~

Iterative Solver convergence

Scientific Computing Winter 2016/2017

Lecture 10

With material from R. S. Varga "Matrix Iterative Analysis"

Jürgen Fuhrmann

juergen.fuhrmann@wias-berlin.de

made wit pandoc

$\sim$

Criteria for the M-Property of a matrix

Iterative methods so far

- main thread ("Roter Faden"):
  - Simple iterative methods converge if the spectral radius of the iteration matrix is less than one
  - If a matrix has the M-Property (positve main diagonal entries, nonpositive off diagonal entries, nonsingular, inverse nonnegative), then methods based regular splittings converge
  - But: how can we see that a matrix has the M-Property?
- This theory is useful in other contexts as well
- Main source: Varga, "Matrix Iterative Analysis"

The Gershgorin Circle Theorem

(everywhere, we assume $n \geq 2$)

Theorem Let A be an $n \times n$ (complex) matrix. Let

$$\Lambda_i = \sum_{\substack{j=1\dots n \\ j \neq i}} |a_{ij}|$$

If λ is an eigenvalue of A then there is r, $1 \leq r \leq n$ such that

$$|\lambda - a_{rr}| \leq \Lambda_r$$

Proof Assume λ is eigenvalue, x a corresponding eigenvector, normalized such that $\max_{i=1\dots n} |x_i| = |x_r| = 1$. From $Ax = \lambda x$ it follows that

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1\dots n \\ j \neq i}} a_{ij}x_j$$

$$|\lambda - a_{rr}| = |\sum_{\substack{j=1\dots n \\ j \neq r}} a_{rj}x_j| \leq \sum_{\substack{j=1\dots n \\ j \neq r}} |a_{rj}||x_j| \leq \sum_{\substack{j=1\dots n \\ j \neq r}} |a_{rj}| = \Lambda_r$$

Gershgorin Circle Corollaries

Corollary: Any eigenvalue of A lies in the union of the disks defined by the Gershgorin cicles

$$\lambda \in \bigcup_{i=1\dots n} \{\mu \in \mathbb{C} : |\mu - |a_{ii}|| \leq \Lambda_i\}$$

Corollary:

$$\rho(A) \leq \max_{i=1\dots n} \sum_{j=1}^{n} |a_{ij}| = ||A||_\infty$$

$$\rho(A) \leq \max_{j=1\dots n} \sum_{i=1}^{n} |a_{ij}| = ||A||_1$$

Proof

$$|\mu - a_{ii}| \leq \Lambda_i \quad \Rightarrow \quad |\mu| \leq \Lambda_i + |a_{ii}| = \sum_{j=1}^{n} |a_{ij}|$$

Furthermore, $\sigma(A) = \sigma(A^T)$. $\square$

Reducible and irreducible matrices

Definition A is *reducible* if there exists a permutation matrix P such that

$$PAP^T = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

A is *irreducible* if it is not reducible.

Directed matrix graph:

- Nodes: $\mathcal{N} = \{N_i\}_{i=1\ldots n}$
- Directed edges: $\mathcal{E} = \{\vec{N_k N_l} | a_{kl} \neq 0\}$

A is irreducible $\Leftrightarrow$ the matrix graph is connected, i.e. for each *ordered* pair N_i, N_j there is a path consisting of directed edges, connecting them.

Equivalently, for each i, j there is a sequence of nonzero matrix entries $a_{ik_1}, a_{k_1 k_2}, \ldots, a_{k_r j}$.

Taussky theorem

Theorem Let A be irreducible. Assume that the eigenvalue λ is a boundary point of the union of all the disks

$$\lambda \in \partial \bigcup_{i=1\ldots n} \{\mu \in \mathbb{C} : |\mu - a_{ii}| \leq \Lambda_i\}$$

Then, all n Gershgorin circles pass through λ, i.e. for $i = 1 \ldots n$,

$$|\lambda - a_{ii}| = \Lambda_i$$

Taussky theorem proof

Proof Assume λ is eigenvalue, x a corresponding eigenvector, normalized such that $\max_{i=1\dots n} |x_i| = |x_r| = 1$. From $Ax = \lambda x$ it follows that

$$|\lambda - a_{rr}| \leq \sum_{\substack{j=1\dots n \\ j\neq r}} |a_{rj}| \cdot |x_j| \leq \sum_{\substack{j=1\dots n \\ j\neq r}} |a_{rj}| = \Lambda_r \qquad (*)$$

Boundary point $\Rightarrow |\lambda - a_{rr}| = \Lambda_r$

$\Rightarrow$ For all $l \neq r$ with $a_{r,p} \neq 0$, $|x_p| = 1$.

