CME 302: NUMERICAL LINEAR ALGEBRA
FALL 2005/06
LECTURE 15

GENE H. GOLUB

1. Convergence of Iterative Methods

Recall the basic iterative methods based on the splitting $A = D + L + U$, the *Jacobi method*

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

and the *Gauss-Seidel method*

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

These are examples of *one-step stationary method*, which is an iteration of the form

$$M\mathbf{x}^{(k+1)} = N\mathbf{x}^{(k)} + \mathbf{b},$$

where $A = M - N$.

Let $B = M^{-1}N$, and define $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$. Then $\mathbf{e}^{(k+1)} = B\mathbf{e}^{(k)} = B^{k+1}\mathbf{e}^{(0)}$. Recall that if $\rho(B^k) < 1$ then $\mathbf{e}^{(k)} \to 0$ for all choices of $\mathbf{x}^{(0)}$. Also, recall that for all consistent norms, $\rho(B) \leq \|B\|$.

Therefore, a sufficient condition for convergence of the Jacobi method is $\|B\|_\infty < 1$ where

$$b_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}} & i \neq j, \\ 0 & i = j. \end{cases}$$

Note that

$$\|B\|_\infty = \max_i \sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| < 1$$

if B is *diagonally dominant*.

Now, define

$$r_i = \sum_{i \neq j} \left| \frac{a_{ij}}{a_{ii}} \right|, \quad r = \max_i r_i.$$

Then we have the following result:

Theorem If $r < 1$, then $\rho(B_{GS}) < 1$. In other words, the Gauss-Seidel iteration converges if A is diagonally dominant.

Proof The proof proceeds using induction on the elements of $\mathbf{e}^{(k)}$. We have

$$(D+L)\mathbf{e}^{(k+1)} = U\mathbf{e}^{(k)},$$

which can be written as

$$\sum_{j=1}^{i} a_{ij} e_j^{(k+1)} = -\sum_{j=i+1}^{N} a_{ij} e_j^{(k)}, \quad i = 1, \ldots, N.$$

Date: November 29, 2005, version 1.0.
Notes originally due to James Lambers. Minor editing by Lek-Heng Lim.

Thus

$$e_i^{(k+1)} = -\sum_{j=i+1}^{N} \frac{a_{ij}}{a_{ii}} e_j^{(k)} - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(k+1)}, \quad i = 1, \ldots, N.$$

For $i = 1$, we have

$$|e_1^{(k+1)}| \leq \sum_{j=2}^{N} \left| \frac{a_{ij}}{a_{ii}} \right| |e_j^{(k)}| \leq \|\mathbf{e}^{(k)}\|_\infty r_1.$$

Assume that for $p = 1, \ldots, i-1$,

$$|e_p^{(k+1)}| \leq \|\mathbf{e}^{(k)}\|_\infty r_p \leq r \|\mathbf{e}^{(k)}\|_\infty.$$

Then,

$$\begin{aligned}
|e_i^{(k+1)}| &\leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| |e_j^{(k+1)}| + \sum_{j=i+1}^{N} \left| \frac{a_{ij}}{a_{ii}} \right| |e_j^{(k)}| \\
&\leq r \|\mathbf{e}^{(k)}\|_\infty \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \|\mathbf{e}\|_\infty \sum_{j=i+1}^{N} \left| \frac{a_{ij}}{a_{ii}} \right| \\
&\leq \|\mathbf{e}^{(k)}\|_\infty \sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| \\
&= r_i \|\mathbf{e}^{(k)}\|_\infty \\
&\leq r \|\mathbf{e}^{(k)}\|_\infty.
\end{aligned}$$

Therefore

$$\|\mathbf{e}^{(k+1)}\|_\infty \leq r \|\mathbf{e}^{(k)}\|_\infty \leq r^{k+1} \|\mathbf{e}^{(0)}\|_\infty,$$

from which it follows that

$$\lim_{k \to \infty} \|\mathbf{e}^{(k)}\| = 0$$

since $r < 1$. $\square$

We see that the Jacobi method and the Gauss-Seidel method both converge if A is diagonally dominant, but convergence can be slow in some cases. For example, if

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}$$

is of size $N \times N$ then

$$-D^{-1}(L+U) = \begin{bmatrix} 0 & 1/2 & & \\ 1/2 & \ddots & \ddots & \\ & \ddots & \ddots & 1/2 \\ & & 1/2 & 0 \end{bmatrix}$$

and therefore

$$\rho(B_J) = \cos \frac{\pi}{N+1} = \cos \pi h \approx 1 - \frac{\pi^2 h^2}{2} + \cdots$$

which is approximately 1 for small $h = \frac{1}{N+1}$. We would like to develop a method where $\rho(B) \approx 1 - ch$.

