \rho(\zeta) - z\sigma(\zeta) \text{ satisfies the root condition}\}$$

where the root condition is that all roots have modulus at most one, and a root with modulus one is simple, and $\rho(\zeta) = \sum_{j=0}^{r} \alpha_j \zeta^k$ and $\sigma(\zeta) = \sum_{j=0}^{r} \beta_j \zeta^j$.

In this case we have,

$$\rho(\zeta) - z\sigma(\zeta) = \zeta^2 - \zeta - z(3/2\zeta - 1/2) = \zeta^2 - (1 + 3z/2)\zeta + z/2$$

Points on the boundary of the region of absolute stability have the form,

$$z = \rho(e^{it})/\sigma(e^{it}), \qquad t \in [0, 2\pi)$$

This looks like an egg to the left of the imaginary axis.

The points on the real axis are $-1, 0$ and the point on the imaginary axis is 0.

Note: double check this and solve analytically?

Since the origin is contained in the region of absolute stability, the method is zero-stable. Since it is also consistent the method is convergent.

Winter 2010, Day 1, Problem 3

Consider the one-step Adams-Moulton method (also known as the trapezoidal rule) to solve the scalar equation $y_0 = f(t, y)$:

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_{n+1}, y_{n+1}) + f(t_n, y_n)\right]$$

- Is this an explicit or an implicit method? Why?
- Show that this method is convergent, find its order, and sketch its region of absolute stability.

Solution

- This is an implicit method since we cannot solve for y_{n+1} without knowing the inverse of f.
- This is a linear multistep method of the form,

$$\sum_{j=0}^{r} \alpha_j U^{n+1} = h \sum_{j=0}^{r} \beta_j f(U^{n+j}, t_{n+j})$$

where $r = 1$, $\alpha_0 = -1$, $\alpha_1 = 1$, $\beta_0 = 1/2$, and $\beta_1 = 1/2$.

We have local truncation error,

$$\tau_n = \frac{y(t_{n+1}) - y(t_n)}{h} - \frac{1}{2}\left[y'(t_{n+1}) + y'(t_n)\right]$$

We expand,

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + hy'(t_n) + \frac{h^2}{2}y''(t_n) + \frac{h^3}{3!}y'''(t_n) + \mathcal{O}(h^4) \\ y'(t_{n+1}) &= y'(t_n) + hy''(t_n) + \frac{h^2}{2}y'''(t_n) + \mathcal{O}(h^3) \end{aligned}$$

Thus,

$$\begin{aligned} \tau_n &= y'(t_n) + \frac{h}{2}y''(t_n) + \frac{h^2}{3!}y'''(t_n) + \mathcal{O}(h^3) - \frac{1}{2}\left[2y'(t_n) + hy''(t_n) + \frac{h^2}{2}y'''(t_n) + \mathcal{O}(h^3)\right] \\ &= -\frac{1}{3}y'''(t_n)h^2 + \mathcal{O}(h^3) \end{aligned}$$

Therefore the method is consistent as $h \to 0$, and is order h^2.

The region of absolute stability is the region,

$$\{z : \rho(\zeta) - z\sigma(\zeta) \text{ satisfies the root condition}\}$$

where the root condition is that all roots have modulus at most one, and a root with modulus one is simple, and $\rho(\zeta) = \sum_{j=0}^{r} \alpha_j \zeta^k$ and $\sigma(\zeta) = \sum_{j=0}^{r} \beta_j \zeta^j$.

In this case we have,

$$\rho(\zeta) - z\sigma(\zeta) = \zeta - 1 - z(\zeta/2 + 1/2) = (1 + z/2)\zeta + z/2 - 1$$

The roots of $\rho(\zeta) - z\sigma(\zeta)$ are,

$$\zeta = (z/2 - 1)/(z/2 + 1)$$

We plot the region where this roott is less than or equal to one in modulus using Mathematica.

This gives the entire left half plane (as expected). Since the origin is contained in the region of absolute stability, the method is zero stable. This along with consistency implies convergence.

Summer 2011, Day 2, Problem 6

Consider the following linear multistep method:

$$y_{n+3} + (2b-3)(y_{n+2} - y_{n+1}) - y_n = hb(f_{n+2} + f_{n+1})$$

to approximate the ordinary differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x) = f(x, y(x)), \qquad y(0) = y_0$$

(where f is a smooth function).

(a) Determine all values of the real parameter $b, b \neq 0$, for which the method is zero-stable.

(b) Show that the truncation error,

$$T_n = \frac{y(x_{n+3}) + (2b-3)(y(x_{n+2}) - y(x_{n+1})) - y(x_n) - hb(f(x_{n+2}) + f(x_{n+1}))}{2hb}$$

(where y is a solution to the ordinary differential equation) is $\mathcal{O}(h^2)$ when the method is zero-stable.

(c) Show that there exists a value of b for which the truncation error is $\mathcal{O}(h^4)$.

Solution

(a) This is a linear multistep method with characteristic polynomial,

$$\begin{aligned}\rho(\zeta) &= \zeta^3 + (2b-3)\zeta^2 - (2b-3)\zeta - 1 \\ &= (\zeta - 1)(\zeta - 1 + b + \sqrt{b^2 - 2b})(\zeta - 1 + b - \sqrt{b^2 - 2b})\end{aligned}$$

The roots are then,

$$1, \qquad 1 - b \pm \sqrt{b^2 - 2b}$$

The method is zero stable if all roots have modulus less than or equal to one, and any roots of modulus one are simple.

First observe $b^2 - 2b \geq 0$ when $b \leq 0$ or $b \geq 2$ so the roots have size larger than one and do not satisfy the root condition.

Now observe that for $b = 1$ and $b = 0$ the roots are repeated and do not satisfy the root condition.

Suppose $b \in (0, 2)$. Then $|\zeta|^2 = (1-b)^2 + i^2((b-1)^2 - 1) = 1$. And the roots are different so the root condition is satisfied.

(b) We do this in Mathematica because fuck Taylor expansions.

First define,

```
S[j_] := Series[f[x + j h], {h, 0, 5}]
```

This will compute the Taylor exapnsion of $f(x + jh)$ up to 5-th order.

Now compute the truncation error and simplify,

```
FullSimplify[(S[3] + (2 b - 3) (S[2] - S[1]) - S[0])/(2 h b) - (D[f[x + 2 h],
    x] + D[f[x + h], x])/2]
```

This shows the truncation error is order h^2.

(c) From this the output of this it is clear that choosing $b = 6$ gives the desired order of accuracy. In particular we have truncation error,

$$T_n = \frac{1}{120} f^{(5)}(x) h^4 + \mathcal{O}(h^5)$$

Summer 2012, Day 2, Problem 3

Consider the scheme

$$u_{n+1} = u_n + hf(t_n + (1-\theta)h, \theta u_n + (1-\theta)u_{n+1})$$

for solving the ODE $u' = f(t,u)$. Here u_n and u_{n+1} are meant to approximate $u(t_n)$ and $u(t_{n+1}) = u(t_n + h)$, respectively.

(a) For all $\theta \in [0,1]$, find the order of this scheme.

(b) Determine for which $\theta \in [0,1]$ the scheme is convergent.

(c) For $\theta_0 = 0$ and $\theta_1 = 1$ determine the stability domain of the scheme.

(d) Consider the system,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} -a & 0 \\ 0 & -1/a \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \frac{\pi}{a}\begin{bmatrix} u_1u_2 \\ u_2^2 - u_1 \end{bmatrix}$$

where a is a positive parameter. Find its equilibrium points and determine their linear stability. Are your results qualitatively the same for all values of $a > 0$? For $a = 10$, draw a phase portrait that is qualitatively consistent with your findings. This plot does not have to be to scale.

(e) Choose a suitable value of θ and using the scheme above, produce a plot of the curve $[u_1(t), u_2(t)]^T$ with initial conditions $[1,-1]^T$ in the phase plane. Run the scheme until you get "reasonably" close to an equilibrium point.

Solution

(a) We have truncation error,

$$\tau_n = \frac{u(t_{n+1}) - u(t_n)}{h} - f(t_n + (1-\theta)h, \theta u(t_n) + (1-\theta)u(t_{n+1}))$$

Write,

$$f(t_{n+1} - \theta h, u(t_{n+1}) - \theta(\Delta u))$$

Write $t = t_n$, $u = u(t_n)$, $\Delta t = (1-\theta)h$, and $\Delta u = (1-\theta)(u(t_{n+1}) - u(t_n))$. Observe,

$$\begin{aligned}\Delta u &= (1-\theta)(u(t_{n+1}) - u(t_n)) \\ &= (1-\theta)\left(hu'(t_n) + \frac{h^2}{2}u''(t_n) + \mathcal{O}(h^3)\right) \\ &= \Delta t u'(t_n) + (1-\theta)\left(\frac{h^2}{2}u''(t_n) + \mathcal{O}(h^3)\right)\end{aligned}$$

Therefore,

$$\mathcal{O}(\Delta t) = \mathcal{O}(\Delta u) = \mathcal{O}(h)$$

Expand and simplify using $u'' = f_t(t,u) + f_u(t,u)u'(t)$,

$$\begin{aligned}f(t+\Delta t, u + \Delta u) &= f(t,u) + \Delta t f_t(t,u) + \Delta u f_u(t,u) + \mathcal{O}(\Delta t^2 + \Delta t\Delta u + \Delta u^2) \\ &= f(t,u) + \Delta t\left(f_t(t,u) + u'(t)f_u(t,u)\right) + \mathcal{O}(h^2) \\ &= u'(t) + \Delta t u''(t) + \mathcal{O}(h^2)\end{aligned}$$

Thus,

$$\tau_n = \frac{h}{2}u''(t_n) - (1-\theta)hu''(t_n) + \mathcal{O}(h^2)$$

Therefore the scheme has local truncation error $\mathcal{O}(h)$, and when $\theta = 1/2$ it is also $\mathcal{O}(h^2)$.

While we have not proved that when $\theta = 1/2$ the scheme is not actually higher order, we guess that it is very unlikely that this is the case.

(b) **Note:** Flesh this argument out

Note: easier to prove general case

We have,

$$u(t_{n+1}) = u(t_n) + hf(t_n + (1-\theta)h, \theta u(t_n) + (1-\theta)u(t_{n+1})) + \tau$$

$$\begin{aligned} d_{n+1} &= u(t_{n+1}) - u_{n+1} \\ &= d_n + h[f(t_n + (1-\theta)h, \theta u(t_n) + (1-\theta)u(t_{n+1})) \\ &\qquad - f(t_n + (1-\theta)h, \theta u_n + (1-\theta)u_{n+1})] \end{aligned}$$

$$|d_{n+1}| \leq |d_n| + hL(\theta|d_n| + (1-\theta)|d_{n+1}|) + h\tau$$

$$|d_{n+1}|(1 - hL(1-\theta)) \leq |d_n|(1 + hL\theta) + h\tau$$

Assume h small enough so that $1 - hL(1-\theta) \geq 1/2$. Then,

$$\begin{aligned} |d_{n+1}| &\leq \frac{1+hL\theta}{1-hL(1-\theta)}|d_n| + \frac{1}{1-hL(1-\theta)}h\tau \\ &:= r|d_n| + sh\tau \\ &\leq r^{n+1}|d_0| + sh\tau\left(\sum_{j=0}^{n} r^j\right) \\ &= r^{n+1}|d_0| + sh\tau\left(\frac{r^{n+1}-1}{r-1}\right) \end{aligned}$$

$$r = \frac{1 + hL\theta - hL + hL}{1 + hL\theta - hL} = 1 + \frac{hL}{1 + hL\theta - hL}$$

Then,

$$r^{n+1} \leq e^{(n+1)hL/(1+hL\theta-hL)} \leq e^{TL/(1+hL\theta-hL)} \leq e^{2TL}$$

Therefore r^{n+1} is bounded as $h \to 0$.

Note that,

$$\frac{sh}{r-1} = \frac{sh}{hLs} = \frac{1}{L}$$

This means $hs(r^{n+1}-1)/(r-1)$ is bounded, so since $\tau \to 0$ the method is convergent.

Therefore for any θ the method is convergent.

(c) When $\theta = 0$ this is backward Euler and has stability region outside of the circle of radius 1 centered at 1. Similarly, when $\theta = 1$ this is forward Euler and has stability region inside the circle of radius 1 centered at -1.

Note: stability domain == region of absolute stability??

(d) **Note:** Not relevant

(e) **Note:** do this later??

Winter 2015, Day 1, Problem 3

Consider the linear multistep method,

$$U^{n+2} = U^{n+1} + \frac{k}{2}(3f(U^{n+1}) - f(U^n)),$$

as a method for solving the ODE initial value problem $u_0(t) = f(u(t))$ with step size k.

(a) Is the method *convergent*? Explain what this term means and justify your answer.
(b) Determine the order of accuracy and the leading term in the local truncation error.
(c) Does the point $z = -1$ lie in the region of absolute stability for this method? Justify your answer.

Solution

(a) Convergent means that the numerical solution at a given value of t converges to the actual solution at this value of t in the limit as $k \to 0$. This is a linear multistep method so we have convergence if and only if we have stability and consistency.

Note: Can I determine this without LTE?

The method is consistent as shown in (b). Since $\rho(\zeta) = \zeta^2 - \zeta$ has roots of, $\zeta = 0$, and $\zeta = 1$ the method is zero-stable (these roots both have modulus less than or equal to one, and the root of modulus one is simple).

This means the method is convergent.