Due to irreducibility there is at least one such p. For this p, equation $(*)$ is valid $\Rightarrow |\lambda - a_{pp}| = \Lambda_p$

Due to irreducibility, this is true for all $p = 1 \dots n$ $\square$

Diagonally dominant matrices

Definition

- A is *diagonally dominant* if for $i = 1 \dots n$,

$$|a_{ii}| \geq \sum_{\substack{j=1\dots n \\ j \neq i}} |a_{ij}|$$

- A is *strictly diagonally dominant* (sdd) if for $i = 1 \dots n$,

$$|a_{ii}| > \sum_{\substack{j=1\dots n \\ j \neq i}} |a_{ij}|$$

- A is *irreducibly diagonally dominant* (idd) if A is irreducible, for $i = 1 \dots n$,

$$|a_{ii}| \geq \sum_{\substack{j=1\dots n \\ j \neq i}} |a_{ij}|$$

  and for at least one r, $1 \leq r \leq n$,

$$|a_{rr}| > \sum_{\substack{j=1\dots n \\ j \neq r}} |a_{rj}|$$

A very practical nonsingularity criterion

Theorem: Let A be strictly diagonally dominant or irreducibly diagonally dominant. Then A is nonsingular.

If in addition, if $a_{ii} > 0$ for $i = 1 \dots n$, then all real parts of the eigenvalues of A are positive:

$$\mathrm{Re}\lambda_i > 0, \quad i = 1 \dots n$$

Proof:

Assume A strictly diagonally dominant. Then the union of the Gershgorin disks does not contain 0 and $\lambda = 0$ cannot be an eigenvalue.

As for the real parts, the union of the disks is

$$\bigcup_{i=1\dots n} \{\mu \in \mathbb{C} : |\mu - a_{ii}| \leq \Lambda_i\}$$

and $\mathrm{Re}\mu$ must be larger than zero if it should be contained.

A very practical nonsingularity criterion II

Assume A irreducibly diagonally dominant. Then, if 0 is an eigenvalue, by the Taussky theorem, we have $|a_{ii}| = \Lambda_i$ for all $i = 1 \dots n$. This is a contradiction as by definition there is at least one i such that $|a_{ii}| > \Lambda_i$

Obviously, all real parts of the eigenvalues must be ≥ 0. Therefore, if a real part is 0, it lies on the boundary of one disk. So by Taussky it must be contained in the boundary of all the disks and the imaginary axis. But there is at least one disk which does not touch the imaginary axis. $\square$

Corollary

Theorem: If A is symmetric, sdd or idd, with positive diagonal entries, it is positive definite.

Proof: All eigenvalues of A are real, and due to the nonsingularity criterion, they must be positive, so A is positive definite. $\square$.

Theorem on Jacobi matrix

Theorem: Let A be sdd or idd, and D its diagonal. Then

$$\rho(|I - D^{-1}A|) < 1$$

Proof: Let $B = (b_{ij}) = I - D^{-1}A$. Then

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

If A is sdd, then for $i = 1 \dots n$,

$$\sum_{j=1\dots n} |b_{ij}| = \sum_{\substack{j=1\dots n \\ j\neq i}} |\frac{a_{ij}}{a_{ii}}| = \frac{\Lambda_i}{|a_{ii}|} < 1$$

Therefore, $\rho(|B|) < 1$.

Theorem on Jacobi matrix II

If A is idd, then for $i = 1 \dots n$,

$$\sum_{j=1\dots n} |b_{ij}| = \sum_{\substack{j=1\dots n \\ j\neq i}} |\frac{a_{ij}}{a_{ii}}| = \frac{\Lambda_i}{|a_{ii}|} \leq 1$$

$$\sum_{j=1\dots n} |b_{rj}| = \frac{\Lambda_r}{|a_{rr}|} < 1 \text{ for at least one } r$$

Therefore, $\rho(|B|) <= 1$. Assume $\rho(|B|) = 1$ By Perron-Frobenius, 1 is an eigenvalue. As it is in the union of the Gershgorin disks

$$|\lambda| = 1 \leq \frac{\Lambda_i}{|a_{ii}|} \leq 1$$

it must lie on the boundary of this union, and by Taussky one has for all i

$$|\lambda| = 1 \leq \frac{\Lambda_i}{|a_{ii}|} = 1$$

which contradicts the idd condition. $\square$

Jacobi method convergence

Corollary: Let A be sdd or idd, and D its diagonal. Assume that $a_{ii} > 0$ and $a_{ij} \leq 0$ for $i \neq j$. Then $\rho(I - D^{-1}A) < 1$, i.e. the Jacobi method converges.

