ITERATIVE METHODS FOR THE SOLUTION OF LINEAR SYSTEMS

A stationary iterative method for solving $Ax=b$ is an iteration of the form

$$\begin{cases} x_{k+1} = Bx_k + c & B \in \mathbb{R}^{n\times n},\ c \in \mathbb{R}^n \\ x_0 \text{ given} \end{cases}$$

Defn. We say that an iterative method, say stationary, for $Ax=b$ is **convergent** if $\lim_{k\to\infty} x_k = x$ for **every** $x_0 \in \mathbb{R}^n$.

Remark It is in fact realistic to expect that the method will converge for all x_0.

Defn A matrix norm $\|\cdot\|$ is a map $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ that satisfies the following properties

(i) $\|A\| \geq 0$ and $\|A\| = 0 \iff A = 0$

(ii) $\|\alpha A\| = |\alpha|\,\|A\|$, $\forall \alpha \in \mathbb{R}$, $\forall A \in \mathbb{R}^{n\times n}$

(iii) $\|A+B\| \leq \|A\| + \|B\|$

(iv) $\|AB\| \leq \|A\|\,\|B\|$

Theorem 1 Let A be a square matrix and $\|\cdot\|$ a matrix norm (induced or otherwise). Then,

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

Furthermore, given a matrix A, given any $\varepsilon > 0$, there exists an induced matrix norm $\|\cdot\|_{\varepsilon,A}$ such that

$$\|A\|_{\varepsilon,A} \leq \rho(A) + \varepsilon .$$

proof.

Let $v \neq 0$ be an eigenvector of A corresponding to a maximal eigenvalue of A : $Av = \lambda v$, $|\lambda| = \rho(A)$. Let u be a vector such that $vu^T \neq 0$. Since

$$\rho(A)\,\|vu^T\| = \|\lambda v u^T\| = \|Av u^T\| \leq \|A\|\,\|vu^T\|,$$

(prop. (iv) of matrix norm)

we see that $\rho(A) \leq \|A\|$.

To prove the 2nd part, let U be an invertible matrix (also unitary) such that $U^{-1}AU$ is upper triangular

$$U^{-1}AU = \begin{pmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ & \lambda_2 & & t_{2n} \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix};$$

note that $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$. For every scalar $\delta > 0$, we let

$$D_\delta = \operatorname{diag}\{1, \delta, \delta^2, \ldots, \delta^{n-1}\}, \text{ so that}$$

$$(UD_\delta)^{-1} A (UD_\delta) = \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-1} t_{1n} \\ & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-2} t_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & & \delta t_{n-1,n} \\ 0 & & & & \lambda_n \end{pmatrix}.$$

Given $\epsilon > 0$, fix δ so that

$$\sum_{j=i+1}^{n} |\delta^{j-i} t_{ij}| \leq \epsilon, \qquad 1 \leq i \leq n-1.$$

Now define the map $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ by

$$\|B\| \equiv \|(UD_\delta)^{-1} B (UD_\delta)\|_\infty.$$

Now

$$\|A\| \le \max_i \left| \lambda_i + \sum_{j=i+1}^{n} \delta^{j-i} t_{ij} \right| \le \rho(A) + \epsilon .$$

Also, it is easy to see that $\|\cdot\|$ satisfies the properties (i) - (iv) of the definition of a matrix norm. To see that it is an induced norm, it suffices to note that $\|\cdot\|$ is the matrix norm induced by the vector norm (verify!)

$$\|v\| \equiv \|(U D_\delta)^{-1} v\|_\infty . \qquad \blacksquare$$

Theorem 2. Let B be a square matrix. The following are equivalent

(1) $\lim_{k\to\infty} B^k = 0$

(2) $\lim_{k\to\infty} B^k v = 0 \quad \forall v \in R^n$

(3) $\rho(B) < 1$

(4) $\|B\| < 1$ for at least one induced matrix norm $\|\cdot\|$.

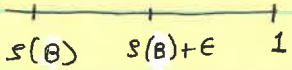
proof. (1) $\Rightarrow$ (2). Let $\|\cdot\|$ be a vector norm and $\|\cdot\|$ the corresp. induced matrix norm. Then $\|B^k v\| \le \|B\|^k \|v\| \to 0$.

(2) $\Rightarrow$ (3). If $\rho(B) \ge 1$, we can find a vector (eigenvector of A) $v \neq 0$ such that $Bv = \lambda v$, $|\lambda| \ge 1$. Now

$$B^k v = \lambda^k v \not\to 0 \quad \text{as} \quad k \to \infty$$

contradicting (2).

(3) => (4). Since $\rho(B) < 1$, $\exists\ \epsilon > 0$



such that $\rho(B) + \epsilon < 1$. Now by Theorem 1, there is an induced matrix norm $\|\cdot\|_{\epsilon,B}$ such that $\|B\|_{\epsilon,B} \leq \rho(B) + \epsilon < 1$.

