Dr A. Iserles
Easter 1997

Numerical Analysis – Lecture 3

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

General sparse matrices Factorization methods of sparse matrices depend on the exploitation of pivoting to minimize *fill-in*. There are modern efficient techniques to produce good pivoting strategies for general sparsity structures. They are based on *graph theory* and are well beyond the scope of this lecture course.

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.* An example of an iterative scheme is to write $A = B + C$, where B & C are $n \times n$, B is nonsingular, the system $B\boldsymbol{x} = \boldsymbol{c}$ is *easy to solve* and 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, (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 $A = A_{\mathrm{L}} + A_{\mathrm{D}} + A_{\mathrm{U}}$, where A_{L} is strictly lower triangular, A_{D} is diagonal and A_{U} is strictly upper triangular. 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

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

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 (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 $\boldsymbol{x}_0 \in \mathbb{R}^n$.
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 $B\boldsymbol{v}_{m+1} = -C\boldsymbol{v}_m$, where $\boldsymbol{v}_m := \boldsymbol{x}_m - \boldsymbol{x}^*$ is the error in the mth iterate. Since B is nonsingular (otherwise we cannot execute (3.1) in the first place), it follows that

$$\boldsymbol{v}_{m+1} = H\boldsymbol{v}_m, \quad m = 0, 1, \ldots \qquad \text{where} \qquad H := -B^{-1}C. \tag{3.4}$$

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 $\lim_{m\to\infty} \boldsymbol{x}_m = \boldsymbol{x}^*$ for all $\boldsymbol{x}_0 \in \mathbb{R}^n$ if and only if $\rho(H) < 1$.

Proof. We commence with the case $\rho(H) \geq 1$ and wish to demonstrate that $\boldsymbol{v}_m$ need not tend to $\mathbf{0}$. Let λ be an eigenvalue of H such that $|\lambda| = \rho(H)$ and let $\boldsymbol{w}$ be a corresponding eigenvector, $H\boldsymbol{w} = \lambda\boldsymbol{w}$. If $\boldsymbol{w}$ is real, we choose $\boldsymbol{x}_0 = \boldsymbol{x}^* + \boldsymbol{w}$, hence $\boldsymbol{v}_0 = \boldsymbol{w}$. It follows at once by induction that $\boldsymbol{v}_m = \lambda^m \boldsymbol{w}$, and this cannot tend to zero since $|\lambda| \geq 1$.

If $\lambda \in \mathbb{C} \setminus \mathbb{R}$ then $\boldsymbol{w}$ is complex. Moreover, also $\bar{\lambda} \neq \lambda$ is an eigenvalue and $\bar{\boldsymbol{w}}$ is its eigenvector (the bar denotes complex conjugation). Note that $\boldsymbol{w}$ and $\bar{\boldsymbol{w}}$ are linearly independent (otherwise they would have corresponded to the same eigenvalue). We denote the *Euclidean length* of $\boldsymbol{p} \in \mathbb{C}^n$ by

$$\|\boldsymbol{p}\| = \left\{ \sum_{k=1}^{n} |p_k|^2 \right\}^{1/2}.$$

Note that $\|\boldsymbol{p}\|$ is a continuous function of the components of $\boldsymbol{p}$. Hence, $\|z\boldsymbol{w} + \bar{z}\bar{\boldsymbol{w}}\|$ is a continuous function of the complex variable z. It is a consequence of the linear independence of $\boldsymbol{w}$ and $\bar{\boldsymbol{w}}$ and of the theorem that a continuous function attains its minimum in a closed interval that

$$\inf_{-\pi\leq\theta\leq\pi} \left\|\mathrm{e}^{\mathrm{i}\theta}\boldsymbol{w} + \mathrm{e}^{-\mathrm{i}\theta}\bar{\boldsymbol{w}}\right\| = \min_{-\pi\leq\theta\leq\pi} \left\|\mathrm{e}^{\mathrm{i}\theta}\boldsymbol{w} + \mathrm{e}^{-\mathrm{i}\theta}\bar{\boldsymbol{w}}\right\| = \nu,$$

say, is positive. By homogeneity, it is true for every $z \in \mathbb{C}$ that

$$\|z\boldsymbol{w} + \bar{z}\bar{\boldsymbol{w}}\| \geq \nu|z|. \tag{3.5}$$

We let $\boldsymbol{x}_0 = \boldsymbol{x}^* + \boldsymbol{w} + \bar{\boldsymbol{w}}$, hence $\boldsymbol{v}_0 = \boldsymbol{w} + \bar{\boldsymbol{w}}$. Note that everything in sight is real! We have by induction on (3.1) that

$$\boldsymbol{v}_m = \lambda^m \boldsymbol{w} + \bar{\lambda}^m \bar{\boldsymbol{w}}, \qquad m = 0, 1, \ldots.$$

Setting $z = \lambda^m$, (3.5) implies that $\|\boldsymbol{v}_m\| \geq \nu|\lambda^m| \geq \nu$. Hence the sequence $\{\boldsymbol{v}_m\}_{m=0}^{\infty}$ is bounded away from zero and $\boldsymbol{v}_m \not\to \mathbf{0}$. This completes the proof of the 'only if' part of the theorem.