(b) We have local truncation error,

$$\tau_n = \frac{u(t_{n+2}) - u(t_{n+1})}{h} - \frac{3u'(t_{n+1}) - u'(t_n)}{2}$$

Using Mathematica we find,

$$\tau_n = \frac{5}{12}u'''(t_n)h^2 + \mathcal{O}(h^3)$$

(c) Consider the polynomial,

$$\pi(\zeta) = \rho(\zeta) + \sigma(\zeta) = \zeta^2 - \zeta + 3\zeta/2 - 1/2 = \zeta^2 - \zeta/2 - 1/2 = (\zeta + 1/2)(\zeta - 1)$$

The roots are then $\zeta = -1/2$ and $\zeta = 1$. These both have modulus less than or equal to one, and the root with modulus one is simple. Therefore $z = -1$ lies in the region of absolute stability.

Winter 2017, Day 1, Problem 2

Consider difference equations of the form,

$$u_{n+2} + a_1 u_{n+1} + a_0 u_n = kbf(u_{n+1})$$

for the initial value problem $u'(t) = f(u(t))$, where $k = t_{n+1} - t_n$ is the timestep.

(a) Determine the coefficients a_0, a_1, and b that give the highest order local truncation error and say what that order is.
(b) Is the resulting method convergnet? Say why or why not.
(c) Determine which, if any, of the points $-1, \pm i, \pm\frac{1}{2}i$ line in the region of absolute stability.

Solution

1. We have local truncation error,

$$\begin{aligned}\tau_n &= \frac{u(t_{n+2}) + a_1 u(t_{n+1}) + a_2 u(t_n)}{k} - bu'(t_{n+1}) \\ &= \frac{1 + a_0 + a_1}{h} f(t_n) + (2 + a_1 - b)f'(t_n) + \frac{1}{2}(4 + a_1 - 2b)f''(t_n)h + \frac{1}{6}(8 + a_1 - 3b) + \mathcal{O}(h^3)\end{aligned}$$

   We can eliminate the first three terms by choosing, $a_0 = -1$, $a_1 = 0$, and $b = 2$.

2. Observe,

$$\rho(\zeta) = \zeta^2 - 1 = (\zeta + 1)(\zeta - 1)$$

   Therefore the roots are $\zeta = \pm 1$, each with modulus less than one and simple. Therefore the method is zero-stable.

   For linear multistep methods we have consistent + zero-stable if and only if convergent.

3. We have stability polynomial,

$$\pi(\zeta; z) = \rho(\zeta) - z\sigma(\zeta) = \zeta^2 - 1 - z(2\zeta) = \zeta^2 - 2z\zeta - 1$$

   This has roots,

$$\zeta = z \pm \sqrt{z^2 + 1}$$

   When $z = -1$, $-1 - \sqrt{(-1)^2 + 1} = -1 - \sqrt{2} < -1$ so the root condition is not satisfied.

   When $z = \pm i$, $\pm i \pm \sqrt{(\pm i)^2 + 1} = \pm i$. However in both cases these are repeated roots so the root condition is not satisfied.

   When $z = \pm i/2$, $\pm i/2 \pm \sqrt{(\pm i/2)^2 + 1} = \pm i/2 \pm \sqrt{3/4}$. These both have modulus one and are simple so the root condition is satisfied.

Winter 2010, Day 2, Problem 1

Consider the advection equation $u_t(x,t) + au_x(x,t) = 0$ defined for all x where $a \neq 0$ is a constant.

(a) What is the solution $u(x,t)$ for $t > 0$ if this equation is solved with initial data $u(x,0) = \eta(x)$?

(b) Discretize in space with fixed mesh width Δx and time step Δt. Explain why the method,

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{2\Delta x}(U_{j+1}^n - U_{j-1}^n)$$

will not converge to the true solution if we refine in space and time with any fixed ratio $\Delta t/\Delta x$.

(c) Under what conditions will the method,

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{\Delta x}(U_{j+1}^n - U_j^n)$$

converge as we refine the grid with fixed ratio $\Delta t/\Delta x$?

Solution

(a) The exact solution is $u(x,t) = \eta(x - at)$.

(b) Write $h = \Delta x$ and $k = \Delta t$. We replace U_j^n by $g(\xi)^n e^{i\xi jh}$ to obtain,

$$g(\xi)^n e^{i\xi jh} = g(\xi)^n e^{i\xi jh} - \frac{ak}{h}\left(g(\xi)^n e^{i\xi(j+1)h} - g(\xi)^n e^{i\xi(j-1)h}\right)$$

Dividing by $g(\xi)^n e^{i\xi jh}$ we find,

$$g(\xi) = 1 - \frac{ak}{2h}\left(e^{i\xi h} - e^{-i\xi h}\right) = 1 - i\frac{ak}{h}\sin(\xi h)$$

Note that the sine can be like $-1 + \mathcal{O}(h)$ for some values of ξ. In this case, assuming that k/h is fixed, we have,

$$|g(\xi)| = \left|1 + i\frac{ak}{h} + \mathcal{O}(k)\right| \geq 1 + \mathcal{O}(k)$$

This proves the method is not convergent.

(c) Write $h = \Delta x$ and $k = \Delta t$. We replace U_j^n by $g(\xi)^n e^{i\xi jh}$ to obtain,

$$g(\xi)^n e^{i\xi jh} = g(\xi)^n e^{i\xi jh} - \frac{ak}{h}\left(g(\xi)^n e^{i\xi(j+1)h} - g(\xi)^n e^{i\xi jh}\right)$$

Dividing by $g(\xi)^n e^{i\xi jh}$ we find,

$$g(\xi) = 1 - \frac{ak}{h}\left(e^{i\xi h} - 1\right)$$

Note that the $g(\xi)$ are centered on a circle of radius ak/h centered at $1 + ak/h$.

Thus it is stable whenever $-1 \leq ah/h \leq 0$.

Note: WHAT ABOUT PLUS $\mathcal{O}(k)$??

Summer 2013, Day 1, Problem 2

Consider the advection equation $u_t + au_x = 0$ and the "skewed" upwind method,

$$U_j^{n+1} = U_{j-1}^n - \left(\frac{a\Delta t}{\Delta x} - 1\right)(U_{j-1}^n - U_{j-2}^n).$$

(a) Show that this method is first order accurate in space and time by computing the local truncation error.

(b) What restriction must be put on the time step Δt in terms of Δx in order for the CFL condition to be satisfied in the case $a > 0$? In the case $a < 0$?

(c) Show that the method is in fact stable provided the CFL condition is satisfied, i.e., with the bounds found in part (b).

Solution

(a) We have local truncation error,

$$\tau = \frac{u(x_j, t_{n+1}) - 2u(x_{j-1}, t_n) + u(x_{j-2}, t_n)}{k} + a\left(\frac{u(x_n, t_{j-1}) - u(x_n, t_{j-2})}{h}\right)$$

In Mathematica we define,

```
U[j_, n_] := Series[Series[u[x + j h, t + n k], {k, 0, 2}], {h, 0, 2}]
```

We compute the LTE using,

```
FullSimplify[(U[0, 1] - 2 U[-1, 0] + U[-2, 0])/k +  a ((U[-1, 0] - U[-2, 0])/h
    ), Assumptions -> {D[u[x, t], t] + a D[u[x, t], x] == 0}]
```

This shows that,

$$\tau_n = \frac{1}{2}u_{tt}k - \frac{3}{2}au_{xx}h + \frac{1}{k}u_{xx} + \mathcal{O}(k^2 + h^2)$$

This proves that the method is first order accurate in space and time.

(b) Write $h = \Delta x$ and $k = \Delta t$. The stencil depends only on the value of U^n at x_{j-1} and x_{j-2}. We therefore require $-2h < -ak < -h$. Equivalently, $ak/h \in [1, 2]$.

Thus we need $a > 0$ and then pick $k \in [h/a, 2h/a]$.

(c) Write $h = \Delta x$ and $k = \Delta t$. We replace U_j^n by $g(\xi)^n e^{i\xi jh}$ to obtain,

$$g(\xi)^{n+1}e^{i\xi jh} = g(\xi)^n e^{i\xi(j-1)h} - \left(\frac{ak}{h} - 1\right)\left(g(\xi)^n e^{i\xi(j-1)h} - g(\xi)^n e^{i\xi(j-2)h}\right)$$

Dividing by $g(\xi)^n e^{i\xi jh}$ we obtain,

$$g(\xi) = e^{i\xi h}\left(1 - \left(\frac{ak}{h} - 1\right)\left(1 - e^{-i\xi h}\right)\right)$$

The part above in the parenthesis is a circle of radius $|ak/h - 1|$ centered at $2 - ak/h$. For $ak/h \in [1, 2]$ the circle will be contained in the unit circle centered at the origin. Therefore $|g(\xi)| \leq 1$ whenever the CFL condition is satisfied.

Summer 2014, Day 2, Problem 1

Consider the advection equation $u_t + au_x = 0$ with periodic bounday conditions.

(a) Derive the Lax-Wendroff method below where $U_j^n \approx u(x_j, t_n)$ with $\Delta t = k$, and $\Delta x = h$:

$$U_j^{n+1} = U_j^n - \frac{ak}{2h}(U_{j+1}^n - U_{j-1}^n) + \frac{a^2k^2}{2h^2}(U_{j-1}^n - 2U_j^n + U_{j+1}^n).$$

(b) What is the order of the method?

(c) Derive the stability condition.

(d) If you program this method, would you expect to see dissipative behavior? Dispersive behavior? Justify your answer by finding a modified PDE on which the method is 3rd order accurate.

Solution

(a) We will derive this by interpolating where the characteristic passing through (x_j, t_{n+1}) intersects $t = t_n$ using the value of u at x_{j-1}, x_j, and x_{j+1}.

This is trivially done in Mathematica by,

```
Expand[InterpolatingPolynomial[{{x - h, Subscript[u, j - 1]}, {x, Subscript[u,
     j]}, {x + h, Subscript[u, j + 1]}}, z] /. {z -> x - a k}]
```

(b) Define,

```
U[j_, n_] := Series[Series[u[x + j h, t + n k], {k, 0, 4}], {h, 0, 4}]
```

Now compute,

```
FullSimplify[ U[0, 1] - (U[0, 0] - (a k)/(2 h) (U[1, 0] - U[-1, 0]) + (a^2 k
    ^2)/(2 h^2) (U[-1, 0] - 2 U[0, 0] + U[1, 0])), Assumptions -> {D[u[x, t],
    t] + a D[u[x, t], x] == 0}]
```

Note $u_{tx} + au_{xx} = 0$ and $u_{tt} + au_{xt} = 0$. If u is smooth enough $u_{xt} = u_{tx}$ so $u_{tt} - au_{xx} = 0$. Using this we see that the one step error is $\mathcal{O}(k^3 + kh^2)$. Dividing by k we find that the LTE is $\mathcal{O}(k^2 + h^2)$.

(c)

$$g(\xi) = 1 - i\frac{ak}{h}\sin(\xi h) + \frac{a^2k^2}{h^2}(\cos(\xi h) - 1)$$

Write $\nu = ak/h$. This is an ellipse centered at $1 - \nu^2$ with vertical (imaginary) axis of length ν and horizontal (real) axis of length ν^2.

Therefore one point on the real axis will always be at the point 1. The other point on the real axis will be at $1 - 2\nu^2$.

This means we need $-1 - \alpha k < 1 - 2\nu^2$ so $\nu^2 < 1 + \alpha k/2$. This implies if $-1 - \alpha k/4 < \nu < 1 + \alpha k/4$ for some constant α the entire ellipse will be contained within $\mathcal{O}(k)$ of the unit circle.

(d) Note that we suppress some of the dependences of v. Observe,

$$\left(\frac{v(t+k) - v(t)}{k}\right) + a\left(\frac{u(x+h) - u(x-h)}{2h}\right) + \frac{a^2k}{2}\left(\frac{u(x+h) - 2u(x) + u(x-h)}{h^2}\right)$$
$$= \left(v_t + v_{tt}\frac{k}{2} + v_{ttt}\frac{k^2}{3!} + \mathcal{O}(k^3)\right) + a\left(v_x + v_{xxx}\frac{h^2}{3!} + \mathcal{O}(h^4)\right)$$
$$+ \frac{a^2k}{2}\left(v_{xx} + v_{xxxx}\frac{h^2}{4!} + \mathcal{O}(h^4)\right)$$

Assuming $k = \mathcal{O}(h)$, v satisfies,

$$v_t + av_x + \frac{k}{2}v_{tt} + \frac{a^2k}{2}v_{xx} + \frac{k^2}{3!}v_{ttt} + \frac{ah^2}{3!}v_{xxx} = \mathcal{O}(k^3)$$

Therefore,

$$\begin{aligned} v_{tt} + av_{xt} + \frac{k}{2}v_{ttt} + \frac{a^2k}{2}v_{xxt} &= \mathcal{O}(k^2) \\ v_{tx} + av_{xx} + \frac{k}{2}v_{ttx} + \frac{a^2k}{2}v_{xxx} &= \mathcal{O}(k^2) \end{aligned}$$

If v is smooth enough that $v_{xt} = v_{tx}$ we have,

$$v_{tt} + a^2v_{xx} + \frac{ak}{2}v_{ttx} + \frac{a^3k}{2}v_{xxx} = \mathcal{O}(k^2)$$

Thus, taking the derivative of both sides with respect to x,

$$v_{ttx} + a^2v_{xxx} = \mathcal{O}(k)$$

Together we have,

$$v_{tt} + a^2v_{xx} + \frac{ak}{2}(-a^2v_{xxx} + \mathcal{O}(k)) + \frac{a^3k}{2}v_{xxx} = \mathcal{O}(k^2)$$

which can be written as,

$$v_{tt} + a^2v_{xx} = \mathcal{O}(k^2)$$

Taking the derivative of both sides with respect to t,

$$v_{ttt} + a^2v_{xxt} = \mathcal{O}(k^2)$$

From before we see that,

$$v_{txx} + av_{xxx} = \mathcal{O}(k)$$

Therefore,

$$v_{ttt} + a^3v_{xxx} = \mathcal{O}(k)$$

We now insert our expressions for v_{tt} and v_{ttt} into the original equation to find,

$$v_t + av_x + \frac{k}{2}(-a^2v_{xx} + \mathcal{O}(k^2)) + \frac{a^2k}{2}v_{xx} + \frac{k^2}{3!}(-a^3v_{xxx} + \mathcal{O}(k)) + \frac{ah^2}{3!}v_{xxx} = \mathcal{O}(k^3)$$

Rearranging gives,

$$v_t + av_x + \frac{a}{3!}\left(h^2 - a^2k^2\right)v_{xxx} = \mathcal{O}(k^3)$$

Therefore, to third order, v satisfies,

$$v_t + av_x = \frac{ah^2}{6}\left(\left(\frac{ak}{h}\right)^2 - 1\right)v_{xxx}$$

Therefore Lax-Wendroff is a 3rd order method when applied to this equation.