(4) => (1) $\quad \|B^k\|_{\epsilon,B} \leq \|B\|^k_{\epsilon,B} \to 0$ as $k \to \infty$. ∎

Theorem 3 Let B be a square matrix and let $\|\cdot\|$ be any matrix norm. Then

$$\lim_{k\to\infty} \|B^k\|^{1/k} = \rho(B)$$

proof.

$$\rho(B) = \rho(B^k)^{1/k} \leq \|B^k\|^{1/k}$$

($\rho(B) = \rho(B^k)^{1/k}$ by $\sigma(B^k) = \{\lambda_1^k, \ldots, \lambda_n^k\}$; $\rho(B^k)^{1/k} \leq \|B^k\|^{1/k}$ by Thm 1)

we shall next establish that for every $\epsilon > 0$, $\exists$ integer $\ell = \ell(\epsilon)$ such that

$$k \geq \ell \Rightarrow \|B^k\|^{1/k} \leq \rho(B) + \epsilon$$

which will give the required result. Indeed, suppose that $\epsilon > 0$ given. Now matrix

$$B_\epsilon = \frac{B}{\rho(B)+\epsilon} \quad \text{satisfies } \rho(B_\epsilon) < 1, \text{ so by}$$

theorem 2, $\lim_{k\to\infty} B_\epsilon^k = 0$. Consequently, $\exists\ \ell(\epsilon)$ such that

$$k \geq \ell \Rightarrow \|B_\epsilon^k\| = \frac{\|B^k\|}{(\rho(B)+\epsilon)^k} \leq 1,$$

which is the required result. ∎

Theorem 4 Concerning the iterative method $x_{k+1} = Bx_k + c$, the following are equivalent

(1) The iterative method is convergent

(2) $\rho(B) < 1$

(3) $\|B\| < 1$ for some matrix norm $\|\cdot\|$.

proof. let $e_k = x_k - x$. Then, $e_k = B^k e_0$, $k \geq 0$.

Thus, the proof is a consequence of theorem 2.0 ∎

Theorem 5 (1) let $\|\cdot\|$ be any vector norm and let x be such that

$$x = Bx + c.$$

Consider the iterative method

$$x_{k+1} = Bx_k + c, \quad k \geq 0.$$

Then

$$\lim_{k\to\infty} \left\{ \sup_{\|x_0 - x\| = 1} \|x_k - x\|^{1/k} \right\} = \rho(B).$$

(2) let x be such that $x = Bx + c = \tilde{B}\tilde{x} + \tilde{c}$. Consider the iterative methods

$$\tilde{x}_{k+1} = \tilde{B}\tilde{x}_k + \tilde{c}, \quad k \geq 0 \quad \text{and} \quad x_{k+1} = Bx_k + c, \quad k \geq 0,$$

with

$$\rho(B) < \rho(\tilde{B}), \qquad x_0 = \tilde{x}_0.$$

Then, for any $\epsilon > 0$, there exists an integer $\ell = \ell(\epsilon)$ such that

$$k \geq \ell \Rightarrow \sup_{\|x_0 - x\| = 1} \left\{ \frac{\|\tilde{x}_k - x\|}{\|x_k - x\|} \right\}^{1/k} \geq \frac{\rho(\tilde{B})}{\rho(B) + \epsilon}.$$

Proof. let $\|\cdot\|$ denote also the induced matrix norm.

For every integer k, one can write

$$(\rho(B))^k = \rho(B^k) \le \|B^k\| = \sup_{\|e_0\|=1} \|B^k e_0\|$$

$$\Rightarrow \quad \rho(B) \le \sup_{\|e_0\|=1} \|B^k e_0\|^{1/k} = \underbrace{\|B^k\|^{1/k}}_{\downarrow \rho(B)},$$

and the statement follows from Thm 3. By the same result, given $\epsilon > 0$, there exists an integer $\ell = \ell(\epsilon)$ such that

$$k \ge \ell \Rightarrow \underbrace{\|B^k\|^{1/k}}_{= \sup_{\|e_0\|=1} \|B^k e_0\|^{1/k}} \le \rho(B) + \epsilon .$$

Moreover, for any integer $k \ge \ell$, there exists a vector $e_0 = e_0(k)$ (*) such that

$$\|e_0\| = 1 \quad \text{and} \quad \|\tilde{B}^k e_0\|^{1/k} = \|\tilde{B}^k\|^{1/k} \ge \rho(\tilde{B}) .$$

$$\sup_{\|e_0\|=1} \left\{ \frac{\|\tilde{x}_k - x\|^{1/k}}{\|x_k - x\|^{1/k}} \right\} \ge \frac{\|\tilde{x}_k - x\|^{1/k}}{\sup_{\|e_0\|=1} \|x_k - x\|^{1/k}} \ge \frac{\rho(\tilde{B})}{\rho(B) + \epsilon} .$$

(e_0 as in (*))

Remark statement (2) if $\rho(B) < \rho(\tilde{B})$ then the iterates $\{x_k\}$ converge "faster" than the iterates $\{\tilde{x}_k\}$.

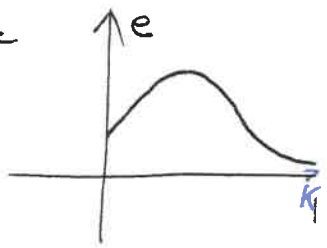
The above results imply that a necessary and sufficient condition for convergence is that $\rho(P)<1$. We now consider some aspects pertinent to the speed of convergence. First, we note that the speed of convergence depends on the vector norm chosen. Also, an interesting question is whether the errors decrease monotonically.

