ΒΑΣΙΚΕΣ ΕΠΑΝΑΛΗΠΤΙΚΕΣ ΜΕΘΟΔΟΙ

Οι παρούσες σημειώσεις αποτελούν βοήθημα στο μάθημα
Αριθμητικές Μέθοδοι του 5ου εξαμήνου του ΤΜΜ

Δημήτρης Βαλουγεώργης
Καθηγητής

Εργαστήριο Φυσικών και Χημικών Διεργασιών
Τμήμα Μηχανολόγων Μηχανικών
Πανεπιστήμιο Θεσσαλίας, 38334 Βόλος

Οκτώβριος 2012
1. Solution of algebraic systems

Systems of algebraic equations arise in the numerical solution of ordinary and partial differential as well as integral equations in time dependent and stationary problems implementing finite difference, finite elements and spectral methods.

In many cases the stability characteristics and properties of the approximation scheme are identical to the convergence characteristics of the implemented algorithm for solving the resulting algebraic system.

Consider the one and two dimensional model boundary value problems

\[ Lu = u_{xx} = -f(x) \text{ where } 0 < x < 1 \]

and

\[ Du = u_{xx} + u_{yy} = -g(x, y) \text{, where } 0 < x, y < 1 \]

with Dirichlet boundary conditions. After discretization the above differential equations are approximated by

\[ Lu = 2u_i - u_{i-1} - u_{i+1} = \frac{f_i}{N^2}, \quad i = 1, 2, ..., N - 1 \]

and

\[ Du = 4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j+1} - u_{i,j-1} = \frac{g_{i,j}}{N^2}, \quad i, j = 1, 2, ..., N - 1 \]

respectively. The differential equations are substituted by systems of algebraic equations, where matrix \( A \) is replaced by \( L \) and \( D \) and vector \( b \) by \( f \) and \( g \).
There are two types of methods: i) direct and ii) iterative.

Some of the most popular direct methods are:

- Gauss elimination (with and without partial and full pivoting)
- Gauss-Jordan elimination
- Thomas algorithm ($O(N)$ operations!)
- LU and LDU factorization ($A=LU$, $A=LDU$)
- Cholesky factorization (for SPD matrices, $A=LL^T$)
- Singular value decomposition (SVD)
- Least square method

Gauss elimination without pivoting is an unstable algorithm (vulnerable to numerical instabilities)

The number of operations for all direct methods applied to non-Hermitian matrices is of $O(N^3)$.

Most of the time, black box subroutines such as (LAPACK LINPACK, IMSL, NAG) are used.

Iterative methods is an alternative more efficient tool for solving large sparse diagonally dominant systems resulting from two and three-dimensional problems.

The number of operations per iteration is of $O(N^2)$.

Some of the most popular iterative methods are:

- Jacobi (J)
- Gauss-Seidel (GS)
- Successive Over-relaxation (S.O.R.)
- Symmetric Over-relaxation (S.S.O.R.)
- Alternative Direction Implicit (ADI) Conjugate Gradient (CG)
- Minimal Residual (MINRES)
- Generalized Minimal Residual (GMRES)
- Chebyshev Iteration (CI)

Here, we will examine the J, GS, SOR, SSOR and ADI schemes.
2. Introduction in iterative methods

Let \( Ax = b \), where \( A \) is a \( N \times N \) nonsingular real matrix and \( b \) a \( N \times 1 \) real column matrix. Then by adding the column vector \( Qx \) (\( Q \) nonsingular) at both sides we have

\[
Ax = b \iff Qx + Ax = Qx + b \iff Qx = (Q - A)x + b \iff x = Q^{-1}(Q - A)x + Q^{-1}b \iff x = (I - Q^{-1}A)x + Q^{-1}b \iff x = Gx + k
\]

which is rewritten in the iterative form

\[
x^{(n+1)} = Gx^{(n)} + k.
\]

Matrix \( G \) is called **iteration matrix** and it is of paramount importance with regard to the iterative characteristics of the implemented method.

All methods based on the above algorithm are linear stationary methods of first degree.