This is dispersive because there is an odd space derivative

Winter 2014, Day 2, Problem 1

Consider the advection equation $u_t + au_x = 0$ with periodic bounday conditions.

(a) Derive the Upwind method below where $U_j^n \approx u(x_j, t_n)$ with $\Delta t = k$ and $\Delta x = h$:

$$U_j^{n+1} = U_j^n - \frac{ak}{h}(U_j^n - U_{j-1}^n), \qquad a > 0$$

(b) Show the method is first order accurate in both time and space.

(c) Use either MOL or Von Neumann analysis to derive the stability condition.

(d) Show this Upwind method is an $\mathcal{O}(k^2)$ accurate method applied to the modified PDE, $v_t + av_x = .5ah(1 - \frac{ak}{h})v_{xx}$ when we keep $k = h$ fixed. Based on this, would you expect to see dissipative or dispersive behavior when the Upwind method is applied to the original advection equation? Explain.

Solution

(a) We approximate $u_x(x,t)$ by $(u(x,t) - u(x-h,t))/k$ and apply forward Euler to obtain the method. More specifically, we apply forward Euler to the system,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix} = -\frac{a}{h}\begin{bmatrix} 1 & & & -1 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}$$

(b) Since forward Euler is first order in time, and our backward difference is first order in space the method is first order accurate in time and space.

(c) Replace U_j^n by $g(\xi)^n e^{i\xi jh}$ to obtain,

$$g(\xi)^{n+1}e^{i\xi jh} = g(\xi)^n e^{i\xi jh} - \frac{ak}{h}\left(g(\xi)^n e^{i\xi jh} - g(\xi)^n e^{i\xi(j-1)h}\right)$$

Dividing by $g(\xi)e^{i\xi jh}$ we obtain,

$$g(\xi) = 1 - \frac{ak}{h}\left(1 - e^{-i\xi h}\right)$$

Then $g(\xi)$ is a circle of radius ak/h centered at $1 - ak/h$. This circle contained in a circle about the origin of radius $1 + \mathcal{O}(k)$ provided,

$$0 - \mathcal{O}(k) \le ak/h \le 1 + \mathcal{O}(k)$$

That is, when $ak/h \in [-\alpha k, 1 + \beta k]$ for some α, β, then $|g(\xi)| \le 1$.

(d) Observe,

$$\frac{v(x,t+k) - v(x,t)}{k} + a\left(\frac{v(x,t) - v(x-h,t)}{h}\right)$$
$$= v_t + v_{tt}\frac{k}{2} + \mathcal{O}(k^2) + a\left(v_x - v_{xx}\frac{h}{2} + \mathcal{O}(h^2)\right)$$

Assuming $k = h$ we have,

$$v_t + av_x + \frac{a}{2}(kv_{tt} - hv_{xx}) = \mathcal{O}(h^2)$$

Therefore,

$$\begin{aligned} v_{tt} + av_{xt} + \frac{a}{2}(kv_{ttt} - hv_{xxt}) &= \mathcal{O}(h^2) \\ v_{tx} + av_{xx} + \frac{a}{2}(kv_{ttx} - hv_{xxx}) &= \mathcal{O}(h^2) \end{aligned}$$

We assume v is smooth enough that $v_{xt} = v_{tx}$. Then,

$$v_{tt} - a^2 v_{xx} + \mathcal{O}(k)$$

Therefore,

$$v_t + av_x + \frac{a}{2}\left(k\left(a^2 v_{xx} + \mathcal{O}(k)\right) - hv_{xx}\right) = v_t + av_x + \frac{ah}{2}\left(\frac{ak}{h} - 1\right)v_{xx}$$

This shows that to second order v satisfies,

$$v_t + av_x = \frac{ah}{k}\left(1 - \frac{ak}{h}\right)v_{xx}$$

We expect dissipative behavior since there is a even derivative.

Summer 2017, Day 1, Problem 5

Consider the advection equation on an infinite domain,

$$u_t + au_x = 0, \qquad x \in (-\infty, \infty)$$

and finite difference schemes of the form,

$$u_j^{n+1} = \alpha u_{j-1}^n + \beta u_{j+1}^n$$

where u_j^n is the approximation to $u(x_j, t_n)$ and $x_j = jh$, $j = 0, \pm 1, \dots$, and $t_n = nk$, $n = 0, 1, \dots$, where u_j^0 is given.

(a) For what values of α and β is the scheme *consistent* with the defined equation, assuming that k/h and h/k remain bounded as $k, h \to 0$?

(b) Use von Neumann analysis or a method of your choice to show that the method is stable if $|\alpha| + |\beta| \leq 1$.

Solution

(a) We expand,

$$\begin{aligned} &u(x, t+k) - \alpha u(x-h, t) - \beta u(x+h, t) \\ &\quad = (1 - \alpha - \beta)u(x,t) + u_t(x,t)k + \mathcal{O}(k^2) + (\alpha - \beta)u_x h + u_{xx}(-\alpha - \beta)\frac{h^2}{2!} + \mathcal{O}(h^3) \end{aligned}$$

This shows we need $\alpha + \beta = 1 + \mathcal{O}(h^2)$ and $(\alpha - \beta)u_x h + u_t k = 0 + \mathcal{O}(h^2)$.

This has solution,

$$\alpha = \frac{1}{2} + \frac{ak}{2h} + \mathcal{O}(h^2), \qquad \beta = \frac{1}{2} - \frac{ak}{2h} + \mathcal{O}(h^2)$$

(b) Replace u_j^n by $g(\xi)^n e^{i\xi jh}$ to obtain,

$$g(\xi)^{n+1} e^{i\xi jh} = \alpha g(\xi)^n e^{i\xi(j-1)h} + \beta g(\xi)^n e^{i\xi(j+1)h}$$

Dividing by $g(\xi)^n e^{i\xi jh}$ we obtain,

$$g(\xi) = \alpha e^{-i\xi h} + \beta e^{i\xi h}$$

Suppose $|\alpha| + |\beta| \leq 1$. Then, by the triangle inequality,

$$|g(\xi)| \leq \left|\alpha e^{-i\xi h}\right| + \left|\beta e^{i\xi h}\right| = |\alpha| + |\beta| \leq 1$$

Therefore the method is stable.

Winter 2010, Day 1, Problem 1

Is it possible for the determinant of a non-triangular matrix to be equal to the product of the diagonal entries? If not, why not. If so, give an example.

Solution

Yes. Any rank defficient matrix has zero determinant. Clearly there exist rank-defficient matrices with a zero on the diagonal which are non-triangular. For instance,

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Winter 2012, Day 2, Problem 1

Denote by $\lambda_j(M)$ the j-th eigenvalue of the real symmetric $N \times N$ matrix $M : \lambda_1(M) \leq \lambda_2(M) \leq \cdots \leq \lambda_N(M)$. In this list, eigenvalues are repeated according to their algebraic multiplicity. Let S and T be $N \times N$ real symmetric matrices. How are $\lambda_1(S+T)$, $\lambda_1(S)+\lambda_1(T)$, and $\lambda_1(S)+\lambda_N(T)$ related (i.e., is one equal to another, is one less than another, or greater, etc.)

Solution

Recall that for any real symmetric matrix A and any vector r,

$$r^*Ar \in [\lambda_1(A), \lambda_N(A)]$$

Then $w^*Sw \geq \lambda_1(S)$ and $w^*Tw \geq \lambda_1(T)$ so that,

$$\lambda_1(S)+\lambda_1(T) \leq w^*Sw + w^*Tw = w^*(S+T)w = \lambda_1(S+T)$$

Let u and v be unit vectors such that,

$$Su = \lambda_1(S)u, \qquad Tv = \lambda_N(T)v$$

Then, since $\lambda_N(T) \geq u^*Tu$,

$$\lambda_1(S+T) \leq u^*(S+T)u = u^*Su + u^*Tu \leq u^*Su + v^*Tv = \lambda_1(S) + \lambda_N(T)$$

Therefore,

$$\lambda_1(S)+\lambda_1(T) \leq \lambda_1(S+T) \leq \lambda_1(S)+\lambda_N(T)$$

All inequalities can be tight (take $T = S = I$) but none are strict equalities.

Note: WHY DID THIS TAKE ME SO LONG. GOD DAMN.

Winter 2013, Day 2, Problem 1

This is the same as Summer 2013, Day 2, Problem 1

Summer 2014, Day 2, Problem 3

Suppose we want to compute,

$$\int_0^\pi \frac{1}{\sin^{1/4}(x)}\mathrm{d}x$$

A standard formula like the trapezoidal rule will break down since $f(x) = (\sin(x))^{-1/4}$ blows up at $x = 0$ and $x = \pi$. A method that avoids evaluating f at its singularities will often still need to evaluate f at many points to capture the singularity and compute a good approximation.

Notice that the singularities $x = 0$ and $x = \pi$ are integrable, which is why the question makes sense in the first place. Break up the integral in a sum of integrals,

$$\int_0^\pi \frac{1}{\sin^{1/4}(x)}\mathrm{d}x = \int_0^\alpha \frac{1}{\sin^{1/4}(x)}\mathrm{d}x + \int_\alpha^\pi \frac{1}{\sin^{1/4}(x)}\mathrm{d}x$$

each one containing only one singularity of the integrand (as an endpoint). Here $\alpha \in (0, \pi)$ is a number you get to pick for your convenience.

Near the singularity $x = 0$ we can write $f(x) = x^{-1/4}g(x)$ where $g(x)$ is a smooth function. We use the fact that the singular part $x^{-1/4}$ can be integrated exactly to find accurate formulas that require evaluating g at only a few points, e.g.

$$\int_0^\alpha x^{-1/4}g(x)\mathrm{d}x \approx \sum_{j=1}^{n} w_j g(x_j)$$

where the x_j are, for example, equally spaced points in the interval. To find such a formula we must determine the weights w_j to use, which depends on the set of points x_j chosen. One way to do this is to require that the above integration formula is exact for the n functions $g(x) = 1, x, x^2, \dots, x^{n-1}$. Note that for these choices of g the integral on the left-hand side can be computed exactly. Thus this results in a linear system of n equations to solve for the weights.

(a) With your choice of α, determine the weights w_1, w_2, w_3 using this approach for the case $n = 3$, using equally spaced points $x_1 = 0, x_2 = \alpha/2, x_3 = \alpha$. Requiring that the above equality holds for $g(x) = 1, x, x^2$ gives a linear system of 3 equations for the w's that can easily be solved however you want (by hand or by computer).

(b) Use these weights to estimate the integral $\int_0^\alpha (\sin(x))^{-1/4}\mathrm{d}x$.

(c) Deal with $\int_\alpha^\pi (\sin(x))^{-1/4}\mathrm{d}x$ however you want.

(d) Use their combination to estimate the original integral. It should agree to several digits with the exact value, which you can get very accurately using your favorite software

---

Solution

(a) **Note:** maybe pick $\alpha = \pi/2$ to use symmetry in (c)

We pick $\alpha = 1$ and solve the system,

$$\frac{4}{3} = \int_0^1 x^{-1/4}\mathrm{d}x = \int_0^\alpha x^{-1/4}\mathrm{d}x = \sum_{j=1}^3 w_j$$
$$\frac{4}{7} = \int_0^1 x^{3/4}\mathrm{d}x = \int_0^\alpha x^{-1/4}x\mathrm{d}x = \sum_{j=1}^3 w_j x_j$$
$$\frac{4}{11} = \int_0^1 x^{7/4}\mathrm{d}x = \int_0^\alpha x^{-1/4}x^2\mathrm{d}x = \sum_{j=1}^3 w_j x_j^2$$

That is,

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1/2 & 1 \\ 0 & 1/4 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 4/3 \\ 3/7 \\ 4/11 \end{bmatrix}$$

This has solution,

$$w_1 = \frac{80}{231}, \qquad w_2 = \frac{64}{77}, \qquad w_3 = \frac{12}{77}$$

(b) Near $x = 0$ we can write,

$$\sin(x)^{-1/4} \approx x^{-1/4}\left(\frac{\sin(x)}{x}\right)^{-1/4}$$

Therefore take $g(x) = \mathrm{sinc}(x)$.

Therefore we have,

$$\int_0^\alpha x^{-1/4} g(x)\mathrm{d}x \approx \sum_{j=1}^n w_j g(x_j) \approx 1.34898$$

(c) We use Mathematica.

(d) It does agree.

Winter 2017, Day 2, Problem 4

Consider the linear ODE system $u'(t) = Au(t)$, with an arbitrary initial vector $u(0)$.