First, we consider the case of a normal iteration matrix and $\|\cdot\|_2$. We have $P = U^* \Lambda U$, where U is unitary and Λ is the diagonal matrix of eigenvalues of P. We have

$$\|P^k\|_2 = \|U^* \Lambda^k U\|_2 = \|\Lambda^k\|_2 = \rho(P)^k.$$

So in this particular case, the speed of convergence is given in terms of $\rho(P)$ and the errors can be seen to decrease monotonically.

If P is not normal and/or $\|\cdot\|$ is not the Euclidean norm, then the errors may not decrease monotonically. Indeed they may behave as in the figure



This phenomenon can be traced to a Jordan block of the form

$$\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}.$$

we have

$$\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}^k = \begin{bmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{bmatrix}.$$

The term $k\lambda^{k-1}$ has a tendency to grow as a function of k "at the beginning", especially for $\lambda \approx 1$. The behaviour of $k\lambda^{k-1}$ is as shown in the figure above.

The Jacobi, Gauss-Seidel and SOR Methods

Consider the regular splitting $A = M - N$.

We assume that

(i) M is invertible

(ii) systems of the form $Mx = c$ are "easy" to solve.

Then,

$$Ax = b \Rightarrow Mx = Nx + b.$$

This motivates the iterative method

$$\begin{cases} M x_{k+1} = N x_k + b, & k \geq 0 \\ x_0 \text{ given.} \end{cases}$$

$$\Rightarrow x_{k+1} = (M^{-1}N)x_k + (M^{-1}b)$$

$$= B x_k + c.$$

B is called the iteration matrix.

Hence for this, we shall consider the splitting

$$A = D - L - U$$

where D is diagonal part of A,

$-L$ is the strictly lower triangular part of A

$-U$ " " " upper " " "

Jacobi : $M = D$, $N = L + U \Rightarrow B = D^{-1}(L+U)$

$$\Rightarrow D x_{k+1} = (L+U)x_k + b, \quad k \geq 0.$$

This can be expressed in the componentwise form

$$a_{11}\left[x_1^{k+1}\right] = -a_{12}x_2^k - \cdots - a_{1n}x_n^k + b_1$$
$$a_{22}\left[x_2^{k+1}\right] = -a_{21}x_1^k - a_{23}x_3^k - \cdots - a_{2n}x_n^k + b_2$$
$$\vdots$$
$$a_{nn}\left[x_n^{k+1}\right] = -a_{n1}x_1^k - \cdots - a_{n,n-1}x_{n-1}^k + b_n .$$

Gauss-Seidel: In this method, $M = D - L$, $N = U$

$$\Rightarrow (D-L)x_{k+1} = Ux_k + b \Rightarrow B = (D-L)^{-1}U .$$

In componentwise form

$$a_{11}\left[x_1^{k+1}\right] = -a_{12}x_2^k - \cdots - a_{1n}x_n^k + b_1$$
$$a_{22}\left[x_2^{k+1}\right] = -a_{21}x_1^{k+1} - a_{23}x_3^k - \cdots - a_{2n}x_n^k + b_2$$
$$\vdots$$
$$a_{nn}\left[x_n^{k+1}\right] = -a_{n1}x_1^{k+1} - \cdots - a_{n,n-1}x_{n-1}^{k+1} + b_n .$$

Unlike in the Jacobi method, the recent updates $x_1^{k+1}, \ldots, x_{i-1}^{k+1}$ are used in the computation of x_i^{k+1}.

SOR (Successive over Relaxation).

Let $\omega \neq 0$ be a real parameter.

we let $M = \frac{D}{\omega} - L$ and $N = \frac{1-\omega}{\omega}D + U$

$$\Rightarrow \left(\frac{D}{\omega} - L\right)x_{k+1} = \left(\frac{1-\omega}{\omega}D + U\right)x_k + b \quad , k \geq 0.$$

$\mathcal{L}_\omega = \left(\frac{D}{\omega} - L\right)^{-1}\left(\frac{1-\omega}{\omega}D + U\right)$ is the iteration matrix.

$\omega = 1 \Rightarrow$ Gauss-Seidel

$\omega > 1 \Rightarrow$ "over-relaxation"

$\omega < 1 \Rightarrow$ "under-relaxation"

The above methods are called point methods. Similar methods can be devised based on block splittings:

$$A = \left[\begin{array}{c|c|c} A_{11} & A_{12} & A_{13} \\ \hline A_{21} & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \end{array}\right]$$

$$D = \left[\begin{array}{c|c|c} A_{11} & & 0 \\ \hline & A_{22} & \\ \hline 0 & & A_{33} \end{array}\right], \quad L = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline -A_{21} & 0 & 0 \\ \hline -A_{31} & -A_{32} & 0 \end{array}\right], \quad U = \left[\begin{array}{c|c|c} 0 & -A_{12} & -A_{13} \\ \hline 0 & 0 & -A_{23} \\ \hline 0 & 0 & 0 \end{array}\right]$$

Block Jacobi

$$A_{11}\left[X_1^{k+1}\right] = -A_{12} X_2^k - A_{13} X_3^k$$

$$A_{22}\left[X_2^{k+1}\right] = -A_{21} X_1^k - A_{23} X_3^k$$

$$A_{33}\left[X_3^{k+1}\right] = -A_{31} X_1^k - A_{32} X_2^k$$

Note that A_{ii} are square matrices and are assumed to be invertible.

