Iterative Methods for Solving Large Linear Systems (I)

Tsung-Ming Huang

Department of Mathematics
National Taiwan Normal University

October 25, 2011

Outline

General procedures for the construction of iterative methods

Given a linear system of nonsingular A

$$Ax = b, \tag{1}$$

we consider the splitting of A

$$A = M - N \tag{2}$$

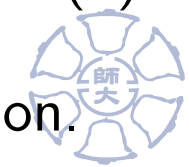
with M nonsingular. Then (1) is equivalent to $Mx = Nx + b$, or

$$x = M^{-1}Nx + M^{-1}b \equiv Tx + f.$$

This suggests an iterative process

$$x_{k+1} = Tx_k + f = M^{-1}Nx_k + M^{-1}b, \tag{3}$$

where x_0 is given. Then the solution x of (1) is determined by iteration.



Remark 1

(a) *Define* $\varepsilon_k = x_k - x$. *Then*

$$\begin{aligned}\varepsilon_{k+1} &= x_{k+1} - x = M^{-1}Nx_k + M^{-1}b - M^{-1}Nx - M^{-1}b \\ &= (M^{-1}N)\varepsilon_k = (M^{-1}N)^k\varepsilon_0\end{aligned}$$

which implies that $\rho(M^{-1}N) < 1$ *if and only if* $\{\varepsilon_k\} \to 0$.

(b) *Let* $r_k = b - Ax_k$. *Then,*

$$\begin{aligned}x_{k+1} &= M^{-1}Nx_k + M^{-1}b \\ &= M^{-1}(M - A)x_k + M^{-1}b \\ &= x_k + M^{-1}(b - Ax_k) \\ &= x_k + z_k\end{aligned}$$

where $Mz_k = r_k$.

Example 1

We consider the standard splitting of A

$$A = D - L - R, \tag{4}$$

where $A = [a_{ij}]_{i,j=1}^{n}$, $D = \mathsf{diag}(a_{11}, a_{22}, \cdots, a_{nn})$,

$$-L = \begin{bmatrix} 0 & & & 0 \\ a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix},$$

$$-R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & 0 & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & & & 0 \end{bmatrix}.$$

For $a_{i,i} \neq 0, i = 1, \ldots, n$, D is nonsingular. If we choose

$$M = D \quad \text{and} \quad N = L + R$$

in (2), we then obtain the Jacobi Method (Total-step Method):

$$x_{k+1} = D^{-1}(L + R)x_k + D^{-1}b$$

or in formula

$$x_{k+1,j} = \frac{1}{a_{jj}}(-\sum_{i \neq j} a_{ji}x_{k,i} + b_j), \ j = 1, \ldots, n, \ k = 0, 1, \ldots.$$

Example 2

If $D - L$ is nonsingular in (4), then we choose

$$M = D - L, \quad N = R$$

as in (2) are possible and yields the so-called Gauss-Seidel Method (Single-Step Method):

$$x_{k+1} = (D - L)^{-1} R x_k + (D - L)^{-1} b$$

or in formula

$$x_{k+1,j} = \frac{1}{a_{jj}}(-\sum_{i<j} a_{ji} x_{k+1,i} - \sum_{i>j} a_{ji} x_{k,i} + b_j), \ j = 1, \ldots, n, \ k = 1, 2, \ldots.$$

- Total-Step Method = TSM = Jacobi method.
- Single-Step Method = SSM = Gauss-Seidel method.

We consider the following points on Examples 1 and 2:

(i) flops counts per iteration step.

(ii) Convergence speed.

Let $\|\cdot\|$ be a vector norm, and $\|T\|$ be the corresponding operator norm. Then

$$\frac{\|\varepsilon_m\|}{\|\varepsilon_0\|} = \frac{\|T^m \varepsilon_0\|}{\|\varepsilon_0\|} \leq \|T^m\|. \tag{5}$$

Here $\|T^m\|^{\frac{1}{m}}$ is a measure for the average of reduction of error ε_m per iteration step. We call

$$R_m(T) = -\ln(\|T^m\|^{\frac{1}{m}}) = -\frac{1}{m}\ln(\|T^m\|) \tag{6}$$

the average of convergence rate for m iterations.

The larger is $R_m(T)$, so the better is convergence rate. Let $\sigma = (\|\varepsilon_m\|/\|\varepsilon_0\|)^{\frac{1}{m}}$. From (5) and (6) we get

$$\sigma \leq \|T^m\|^{\frac{1}{m}} \leq e^{-R_m(T)},$$

or

$$\sigma^{1/R_m(T)} \leq \frac{1}{e}.$$

That is, after $1/R_m(T)$ steps in average the error is reduced by a factor of $1/e$. Since $R_m(T)$ is not easy to determine, we consider $m \to \infty$. Since

$$\lim_{m\to\infty} \|T^m\|^{\frac{1}{m}} = \rho(T),$$

it follows

$$R_\infty(T) = \lim_{m\to\infty} R_m(T) = -\ln \rho(T).$$

R_∞ is called the asymptotic convergence rate. It holds always $R_m(T) \leq R_\infty(T)$.

