From Direct to Iterative Solvers

Scientific Computing Winter 2016/2017

Lecture 8

With material from Y. Saad "Iterative Methods for Sparse Linear Systems"

Jürgen Fuhrmann

juergen.fuhrmann@wias-berlin.de

made wit pandoc

Recap from last time

Matrices from PDE: a first example

- "Drosophila": Poisson boundary value problem in rectangular domain

Given:

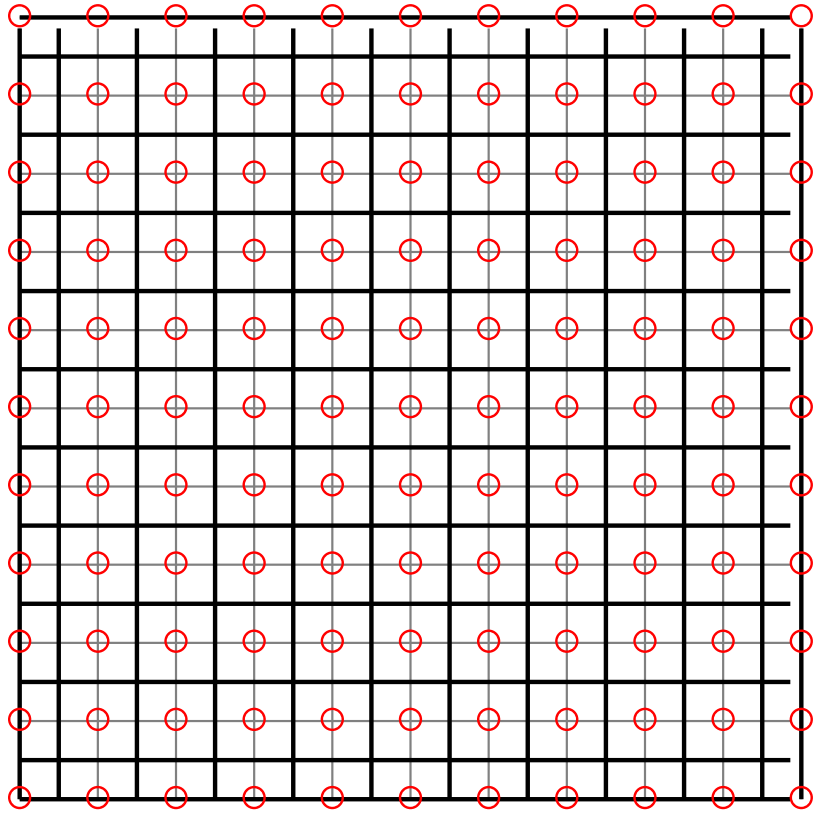
- Domain $\Omega = (0, X) \times (0, Y) \subset \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$, outer normal $\mathbf{n}$
- Right hand side $f : \Omega \to \mathbb{R}$
- "Conductivity" λ
- Boundary value $v : \Gamma \to \mathbb{R}$
- Transfer coefficient α

Search function $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot \lambda \nabla u &= f \quad \text{in} \Omega \\ -\lambda \nabla u \cdot \mathbf{n} + \alpha(u - v) &= 0 \quad \text{on} \Gamma \end{aligned}$$

- Example: heat conduction:
  - u: temperature
  - f: volume heat source
  - λ: heat conduction coefficient
  - v: Ambient temperature
  - α: Heat transfer coefficient

2D finite volume grid



- Red circles: discretization nodes
- Thin lines: original “grid”
- Thick lines: boundaries of control volumes
- Each discretization point has not more then 4 neighbours

Discretization matrix (2D)

Assume $\lambda = 1$, $h_{kl} = h$ and we count collocation points in each direction from $1 \dots N$. For $i = 2 \dots N-1$, $j = 2 \dots N-1$, $k = N*(j-1)+i$ one has$|\omega_K| = h^2$, $|\sigma_{KL}| = h$, and

$$\sum_{L \in \mathcal{N}_k} \frac{|\sigma_{kl}|}{h_{kl}}(u_k - u_l) = -u_{k-N} - u_{k-1} + 2u_k - u_{k+1} - u_{k+N}$$

The linear system then has 5 nonzero diagonals

Sparse matrices

- Regardless of number of unknowns n, the number of non-zero entries per row remains limited by n_r
- If we find a scheme which allows to store only the non-zero matrix entries, we would need $nn_r = O(n)$ storage locations instead of n^2
- The same would be true for the matrix-vector multiplication if we program it in such a way that we use every nonzero element just once:
  martrix-vector multiplication uses $O(n)$ instead of $O(n^2)$ operartions

Compressed Row Storage (CRS) format with 0-based indexing

(aka Compressed Sparse Row (CSR) or IA-JA etc.)

