ITERATIVE PROJECTION METHODS FOR SPARSE LINEAR SYSTEMS AND EIGENPROBLEMS
CHAPTER 2 : SPLITTING METHODS

Heinrich Voss

voss@tu-harburg.de

Hamburg University of Technology
Institute of Numerical Simulation

TUHH
Technische Universität Hamburg-Harburg

Jacobi method

The simplest iterative scheme for linear systems of equations

$$Ax = b, \quad A = (a_{ij})_{i,j=1,\ldots,n} \in \mathbb{C}^{n\times n},\ b \in \mathbb{C}^n,$$

is the Jacobi method which is defined for system matrices which have nonzero diagonal elements.

The iteration step $x^k \to x^{k+1}$ can be motivated by solving the i-th equation for x_i and substituting the known $x_j^{(k)}$'s on the right hand side to obtain

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\right), \quad i = 1,\ldots,n.$$

Jacobi sweep

1: **for** $i = 1 : n$ **do**
2: $xnew(i) = b(i)$
3: **for** j=1:i-1 **do**
4: $xnew(i) = xnew(i) - a(i,j) * xold(j)$
5: **end for**
6: **for** j=i+1:n **do**
7: $xnew(i) = xnew(i) - a(i,j) * xold(j)$
8: **end for**
9: $xnew(i) := xnew(i)/a(i,i)$
10: **end for**
11: $xold = xnew$

Gauß–Seidel method

Notice that in the Jacobi method one does not use the newest information that is available.

When the i-th component $x_i^{(k+1)}$ is evaluated then the improved components $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ of x^{k+1} are already known and it is natural to take advantage of this information.

If we always use the newest information we obtain the Gauß-Seidel method

Gauß–Seidel swep

```
1: for i = 1 : n do
2:     x(i) = b(i)
3:     for j=1:i-1 do
4:         x(i) = x(i) − a(i, j) ∗ x(j)
5:     end for
6:     for j=i+1:n do
7:         x(i) = x(i) − a(i, j) ∗ x(j)
8:     end for
9:     x(i) := x(i)/a(i, i)
10: end for
```

Observe that in this case we do not have to use two vectors to store x^k and x^{k+1} but the memory of the old approximation can be overwritten with the new data currently.

Example

Consider

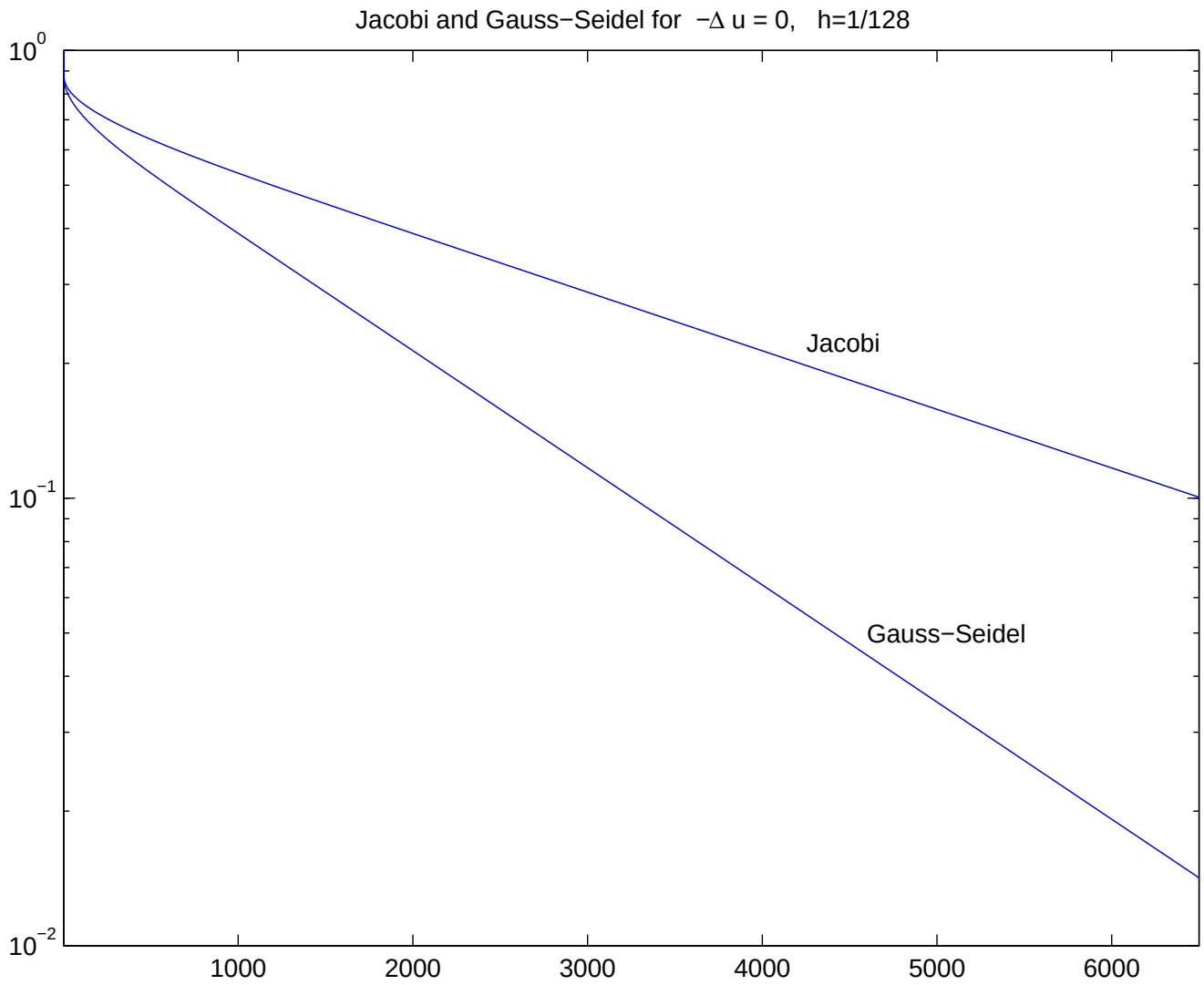
$$-\Delta u = 0 \text{ in } \Omega = (0,1) \times (0,1), \quad u = 0 \text{ on } \partial\Omega$$

Discretizing Δ with central differences with stepsize $h = 1/128$ yields a linear system

$$4U_{ij} - U_{i-1,j} - U_{i,j-1} - U_{i,j+1} - U_{i+1,j} = 0, \ i,j = 1,\ldots,127,$$

of dimension $n = 127^2 = 16129$.

Convergence history



Splitting methods

The Jacobi and the Gauß-Seidel methods are typical members of a large class of methods for the iterative solution of the linear system of equations

$$Ax = b, \quad A \in \mathbb{C}^{n\times n},\ b \in \mathbb{C}^n,$$

which are obtained by splitting of the system matrix $A = B + (A - B)$, where B is regular, and rewriting $Ax = b$ as

$$Bx = (B - A)x + b.$$

This system is solved iteratively by the so called splitting method

$$x^{k+1} := B^{-1}\Big((B - A)x^k + b\Big).$$

B is called the splitting matrix and $G := B^{-1}(B - A)$ is called the iteration matrix.

Examples

Jacobi method

$$B := \mathrm{diag}(A)$$

Gauß-Seidel method

$$B := \mathrm{diag}(A) - E$$

where $-E$ denotes the lower triangular part of A

Convergence

By the fixed point theorem for contracting mappings the sequence $\{x^k\}$ generated by

$$x^{k+1} := B^{-1}\Big((B-A)x^k + b\Big).$$

converges for every initial vector x^0 and every right hand side b to the unique solution x^* of $Ax = b$ if there exists a vector norm $\|\cdot\|$ such that for the corresponding matrix norm it holds

$$\|G\| = \|B^{-1}(B-A)\| =: \alpha < 1 \qquad (*).$$

In this case the error of the k-th iterate can be estimated as

$$\|x^k - x^*\| \le \frac{\alpha}{1-\alpha}\|x^k - x^{k-1}\| \le \frac{\alpha^k}{1-\alpha}\|x^1 - x^0\|.$$

Convergence ct.

Conversely it can be shown that from the convergence of the splitting method for every initial vector x^0 and for every b it follows that there exists a vector norm $\|\cdot\|$ such that for the corresponding matrix norm (*) holds.

