Mathematical Tripos Part IB: Easter 2001

Numerical Analysis – Lecture 3[1]

Band matrices. The matrix A is a *band matrix* if there exists an integer $r < n$ such that $A_{i,j} = 0$ for $|i-j| > r$, $i,j = 1,2,\ldots,n$. In other words, all the nonzero elements of A reside in a band of width $2r+1$ along the main diagonal. In that case, according to the statement from the end of the last lecture, $A = LU$ implies that $L_{i,j} = U_{i,j} = 0 \;\forall\; |i-j| > r$ and sparsity structure is inherited by the factorization.

In general, the expense of calculating an LU factorization of an $n \times n$ *dense* matrix A is $\mathcal{O}(n^3)$ operations and the expense of solving $A\boldsymbol{x} = \boldsymbol{b}$, provided that the factorization is known, is $\mathcal{O}(n^2)$. However, in the case of a banded A, we need just $\mathcal{O}(r^2 n)$ operations to factorize and $\mathcal{O}(rn)$ operations to solve a linear system. If $r \ll n$ this represents a very substantial saving!

3 Iterative methods for linear systems

3.1 Basic iterative schemes

Solution of $A\boldsymbol{x} = \boldsymbol{b}$ by factorization is frequently very expensive for large n, even if we exploit sparsity. An alternative is to use *iterative methods* which are very efficient. An example of an iterative scheme is to write $A = B - C$, where ***(1)*** B & C are $n \times n$ matrices; ***(2)*** B is nonsingular; ***(3)*** the system $B\boldsymbol{x} = \boldsymbol{c}$ is *easy to solve* and ***(4)*** the matrix C is somehow 'small' in comparison with B. We write the original system in the form $B\boldsymbol{x} = C\boldsymbol{x} + \boldsymbol{b}$ and consider solving it by iteration. Choose an arbitrary $\boldsymbol{x}_0 \in \mathbb{R}^n$ and define $\boldsymbol{x}_{m+1}$, $m = 0,1,\ldots$, by solving

$$B\boldsymbol{x}_{m+1} = C\boldsymbol{x}_m + \boldsymbol{b}. \tag{3.1}$$

Provided that B is, for example, banded, the solution of (3.1) is cheap (and the LU factorization of B can be re-used – an example of why the LU formalism is superior to Gaussian elimination). Often the sequence $\{\boldsymbol{x}_m\}_{m=0}^{\infty}$ converges to the solution of $A\boldsymbol{x} = \boldsymbol{b}$.

The Jacobi iteration. We write $A = A_{\mathrm{D}} + A_{\mathrm{L}} + A_{\mathrm{U}}$, where A_{L}, A_{D} and A_{U} are portions of A: subdiagonal (strictly lower triangular), diagonal and supdiagonal (strictly upper triangular), respectively. Suppose that no diagonal element of A is zero. The *Jacobi iteration* is

$$A_{\mathrm{D}}\boldsymbol{x}_{m+1} = -(A_{\mathrm{L}} + A_{\mathrm{U}})\boldsymbol{x}_m + \boldsymbol{b}, \qquad m = 0,1,\ldots. \tag{3.2}$$

The Gauss–Seidel iteration. In the above notation, it takes the form

$$(A_{\mathrm{D}} + A_{\mathrm{L}})\boldsymbol{x}_{m+1} = -A_{\mathrm{U}}\boldsymbol{x}_m + \boldsymbol{b}, \qquad m = 0,1,\ldots. \tag{3.3}$$

Note that $A_{\mathrm{L}} + A_{\mathrm{D}}$ is lower triangular, hence the solution of (3.3) is cheap.

3.2 Necessary and sufficient conditions for convergence

Suppose that A is nonsingular and denote by $\boldsymbol{x}^*$ the solution of $A\boldsymbol{x} = \boldsymbol{b}$. Having written $A = B - C$, we examine the iterative scheme (3.1). (Note that both (3.2) and (3.3) can be cast in this form.) Our goal is to identify conditions so that $\boldsymbol{x}_m \to \boldsymbol{x}^*$, regardless of the choice of $\boldsymbol{x}_0 \in \mathbb{R}^n$.

[1]Corrections and suggestions to these notes should be emailed to `a.shadrin@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartIB/Handouts.html`.

Subtract $B\boldsymbol{x}^* = C\boldsymbol{x}^* + \boldsymbol{b}$ from (3.1). This gives $B(\boldsymbol{x}_{m+1} - \boldsymbol{x}^*) = C(\boldsymbol{x}_m - \boldsymbol{x}^*)$, hence $B\boldsymbol{\varepsilon}_{m+1} = C\boldsymbol{\varepsilon}_m$, where $\boldsymbol{\varepsilon}_m := \boldsymbol{x}_m - \boldsymbol{x}^*$ is the error in the mth iterate. Since B is nonsingular (otherwise we cannot execute (3.1) in the first place), it follows that

$$\boldsymbol{\varepsilon}_{m+1} = H\boldsymbol{\varepsilon}_m = \cdots = H^{m+1}\boldsymbol{\varepsilon}_0, \quad m = 0, 1, \ldots \quad \text{where} \quad H = B^{-1}C \quad \text{is the } \textit{iteration matrix}.. \tag{3.4}$$

This indicates that the errors tend to zero as $m \to \infty$ (regardless of the choice of $\boldsymbol{x}_0$) provided that $\lim_{m\to\infty} H^m = O$. The next theorem establish a criterium for such a convergence.