Convergence of the Jacobi, Gauss-Seidel and relaxation methods

Given a linear system $Ax = b$, there is no guarantee that any one of the above three methods will be convergent. As a matter of fact, the method may not even be well defined. For example, the Jacobi and Gauss-Seidel methods are defined only if the diagonal elements of A are nonzero.

So the task at hand is to identify classes of matrices A for which a particular method is not only well-defined but also convergent. For the relaxation method, there exists the additional issue of identifying not only a range of values of the parameter ω, but also to identify specific values of ω for which the convergence will be rapid.

At this point the results of the previous section tell us that the iterative method

$$M x^{(k+1)} = N x^{(k)} + b, \quad k = 0, 1, \ldots$$

will be convergent iff $\rho(M^{-1}N) < 1$. Also, the smaller the value of $\rho(M^{-1}N)$ is, the faster the convergence rate will be in an asymptotic sense.

For (really) general matrices, there is no connection between the convergence of the Jacobi and Gauss-Seidel methods. Even though we suspect that the Gauss-Seidel method should have better convergence properties since it uses more "fresh" (updated) information than the Jacobi method.

Indeed the following two examples are instructive:

For

$$A = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}, \quad \text{we have} \quad \rho(J) < 1 < \rho(\mathcal{L}_1)$$

i.e. the Jacobi method converges whereas the Gauss-Seidel method does not. The next example in some sense "returns the favor". For

$A = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{bmatrix}$, we have $\rho(\mathcal{L}_1) < 1 < \rho(J)$.

Theorem 6 (Stein-Rosenberg) Let the Jacobi matrix $J = D^{-1}(L+U)$ be nonnegative, i.e. $J_{ij} \geq 0$. Let $\mathcal{L}_1 = (D-L)^{-1}U$ denote the Gauss-Seidel matrix. Then, one and only one of the following mutually exclusive relations hold

(i) $\rho(J) = \rho(\mathcal{L}_1) = 0$
(ii) $0 < \rho(\mathcal{L}_1) < \rho(J) < 1$
(iii) $1 = \rho(J) = \rho(\mathcal{L}_1)$
(iv) $1 < \rho(J) < \rho(\mathcal{L}_1)$

Theorem 7 Let A be a rowwise strictly diagonally dominant matrix. Then the Jacobi and Gauss-Seidel methods are convergent.

proof: We first consider the Jacobi method for which $M = D$ and $N = L+U$. We set $J = M^{-1}N = D^{-1}(L+U)$. We will show that $\|J\|_\infty < 1$ implying $\rho(J) < 1$. Indeed

$$\|J\|_\infty = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |J_{ij}| = \max_{1 \leq i \leq n} \frac{1}{|a_{ii}|} \left(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \right)$$

$$= \max_{1 \leq i \leq n} \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < 1$$

in view of strict rowwise diagonal dominance.

For the Gauss-Seidel method, we have $M = D - L$, $N = U$. So we set, adopting commonly used notation

$$\mathcal{L}_1 = M^{-1}N = (D-L)^{-1}U.$$

Let v be an eigenvector of $\mathcal{L}_1$ with corresponding eigenvalue λ. Assume that $\|v\|_\infty = 1 = |v_k|$ for some $k \in \{1, \dots, n\}$. We have

$$\lambda v = \mathcal{L}_1 v = (D-L)^{-1}Uv \Rightarrow \lambda Dv = \lambda Lv + Uv.$$

Of course if $\lambda = 0$ then we are fine since the goal is to show that $\rho(\mathcal{L}_1) < 1$. So assume that $|\lambda| \geq 1$. We have

$$Dv = Lv + \frac{1}{\lambda}Uv \Rightarrow a_{kk}v_k = \sum_{j=1}^{k-1} a_{kj}v_j + \frac{1}{\lambda}\sum_{j=k+1}^{n} a_{kj}v_j.$$

$$\Rightarrow |a_{kk}|\underbrace{|v_k|}_{=1} \leq \left(\sum_{j=1}^{k-1}|a_{kj}| + \frac{1}{|\lambda|}\sum_{j=k+1}^{n}|a_{kj}|\right)|v_k|.$$

This in turn implies that $\left(\frac{1}{|\lambda|} \leq 1\right)$

$$|a_{kk}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|$$

which contradicts the strict diagonal dominance of A. ∎

Remark The proof applied to the Gauss-Seidel case can be adapted to show that for $0 < \omega < 2$ the relaxation method is convergent for rowwise strictly diagonally dominant matrices.

Remark It is unfortunate that, (of the matrices in) many linear systems arising from finite difference or finite element discretizations of partial differential equations do not possess the strict diagonal dominance property. That this is so is in the very nature of the problem. Indeed, the approximation of partial derivatives is undertaken and therefore for almost all the rows of the matrix it will be the case that the elements add up to zero. (constant vectors are in the null space of the operator)

Fortunately, it turns out that by relaxing the strict diagonal dominance but adding other properties, one can arrive at conditions which are satisfied in many practical situations and also lead to satisfactory convergence results.