(a) Show that for a general real, constant matrix A, the necessary and sufficient condition for $\|u(t)\|_2$ to decrease monotonically (for initial vector $u(0)$) is that the eigenvalues of $A + A^T$ be negative.

(b) Give an example to show that in general it is not sufficient to have the eigenvalues of A in the left half-plane; that is, write down a matrix A whose eigenvalues are all in the left half-plane but for which the 2-norm of the solution to $u' = Au$ with some initial vector $u(0)$ does not decay monotonically.

Solution

(a) We suppress the explicit dependence on t. Observe,

$$\frac{\mathrm{d}}{\mathrm{d}t}u^T u = \frac{\mathrm{d}}{\mathrm{d}t}\left(u_1^2 + \cdots u_n^2\right) = 2u_1u_1' + \cdots 2u_nu_n' = 2u^Tu' = 2u^TAu$$

Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|_2 = \frac{\mathrm{d}}{\mathrm{d}t}(u^Tu)^{1/2} = \left(\frac{1}{2}(u^Tu)^{-1/2}\right)\left(\frac{\mathrm{d}}{\mathrm{d}t}u^Tu\right) = \frac{u^TAu}{\|u\|_2}$$

For $\|u(t)\|_2$ to decrease monotonically, we need $\frac{\mathrm{d}}{\mathrm{d}t}\|u(t)\| \leq 0$. This happens if and only if,

$$u^TAu \leq 0$$

Since A is real,

$$u^TAu = (u^TAu)^T = u^TA^Tu$$

Therefore,

$$u^TAu = \frac{1}{2}\left(u^TAu + u^TA^Tu\right) = \frac{1}{2}u^T(A + A^T)u$$

Recall that B is symmetric negative definite (all negative eigenvalues) if and only if $v^TBv < 0$ for all v.

Therefore we find that $\|u(t)\|_2$ decreases monotonically if and only if $A + A^T$ has non-positive eigenvalues, and decreases strictly monotonically if $A + A^T$ has negative eigenvalues.

(b) **Note:** Fix some errors and simplify

Consider the system,

$$u'(t) = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix} u(t)$$

Note: TODO

Summer 2010, Day 2, Problem 1

Suppose we wish to compute a set of points (x_j, y_j) that can be connected to form a circle of radius 1. We could simply set $x_j = \cos(\theta_j)$ and $y_j = \sin(\theta_j)$ for some points $\theta_j = j\Delta\theta$. But that would be too easy. Instead, suppose we decide to numerically solve the system of ODEs:

$$\begin{aligned} x'(\theta) &= -y(\theta), & x(0) &= 1 \\ y'(\theta) &= x(\theta), & y(0) &= 0 \end{aligned}$$

The exact solution, when plotted in the x-y plane, traces out the desired circle. If we use the Forward Euler method, however, we obtain a figure like this:

[Figure of circle spirialing out]

This shows the points computed with $\Delta\theta = 2\pi/50$ for $j = 0, 1, \ldots, 100$ The computed solution spirals outwards rather than tracing a circle.

Let $u = [x, y]^T$, so that the Forward Euler method can be written as $u_{j+1} = Cu_j$ for some matrix C.

(a) Explain why $\|u_j\|_2$ increases with j based on the eigenvalues of C.
(b) Using the eigenstructure of C, determine what x_{1000} and y_{1000} would be, with $\Delta\theta$ as above.
(c) If Backward Euler is used instead, the computed curve will spiral inward instead of outward. Explain this using eigenvalue analysis.
(d) How will the curve behave if the trapezoidal method is used?
(e) Produce plots on the computer analogous to the figure above for the Backward Euler and Trapezoidal methods by programming this in Matlab or another language.

Solution

(a) We have,

$$u_{j+1} = u_j + k \begin{bmatrix} & -1 \\ 1 & \end{bmatrix} u_j = \begin{bmatrix} 1 & -k \\ k & 1 \end{bmatrix} u_j$$

Therefore C has eigenvalue/vector pairs,

$$\left(1 + ik, \begin{bmatrix} -i \\ 1 \end{bmatrix}\right), \qquad \left(1 - ik, \begin{bmatrix} i \\ 1 \end{bmatrix}\right)$$

Both eigenvalues have modulus greater than 1, and the eigenvectors are orthogonal.

Note: not enough to say things just about eigenvalues

(b) The eigenvectors are $v = [-i, 1]^T$ and $w = [i, 1]^T$. Note that $u_0 = [1, 0]^T = (-iv + iw)/2$. Therefore,

$$u_{1000} = C^{1000}u_0 = C^{1000}(-iv + iw)/2 = -\frac{i}{2}(1 + ik)^{1000}v + \frac{i}{2}(1 - ik)^{1000}w$$

Thus,

$$x_{1000} = -\frac{i}{2}(1 + ik)^{1000}(-i) + \frac{i}{2}(1 - ik)^{1000}(i) = -\frac{(1 + ik)^{1000} + (1 - ik)^{1000}}{2}$$

Similarly,

$$x_{1000} = -\frac{i}{2}(1 + ik)^{1000}(1) + \frac{i}{2}(1 - ik)^{1000}(1) = i\frac{(1 - ik)^{1000} - (1 + ik)^{1000}}{2}$$

(c) For backward Euler we have,

$$u_{j+1} = u_j + k \begin{bmatrix} & -1 \\ 1 & \end{bmatrix} u_{j+1} = \begin{bmatrix} 1 & k \\ -k & 1 \end{bmatrix}^{-1} u_j$$

The iteration matrix has eigenvalue/vector pairs,

$$\left(\frac{1}{1+ik}, \begin{bmatrix} -i \\ 1 \end{bmatrix} \right), \qquad \left(\frac{1}{1-ik}, \begin{bmatrix} i \\ 1 \end{bmatrix} \right)$$

Both eigenvalues have modulus less than one, and the eigenvectors are orthogonal.

(d) For the trapezoid rule we have,

$$u_{j+1} = u_j + \frac{k}{2} \begin{bmatrix} & -1 \\ 1 & \end{bmatrix} (u_j + u_{j+1}) = \begin{bmatrix} 1 & k/2 \\ -k/2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -k/2 \\ k/2 & 1 \end{bmatrix} u_j$$

The iteration matrix has eigenvalue/vector pairs,

$$\left(\frac{4-k^2+4ik}{4+k^2}, \begin{bmatrix} -i \\ 1 \end{bmatrix} \right), \qquad \left(\frac{4-k^2-4ik}{4+k^2}, \begin{bmatrix} -i \\ 1 \end{bmatrix} \right)$$

Both eigenvalues have modulus exactly equal to one, and the eigenvectors are orthogonal.

(e) **Note:** left as an exercise to the reader

Winter 2011, Day 1, Problem 2

Consider the trapezoid method to solve the scalar equation $y' = f(t, y)$:

$$y_{n+1} = y_n + \frac{h}{2}\left(f(t_{n+1}, y_{n+1}) + f(t_n, y_n)\right)$$

Show that this method is convergent, find its order, and sketch its region of absolute stability. In particular, determine where the region of absolute stability intersects the real and imaginary axes.

We have local truncation error,

$$\begin{aligned}
\tau_n &= \frac{y(t_n+h)-y(t_n)}{h} - \frac{f(t_n+h, y(t_n+h)) + f(t_n, y(t_n))}{2} \\
&= \frac{y(t_n+h)-y(t_n)}{h} - \frac{y'(t_n+h)+y'(t_n)}{2} \\
&= y'(t_n) + y''(t_n)\frac{h}{2} + y'''(t_n)\frac{h^2}{3!} + \mathcal{O}(h^3) - y'(t_n) - \frac{1}{2}\left(y''(t_n)h + y'''(t_n)\frac{h^2}{2} + \mathcal{O}(h^3)\right) \\
&= -\frac{y'''(t_n)}{12}h^2
\end{aligned}$$

Therefore the method is consistent and second order accurate.

This is a LMM with characteristic polynomials,

$$\rho(\zeta) = \zeta - 1, \qquad \sigma(\zeta) = (\zeta+1)/2$$

We then have stability polynomial,

$$\pi(\zeta; z) = \rho(\zeta) - z\sigma(\zeta) = (1 - z/2)\,\zeta - (1 + z/2)$$

This has root,

$$\zeta = \frac{1 + z/2}{1 - z/2}$$

Since there is only a single root, the root condition is satisfied when $|\zeta| = 1$. This happens when,

$$\operatorname{Re}(z) \le 0$$

That is, the region of absolute stability is the entire left half plane. In particular, it contains the entire imaginary axis, and the part of the real axis, $(-\infty, 0]$.

In particular, $z = 0$ is contained in the region of absolute stability. This means the method is zero-stable.

Zero stability and consistency imply convergence.

Winter 2013 Day 1, Problem 3

Consider the advection equation $u_t + au_x = 0$ and the "skewed" upwind method,

$$U_j^{n+1} = U_{j+1}^n - \left(\frac{a\Delta t}{\Delta x} + 1\right)(U_{j+2}^n - U_{j+1}^n).$$

(a) Show that this method is first order accurate in space and time by computing the local truncation error or by showing that the error after 1 step is $\mathcal{O}(\Delta t^2 + \Delta x^2)$ when applied to a sufficiently smooth function.

(b) For what values of $\Delta t/\Delta x$ does this method reduce to an "exact" solver, in the sense that if $U_j^n = u(x_j, t_n)$ for all j at time n then $U_j^{n+1} = u(x_j, t_{n+1})$ at the next time as well?

(c) What restriction must be put on the time step Δt in terms of Δx in order for the CFL condition to be satisfied in the case $a > 0$? In the case $a < 0$?

(d) Show that the method is in fact stable provided the CFL condition is satisfied, i.e., with the bounds found in part (c).

Solution

(a) Write $h = \Delta x$ and $k = \Delta t$.

We have local truncation error,

$$\tau = \frac{U_j^{n+1} - U_{j+1}^n - U_{j+2}^n + U_{j-1}^n}{k} + \frac{a}{h}\left(U_{j+2}^n - U_{j-1}^n\right)$$

Define,

```
U[j_, n_] := Series[Series[u[x + j h, t + n k], {k, 0, 4}], {h, 0, 4}]
```

Compute LTE,

```
FullSimplify[(U[0, 1] - U[1, 0] + U[2, 0] - U[1, 0])/k + a/h (U[2, 0] - U[1,
    0]),  Assumptions -> {D[#, t] + a D[#, x] &[u[x, t]] == 0}]
```

This gives LTE,

$$\tau = \frac{1}{2}u_{tt}k + \frac{3}{2}au_{xx}h + \mathcal{O}(h^2 + k^2)$$

This proves the method is first order accurate in space and time.

(b) The exact solution satisfies $u(x, t + k) = u(x - ak, t)$.

Suppose $ak/h = -2$. Then $u(x_j, t_{n+1}) = u(x_j - ak, t_n) = u(x_j + 2h, t_n) = u(x_{j+2}, t_n) = U_{j+2}^n = U_j{+}^{n+1}$.

Similarly, suppose $ak/h = -1$. Then $u(x_j, t_{n+1}) = u(x_j - ak, t_n) = u(x_j + h, t_n) = u(x_{j+1}, t_n) = U_{j+1}^n = U_j^{n+1}$.

Therefore, if $ak/h = \pm 1$ then the method is exact.

(c) Note that the stencil depends on the value of U^n at x_{j+1}, and x_{j+2}. We therefore require $x_j + h \leq x_j - ak \leq x_j + 2h$. Equivalently, $ak/h \in [-2, -1]$. If $a > 0$ we require $k \in [-2h/a, -h/a]$ and if $a < 0$ we require $k \in [-h/a, -2h/a] = [k/|a|, 2h/|a|]$.

(d) We replace U_j^n by $g(\xi)^n e^{i\xi jh}$ to obtain,

$$g(\xi)^n e^{i\xi jh} = g(\xi)^n e^{i\xi(j+1)h} - \left(\frac{ak}{h}+1\right)\left(g(\xi)^n e^{i\xi(j+2)h} - g(\xi)^n e^{i\xi(j+1)h}\right)$$

Dividing by $g(\xi)^n e^{i\xi jh}$ we obtain,

$$g(\xi) = e^{i\xi h} - \left(\frac{ak}{h}+1\right)\left(e^{2i\xi h} - e^{i\xi h}\right) g(\xi) = e^{i\xi h}\left(1 - \left(\frac{ak}{h}+1\right)\left(e^{i\xi h}-1\right)\right)$$

The part in the parenthesis is is a circle of radius $|ak/h+1|$ centered at $2+ak/h$.

This circle is contained in the unit circle when $ak/h \in [-2,-1]$. Therefore, whenever the CFL condition is satisfied, $|g(\xi)| \leq 1$ so the method is stable.

Summer 2017, Day 1, Problem 2

Let $A \in \mathbb{C}^{n\times n}$ be any $n \times n$ matric and let $v \in \mathbb{C}^n$ be a nonzero n-vector.

(a) Show that if $v \in \operatorname{span}\{Av, A^2v\}$ then $v \in \operatorname{span}\{A^2v, A^3v\}$.
(b) Given an example to show that the converse is not necessarily true; that is $v \in \operatorname{span}\{A^2v, A^3v\}$ does *not* necessarily imply that $v \in \operatorname{span}\{Av, A^2v\}$.

Solution

(a) Suppose $v \in \operatorname{span}\{Av, A^2v\}$. That is, there are scalars a and b such that,

$$v = aAv + bA^2v$$

Then,

$$Av = aA^2v + bA^3v$$

Therefore,

$$v = a(aA^2v + bA^3v) + bA^2v = (a^2 + b)A^2v + abA^3v \in \operatorname{span}\{A^2v, A^3v\} \qquad \square$$

(b) Let,

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad v = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Then clearly $A^3v = v$ so $v \in \operatorname{span}\{A^2v, A^3v\}$. However, $Av = [0, 1, 0]^T$ and $A^2v = [0, 0, 1]^T$ so clearly $v \notin \operatorname{span}\{Av, A^2v\}$.

