Iterative Methods for Ax=b

1 Fundamentals

consider the solution of the set of simultaneous equations $Ax = b$ where $A$ is a square matrix, $n \times n$ and $b$ is a right hand vector. We write the iterative method as

$$x^{k+1} = F_k(A, b, x^k, x^{k-1}, \ldots, x^{k-m+1})$$ (1)

where $x^k$ is the vector found from the $k$th iteration. Obviously we must have that $x^k \to x_t$ as $k \to \infty$, where $x_t$ is the true solution. The iteration process is called stationary if $F_k$ is independent of $k$, linear if a linear function of $x^k, x^{k-1}, \ldots, x^1$, and of degree $m$.

1.1 Linear Iterations

Restricting ourselves to one point (i.e. $m=1$) linear iterations we have

$$x^{k+1} = T^k x^k + q^k$$ (2)

As an example, consider

$$Ax = b$$ (3a)

$$(I + A)x = x + b$$ (3b)

$$x = (I + A)x - b$$ (3c)

giving

$$T = I + A \text{ and } q = -b$$ (3d)

Clearly Eq.(2) is just a generalization of Eq.(3c). Now at convergence we must have

$$x_t = T^k x_t + q^k$$ (4)

for all $k$. Since $x_t = A^{-1}b$ we can write

$$A^{-1}b = T^k A^{-1}b + q^k$$ (5a)

or

$$q^k = (I - T^k)A^{-1}b \equiv Q^kb$$ (5b)

Thus our final form for the iterative process is

$$x^{k+1} = T^k x^k + Q^kb$$ (6)

1.2 Convergence

Defining the error, $e^k = x^k - x_t$ we have

$$e^{k+1} = T^k x^k + Q^kb - x_t$$ (7a)

$$= T^k x^k + Q^kb - A^{-1}b$$ (7b)

$$= T^k e^k \text{ using Eq.}(5b)$$ (7c)

$$= (T^k T^{k-1} \ldots T^1) e^1$$ (7d)

$$= M^k e^1$$ (7e)
We know that the residual, $r^k = b - Ax^k = Ae^k$ is not a good measure of the accuracy of the solution if the matrix $A$ is ill conditioned, but of course $e^k$ is a good measure. If all of the eigenvectors of $M^k$ are less than 1 in absolute value, the iteration will converge. Letting $\lambda(M^k)$ be the magnitude of the largest eigenvalue of $M^k$ the rate of convergence, $R(M^k)$ is given by

$$R(M^k) = -\frac{1}{k} \log(\lambda)$$

If the iteration is stationary $\lambda(M^k) = \lambda(T)^k$ giving $R(M^k) = -\log_{10}(T)$. That is $R$ indicates how many digits of accuracy are gained with each iteration.

It is important to estimate the value of $\lambda$. Several bounds are (see the notes for Lecture 3 on Condition Numbers)

1. $|\lambda_{\text{max}}| \leq \sum |a_{ij}|$
2. $|\lambda_{\text{max}}| \leq (\sum a^2_{ij})^{1/2}$
3. $|\lambda_{\text{max}}| \leq \max_i \sum |a_{ij}|$
4. $|\lambda_{\text{max}}| \leq \max_j \sum |a_{ij}|$

Since

$$\max_{e^1} \frac{||e^k||_E}{||e^1||_E} = \max_{e^1} \frac{||M^k e^1||}{||e^1||} = ||M^k||_S$$

we must have

$$\lim_{k \to \infty} ||M^k||_S = 0$$

If the process is stationary, it is sufficient for

$$\lim_{k \to \infty} \rho(M^k) = 0$$

2 Jacobi’s Method

We write $A = L + D + U$ and define our algorithm by

$$(L + D + U)x = b$$

$$x^{k+1} = -D^{-1}(L + U)x^k + D^{-1}b$$

and using $\delta x = x^{k+1} - x^k$

$$\delta x = -D^{-1}(L + U)x^k - x^k + D^{-1}b$$

$$= -D^{-1}Ax^k + D^{-1}b$$

$$= D^{-1}r^k$$

The method is also known as the method of simultaneous displacements. The method converges if all of the eigenvalues of $D^{-1}(L + U)$ are less than 1 in absolute value. Unfortunately it is not easy to determine if this is true. Instead, for most engineering problems we rely upon the following theorem which is true for matrices that are diagonally dominant and irreducible.
2 JACOBI’S METHOD

Irreducibility

A matrix $A$ is said to be reducible if it can be written as

$$A = \begin{pmatrix} A_1 & 0 \\ A_2 & A_3 \end{pmatrix}$$

(13)

A reducible form means that some of the $n_1$ components of $x$ are uniquely determined by $n_1$ components of $b$, not by all $n$ components. In a boundary value problem this means that some of the field variables are independent of the boundary conditions. Obviously for boundary value problems the matrix $A$ must be irreducible.