- real array `AA`, length nnz, containing all nonzero elements row by row
- integer array `JA`, length nnz, containing the column indices of the elements of `AA`
- integer array `IA`, length n+1, containing the start indizes of each row in the arrays `IA` and `JA` and `IA(n)=nnz`

$$A = \begin{pmatrix} 1. & 0. & 0. & 2. & 0. \\ 3. & 4. & 0. & 5. & 0. \\ 6. & 0. & 7. & 8. & 9. \\ 0. & 0. & 10. & 11. & 0. \\ 0. & 0. & 0. & 0. & 12. \end{pmatrix}$$

```
AA: 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12.
JA: 0 3 0 1 3 0 2 3 4 2 3 4
IA: 0 2 4 0 11 12
```

Sparse direct solvers

- Sparse direct solvers implement Gaussian elimination with different pivoting strategies
  - UMFPACK
  - Pardiso (omp + MPI parallel)
  - SuperLU
  - MUMPS (MPI parallel)
  - Pastix
- Quite efficient for 1D/2D problems
- They suffer from *fill-in*: $\Rightarrow$ huge memory usage for 3D

Using sparse direct solvers

Sparse direct solvers: solution steps (Saad Ch. 3.6)

1. Pre-ordering
   - The amount of non-zero elements generated by fill-in can be decreases by re-ordering of the matrix
   - Several, graph theory based heuristic algorithms exist
2. Symbolic factorization
   - If pivoting is ignored, the indices of the non-zero elements are calculated and stored
   - Most expensive step wrt. computation time
3. Numerical factorization
   - Calculation of the numerical values of the nonzero entries
   - Not very expensive, once the symbolic factors are available
4. Upper/lower triangular system solution
   - Fairly quick in comparison to the other steps

- Separation of steps 2 and 3 allows to save computational costs for problems where the sparsity structure remains unchanged, e.g. time dependent problems on fixed computational grids
- With pivoting, steps 2 and 3 have to be performed together
- Instead of pivoting, *iterative refinement* may be used in order to maintain accuracy of the solution

Interfacing UMFPACK from C++ (numcxx)

(shortened version of the code)

```
#include <suitesparse/umfpack.h>

// Calculate LU factorization
template<> inline void TSolverUMFPACK<double>::update()
{
    pMatrix->flush(); // Update matrix, adding newly created elements
    int n=pMatrix->shape(0);
    double *control=nullptr;

    //Calculate symbolic factorization only if matrix patter
    //has changed
    if (pMatrix->pattern_changed())
    {
       umfpack_di_symbolic (n, n, pMatrix->pIA->data(), pMatrix->pJA->data(), pMatrix->pA->data(),
       &Symbolic, 0, 0);
    }

    umfpack_di_numeric (pMatrix->pIA->data(), pMatrix->pJA->data(), pMatrix->pA->data(),
    Symbolic, &Numeric, control, 0) ;

   pMatrix->pattern_changed(false);
}

// Solve LU factorized system
template<> inline void TSolverUMFPACK<double>::solve( TArray<T> & Sol,  const TArray<T> & Rhs)
{
    umfpack_di_solve (UMFPACK_At,pMatrix->pIA->data(), pMatrix->pJA->data(), pMatrix->pA->data(),
                      Sol.data(), Rhs.data(),
                      Numeric, control, 0 ) ;
}
```

How to use ?

```
#include <numcxx/numcxx.h>
auto pM=numcxx::DSparseMatrix::create(n,n);
auto pF=numcxx::DArray1::create(n);
auto pU=numcxx::DArray1::create(n);

auto &M=*pM;
auto &F=*pF;
auto &U=*pU;

F=1.0;
for (int i=0;i<n;i++)
{
    M(i,i)=3.0;
    if (i>0) M(i,i-1)=-1;
    if (i<n-1) M(i,i+1)=-1;
}

auto pUmfpack=numcxx::DSolverUMFPACK::create(pM);
pUmfpack->solve(U,F);
```

$\sim$

Towards iterative methodsx

Elements of iterative methods (Saad Ch.4)