Proof In this case, $|B| = B$. $\square$.

Main Practical M-Matrix Criterion

Corollary: Let A be sdd or idd. Assume that $a_{ii} > 0$ and $a_{ij} \leq 0$ for $i \neq j$. Then A is an M-Matrix, i.e. A is nonsingular and $A^{-1} \geq 0$.
Proof: Let $B = \rho(I - D^{-1}A)$. Then $\rho(B) < 1$, therefore $I - B$ is nonsingular.

We have for $k > 0$:

$$
\begin{aligned}
I - B^{k+1} &= (I - B)(I + B + B^2 + \cdots + B^k) \\
(I - B)^{-1}(I - B^{k+1}) &= (I + B + B^2 + \cdots + B^k)
\end{aligned}
$$

The left hand side for $k \to \infty$ converges to $(I - B)^{-1}$, therefore

$$
(I - B)^{-1} = \sum_{k=0}^{\infty} B^k
$$

As $B \geq 0$, we have $(I - B)^{-1} = A^{-1}D \geq 0$. As $D > 0$ we must have $A^{-1} \geq 0$. $\square$

Regular splittings

- $A = M - N$ is a regular splitting if
  - M is nonsingular
  - M^{-1}, N are nonnegative, i.e. have nonnegative entries
- Regard the iteration $u_{k+1} = M^{-1}Nu_k + M^{-1}b$.
- We have $I - M^{-1}A = M^{-1}N$.

Convergence theorem for regular splitting

Theorem: Assume A is nonsingular, $A^{-1} \geq 0$, and $A = M - N$ is a regular splitting. Then $\rho(M^{-1}N) < 1$.

Proof: Let $G = M^{-1}N$. Then $A = M(I - G)$, therefore $I - G$ is nonsingular.

In addition

$$A^{-1}N = (M(I - M^{-1}N))^{-1}N = (I - M^{-1}N)^{-1}M^{-1}N = (I - G)^{-1}G$$

By Perron-Frobenius, there $\rho(G)$ is an eigenalue with a nonnegative eigenvector x . Thus,

$$0 \leq A^{-1}Nx = \frac{\rho(G)}{1 - \rho(G)}x$$

Therefore $0 \leq \rho(G) \leq 1$. As $I - G$ is nonsingular, $\rho(G) < 1$ $\square$.

Convergence rate

Corollary: $\rho(M^{-1}N) = \frac{\tau}{1+\tau}$ where $\tau = \rho(A^{-1}N)$.

Proof: Rearrange $\tau = \frac{\rho(G)}{1-\rho(G)}$ $\square$

Corollary: Let $A \geq 0$, $A = M_1 - N_1$ and $A = M_2 - N_2$ be regular splittings. If $N_2 \geq N_1 \geq 0$, then $1 > \rho(M_2^{-1}N_2) \geq \rho(M_1^{-1}N_1)$.

Proof: $\tau_2 = \rho(A^{-1}N_2) \geq \rho(A^{-1}N_1) = \tau_1$, $\frac{\tau}{1+\tau}$ is strictly increasing.

Application

Let A be an M-Matrix. Assume $A = D - E - F$.

- Jacobi method: $M = D$ is nonsingular, $M^{-1} \geq 0$. $N = E + F$ nonnegative $\Rightarrow$ convergence
- Gauss-Seidel: $M = D - E$ is an M-Matrix as $A \leq M$ and M has non-positive off-digonal entries. $N = F \geq 0$. $\Rightarrow$ convergence
- Comparison: $N_J \geq N_{GS} \Rightarrow$ Gauss-Seidel converges faster.