Now, suppose $B = B^\top$. Then

$$\frac{\|\mathbf{e}^{(k)}\|_2}{\|\mathbf{e}^{(0)}\|_2} \leq \|B\|_2^k = \rho(B)^k.$$

We want $\|\mathbf{e}^{(k)}\|_2/\|\mathbf{e}^{(0)}\|_2 \leq \epsilon$, so if we let $\rho^k = \epsilon$, then

$$k = \frac{-\log \epsilon}{-\log \rho}$$

is the number of iterations necessary for convergence. The quantity $-\log \rho$ is called the rate of convergence.

2. The SOR Method

The method of *successive overrelaxation* (SOR) is the iteration

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij}x_j^{(k)}\right] + (1-\omega)x_i^{(k)}.$$

The quantity ω is called the *relaxation parameter*. If $\omega = 1$, then the SOR method reduces to the Gauss-Seidel method.

In matrix form, the iteration can be written as

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

which can be rearranged to obtain

$$(D + \omega L)\mathbf{x}^{(k+1)} = \omega\mathbf{b} + [(1-\omega)D - \omega U]\mathbf{x}^{(k)}$$

or

$$\mathbf{x}^{(k+1)} = \left(\frac{1}{\omega}D + L\right)^{-1}\left[\left(\frac{1}{\omega} - 1\right)D - U\right]\mathbf{x}^{(k)} + \left(\frac{1}{\omega}D + L\right)^{-1}\mathbf{b}.$$

Define

$$\mathcal{L}_\omega = \left(\frac{1}{\omega}D + L\right)^{-1}\left[\left(\frac{1}{\omega} - 1\right)D - U\right].$$

Then

$$\begin{aligned}
\det \mathcal{L}_\omega &= \det\left(\frac{1}{\omega}D + L\right)^{-1}\det\left[\left(\frac{1}{\omega} - 1\right)D - U\right] \\
&= \frac{1}{\det\left(\frac{1}{\omega}D + L\right)}\det\left[\left(\frac{1}{\omega} - 1\right)D - U\right] \\
&= \frac{\omega^n}{\prod_{i=1}^n a_{ii}}\frac{(1-\omega)^n\prod_{i=1}^n a_{ii}}{\omega^n} \\
&= (1-\omega)^n.
\end{aligned}$$

Therefore, $\prod_{i=1}^n \lambda_i = (1-\omega)^n$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathcal{L}_\omega$, with $|\lambda_1| \geq \cdots \geq |\lambda_n|$. Therefore $|\lambda_1|^n \geq (1-\omega)^n$. Since we must have $|\lambda_1| < 1$ for convergence, it follows that a necessary condition for convergence of SOR is

$$0 < \omega < 2.$$

3. Block Methods

Recall that in solving Poisson's equation on a rectangle, we needed to solve systems of the form

$$-\mathbf{v}_j + T\mathbf{v}_j - \mathbf{v}_{j+1} = \mathbf{g}_j.$$

This can be accomplished using an iteration

$$T\mathbf{v}^{(k+1)} = \mathbf{g}_j + \mathbf{v}_{j-1}^{(k)} + \mathbf{v}_{j+1}^{(k)},$$

which is an example of a *block Jacobi* iteration, since it involves solving the system $A\mathbf{u} = \mathbf{g}$ by applying the Jacobi method to A, except each block of size $N \times N$ is treated as a single element. Similarly, we can use the *block Gauss-Seidel* iteration

$$T\mathbf{v}_j^{(k+1)} = \mathbf{g}_j + \mathbf{v}_{j-1}^{(k+1)} + \mathbf{v}_j^{(k)}.$$

4. Richardson Method

Consider the iteration

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= (I - \alpha A)\mathbf{x}^{(k)} + \alpha\mathbf{b} \\
&= \mathbf{x}^{(k)} + \alpha(\mathbf{b} - A\mathbf{x}^{(k)}) \\
&= \mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)}
\end{aligned}$$

This is known as the *Richardson method.* If we define the error $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$, then $\mathbf{e}^{(k+1)} = B_\alpha\mathbf{e}^{(k)}$ where $B_\alpha = I - \alpha A$; we want to choose the parameter α a priori so as to minimize $\|B_\alpha\|$.