Example 3

Consider the Dirichlet boundary-value problem (Model problem):

$$-\Delta u \equiv -u_{xx} - u_{yy} = f(x,y), \quad 0 < x, y < 1, \tag{7}$$

$$u(x,y) = 0 \ \ (x,y) \in \partial\Omega,$$

for the unit square $\Omega := \{x, y | 0 < x, y < 1\} \subseteq \mathbb{R}^2$ *with boundary* $\partial\Omega$.

To solve (7) by means of a difference methods, one replaces the differential operator by a difference operator. Let

$$\begin{aligned} \Omega_h &:= \{(x_i, y_i) | i, j = 1, \ldots, N+1\}, \\ \partial\Omega_h &:= \{(x_i, 0), (x_i, 1), (0, y_j), (1, y_j) | i, j = 0, 1, \ldots, N+1\}, \end{aligned}$$

where
$x_i = ih,\ y_j = jh,\ i, j = 0, 1, \ldots, N+1,\ h := \frac{1}{N+1},\ N \geq 1,$ an integer.

The differential operator $-u_{xx} - u_{yy}$ can be replaced for all $(x_i, y_i) \in \Omega_h$ by the difference operator:

$$\frac{4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} \tag{8}$$

up to an error $\tau_{i,j}$. For small h one can expect that the solution $z_{i,j}$, for $i, j = 1, \ldots, N$, of the linear system

$$\begin{aligned} &4z_{i,j} - z_{i-1,j} - z_{i+1,j} - z_{i,j-1} - z_{i,j+1} = h^2 f_{i,j}, i, j = 1, \ldots, N, \quad (9) \\ &z_{0,j} = z_{N+1,j} = z_{i,0} = z_{i,N+1} = 0, i, j = 0, 1, \ldots, N+1, \end{aligned}$$

obtained from (8) by omitting the error $\tau_{i,j}$, agrees approximately with the $u_{i,j}$. Let

$$z = [z_{1,1}, z_{2,1}, \cdots, z_{N,1}, z_{1,2}, \cdots, z_{N,2}, \cdots, z_{1,N}, \cdots, z_{N,N}]^T$$

and

$$b = h^2[f_{1,1}, \cdots, f_{N,1}, f_{1,2}, \cdots, f_{N,2}, \cdots, f_{1,N}, \cdots, f_{N,N}]^T.$$

Then (9) is equivalent to a linear system $Az = b$ with the $N^2 \times N^2$ matrix.

$$
\left[\begin{array}{cccc|cccc|cccc|cccc}
4 & -1 & & & -1 & & & & & & & & & & & \\
-1 & \ddots & \ddots & & & \ddots & & & & & & & & & & \\
 & \ddots & \ddots & -1 & & & \ddots & & & & & & & & & \\
 & & -1 & 4 & & & & -1 & & & & & & & & \\
\hline
-1 & & & & 4 & -1 & & & \ddots & & & & & & & \\
 & \ddots & & & -1 & \ddots & \ddots & & & \ddots & & & & & & \\
 & & \ddots & & & & \ddots & -1 & & & \ddots & & & & & \\
 & & & -1 & & & -1 & 4 & & & & \ddots & & & & \\
\hline
 & & & & \ddots & & & & \ddots & & & & -1 & & & \\
 & & & & & \ddots & & & & \ddots & & & & \ddots & & \\
 & & & & & & \ddots & & & & \ddots & & & & \ddots & \\
 & & & & & & & \ddots & & & & \ddots & & & & -1 \\
\hline
 & & & & & & & & -1 & & & & 4 & -1 & & \\
 & & & & & & & & & \ddots & & & -1 & \ddots & \ddots & \\
 & & & & & & & & & & \ddots & & & \ddots & \ddots & -1 \\
 & & & & & & & & & & & -1 & & -1 & 4 &
\end{array}\right] \tag{11}
$$

Let $A = D - L - R$. The matrix $J = D^{-1}(L + R)$ belongs to the Jacobi method (TSM). The N^2 eigenvalues and eigenvectors of J can be determined explicitly. We can verify at once, by substitution, that N^2 vectors $z^{(k,l)}$, $k, l = 1, \ldots, N$ with components

$$z_{i,j}^{(k,l)} := \sin \frac{k\pi i}{N+1} \sin \frac{l\pi j}{N+1}, \quad 1 \le i, j \le N,$$

satisfy

$$Jz^{(k,l)} = \lambda^{(k,l)} z^{(k,l)}$$

with

$$\lambda^{(k,l)} := \frac{1}{2}(\cos \frac{k\pi}{N+1} + \cos \frac{l\pi}{N+1}), \quad 1 \le k, l \le N.$$

J thus has eigenvalues $\lambda^{(k,l)}$, $1 \le k, l \le N$. Then we have

$$\rho(J) = \lambda_{1,1} = \cos\frac{\pi}{N+1} = 1 - \frac{\pi^2 h^2}{2} + O(h^4) \tag{12}$$

and

$$R_\infty(J) = -\ln(1 - \frac{\pi^2 h^2}{2} + O(h^4)) = \frac{\pi^2 h^2}{2} + O(h^4).$$

These show that

(i) TSM converges;

(ii) Diminution of h will not only enlarge the flop counts per step, but also the convergence speed will drastically make smaller.

