Dr H. Fawzi

Mathematical Tripos Part II: Michaelmas Term 2023

Numerical Analysis – Lecture 16

Theorem 4.10 (The Householder–John theorem) *If A and B are real matrices such that both A and $A-B-B^T$ are symmetric positive definite, then the spectral radius of $H = -(A-B)^{-1}B$ is strictly less than one.*

Proof. Let λ be an eigenvalue of H, so $H\boldsymbol{w} = \lambda\boldsymbol{w}$ holds, where $\boldsymbol{w} \neq \boldsymbol{0}$ is an eigenvector. (Note that both λ and $\boldsymbol{w}$ may have nonzero imaginary parts when H is not symmetric, e.g. in the Gauss–Seidel method.) By definition of H we have $-B\boldsymbol{w} = \lambda(A-B)\boldsymbol{w}$, and we note that $\lambda \neq 1$ since otherwise A would be singular (which it is not). Thus, we deduce

$$\overline{\boldsymbol{w}}^T B\boldsymbol{w} = \frac{\lambda}{\lambda - 1}\overline{\boldsymbol{w}}^T A\boldsymbol{w}, \tag{4.3}$$

where the bar means complex conjugation. Moreover, writing $\boldsymbol{w} = \boldsymbol{u} + i\boldsymbol{v}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are real, we find (for $C = C^T$) the identity $\overline{\boldsymbol{w}}^T C\boldsymbol{w} = \boldsymbol{u}^T C\boldsymbol{u} + \boldsymbol{v}^T C\boldsymbol{v}$, so symmetric positive definiteness in the assumption implies $\overline{\boldsymbol{w}}^T A\boldsymbol{w} > 0$ and $\overline{\boldsymbol{w}}^T(A - B - B^T)\boldsymbol{w} > 0$. In the latter inequality, we use relation (4.3) and its conjugate transpose to obtain

$$0 < \overline{\boldsymbol{w}}^T A\boldsymbol{w} - \overline{\boldsymbol{w}}^T B\boldsymbol{w} - \overline{\boldsymbol{w}}^T B^T\boldsymbol{w} = \left(1 - \frac{\lambda}{\lambda-1} - \frac{\overline{\lambda}}{\overline{\lambda}-1}\right)\overline{\boldsymbol{w}}^T A\boldsymbol{w} = \frac{1-|\lambda|^2}{|\lambda-1|^2}\,\overline{\boldsymbol{w}}^T A\boldsymbol{w}.$$

Now $\lambda \neq 1$ implies $|\lambda - 1|^2 > 0$. Hence, recalling that $\overline{\boldsymbol{w}}^T A\boldsymbol{w} > 0$, we see that $1 - |\lambda|^2$ is positive. Therefore every eigenvalue of H satisfies $|\lambda| < 1$ as required. □

Corollary 4.11 1) *If A is symmetric positive definite, then the Gauss-Seidel method converges.*
2) *If both A and $2D - A$ are symmetric positive definite, then the Jacobi method converges.*

Proof. 1) For the Gauss-Seidel method, B is the superdiagonal part of symmetric A, hence $A-B-B^T$ is equal to D, the diagonal part of A, and if A is positive definite, then D is positive definite too (this is the first part of the Exercise 23 from Example Sheets).

2) For the Jacobi method, we have $B = A-D$, and if A is symmetric, then $A-B-B^T = 2D-A$. (The latter matrix is the same as A except that the signs of the off-diagonal elements are reversed.) □

Example 4.12 (Poisson's equation on a square) As we have seen in the previous sections linear systems $A\boldsymbol{x} = \boldsymbol{b}$, where A is a real symmetric positive (negative) definite matrix, frequently occur in numerical methods for solving elliptic partial differential equations. A typical example we already encountered is Poisson's equation on a square where the *five-point formula* approximation yields an $n \times n$ system of linear equations with $n = m^2$ unknowns $u_{p,q}$:

$$u_{p-1,q} + u_{p+1,q} + u_{p,q-1} + u_{p,q+1} - 4u_{p,q} = h^2 f(ph, qh) \tag{4.4}$$

(Note that when p or q is equal to 1 or m, then the values $u_{0,q}$, $u_{p,0}$ or $u_{p,m+1}$, $u_{m+1,q}$ are known boundary values and they should be moved to the right-hand side, thus leaving fewer unknowns on the left.)