The adequate measure of the size of the iteration matrix G is the spectral radius

$$\rho(G) := \max\{|\lambda| \,:\, \lambda \in \text{spec}(G)\}$$

where spec(G) denotes the spectrum of G, i.e. the set of all eigenvalues of G. This follows from the following theorem:

Theorem 2.1: *The splitting iteration*

$$x^{k+1} := B^{-1}\Big((B-A)x^k + b\Big)$$

converges for every initial vector $x^0 \in \mathbb{C}^n$ and for every vector $b \in \mathbb{C}^n$ to the unique solution of $Ax = b$ if and only if the iteration matrix $G := B^{-1}(B-A)$ satisfies the condition

$$\rho(G) < 1.$$

Proof for special case

Before we prove Theorem 2.1 for the general case we first consider the special case of a real symmetric iteration matrix G which is more transparent. In this case there exists an orthonormal set $v^1, \ldots, v^n \in \mathbb{C}^n$ of eigenvectors of G:

$$Gv^j = \lambda_j v^j, \quad (v^j)^H v^k = \delta_{jk}, \quad j,k = 1,\ldots,n,$$

where δ_{jk} denotes the Kronecker symbol.

We represent x^k and $\tilde{b} := B^{-1}b$ as linear combinations of the v^j's

$$x^k =: \sum_{j=1}^{n} \alpha_{kj} v^j, \quad \tilde{b} =: \sum_{j=1}^{n} \beta_j v^j.$$

From the splitting iteration we obtain

$$\sum_{j=1}^{n} \alpha_{k+1,j} v^j = x^{k+1} = Gx^k + \tilde{b} = \sum_{j=1}^{n} (\alpha_{kj}\lambda_j + \beta_j) v^j.$$

Proof for special case ct.

Hence, the coefficients α_{kj} satisfy the recurrence formula

$$\alpha_{k+1,j} = \lambda_j \alpha_{kj} + \beta_j, \quad j = 1, \ldots, n,$$

from which we easily get by induction

$$\alpha_{kj} = \begin{cases} \lambda_j^k \alpha_{0j} + \frac{1-\lambda_j^k}{1-\lambda_j} \beta_j & \text{for} \quad \lambda_j \neq 1 \\ \alpha_{0j} + k\beta_j & \text{for} \quad \lambda_j = 1 \end{cases}, \quad j = 1, \ldots, n.$$

Now it is obvious that the iteration converges for every initial vector x^0 and for every vector $\tilde{b}$ (i.e. for every α_{0j} and every β_j, $j = 1, \ldots, n$) if and only if all eigenvalues λ_j of G are smaller than 1 in modulus, i.e. $\rho(G) < 1$. □

Lemma

To prove Theorem 2.1 in its full generality we need the following lemma:

Lemma 2.2
Let $C \in \mathbb{C}^{n\times n}$. Then for every $\varepsilon > 0$ there exists a norm (which depends on C and on ε) such that

$$\|C\| \leq \rho(C) + \varepsilon.$$

Proof of Lemma 2.2

Because it is more transparent than the general case we first consider the case that C is diagonalizable.

Let $T \in \mathbb{C}^{n\times n}$ be a regular matrix such that

$$T^{-1}CT = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}.$$

Obviously, we then have

$$\|T^{-1}CT\|_\infty = \rho(C).$$

Consider the vector norm $\|x\|_{T^{-1}} := \|T^{-1}x\|_\infty$. Then we obtain for the corresponding matrix norm

$$\begin{aligned}\|C\|_{T^{-1}} &= \max_{x\neq 0} \frac{\|Cx\|_{T^{-1}}}{\|x\|_{T^{-1}}} = \max_{x\neq 0} \frac{\|T^{-1}CTT^{-1}x\|_\infty}{\|T^{-1}x\|_\infty} \\ &= \max_{y\neq 0} \frac{\|T^{-1}CTy\|_\infty}{\|y\|_\infty} = \|T^{-1}CT\|_\infty,\end{aligned}$$

which proves the lemma. In this case it even holds $\|C\|_T = \rho(C)$.

Proof of Lemma 2.2 ct.

In the general case we consider the transformation of C to Jordan's canonical form

$$T^{-1}CT = \operatorname{diag}\{J_1, J_2, \ldots, J_m\} =: J,$$

where

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & \ldots & 0 & 0 \\ 0 & \lambda_i & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_i & 1 \\ 0 & 0 & 0 & \ldots & 0 & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_i \times n_i}, \quad \sum_{i=1}^{m} n_i = n.$$

Proof of Lemma 2.2

Let

$$D_\varepsilon := \operatorname{diag}\{1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1}\}.$$

Then

$$D_\varepsilon^{-1} J D_\varepsilon = \operatorname{diag}\{J_1^\varepsilon, \ldots, J_m^\varepsilon\}$$

and

$$J_i^\varepsilon = \begin{pmatrix} \lambda_i & \varepsilon & 0 & \ldots & 0 & 0 \\ 0 & \lambda_i & \varepsilon & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_i & \varepsilon \\ 0 & 0 & 0 & \ldots & 0 & \lambda_i \end{pmatrix}.$$

Hence, we obtain

$$\|D_\varepsilon^{-1} T^{-1} C T D_\varepsilon\|_\infty \le \rho(C) + \varepsilon,$$

and the rest follows as in the case of a matrix which can be transformed to diagonal form. □

Proof of Theorem 2.1

Let $\rho(G) < 1$. Then by Lemma 2.2 there exists a norm $\|\cdot\|$ such that $\|G\| < 1$ and the convergence follows by the contracting mapping theorem.

Suppose that $\rho(G) \geq 1$. Let $\tilde{\lambda}$ be an eigenvalue of G such that $|\tilde{\lambda}| = \rho(G)$, and let y be a corresponding eigenvector. Then for the choice

$$\begin{cases} x^0 = y,\ b = 0 & \text{for} \quad \tilde{\lambda} \neq 1 \\ x^0 = y,\ \tilde{b} = y & \text{for} \quad \tilde{\lambda} = 1 \end{cases}$$

we obtain

$$\begin{cases} x^k = \tilde{\lambda}^k y & \text{for} \quad \tilde{\lambda} \neq 1 \\ x^k = (k+1)y & \text{for} \quad \tilde{\lambda} = 1 \end{cases}$$

and in both cases the sequence $\{x^k\}$ is not convergent. □

Convergence Rate

By Theorem 2.1 we have to choose the splitting matrix B such that the spectral radius $\rho(G)$ is less than 1 to obtain convergence of the iteration. Actually, we should choose B such that $\rho(G)$ is as small as possible because $\rho(G)$ is the final convergence rate of the iteration:

Theorem 2.3
Suppose that $\rho(G) < 1$ and denote by $e^k = x^k - x^*$ the error of the k-th iterate of the splitting method. Then for every vector norm

$$\ell := \sup_{x^0 \neq x^*} \limsup_{k \to \infty} \sqrt[k]{\frac{\|e^k\|}{\|e^0\|}} = \rho(G).$$

Proof of Theorem 2.3

Let $\tilde{\lambda}$ be an eigenvalue of G such that $|\tilde{\lambda}| = \rho(G)$ and let e^0 be an eigenvector of G which corresponds to $\tilde{\lambda}$. Then obviously,

$$\frac{\|e^k\|}{\|e^0\|} = \rho(G)^k,$$

and the inequality

$$\ell \geq \rho(G)$$

holds.

To prove the converse inequality we choose $\delta > 0$ arbitrarily. Then by Lemma 2.2 there exists a vector norm $\|\cdot\|_\delta$ such that

$$\|G\|_\delta \leq \rho(G) + \delta.$$

Proof of Theorem 2.3 ct.

The equivalence of all norms in $\mathbb{C}^n$ yields the existence of positive numbers $M \geq m > 0$ such that

$$m\|x\| \leq \|x\|_\delta \leq M\|x\| \quad \text{for every } x \in \mathbb{C}^n.$$

Hence, for every initial error e^0 we obtain

$$\|e^k\| \leq \frac{1}{m}\|e^k\|_\delta = \frac{1}{m}\|G^k e^0\|_\delta \leq \frac{1}{m}(\rho(G)+\delta)^k\|e^0\|_\delta \leq \frac{M}{m}(\rho(G)+\delta)^k\|e^0\|,$$

i.e.

$$\sqrt[k]{\frac{\|e^k\|}{\|e^0\|}} \leq (\rho(G)+\delta)\sqrt[k]{\frac{M}{m}}.$$

For $k \to \infty$ we get $\ell \leq \rho(G)+\delta$, and because $\delta > 0$ can be chosen arbitrarily the inequality $\ell \leq \rho(G)$ holds. This completes the proof. □

Convergence rate of Jacobi method

For the Jacobi method for the model problem $-\Delta u = f$ in the unit square with Dirichlet's boundary conditions the spectral radius of the iteration matrix G_J is

$$\rho(G_J) = \cos\frac{\pi}{m+1} \quad (\approx 0.9997 \text{ for } m = 128).$$

For the discrete model problem with stepsize $h = 1/(m+1)$ the spectral radius of the iteration matrix of the Jacobi method satisfies $1 - \rho(G_J) = O(h^2)$. Hence, the convergence is very slow for fine discretizations.

