math 600

L#4

Sept 8, 2011

Matrix norms, condition number etc

Recall the following facts from linear algebra for $x \in \mathbb{R}^n$ we have several norms $\|x\|$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|$$

unit square

$$\|x\|_2 = \sum_{i=1}^{n} |x_i|^2$$

unit ball

Matrix norms could be introduced as operator norms since $A \in \mathbb{R}^{n \times n}$ is an operator $\mathbb{R}^n \to \mathbb{R}$

$$\|A\| = \sup\{\|Ax\| : x \in \mathbb{R}^n \ \|x\| = 1\} = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|}{\|x\|}$$

this is called matrix norm subordinate to the vector norm $\|\cdot\|$.

<u>Remark:</u> Frobenius norm $\|A\|_F = \left(\sum |a_{ij}|^2\right)^{1/2}$ is not an operator norm, i.e. it is not subordinate to a vector norm

One can deduce various properties of the matrix norm

$$\|\lambda A\| = |\lambda| \, \|A\|$$

$$\|A+B\| \leq \|A\| + \|B\|$$

$$\|AB\| \leq \|A\| \, \|B\|$$

<u>Condition Number</u>: Motivation and importance

Consider the system $Ax=b$

$$\varkappa(A) = \|A\| \, \|A^{-1}\|$$

Assume A invertible and the systems $Ax=b$ & $A\tilde{x}=\tilde{b}$ where $b-\tilde{b}$ is "small" (in some norm). Say $\frac{\|b-\tilde{b}\|}{\|b\|} < \varepsilon = 10$

What can be say about

$$\frac{\|x-\tilde{x}\|}{\|x\|} \; ?$$

E.g. if we compute $\tilde{b}$ with 6 decimal digits how accurate $\tilde{x}$ will be assuming exact calculations in the Gauss process? Or what is the worst error $\|x-\tilde{x}\|$?

$$\|x-\tilde{x}\| = \|A^{-1}b - A^{-1}\tilde{b}\| \leq \|A^{-1}\| \, \|b-\tilde{b}\|$$

$$\leq \|A^{-1}\| \frac{\|b-\tilde{b}\|}{\|b\|} \underbrace{\|b\|}_{Ax} \leq \|A^{-1}\| \, \|A\| \, \|x\| \frac{\|b-\tilde{b}\|}{\|b\|}$$

$$\Rightarrow \quad \frac{\|x-\tilde{x}\|}{\|x\|} \leq \underbrace{\|A^{-1}\| \, \|A\|}_{} \frac{\|b-\tilde{b}\|}{\|b\|} \equiv \varkappa(A) \frac{\|b-\tilde{b}\|}{\|b\|}$$

<u>Theorem</u>: Let $e = x - \tilde{x}$ $\quad r = b - A\tilde{x}$ - residual
error

Then

$$\frac{1}{\varkappa(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \varkappa(A) \frac{\|r\|}{\|b\|}$$

<u>Proof</u>. The second inequality has been already proven. Now the lower one

$$\|r\| \|x\| = \|A\tilde{x} - b\| \|x\| = \|A(\tilde{x} - x)\| \|A^{-1}b\|$$

$$\leq \|A\| \|e\| \|A^{-1}\| \|b\|$$

$$\Rightarrow \frac{1}{\varkappa(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \qquad \boxtimes$$

This is not very reliable estimate, but it says that the error of the r.h.s. could be magnified by the condition number compared with the error of the solution.

<u>Def</u>. We say that the system is ill-conditioned if the condition number is very large.

Hilbert matrix $\qquad a_{ij} = \dfrac{1}{i+j-1} \qquad i,j = 1, \dots, n$

$$\varkappa(A) \cong e^{cn}$$

The idea of iterative improvement (refinement)

$A \approx LU$ $\quad (LU)^{-1}$ is easy to compute with

$LUx = b$ $\quad x^{(0)}$ which differs from x

Then compute the residual $r^{(0)} = -Ax^{0} + b$. Maybe using this residual we can improve the solution $x^{(0)}$ by correcting it

$$e^{(0)} = \overset{B}{(LU)^{-1}} r^{(0)} \Rightarrow x^{(1)} = x^{(0)} + e^{(0)}$$

iterative improvement

Now we formalize this into the following: Let

$x^{(0)} = Bb$ where B is easy to compute with

obviously if $B = A^{-1}$ we are done

But A^{-1} is very costly. We take

B to be some cheaply computed

Obviously, $x^{(0)}$ is not a solution. Therefore we can find

$$r^{(0)} = b - Ax^{(0)} \qquad e^{(0)} = Br^{(0)} \qquad x^{(1)} = x^{(0)} + e^{(0)}$$

Now we make this in many steps

$$(1) \qquad \boxed{r^{(k)} = b - Ax^{(k)} \quad \Big| \quad x^{(k+1)} = x^{(k)} + Br^{(k)} \qquad k = 0, 1, \ldots}$$

with the hope that this in a few steps will (produce) converge to x

$$x^{(k+1)} = x^{(k)} + B(b - Ax^{(k)}) = (I - BA)x^{(k)} + Bb$$

Assume for a while that $x^{(k)} \to y$

$$\cancel{y} = \cancel{y} - BAy + Bb \qquad BAy = Bb$$

$$B \text{ nonsingular} \Rightarrow Ay = b \quad y = x$$

If the process converges, then it converges to the desired solution of the system $Ax = b$

$$x = x - BAx + Bb$$

$$\boxed{e^{(k+1)} = (I - BA)e^{(k)} \quad \text{error equation}}$$

$$\|e^{(k)}\| \xrightarrow[k \to \infty]{} 0$$

$$\|e^{(k+1)}\| \le \underbrace{\|I - BA\|}_{\delta} \|e^{(k)}\| \le \dots \le \delta^{k+1} \|e^{(0)}\|$$

if $\delta < 1$ then the process converges

$$\boxed{e^{(k)} = (I - BA)^k e^{(0)}}$$

$e^{(0)}$ is some fixed vector

Obviously $\|I - BA\| = \delta < 1$ is sufficient for the convergence of the iterative method (1).

