Mathematical Tripos Part II
Lent 2005
Professor A. Iserles

# Numerical Analysis – Lecture 1[1]

## 1 Iterative methods for linear algebraic systems

**Problem 1.1** (Positive definite systems). We consider linear systems of equations of the form $A\boldsymbol{x} = \boldsymbol{b}$, where $A$ is an $n \times n$ real positive definite symmetric matrix and $\boldsymbol{b} \in \mathbb{R}^n$ is known. Such systems occur, for example, in numerical methods for solving elliptic partial differential equations.

**Example 1.2** (Poisson's equation on a square). We seek the function $u(x, y)$, $0 \le x, y \le 1$, that satisfies $\nabla^2 u = f$ on the square, where $f$ is a given function and where $u$ is given on the boundary of the square. The best known finite difference method (the *five-point formula*) results in the approximation

$$(\Delta x)^2 \nabla^2 u(x, y) \approx u(x - \Delta x, y) + u(x + \Delta x, y) + u(x, y - \Delta x) + u(x, y + \Delta x) - 4u(x, y), \quad (1.1)$$

where $\Delta x = 1/(m + 1)$ for some positive integer $m$. Thus, we estimate the $m^2$ unknown function values $u(p\Delta x, q\Delta x)$, $1 \le p, q \le m$, by letting the approximation to $-(\Delta x)^2 \nabla^2 u(p\Delta x, q\Delta x)$ equal $-(\Delta x)^2 f(p\Delta x, q\Delta x)$ at these values of $p$ and $q$. This yields an $n \times n$ system of linear equations, where $n = m^2$. The matrix of this system, $A$ say, has the following properties.

***Lemma 1.3*** *The matrix $A$ of Example 1.2 is symmetric and positive definite.*

*Proof* Equation (1.1) implies that if $A_{i,j} \neq 0$ for $i \neq j$ then the $i$th and $j$th points of the grid $(p\Delta x, q\Delta x)$, $1 \le p, q \le m$, are nearest neighbours. Hence $A_{i,j} \neq 0$ implies $A_{i,j} = A_{j,i} = -1$, which proves the symmetry of $A$.

Therefore $A$ has real eigenvalues and eigenvectors. We consider the eigenvalue equation $A\boldsymbol{x} = \lambda\boldsymbol{x}$, and let $k$ be an integer in $\{1, \ldots, n\}$ such that $|x_k| = \max\{|x_i| : 1 \le i \le n\}$, where $x_i$ is the $i$th component of $\boldsymbol{x}$. Then we address the identity

$$(A_{k,k} - \lambda)x_k = -\sum_{\substack{j=1 \\ j \neq k}}^{n} A_{k,j} x_j.$$

Assume first that $\lambda < 0$. Since $A_{k,k} = 4$, $A_{k,j} \in \{0, -1\}$ for $k \neq j$ and at most four off-diagonal elements are nonzero for each $k$, it follows by the triangle inequality that

$$|(A_{k,k}-\lambda)x_k| = \left|\sum_{\substack{j=1 \\ j \neq k}}^{n} A_{k,j} x_j\right| \le \sum_{\substack{j=1 \\ j \neq k}}^{n} |A_{k,j}||x_j| \le 4 \max_{j=1,\ldots,n} |x_j| = 4|x_k| < (4+|\lambda|)|x_k| = |(A_{k,k}-\lambda)x_k|,$$

a contradiction. Therefore $\lambda \ge 0$. If $\lambda = 0$ then

$$(A_{k,k} - \lambda)|x_k| = 4|x_k| = \sum_{\substack{j=1 \\ j \neq k}}^{n} |A_{k,j}||x_j|$$

implies that $|x_j| = |x_k|$ whenever $A_{k,j} = -1$. Continuing by induction, this implies that $|x_j| \equiv |x_k|$ for all $j$, but this leads to contradiction in the equations that have fewer than four off-diagonal terms: such

[1] Please email all corrections and suggestions to these notes to `A.Iserles@damtp.cam.ac.uk`. All handouts are available on the WWW at the URL `http://www.damtp.cam.ac.uk/user/na/PartII/Handouts.html`.

equations occur at the boundary of the grid. It follows that all the eigenvalues are positive and $A$ is positive definite. □

