Dr H. Fawzi

Mathematical Tripos Part II: Michaelmas Term 2023

Numerical Analysis – Lecture 15

4 Iterative methods for linear systems

A general *iterative* method for solving $A\boldsymbol{x} = \boldsymbol{b}$ is a rule $\boldsymbol{x}^{k+1} = f_k(\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^k)$. We will consider the simplest ones: *linear, one-step, stationary* iterative schemes:

$$\boldsymbol{x}^{k+1} = H\boldsymbol{x}^k + \boldsymbol{v}, \qquad \boldsymbol{x}^0, \boldsymbol{v} \in \mathbb{R}^n. \tag{4.1}$$

Here one chooses H and $\boldsymbol{v}$ so that $\boldsymbol{x}^*$, a solution of $A\boldsymbol{x} = \boldsymbol{b}$, satisfies $\boldsymbol{x}^* = H\boldsymbol{x}^* + \boldsymbol{v}$, i.e. it is the fixed point of the iteration (4.1) (if the scheme converges). Standard terminology:

the *iteration matrix* H, the *error* $\boldsymbol{e}^k := \boldsymbol{x}^* - \boldsymbol{x}^k$, the *residual* $\boldsymbol{r}^k := A\boldsymbol{e}^k = \boldsymbol{b} - A\boldsymbol{x}^k$.

For a given class of matrices A (e.g. positive definite matrices, or even a single particular matrix), we are interested in *convergent* methods, i.e. the methods such that $\boldsymbol{x}^k \to \boldsymbol{x}^* = A^{-1}\boldsymbol{b}$ for every starting value $\boldsymbol{x}^0$. Subtracting $\boldsymbol{x}^* = H\boldsymbol{x}^* + \boldsymbol{v}$ from (4.1) we obtain

$$\boldsymbol{e}^{k+1} = H\boldsymbol{e}^k = \cdots = H^{k+1}\boldsymbol{e}^0, \tag{4.2}$$

i.e., a method is convergent if $\boldsymbol{e}^k = H^k\boldsymbol{e}^0 \to 0$ for any $\boldsymbol{e}^0 \in \mathbb{R}^n$.

Scheme 4.1 (Iterative refinement) This is the scheme

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - S(A\boldsymbol{x}^k - \boldsymbol{b}).$$

If $S = A^{-1}$, then $\boldsymbol{x}^{k+1} = A^{-1}\boldsymbol{b} = \boldsymbol{x}^*$, so it is suggestive to choose S as an approximation to A^{-1}. The iteration matrix for this scheme is $H_S = I - SA$.

Scheme 4.2 (Splitting) We assume $A = B + C$ in such a way that solving a linear system with the matrix C is "easy". We consider the scheme which can be written as $B\boldsymbol{x}^k + C\boldsymbol{x}^{k+1} = b$, i.e., eliminating C

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

with the iteration matrix $H = -(A - B)^{-1}B$. Any splitting can be viewed as an iterative refinement (and vice versa) because

$$\begin{aligned}(A - B)\boldsymbol{x}^{k+1} = -B\boldsymbol{x}^k + \boldsymbol{b} \quad &\Leftrightarrow \quad (A - B)\boldsymbol{x}^{k+1} = (A - B)\boldsymbol{x}^k - (A\boldsymbol{x}^k - \boldsymbol{b}) \\ &\Leftrightarrow \quad \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - (A - B)^{-1}(A\boldsymbol{x}^k - \boldsymbol{b}),\end{aligned}$$

so we should seek a splitting such that $S = (A - B)^{-1}$ approximates A^{-1}.

Theorem 4.3 *Let $H \in \mathbb{R}^{n\times n}$. Then $\lim_{k\to\infty} H^k\boldsymbol{z} = 0$ for any $\boldsymbol{z} \in \mathbb{R}^n$ if and only if $\rho(H) < 1$.*

Proof. 1) Let λ be an eigenvalue of (the real) H, real or complex, such that $|\lambda| = \rho(H) \geq 1$, and let $\boldsymbol{w}$ be a corresponding eigenvector, i.e., $H\boldsymbol{w} = \lambda\boldsymbol{w}$. Then $H^k\boldsymbol{w} = \lambda^k\boldsymbol{w}$, and

$$\|H^k\boldsymbol{w}\|_\infty = |\lambda|^k\|\boldsymbol{w}\|_\infty \geq \|\boldsymbol{w}\|_\infty =: \gamma > 0. \tag{4.3}$$

If $\boldsymbol{w}$ is real, we choose $\boldsymbol{z} = w$, hence $\|H^k\boldsymbol{z}\|_\infty \geq \gamma$, and this cannot tend to zero.