Suppose A is symmetric positive definite, with eigenvalues

$$\mu_1 \geq \mu_2 \geq \cdots \mu_n > 0.$$

Since $B = I - \alpha A$, $\lambda_i = 1 - \alpha\mu_i$. We want to choose α so that $\|B_\alpha\|_2$ is minimized; i.e.

$$\min_\alpha \max_{1\leq i\leq n} |\lambda_i(\alpha)| = \min_\alpha \max_{1\leq i\leq n} |1 - \alpha\mu_i|.$$

The optimal parameter $\hat{\alpha}$ is found by solving

$$1 - \hat{\alpha}\mu_n = -(1 - \hat{\alpha}\mu_1)$$

which yields

$$\hat{\alpha} = \frac{2}{\mu_1 + \mu_n}.$$

Note that When $1 - \alpha\mu_n = -1$ that the iteration diverges, from which it follows that the method converges for $0 < \alpha < 2/\mu_n$. However, this iteration is sensitive to perturbation, and therefore bad numerically. For example, if $\mu_1 = 10$ and $\mu_n = 10^{-4}$, then the optimal α is $2/(10 + 10^{-4}$, but this is close to a value of α for which the iteration diverges, $\alpha = 2/10$.

Also, note that

$$\lambda_1(\hat{\alpha}) = 1 - \frac{2}{\mu_1 + \mu_n}\mu_1 = \frac{\mu_n - \mu_1}{\mu_1 + \mu_n},$$

and similarly,

$$\lambda_n(\hat{\alpha}) = \frac{\mu_1 - \mu_n}{\mu_1 + \mu_n} = \frac{\frac{\mu_1}{\mu_n} - 1}{\frac{\mu_1}{\mu_n} + 1} = \frac{\kappa(A) - 1}{\kappa(A) + 1}.$$

Therefore the convergence rate depends on $\kappa(A)$.

For example, consider the Helmholtz equation on a rectangle R,

$$\begin{aligned}
-\Delta\mathbf{u}^{(k+1)} + \sigma(x,y)\mathbf{u}^{(k)} &= \mathbf{f}, \quad (x,y) \in R \\
\mathbf{u} &= \mathbf{g}, \quad (x,y) \in \partial R
\end{aligned}$$

Using a finite difference approximation for Δ gives

$$A = \begin{bmatrix} T & -I & & \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & T \end{bmatrix}$$

and thus the iteration has the form

$$A\mathbf{u}^{(k+1)} + h^2\Sigma\mathbf{u}^{(k)} = \mathbf{f}$$

where

$$\Sigma = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{nn} \end{bmatrix}, \quad \sigma_{ij} = \sigma(x_i, y_j).$$

We wish to determine the rate of convergence. We define the *error operator* by

$$\mathbf{e}^{(k+1)} = (h^2 A^{-1} \Sigma)\mathbf{e}^{(k)}.$$

Therefore

$$\|\mathbf{e}^{(k+1)}\|_2 \leq h^2 \|A^{-1}\|_2 \|\Sigma\|_2 \|\mathbf{e}^{(k)}\|_2.$$

But

$$\|\Sigma\|_2 = \max_{i,j} |\sigma_{ij}|$$

and

$$\begin{aligned} \lambda_{\min} &= 4 - 4\cos \pi h \\ &= 4(1 - \cos \pi h) \\ &= 8 \sin^2 \left(\frac{\pi h}{2}\right) \end{aligned}$$

Therefore

$$\|\mathbf{e}^{(k+1)}\|_2 \leq \frac{\max_{i,j} |\sigma_{ij}|}{2\left(\frac{\sin xh/2}{h/2}\right)^2} \|\mathbf{e}\|_2 \approx \frac{\max_{i,j} |\sigma_{ij}|}{2\pi^2} \|\mathbf{e}^{(k)}\|_2$$

and thus the size of the problem mesh has disappeared, and the method converges if $\max_{i,j} |\sigma_{ij}| \leq 20$. The rate of convergence is essentially independent of h, which is very desirable.

Department of Computer Science, Gates Building 2B, Room 280, Stanford, CA 94305-9025
E-mail address: `golub@stanford.edu`