Diagonal Dominance

A matrix $A$ is said to be diagonally dominant if its components satisfy

$$\sum_{j=1}^{n} |a_{i,j}| \leq |a_{i,i}|$$

(14)

for all $i$, with strict inequality for at least one value of $i$.

Theorem

If $A$ has diagonal dominance and is not reducible, then Jacobi’s method converges.

2.1 Example of Failure

If the diagonal fails to dominate, the method may fail to converge. For example consider the finite difference representation of

$$\frac{d^2u}{dx^2} = f$$

(15)

over a series of equally spaced points, each $\delta$ apart. Let the leftmost point have the value $u_0$ while the right hand point, $u_1$ is located between the points $M$ and $M-1$ at a distance $p\delta$ from the point $M$. Using linear interpolation, the appropriate matrix is

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 & 0 \\ \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -2 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1-p \end{pmatrix}$$

(16)

The eigenvalues are found from the determinant of

$$det(L + \lambda D + U) = 0$$

$$= p\lambda P_{n+1}(\lambda) + (1-p)\lambda P_{n-2}(\lambda)$$

(17)

where $P_{n-1}(\lambda)$ is a $n-1$ order polynomial. As $p \to 0$ the determinant reduces to a polynomial of order $n-2$ and thus has only $n-1$ roots. This implies that one of the roots of the equation tends to $\infty$. Hence the iteration will diverge.
2.2 Equivalence to a transient problem

Consider the transient problem

\[ \frac{du}{dt} = cu + b \]

which can be represented as

\[ \frac{u^{k+1} - u^k}{\delta t} = b + cu^k \] (18)

If the time step is \( \delta t = 1 \), then this equation is identical to Jacobi’s method. In other words, Jacobi’s method represents the asymptotic behavior of a time dependent problem. This will help us understand the behavior of the solution as the iterations proceed.

3 Gauss-Seidel Method

Suppose that in the Jacobi method that we use the new value of a component of \( x^{k+1} \) as soon as it is known. That is we write

\[ a_{j,1}x_1^{k+1} + \ldots + a_{j,j-1}x_{j-1}^{k+1} + a_{j,j}x_j^k + a_{j,j+1}x_{j+1}^k + \ldots + a_{j,n}x_n^k = b_j \] (19)

so that

\[ x_j^{k+1} = \frac{b_j - a_{j,1}x_1^k - \ldots - a_{j,j-1}x_{j-1}^k - a_{j,j+1}x_{j+1}^k - \ldots - a_{j,n}x_n^k}{a_{j,j}} \] (20)

This is represented as

\[ Lx^{k+1} + Dx^{k+1} + Ux^k = b \]

or

\[ x^{k+1} = (D + L)^{-1}(b - Ux^k) \] (21)

where now \( T = -(D + L)^{-1}U \)

Using \( \delta x = x^{k+1} - x^k \) we have for the \( J \)th component of \( x \)

\[ \delta x_J = \frac{b_J - \sum_{j=1}^{J-1} a_{j,j}x_j^{k+1} - \sum_{j=J}^{n} a_{j,j}x_j^k}{a_{j,j}} \] (22a)

\[ x_j^{k+1} = x_j^k + \delta x_J \] (22b)

1. If Jacobi converges, Gauss-Seidel will converge

2. If \( A \) is symmetric and positive definite and \( a_{i,i} \geq 0 \) for all \( i \), then Gauss-Seidel converges for all \( x^0 \)

4 Over-Relaxation, SOR

It has been observed that an improved method can be achieved if the change used is larger than the computed value. That is in the Gauss-Seidel method we replace Eq.(22b) with

\[ x_j^{k+1} = x_j^k + \omega \delta x_J \] (23)

If \( \omega > 1 \) the method is called Successive Over Relaxation. If \( \omega < 1 \) it is termed under relaxation. The question is how to determine an optimal value of \( \omega \).
4 OVER-RELAXATION, SOR

4.1 Finding $\omega_{opt}$

Let $\lambda$ be the eigenvalue of Jacobi’s method and $\eta$ be of the SOR method. Then the relationship between $\lambda, \eta, \omega$ is

\[
\eta^{1/2} = \frac{\omega \lambda \pm \sqrt{\omega^2 \lambda^2 - 4(\omega - 1)}}{2}
\]

(24a)

\[
\omega_{opt} = \frac{2}{1 + \sqrt{1 - \lambda^2}}
\]

(24b)

Your choices are then:

1. To run with $\omega = 1$, estimate $\eta$, and then use the fact that $\eta = \lambda^2$ for Gauss-Seidel’s method (SOR with $\omega = 1$).