Intermediate Summary

- Given some matrix, we now have some nice recipies to establish nonsingularity and iterative method convergence:
- **Check if the matrix is irreducible.**
  This is mostly the case for elliptic and parabolic PDEs.
- **Check for if matrix is strictly or irreducibly diagonally dominant**.
  If yes, it is in addition nonsingular.
- **Check if main diagonal entries are positive and off-diagonal entries are nonpositive.**
  If yes, in addition, the matrix is an M-Matrix, its inverse is nonnegative, and elementary iterative methods converge.

Example: 1D finite volume matrix:

We assume $\alpha > 0$.

$$\begin{pmatrix} \alpha + \frac{1}{h} & -\frac{1}{h} & & & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ & & & & -\frac{1}{h} & \frac{1}{h} + \alpha \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-2} \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} \frac{h}{2} f_1 + \alpha v_1 \\ h f_2 \\ h f_3 \\ \vdots \\ h f_{N-2} \\ h f_{N-1} \\ \frac{h}{2} f_N + \alpha v_n \end{pmatrix}$$

- idd
- main diagonal entries are positive and off-diagonal entries are nonpositive

So this matrix is nonsingular, has the M-property, and we can e.g. apply the Jacobi iterative method to solve it.

Moreover, due to $A^{-1} \geq 0$, for $f \geq 0$ and $v \geq 0$ it follows that $u \geq 0$.

Incomplete LU factorizations (ILU)

Idea (Varga, Buleev, 1960):

- fix a predefined zero pattern
- apply the standard LU factorization method, but calculate only those elements, which do not correspond to the given zero pattern
- Result: incomplete LU factors L, U, remainder R:

$$A = LU - R$$

- Problem: with complete LU factorization procedure, for any nonsingular matrix, the method is stable, i.e. zero pivots never occur. Is this true for the incomplete LU Factorization as well ?

Stability of ILU

Theorem (Saad, Th. 10.2): If A is an M-Matrix, then the algorithm to compute the incomplete LU factorization with a given nonzero pattern

$$A = LU - R$$

is stable. Moreover, $A = LU - R$ is a regular splitting.

ILU(0)

- Special case of ILU: ignore any fill-in.
- Representation:

$$M = (\tilde{D} - E)\tilde{D}^{-1}(\tilde{D} - F)$$

- $\tilde{D}$ is a diagonal matrix (wich can be stored in one vector) which is calculated by the incomplete factorization algorithm.
- Setup:

```
for i=1...n do
    d(i)=a(i,i)
end

for i=1...n do
     d(i)=1.0/d(i)
     for j=i+1 ... n do
         d(j)=d(j)-a(i,j)*d(i)*a(j,i)
     end
end
```

ILU(0)

Solve $Mu = v$

```
for i=1...n do
     x=0
     for j=1 ... i-1 do
        x=x+a(i,j)*u(j)
     end
     u(i)=d(i)*(v(i)-x)
end

for i=n...1 do
     x=0
     for j=i+1...n do
        x=x+a(i,j)*u(j)
     end
     u(i)=u(i)-d(i)*x
```

ILU(0)

- Generally better convergence properties than Jacobi, Gauss-Seidel
- One can develop block variants
- Alternatives:
  - ILUM: ("modified"): add ignored off-diagonal entries to $\tilde{D}$
  - ILUT: zero pattern calculated dynamically based on drop tolerance
- Dependence on ordering
- Can be parallelized using graph coloring
- Not much theory: experiment for particular systems
- I recommend it as the default initial guess for a sensible preconditioner
- Incomplete Cholesky: symmetric variant of ILU

Preconditioners

- Leave this topic for a while now
- Hopefully, we well be able to discuss
  - Multigrid: gives $O(n)$ complexity in optimal situations
  - Domain decomposition: Structurally well suited for large scale parallelization

$\sim$

More general iteration schemes

Generalization of iteration schemes

- Simple iterations converge slowly
- For most practical purposes, Krylov subspace methods are used.
- We will introduce one special case and give hints on practically useful more general cases
- Material after J. Shewchuk: !An Introduction to the Conjugate Gradient Method Without the Agonizing Pain“

Solution of SPD system as a minimization procedure

Regard $Au = f$,where A is symmetric, positive definite. Then it defines a bilinear form $a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

$$a(u, v) = (Au, v) = v^T Au = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} v_i u_j$$

As A is SPD, for all $u \neq 0$ we have $(Au, u) > 0$.