We employ the notation $\rho(P)$ for the magnitude of the largest (in absolute value) eigenvalue of the $n \times n$ matrix P. The quantity $\rho(P)$ is called the *spectral radius* of the matrix P. (*Note:* Recall that, even if P is real, its eigenvalues might be complex.)

Theorem. We have $\lim\limits_{m\to\infty} H^m\boldsymbol{\varepsilon}_0 = 0$ for all $\boldsymbol{\varepsilon}_0 \in \mathbb{R}^n$ if and only if $\rho(H) < 1$.

Proof. We commence with the case $\rho(H) \geq 1$ and wish to demonstrate that, for some *real* $\boldsymbol{\varepsilon}_0$, the vector $\boldsymbol{\varepsilon}_m$ need not tend to $\mathbf{0}$. Let λ be an eigenvalue of (the real) H, real or complex, such that $|\lambda| = \rho(H)$, and let $\boldsymbol{w}$ be a corresponding eigenvector, i.e., $H\boldsymbol{w} = \lambda\boldsymbol{w}$. Denote the *Euclidean length* of $\boldsymbol{p} \in \mathbb{C}^n$ by $\|\boldsymbol{p}\| = \left(\sum_{k=1}^n |p_k|^2\right)^{1/2}$. Then the vector $H^m\boldsymbol{w} = \lambda^m\boldsymbol{w}$ is bounded away from zero, because

$$\|H^m\boldsymbol{w}\| = |\lambda|^m\|\boldsymbol{w}\| \geq \|\boldsymbol{w}\| \geq \max_k |\boldsymbol{w}_k| = \gamma > 0. \tag{3.5}$$

If $\boldsymbol{w}$ is real, we choose $\boldsymbol{\varepsilon}_0 = \boldsymbol{w}$, hence $\boldsymbol{\varepsilon}_m = H^m\boldsymbol{\varepsilon}_0 \geq \gamma$, and this cannot tend to zero.

If $\boldsymbol{w}$ is complex, then $\boldsymbol{w} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v}$ with $\boldsymbol{u}, \boldsymbol{v}$ some real vectors. But then at least one of the sequences $(H^m\boldsymbol{u})$, $(H^m\boldsymbol{v})$ does not tend to zero. For if both do, then also

$$H^m\boldsymbol{w} = H^m\boldsymbol{u} + \mathrm{i}H^m\boldsymbol{v} \to \mathbf{0},$$

and this contradicts (3.5). This completes the proof of the 'only if' part of the theorem.

Now we turn to the 'if' case, i.e., $\rho(H) < 1$, where we assume for simplicity that H possesses n linearly independent eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n$, say. Hence $H\boldsymbol{w}_j = \lambda_j\boldsymbol{w}_j$, $|\lambda_j| < 1$, $j = 1, 2, \ldots, n$. Linear independence means that every $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ can be expressed as a linear combination of the eigenvectors. Therefore, given $\boldsymbol{\varepsilon}_0 \in \mathbb{R}^n$, there exist $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{C}$ such that $\boldsymbol{\varepsilon}_0 = \sum_{j=1}^n \alpha_j\boldsymbol{w}_j$. Thus,

$$\boldsymbol{\varepsilon}_1 = H\boldsymbol{\varepsilon}_0 = \sum_{j=1}^n \alpha_j\lambda_j\boldsymbol{w}_j \qquad \text{and, by induction,} \qquad \boldsymbol{\varepsilon}_m = \sum_{j=1}^n \alpha_j\lambda_j^m\boldsymbol{w}_j$$

for all $m = 0, 1, \ldots$. Since $\rho(H) < 1$, it follows that $\lim_{m\to\infty} \boldsymbol{\varepsilon}_m = \mathbf{0}$, as required. □

Remark. The complete proof of the 'if' case exploits the so-called Jordan normal form of the matrix H, $H = SJS^{-1}$, where J is a block diagonal matrix consisting of the Jordan blocks

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix}, \qquad J_i \in \mathbb{R}^{n_i \times n_i}.$$

To prove that $J_i^m \to 0$ if $|\lambda_i| < 1$, one can split $J_i = \lambda_i I + P$, notice that $P^k = 0$ for $k \geq n_i$, and evaluate the terms of expansion $(\lambda_i I + P)^m$.

Application 1. If $A = A^{\mathrm{T}}$ is positive definite, then the Gauss–Seidel method converges [to a solution of $A\boldsymbol{x} = \boldsymbol{b}$ for any starting vector $\boldsymbol{x}_0$] (for then $\rho(H_{\mathrm{GS}}) < 1$).

Application 2. If both $A = A^{\mathrm{T}}$ and $2A_{\mathrm{D}} - A$ are positive definite, then the Jacobi method converges (for then $\rho(H_{\mathrm{J}}) < 1$ as well).