If $\boldsymbol{w}$ is complex, then $\boldsymbol{w} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v}$ with some real vectors $\boldsymbol{u}, \boldsymbol{v}$. But then at least one of the sequences $(H^k\boldsymbol{u})$, $(H^k\boldsymbol{v})$ does not tend to zero. For if both do, then also $H^k\boldsymbol{w} = H^k\boldsymbol{u} + \mathrm{i}H^k\boldsymbol{v} \to 0$, and this contradicts (4.3).

2) Now, let $\rho(H) < 1$, and assume for simplicity that H possesses n linearly independent eigenvectors $(\boldsymbol{w}_j)$ such that $H\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j$. Linear independence means that every $\boldsymbol{z} \in \mathbb{R}^n$ can be expressed as a linear combination of the eigenvectors, i.e., there exist $(c_j) \in \mathbb{C}$ such that $\boldsymbol{z} = \sum_{j=1}^n c_j \boldsymbol{w}_j$. Thus,

$$H^k \boldsymbol{z} = \textstyle\sum_{j=1}^n c_j \lambda_j^k \boldsymbol{w}_j \,,$$

and since $|\lambda_j| \le \rho(H) < 1$ we have $\lim_{k\to\infty} H^k \boldsymbol{z} = 0$, as required. □

Remark 4.4 The complete proof of case (2) of Theorem 4.3 exploits the so-called Jordan normal form of the matrix H, namely $H = SJS^{-1}$, where J is a block diagonal matrix consisting of the Jordan blocks,

$$J = \begin{bmatrix} \boxed{J_1} & & & \\ & \boxed{J_2} & & \\ & & \ddots & \\ & & & \boxed{J_r} \end{bmatrix}, \qquad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}, \qquad J_i \in \mathbb{R}^{n_i \times n_i}, \qquad \textstyle\sum_i n_i = n\,.$$

To prove that $J_i^k \to 0$ if $|\lambda_i| < 1$ one should split $J_i = \lambda_i I + P$, notice that $P^m = 0$ for $m \ge n_i$, and evaluate the terms of expansion $(\lambda_i I + P)^k = \sum_{m=0}^{n_i-1} \binom{k}{m} \lambda_i^{k-m} P^m$.

Applying Theorem 4.3 to the error estimate (4.2), we arrive at the following statement.

Theorem 4.5 *Let $\boldsymbol{x}^*$, a solution of $A\boldsymbol{x} = \boldsymbol{b}$, satisfy $\boldsymbol{x}^* = H\boldsymbol{x}^* + \boldsymbol{v}$ and we are given the scheme*

$$\boldsymbol{x}^{k+1} = H\boldsymbol{x}^k + \boldsymbol{v}, \qquad \boldsymbol{x}^0, \boldsymbol{v} \in \mathbb{R}^n. \tag{4.4}$$

Then $\boldsymbol{x}^k \to \boldsymbol{x}^*$ for any choice of $\boldsymbol{x}^0$ if and only if $\rho(H) < 1$.

Note: Of course, we would like to know not just convergence but the rate of it. For example, we achieve convergence with

$$H = \begin{bmatrix} 0.99 & 10^6 \\ 0 & 0.99 \end{bmatrix},$$

but it will take quite a long time. We will discuss this topic briefly later on.

Method 4.6 (Jacobi and Gauss–Seidel) Both of these methods are versions of splitting which can be applied to any A with nonzero diagonal elements. We write A as the sum of three matrices $L_0 + D + U_0$: subdiagonal (strictly lower-triangular), diagonal and superdiagonal (strictly upper-triangular) portions of A, respectively.