Two properties are of main interest: i) Consistency and ii) Convergence

The method is completely consistent when the solution of the related system \((I - G)x = k\) is also a solution of the initial system. A necessary and sufficient condition for convergence is:

\[
\rho(G) < 1.
\]

To measure the rate of convergence of the linear stationary iterative method we define the error vector

\[
\varepsilon^{(n)} = x^{(n)} - \overline{x}
\]

and then we easily find that \( \varepsilon^{(0)} = G\varepsilon^{(n-1)} = \ldots = G^n\varepsilon^{(0)} \).

Therefore by taking, any norm \( \beta \) of the above expression, a measure of the error after \( (n) \) iterations is obtained.

\[
\| \varepsilon^{(n)} \|_\beta = \| G^n \varepsilon^{(0)} \|_\beta \leq \| G^n \|_\beta \| \varepsilon^{(0)} \|_\beta
\]

Thus the norm \( \| G^n \|_\beta \) gives a measure by which the norm of the error has been reduced after \( (n) \) iterations.
The average rate of convergence is defined by

\[ R_n(G) = -\frac{1}{n} \ln \|G^n\|_\beta \]

It can be shown that if \( \rho(G) < 1 \) then

\[ \lim_{n \to \infty} \left( \|G^n\|_\beta \right)^{1/n} = \rho(G). \]

Hence we define the asymptotic rate of convergence by

\[ R_\infty(G) = \lim_{n \to \infty} R_n(G) = -\ln \left[ \rho(G) \right]. \]

Another way to approach the iterative methods is to introduce the idea of preconditioner. Reconsider the iterative algorithm in the form

\[ x^{(n+1)} = (I - Q^{-1}A)x^{(n)} + Q^{-1}b = x^{(n)} + Q^{-1}\left(b - Ax^{(n)}\right). \]

Thus by taking an initial guess for \( x^{(0)} \) then we find a matrix \( Q \) such that the error \( b - Ax^{(0)} \) is improved by applying the above iteration algorithm. Then matrix \( Q \) is called a preconditioner of the system \( Ax = b \).

This procedure is sometimes called simple iteration, but more often it goes by different names according to the choice of \( Q \). For \( M \) equal to the diagonal \( D \) of \( A \) it is called Jacobi iteration (J), while for \( Q \) equal to the lower triangle \( L \) of \( A \) it is the Gauss Seidel (GS) method. Even more for \( Q \) of the form \( \omega^{-1}D - L \) is the successive over relaxation method (SOR).

Finally it is important to note that the behavior of an iterative method depends on the properties of the preconditioned matrix \( Q' A \), since the eigenvalues of \( G \) are \( 1 - \lambda_i \), where \( \lambda_i \) are the eigenvalues of the preconditioned matrix \( Q' A \).

Preconditioners can be divided into three categories:
• Preconditioners designed for general classes of matrices: e.g.,
matrices with nonzero diagonal entries, positive definite matrices.
Examples of such preconditioners are the Jacobi, Gauss Seidel and
SOR preconditioners and the incomplete Cholesky preconditioners.

• Preconditioners designed for broad classes of underlying problems:
e.g., elliptic partial differential equations. Examples are multigrid
and domain decomposition preconditioners (ADI).

• Preconditioners designed for a specific matrix or underlying
problem: e.g., the transport equation. An example is the diffusion
synthetic acceleration (DSA) preconditioner.

3. Extrapolation

An iterative method is symmetrical if for some nonsingular matrix \( W \) the
matrix \( W(I-G)W^{-1} \) is SPD.

If an iterative method is symmetrical then

1) The eigenvalues of \( G \) are real
2) The algebraically largest eigenvalue \( M(G) \) is less than unity
3) The set of eigenvectors of \( G \) includes a basis of the associated vector
space.

The symmetrization property does not imply convergence. However, if
the method is symmetrizable then there always exists a so-called
extrapolated corresponding method, which is convergent.

The extrapolated method is defined by

\[
x^{(n+1)} = \gamma(Gu^{(n)} + k) + (1 - \gamma)u^{(n)} = \left[\gamma G + (1 - \gamma)I\right]u^{(n)} + \gamma k = G[\gamma]u^{(n)} + \gamma k
\]

If the iterative method is symmetrizable the optimum value for \( \gamma \), is the
one, which minimizes the spectral radius of \( G \) and it is given by

\[
\gamma_{opt} = \frac{2}{2 - M(G) - m(G)}
\]

where \( M(G) \) and \( m(G) \) are the largest and smallest eigenvalues og \( G \)
respectively. It follows that

\[
\rho\left(G_{\gamma_{opt}}\right) = \frac{M(G) - m(G)}{2 - M(G) - m(G)} < 1.
\]

Thus the optimum extrapolated method is convergent.
4. Stationary iterative methods

Stationary are the iterative methods for which the iteration matrix $G$ does not depend on the iteration count $(n)$.