Some theorems and definitions

$\rho(T)$: A measure of quality for convergence.

Definition 4

A real $m \times n$-matrix $A = (a_{ik})$ is called nonnegative (positive), denoted by $A \geq 0$ $(A > 0)$, if $a_{ik} \geq 0$ (> 0), $i = 1, \ldots, m$, $k = 1, \ldots, n$.

Definition 5

An $m \times n$-matrix A is called reducible, if there is a subset $I \subset \tilde{N} \equiv \{1, 2, \ldots, n\}$, $I \neq \phi$, $I \neq \tilde{N}$ such that $i \in I$, $j \notin I \Rightarrow a_{ij} = 0$.
A is not reducible $\Leftrightarrow$ A is irreducible.

Remark 2

$G(A)$ is the directed graph associated with the matrix A. If A is an $n \times n$-matrix, then $G(A)$ consists of n vertices $P_1, \cdots, P_n$ and there is an (oriented) arc $P_i \to P_j$ in $G(A)$ precisely if $a_{ij} \neq 0$.

It is easily shown that A is irreducible if and only if the graph $G(A)$ is connected in the sense that for each pair of vertices (P_i, P_j) in $G(A)$ there is an oriented path from P_i to P_j. i.e., if $i \neq j$, there is a sequence of indices $i = i_1, i_2, \cdots, i_s = j$ such that $(a_{i_1,i_2} \cdots a_{i_{s-1},i_s}) \neq 0$.

Theorem 6 (Perron-Frobenius)

Let $A \geq 0$ irreducible. Then

(i) $\rho = \rho(A)$ *is a simple eigenvalue;*

(ii) *There is a positive eigenvector z associated to ρ, i.e., $Az = \rho z, z > 0$;*

(iii) *If $Ax = \lambda x$, $x \geq 0$, then $\lambda = \rho$, $x = \alpha z$, $\alpha > 0$;*

(iv) $A \leq B, A \neq B \Longrightarrow \rho(A) < \rho(B)$.

Theorem 7

Let $A \geq 0, x > 0$. Define the quotients:

$$q_i(x) \equiv \frac{(Ax)_i}{x_i} = \frac{1}{x_i} \sum_{k=1}^{n} a_{ik} x_k, \quad \text{for } i = 1, \ldots, n.$$

Then

$$\min_{1 \leq i \leq n} q_i(x) \leq \rho(A) \leq \max_{1 \leq i \leq n} q_i(x). \tag{13}$$

If A is irreducible, then it holds additionally, either

$$q_1 = q_2 = \cdots = q_n \ (\text{then } x = \mu z, \quad q_i = \rho(A))$$

or

$$\min_{1 \leq i \leq n} q_i(x) < \rho(A) < \max_{1 \leq i \leq n} q_i(x). \tag{14}$$

Theorem 8

The statements in Theorem 7 can be formulated as: Let $A \geq 0, x > 0$. *(13) corresponds:*

$$\begin{cases} Ax \leq \mu x & \Rightarrow \quad \rho \leq \mu, \\ \\ Ax \geq \nu x & \Rightarrow \quad \nu \leq \rho. \end{cases} \tag{15}$$

Let $A \geq 0$, *irreducible,* $x > 0$. *(14) corresponds :*

$$\begin{cases} Ax \leq \mu x, \ \ Ax \neq \mu x & \Rightarrow \quad \rho < \mu, \\ Ax \geq \nu x, \ \ Ax \neq \nu x & \Rightarrow \quad \nu < \rho. \end{cases} \tag{16}$$

Definition 9

A real matrix B is called an M-matrix if $b_{ij} \leq 0, i \neq j$ and B^{-1} exists with $B^{-1} \geq 0$.

Sufficient conditions for convergence of TSM and SSM

Theorem 10

Let B be a real matrix with $b_{ij} \leq 0$ for $i \neq j$. Then the following statements are equivalent.

(i) *B is an $M-$matrix.*

(ii) *There exists a vector $v > 0$ so that $Bv > 0$.*

(iii) *B has a decomposition $B = sI - C$ with $C \geq 0$ and $\rho(C) < s$.*

(iv) *For each decomposition $B = D - C$ with $D = \text{diag}\,(d_i)$ and $C \geq 0$, it holds: $d_i > 0$, $i = 1, 2, \ldots, n$, and $\rho(D^{-1}C) < 1$.*

(v) *There is a decomposition $B = D - C$, with $D = diag(d_i)$ and $C \geq 0$ it holds: $d_i > 0, i = 1, 2, \ldots, n$ and $\rho(D^{-1}C) < 1$.*
Further, if B is irreducible, then (vi) is equivalent to (i)-(v).

(vi) *There exists a vector $v > 0$ so that $Bv \geq 0, \neq 0$.*

▸ Proof

Lemma 11

Let A be an arbitrary complex matrix and define $|A| = [|a_{ij}|]$. If $|A| \le C$, then $\rho(A) \le \rho(C)$. Especially $\rho(A) \le \rho(|A|)$.