2. To run with $\omega \neq 1$, estimate $\eta$, use Eq.(24a) to determine $\lambda$, and then Eq.(24b) to find $\omega_{opt}$.

The following figure shows the characteristic shape of the spectral radius as a function of $\omega$.

![Figure 1: Effect of $\omega$](image)

Note that underestimating $\omega(\text{opt})$ will lead to a severe lack of performance while a small overestimate will be of little effect. Consequently, after estimating $\omega(\text{opt})$, choose a value slightly larger, for example $\omega(\text{opt}) + 0.1 \times (2 - \omega(\text{opt}))$.

Two excellent books for iterative methods, with discussion of rates of convergence and the choice of optimal values of $\omega$ are:

1. Iterative Solution of Large Linear Systems by D. M. Young, Academic Press.

4.2 Non-symmetric matrices, SSOR

Most engineering problems involve symmetric matrices. If $A$ is not symmetric, a two step process is often used. We solve for new components, $x_{k+1}^i$ by going from $i = 1$ to $i = n$. Then we go from $i = n$ to $i = 1$. If we do this with SOR, the method is called Symmetric Successive Over Relaxation (SSOR).

5 Acceleration

Most iterative methods have a spectral radius, $\rho = 1 - \epsilon$, with very slow convergence and their success depends upon special tricks to accelerate them.

5.1 Basic Acceleration

Let

$$x^{k+1} = x_t + e^{k+1} = A^{-1}b + e^{k+1} \approx A^{-1}b + \rho e^k$$

$$x^k = A^{-1}b + e^k$$

Solving for $e^k$ and substituting gives

$$A^{-1}b = x_t \approx \frac{x^{k+1} - x^k \rho}{1 - \rho}$$

How do we find $\rho$? The residuals, $r^k = b - Ax^k$ can be shown to follow

$$\frac{||r^{k+1}||}{||r^k||} \rightarrow \rho \text{ as } k \rightarrow \infty$$

so by watching their behavior we can estimate $\rho$ and use Eq.(25)

5.2 Aitken’s Method

The iterative methods, Jacobi and Gauss-Seidel) are linear in nature. That is if we express the error at any step $k$ we can write

$$e^{k+1} = T^k e^k$$

and the new error is linearly related to the previous error. Now near the solution we expect (assume) that the new error is a constant times the previous error. We write

$$x_t - x^{k+1} = B(x_t - x^k)$$

$$x_t - x^{k+2} = B(x_t - x^{k+1})$$

We then have two equations for the two unknowns, $B$ and $x_t$. Solving we obtain

$$x_t = \frac{x^k x^{k+2} - x^{k+1} x^{k+1}}{x^k - 2 x^{k+1} + x^{k+2}}$$

$$= x^{k+2} - \frac{\Delta(x^{k+1})^2}{\Delta^2 x^{k+1}(k + 1)}$$

Note that we need not estimate the spectral radius, $\rho(M^k)$. The method is referred to as Aitken’s $\Delta^2$ method.
5.3 Acceleration with Relaxation

You should not accelerate the over relaxed iterations because they are already being accelerated through the use of $\omega$. Using Aitken’s method on SOR will probably slow down the convergence.

6 Rates of Convergence

When using the different methods for solving Laplace’s equation on a square grid, with the spacing between the nodes of $h$ and $n$ points on a line, the approximate rates of convergence are:

<table>
<thead>
<tr>
<th>Method</th>
<th>Rate of Convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacobi</td>
<td>$h^2/2$</td>
</tr>
<tr>
<td>Gauss Seidel</td>
<td>$h^2$</td>
</tr>
<tr>
<td>Optimal SOR</td>
<td>$2h$</td>
</tr>
<tr>
<td>Jacobi, implicit by lines</td>
<td>$h^2$</td>
</tr>
<tr>
<td>Gauss – Seidel, implicit by lines</td>
<td>$2h^2$</td>
</tr>
<tr>
<td>Optimal SOR, implicit by lines</td>
<td>$2\sqrt{2}h$</td>
</tr>
</tbody>
</table>

For a general matrix $A$ with a condition number of $P$, the rates of convergence are

\[
 r = \ln \left( \frac{P + 1}{P - 1} - \frac{2\sqrt{P}}{P - 1} \right) \quad \text{Richardson}
\]

\[
 \approx \frac{2}{\sqrt{P}} \quad \text{(31a)}
\]

\[
 = \frac{2}{P} \quad \text{Jacobi} \quad \text{(31b)}
\]