

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 $|i - j| > r$, $i, j = 1, 2, \dots, n$. 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 $\mathcal{O}(n^3)$ operations and the expense of solving $A\mathbf{x} = \mathbf{b}$, provided that the factorization is known, is $\mathcal{O}(n^2)$. However, in the case of a banded A , we need just $\mathcal{O}(r^2n)$ 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 $A\mathbf{x} = \mathbf{b}$ 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 $B\mathbf{x} = \mathbf{c}$ is *easy to solve* and the matrix C is somehow ‘small’ in comparison with B . We write the original system in the form $B\mathbf{x} = -C\mathbf{x} + \mathbf{b}$ and consider solving it by iteration. Choose an arbitrary $\mathbf{x}_0 \in \mathbb{R}^n$ and define \mathbf{x}_{m+1} , $m = 0, 1, \dots$, by solving

$$B\mathbf{x}_{m+1} = -C\mathbf{x}_m + \mathbf{b}. \quad (3.1)$$

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 $\{\mathbf{x}_m\}_{m=0}^{\infty}$ converges to the solution of $A\mathbf{x} = \mathbf{b}$.

The Jacobi iteration We write $A = A_L + A_D + A_U$, 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

$$A_D\mathbf{x}_{m+1} = -(A_L + A_U)\mathbf{x}_m + \mathbf{b}, \quad m = 0, 1, \dots \quad (3.2)$$

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

$$(A_L + A_D)\mathbf{x}_{m+1} = -A_U\mathbf{x}_m + \mathbf{b}, \quad m = 0, 1, \dots \quad (3.3)$$

Note that $A_L + A_D$ 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 \mathbf{x}^* the solution of $A\mathbf{x} = \mathbf{b}$. 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 $\mathbf{x}_m \rightarrow \mathbf{x}^*$, regardless of the choice of $\mathbf{x}_0 \in \mathbb{R}^n$. Subtract $B\mathbf{x}^* = -C\mathbf{x}^* + \mathbf{b}$ from (3.1). This gives $B(\mathbf{x}_{m+1} - \mathbf{x}^*) = -C(\mathbf{x}_m - \mathbf{x}^*)$, hence $B\mathbf{v}_{m+1} = -C\mathbf{v}_m$, where $\mathbf{v}_m := \mathbf{x}_m - \mathbf{x}^*$ is the error in the m th iterate. Since B is nonsingular (otherwise we cannot execute (3.1) in the first place), it follows that

$$\mathbf{v}_{m+1} = H\mathbf{v}_m, \quad m = 0, 1, \dots \quad \text{where} \quad H := -B^{-1}C. \quad (3.4)$$

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

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

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

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

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

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

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

$$\|z\mathbf{w} + \bar{z}\bar{\mathbf{w}}\| \geq \nu|z|. \quad (3.5)$$

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

$$\mathbf{v}_m = \lambda^m\mathbf{w} + \bar{\lambda}^m\bar{\mathbf{w}}, \quad m = 0, 1, \dots$$

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