The same asymptotic holds true for the Gauß-Seidel method.

Remark

The contracting mapping theorem guarantees that the error is reduced in each step by a factor close to $\rho(G)$. This is true with respect to the norm constructed in Lemma 2.2.

It is usually the 2-norm or the ∞-norm or some closely related norm of the error that is of interest.

For matrices with a complete set of orthonormal eigenvectors (i.e. for normal matrices) the 2-norm and the spectral radius coincide. Thus, if $I - B^{-1}A$ is a normal matrix, then the 2-norm of the error is reduced at least by the factor $\rho(I - B^{-1}A)$ at each step.

For nonnormal matrices it is often the case that

$$\rho(I - B^{-1}A) < 1 < \|I - B^{-1}A\|_2.$$

In this case the 2-norm of the error may grow over some finite number of steps before it is reduced and convergence occurs.

Example

Consider the homogeneous problem $Ax = 0$ where

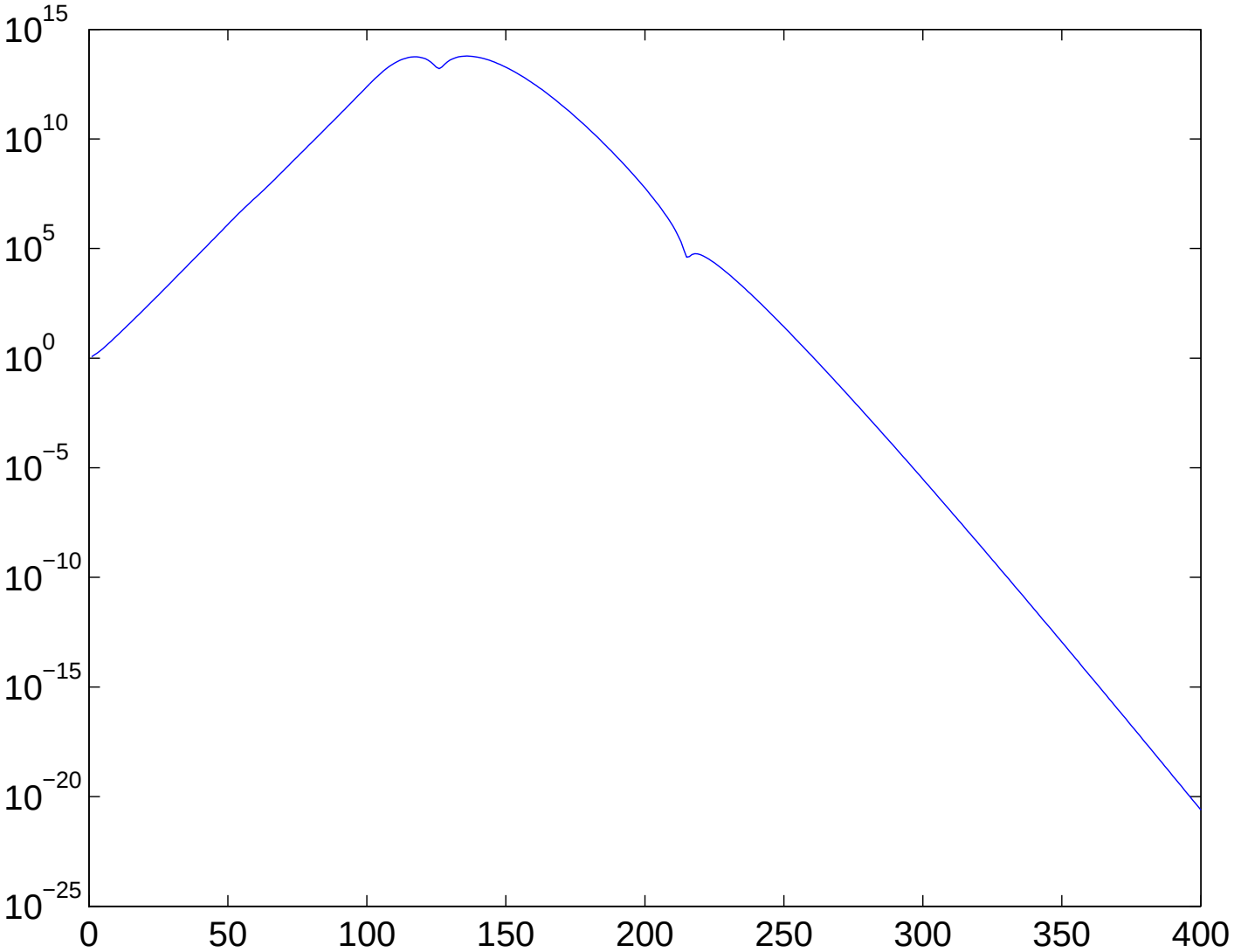
$$A = \begin{pmatrix} 1 & -1.15 & & & \\ 0.15 & 1 & -1.15 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1.15 \\ & & & 0.15 & 1 \end{pmatrix} \in \mathbb{R}^{100\times 100}$$

Choose B as the lower triangle of A (Gauß-Seidel method), and let x^0 be a random vector.

The following figure demonstrates that the error increases by about 15 orders of magnitude before starting to decrease.

The spectral norm of $B^{-1}(B - A)$ is 1.352 while its spectral radius is 0.6892.

Example ct.


10^{15}
10^{10}
10^{5}
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
10^{-25}
0
50
100
150
200
250
300
350
400

Remark

Without loss of generality we considered a homogeneous problem $Ax = 0$

The iteration

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

for the inhomogeneous problem $Ax = b$ and

$$y^{k+1} = B^{-1}(B - A)y^k,\ y^0 := x^0 - A^{-1}b$$

for the homogeneous problem $Ay = 0$ exhibit the same convergence behavior.

By induction its follows $y^k = x^k - A^{-1}b$: For $k = 0$ this is trivial, and if it holds for some k then it follows

$$\begin{aligned} y^{k+1} &= B^{-1}(B - A)y^k = B^{-1}(B - A)(x^k - A^{-1}b) \\ &= B^{-1}(B - A)x^k + B^{-1}b - A^{-1}b = y^k - A^{-1}b. \end{aligned}$$

Accelerating the Gauß–Seidel method

The convergence properties of the Gauß-Seidel method can be improved substantially if extrapolation is used.

The Gauß-Seidel method yields the i-th component of the new iterate

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big),$$

which is modified as

$$x_i^{(k+1)} := x_i^{(k)} + \omega(\tilde{x}_i^{(k+1)} - x_i^{(k)}),$$

where $\omega > 0$ is a fixed relaxation factor.

Inserting $\tilde{x}_i^{(k+1)}$ into the last equation one gets

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)}\Big).$$

Successive Overrelaxation (SOR)

If we use the decomposition $A = D - E - F$ where D denotes the diagonal of A, and $-E$ and $-F$ are the lower and upper triangle of A, respectively, then the iteration reads in matrix notation

$$(D - \omega E)x^{k+1} = (1 - \omega)Dx^k + \omega Fx^k + \omega b,$$

i.e.

$$x^{k+1} := (D - \omega E)^{-1}((1 - \omega)D + \omega F)x^k + \omega(D - \omega E)^{-1}b. \qquad (*)$$

The iterative method which is defined by (*) is called successive overrelaxation method or for short SOR method. Its iteration matrix is

$$G_\omega := (D - \omega E)^{-1}((1 - \omega)D + \omega F) = (I - \omega L)^{-1}((1 - \omega)I + \omega R),$$

where $L := D^{-1}E$ and $R := D^{-1}F$.

Convergence of SOR

Theorem 2.4:
Let $A \in \mathbb{C}^{n\times n}$ be an arbitrary matrix with regular diagonal D. Then for every $\omega \in \mathbb{R}$

$$\rho(G_\omega) \geq |\omega - 1|.$$

Hence, we can expect convergence of the SOR method only for relaxation factors $\omega \in (0, 2)$.

Proof: Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the iteration matrix G_ω.
Since $(D - \omega E)^{-1}$ and $(1 - \omega)D + \omega F$ are triangular matrices

$$\begin{aligned} \prod_{j=1}^n \lambda_j &= \det G_\omega = \det(D - \omega E)^{-1} \cdot \det((1 - \omega)D + \omega F) \\ &= \prod_{j=1}^n \frac{1}{a_{jj}} \cdot (1 - \omega)^n \prod_{j=1}^n a_{jj} = (1 - \omega)^n, \end{aligned}$$

from which we obtain the assertion. □

Convergence of SOR ct.