▸ Proof

Theorem 12

Let A be an arbitrary complex matrix. It satisfies
either (Strong Row Sum Criterion)*:*

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|, \quad i = 1, \ldots, n. \tag{17}$$

or (Weak Row Sum Criterion)*:*

$$\begin{aligned} \sum_{j \neq i} |a_{ij}| &\le |a_{ii}|, \ i = 1, \ldots, n, \\ &< |a_{i_0 i_0}|, \ \textit{at least one } i_0, \end{aligned} \tag{18}$$

for A irreducible. Then TSM(Jacobi) and SSM(GS) both are convergent.

▸ Proof

Relaxation Methods (Successive Over-Relaxation (SOR) Method)

Consider the parametrized linear system $\omega Ax = \omega b$ and consider the splitting

$$\begin{aligned} \omega A &= \omega D - \omega L - \omega R + D - D \\ &= (D - \omega L) - ((1-\omega)D + \omega R) \equiv M - N. \end{aligned}$$

From (3) we have the iteration

$$x_{k+1} = (D - \omega L)^{-1} \left((1-\omega)D + \omega R\right) x_k + \omega (D - \omega L)^{-1} b. \tag{19}$$

From Remark 1 (b) the iteration (19) is equivalent to

$$x_{k+1} = x_k + \omega z_k$$

where

$$(D - \omega L) z_k = r_k \equiv b - A x_k.$$

Define

$$L_\omega := (D - \omega L)^{-1} \left((1-\omega) D + \omega R \right).$$

We may assume $D = I$, i.e.,

$$L_\omega := (I - \omega L)^{-1} \left((1-\omega) I + \omega R \right).$$

Otherwise, we can let $\tilde{A} = D^{-1}A$, $\tilde{L} = D^{-1}L$, $\tilde{R} = D^{-1}R$. Then it holds that

$$\tilde{A} = I - \tilde{L} - \tilde{R}.$$

$\omega < 1$: under relaxation

$\omega = 1$: single-step method (GS)

$\omega > 1$: over relaxation.

We now try to choose an ω such that $\rho(L_\omega)$ is small as possible. But this is only under some special assumptions possible. we first list a few qualitative results about $\rho(L_\omega)$.

Theorem 13

Let $A = D - L - L^*$ be Hermitian and positive definite. Then the relaxation method is convergent for $0 < \omega < 2$.

Theorem 14

Let A be Hermitian and nonsingular with positive diagonal. If SSM converges, then A is positive definite.

Determination of the Optimal Parameter ω for 2-consistly Ordered Matrices

For an important class of matrices the more qualitative assertions of Theorems 13 and 14 can be considerably sharpened. This is the class of consistly ordered matrices. The optimal parameter ω_b with

$$\rho(L_{\omega_b}) = \min_{\omega} \rho(L_\omega)$$

can be determined. We consider $A = I - L - R$.

Definition 15

A is called 2-consistly ordered, if the eigenvalues of $\alpha L + \alpha^{-1} R$ are independent of α.

If A is 2-consistly ordered, then $L+R$ and $-(L+R)$ $(\alpha=-1)$ has the same eigenvalues. The nonzero eigenvalues of $L+R$ appear in pairs. Hence

$$\det(\lambda I - L - R) = \lambda^m \prod_{i=1}^{r}(\lambda^2 - \mu_i^2),\ n = 2r+m\ (m=0,\ \text{possible}). \quad (20)$$

Theorem 16

Let A be 2-consistly ordered, $a_{ii}=1$, $\omega \neq 0$. Then hold:

(i) *If $\lambda \neq 0$ is an eigenvalue of L_ω and μ satisfies the equation*

$$(\lambda+\omega-1)^2 = \lambda\mu^2\omega^2, \quad (21)$$

then μ is an eigenvalue of $L+R$ (so is $-\mu$).

(ii) *If μ is an eigenvalue of $L+R$ and λ satisfies the equation (21), then λ is an eigenvalue of L_ω.*

Remark 3

If $\omega = 1$, then $\lambda = \mu^2$, and $\rho((I-L)^{-1}R) = (\rho(L+R))^2$.

Proof: We first prove the identity

$$\det(\lambda I - sL - rR) = \det(\lambda I - \sqrt{sr}(L+R)). \tag{22}$$

Since both sides are polynomials of the form $\lambda^n + \cdots$ and

$$sL + rR = \sqrt{sr}(\sqrt{\frac{s}{r}}L + \sqrt{\frac{r}{s}}R) = \sqrt{sr}(\alpha L + \alpha^{-1}R),$$

if $sr \neq 0$, then $sL + rR$ and $\sqrt{sr}(L+R)$ have the same eigenvalues. It is obviously also for the case $sr = 0$. The both polynomials in (22) have the same roots, so they are identical.