1) *Jacobi method*. We set $A - B = D$, the diagonal part of A, and we obtain the next iteration by solving the diagonal system

$$D\boldsymbol{x}^{k+1} = -(L_0 + U_0)\boldsymbol{x}^k + \boldsymbol{b}, \qquad H_{\mathrm{J}} = -D^{-1}(L_0 + U_0)\,.$$

2) *Gauss–Seidel method*. We take $A - B = L_0 + D = L$, the lower-triangular part of A, and we generate the sequence $(\boldsymbol{x}^{(k)})$ by solving the triangular system

$$(L_0 + D)\,\boldsymbol{x}^{k+1} = -U_0 \boldsymbol{x}^k + \boldsymbol{b}, \qquad H_{\mathrm{GS}} = -(L_0 + D)^{-1} U_0\,.$$

There is no need to invert $(L_0 + D)$, we calculate the components of $\boldsymbol{x}^{(k+1)}$ in sequence by forward substitution:

$$a_{ii} x_i^{k+1} = -\textstyle\sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j>i} a_{ij} x_j^k + b_i, \qquad i = 1..n.$$

As we mentioned above, the sequence $\boldsymbol{x}^k$ converges to solution of $A\boldsymbol{x} = \boldsymbol{b}$ if the spectral radius of the iteration matrix, $H_{\mathrm{J}} = -D^{-1}(L_0 + U_0)$ or $H_{\mathrm{GS}} = -(L_0 + D)^{-1} U_0$, respectively, is less than one. Our next goal is to prove that this is the case for two important classes of matrices A:

a) diagonally dominant and b) positive definite matrices.

We start with recalling the simple, but very useful Gershgorin theorem.

Revision 4.7 (Gershgorin theorem) *All eigenvalues of an $n \times n$ matrix A are contained in the union of the Gershgorin discs in the complex plane:*

$$\sigma(A) \subset \cup_{i=1}^{n} \Gamma_i\,, \qquad \Gamma_i := \{z \in \mathbb{C} : |z - a_{ii}| \le r_i\}, \qquad r_i := \textstyle\sum_{j \neq i} |a_{ij}|\,.$$

Definition 4.8 (Strictly diagonally dominant matrices) A matrix A is called strictly diagonally dominant by rows (resp. by columns) if

$$|a_{ii}| > \textstyle\sum_{j \neq i} |a_{ij}|, \quad i = 1..n \qquad (\text{resp.} \quad |a_{jj}| > \textstyle\sum_{i \neq j} |a_{ij}|, \quad j = 1..n\,).$$

From Gershgorin theorem, it follows that strictly diagonally dominant matrices are nonsingular.

Theorem 4.9 *If A is strictly diagonally dominant (either by rows or columns), then both the Jacobi and the Gauss-Seidel methods converge.*

Proof. Jacobi method: We have $H_{\mathrm{J}} = -D^{-1}(A - D) = I - D^{-1}A$. The diagonal elements of H_{J} are all zero, and the sum of the off-diagonal entries on the i'th row is $\sum_{j \neq i} |(H_{\mathrm{J}})_{ij}| = \sum_{j \neq i} |A_{ij}|/|A_{ii}| < 1$ if A is strictly diagonally dominant by rows. Applying Gersgorin's theorem to H_J, we get that all the eigenvalues of H_{J} have modulus < 1, which is what we wanted. If A is strictly diagonally dominant by columns (instead of by rows), we get that $\rho(I - AD^{-1}) < 1$ using the same argument, and use the fact that $I - D^{-1}A$ and $I - AD^{-1}$ have the same eigenvalues (since $I - D^{-1}A = D^{-1}(I - AD^{-1})D$).

Gauss-Seidel: If λ is an eigenvalue of $H_{\mathrm{GS}} = -(L_0 + D)^{-1}U_0$, then this means that $H_{\mathrm{GS}} - \lambda I = -(L_0 + D)^{-1}U_0 - \lambda I$ is singular, which in turn implies that $U_0 + \lambda(L_0 + D)$ is singular. It is easy to see that if $A = L_0 + D + U_0$ is strictly diagonally dominant, then the same is true for $A_\lambda = U_0 + \lambda(L_0 + D)$ for all $|\lambda| \ge 1$, and in particular A_λ is nonsingular in this case. This implies that any eigenvalue λ of $-(L_0{+}D)^{-1}U_0$ must satisfy $|\lambda| < 1$. This shows convergence of the Gauss-Seidel method. (Note: a similar argument can also be used for Jacobi.) $\square$