We assume throughout this section that the matrix $A$ is SPD and we use the two dimensional Dirichlet problem as the model problem.

4.1 Richardson method (RF)

The RF algorithm is based on a variant of the method of Richardson (1910) and is defined by

$$x^{(n+1)} = (I - A)x^{(n)} + b$$

The iteration matrix $G = I - A$ and the associated preconditioner or splitting matrix is the identity matrix $I$. Since $A$ is SPD the RF method is symmetrizable and $W = I$ is the symmetrization matrix.

The eigenvalues of $G$ are $1 - \mu_i$, where $\mu_i$ are the eigenvalues of $A$. Thus

$$\rho(I - A) = \max \{|1 - m(A)|, |1 - M(A)|\}.$$  

Since $A$ is SPD its eigenvalues are positive and consequently the RF method is convergent if and only if $M(A) < 2$.

Applying the optimal extrapolation at the RF method we have

$$x^{(n+1)} = (I - \gamma_{opt} A)x^{(n)} + \gamma_{opt} b$$

where

$$\gamma_{opt} = \frac{2}{2 - M(G) - m(G)} = \frac{2}{M(A) + m(A)}.$$
The spectral radius of the corresponding iteration matrix is

\[ \rho(I - \gamma_{opt} A) = \frac{M(A) - m(A)}{M(A) + m(A)} = \frac{\kappa(A) - 1}{\kappa(A) + 1} < 1 \]

where \( \kappa(A) \) is the spectral condition number of matrix \( A \). Therefore, the asymptotic rate of convergence of the extrapolated RF method is

\[ R_\infty(I - \gamma_{opt} A) = -\ln \frac{\kappa(A) - 1}{\kappa(A) + 1} \sim -\frac{2}{\kappa(A)}, \text{ as } \kappa(A) \to \infty. \]

For the model problem it can be shown that

\[ m(A) = 8 \sin^2 \left( \frac{1}{2} \pi h \right) \quad \text{and} \quad M(A) = 8 \cos^2 \left( \frac{1}{2} \pi h \right). \]

Then

\[ R_\infty(I - \gamma_{opt} A) = -\ln \rho(I - \gamma_{opt} A) = -\ln \frac{\cot^2 \left( \frac{1}{2} \pi h \right) - 1}{\cot^2 \left( \frac{1}{2} \pi h \right) + 1} \sim -\frac{1}{2} \pi^2 h^2, \quad h \to 0. \]

**4.2 Jacobi method (J)**

The J method is easily derived if we write the system in the form

\[ \sum_{j=1}^{N} a_{ij} x_j = b_i \]

and then we solve for the \( x_i \), while we assume that the other entries of \( x \) remain fixed. Then the J iterative method is written as

\[ x_i^{(n+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(n)} \right) \]
In matrix form, the algorithm is written as

\[ x^{(n+1)} = D^{-1}(L + U)x^{(n)} + D^{-1}b = x^{(n)} + D^{-1}(b - Ax^{(n)}), \]

where \( D \), -\( L \) and -\( U \) represent the diagonal, the strictly lower triangular and the strictly upper triangular parts of \( A (A=D-L-U) \).

The iteration matrix will be denoted by \( B=I-D^{-1}A=D^{-1}(L+U) \) and the associated preconditioner or splitting matrix is the diagonal matrix \( D \). The J method is convergent if and only if \( \rho(B) < 1 \).

A rigorous analysis of the convergence of the J method can be given if we consider the one dimensional Dirichlet boundary value problem:

\[ \tilde{L}u = u_{xx} = -f(x) \text{ where } 0 < x < 1 \]

discretized by

\[
Lu = 2u_i - u_{i-1} - u_{i+1} = \frac{f_i}{N^2}, \quad i = 1, 2, \ldots, N - 1.
\]

The eigenvalues \( \lambda = 4\sin^2 \frac{n\pi}{2N} \) and the corresponding eigenfunctions \( u_n(x) = \sin(n\pi x), n=1,2,\ldots,N-1 \), of the \( \tilde{L} \) and \( L \) operators are the same.

