Computer Science Department
Stanford University
Comprehensive Examination in Numerical Analysis
Fall 2003

1. **Vector and Matrix Norms [8 pts]**

The following definitions hold for the norm and condition number of a *rectangular* $m \times n$ matrix A with respect to a specific matrix norm

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \qquad cond(A) = \left(\max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \right) \cdot \left(\min_{x \neq 0} \frac{\|Ax\|}{\|x\|} \right)^{-1}$$

Given the singular value decomposition $A = U \Sigma V^T$ of the matrix A (where U and V are orthogonal and Σ is the $m \times n$ diagonal matrix containing the singular values of A)

i. [3 pts] Prove that $\|A\|_2 = \|\Sigma\|_2$ using the definition above

ii. [5 pts] Prove that

$$\|A\|_2 = \sigma_{max} \qquad \text{and} \qquad cond_2(A) = \sigma_{max} / \sigma_{min}$$

where σ_{max} is the largest singular value of A and σ_{min} the smallest one.

Solutions

i. For any $x \in R^n \setminus \{\vec{0}\}$ we have

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{(Ax)^T Ax}{x^T x} = \frac{x^T A^T A x}{x^T x} = \frac{x^T \left(U \Sigma V^T\right)^T U \Sigma V^T x}{x^T x} = \frac{x^T V \Sigma^T U^T U \Sigma V^T x}{x^T x} = \frac{x^T V \Sigma^T \Sigma V^T x}{x^T V V^T x} =$$

$$= \frac{\left(\Sigma V^T x\right)^T \Sigma V^T x}{\left(V^T x\right)^T V^T x} = \frac{\|\Sigma V^T x\|_2^2}{\|V^T x\|_2^2} \Rightarrow \frac{\|Ax\|_2}{\|x\|_2} = \frac{\|\Sigma V^T x\|_2}{\|V^T x\|_2}$$

Since the mapping $x \mapsto V^T x$ is an isomorphism on $R^n \setminus \{\vec{0}\}$ (its inverse is simply $x \mapsto Vx$) we have

$$\max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \neq 0} \frac{\|\Sigma V^T x\|_2}{\|V^T x\|_2} = \max_{V^T x \neq 0} \frac{\|\Sigma (V^T x)\|_2}{\|V^T x\|_2} = \max_{y \neq 0} \frac{\|\Sigma y\|_2}{\|y\|_2}$$

thus $\|A\|_2 = \|\Sigma\|_2$

ii. Let σ_i, $i=1,\ldots,n$ be the singular values of A, forming the diagonal of the matrix Σ. Let $\sigma_k=\sigma_{\max}$ be the largest and $\sigma_l=\sigma_{\min}$ the smallest among them. Then

$$\left.\begin{aligned}\max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\max_{x\neq 0}\frac{\|\Sigma x\|_2}{\|x\|_2}=\max_{x\neq 0}\sqrt{\frac{\sum_i \sigma_i^2 x_i^2}{\sum_i x_i^2}}\le \max_{x\neq 0}\sqrt{\frac{\sum_i \sigma_{\max}^2 x_i^2}{\sum_i x_i^2}}=\sigma_{\max}\\ \max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\max_{x\neq 0}\frac{\|\Sigma x\|_2}{\|x\|_2}\ge\frac{\|\Sigma e_k\|_2}{\|e_k\|_2}=\frac{\|\sigma_k e_k\|_2}{1}=\sigma_k=\sigma_{\max}\end{aligned}\right\}\Rightarrow \max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\sigma_{\max}$$

and

$$\left.\begin{aligned}\min_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\min_{x\neq 0}\frac{\|\Sigma x\|_2}{\|x\|_2}=\min_{x\neq 0}\sqrt{\frac{\sum_i \sigma_i^2 x_i^2}{\sum_i x_i^2}}\ge \min_{x\neq 0}\sqrt{\frac{\sum_i \sigma_{\min}^2 x_i^2}{\sum_i x_i^2}}=\sigma_{\min}\\ \min_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\min_{x\neq 0}\frac{\|\Sigma x\|_2}{\|x\|_2}\le\frac{\|\Sigma e_l\|_2}{\|e_l\|_2}=\frac{\|\sigma_l e_l\|_2}{1}=\sigma_l=\sigma_{\min}\end{aligned}\right\}\Rightarrow \min_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\sigma_{\min}$$

Therefore, using the definition we have $\|A\|_2=\sigma_{\max}$ and $cond_2(A)=\sigma_{\max}/\sigma_{\min}$