Winter 2017, Day 1, Problem 1

An undirected graph consists of N nodes and a set of edges that connect given nodes. The adjacency matrix for the graph, A, has entries $A_{ij} = 1$ if there is an edge between nodes i and j and $A_{ij} = 0$ otherwise. Note that $A_{ij} = A_{ji}$, because we say if there is an edge between nodes i and j, then there is also an edge between nodes j and i. We do not allow self-connections, so that diagonal entries A_{ii} are always 0.

The degree of a node is the number of edges that connect to it. A d-regular random graph is a graph of which every node has degree d, for some integer $d \le N$, but the edges are otherwise in random positions. For the adjacency matrix A of a d-regular, undirected random graph with N nodes, answer the following questions.

(a) Can A ever have negative eigenvalues? If yes, give an example, if not, explain why not.
(b) Can A ever have imaginary eigenvalues? If yes, give an example, if not, explain why not.
(c) Identify a vector v and value λ that is an eigenvector-eigenvalue pair for the adjacency matrix of every d-regular undirected random graph with N nodes.
(d) Give a lower bound on the eigenvalues of the adjacency matrix for any d-regular undirected random graph.

Solution

(a) Yes. Consider the complete graph on 2 nodes. This has adjacency matrix,

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

which has eigenvalues ±1.

(b) The adjacency matrix A of any undirected graph is real and symmetric, so all eigenvalues are real.

(c) The vector $v = [1, 1, \dots, 1]^T$ has eigenvalue d since the sum of each row is d (the number of edges).

(d) For any matrix, the spectral radius bounded above by any matrix norm. The infinity and one norms of the adjacency matrix of a d-regular graph are both d. Therefore $\rho(A) \le d$ so all eigenvalues are bounded below by d.

Practice 2010, Day 2, Problem 5

This is basically the same as Summer 2014, Day 2, Problem 3

Solution

Practice 2010, Day 3, Problem 1

Consider the pendulum equation,

$$\theta'' + \frac{g}{L}\sin\theta = 0$$

where θ denotes the angular position of the pendulum away from the downard vertical position.

(a) Rewriting the equation in the phase plane with $x = \theta$ and $y = \theta'$, analyze the dynamics (equilibrium points, stability, phase plane plot).

(b) Find a conserved quantity for the pendulum equation, and rewrite it in terms of your phase plane variables. Call this quantity $H(x, y)$.

(c) Let $g = 1$, $L = 1$. Solve the pendulum equation numerically, using your (self-programmed) numerical solver of choice. Use initial conditions $\theta(0) = \pi/2$, $\theta'(0) = 0$. Provide the solution at $t = n$, $n \in \{0, 1, \ldots, 20\}$.

(d) At each of these times, give the value of $H(x, y)$. Is it conserved? Discuss your results.

Solution

(a) **Note:** Not relevant

(b) Energy is conserved as is energy/mass. We therefore have a conserved quantity,

$$H(x, y) = -g\cos(x) + \frac{1}{2}L^2y^2$$

(c) We write this as a system of the form $u' = G(u)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} -\frac{g}{L}\sin x \\ y \end{bmatrix}$$

We will solve this using forward Euler. That is,

$$u(t_{n+1}) = u(t_n) + kf(u(t_n))$$

Note: maybe do bw euler for practice.

(d)

Winter 2011, Day 3, Problem 1

Note: This is very similar to Winter 2011, Day 2, Problem 5.

Consider the Helmholtz problem,

$$u_{xx} + u_{yy} + k^2 u = f(x, y) = (k^2 - 5\pi^2)\sin(\pi x)\sin(2\pi y)$$

with $u(x, y) = 0$ on the boundary of the unit square $0 \le x \le 1$, $0 \le y \le 1$.

(a) Find the exact solution to this Helmholtz problem.

(b) Solve the Helmholtz problem using the 5-point Laplacian (second-order finite difference) and the backslash as linear system solver.

(c) Plot the maximum nodal error as a function of h for $k = 5$, $k = 10$, and $k = 60$ (using log-log scale).

(d) Suppose we change the solver to Jacobi (say take 200 iterations of Jacobis method). Program this in your code using the matrix version of Jacobi. Derive the spectral radius of Jacobis iteration matrix in terms of h and k. Recall the eigenvalues λ_{pq} of the 5-point Laplacian are,

$$\frac{2}{h^2}(\cos(p\pi h) + \cos(q\pi h) - 2)$$

where $h = 1/(m + 1)$, $p = 1, 2, \dots, m$, $q = 1, 2, \dots, m$.

(e) If we fix $h = 1/21$, for what values of k will Jacobis method converge? Verify this in your code by trying $k = 5$, $k = 10$, and $k = 60$, and any other k values, say $k = 0$ for example, you deem appropriate.

Solution

(a) We have solution,

$$u(x, y) = \sin(\pi x)\sin(2\pi y)$$

(b)

(c)

(d)

(e)

Winter 2010, Day 1, Problem 5

Verify that $y(t) = t^2/4$ solves the initial value problem,

$$y' = \sqrt{y}, \qquad\qquad y(0) = 0$$

Apply Eulers method to this problem and explain why the approximation obtained differs from the solution $t^2/4$.

Solution

Clearly,

$$y'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{t^2}{4} = \frac{t}{2} = \sqrt{\frac{t^2}{4}} = \sqrt{y(t)}, \qquad\qquad y(0) = 0$$

Euler's method gives an approximation to y one some mesh $\{t_n = nk\}_{n=0}^{N}$ by computing,

$$\hat{y}_{n+1} = \hat{y}_n + k\sqrt{\hat{y}_n}$$

where $\hat{y}_n$ is meant to approximate $y(t_n)$, and $y_0 = 0$.

We get zero at every step when using Euler's method regardless of the mesh.

Note: More about uniqueness of solution

Over one step we have,

$$y(t+k) = y(t) + \int_t^{t+k} y'(s)\mathrm{d}s$$

We approximate this integral using a single Riemann sum. Therefore some error is introduced and our estimated solution diverges from the actual solution. Moreover, the errors compound.

More specifically, even assuming $\hat{y}_n = y_n$, in one step we introduce an error,

$$\begin{aligned} y(t_n + k) - \hat{y}_{n+1} &= y(t_n + k) - \hat{y}_n - k\sqrt{\hat{y}_n} \\ &= ky'(t) + \frac{k^2}{2}y''(t) + \mathcal{O}(k^3) - k\sqrt{y(t)} \\ &= -\frac{k^2}{2}\frac{1}{\sqrt{y(t)}} + \mathcal{O}(k^3) \end{aligned}$$

These errors are all in the same direction as $\sqrt{y(t)}$ does not chance signs for $t > 0$. Therefore the errors compound.

Winter 2011, Day 2, Problem 4

Suppose you wish to compute the discrete Fourier transform (DFT) $F = [F_0, F_1, \ldots, F_{N-1}]^T$ of a vector $f = [f_0, f_1, \ldots, f_{N-1}]^T$ defined by,

$$F_k = \sum_{j=0}^{N-1} e^{2\pi ijk/N} f_j, \qquad k = 0, 1, \ldots, N-1$$

Here $i = \sqrt{-1}$, and you may assume N is a power of 2.

(a) Suppose you do not know $f_0, \ldots, f_{N-1}$, but you do know the DFT $F^{(e)}$ of the even numbered terms and the DFT $F^{(o)}$ of the odd numbered terms; i.e.,

$$F_k^{(e)} = \sum_{j=0}^{N/2-1} e^{2\pi ijk/(N/2)} f_{2j}, \qquad F_k^{(o)} = \sum_{j=0}^{N/2-1} e^{2\pi ijk/(N/2)} f_{2j+1}$$

where $k = 0, 1, \ldots, N/2-1$. Explain how you can compute F from $F^{(e)}$ and $F^{(o)}$. Be sure to show how you determine the entries $F_{N/2}, \ldots, F_{N-1}$, as well as $F_0, \ldots, F_{N/2-1}$.

(b) About how many operations (additions, subtractions, multiplications, divisions) are required to compute F, given $F^{(e)}$ and $F^{(o)}$?
Suppose this process is repeated and the length $N/2$ transforms $F^{(e)}$ and $F^{(o)}$ are computed from the DFTs of their even and old entries. How many operations would be required for this computation? Suppose the process is repeated until one reaches vectors of length 1 (for which the DFT is the identity). About how many total operations would be required?

Solution

(a) First split the sum into odd and even terms as,

$$F_k = \sum_{j=0}^{N-1} e^{2\pi ijk/N} f_j = \sum_{j=0,2,\ldots}^{N-1} e^{2\pi ijk/N} f_j + \sum_{j=1,3,\ldots}^{N-1} e^{2\pi ijk/N} f_j$$

Now reindex as,

$$\begin{aligned} F_k &= \sum_{j=0}^{N/2-1} e^{2\pi i(2j)k/N} f_{2j} + \sum_{j=0}^{N/2-1} e^{2\pi i(2j+1)k/N} f_{2j+1} \\ &= \sum_{j=0}^{N/2-1} e^{2\pi ijk/(N/2)} f_{2j} + e^{2\pi ik/N} \sum_{j=0}^{N/2-1} e^{2\pi ijk/(N/2)} f_{2j+1} \end{aligned}$$

For $k = 0, \ldots, N/2-1$, observe that by our definitions of $F^{(e)}$ and $F^{(o)}$ we have,

$$F_k = F_k^{(e)} + e^{2\pi ik/N} F_k^{(o)}$$

For $k = N/2, \dots, N-1$, since $e^{2\pi ij} = 1$ for all $j \in \mathbb{Z}$,

$$\begin{aligned}
F_k &= \sum_{j=0}^{N/2-1} e^{2\pi ijk/(N/2)} f_{2j} + e^{2\pi ik/N} \sum_{j=0}^{N/2-1} e^{2\pi ijk/(N/2)} f_{2j+1} \\
&= \sum_{j=0}^{N/2-1} e^{2\pi ij} e^{2\pi ij(k-N/2)/(N/2)} f_{2j} + e^{2\pi ik/N} \sum_{j=0}^{N/2-1} e^{2\pi ij} e^{2\pi ij(k-N/2)/(N/2)} f_{2j+1} \\
&= \sum_{j=0}^{N/2-1} e^{2\pi ij(k-N/2)/(N/2)} f_{2j} + e^{2\pi ik/N} \sum_{j=0}^{N/2-1} e^{2\pi ij(k-N/2)/(N/2)} f_{2j+1} \\
&= F_k^{(e)} + e^{2\pi ik/N} F_k^{(o)}
\end{aligned}$$

(b) We must do n additions and multiplications (assuming we can evaluate $e^{2\pi ik/N}$ in $\mathcal{O}(1)$ time). This is true at every step since the number of pieces which need added doubles at each step while the length of each piece is cut in half.

We can repeat the bisection $\mathcal{O}(\log_2(n))$ times so the total cost is $\sim 2\log_2(n) = \mathcal{O}(n\log(n))$ (in comparison to the $\mathcal{O}(n^2)$ operations required to compute the definition directly).

Summer 2012, Day 1, Problem 3

Consider the following difference scheme:

$$u(x,t+k) - u(x,t) = \frac{\sigma k}{h^2}(u(x+h,t) - u(x,t) - u(x,t+k) + u(x-h,t+k))$$

(a) In the limit $h \to 0$, $k \to 0$, what equation is this scheme consistent with? Are there any conditions you need to impose on h and k for this to be true?
(b) Discuss the stability of this scheme.

Solution

(a) We define the local truncation error by,

$$\tau = \frac{u(x,t+k) - u(x,t)}{k} - \frac{\sigma}{h^2}(u(x+h,t) - u(x,t) - u(x,t+k) + u(x-h,t+k))$$

We compute this in Mathematica,

```
U[n_, j_] :=  Normal[Series[u[x + z n h, t + z j k], {z, 0, 3}]] /. {z -> 1}
```

```
FullSimplify[(U[0, 1] - U[0, 0])/   k - \[Sigma] /h^2 (U[1, 0] - U[0, 0] - U
    [0, 1] + U[-1, 1]),   Assumptions -> {h > 0, k > 0}] // ExpandAll
```

This gives the result,

$$\tau = u_t(x,t) - \sigma u_{xx}(x,t) + \frac{k}{h}\sigma u_{tx}(x,t) + \mathcal{O}(h+k)$$

The local truncation error goes to zero as $h, k \to 0$ if $k/h \to 0$ and we are solving,

$$u_t(x,t) - \sigma u_{xx}(x,t) = 0$$

(b) Von Neumann? **Note:** Come back and do this

Summer 2015, Day 2, Problem 1

Let A be a complex square matrix. A coneigenvalue of A is a number $\lambda \in \mathbb{C}$ such that,

$$A\overline{x} = \lambda x$$

for some nonzero vector x, where $\overline{x}$ denotes the complex conjugate of x. The vector x is called a coneigenvector.

(a) Show that if λ is a coneigenvalue of A then so is $e^{i\theta}$ for any $\theta \in [0, 2\pi)$. (Here $i = \sqrt{-1}$.)
(b) Show that if λ is a coneignevalue of A then $|\lambda|^2$ is an eigenvalue of $\overline{A}A$.
(c) Write down a 2 by 2 matrix that has no coneigenvalues and explain how you know this.