For

$$\begin{aligned}\det(I-\omega L)\det(\lambda I - L_\omega) &= \det(\lambda(I-\omega L)-(1-\omega)I-\omega R)\\ &= \det((\lambda+\omega-1)I-\omega\lambda L-\omega R) = \Phi(\lambda)\end{aligned}$$

and $\det(I-\omega L)\neq 0$, λ is an eigenvalue of L_ω if and only if $\Phi(\lambda)=0$. From (22) follows

$$\Phi(\lambda)=\det((\lambda+\omega-1)I-\omega\sqrt{\lambda}(L+R))$$

and that is (from (20))

$$\Phi(\lambda)=(\lambda+\omega-1)^m\prod_{i=1}^{r}((\lambda+\omega-1)^2-\lambda\mu_i^2\omega^2), \tag{23}$$

where μ_i is an eigenvalue of $L+R$. Therefore, if μ is an eigenvalue of $(L+R)$ and λ satisfies (21), so is $\Phi(\lambda)=0$, then λ is eigenvalue of L_ω. This shows (ii).

Now if $\lambda \neq 0$ an eigenvalue of L_ω, then one factor in (23) must be zero. Let μ satisfy (21). Then

(i) $\mu \neq 0$: From (21) follows $\lambda + \omega - 1 \neq 0$, so

$$\begin{aligned} (\lambda + \omega - 1)^2 &= \lambda\omega^2\mu_i^2, \quad \text{for one } i \text{ (from (23))}, \\ &= \lambda\omega^2\mu^2, \quad \text{(from (21))}. \end{aligned}$$

This shows that $\mu = \pm\mu_i$, so μ is an eigenvalue of $L + R$.

(ii) $\mu = 0$: We have $\lambda + \omega - 1 = 0$ and

$$0 = \Phi(\lambda) = \det((\lambda + \omega - 1)I - \omega\sqrt{\lambda}(L + R)) = \det(-\omega\sqrt{\lambda}(L + R)),$$

i.e., $L + R$ is singular, so $\mu = 0$ is eigenvalue of $L + R$. ■

Theorem 17

Let $A = I - L - R$ be 2-consistly ordered. If $L + R$ has only real eigenvalues and satisfies $\rho(L + R) < 1$, then it holds

$$\rho(L_{\omega_b}) = \omega_b - 1 < \rho(L_\omega), \quad \text{for } \omega \neq \omega_b,$$

where

$$\omega_b = \frac{2}{1 + \sqrt{1 - \rho^2(L + R)}} \quad \text{(solve } \omega_b \text{ in (21)).}$$

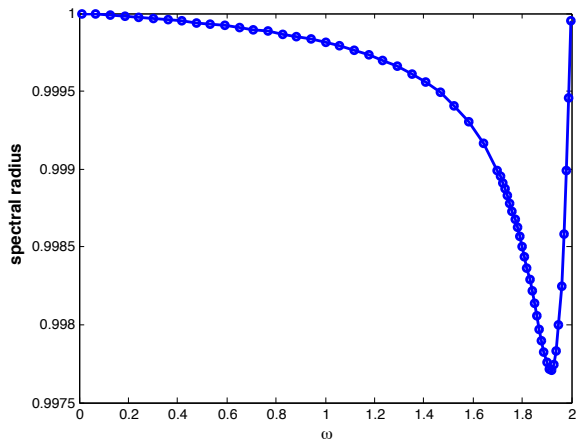


Figure: figure of $\rho(L_{\omega_b})$

Theorem 18

One has in general,

$$\rho(L_\omega) = \begin{cases} \omega - 1, & \text{for } \omega_b \leq \omega \leq 2 \\ 1 - \omega + \frac{1}{2}\omega^2\mu^2 + \omega\mu\sqrt{1 - \omega + \frac{1}{4}\omega^2\mu^2}, & \text{for } 0 < \omega \leq \omega_b \end{cases} \tag{24}$$

Remark: We first prove the following Theorem proposed by Kahan: For arbitrary matrices A it holds

$$\rho(L_\omega) \geq |\omega - 1|, \text{ for all } \omega. \tag{25}$$

Since $\det(I - \omega L) = 1$ for all ω, the characteristic polynomial $\Phi(\lambda)$ of L_ω is

$$\begin{aligned} \Phi(\lambda) &= \det(\lambda I - L_\omega) = \det((I - \omega L)(\lambda I - L_\omega)) \\ &= \det((\lambda + \omega - 1)I - \omega\lambda L - \omega R). \end{aligned}$$

For $\prod_{i=1}^{n} \lambda_i(L_\omega) = \Phi(0) = \det((\omega - 1)I - \omega R) = (\omega - 1)^n$, it follows immediately that

$$\rho(L_\omega) = \max_i |\lambda_i(L_\omega)| \geq |\omega - 1|.$$

Proof of Theorem: By assumption the eigenvalues μ_i of $L+R$ are real and $-\rho(L+R) \le \mu_i \le \rho(L+R) < 1$. For a fixed $\omega \in (0,2)$ (by (25) in the Remark it suffices to consider the interval (0,2)) and for each μ_i there are two eigenvalues $\lambda_i^{(1)}(\omega,\mu_i)$ and $\lambda_i^{(2)}(\omega,\mu_i)$ of L_ω, which are obtained by solving the quadratic equation (21) in λ.