Defn. A square matrix A is said to be **reducible** if there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

where B_{11} and B_{22} are square matrices but not necessarily of the same size.

A matrix which is not reducible is called **irreducible**.

Defn. A matrix A is said to be irreducibly diagonally dominant (i.d.d.) if

(i) It is irreducible

(ii) It is diagonally dominant, i.e.

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad , \quad i = 1, \ldots, n$$

(iii) the inequality above is strict for at least one index i.

Ex. The symmetric, tridiagonal matrix

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}$$

is irreducibly diagonally dominant.

Ex. The $n^2 \times n^2$ block-tridiagonal matrix

$$A = \begin{bmatrix} T & -I & & & \\ -I & T & -I & & 0 \\ & \ddots & \ddots & \ddots & \\ & 0 & & & -I \\ & & & -I & T \end{bmatrix}$$

where I is the $n \times n$ identity matrix and T is the $n \times n$ tridiagonal matrix

$$T = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & 0 \\ & \ddots & \ddots & \ddots & \\ & 0 & & & -1 \\ & & & -1 & 4 \end{bmatrix}$$

Can also be shown to be irreducibly diagonally dominant.

Theorem 8 Let A be irreducibly diagonally dominant. Then the Jacobi and Gauss-Seidel methods are convergent. □

proof See J. Ortega, "Numerical Analysis, A second course"

Theorem 8 (Stein-Rosenberg) Let the Jacobi matrix $J = D^{-1}(L+U)$ be nonnegative, i.e. $J_{ij} \geq 0$. Let $\mathcal{L}_1 = (D-L)^{-1}U$ denote the associated Gauss-Seidel matrix. Then, one and only one of the following relations is valid

we shall see below that the Gauss-Seidel method is convergent for symmetric, positive definite matrices. On the other hand the example of

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

shows that this is not always true for the Jacobi method.

Theorem 9 Let A be a Hermitian, positive definite matrix and let $A = M - N$, M invertible. If the Hermitian matrix $M^* + N$ is positive definite, then

$$\rho(M^{-1}N) < 1 .$$

proof.

$M^* + N$ is indeed Hermitian since

$$(M^* + N)^* = M + N^* = M + (M - A)^* = M + M^* - A^*$$
$$= M^* + (M^* - A^*)^* = M^* + N^{**} = M^* + N .$$

we shall next establish the inequality $\|M^{-1}N\| < 1$ where $\|\cdot\|$ denotes the matrix norm induced by the vector norm

$$\|v\| = (v^* A v)^{1/2} .$$

This is indeed a vector norm since A is Hermitian, positive definite

Now

$$\|M^{-1}N\| = \|I - M^{-1}A\| = \sup_{\|v\|=1} \|v - M^{-1}Av\| .$$

for simplicity, let $w = M^{-1}Av$. Then

$$\|v - M^{-1}Av\|^2 = \|v - w\|^2 = (v - w)^* A (v - w)$$
$$= v^* A v - v^* A w - w^* A v + w^* A w$$
$$= 1 - w^* M^* A^{-*} A w - w^* A A^{-1} M w + w^* A w \qquad (v = A^{-1} M w)$$
$$= 1 - w^* (M^* + N) w .$$
$$< 1$$

Since $M^* + N$ is Hermitian and positive definite.

Now, the set $S = \{v \in \mathbb{C}^n : \|v\| = 1\}$ is compact and the function $v \longmapsto \|v - M^{-1}Av\|$ is continuous. Hence it attains its maximum at some point (vector) $\tilde{v} \in S$. By the above, $\|\tilde{v}\| = 1 \Rightarrow \sup_{\|v\|=1} \|v - M^{-1}Av\| = \|\tilde{v} - M^{-1}A\tilde{v}\| < 1$.
By Theorem 1, $\rho(M^{-1}N) \leq \|M^{-1}N\| < 1$. ∎

Theorem 10 (Ostrowski-Reich) If the matrix A is Hermitian and positive definite, then the point or block relaxation method converges if $0 < \omega < 2$.

proof.

$$A = M - N = \underbrace{\left\{\frac{D}{\omega} - L\right\}}_{M} - \underbrace{\left\{\frac{1-\omega}{\omega} D + U\right\}}_{N}$$

so

$$M^* + N = \frac{D^*}{\omega} - L^* + \frac{1-\omega}{\omega} D + U = \frac{2-\omega}{\omega} D$$

since $A = A^* \Rightarrow D = D^*$ and $L^* = U$.

Since A is positive definite, it easily follows that D is also positive definite. Also, $\frac{2-\omega}{\omega} D$ will remain positive definite as long as $\frac{2-\omega}{\omega} > 0 \Leftrightarrow 0 < \omega < 2$.
The result now follows from Theorem 6. ∎

The next result shows that the converse of Theorem 10 is true regardless of any hypotheses on A.