Theorem 2.5
Let $A := D - E - E^T$ be a Hermitean and positive definite matrix. Then the SOR method converges if and only if $\omega \in (0, 2)$.

Proof: By Theorem 2.4 we only have to prove that the SOR method converges for every $\omega \in (0, 2)$.
Let λ be an eigenvalue of G_ω and let $x \neq 0$ be a corresponding eigenvector, i.e.

$$((1 - \omega)D + \omega E^H)x = \lambda(D - \omega E)x. \qquad (*)$$

Convergence of SOR

Obviously, the following two equations hold:

$$\begin{aligned} 2((1-\omega)D+\omega E^H) &= (2-\omega)D+\omega(-D+2E^H) \\ &= (2-\omega)D-\omega A+\omega(E^H-E)), \end{aligned}$$

$$\begin{aligned} 2\lambda(D-\omega E) &= \lambda((2-\omega)D+\omega(D-2E)) \\ &= \lambda((2-\omega)D+\omega A+\omega(E^H-E)). \end{aligned}$$

Hence, we obtain from equation (*) ($((1-\omega)D+\omega E^H)x = \lambda(D-\omega E)x$)

$$\begin{aligned} &\lambda\Big((2-\omega)x^HDx+\omega x^HAx+\omega x^H(E^H-E)x\Big) \\ &= (2-\omega)x^HDx-\omega x^HAx+\omega x^H(E^H-E)x, \end{aligned}$$

Convergence of SOR

or using the abbreviations $d := x^H Dx > 0$, $a := x^H Ax$ and $x^H(E^H - E)x = is$, $s \in \mathbb{R}$:

$$\lambda((2-\omega)d + \omega a + i\omega s) = (2-\omega)d - \omega a + i\omega s.$$

Division by ω yields

$$\lambda(\mu d + a + is) = \mu d - a + is,$$

where $\mu := (2-\omega)/\omega$.

If $\omega \in (0,2)$ then $\mu > 0$ and $\mu d + is$ is contained in the right half of the complex plane. Hence, the distance to a is smaller than the one to $-a$, and therefore, we get

$$|\lambda| = \frac{|\mu d + is - a|}{|\mu d + is + a|} < 1. \qquad \square$$

Consistently ordered matrices

For the following class of matrices the optimal relaxation parameter can be determined.

Definition:
Let $A = D(I - L - R)$ the standard decomposition of $A \in \mathbb{C}^{n\times n}$.
A is 2-consistently ordered if the eigenvalues of the matrix

$$J(\alpha) := \alpha L + \alpha^{-1} R$$

are independent of $\alpha \neq 0$.

Consistently ordered matrices ct.

Example

Let

$$A = I - \begin{pmatrix} O & \tilde{R} \\ \tilde{L} & O \end{pmatrix}.$$

Then it holds that

$$\begin{aligned} \alpha L + \alpha^{-1} R &= \begin{pmatrix} O & \alpha^{-1}\tilde{R} \\ \alpha\tilde{L} & O \end{pmatrix} \\ &= \begin{pmatrix} I & O \\ O & \alpha I \end{pmatrix} \begin{pmatrix} O & \tilde{R} \\ \tilde{L} & O \end{pmatrix} \begin{pmatrix} I & O \\ O & \alpha^{-1} I \end{pmatrix}. \end{aligned}$$

Hence, $\alpha L + \alpha^{-1} R$ and $L + R$ are similar matrices, and A is 2-consistently ordered.

Consistently ordered matrices ct.

Example

Consider the finite difference approximation of the model problem where the unknowns are ordered in the following way.

The grid points are subdivided into two classes, red and black (like on a checkerboard), such that for every row and for every column in Ω neighboring points belong to different classes, and then they are ordered within each class.

Then the system matrix obtains the form considered in the last example, and hence, it is 2-consistently ordered.

Consistently ordered matrices ct.

If A is 2-consistently ordered then for $\alpha = 1$ and $\alpha = -1$ we obtain that the matrices $L + R$ and $-L - R$ have identical eigenvalues. Hence, the eigenvalues μ_j appear in pairs which differ by their sign, i.e.

$$\det(\lambda I - L - R) = \lambda^m \prod_{j=1}^{r} (\lambda^2 - \mu_j^2), \quad n = 2r + m. \qquad (1)$$

The following theorem relates the eigenvalues of the iteration matrices of the Jacobi method and of the SOR method for 2-consistently ordered matrices.

Theorem of Young

Let the matrix A be 2-consistently ordered where $a_{jj} = 1$ for every j, and assume that $\omega \neq 0$. Then the following assertions hold:

(i) *If $\lambda \neq 0$ is an eigenvalue of the iteration matrix G_ω of the SOR method and if μ satisfies the equation*

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

then μ is an eigenvalue of $L + R$.

(ii) *If μ is an eigenvalue of $L + R$ and if λ satisfies the equation (2), then λ is an eigenvalue of G_ω.*

Proof of Young's Theorem

We first prove the identity

$$\det(\lambda I - sL - rR) = \det(\lambda I - \sqrt{rs}(L + R)). \qquad (3)$$

Both sides of (3) contain polynomials of degree n which are of the form $\lambda^n + \ldots$.

From

$$sL + rR = \sqrt{rs}\Big(\sqrt{\frac{s}{r}}L + \sqrt{\frac{r}{s}}R\Big), \quad rs \neq 0,$$

it follows that $sL + rR$ and $\sqrt{rs}(L + R)$ have the same eigenvalues, and this is true for $rs = 0$ as well.

Therefore, both polynomials have the same roots, and hence, they are identical.

Proof of Young's Theorem ct.

Because $\det(I - \omega L) \neq 0$ the eigenvalues of G_ω are the roots of

$$\begin{aligned}\det(I - \omega L) \cdot \det(\lambda I - G_\omega) &= \det(\lambda(I - \omega L) - (1 - \omega)I - \omega R) \\ &= \det((\lambda + \omega - 1)I - \omega\lambda L - \omega R) =: \phi(\lambda).\end{aligned}$$

From (3) it follows that

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

and from (1) one obtains

$$\phi(\lambda) = (\lambda + \omega - 1)^m \prod_{j=1}^{r} \left((\lambda + \omega - 1)^2 - \omega^2 \lambda \mu_j^2\right), \qquad (4)$$

where μ_j denote the eigenvalues of $L + R$.

Proof of Young's Theorem ct.

If μ is an eigenvalue of $L+R$ and if λ satisfies equation (2), then $\phi(\lambda)=0$, and λ is an eigenvalue of G_ω. This proves the statement (*ii*).

If conversely $\lambda \neq 0$ is an eigenvalue of G_ω, then one of the factors in (4) vanishes. We assume that μ satisfies (2).

If $\mu \neq 0$ then $\lambda+\omega-1 \neq 0$. From (4) one obtains $(\lambda+\omega-1)^2=\lambda\omega^2\mu_j^2$ for some eigenvalue μ_j of $L+R$, and (2) yields $(\lambda+\omega-1)^2=\lambda\omega^2\mu^2$.

Thus, $\mu=\pm\mu_j$, and because the eigenvalues of $R+L$ appear in pairs of opposite sign, μ is an eigenvalue of $L+R$.

If $\mu=0$, then it follows from (2) that $\lambda+\omega-1=0$, and from (4)

$$0=\phi(\lambda)=\det(-\omega\sqrt{\lambda}(L+R)).$$

Hence, $L+R$ is singular, and $\mu=0$ is an eigenvalue of $L+R$. This completes the proof. □

Corolary

As a direct consequence of Young's Theorem we obtain for the case $\omega = 1$:

Corollary
If A is 2-consistently ordered then

$$\rho(G_{GS}) = \rho(G_J)^2,$$

where G_{GS} and G_J are the iteration matrices of the Gauss-Seidel method and the Jacobi method, respectively.

Hence, the Gauss-Seidel method needs only half as much iteration steps as the Jacobi method to yield the same accuracy.