2. **Differential Equations [10 pts]**

i. [4 pts] Given a square matrix A whose eigenvalues have negative real parts, show that the matrix $I-A$ is invertible and the eigenvalues of $B=(I-A)^{-1}(I+A)$ are given by the formula $\lambda_i^B=\frac{1+\lambda_i^A}{1-\lambda_i^A}$, where λ_i^A are the eigenvalues of A

ii. [6 pts] Consider the vector ordinary differential equation $\vec{y}'=f(x,\vec{y})$ and the implicit trapezoidal method for solving it:

$$\vec{y}_{k+1}=\vec{y}_k+h\frac{f(x_k,\vec{y}_k)+f(x_{k+1},\vec{y}_{k+1})}{2}$$

Prove that this method is unconditionally stable when applied to the model vector ODE $\vec{y}'=A\vec{y}$ for a matrix A whose eigenvalues have negative real parts (that is, show that $\|\vec{y}_k\|\to 0$ as $k\to\infty$, regardless of the value of the step size h)

Solutions

i. The matrix $I-A$ is singular if and only if $\det(I-A)=0$. But this implies that $\lambda=1$ is an eigenvalue of A, which is not the case since all the eigenvalues of A have negative real parts. Furthermore, we have

$$\det\left[(I-A)^{-1}(I+A)-\lambda I\right]=0 \Leftrightarrow \det\left[(I-A)^{-1}(I+A)-\lambda(I-A)^{-1}(I-A)\right]=0 \Leftrightarrow$$
$$\Leftrightarrow \det\left[(I-A)^{-1}\right]\det\left[(I+A)-\lambda(I-A)\right]=0 \Leftrightarrow \det\left[(I+A)-\lambda(I-A)\right]=0 \Leftrightarrow$$
$$\Leftrightarrow \det\left[(\lambda+1)A-(\lambda-1)I\right]=0 \Leftrightarrow \det\left[A-\frac{\lambda-1}{\lambda+1}I\right]=0$$

Therefore, the eigenvalues of A and $B=(I-A)^{-1}(I+A)$ are associated as follows

$$\lambda_i^A=\frac{\lambda_i^B-1}{\lambda_i^B+1} \Leftrightarrow \lambda_i^B=\frac{1+\lambda_i^A}{1-\lambda_i^A}$$

ii. Application of the trapezoidal rule in the model equation yields

$$\vec{y}_{k+1}=\vec{y}_k+\frac{h}{2}(A\vec{y}_k+A\vec{y}_{k+1}) \Rightarrow$$
$$\Rightarrow\left(I-\frac{h}{2}A\right)\vec{y}_{k+1}=\left(I+\frac{h}{2}A\right)\vec{y}_k$$
$$\Rightarrow \vec{y}_{k+1}=\left(I-\frac{h}{2}A\right)^{-1}\left(I+\frac{h}{2}A\right)\vec{y}_k$$
$$\Rightarrow \vec{y}_k=\left[\left(I-\frac{h}{2}A\right)^{-1}\left(I+\frac{h}{2}A\right)\right]^k\vec{y}_0$$

The method is stable if and only if the spectral radius of the matrix $\left(I-\frac{h}{2}A\right)^{-1}\left(I+\frac{h}{2}A\right)$ is less than 1. Using the result of (i) the eigenvalues of the matrix above are given as $\hat{\lambda}_i=\frac{1+\frac{h}{2}\lambda_i^A}{1-\frac{h}{2}\lambda_i^A}$ which all have magnitude less than 1 for $\mathrm{Re}\{\lambda_i^A\}<0$ (since each λ_i^A lies closer to -1 than to 1 on the complex plane) regardless of the value of h. Therefore, the trapezoidal rule is unconditionally stable for this model equation.

3. **Numerical Quadrature [12 pts]**

Consider a real function f that is differentiable on an interval $[a,b]$

i. [3 pts] Find a *quadratic* polynomial $g(x)$ that approximates $f(x)$ on $[a,b]$ in that $f'(a)=g'(a)$, $f'(b)=g'(b)$ and $f\left(\frac{a+b}{2}\right)=g\left(\frac{a+b}{2}\right)$ [Hint : Consider expressing $g(x)$ as a quadratic polynomial of $\left(x-\frac{a+b}{2}\right)$]

ii. [2 pts] Define a numerical quadrature rule for $\int_a^b f(x)dx$ by integrating the interpolant $g(x)$ on $[a,b]$

iii. [3 pts] Prove that this integration scheme has degree of accuracy equal to 3.

iv. [2 pts] Define the corresponding composite quadrature rule for $\int_a^b f(x)dx$ we obtain by subdividing $[a,b]$ into the n sub-intervals $\left[a+k\frac{b-a}{n}, a+(k+1)\frac{b-a}{n}\right]$, $k=0,1,\ldots,n-1$ and applying the basic integration rule on each of them

v. [2 pts] Consider the composite rule of (iv), the composite midpoint rule and the composite Simpson's rule. Under which circumstances would you prefer to use each one of them?