Theorem 11 (Kahane) The spectral radius of the point or block relaxation method satisfies the inequality

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1|, \quad \omega \neq 0.$$

Consequently, the method can converge only if $0 < \omega < 2$.

proof

$$\prod_{i=1}^{n} \lambda_i(\mathcal{L}_\omega) = \frac{\det\left(\frac{1-\omega}{\omega} D + U\right)}{\det\left(\frac{D}{\omega} - L\right)} = \frac{\left(\frac{1-\omega}{\omega}\right)^n \det(D)}{\left(\frac{1}{\omega}\right)^n \det(D)} = (1-\omega)^n,$$

given the particular structure of the matrices involved. Consequently,

$$\rho(\mathcal{L}_\omega) \geq \left|\prod_{i=1}^{n} \lambda_i(\mathcal{L}_\omega)\right|^{1/n} \geq |\omega - 1|$$

with equality holding if and only if all the eigenvalues have the same modulus $|1-\omega|$. ∎

Theorem 12 (Comparison of Jacobi and Gauss-Seidel methods)
Let A be a block tridiagonal matrix. Then, the spectral radii of the corresponding block Jacobi and block Gauss-Seidel matrices are related by

$$(\rho(T))^2 = \rho(\mathcal{L}_1),$$

so that the two methods converge or diverge simultaneously. If they converge, the Gauss-Seidel method converges "twice" faster.

proof

Let A be any block tridiagonal matrix of the form

$$A = \begin{bmatrix} B_1 & C_1 & & & \\ A_2 & B_2 & C_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & C_{N-1} \\ 0 & & & A_N & B_N \end{bmatrix}$$

and consider the one-parameter family of block tridiagonal matrices $A(\mu)$, $Q(\mu)$, $\mu \neq 0$.

$$A(\mu) = \begin{bmatrix} B_1 & \mu^{-1}C_1 & & & \\ \mu A_2 & B_2 & \mu^{-1}C_2 & 0 & \\ & \ddots & \ddots & \ddots & \\ & & & \mu^{-1}C_{N-1} & \\ 0 & & & \ddots & \\ & & & \mu A_N & B_N \end{bmatrix} ; \quad Q(\mu) = \begin{bmatrix} \mu I_1 & & & \\ & \mu^2 I_2 & & 0 \\ & & \ddots & \\ & 0 & & \\ & & & \mu^N I_N \end{bmatrix}.$$

(i) Clearly $A(1) = A$. Also,

It is easy to see that $A(\mu) = Q(\mu)\, A(1) \{Q(\mu)\}^{-1}$.

Thus,

$$\det(A(\mu)) = \det(A(1)) \quad \forall\, \mu \neq 0.$$

(ii) The eigenvalues of the Jacobi matrix $J = D^{-1}(L+U)$ are the roots of the characteristic polynomial $p_J = \det[D^{-1}(L+U) - \lambda I]$.

since $D^{-1}(L+U) - \lambda I = -D^{-1}[\lambda D - L - U]$, these are also the roots of the polynomial $q_J(\lambda) = \det[\lambda D - L - U] =$

Similarly, the eigenvalues of the Gauss-Seidel matrix $\mathcal{L}_1$ are roots of the characteristic polynomial $p_{\mathcal{L}_1} = \det[(D-L)^{-1}U - \lambda I]$.

Again, since $(D-L)^{-1}U - \lambda I = -(D-L)^{-1}[\lambda D - \lambda L - U]$, these are also the roots of the polynomial $q_{\mathcal{L}_1}(\lambda) = \det[\lambda D - \lambda L - U]$.

Now, observe that by (i), for $\lambda \neq 0$

$$q_{\mathcal{L}_1}(\lambda^2) = \det[\lambda^2 D - \lambda^2 L - U] = \det[\lambda^2 D - \lambda L - \lambda U]$$
$$= \lambda^N \det[\lambda D - L - U] = \lambda^N q_J(\lambda), \quad \lambda \neq 0.$$

By continuity of $q_{\mathcal{L}_1}$, this identity also holds for $\lambda = 0$. Thus

$$q_{\mathcal{L}_1}(\lambda^2) = \lambda^N q_J(\lambda) \qquad \forall\, \lambda.$$

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

$$\equiv \left[\det\left(L-\frac{D}{\omega}\right)\right]^{-1} \; q_{\mathcal{L}_\omega}(\lambda) .$$

The block tridiagonal structure of A allows us to write for $\boxed{\lambda \neq 0,}$

$$q_{\mathcal{L}_\omega}(\lambda^2) = \det\left[\frac{\lambda^2+\omega-1}{\omega} D - \lambda^2 L - U\right]$$

$$= \det\left[\frac{\lambda^2+\omega-1}{\omega} D - \lambda L - \lambda U\right] = \lambda^n \det\left[\frac{\lambda^2+\omega-1}{\lambda\omega} D - L - U\right]$$

$$= \lambda^n \, q_J\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right) = \lambda^n \, p_J\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right) \det(-D)$$

$\Rightarrow$

$$p_{\mathcal{L}_\omega}(\lambda^2) = \lambda^n \, p_J\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right) \underbrace{\left[\det(-D)\det\left(L-\frac{D}{\omega}\right)^{-1}\right]}_{\text{constant.}} .$$