Optimum parameter

Theorem 2.6: *Let $A = I - L - R$ be 2-consistently ordered, and suppose that the matrix $L + R$ has only real eigenvalues and that $\hat{\rho} := \rho(L + R) < 1$. Then for $\omega \in (0, 2)$ the spectral radius of the iteration matrix G_ω attains its minimum value for*

$$\omega^* = \frac{2}{1 + \sqrt{1 - \hat{\rho}^2}},$$

and the minimum spectral radius is given by

$$\rho^* = \omega^* - 1 = \left(\frac{\hat{\rho}}{1 + \sqrt{1 - \hat{\rho}^2}} \right)^2.$$

Remark: The spectral radius of G_ω is given by

$$\rho(G_\omega) = \begin{cases} \omega - 1 & \text{for} \quad \omega \geq \omega^* \\ 1 - \omega + \frac{1}{2}\omega^2\hat{\rho}^2 + \frac{1}{2}\omega\hat{\rho}\sqrt{\omega^2\hat{\rho}^2 + 4(1-\omega)} & \text{for} \quad \omega \leq \omega^* \end{cases}$$

Proof

From Young's theorem we know that all non-vanishing eigenvalues λ of G_ω satisfy

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2, \qquad (5)$$

where the eigenvalues μ appear in pairs of opposite sign and are real by our assumptions. Hence, we have to study (5) only for non-negative values of μ.

For fixed $\mu \geq 0$ we first determine ω such that the maximum of the moduli of the two roots of (5) is minimum.

If equation (5) has no real root then the roots are complex conjugate, and their absolute value is $|\lambda| = |\omega - 1|$.

Proof ct.

If equation (5) has a double real root, then we obtain from

$$\lambda^2 + \lambda(2\omega - 2 - \omega^2\mu^2) + (\omega - 1)^2 = 0$$

the condition

$$\frac{1}{4}(2\omega - 2 - \omega^2\mu^2)^2 = (\omega - 1)^2,$$

i.e.

$$\mu^2\omega^2\Big(4(\omega - 1) - \mu^2\omega^2\Big) = 0.$$

Hence, one obtains $\mu = 0$ or

$$\omega = \tilde{\omega} = \frac{2}{1 \pm \sqrt{1 - \mu^2}}, \qquad (6)$$

and the corresponding root is as in the case of non-real roots

$$\lambda = |\tilde{\omega} - 1| = \tilde{\omega} - 1.$$

Proof ct.

Finally, if equation (5) has two real roots, then the straight line $g_\omega(\lambda) := \frac{1}{\omega}(\lambda + \omega - 1)$ and the parabola $p(\lambda) = \pm\sqrt{\lambda}\mu$ have exactly two points of intersection.

The bigger of these points is minimum, if they coalesce, i.e. if the second case appears.

Moreover, in (6) the $+$-sign has to be chosen, because in this case the root is situated in the interval $(0, 1)$.

Proof ct.

If $\omega > 0$ is small, then (5) has two real solutions.

For increasing ω the maximum root of (5) decreases until $\omega = \tilde{\omega}$ and (5) has a double root $\lambda = \tilde{\omega} - 1$.

If ω is increased further, the roots become complex and $|\lambda| = |\omega - 1|$ starts growing. Hence, for fixed μ the parameter

$$\tilde{\omega} = \frac{2}{1 + \sqrt{1 - \mu^2}}$$

is optimum.

Proof ct.

We now demonstrate that for $\omega = \omega^*$ the spectral radius of G_ω satisfies $\rho^* = \omega^* - 1$.

$\mu := \hat{\rho}$ is an eigenvalue of $L + R$. Hence, $\lambda = \omega^* - 1 > 0$ is an eigenvalue of G_{ω^*}.

If $\mu \in [0, \hat{\rho})$ is an eigenvalue of $L + R$ then the parabola $P : \lambda \mapsto \pm\sqrt{\lambda}\mu$ is contained in the parabola corresponding to $\hat{\rho}$.

Hence, the straight line $\lambda \mapsto g_{\omega^*}(\lambda)$ does not meet the parabola P, and the eigenvalues of G_{ω^*} corresponding to μ are non-real with absolute value $\omega^* - 1$.

Proof ct.

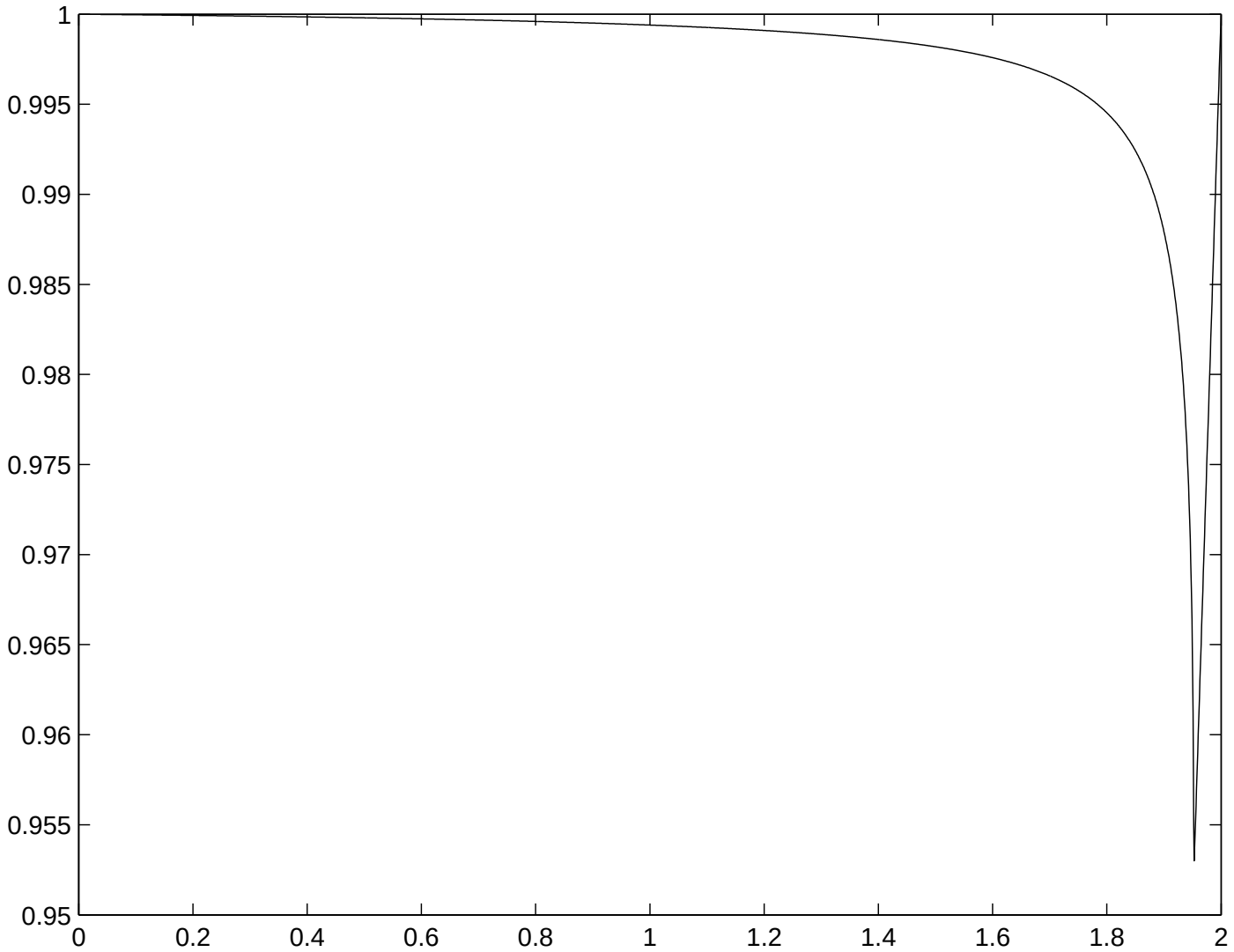
Therefore, all eigenvalues of G_{ω^*} have the same absolute value ρ^*.

It remains to show that $\rho(G_\omega) > \rho^*$ for $\omega \neq \omega^*$.