<u>Theorem:</u> Assume that $\|I-BA\|<1$ for some matrix norm subordinate to a vector norm. Then

$$x^{(0)}=Bb \qquad x^{(k+1)}=(I-BA)x^{(k)}+Bb\ ,\ k=0,1,\dots$$

converges to the solution of $Ax=b$.

<u>Note:</u> We did not assume $Ax=b$ has a solution. In fact $\|I-BA\|<1$ ensures us that A is a nonsingular matrix. Indeed if A were singular then $Ac=0$ $c\neq 0$

Then

$$1>\|I-BA\|=\sup_{\substack{x\in\mathbb{R}^n\\ x\neq 0}}\frac{\|(I-BA)x\|}{\|x\|}\geq\frac{\|(I-BA)c\|}{\|c\|}=$$

$$=\frac{\|c\|}{\|c\|}=1 \quad\Rightarrow\quad 1>1$$

contradiction

<u>Proof:</u> $x^{(k+1)}-x=x^{(k)}-x+B(\underset{Ax}{b}-Ax^{(k)})$ $\qquad \|I-BA\|=\delta<1$

$$e^{(k+1)}=e^{(k)}-BAe^{(k)}=(I-BA)e^{(k)}$$

$$\|e^{(k+1)}\|\leq\delta^{k+1}\|e^{(0)}\| \qquad e^{(k)}\to 0 \quad k\to\infty$$

One more thing

$$x^{k+1}-x=(I-BA)(x^k\underline{-x+x^{k+1}}-x^{k+1})$$

$$\|x^{k+1}-x\|\leq\delta\|x^k-x^{k+1}\|+\delta\|x^{k+1}-x\|$$

$$(1-\delta)\|x^{k+1}-x\|\leq\delta\|x^k-x^{k+1}\|$$

$$\boxed{\|x^{k+1}-x\|\leq\frac{\delta}{1-\delta}\|x^k-x^{k+1}\|}$$

important

Denoting by $Q = B^{-1} \sim A$ Q called
aim Splitting of A

$$x^{(k+1)} = x^{(k)} + B(b - Ax^{(k)})$$

$$Q(x^{k+1} - x^{k}) = b - Ax^{(k)} = r^{(k)}$$

$$Q\,x^{(k+1)} = (Q - A)\,x^{k} + b$$

we can look at Q as some piece of A
this is the reason Q is called splitting of A

Basic concept of standard iterative methods is to find a nonsingular splitting of A such that

(1) $x^{(k)}$ is easy to compute $\equiv Qx^{(k)} =$ date easy to solve

(2) $x^{(k)}$ converges rapidly to $x \equiv Q^{-1}$ good approx A^{-1}

we know that $\| I - Q^{-1}A \| = \delta < 1$ is sufficient
smaller δ provides faster convergence

More facts from linear algebra

① $\det(A)$ $\quad \det(AB) = \det(A)\det(B)$

② eigenvalues and eigenvectors of $A \in \mathbb{R}^{n\times n}$

$$\det(A-\lambda I) = (-1)\lambda^n + (a_{11}+\dots+a_{nn})\lambda^{n-1} + \dots + \det(A) = 0$$

characteristic polynomial of A

counting their multiplicities $\lambda_1, \dots, \lambda_n$ in the complex plane!

<u>Example:</u> $\det(I-\lambda I) = (1-\lambda)^n \qquad \lambda_1 = \lambda_2 = \dots = \lambda_n = 1$

Take λ_j $\quad A\psi_j = \lambda_j \cdot \psi_j \quad \psi_j$ eigenvector

One of the most difficult questions in LA is to answer how many linearly independent eigenvectors we can get?

$$D = \begin{bmatrix} d_{11} & & 0 \\ & \ddots & \\ 0 & & d_{nn} \end{bmatrix} \qquad \det(D-\lambda I) = \prod_{i=1}^{n}(d_{ii}-\lambda)$$

$$\boxed{\lambda_i = d_{ii} \quad i=1,\dots,n}$$

eigenvectors will be e_i

(λ_i, e_i) eigenpair

There is a class of matrices that this question is fully understood and solved, the class of symmetric matrices.

If A a symmetric matrix in $\mathbb{R}^{n\times n}$, then

1. it has real eigenvalues
2. it has full set of eigenvectors, i.e. $(\psi_1, \dots, \psi_n)$ span the space $\mathbb{R}^n$ (or form a basis for $\mathbb{R}^n$)

we can orthonormalize them so that $(\psi_i, \psi_j) = \delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$

$x = \sum_{j=1}^{n} c_j \psi_j \quad c_j = (x, \psi_j)$ "Fourier" representation

$\|x\|_2^2 = \sum c_j^2 \quad \ell_2$-norm

<u>Def</u>: The matrix W is unitary (complex elements) if $W^* = W^{-1}$, $\overline{W^T} \equiv W^*$

<u>Example</u>: $W = \begin{vmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix} \quad W^* = \begin{vmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{vmatrix}$

$$WW^* = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$