In view of this, $(\beta = \lambda^2)$

$$0 \neq \beta \text{ , and } \beta \in \sigma(\mathcal{L}_\omega) \Rightarrow \left\{\frac{\beta\omega-1}{\sqrt{\beta}\,\omega}, -\frac{\beta+\omega-1}{\sqrt{\beta}\,\omega}\right\} \subseteq \sigma(J)$$

$$\alpha \in \sigma(J) \Leftrightarrow -\alpha \in \sigma(J) \Rightarrow \{\mu_+(\alpha,\omega), \mu_-(\alpha,\omega)\} \subseteq \sigma(\mathcal{L}_\omega)$$

where

$$\mu_+(\alpha,\omega) = \frac{1}{2}(\alpha^2\omega^2 - 2\omega + 2) + \frac{\alpha\omega}{2}\sqrt{\alpha^2\omega^2 - 4\omega + 4}$$

and

$$\mu_-(\alpha,\omega) = \frac{1}{2}(\alpha^2\omega^2 - 2\omega + 2) - \frac{\alpha\omega}{2}\sqrt{\alpha^2\omega^2 - 4\omega + 4}$$

are the squares of the two roots of the quadratic (in λ)

$$\lambda^2 - \lambda\alpha\omega + (\omega-1) = 0 \Leftrightarrow \boxed{\frac{\lambda^2+\omega-1}{\lambda\omega} = \alpha} \quad \text{for } \lambda \neq 0 .$$

Note that the above relations are not valid for $\lambda = 0$ so that the 2nd implication is not valid for $\omega = 1$. But this case

was considered in theorem 8.

(ii) we have $$\rho(\mathcal{C}_\omega) = \max_{\alpha \in \sigma(J)} \left\{ \max \left(|\mu_+(\alpha,\omega)|, |\mu_-(\alpha,\omega)| \right) \right\}.$$

Hence, on making the assumption that all the eigenvalues of J are real, we are led to a study of the function

$$M : (\alpha,\omega) \in \mathbb{R} \times (0,2) \longrightarrow M(\alpha,\omega) \equiv \max\left(|\mu_+(\alpha,\omega), \mu_-(\alpha,\omega)| \right).$$

We also note that $M(-\alpha,\omega) = M(\alpha,\omega)$. So it is enough to study M on $\mathbb{R}_+ \times (0,2)$.

(iii) Consider first the case $\boxed{0 \le \alpha < 1}$. For $\alpha = 0$,

$$M(0,\omega) = |\omega - 1|.$$

For $0 < \alpha < 1$, the quadratic $\omega \longrightarrow \alpha^2\omega^2 - 4\omega + 4$ has two real roots $\omega_0(\alpha)$ and $\omega_1(\alpha)$ satisfying

$$1 < \omega_0(\alpha) = \frac{2}{1+\sqrt{1-\alpha^2}} < 2 < \omega_1(\alpha).$$

If, then, $\omega_0(\alpha) < \omega < 2$, the complex numbers $\mu_+(\alpha,\omega)$ and $\mu_-(\alpha,\omega)$ are conjugate. As these are the squares of the roots of the quadratic $a \longrightarrow \lambda^2 - 2\alpha\omega + (\omega - 1)$, the product of the roots being $\omega - 1$, it follows that

$$1 < \omega_0(\alpha) < \omega < 2 \Rightarrow M(\alpha,\omega) = |\mu_+(\alpha,\omega)| = |\mu_-(\alpha,\omega)| = \omega - 1.$$

Now suppose that $0<\omega<\omega_0(\alpha)$. It is easy to see that

$$M(\alpha,\omega) = M_+(\alpha,\omega) \equiv v^2(\alpha,\omega), \text{ where } v(\alpha,\omega) = \frac{1}{2}\alpha\omega + \frac{1}{2}\sqrt{\alpha^2\omega^2 - 4\omega + 4}.$$

Note that

$$0<\alpha<1 \text{ and } 0<\omega<2 \Rightarrow v^2 = \frac{1}{2}(\alpha^2\omega^2 - 2\omega + 2) + \frac{\alpha\omega}{2}\sqrt{\alpha^2\omega^2 - 4\omega + 4}$$

$$< \frac{1}{2}(\omega^2 - 2\omega + 2) + \frac{\omega}{2}(2-\omega) = 1, \quad (*)$$

and

$$\lim_{\omega\to 0^+} v(\alpha,\omega) = 1 \quad (\dagger) \quad \text{and} \quad \lim_{\omega\to\omega_0(\alpha)^-} v(\alpha,\omega) = \frac{\alpha}{2}\omega_0(\alpha) = \frac{\alpha}{1+\sqrt{1-\alpha^2}}$$

$$= \sqrt{\omega_0(\alpha) - 1},$$

after some elementary calculations.