If λ is the root of (5) corresponding to $\mu = \hat{\rho}$ of maximum modulus, then $|\lambda| \geq |\omega - 1| > \omega^* - 1$, and one obtains $\rho(G_\omega) > \rho^*$. $\square$

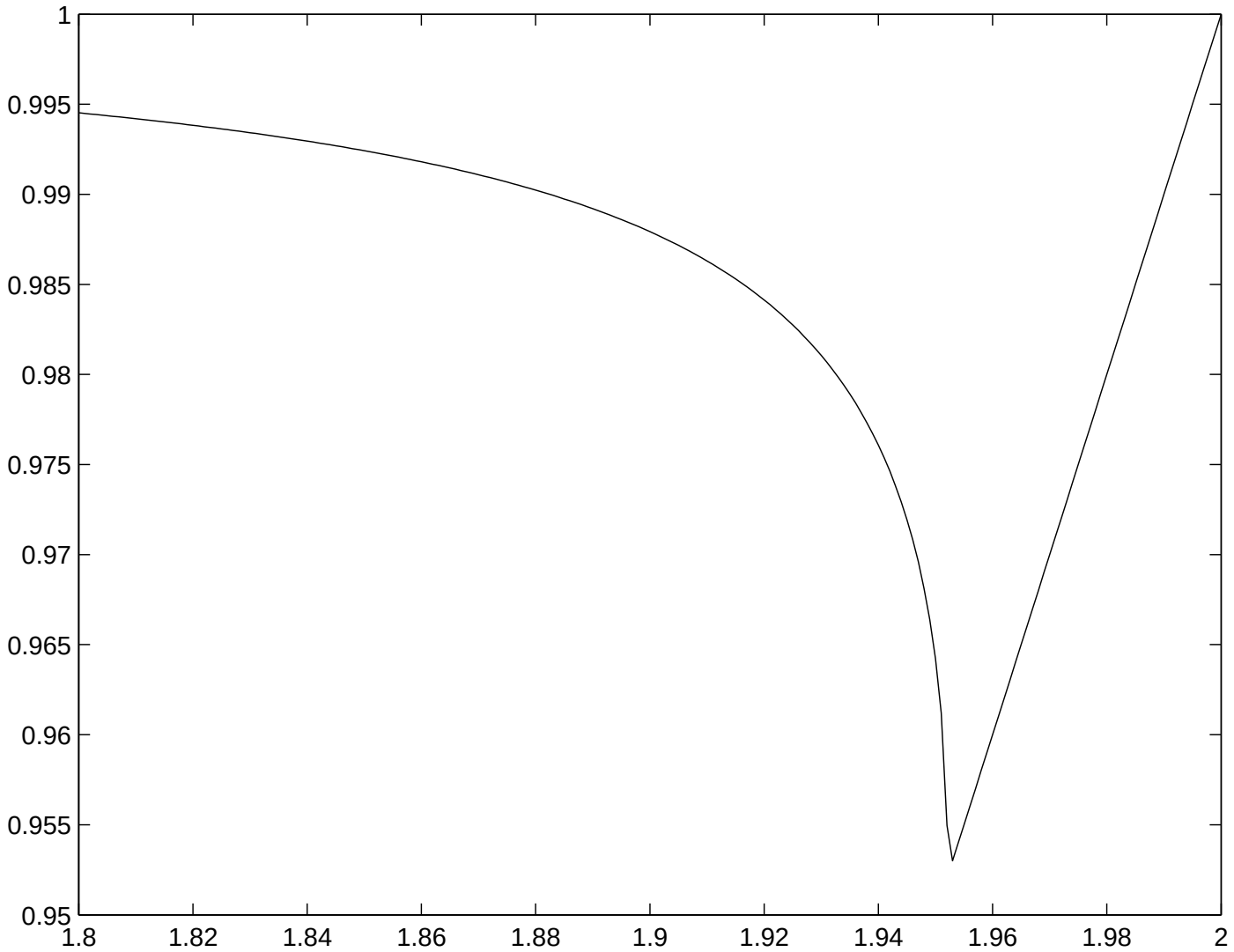
Consistently ordered matrices

Graph of $\omega \mapsto \rho(G_\omega)$



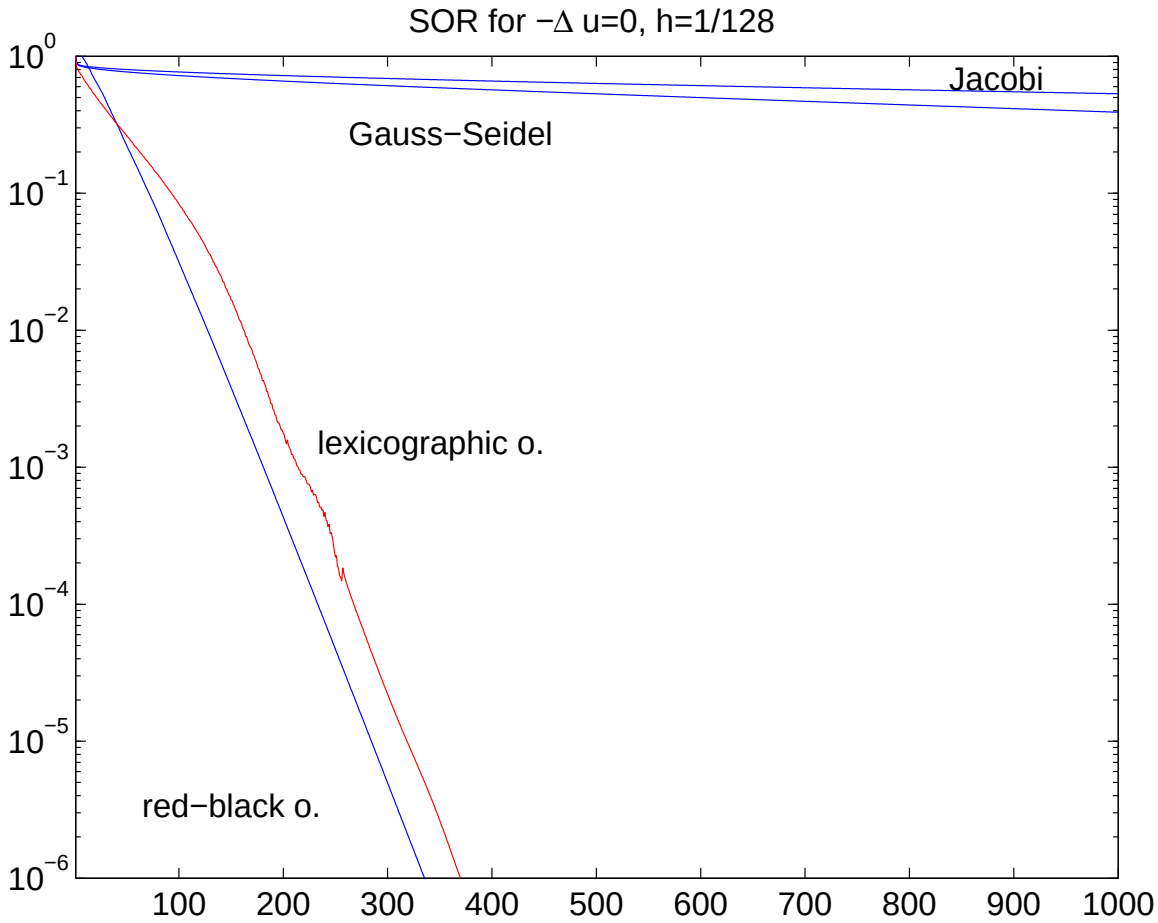
Consistently ordered matrices

Close-up of graph of $\omega \mapsto \rho(G_\omega)$



Consistently ordered matrices

Convergence history for model problem



For the model problem with red-black ordering and stepsize $h = 1/128$ one gets $\omega^* = 1.952$ and $\rho^* = 0.951$.

Symmetric splitting methods

In the next chapter when analyzing the acceleration by polynomial methods we will assume that the iteration matrix G has real eigenvalues.

For a symmetric system matrix A this can be accomplished by the requirement that the iteration matrix G has the form $G = B^{-1}(B - A)$, where B is symmetric and positive definite, because in this case G is similar to the symmetric matrix $B^{1/2}GB^{-1/2} = B^{-1/2}(B - A)B^{-1/2}$.

Moreover, in Chapter 4 we will need iteration matrices of this type with good convergence properties for the preconditioning of the conjugate gradient method.

Symmetric splitting methods ct.

Definition

The iteration method $x^{k+1} := Gx^k + \tilde{b}$ is called symmetric, if the iteration matrix can be written as

$$G = B^{-1}(B - A),$$

where the splitting matrix B is symmetric and positive definite whenever A is symmetric and positive definite.

Notice that in general the iteration matrix of a symmetric method is not symmetric.

Symmetric splitting methods ct.

Obviously, the Gauß-Seidel method and the successive overrelaxation are not symmetric (the eigenvalues of the SOR iteration matrix are not even real). It is possible, however, to symmetrize the Gauß-Seidel and the SOR method by coupling it with the corresponding backward scheme.

This backward iteration is obtained by updating the unknowns in reverse order. Using the partition $A := D - E - F = D(I - L - R)$ the backward Gauß-Seidel method is given by the iteration matrix

$$G_{GS}^b := (D - F)^{-1}E = (I - R)^{-1}L$$

and the backward SOR method by

$$G_\omega^b := (D - \omega F)^{-1}((1 - \omega)D + \omega E).$$

Backward Gauß-Seidel method

Given x^k the determination of x^{k+1} by the backward Gauß–Seidel methods reads

1: **for** $i = n : -1 : 1$ **do**
2: $\quad z = b(i)$
3: $\quad$ **for** j=1:n **do**
4: $\quad\quad z = z - a(i,j) * x(j)$
5: $\quad$ **end for**
6: $\quad x(i) = x(i) + \omega z / a(i,i)$
7: **end for**

Symmetric splitting methods ct.