For a given vector b, regard the function

$$f(u) = \frac{1}{2} a(u, u) - b^T u$$

What is the minimizer of f ?

$$f'(u) = Au - b = 0$$

- Solution of SPD system $\equiv$ minimization of f.

Method of steepest descent

- Given some vector u_i look for a new iterate u_{i+1}.
- The direction of steepest descend is given by $-f'(u_i)$.
- So look for u_{i+1} in the direction of $-f'(u_i) = r_i = b - Au_i$ such that it minimizes f in this direction, i.e. set $u_{i+1} = u_i + \alpha r_i$ with α choosen from

$$
\begin{aligned}
0 &= \frac{d}{d\alpha} f(u_i + \alpha r_i) = f'(u_i + \alpha r_i) \cdot r_i \\
&= (b - A(u_i + \alpha r_i), r_i) \\
&= (b - Au_i, r_i) - \alpha(Ar_i, r_i) \\
&= (r_i, r_i) - \alpha(Ar_i, r_i) \\
\alpha &= \frac{(r_i, r_i)}{(Ar_i, r_i)}
\end{aligned}
$$

Method of steepest descent: iteration scheme

$$r_i = b - Au_i$$
$$\alpha_i = \frac{(r_i, r_i)}{(Ar_i, r_i)}$$
$$u_{i+1} = u_i + \alpha_i r_i$$

Let $\hat{u}$ the exact solution. Define $e_i = u_i - \hat{u}$. Let $||u||_A = (Au, u)^{\frac{1}{2}}$ be the *energy norm* wrt. A.

Theorem The convergence rate of the method is

$$||e_i||_A \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^i ||e_0||_A$$

Conjugate directions

For steepest descent, there is no guarantee that a search direction $d_i = r_i = Ae_i$ is not used several times. If all search directions would be orthogonal, or, indeed, A-orthogonal, one could control this situation.

So, let $d_0, d_1 \dots d_{n-1}$ be a series of A-orthogonal (or conjugate) search directions, i.e. $(Ad_i, d_j) = 0,\ i \neq j$.

- Look for u_{i+1} in the direction of d_i such that it minimizes f in this direction, i.e. set $u_{i+1} = u_i + \alpha d_i$ with α choosen from

$$
\begin{aligned}
0 &= \frac{d}{d\alpha} f(u_i + \alpha d_i) = f'(u_i + \alpha d_i) \cdot d_i \\
&= (b - A(u_i + \alpha d_i), d_i) \\
&= (b - Au_i, d_i) - \alpha(Ad_i, d_i) \\
&= (r_i, d_i) - \alpha(Ad_i, d_i) \\
\alpha &= \frac{(r_i, d_i)}{(Ad_i, d_i)}
\end{aligned}
$$

Conjugate directions II

$e_0 = u_0 - \hat{u}$ (such that $Ae_0 = -r_0$) can be represented in the basis of the search directions:

$$e_0 = \sum_{i=0}^{n-1} \delta_j d_j$$

Projecting onto d_k in the A scalar product gives

$$\begin{aligned}
(Ae_0, d_k) &= \sum_{i=0}^{n-1} \delta_j (Ad_j, d_k) \\
(Ae_0, d_k) &= \delta_k (Ad_k, d_k) \\
\delta_k &= \frac{(Ae_0, d_k)}{(Ad_k, d_k)} = \frac{(Ae_0 + \sum_{i<k} \alpha_i d_i, d_k)}{(Ad_k, d_k)} = \frac{(Ae_k, d_k}{(Ad_k, d_k)} \\
&= \frac{(r_k, d_k)}{(Ad_k, d_k)} \\
&= -\alpha_k
\end{aligned}$$

Conjugate directions III

Then,

$$
\begin{aligned}
e_i &= e_0 + \sum_{j=0}^{i-1} \alpha_j \boldsymbol{d}_j \\
&= -\sum_{j=0}^{n-1} \alpha_j \boldsymbol{d}_j + \sum_{j=0}^{i-1} \alpha_j \boldsymbol{d}_j \\
&= -\sum_{j=i}^{n-1} \alpha_j \boldsymbol{d}_j
\end{aligned}
$$

So, the iteration consists in component-wise suppression of the error, and it must converge after n steps.

But by what magic we can obtain these d_i?