For any ordering of the grid points (ph, qh) we have shown in Lemma 1.11 that the matrix A of this linear system is symmetric and negative definite.

Corollary 4.13 *For linear system (4.4), both Jacobi and Gauss-Seidel methods converge.*

Proof. By Lemma 1.9 (Lecture 2), A is symmetric and negative definite, hence convergence of Gauss-Seidel. To prove convergence of the Jacobi method, we need negative definiteness of the matrix $2D - A$, and that follows by the same arguments as in Lemma 1.9: recall that the proof operates with the modulus of the off-diagonal elements and does not depend on their sign. □

Relaxation It is often possible to improve the efficiency of the recursive schemes above by *relaxation*. Specifically, instead of letting $\boldsymbol{x}^{(k+1)} = H\boldsymbol{x}^{(k)} + \boldsymbol{v}$, we let

$$\widehat{\boldsymbol{x}}^{(k+1)} = H\boldsymbol{x}^{(k)} + \boldsymbol{v}, \quad \text{and then} \quad \begin{aligned} \boldsymbol{x}^{(k+1)} &= \omega\widehat{\boldsymbol{x}}^{(k+1)} + (1-\omega)\boldsymbol{x}^{(k)} \\ &= H_\omega \boldsymbol{x}^{(k)} + \omega\boldsymbol{v} \end{aligned}$$

with

$$H_\omega = \omega H + (1-\omega)I,$$

where ω is a real constant called the *relaxation parameter*. (Note that $\omega = 1$ corresponds to the standard "unrelaxed" iteration.) Good choice of ω leads to a smaller spectral radius of the iteration matrix (compared with the "unrelaxed" method), and the smaller the spectral radius, the faster the iteration converges.

The eigenvalues of H_ω and H are related by the rule $\lambda_\omega = \omega\lambda + (1-\omega)$, therefore one may try to choose $\omega \in \mathbb{R}$ to minimize

$$\rho(H_\omega) = \max\left\{|\omega\lambda + (1-\omega)| : \lambda \in \sigma(H)\right\}$$

where $\sigma(H)$ is the spectrum of H. In general, $\sigma(H)$ is unknown, but often we have some information about it which can be utilized to find a "good" (rather than "best") value of ω. For example, suppose that it is known that $\sigma(H)$ is real and resides in the interval $[\alpha, \beta]$ where $-1 < \alpha < \beta < 1$. In that case we seek ω to minimize

$$\max\left\{|\omega\lambda + (1-\omega)| : \lambda \in [\alpha, \beta]\right\}.$$

It is readily seen that, for a fixed $\lambda < 1$, the function $f(\omega) = \omega\lambda + (1-\omega)$ is decreasing, therefore, as ω increases (decreases) from 1 the spectrum of H_ω moves to the left (to the right) of the spectrum of H. It is clear that the optimal location of the spectrum $\sigma(H_\omega)$ (or of the interval $[\alpha_\omega, \beta_\omega]$ that contains $\sigma(H_\omega)$) is the one which is centralized around the origin:

$$-[\omega\alpha + (1-\omega)] = \omega\beta + (1-\omega) \quad \Rightarrow \quad \omega_{\text{opt}} = \tfrac{2}{2-(\alpha+\beta)}\,, \quad -\alpha_{\omega_{\text{opt}}} = \beta_{\omega_{\text{opt}}} = \tfrac{\beta-\alpha}{2-(\alpha+\beta)}\,.$$