Solution

(a) Suppose λ is a coneigenavlue. Then there is some x such that,

$$A\overline{x} = \lambda x$$

Therefore,

$$A\overline{e^{-i\theta/2}x} = e^{i\theta/2}A\overline{x} = e^{i\theta/2}\lambda x = e^{i\theta}\lambda\left(e^{-i\theta/2}x\right)$$

This proves $e^{i\theta}\lambda$ is a coneigenvalue with coneigenvector $e^{-i\theta/2}x$.

(b) Similarly,

$$\overline{A}A\overline{x} = \overline{A}\lambda x = \overline{\overline{\lambda\overline{A}x}} = \overline{\overline{\lambda}A\overline{x}} = \overline{\overline{\lambda}\lambda x} = \lambda\overline{\lambda}\overline{x} = |\lambda|^2\overline{x}$$

This proves $|\lambda|^2$ is an eigenvalue of $\overline{A}A$ with eigenvector $\overline{x}$.

(c) We note that we would like A such that $\overline{A}A$ has all negative eignevalues. In this case, by the contrapositive of (b) thre can be no coneigenvalues.

Observe also that the eigenvectors of A cannot be real or A will have coneigenvalues equal to the eignevalues of A. Moreover, we need that.

We now construct A with

Let,

$$A = \begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}^{-1} = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$$

Then,

$$\overline{A}A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Since $\overline{A}A$ has negative eigenvalues, so A has no coneigenvalues.

Winter 2015, Day 1, Problem 2

Let A be a $n \times n$ matrix and defined $\cos(A)$ by the Taylor series,

$$\cos(A) = I - \frac{1}{2!}A^2 + \frac{1}{4!}A^4 - \frac{1}{6!}A^6 + \cdots$$

(a) Show that this series converges in $\|\cdot\|_\infty$ for any matrix.

(b) If A is diagonalizable, show how $\cos(A)$ can be expressed in terms of the eigenvalues and eigenvectors of A.

(c) In particular, calculate $\cos(A)$ for,

$$A = \frac{2\pi}{5}\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

(d) Write out explicitly the matrix $\cos(A)$ if A is a 2×2 Jordan block with eigenvalue λ.

Solution

(a) **Note:** Not Relevant

(b) Suppose $A = X\Lambda X^{-1}$. Then $A^n = X\Lambda^n X^{-1}$. Therefore the partial sums are,

$$\sum_{i=0}^{n}(-1)^n \frac{1}{(2n)!}A^{2n} = \sum_{i=0}^{n}(-1)^n \frac{1}{(2n)!}X\Lambda^{2n}X^{-1} = X\left(\sum_{i=0}^{n}(-1)^n \frac{1}{(2n)!}\Lambda^{2n}\right)X^{-1}$$

We now move the sum into each component so that the i, i entry of the quantity in the parenthesis is,

$$\sum_{i=0}^{n}(-1)^n \frac{1}{(2n)!}\lambda_i^{2n}$$

As $n \to \infty$, the i, i entry converges to $\cos(\lambda_i)$ so that,

$$\cos(A) = X\begin{bmatrix} \cos(\lambda_0) & & & \\ & \cos(\lambda_2) & & \\ & & \ddots & \\ & & & \cos(\lambda_n) \end{bmatrix}X^{-1}$$

(c) We can write,

$$A = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 2\pi & \\ & 0 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}^{-1}$$

Therefore,

$$\cos(A) = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} \cos(2\pi) & \\ & \cos(0) \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$$

(d) Suppose,

$$A = \begin{bmatrix} \lambda & 1 \\ & \lambda \end{bmatrix}$$

Then,

$$A^n = \left[\begin{array}{cc} \lambda^n & n\lambda^{n-1} \\ & \lambda^n \end{array} \right]$$

By definition the diagonal entries of $\cos(A)$ are then,

$$1 - \frac{1}{2!}\lambda^2 + \frac{1}{4!}\lambda^4 - \frac{1}{6!}\lambda^6 + \cdots = \cos(\lambda)$$

The $2, 1$ entry is cleary 0. Finally, the $1, 2$ entry is,

$$1 - \frac{1}{2!}(2\lambda) + \frac{1}{4!}(3\lambda^3) - \frac{1}{6!}(6\lambda^5) + \cdots = -\sin(\lambda)$$

Explicitly,

$$\cos(A) = \left[\begin{array}{cc} \cos(\lambda) & -\sin(\lambda) \\ & \cos(\lambda) \end{array} \right]$$

Winter 2015, Day 2, Problem 1

Let V be the function space consisting of all functions of the form $g(x) = c_1 + c_2x^2$ on $0 \leq x \leq 1$ where $c_1, c_2 \in \mathbb{R}$. Define the L_2 inner product as usual by $(u, v) = \int_0^1 u(x)v(x)\mathrm{d}x$ and let $\|\cdot\|$ be the corresponding norm.

(a) Determine an orthonormal basis for V.
(b) Let P denote the orthogonal projection operator from functions $L_2(0, 1)$ to the subspace V. Determine $\|P\|$ and $\|I - P\|$.
(c) Let $f(x) = x$ and determine the projection $g = Pf$.
(d) Plot the functions f and g on the same axes.

Solution

(a) Let $p = 1$ and $\tilde{q} = x^2$. These are clearly linearly independent elements of V. It is also clear that V has dimension at most 2 so these must span V. We do Gram-Schmidt. Indeed, define,

$$\hat{q} = q - \frac{\langle p, q\rangle}{\langle p, p\rangle} p = x^2 - \frac{1/3}{1} 1 = x^2 - \frac{1}{3}$$

Finally, we normalize,

$$q = \frac{1}{\langle \hat{q}, \hat{q}\rangle^{1/2}} \hat{q} = \frac{3\sqrt{5}}{2}\left(x^2 - \frac{1}{3}\right)$$

We easily verify that p and q are orthonormal.

(b) Since p and q are orthonormal, an orthogonal projector P onto their span is defined as,

$$Pf = \langle f, p\rangle\, p + \langle f, q\rangle\, q$$

Similarly, $I - P$ is defined as,

$$(I - P)f = f - \langle f, p\rangle\, p + \langle f, q\rangle\, q$$

All orthogonal projectors have norm 1 and both P and $I - P$ are orthogonal projectors.

(c) We have,

$$Pf = Px = \langle x, p\rangle\, p + \langle x, q\rangle\, q = \frac{3}{16}\left(1 + 5x^2\right)$$

(d) The plot looks reasonable and is not included.

Winter 2015, Day 2, Problem 2

(a) Let A and B be real $n \times n$ matrices and define the commutator $C = AB - BA$. Show that trace$(C) = 0$.

(b) Suppose C has all diagonal entries equal to 0. Find a diagonal matrix X and a matrix Y such that C is the commutator for X and Y: $C = XY - YX$.

Solution

(a) Note that,

$$(AB)_{i,i} = \sum_{k=1}^{n} A_{i,k}B_{k,i}, \qquad\qquad (BA)_{i,i} = \sum_{k=1}^{n} B_{i,k}A_{k,i}$$

Therefore,

$$\sum_{i=1}^{n} C_{i,i} = \sum_{i=1}^{n}\sum_{k=1}^{n}(A_{i,k}B_{k,i} - B_{i,k}A_{k,i}) = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{i,k}B_{k,i} - \sum_{k=1}^{n}\sum_{i=1}^{n} A_{k,i}B_{i,k} = 0$$

(b) We assume we can write $C = XY - YX$ for some X diagonal. Then,

$$C_{i,j} = \sum_{k=1}^{n} X_{i,k}Y_{k,j} - Y_{i,k}X_{k,j} = X_{i,i}Y_{i,j} - Y_{i,j}X_{j,j} = (X_{i,i} - X_{j,j})Y_{i,j}$$

Therefore one choice of X and Y is to pick,

$$X_{i,i} = i, \qquad\qquad Y_{i,j} = \frac{C_{i,j}}{i-j}, \qquad\qquad Y_{i,i} = \text{anything}$$

Summer 2016, Day 1, Problem 2

Consider the predictor-corrector scheme,

$$\begin{aligned}\hat{U}_{n+2} &= U_{n+1} + hf(t_{n+1}, U_{n+1}),\\ U_{n+2} &= U_n + hf(t_n, U_n) + hf(t_{n+2}, \hat{U}_{n+2})\end{aligned}$$

for solving the ODE initial value problem $u' = f(t,u)$, $u(0) = u_0$, using time step size h. Assume that f is C^∞ in both variables.

(a) Determine the order of the local truncation error of the method.

(b) Suppose this scheme is applied to the test equation $u' = \lambda u$. Which, if any, of the values $h\lambda = -1, -2, -3$ lie in the *region of absolute stability* of the method (i.e., values of $h\lambda$ for which $U_n \to 0$ as $n \to \infty$, for all initial values U_0 and U_1). Justify your answer.

Solution

(a) The local truncation error is defined as,

$$\frac{u(t_{n+2}) - u(t_n)}{2h} - \frac{1}{2}(f(t_n, u(t_n)) + f(t_{n+2}, u(t_{n+1}) + hf(t_{n+1}, u(t_{n+1}))))$$

We first compute how much $\hat{u} = u(t_{n+1}) + hf(t_{n+1}, u(t_{n+1}))$ varies from $u(t_{n+2})$. Indeed, making the substitution $u' = f(t,u)$,

$$u(t_{n+2} - h) + hu'(t_{n+2} - h) = u(t_{n+2}) - \frac{1}{2}u''(t_{n+2})h^2 + \mathcal{O}(h^3)$$

Write $\Delta = -u''(t_{n+2})h^2/2 + \mathcal{O}(h^3)$ and compute,

$$f(t_{n+2}, \hat{u}) = f(t_{n+2}, u(t_{n+2}) + \Delta) = f(t_{n+2}, u(t_{n+2})) + f_u(t_{n+2}, u(t_{n+2}))\Delta + \mathcal{O}(\Delta^2)$$

Therefore,

$$\begin{aligned}f(t_{n+2}, \hat{u}) &= u'(t_{n+2}) + f_u(t_{n+2}, u(t_{n+2}))\Delta + \mathcal{O}(\Delta^4)\\ &= u'(t_n) + 2hu''(t_n) + f_u(t_{n+2}, u(t_{n+2}))\Delta + \mathcal{O}(\Delta^2)\\ &= u'(t_n) + 2hu''(t_n) - f_u(t_{n+2}, u(t_{n+2}))u''(t_{n+2})h^2/2 + \mathcal{O}(h^4)\end{aligned}$$

Finally,

$$u(t_{n+2}) - u(t_n)2h = u'(t_n) + u''(t_n)h + \frac{2}{3}u'''(t_n)h^2 + \mathcal{O}(h^3)$$

Putting this all together we have,

$$\tau = \frac{2}{3}u'''(t_n)h^2 + \frac{1}{4}f_u(t_{n+2}, u(t_{n+2}))u''(t_{n+2})h^2$$

Consider the equation $u' = \lambda u$. Then $f_u = \lambda$ and $u''' = \lambda u''$. This term does not disappear.

The method is therefore second order accurate.

(b) Applied to the test equation, writing $z = h\lambda$, we have,

$$\begin{aligned} \hat{U}_{n+2} &= U_{n+1}h\lambda U_{n+1} = (1+z)U_{n+1} \\ U_{n+2} &= U_n + h\lambda U_n + h\lambda \hat{U}_{n+2} = (1+z)U_n + z(1+z)U_{n+1} \end{aligned}$$

We can write this as,

$$U_{n+2} - aU_{n+1} - bU_n = 0, \qquad a = z(1+z), \qquad b = (1+z)$$

This is a linear difference equation with characteristic polynomial,

$$\rho(x) = x^2 - ax - b = \left(x - \frac{1}{2}\left(a - \sqrt{a^2 - 4b}\right)\right)\left(x - \frac{1}{2}\left(a + \sqrt{a^2 - 4b}\right)\right)$$

We need the root condition to be satisfied.

If $z = -1$ then $a, b = 0$ so $x_1 = x_2 = 0$ so the root condition is satisfied.

If $z = -2$ then $a = 2$, $b = -1$ so $x_1 = x_2 = 1$ so the root condition is not satisfied.

If $z = -3$ then $a = 6$, $b = -2$ so $x_1 = 3 - \sqrt{7}$, $x_2 = 3 + \sqrt{7} > 1$ so the root condition is not satisfied.

Therefore $z = -1$ is in the region of absolute stability, $z = -2$ is on the boundary, and $z = -3$ is on the exterior.

Summer 2011, Day 3, Problem 2

In his book on symmetric eigenvalue problems, B. Partlett proves the result:

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix, y a nonzer vector in $\mathbb{R}^n$, θ a real number, and r the residual vector,

$$r = Ay - y\theta$$

If α is the eigenvalue of A closest to θ, where $Az = z\alpha$ and $\|z\| = 1$ then,

$$|\theta - \alpha| \le \frac{\|r\|}{\|y\|}, \qquad\qquad |\sin\angle(y,z)| \le \frac{1}{\min_{\lambda_i \neq \alpha} |\lambda_i - \alpha|}\frac{\|r\|}{\|y\|}$$

where λ_i is an eigenvalue of A and $\|\cdot\|$ is the Euclidian norm.

Derive similar estimates for the eigenvalue problem,

$$\begin{bmatrix} K & B \\ B^T & 0 \end{bmatrix} x = \lambda \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} x$$

where the matrices K and M belong to $\mathbb{R}^{n\times n}$ and are symmetric positive definite. The matrix B belongs to $\mathbb{R}^{n\times p}$, $(p < n)$ and is full rank.

Solution

Note: Not really sure what they mean "similar bounds". I don't think I've understood the problem

---

Winter 2017, Day 3, Problem 1

Let $A = \{a_{ij}\}$ be a real n by n matrix.