Geometrically, $\lambda_i^{(1)}(\omega)$ and $\lambda_i^{(2)}(\omega)$ are obtained as abscissae of the points of intersection of

$$\text{the straight line } g_\omega(\lambda) = \frac{\lambda+\omega-1}{\omega}$$

and

$$\text{the parabola } m_i(\lambda) := \pm\sqrt{\lambda}\mu_i$$

(see Figure 2). The line $g_\omega(\lambda)$ has the slope $1/\omega$ and passes through the point (1,1). If $g_\omega(\lambda) \cap m_i(\lambda) = \phi$, then $\lambda_i^{(1)}(\omega)$ and $\lambda_i^{(2)}(\omega)$ are conjugate complex with modulus $|\omega-1|$ (from (21)).

Evidently

$$\rho(L_\omega) = \max_i(|\lambda_i^{(1)}(\omega)|, |\lambda_i^{(2)}(\omega)|) = \max(|\lambda^{(1)}(\omega)|, |\lambda^{(2)}(\omega)|),$$

where $\lambda^{(1)}(\omega)$, $\lambda^{(2)}(\omega)$ being obtained by intersecting $g_\omega(\lambda)$ with $m(\lambda) := \pm\sqrt{\lambda}\mu$, with $\mu = \rho(L+R) = \max_i |\mu_i|$. By solving (21) with $\mu = \rho(L+R)$ for λ, one verifies (24) immediately, and thus also the remaining assertions of the theorem. ■

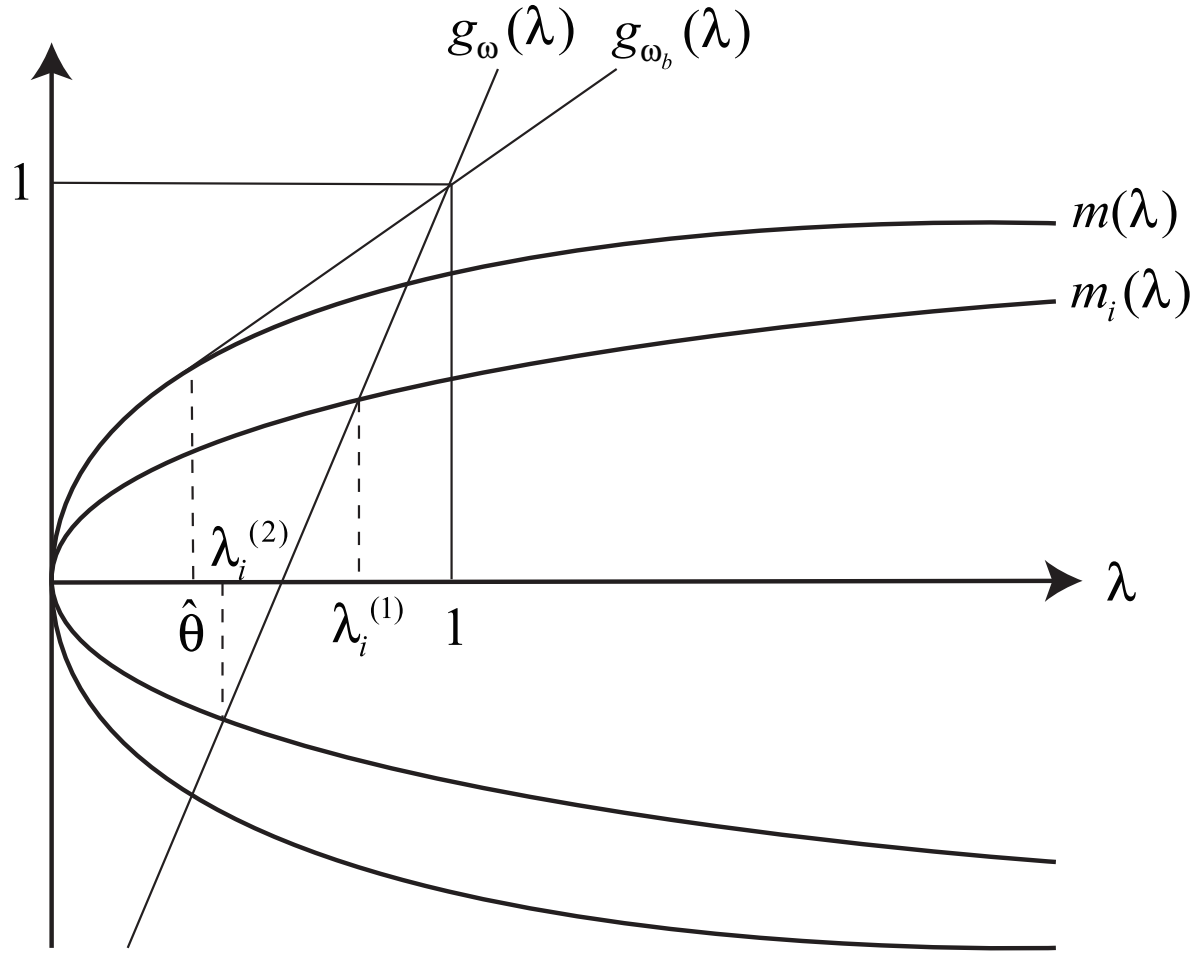


Figure: Geometrical view of $\lambda_i^{(1)}(\omega)$ and $\lambda_i^{(2)}(\omega)$.