**Method 1.4** (Simple iteration). We write $A\boldsymbol{x} = \boldsymbol{b}$ in the form

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

where the matrix $B$ is chosen so that it is easy to solve the system $(A - B)\boldsymbol{x} = \boldsymbol{y}$ for any given $\boldsymbol{y}$. Then *simple iteration* commences with an estimate $\boldsymbol{x}^{(0)}$ of the required solution, and generates the sequence $\boldsymbol{x}^{(k+1)}$, $k = 0, 1, 2, \ldots$, by solving

$$(A - B)\boldsymbol{x}^{(k+1)} = -B\boldsymbol{x}^{(k)} + \boldsymbol{b}, \qquad k = 0, 1, 2, \ldots.$$

If the sequence converges to a limit, $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \hat{\boldsymbol{x}}$, say, then the limit has the property $(A - B)\hat{\boldsymbol{x}} = -B\hat{\boldsymbol{x}} + \boldsymbol{b}$. Therefore $\hat{\boldsymbol{x}}$ is a solution of $A\boldsymbol{x} = \boldsymbol{b}$ as required.

**Revision 1.5** (Conditions for convergence). Method 1.4 requires both $A$ and $A - B$ to be nonsingular. Moreover, defining $\hat{\boldsymbol{x}}$ by $A\hat{\boldsymbol{x}} = \boldsymbol{b}$, we recall from the IB Numerical Analysis course that it is instructive to write the expression (1.4) in the form $(A - B)(\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}}) = -B(\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}})$. Indeed we deduce the relation

$$\boldsymbol{x}^{(k+1)} - \hat{\boldsymbol{x}} = H(\boldsymbol{x}^{(k)} - \hat{\boldsymbol{x}}) = H^{k+1}(\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}),$$

where $H = -(A - B)^{-1}B$. We found in Part IB that the required convergence of the sequence $\boldsymbol{x}^{(k)}$, $k = 0, 1, 2, \ldots$, is achieved for all choices of $\boldsymbol{x}^{(0)}$ if and only if $H$ has the property $\rho(H) < 1$. Here $\rho(H)$ is the *spectral radius* of $H$, which means the largest modulus of an eigenvalue of $H$. (Some of the eigenvalues may have nonzero imaginary parts.)

**Methods 1.6** (Jacobi and Gauss–Seidel). Both of these methods are versions of simple iteration for the case when $A$ is symmetric and positive definite.[2] In the *Jacobi method* the matrix $B$ has a zero diagonal but the off-diagonal elements of $B$ are those of $A$. In other words, $B$ is defined by the condition that $A - B$ is the diagonal matrix whose diagonal elements are the nonzero numbers $A_{i,i}$, $i = 1, 2, \ldots, n$.

In the *Gauss–Seidel method* one sets $B_{i,j} = 0$ for $j \le i$ and $B_{i,j} = A_{i,j}$ for $j > i$, so $A - B$ is a lower-triangular matrix with nonzero diagonal elements. The components of $\boldsymbol{x}^{(k+1)}$ satisfy

$$\sum_{j=1}^{i} A_{i,j}x_j^{(k+1)} = -\sum_{j=i+1}^{n} A_{i,j}x_j^{(k)} + b_i, \qquad i = 1, 2, \ldots, n,$$

and it is straightforward to calculate them in sequence by forward substitution.

Let us return to Example 1.2. Denoting by $u_{p,q}$ our approximation to $u(p\Delta x, q\Delta x)$, we have the equations

$$u_{p-1,q} + u_{p+1,q} + u_{p,q-1} + u_{p,q+1} - 4u_{p,q} = (\Delta x)^2 f(p\Delta x, q\Delta x).$$

Arranging grid points by columns (so called *natural ordering*), we obtain

$$\begin{aligned} \text{Jacobi:} \quad & u_{p-1,q}^{(k)} + u_{p+1,q}^{(k)} + u_{p,q-1}^{(k)} + u_{p,q+1}^{(k)} - 4u_{p,q}^{(k+1)} = (\Delta x)^2 f(p\Delta x, q\Delta x); \\ \text{Gauss–Seidel:} \quad & u_{p-1,q}^{(k+1)} + u_{p+1,q}^{(k)} + u_{p,q-1}^{(k+1)} + u_{p,q+1}^{(k)} - 4u_{p,q}^{(k+1)} = (\Delta x)^2 f(p\Delta x, q\Delta x). \end{aligned}$$

We will find that both methods converge for Example 1.2 by applying the following theorem. Its proof will be given in the next handout.

***Theorem 1.7*** *(The Householder–John theorem). If the real symmetric matrices $A$ and $A - B - B^{\top}$ are both positive definite, then the spectral radius of $H = -(A - B)^{-1}B$ is strictly less than one.*

[2] Actually, it is sufficient for all diagonal elements of $A$ to be nonzero.