(a) Suppose,

$$\|A\|_F \leq \|A + tI\|_F \quad \forall t \in \mathbb{R}$$

where $\|\cdot\|_F$ denotes the Frobenius norm: $\|A\| = \left(\sum_{i,j=1}^n a_{ij}^2\right)^{1/2}$. Show that the trace of A is 0.

If you dont see how to do this, write a code to test the result numerically. Perhaps you can get some insight from your numerical experiments.

(b) Suppose,

$$\|A\|_2 \leq \|A + tI\|_2 \quad \forall t \in \mathbb{R}$$

where $\|\cdot\|_2$ is the spectral norm; i.e., the largest singular value. Show that the left and right singular vectors of A corresponding to the largest singular value are orthogonal to each other.

If you dont see how to do this, write a code to test the result numerically. Perhaps you can get some insight from your numerical experiments.

Solution

(a) Suppose that $\|A\|_F \leq \|A + tI\|_F$ for all t. Then $\|A\|_F^2 \leq \|A + tI\|_F^2$ for all t.

Define $f(t) = \|A + tI\|_F^2 - \|A\|_F^2$. Then $f(t) \geq 0$ for all t and since $f(0) = 0$ we must have $f'(0) = 0$

Now note that,

$$f(t) := \|A + tI\|_F^2 - \|A\|_F^2 = \sum_{i=1}^n (a_{i,i} + t)^2 + \sum_{i=1}^n \sum_{j \neq i} a_{i,j}^2 - \sum_{i,j=1}^n a_{i,j}^2 = \sum_{i=1}^n (a_{i,i} + t)^2 - a_{i,i}^2$$

Therefore,

$$0 = f'(0) = \left[\sum_{i=1}^n 2(a_{i,i} + t)\right]_{t=0} = \sum_{i=1}^n 2a_{i,i} = 2\operatorname{tr}(A)$$

This proves the result. □

(b) Suppose that $\|A\|_2 \leq \|A + tI\|_2$ for all t. Then $\|A\|_2^2 \leq \|A + tI\|_2^2$ for all t.

Let v_t be a unit vector such that $\|(A + tI)v_t\|_2 = \|A + tI\|_2$, i.e. v_t is the first right singular vector of $A + tI$. Note that $v_0 = v$, the first right singular vector of A.

Then,

$$\|A + tI\|_2^2 = \|(A + tI)v_t\|_2^2 \geq \|A\|_2^2 \geq \|Av_t\|_2^2$$

Define,

$$\begin{aligned} g(t) &= \|(A + tI)v_t\|_t^2 - \|Av_t\|_2^2 \\ &= \langle Av_t, Av_t\rangle + 2t\langle Av_t, v_t\rangle + t^2\langle v_t, v_t\rangle - \langle Av_t, Av_t\rangle \\ &= 2t\langle Av_t, v_t\rangle + t^2 \end{aligned}$$

As in (a), by hypothesis we have $g(t) \geq 0$, and since $g(0) = 0$ we have $g'(0) = 0$. That is,

$$g'(0) = \left[2 \langle Av_t, v_t \rangle + 2t \frac{\mathrm{d}}{\mathrm{d}t} \langle Av_t, v_t \rangle + 2t \right]_{t=0} = 2 \langle Av_0, v_0 \rangle = 0$$

Again, $v_0 = v$, the first right singular vector, and $Av = u$, the first left singular vector. Therefore we have $\langle Av, v \rangle = \langle \sigma u, v \rangle = \sigma \langle u, v \rangle$, where σ is the largest singular value of A.

This proves the result. ☐

Practice 2010, Day 1, Problem 3

Consider the ordinary differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -100y$$

Your unsophisticated friend is using the forward Euler method to solve this equation over the time interval $[0, 5]$. He observes a discrete approximation to the solution $y(t)$ whose absolute value increases over time. Is this possible? If so, characterize the time step that he is using. Now answer the same question using the backward Euler method.

Solution

Forward Euler gives,

$$U_{n+1} = U_n + kf(U_n) = U_n + k(-100U_n) = (1 - 100k)U_n$$

The absolute value of the solution will increase in time if $|1 - 100k| > 1$. This happens if $k \in (0, 1/50)$.

Backward Euler gives,

$$U_{n+1} = U_n + kf(U_{n+1}) = U_n + k(-100U_{n+1}) = (1 + 100k)^{-1}U_n$$

The absolute value of the solution will increase in time if $|1 + 100k|^{-1} > 1$. This never happens for a positive time step, and we do not care about negative timesteps.

Winter 2012, Day 2, Problem 3

For this question: make sure to attempt all parts, even if youre stuck on an earlier part. Not all parts depend on all previous parts. Please make sure to indicate which part is answered where.

Consider the scheme,

$$u_{n+1} = u_n + hf(t_n + (1-\theta)h, \theta u_n + (1-\theta)u_{n+1})$$

for solving the ODE $u' = f(t,u)$. Here u_n and u_{n+1} are meant to approximate $u(t_n)$ and $u(t_{n+1}) = u(t_n + h)$, rrespectively.

(a) For all $\theta \in [0,1]$, find the order of this scheme.
(b) Determine for which $\theta \in [0,1]$ the scheme is convergent.
(c) For $\theta_0 = 0$ and $\theta_1 = 1$ determine the stability domain of the scheme.
(d) Consider the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u_1 \\ u_1 \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} \epsilon^2 u_1(u_1^2+u_2^2) \\ -u_2(u_1^2+u_2^2) \end{bmatrix},$$

where $\epsilon > 0$ is a parameter. Let $\epsilon = 0.5$. Find the equilibrium points of this system and determine their linear stability. Draw a phase portrait that is qualitatively consistent with your findings. This plot should be big and approximately to scale.

(e) Choose a suitable value of θ and using the scheme above, produce a plot of the curve $(u_1(t), u_2(t))^T$ with initial conditions $(-1,-1)^T$ in the phase plane. If this initial condition is in the basin of attraction of one of the equilibrium points, run the scheme until you get "reasonably" close to that equilibrium point.
(f) Discuss how your numerical results from (e) agree or disagree with your linear stability results from (d).

Solution

(a)

(b)

(c)

(d)

Summer 2017, Day 2, Problem 4

On Day 1, you should have found that the finite difference scheme,

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

is a consistent and stable method for the advection equation

$$u_t + au_x = 0, \qquad -\infty < x < \infty, \qquad t > 0$$

provided k/h and h/k remain bounded as $k, h \to 0$.

(a) Show that this difference scheme is a second-order in space and first-order in time approximation for the advection-diffusion equation,

$$u_t + au_x = \epsilon u_{xx}, \qquad \epsilon = \frac{h^2}{2k}$$

(b) nstead of an infinite domain, assume a finite domain $-1 \leq x \leq 1$ with periodic boundary condtions $u(-1,t) = u(1,t)$. how that the above difference scheme is the forward Euler method applied to the system of ODEs resulting from discretizing equation (1) in space. Write out the system of ODEs in the form $U'(t) = A_\epsilon U(t)$ and show that the eigenvalues of A_ϵ line in the left half-plane. [Hint: It may help to recall that all n by n circulant matrices have the same eigenvectors: $v_j = (1, \omega_j, \omega_j^2, \ldots, \omega_h^{n-1})^T$, $j = 0, 1, \ldots, n-1$, where $\omega_j = \exp(2\pi ij/n)$.]

Solution

(a)

(b)

Summer 2015, Day 2, Problem 3

Let A be a $n \times n$ matrix of the form,

$$A = \frac{1}{4}\begin{bmatrix} 2 & 1 & & & & 1 \\ 1 & 2 & 1 & & & \\ & 1 & 2 & 1 & & \\ & & 1 & 2 & 1 & \\ & & & 1 & 2 & 1 \\ 1 & & & & 1 & 2 \end{bmatrix}$$

Note that the diagonal entries are $+2$, not -2, so this is not a difference operator, but has similar eigenstructure.

Supopse we chose starting vector $y^{[0]} \in \mathbb{R}^n$ and we define $y^{[k+1]} = Ay^{[k]}$ for $k = 0, 1, 2, \ldots$.

(a) Do you expect this to converge to a limit as $k \to 0$ in general? How does it depend on the starting vector $y^{[0]}$?
(b) In cases when it does converge, what can you say about the rate of convergence in general?
(c) Take $n = 10$ and $y^{[0]} = [1, 2, \ldots, 10]^T$. What is the limit $\hat{y}$ in this case? Estimate by analysis if possible, not (only) experimentation, how many iterations m would be needed to guarantee $\left\| y^{[m]} - \hat{u} \right\|_\infty < 10^{-8}$.
(d) Now let B be the $n \times n$ matrix of the form,

$$B = \frac{1}{2}\begin{bmatrix} 0 & 1 & & & & 1 \\ 1 & 0 & 1 & & & \\ & 1 & 0 & 1 & & \\ & & 1 & 0 & 1 & \\ & & & 1 & 0 & 1 \\ 1 & & & & 1 & 0 \end{bmatrix}$$

What can you say about the convergence of the iteration $y^{[k+1]} = By^{[k]}$ for $k = 0, 1, 2 \ldots$? In particular, why are the cases n even and n odd different?

Solution

(a) For $j = 1, 2, \ldots, 10$, the eigenvectors for A are,

$$v_i^j = \exp\left(2\pi i j k / n\right), \qquad k = 1, 2, \ldots, n$$

and the corresponding eigenvalues are,

$$\lambda_j = \cos\left(2\pi j/n\right)/2 + 1/2$$

Therefore the spectral radius of A is 1, and there are no noreal eigenvalues with modulus 1. This means we expect convergence.

Let $v = v^n$ be the eigenvector corresopnding to an eigenvalue of 1. We expect the solution to converge to $(\langle y, v\rangle / \langle v, v\rangle)v$. If $\langle y, v\rangle = 0$ then this will be zero. Otherwise it will be nonzero.

(b) The rate of convergence is governed by the second largest eigenvalue. In each step all components of $y^{[k]}$ in directions other than the dominant eigenvector will be decreased by at lest the second largest eigenvalue.

(c) We have $v = v^n = [1, 1, \ldots, 1]^T$ so $\hat{y} \left\langle v, y^{[0]} \right\rangle / \langle v, v \rangle \, v = 55/10v = 11/2v$.

Note that $A^k y^{[0]}$ projected onto v is always equal to $\hat{y}$. The next largest stretch direction is v^{n-1} which is decreased by a factor of λ_{n-1} each iteration.

Therefore, lazily bounding the projection of $y^{[0]}$ onto any eigenvector by $\left\| y^[0] \right\|_\infty = 10$,

$$\begin{aligned} \left\| y^{[k]} - \hat{y} \right\|_\infty &= \left\| \sum_{\ell \neq n} \lambda_\ell^k \frac{\left\langle y^{[0]}, v_\ell \right\rangle}{\langle v_\ell, v_\ell \rangle} v_\ell \right\|_\infty \\ &\leq \sum_{\ell \neq n} |\lambda_\ell|^k \left| \frac{\left\langle y^{[0]}, v_\ell \right\rangle}{\langle v_\ell, v_\ell \rangle} \right| \|v_\ell\|_\infty \\ &\leq n |\lambda_{n-1}|^k \left\| y^{[0]} \right\|_\infty \\ &= 10n \left(\frac{1}{2} \left(\frac{1}{4} \left(\sqrt{5} + 1 \right) + 1 \right) \right)^k \end{aligned}$$

This is below 10^{-8} in about $k = \log_{10}(10^{-8}/100) / \log_{10}(100\lambda_{n-1}) \approx 138$ iterations.

(d) The eigenvalues are now,

$$\lambda_j = \cos(2\pi j/n)$$

Again the spectral radius is 1. However, if n is even, then ±1 are both eigenvalues, so if $y^{[0]}$ has non-zero projection onto the eigenvector correspoding to eigenvalue -1 the method will not converge. However, if n is even, there is no eigenvalue -1 so the method will always converge.

Summer 2016, Day 1. Problem 1

Let,

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 \\ 0 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ ) \\ 0 \\ 0 \end{bmatrix}$$

(a) Find the orthogonal projection of b onto the span of $\{Ab, A^2b, A^3b\}$.

(b) Find a 4 by 4 matrix M whose eigenvalues are all ones such that $b \perp \text{span}(\{Mb, M^2b, M^3b\})$. [Hint: You might want to use a companion matrix, which has the form,

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ * & * & * & * \end{bmatrix}$$

where the $*$'s can be nonzero.] Could such a matrix be diagonalizable? Explain why or why not.

Solution

(a) We compute,

$$Ab = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 7 \end{bmatrix}, \qquad A^2b = \begin{bmatrix} 0 \\ 0 \\ 42 \\ 70 \end{bmatrix}, \qquad A^3b = \begin{bmatrix} 0 \\ 126 \\ 630 \\ 1078 \end{bmatrix}$$

Then clearly $b \notin \text{span}(\{Ab, A^2b, A^3b\})$. Therefore the orthogonal projection to this space is 0.

(b) Note that,

$$(x-1)^4 = x^4 - 4x^3 + 6x^2 - 4x + 1$$

Define,

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 4 & -6 & 4 \end{bmatrix}$$

Then M has all eigenvalues equal to 1.

Finally,

$$Mb = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \qquad M^2b = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -4 \end{bmatrix}, \qquad M^3b = \begin{bmatrix} 0 \\ -1 \\ -4 \\ -10 \end{bmatrix}$$

Then clearly $b \perp \text{span}(\{Mb, M^2b, M^3b\})$.

If a matrix has all 1s as eigenvaues and is diagonalizable it is the identity as $XIX^{-1} = I$. Clearly b is in the span of Mb so this is not possible.