The eigenvalues of the J iteration matrix \( B \) are then

\[
\lambda(B) = 1 - \frac{\lambda(L)}{2} = 1 - 2\sin^2 \frac{n\pi}{2N} = \cos \frac{n\pi}{N}.
\]

From this it is easy to see that high frequency modes (i.e. eigenfunctions with \( n \) large) are dumped quickly, while the damping factor for modes with \( n \) small is close to 1. The spectral radius of the J method is for \( n=1 \), proportional to \( 1-O(h^2) \) and it is attained for the eigenfunction \( u_1(x)=\sin(n\pi x) \) (\( h=N^d \) where \( N \) is the number of variables and \( d \) is the number of space dimensions).

Also \( R_o(B) \sim \frac{1}{2}\pi^2h^2, \quad h \to 0 \).
4.3 Gauss-Seidel method (GS)

If we proceed as with the J method but now assume that the equations are examined one at a time in sequence, and that previously computed results are used as soon as they are available, we obtain the GS method:

\[ x_i^{(n+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j<i} a_{ij} x_j^{(n+1)} - \sum_{j>i} a_{ij} x_j^{(n)} \right) \]

or in matrix form

\[ x^{(n+1)} = (D-L)^{-1}(Ux^{(n)} + b) = \left( (I-D^{-1}L)^{-1}D^{-1}U \right) u^{(n)} + (I-D^{-1}L)^{-1}D^{-1}b \]

Important facts about the GS method:

- The computations are serial (not easily parallelized).
- The new iterate \( x^{(n+1)} \) depends upon the order in which the equations are examined (from the first to the last or vice versa).
- The preconditioner or splitting matrix \( D-L \), is not SPD and the GS method is not symmetrizable.
- Since \( A \) and \( D \) are SPD, it can be shown that the GS method always converge.
- In general, the eigenvalues of the iteration matrix, though less than unity in modulus may be complex and the set of eigenvectors of the iteration matrix may not include a basis for the associated vector space. In this case the extrapolation method is not applicable.

If \( G \) and \( B \) denote the iteration matrices of the GS and the J methods for the model problem, then it can be shown that

\[ \rho(G) = \left[ M(B) \right]^2 = \cos^2 \pi h, \]

\[ m(G) = 0 \]

and

\[ R_e(G) = \pi^2 h^2, \quad h \to \infty. \]
4.4 Successive over-relaxation method (SOR)

A tremendous improvement of the GS method is the SOR algorithm given by

\[ a_{ii}x_i^{(n+1)} = \omega \left( b_i - \sum_{j<i} a_{ij}x_j^{(n+1)} - \sum_{j>i} a_{ij}x_j^{(n)} \right) + (1 - \omega) a_{ii}x_i^{(n)}, \]

where \( \omega \in (0, 2] \) is a real number known as the relaxation factor. With \( \omega = 1 \), the SOR method reduces to the GS method. If \( \omega > 1 \) or \( \omega < 1 \), we have over-relaxation or under-relaxation respectively. Usually we are using over-relaxation.

The interesting part of the SOR method is the estimation of the optimal value of \( \omega \) and the spectral radius of the SOR method in terms of the spectral radius of the J method. This can be done easily for the model problems, where the eigenvalues are known. It can be shown that the spectral radius of the SOR method is \( 1 - O(h) \) instead of \( 1 - O(h^2) \) as it is for \( \omega = 1 \). This is a tremendous improvement, because it implies that the required number of iterations for convergence is reduced from \( O(h^{-2}) \) to \( O(h^{-1}) \).

In matrix form the algorithm is written as

\[ x^{(n+1)} = (D - \omega L)^{-1} \left[ \omega U + (1 - \omega) D \right] x^{(n)} + \omega(D - \omega L)^{-1} b = \]

\[ = (I - \omega L)^{-1} \left[ \omega U + (1 - \omega) I \right] x^{(n)} + \omega(I - \omega L)^{-1} D^{-1} b \]

The preconditioner or splitting matrix for the SOR method is \( \omega^{-1}D - L \), which as in the case of the GS method is not SPD and the method is not symmetrizable. Extrapolation is not applicable.