Application to Finite Difference Methods: Model Problem

We consider the Dirichlet boundary-value problem (Model problem) as in Example 3. We shall solve a linear system $Az = b$ of the $N^2 \times N^2$ matrix A as in (11).

To Jacobi method: The iterative matrix is

$$J = L + R = \frac{1}{4}(4I - A).$$

It is easily seen that A is 2-consistly ordered (Exercise!).

To Gauss-Seidel method: The iterative matrix is

$$H = (I - L)^{-1}R.$$

From the Remark of Theorem 16 and (12) follows that

$$\rho(H) = \rho(J)^2 = \cos^2 \frac{\pi}{N+1}.$$

According to Theorem 17 the optimal relaxation parameter ω_b and $\rho(L_{\omega_b})$ are given by

$$\omega_b = \frac{2}{1 + \sqrt{1 - \cos^2 \frac{\pi}{N+1}}} = \frac{2}{1 + \sin \frac{\pi}{N+1}}$$

and

$$\rho(L_{\omega_b}) = \frac{\cos^2 \frac{\pi}{N+1}}{(1 + \sin \frac{\pi}{N+1})^2}.$$

The number $k = k(N)$ with $\rho(J)^k = \rho(L_{\omega_b})$ indicates that the k steps of Jacobi method produce the same reduction as one step of the optimal relaxation method. Clearly,

$$k = \ln \rho(L_{\omega_b}) / \ln \rho(J).$$

Now for small z one has $\ln(1+z) = z - z^2/2 + O(z^3)$ and for large N we have

$$\cos\left(\frac{\pi}{N+1}\right) = 1 - \frac{\pi^2}{2(N+1)^2} + O(\frac{1}{N^4}).$$

Thus that

$$\ln \rho(J) = \frac{\pi^2}{2(N+1)^2} + O(\frac{1}{N^4}).$$

Similarly,

$$\begin{aligned}
\ln \rho(L_{\omega_b}) &= 2[\ln \rho(J) - \ln(1 + \sin \frac{\pi}{N+1})] \\
&= 2[-\frac{\pi^2}{2(N+1)^2} - \frac{\pi}{N+1} + \frac{\pi^2}{2(N+1)^2} + O(\frac{1}{N^3})] \\
&= -\frac{2\pi}{N+1} + O(\frac{1}{N^3}) \text{ (for large } N).
\end{aligned}$$

and

$$k = k(N) \approx \frac{4(N+1)}{\pi}.$$

The optimal relaxation method is more than N times as fast as the Jacobi method. The quantities

$$R_J := \frac{-\ln 10}{\ln \rho(J)} \approx 0.467(N+1)^2. \tag{26}$$

$$R_H := \frac{1}{2} R_J \approx 0.234(N+1)^2 \tag{27}$$

$$R_{L_{\omega_b}} := -\frac{\ln 10}{\ln \rho(L_{\omega_b})} \approx 0.367(N+1) \tag{28}$$

indicate the number of iterations required in the Jacobi, the Gauss-Seidel method, and the optimal relaxation method, respectively, in order to reduce the error by a factor of $1/10$.

SSOR (Symmetric Successive Over Relaxation):

A is symmetric and $A = D - L - L^T$. Let

$$\begin{cases} M_\omega := D - \omega L, \\ N_\omega := (1-\omega)D + \omega L^T, \end{cases} \quad \text{and} \quad \begin{cases} M_\omega^T = D - \omega L^T, \\ N_\omega^T = (1-\omega)D + \omega L. \end{cases}$$

Then from the iterations

$$\begin{aligned} M_\omega x_{i+1/2} &= N_\omega x_i + \omega b, \\ M_\omega^T x_{i+1} &= N_\omega^T x_{i+1/2} + \omega b, \end{aligned}$$

follows that

$$\begin{aligned} x_{i+1} &= \left(M_\omega^{-T} N_\omega^T M_\omega^{-1} N_\omega\right) x_i + \tilde{b} \\ &\equiv G x_i + \omega \left(M_\omega^{-T} N_\omega^T M_\omega^{-1} + M_\omega^{-T}\right) b \\ &\equiv G x_i + M(\omega)^{-1} b. \end{aligned}$$

It holds that

$$\begin{aligned}
&((1-\omega)D+\omega L)\,(D-\omega L)^{-1}+I \\
&=(\omega L-D-\omega D+2D)(D-\omega L)^{-1}+I \\
&=-I+(2-\omega)D(D-\omega L)^{-1}+I \\
&=(2-\omega)D(D-\omega L)^{-1},
\end{aligned}$$

Thus

$$M(\omega)^{-1}=\omega\left(D-\omega L^{T}\right)^{-1}(2-\omega)D(D-\omega L)^{-1},$$

then

$$\begin{aligned}
M(\omega) &= \frac{1}{\omega(2-\omega)}(D-\omega L)D^{-1}\left(D-\omega L^{T}\right) \qquad (29)\\
&\approx (D-L)D^{-1}\left(D-L^{T}\right),\ (\omega=1).
\end{aligned}$$

Proof of Theorem

(1) $\Longrightarrow$ (2): Let $e = (1, \cdots, 1)^T$. Since $B^{-1} \geq 0$ is nonsingular it follows $\nu = B^{-1}e > 0$ and $B\nu = B(B^{-1}e) = e > 0$.