Summer 2016, Day 2, Problem 1

Let $A \in \mathbb{R}^{m \times m}$ be a given matrix whose real singular value decomposition is $A = U\Sigma V^T$. Show that $Q = UV^T$ is the nearest orthogonal matrix to A in the Frobenius norm:

$$\arg \min_{Q^T Q = I} \|A - Q\|_F = UV^T$$

Discuss circumstances in which the nearest orthogonal matrix is or is not unique. In particular, consider the case where the singular values of A are all positive and distinct.

Solution

Recall that the Frobenius norm of a matrix is the square root of the sum of the squares of the singular values.

Recall also that a unitary linear transform does not stretch any vectors (wrt. Euclidian distance).

Our intuition is roughly: Minimizing $\|A - Q\|_F$ means we want to "cancel off" as much of the stretching which A does. However, in each direction we can "cancel off" the stretching from A and the unitary transform from Q as much as possible. By the triangle inequality we want the directions of the images of the right singular vectors of A to line up with those of Q. In this case we will have $Q = UV^T$.

Since the Frobenius norm is unitarily invariant, calling $\tilde{Q} = U^T QV$ we have,

$$\|A - Q\|_F = \left\|U(\Sigma - U^T QV)V^T\right\|_F = \left\|\Sigma - \tilde{Q}\right\|_F$$

Note that if we can solve the new minimization problem for some unitary $\tilde{Q}$ then we can obtain $Q = U\tilde{Q}V^T$, the solution to the original problem. We therefore look to solve the simpler problem,

$$\arg \min_{\tilde{Q}^T \tilde{Q} = I} \left\|\Sigma - \tilde{Q}\right\|_F^2 = \sum_{i=1}^{n} (\sigma_i - \tilde{q}_{i,i})^2 + \sum_{i=1}^{n} \sum_{j \neq i} \tilde{q}_{i,j}^2 = \sum_{i=1}^{n} \sigma_i^2 - 2\sigma \tilde{q}_{i,i} + \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{q}_{i,j}^2$$

Now note that for any unitary matrix,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{q}_{i,j}^2 = \left\|\tilde{Q}\right\|_F^2 = n$$

Therefore we seek $\tilde{q}_{i,i}$ minimizing,

$$\sum_{i=1}^{n} \sigma_i^2 - 2\sigma_i \tilde{q}_i$$

Since $\tilde{Q}$ is unitary the columns are normal. This means $|\tilde{q}_{i,j}| \leq 1$ for all i, j. Since $\sigma_i \geq 0$, we pick $\tilde{q}_{i,i} = 1$ to minimize the above sum. This forces $\tilde{q}_{i,j} = 0$ for all $i \neq j$.

Now note that this choice of $\tilde{Q} = I$ is in fact unitary. Therefore $\tilde{Q} = I$ minimizes $\left\|\Sigma - \tilde{Q}\right\|_F$ over unitary $\tilde{Q}$. Then $Q = U\tilde{Q}V^T = UIV^T = UV^T$ minimizes the original problem.

The uniequness of the nearest orthogonal matrix is the same as the uniquness of UV^T. In particular, this means that if all the singular values are distinct then the solution is unique. If there are repeat singular values, then we can move around the columns of U and V a bit, and so the solution may not be unique.

Summer 2015, Day 3, Problem 1

Consider the ODE BVP from Day 1 with a particular choice of ϵ:

$$\epsilon x^2 y'' = y, \qquad y(0) \text{ bounded}, \qquad y(1) = 1, \epsilon = 0.1$$

Implement a finite difference method using $N = 50$ grid points (your choice how to distribute them across the computational domain) for solving this problem and show that your numerical solution at least visually matches the exact solution (which has the form $y(x) = xp$ for some p). Discuss how your method enforces the boundary condition at $x = 0$ as stated in the problem. [Note that it would be easier to solve the problem numerically if the boundary condition were $y(0) = 0$. You could instead solve this modified problem for partial credit.]

Solution

Summer 2015, Day 1, Problem 2

Let A denote a real sequence with m rows.

(a)
(b)
(c) Derive another expression for the matrix exponential by applying the forward Euler method to the equation $y'(t) = Ay$, $y(0) = y_0$, whose solution at time $t = 1$ is $y(1) = e^A y_0$. That is, using step size $h = 1/N$, write the approximate solution y_N as an N-th degree polynomial in A times the initial vector $y0$ and show that as $N \to \infty$, $y_N \to e^A y_0$.
(d) Suppose the 2-step method,

$$y_{n+2} = y_{n+1} + \frac{k}{2}(3Ay_{n+1} - Ay_n)$$

with step size k is applied to the equation $y'(t) = Ay$, $y(0) = y_0$. Determine if the method is convergent and find the order of accuracy and the leading term in the local truncaiton error.

Solution

(a)

(b)

(c) Forward Euler gives,

$$y^{[k+1]} = y^{[k]} + hAy^{[k]} = (I + hA)y^{[k]} = (I + A)^{[k+1]}y^{[0]}$$

Therefore,

$$y(1) \approx y^{[N]} = (I + hA)^N y^{[0]} = \left(I + \frac{1}{N}A\right)^N y(0)$$

Forward Euler is convergent, so as $N \to \infty$, $y^{[N]} \to e^A y_0$, the exact solution at $hN = 1$.

(d) This is a linear multistep method. However we will just use Mathematica. The local truncation error is of the form,

$$\tau = \frac{y(t+2h) - y(t+h)}{k} - \frac{3f(y(t+h)) - f(y(t))}{2} = \frac{y(t+2h) - y(t+h)}{k} - \frac{3y'(t+h) - y'(t)}{2}$$

We expand about t to find,

$$\tau = \frac{5}{12}y'''(t)h^2 + \mathcal{O}(h^3)$$

Therefore the method is convergent, is second order accurate, and has leading term $(5/12)y'''(t)h^2$.

Summer 2015, Day 2, Problem 4

Consider the family of 2-stage Runge-Kutta methods of the form

$$\begin{aligned} Y_1 &= U^n + \alpha k f(U^n) + \beta k f(Y_1) \\ U^{n+1} &= U^n + \gamma k f(Y_1) \end{aligned}$$

for solving the ODE initial value problem $u'(t) = f(u(t))$ with step size k. A particular method is obtained by choosing $\alpha, \beta, \gamma \in \mathbb{R}$.

(a) What can you say about the consistency, stability, and covergence of methods from this family (for various values of the parameters α, β, γ)?
(b) By considering the behavior of these methods on the test problem $y_0 = \lambda y$, what can you say about the order of accuracy for various choices of the parameters?
(c) Are there methods from this family that would be reasonable methods to use for solving stiff ODEs? Justify your answer.
(d) If you were forced to use one of these methods to solve an ODE, what choice of α, β, γ would you use and why? (Your answer might depend on what sort of ODE it is.)

Solution

(a) Note that,

$$Y_1 = U^n + \mathcal{O}(k)$$

Making the substitution $f(u(t)) = u'(t)$, the local truncation error for the method is,

$$\frac{u(t+h) - u(t)}{k} - \gamma f(u(t) + \mathcal{O}(k)) = u'(t) + \mathcal{O}(k) - \gamma u'(t) + \mathcal{O}(k)$$

Therefore, the method is consistent when $\gamma = 1 + \mathcal{O}(k)$.

Alternate method for convergence using general RK method facts

This is a Runge-Kutta method of the form,

$$U^{n+1} = U^n + k\sum_{j=1}^{r} b_j f(Y_j), \qquad Y_i = U^n + k\sum_{j=1}^{r} a_{i,j} f(Y_j)$$

For convenience define $c_i = \sum_{j=1}^{r} a_{i,j}$. In our case we have,

$$r = 2, \qquad \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma \end{bmatrix}, \qquad \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \alpha & \beta \end{bmatrix}$$

The condition for consistency is that,

$$\gamma = b_1 + b_2 = 1$$

The condition for second order convergence is,

$$\alpha + \beta = \gamma(\alpha + \beta) = b_1 c_1 + b_2 c_2 = \frac{1}{2}$$

We have,

$$Y_1 = U^n + \alpha k\lambda U^n + \beta k\lambda Y_1 = \frac{1+k\lambda\alpha}{1-k\lambda\beta}U^n$$

$$U^{n+1} = U^n + k\gamma\lambda Y_1 = \left(1 + k\lambda\gamma\frac{1+k\lambda\alpha}{1-k\lambda\beta}\right)U^n$$

Taking $z = k\lambda$, this is in the form, $U^{n+1} = R(z)U^n$ where,

$$R(z) = 1 + \gamma z\frac{1+\alpha z}{1-\beta z}$$

The stability region is points $z = k\lambda$ such that $|R(z)| \leq 1$.

Clearly $z = 0$ is such a point so the method is zero stable and therefore convergent when it is consistent.

(b) From above we have computed,

$$U^{n+1} = U^n + k\lambda\gamma\frac{1+k\lambda\alpha}{1-k\lambda\beta}U^n$$

The local truncation error is then,

$$\tau = \frac{u(t+k)-u(t)}{k} - \lambda\gamma\frac{1+k\lambda\alpha}{1-k\lambda\beta}u(t)$$

We expand this in Mathematica using,

```
FullSimplify[ ReplaceRepeated[ Series[(u[t + k] - u[t])/k - \[Lambda] \[Gamma]
     ((1 + k \[Lambda] \[Alpha])/(1 - k \[Lambda] \[Beta])) u[t], {k, 0, 3}],
    {u’[t] -> \[Lambda] u[t], D[u’[t] -> \[Lambda] u[t], t], D[u’[t] -> \[
    Lambda] u[t], {t, 2}], D[u’[t] -> \[Lambda] u[t], {t, 3}]}]]
```

This gives the result,

$$\begin{aligned}\tau = (\gamma-1)\lambda u(t) \\ + \frac{1}{2}(1-2(\alpha+\beta)\gamma)\lambda^2 u(t)k \\ + \frac{1}{6}(1-6(\beta(\alpha+\beta)\gamma)\lambda^3 u(t)k^2 \\ + \frac{1}{24}(1-24\beta^2(\alpha+\beta)\gamma)\lambda^4 u(t)k^3 + \mathcal{O}(k^4)\end{aligned}$$

We therefore see that we require $\gamma = 1$ for consistency, $(\alpha+\beta)\gamma = 1/2$ for second order convergence, and $\beta(\alpha+\beta)\gamma = 1/6$ for third order convergence.

(c) When $\alpha = 0$, $\beta = 1$, $\gamma = 1$ this is close to backward Euler, a method which is ok a solving stiff equations. In some sense this is the “most implcit” method of the given form. However, Since it is always implicit if $\beta \neq 0$ it will probably be ok for stiff equations (or about as good as a 2nd order method can be).

(d) Clearly if the ODE is stiff we want to pick $\beta \neq 0$ so that the method is implicit. If the ODE is simple I would use backward Euler for simplicity. Otherwise, chosing $\alpha = \beta = 1/2$ means the first equation is the midpoint method, and then the seconde quation is forward Euler applied to that. That might be decent since the midpoint method is decent. If the ODE is really simple, then we can take $\beta = 0$ and get an explicit method so we do not need to use root finders.

Summer 2012, Day 3, Problem 1

Let P be an orthogonal projection operator in $\mathbb{R}^n$, $n > 1$. In other words, every $u \in \mathbb{R}^n$ can be written uniquely as $u = v + w$, where v is in R_P, the range of P, and w is in the null space of P which is the orthogonal complement of R_P. Thus $Pu = v$, $(I - P)u = w$ and $Pv = v$. Let $P$and Q be two orthogonal projections in $\mathbb{R}^n$. What can you say about the 2-norm of $P - Q$?

Solution

Trivially we have,

$$0 \leq \|P - Q\| \leq 2$$

since both P and Q have norm 1.

However, it is clear that the upper bound cannot be attained, since this would require P and Q point in opposite directions. This only happens when they act on a vector which is in the null space of both of them, which clearly would not be the vector in the direction of max stretch.

We claim that,

$$0 \leq \|P - Q\| \leq 1$$

Clearly this upper bound can be attained with $Q = 0$.

Fix P. Suppose R_Q is orthogonal to R_P. For any unit vector v, the distance between Pv and Qv is one.

A slight perturbation of R_Q knocking it out of orthogonality with R_P will decrease $\|P - Q\|$ as now Qv has a bit in the direction of P.

Therefore having R_Q orthogonal to R_P at least gives a local minimum to $\|P_Q\|$.

Note: GLOBAL because convexity or something?

Let $u = v + w$ as above. Then,

$$\begin{aligned}\|(P - Q)u\|^2 &= \langle Pu, Pu\rangle + \langle Qu, Qu\rangle - 2\langle Pu, Qu\rangle \\ &= \langle v, v\rangle + (\langle Qv, Qv\rangle + \langle Qw, Qw\rangle + 2\langle Qv, Qw\rangle) - 2\langle v, Qu\rangle \\ &= \langle v, v\rangle + (\langle Qv, Qv\rangle + \langle Qw, Qw\rangle + 2\langle Qv, Qw\rangle) - 2\langle v, Qv\rangle - 2\langle v, Qw\rangle\end{aligned}$$

Since Q is orthogonal, $\langle Qw, Qv\rangle = \langle v, Qw\rangle$ and $\langle Qv, Qv\rangle = \langle v, Qv\rangle$. Therefore,

$$\|(P - Q)u\|^2 = \langle v, v\rangle - \langle Qv, Qv\rangle + \langle Qw, Qw\rangle$$

We seek Q and u maximizing $\|(P - Q)u\|$ such that $\|u\| = 1$.

Its obvious that with $Qw = w$ and $Qv = 0$ and $u = w$ that we can reach 1.

Note: STILL STUCK ON HIGHER