Solutions

i. Let $g(x)=c_2\left(x-\frac{a+b}{2}\right)^2+c_1\left(x-\frac{a+b}{2}\right)+c_0$. Using the given constraints we have

$$\left\{\begin{array}{c} g'(a)=f'(a) \\ g'(b)=f'(b) \\ g\left(\frac{a+b}{2}\right)=f\left(\frac{a+b}{2}\right)\end{array}\right\}\Rightarrow\left\{\begin{array}{c} c_2(a-b)+c_1=f'(a) \\ c_2(b-a)+c_1=f'(b) \\ c_0=f\left(\frac{a+b}{2}\right)\end{array}\right\}\Rightarrow\left\{\begin{array}{c} c_2=\frac{f'(b)-f'(a)}{2(b-a)} \\ c_1=\frac{f'(a)+f'(b)}{2} \\ c_0=f\left(\frac{a+b}{2}\right)\end{array}\right\}$$

Therefore

$$g(x)=\frac{f'(b)-f'(a)}{2(b-a)}\left(x-\frac{a+b}{2}\right)^2+\frac{f'(a)+f'(b)}{2}\left(x-\frac{a+b}{2}\right)+f\left(\frac{a+b}{2}\right)$$

ii. We have

$$\int_a^b f(x)dx\approx\int_a^b g(x)dx=\int_a^b\left[c_2\left(x-\frac{a+b}{2}\right)^2+c_1\left(x-\frac{a+b}{2}\right)+c_0\right]dx=c_2\frac{(b-a)^3}{12}+c_0(b-a)$$

$$\Rightarrow\int_a^b f(x)dx\approx(b-a)f\left(\frac{a+b}{2}\right)+\frac{(b-a)^2}{24}[f'(b)-f'(a)]$$

iii. The interpolant used approximates exactly polynomials of degree up to 2, thus the degree of accuracy is at least 2. We also have

$$\int_a^b x^3 dx = (b-a)\left(\frac{a+b}{2}\right)^3 + \frac{(b-a)^2}{24}\left[3b^2 - 3a^2\right] = \frac{(b-a)(a+b)^3}{8} + \frac{(b-a)^3(a+b)}{8} =$$

$$= \frac{(b-a)(a+b)}{8}\left[(a+b)^2 + (a-b)^2\right] = \frac{(b^2-a^2)}{4}\left(b^2+a^2\right) = \frac{b^4-a^4}{4}$$

which is the exact result. To show that the degree of accuracy is <u>exactly</u> 3, we give the counterexample $f(x) = x^4$ on the interval $[-a, a]$

$$\int_{-a}^{a} x^4 dx = 2a \cdot 0^4 + \frac{(2a)^2}{24}\left[4a^3 + 4a^3\right] = \frac{4}{3}a^5$$

which is <u>not</u> the exact result ($\frac{2}{5}a^5$). Therefore the method is third order accurate

iv. The composite rule is

$$\int_a^b f(x)dx = \sum_{k=0}^{n-1} \int_{a+k\frac{b-a}{n}}^{a+(k+1)\frac{b-a}{n}} f(x) \approx$$

$$\approx \sum_{k=0}^{n-1}\left\{\frac{b-a}{n} f\left(a + (2k+1)\frac{b-a}{2n}\right) + \frac{(b-a)^2}{24n^2}\left[f'\left(a + (k+1)\frac{b-a}{n}\right) - f'\left(a + k\frac{b-a}{n}\right)\right]\right\}$$

$$= \left\{\frac{b-a}{n}\sum_{k=0}^{n-1} f\left(a + (2k+1)\frac{b-a}{2n}\right)\right\} + \frac{(b-a)^2}{24n^2}\left[f'(b) - f'(a)\right]$$

v. If we know the *exact* value of $f'(a)$ and $f'(b)$, the rule we proved in (iv) is third order accurate while only slightly more complex than the midpoint rule and should be preferred. Note that this wouldn't work if we tried to approximate $f'(a)$ and $f'(b)$ from nearby values of f, since this approximation would have an $O(h)$ error leading to an $O(h^3)$ error in the integration formula (same as the midpoint rule).

If we don't know $f'(a)$ and $f'(b)$ and third order accuracy is desired, Simpson's rule is the only option. Nevertheless, if first order accuracy is sufficient (for example if f is very smooth or if the discretization step h is already very small) the midpoint rule is simpler and requires much fewer floating point operations.