If we combine the forward and the backward scheme we obtain the symmetric Gauß-Seidel method which is defined by the iteration matrix

$$G_{GS}^s = (D-F)^{-1}E(D-E)^{-1}F$$

and the symmetric SOR method or SSOR method given by

$$G_\omega^s = (D-\omega F)^{-1}((1-\omega)D+\omega E)(D-\omega E)^{-1}((1-\omega)D+\omega F).$$

It can be shown easily that the splitting matrices of the symmetric Gauß-Seidel method and the symmetric SOR method are given by

$$B_{GS}^s = (D-E)D^{-1}(D-F)$$

and

$$B_\omega^s = \frac{1}{\omega(2-\omega)}(D-\omega E)D^{-1}(D-\omega F),$$

respectively.

Symmetric splitting methods

If A is symmetric and positive definite then $F = E^H$ and the diagonal part D is positive definite, and the symmetric Gauß-Seidel method and the SSOR are indeed symmetric iteration schemes. Their convergence properties with respect to the energy norm are given in the following theorem.

Theorem 2.7
Let $A = D - E - E^H$ be Hermitean and positive definite. Then for every $\omega \in (0, 2)$ the SSOR method is convergent, and

$$\rho(G_\omega^s) = \|G_\omega^s\|_A = \|G_\omega\|_A^2.$$

Moreover, the spectrum of G_ω^s is contained in $[0, \rho(G_\omega^s)]$.

Proof of Theorem 2.7

The matrix G_ω^s is similar to

$$A^{1/2} G_\omega^s A^{-1/2} = \left(A^{1/2} G_\omega^b A^{-1/2}\right)\left(A^{1/2} G_\omega A^{-1/2}\right).$$

From $G_\omega = I - (D - \omega E)^{-1} A$ and $G_\omega^b = I - (D - \omega E^H)^{-1} A$ one obtains for the first factor

$$\begin{aligned} A^{1/2} G_\omega^b A^{-1/2} &= I - A^{1/2}(D - \omega E^H)^{-1} A^{1/2} \\ &= \left(I - A^{1/2}(D - \omega E)^{-1} A^{1/2}\right)^H = \left(A^{1/2} G_\omega A^{-1/2}\right)^H. \end{aligned}$$

Hence, $A^{1/2} G_\omega^s A^{-1/2}$ is Hermitean and positive semidefinite, and therefore, $\operatorname{spec}(G_\omega^s) \subset [0, \rho(G_\omega^s)]$.

Proof of Theorem 2.7 ct.

$\rho(G_\omega^s) = \|G_\omega^s\|_A = \|G_\omega\|_A^2$ follows from

$$\rho(G_\omega^s) = \rho(A^{1/2} G_\omega^s A^{-1/2}) = \|A^{1/2} G_\omega^s A^{-1/2}\|_2 = \|G_\omega^s\|_A$$

and

$$\begin{aligned} \|A^{1/2} G_\omega^s A^{-1/2}\|_2 &= \|(A^{1/2} G_\omega A^{-1/2})^H (A^{1/2} G_\omega A^{-1/2})\|_2 \\ &= \|A^{1/2} G_\omega A^{-1/2}\|_2^2 = \|G_\omega\|_A^2. \end{aligned}$$

This completes the proof. □

SSOR

At a first glance, two properties of a symmetrized method seem to be at hand.

Firstly, it should require twice as much work as the underlying simple method,

and secondly, it should converge faster because it is the composition of the base method and a second convergent method.

BOTH CONJECTURES ARE FALSE.

SSOR ct.

Due to an observation of Niethammer (1964) (which was rediscovered by Conrad & Wallach (1977)) a symmetrized method can be performed with nearly the same effort as the base method. We demonstrate this for the SSOR method.

The first halfstep (forward sweep) can be written as

$$x^{k+1/2} := x^k + \omega D^{-1}(b - Dx^k + Fx^k + Ex^{k+1/2}) \qquad (*)$$

and the second halfstep (backward sweep) is given by

$$x^{k+1} := x^{k+1/2} + \omega D^{-1}(b - Dx^{k+1/2} + Ex^{k+1/2} + Fx^{k+1}) \qquad (**).$$

In the first halfstep $Ex^{k+1/2}$ has already been evaluated and can be used in (**). Moreover, Fx^{k+1}, which is computed in (**), can be used in the first half of (*) of the following iteration step. Hence, in every SSOR step we have to perform only one matrix-vector multiplication.

SSOR ct.

From Theorem 2.7 one obtains that with respect to the energy norm the SSOR method indeed converges twice as fast as the SOR method. The following example shows that with respect to the spectral radius, the norm independent measure of convergence, in general this is not true.

Example

We consider again the difference approximation of the model problem with lexicographical order of the unknowns and $n = 128$. The following table contains the spectral radii of the iteration matrices of the SSOR method for different values of the relaxation parameter. Obviously, the optimum parameter is $\omega^* \approx 1.959$ and the corresponding spectral radius is $\rho(G^S_{\omega^*}) \approx 0.96819$.

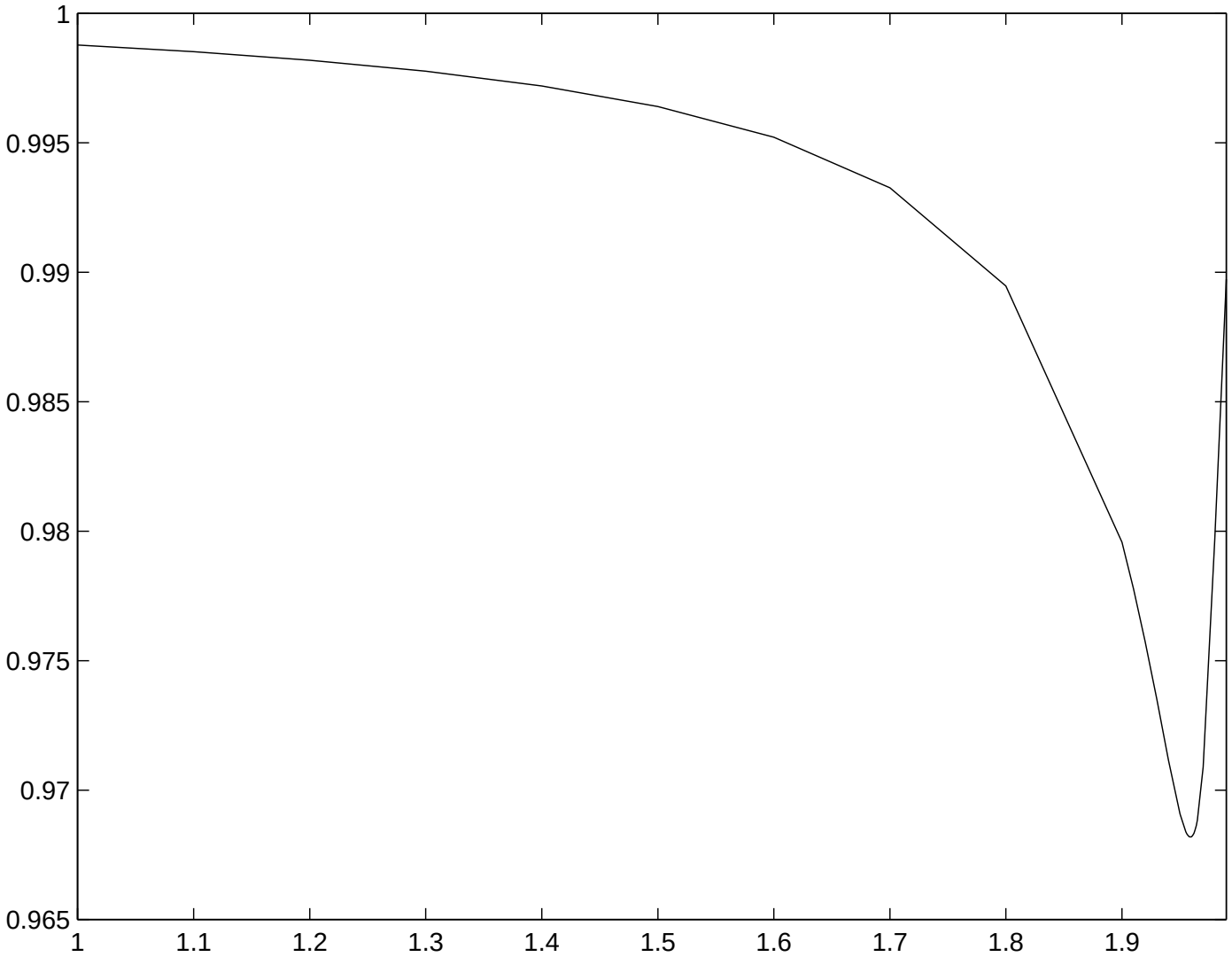
SSOR ct.