Also, note that

$$\frac{\partial M}{\partial \omega} = 2v\frac{\partial v}{\partial \omega} = 2v\left\{\frac{v\alpha - 1}{2v - \alpha\omega}\right\} \text{ for } 0<\omega<\omega_0(\alpha).$$

We want to show that $\frac{\partial M}{\partial \omega} < 0$ for $0<\omega<\omega_0(\alpha)$. Indeed,

$$\begin{cases} 2v(v\alpha - 1) < 0 & \text{for } 0<\omega<\omega_0(\alpha) \quad (v<1 \text{ by } (*) \text{ and } 0<\alpha<1) \\ \lim\limits_{\omega\to\omega_0(\alpha)^-}\{2v(v\alpha-1)\} < 0, \quad \lim\limits_{\omega\to 0^+}\{2v(v\alpha-1)\} < 0 \end{cases}$$

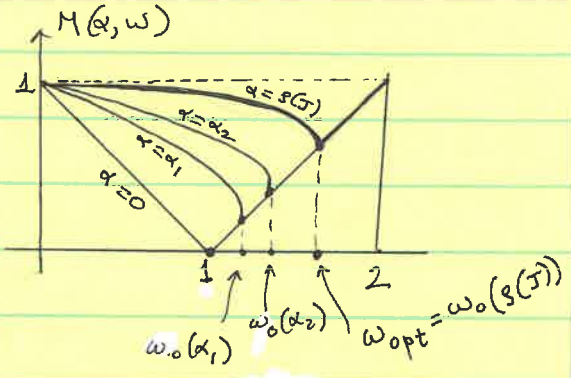
and

$$\begin{cases} 2v - \alpha\omega = \sqrt{\alpha^2\omega^2 - 4\omega + 4} > 0 & \text{for } 0<\omega<\omega_0(\alpha) \\ \lim\limits_{\omega\to\omega_0(\alpha)^-}\{2v - \alpha\omega\} > 0, \quad \lim\limits_{\omega\to 0^+}\{2v - \alpha\omega\} = 2 \quad (\text{from } (\dagger)) \end{cases}$$

This also shows that $\lim\limits_{\omega\to\omega_0(\alpha)^-} \frac{\partial M}{\partial \omega} = -\infty$.

Lastly, we will note that for fixed $\omega > 0$, the function $\alpha \longrightarrow M(\alpha, \omega)$ is strictly increasing for $\alpha > 0$.

With the information at hand, we see that the family of curves $M(\alpha, \omega)$, $0 \le \alpha < 1$ & $0 < \omega < 2$ has the behaviour shown in the figure.



Observe in particular that

$$\rho(T) < 1 \Rightarrow \rho(\mathcal{L}_\omega) = M(\rho(T), \omega) < 1 \quad \text{for} \quad 0 < \omega < 2,$$

and that

$$\rho(\mathcal{L}_{\omega_{opt}}) = \inf_{0<\omega<2} \rho(\mathcal{L}_\omega) = \omega_{opt} - 1,$$

with

$$\omega_{opt} = \omega(\rho(T)) = \frac{2}{1 + \sqrt{1 - (\rho(T))^2}}.$$

(iv.) Finally, we look at the case $\alpha \ge 1$. Here, the quadratic $\omega \longmapsto \alpha^2\omega^2 - 4\omega + 4$ is always nonnegative, and

$$M_+(\alpha, \omega) = \frac{1}{2}(\alpha^2\omega^2 - 2\omega + 2) + \frac{\alpha\omega}{2}(\alpha^2\omega^2 - 4\omega + 4)^{1/2}$$

$$\ge \frac{1}{2}(\omega^2 - 2\omega + 2) + \frac{\omega}{2}(2 - \omega) = 1 \quad \text{for} \quad 0 < \omega < 2,$$

which shows in particular that

$$\rho(T) \ge 1 \Rightarrow \rho(\mathcal{L}_\omega) \ge 1 \quad \text{for} \quad 0 < \omega < 2. \qquad \blacksquare$$

Theorem 14 Let A be a Hermitian, positive definite, block tridiagonal matrix. Then the block Jacobi and Gauss-Seidel methods and the relaxation method with $0<\omega<2$ converge, the function $\omega \in (0,2) \longmapsto \rho(\mathcal{L}_\omega)$ having the form indicated in the figure. There exists, in particular, a unique optimal relaxation parameter

$$\omega_0 = \frac{2}{1+\sqrt{1-(\rho(J))^2}},$$

such that for $\rho(J)>0$

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{0<\omega<2} \rho(\mathcal{L}_\omega) = \omega_0 - 1 < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J).$$

If $\rho(J)=0$, then $\omega_0 = 1$ and $\rho(\mathcal{L}_1) = \rho(J) = 0$.

proof. By Theorem 10, (Ostrowski-Reich) we know that the relaxation method (both point or block) with $0<\omega<2$ converges. In order to show that the block Jacobi method converges we need to show, in view of Theorem 13, that the eigenvalues of $J = D^{-1}(L+U)$ are real. Indeed,

$$D^{-1}(L+U)v = \alpha v \Rightarrow (L+U)v = \alpha D v$$

$$\Rightarrow Av = (1-\alpha)Dv \Rightarrow v^* A v = (1-\alpha) v^* D v.$$

Now A h.p.d. $\Rightarrow$ D is h.p.d. Thus v^*Av and v^*Dv are both positive if $v \neq 0$. Thus $\alpha \in \sigma(J) \Rightarrow \alpha \in \mathbb{R}$.

The conclusion now follows from theorem 13 and its proof. ∎