Since \( A \) and \( D \) are SPD it can be shown that the SOR method converges for any \( 0 < \omega < 2 \).
If $G$ and $B$ denote the iteration matrices of the SOR and J methods for the model problem, then it can be shown that

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - M(B)^2}} = \frac{2}{1 + \sin \pi h}$$

and the corresponding spectral radius and asymptotic convergence rate of the SOR are

$$\rho(G) = \frac{1 - \sqrt{1 - M(B)^2}}{1 + \sqrt{1 - M(B)^2}} = 1 - \frac{\sin \pi h}{1 + \sin \pi h},$$

and

$$R_x(G) = 2\pi h, \quad h \to \infty.$$ 

Note that for $\rho(B) = 0.99$ we find $\rho(G_{\text{GS}}) = 0.98$, while $\rho(G_{\text{SOR}_{\text{opt}}}) = 0.75 = 0.99^{30}$. In other words 30 Jacobi iterations produce the same result with 15 Gauss Seidel iterations and 1 SOR iteration!!!

### 4.5 Successive over relaxation method (SSOR)

If we assume that the coefficient matrix $A$ is symmetric, then the SSOR method combines two SOR sweeps together in such a way that the resulting iteration matrix is similar to a symmetric matrix. That is, SSOR is a forward SOR sweep followed by a backward SOR sweep. The similarity of the SSOR iteration matrix to a symmetric matrix permits the application of SSOR as a preconditioner for other iterative schemes for symmetric matrices. Indeed, this is the primary motivation for SSOR, since its convergence rate with $\omega_{\text{opt}}$ is slower than the convergence rate of SOR with $\omega_{\text{opt}}$.

The SSOR is defined as

$$a_{ii}x_i^{(n+1/2)} = \omega \left( b_i - \sum_{j<i} a_{ij}x_j^{(n+1/2)} - \sum_{j>i} a_{ij}x_j^{(n)} \right) + (1 - \omega) a_{ii}x_i^{(n)},$$

with $i = 1, 2, \ldots, n$. The new coefficients are

$$a_{ii}^{(n+1/2)} = \omega \sum_{j<i} a_{ij}^{(n+1/2)} + (1 - \omega) a_{ii}^{(n)}, \quad i = 1, 2, \ldots, n.$$
and
\[ a_{ii}x_i^{(n+1)} = \omega \left( b_i - \sum_{j<i} a_{ij}x_j^{(n+1/2)} - \sum_{j>i} a_{ij}x_j^{(n+1)} \right) + (1 - \omega) a_{ii}x_i^{(n+1/2)}, \]

or in matrix form as
\[ Dx^{(n+1/2)} = \omega \left( Lx^{(n+1/2)} + Ux^{(n)} \right) + (1 - \omega) Dx^{(n)} \]
and
\[ Dx^{(n+1)} = \omega \left( Lx^{(n+1/2)} + Ux^{(n+1)} \right) + (1 - \omega) Dx^{(n+1/2)} \]

Combining the two equations we find
\[ x^{(n+1)} = Gx^{(n)} + k \]

where
\[ G = (I - \omega U)^{-1} \left[ \omega L + (1 - \omega) I \right] (I - \omega L)^{-1} \left[ \omega U + (1 - \omega) I \right] \]
and
\[ k = \omega (2 - \omega) (I - \omega U)^{-1} (I - \omega L)^{-1} D^{-1} b. \]

The preconditioner or splitting matrix for the SSOR method is
\[ Q = \frac{\omega}{2 - \omega} \left( \frac{1}{\omega} D - L \right) D^{-1} \left( \frac{1}{\omega} D - U \right). \]

Matrix $Q$ is SPD and hence SSOR is symmetrizable.

The rate of convergence of the SSOR is practically insensitive of the exact choice of $\omega$.

A good value of $\omega$ is $\omega = \frac{2}{1 + \sqrt{2(1 - M(B))}}$ with the corresponding asymptotic rate of convergence $R_{\infty}(G) \geq \pi h$, as $h \to 0$. 

13