ω	$\rho(G_\omega)$	ω	$\rho(G_\omega)$	ω	$\rho(G_\omega)$
1.0	0.99878			1.955	0.96839
1.1	0.99852	1.91	0.97777	1.956	0.96831
1.2	0.99819	1.92	0.97574	1.957	0.96824
1.3	0.99777	1.93	0.97350	1.958	0.96820
1.4	0.99720	1.94	0.97116	1.959	0.96819
1.5	0.99640	1.95	0.96908	1.960	0.96820
1.6	0.99522	1.96	0.96820	1.961	0.96825
1.7	0.99326	1.97	0.97089	1.962	0.96833
1.8	0.98947	1.98	0.97966	1.963	0.96846
1.9	0.97958	1.99	0.98973	1.964	0.96863

For the (forward) SOR method with lexicographic ordering the optimal ω (obtained experimentally) is $\tilde{\omega} = 1.956$ and $\rho(G_{\tilde{\omega}}) = 0.959$.

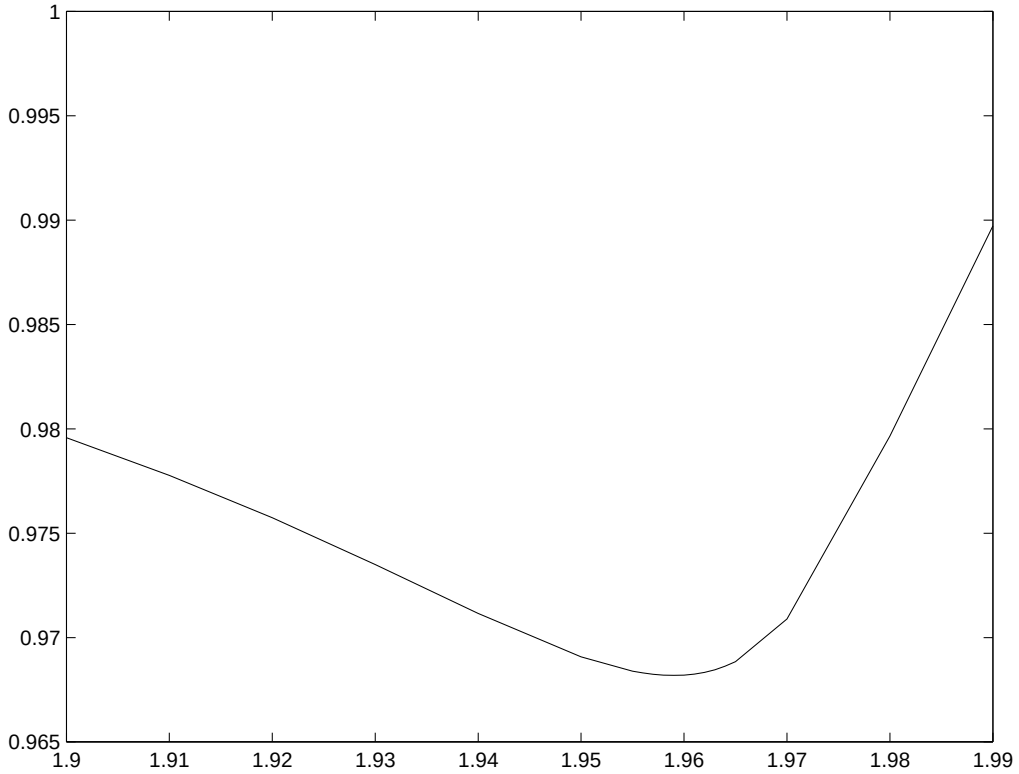
SSOR ct.

Graph of the mapping $\omega \mapsto \rho(G_\omega^s)$.



SSOR ct.

Close up of graph of the mapping $\omega \mapsto \rho(G_\omega^s)$.



Close to the optimum parameter it seems to be smoother than that of the SOR method. Hence, the choice of the parameter is not as critical as in the SOR case.

2-cyclic matrix

Even more surprising is the analysis of the SSOR method for the following class of matrices:

Definition
The matrix $A \in \mathbb{C}^{n\times n}$ is 2-cyclic if it has the block form

$$A = \begin{pmatrix} D_1 & \tilde{F} \\ \tilde{E} & D_2 \end{pmatrix},$$

where D_1 and D_2 are diagonal matrices.

Notice that the matrices $\tilde{E}$ and $\tilde{F}$ need not be quadratic.

2-cyclic matrix ct.

If A is a regular and 2-cyclic matrix then the iteration matrix of the Gauß-Seidel method is given by

$$\begin{aligned} G_{GS} &= (D-E)^{-1}F = \begin{pmatrix} D_1 & O \\ \tilde{E} & D_2 \end{pmatrix}^{-1} \begin{pmatrix} O & -\tilde{F} \\ O & O \end{pmatrix} \\ &= \begin{pmatrix} D_1^{-1} & O \\ -D_2^{-1}\tilde{E}D_1^{-1} & D_2^{-1} \end{pmatrix} \begin{pmatrix} O & -\tilde{F} \\ O & O \end{pmatrix} = \begin{pmatrix} O & -D_1^{-1}\tilde{F} \\ O & D_2^{-1}\tilde{E}D_1^{-1}\tilde{F} \end{pmatrix}. \quad (+) \end{aligned}$$

Correspondingly, we obtain the iteration matrix of the backward Gauß-Seidel method

$$G_{GS}^b = \begin{pmatrix} D_1^{-1}\tilde{F}D_2^{-1}\tilde{E} & O \\ -D_2^{-1}\tilde{F} & O \end{pmatrix},$$

2-cyclic matrix ct.

and therefore, the iteration matrix of the symmetric Gauß-Seidel method is given by

$$G_{GS}^{s} = G_{GS}^{b} G_{GS} = \begin{pmatrix} O & -D_1^{-1}\tilde{F}D_2^{-1}\tilde{E}D_1^{-1}\tilde{F} \\ O & D_2^{-1}\tilde{E}D_1^{-1}\tilde{F} \end{pmatrix}. \qquad (++)$$

From (+) and (++) it follows that the Gauß-Seidel method and its symmetric version have the same speed of convergence because the spectral radii of their iteration matrices coincide.

2-cyclic matrix ct.

We now consider the SSOR method in the 2-cyclic case. The iteration matrices of the two halfsteps of the SOR method can be written as

$$G_\omega^{(1)} = \begin{pmatrix} (1-\omega)I & \omega D_1^{-1}\tilde{F} \\ O & I \end{pmatrix}, \quad G_\omega^{(2)} = \begin{pmatrix} I & O \\ \omega D_2^{-1}\tilde{E} & (1-\omega)I \end{pmatrix}.$$

Hence, the iteration matrix of the SSOR method is given by

$$G_\omega^s = G_\omega^{(1)} G_\omega^{(2)} G_\omega^{(2)} G_\omega^{(1)}$$

which is similar to

$$(G_\omega^{(1)})^{-1} G_\omega^s G_\omega^{(1)} = G_\omega^{(2)} G_\omega^{(2)} G_\omega^{(1)} G_\omega^{(1)}.$$

2-cyclic matrix ct.

Moreover, it holds that

$$G_\omega^{(1)} G_\omega^{(1)} = \begin{pmatrix} (1-\omega)^2 I & (2-\omega)\omega D_1^{-1}\tilde{F} \\ O & I \end{pmatrix} = G_{\omega(2-\omega)}^{(1)}$$

and $G_\omega^{(2)} G_\omega^{(2)} = G_{\omega(2-\omega)}^{(2)}$, and therefore,

$$\rho(G_\omega^s) = \rho(G_{\omega(2-\omega)}).$$

Obviously, $\omega(2-\omega) < 1$ for every $\omega \in (0,2)$, $\omega \neq 1$, and the spectral radius of the iteration matrix of the SOR method is strictly decreasing in the parameter intervall $(0,1]$.

Hence, the optimum SSOR method in the 2-cyclic case is the symmetric Gauß-Seidel method, and the speed of convergence is identical to that of the simple Gauß-Seidel method.