(2) $\Longrightarrow$ (3): Let $s > \max(b_{ii})$. It follows $B = sI - C$ with $C \geq 0$. There exists a $\nu > 0$ with $B\nu = s\nu - C\nu$ (via (2)), also $s\nu > C\nu$. From the statement (15) in Theorem 8 follows $\rho(C) < s$.

(3) $\Longrightarrow$ (1): $B = sI - C = s(I - \frac{1}{s}C)$. For $\rho(\frac{1}{s}C) < 1$ and from Theorem 2.6 $(I - \frac{1}{s}C)^{-1}$ follows that there exists a series expansion $\sum_{\nu=0}^{\infty}(\frac{1}{s}C)^k$. Since the terms in sum are nonnegtive, we get $B^{-1} = \frac{1}{s}(I - \frac{1}{s}C)^{-1} \geq 0$.

(2) $\Longrightarrow$ (4): From $B\nu = D\nu - C\nu > 0$ follows $D\nu > C\nu \geq 0$ and $d_i > 0$, for $i = 1, 2, \cdots, n$. Hence $D^{-1} \geq 0$ and $\nu > D^{-1}C\nu \geq 0$. From (15) follows that $\rho(D^{-1}C) < 1$.

(4) $\Longrightarrow$ (5): Trivial.

(5) $\Longrightarrow$ (1): Since $\rho(D^{-1}C) < 1$, it follows from Theorem 2.6 that $(I - D^{-1}C)^{-1}$ exists and equals to $\sum_{k=0}^{\infty}(D^{-1}C)^k$. Since the terms in sum are nonnegative, we have $(I - D^{-1}C)^{-1}$ is nonnegative and $B^{-1} = (I - D^{-1}C)^{-1}D^{-1} \geq 0$.

(2) $\Longrightarrow$ (6): Trivial.

(6) $\Longrightarrow$ (5): Consider the decomposition $B = D - C$, with $d_i = b_{ii}$. Let $\{I = i \mid d_i \leq 0\}$. From $d_i\nu_i - \sum_{k\neq i} c_{ik}\nu_k \geq 0$ follows $c_{ik} = 0$ for $i \in I$, and $k \neq i$. Since $B\nu \geq 0, \neq 0 \Longrightarrow I \neq \{1, \cdots, n\}$. But B is irreducible $\Longrightarrow I = \emptyset$ and $d_i > 0$. Hence for $D\nu >, \neq C\nu$ also $\nu >, \neq D^{-1}C\nu$ and (16) show that $\rho(D^{-1}C) < 1$. ■

return

Proof of Lemma 11 There is a $x \neq 0$ with $Ax = \lambda x$ and $|\lambda| = \rho(A)$. Hence

$$\rho(A)|x_i| = |\sum_{k=1}^{n} a_{ik}x_k| \leq \sum_{k=1}^{n} |a_{ik}||x_k| \leq \sum_{k=1}^{n} c_{ik}|x_k|.$$

Thus,

$$\rho(A)|x| \leq C|x|.$$

If $|x| > 0$, then from (15) we have $\rho(A) \leq \rho(C)$. Otherwise, let $I = \{i \mid x_i \neq 0\}$ and C_I be the matrix, which consists of the ith row and ith column of C with $i \in I$. Then we have $\rho(A)|x_I| \leq C_I|x_I|$. Here $|x_I|$ consists of ith component of $|x|$ with $i \in I$. Then from $|x_I| > 0$ and (15) follows $\rho(A) \leq \rho(C_I)$. We now fill C_I with zero up to an $n \times n$ matrix $\tilde{C}_I$. Then $\tilde{C}_I \leq C$. Thus, $\rho(C_I) = \rho(\tilde{C}_I) \leq \rho(C)$ (by Theorem **??**(3)). ■

return

Proof of Theorem ?? Let $A = D - L - R$. From (17) and (18) D must be nonnsingular and then as in Remark 9.3 we can w.l.o.g. assume that $D = I$. Now let $B = I - |L| - |R|$. Then (17) can be written as $Be > 0$. From Theorem 10(2) and (1) follows that B is an M-matrix.
(18) can be written as $Be \geq 0$, $Be \neq 0$. Since A is irreducible, also B, from Theorem 10 (6) and (1) follows that B is an M-matrix.
Especially, from theorem 10(1), (4) and Theorem **??** follows that

$$\rho(|L| + |R|) < 1 \text{ and } \rho((I - |L|)^{-1}|R|) < 1.$$

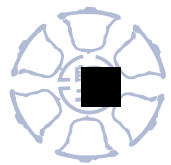
Now Lemma 11 shows that

$$\rho(L + R) \leq \rho(|L| + |R|) < 1.$$

So TSM is convergent. Similarly,

$$\begin{aligned} \rho((I - L)^{-1}R) &= \rho(R + LR + \cdots + L^{n-1}R) \\ &\leq \rho(|R| + |L||R| + \cdots + |L|^{n-1}|R|) \\ &= \rho((I - |L|)^{-1}|R|) < 1. \end{aligned}$$

So SSM is convergent.



return