Solve $Au = b$ iteratively

- Preconditioner: a matrix $M \approx A$ "approximating" the matrix A but with the property that the system $Mv = f$ is easy to solve
- Iteration scheme: algorithmic sequence using M and A which updates the solution step by step

Simple iteration with preconditioning

Idea: $A\hat{u} = b \Rightarrow$

$$\hat{u} = \hat{u} - M^{-1}(A\hat{u} - b)$$

$\Rightarrow$ iterative scheme

$$u_{k+1} = u_k - M^{-1}(Au_k - b) \quad (k = 0, 1 \dots)$$

1. Choose initial value u_0, tolerance ε, set $k = 0$
2. Calculate *residuum* $r_k = Au_k - b$
3. Test convergence: if $||r_k|| < \varepsilon$ set $u = u_k$, finish
4. Calculate *update*: solve $Mv_k = r_k$
5. Update solution: $u_{k+1} = u_k - v_k$, set $k = i + 1$, repeat with step 2.

The Jacobi method

- Let $A = D - E - F$, where D: main diagonal, E: negative lower triangular part F: negative upper triangular part
- Jacobi: $M = D$, where D is the main diagonal of A.

$$u_{k+1,i} = u_{k,i} - \frac{1}{a_{ii}} \left(\sum_{j=1 \ldots n} a_{ij} u_{k,j} - b_i \right) \qquad (i = 1 \ldots n)$$

$$a_{ii} u_{k+1,i} + \sum_{j=1 \ldots n, j \neq i} a_{ij} u_{k,j} = b_i \qquad (i = 1 \ldots n)$$

- Alternative formulation:

$$u_{k+1} = D^{-1}(E + F)u_k + D^{-1}b$$

- Essentially, solve for main diagonal element row by row
- Already calculated results not taken into account
- Variable ordering does not matter

The Gauss-Seidel method

- Solve for main diagonal element row by row
- Take already calculated results into account

$$a_{ii}u_{k+1,i} + \sum_{j<i} a_{ij}u_{k+1,j} + \sum_{j>i} a_{ij}u_{k,j} = b_i \qquad (i = 1 \ldots n)$$

$$(D - E)u_{k+1} - Fu_k = b$$

$$u_{k+1} = (D - E)^{-1}Fu_k + (D - E)^{-1}b$$

- May be it is faster
- Variable order probably matters
- The preconditioner is $M = D - E$
- Backward Gauss-Seidel: $M = D - F$
- Splitting formulation: $A = M - N$, then

$$u_{k+1} = M^{-1}Nu_k + M^{-1}b$$

Gauss an Gerling I

[6.]

[Über Stationsausgleichungen.]

GAUSS an GERLING. Göttingen, 26. December 1823.

Mein Brief ist zu spät zur Post gekommen und mir zurückgebracht. Ich erbreche ihn daher wieder, um noch die praktische Anweisung zur Elimination beizufügen. Freilich gibt es dabei vielfache kleine Localvortheile, die sich nur ex usu lernen lassen.

Ich nehme Ihre Messungen auf Orber-Reisig zum Beispiel[*].

Ich mache zuerst

$$[\text{Richtung nach}]\ 1 = 0,$$

nachher aus 1.3

$$3 = 77^0\,57'\,53{,}''107$$

(ich ziehe dies vor, weil 1.3 mehr Gewicht hat als 1.2);

dann aus

$$\left.\begin{array}{c|c|l} 13 & 1.2 & 2 = 26^0\,44'\ 7{,}''423 \\ 50 & 2.3 & 2 = \quad\ 6{,}''507 \end{array}\right\} 2 = 26^0\,44'\ 6{,}''696;$$

endlich aus

$$\left.\begin{array}{c|c|l} 26 & 1.4 & 4 = 136^0\,21'\,13{,}''481 \\ 6 & 2.4 & 4 = \quad\ 8{,}''529 \\ 78 & 3.4 & 4 = \quad 11{,}''268 \end{array}\right\} 4 = 136^0\,21'\,11{,}''641.$$

Ich suche, um die Annäherung erst noch zu vergrössern, aus

[*] Die von GERLING mitgetheilten Winkelmessungen waren (nach einem in GAUSS' Nachlass befindlichen Blatte), wenn 1 Berger Warte, 2 Johannisberg, 3 Taufstein und 4 Milseburg bezeichnet:

Rep.	Winkel
13	1.2 = 26° 44′ 7,″423
28	1.3 = 77 57 53, 107
26	1.4 = 136 21 13, 481
50	2.3 = 51 13 46, 600
6	2.4 = 109 37 1, 833
78	3.4 = 58 23 18, 161.]

$$\left.\begin{array}{c|c|l} 13 & 1.2 & 1 = -0{,}''727 \\ 28 & 1.3 & 1 = \quad 0 \\ 26 & 1.4 & 1 = -1{,}840 \end{array}\right\} 1 = -0{,}''855.$$

Da jede gemeinschaftliche Änderung aller Richtungen erlaubt ist, so lange es nur die relative Lage gilt, so ändere ich alle vier um $+0{,}''855$ und setze

$$\begin{aligned} 1 &= \quad 0^0\ 0'\ 0{,}''000 + a \\ 2 &= \quad 26\ 44\ 7{,}551 + b \\ 3 &= \quad 77\ 57\ 53{,}962 + c \\ 4 &= 136\ 21\ 12{,}496 + d. \end{aligned}$$

Es ist beim indirecten Verfahren sehr vortheilhaft, jeder Richtung eine Veränderung beizulegen. Sie können sich davon leicht überzeugen, wenn Sie dasselbe Beispiel ohne diesen Kunstgriff durchrechnen, wo Sie überdies die grosse Bequemlichkeit, an der Summe der absoluten Glieder $= 0$ immer eine Controlle zu haben, verlieren. Jetzt formire ich die vier Bedingungsgleichungen und zwar nach diesem Schema (bei eigener Anwendung und wenn die Glieder zahlreicher sind, trenne ich wohl die positiven und negativen Glieder), [wobei die Constanten in Einheiten der dritten Decimalstelle angesetzt sind:]

$ab - 1664$	$ba + 1664$	$ca + 23940$	$da - 25610$
$ac - 23940$	$bc + 9450$	$cb - 9450$	$db + 18672$
$ad + 25610$	$bd - 18672$	$cd - 29094$	$dc + 29094.$

Die Bedingungsgleichungen sind also:

$$\begin{aligned} 0 &= + \quad 6 + 67a - 13b - 28c - 26d \\ 0 &= - \ 7558 - 13a + 69b - 50c - 6d \\ 0 &= -14604 - 28a - 50b + 156c - 78d \\ 0 &= +22156 - 26a - 6b - 78c + 110d; \end{aligned}$$

$$\text{Summe} = 0.$$

Um nun indirect zu eliminiren, bemerke ich, dass, wenn 3 der Grössen a, b, c, d gleich 0 gesetzt werden, die vierte den grössten Werth bekommt, wenn d dafür gewählt wird. Natürlich muss jede Grösse aus ihrer eigenen Gleichung, also d aus der vierten, bestimmt werden. Ich setze also $d = -201$

Gauss an Gerling II

und substituire diesen Werth. Die absoluten Theile werden dann: +5232, −6352, +1074, +46; das Übrige bleibt dasselbe.

Jetzt lasse ich b an die Reihe kommen, finde $b = +92$, substituire und finde die absoluten Theile: +4036, −4, −3526, −506. So fahre ich fort, bis nichts mehr zu corrigiren ist. Von dieser ganzen Rechnung schreibe ich aber in der Wirklichkeit bloss folgendes Schema:

	$d=-201$	$b=+92$	$a=-60$	$c=+12$	$a=+5$	$b=-2$	$a=-1$
+ 6	+5232	+4036	+ 16	−320	+ 15	+41	−26
− 7558	−6352	− 4	+ 776	+176	+111	−27	−14
−14604	+1074	−3526	−1846	+ 26	−114	−14	+14
+22156	+ 46	− 506	+1054	+118	− 12	0	+26.

Insofern ich die Rechnung nur auf das nächste 2000^{tel} [der] Secunde führe, sehe ich, dass jetzt nichts mehr zu corrigiren ist. Ich sammle daher

$a=-60$	$b=+92$	$c=+12$	$d=-201$
+ 5	− 2		
− 1			
−56	+ 90	+12	−201

und füge die Correctio communis +56 bei, wodurch wird:

$$a = 0 \qquad b = +146 \qquad c = +68 \qquad d = -145,$$

also die Werthe [der Richtungen]

1	0° 0′ 0″,000
2	26 44 7,697
3	77 57 54,030
4	136 21 12,351.

Fast jeden Abend mache ich eine neue Auflage des Tableaus, wo immer leicht nachzuhelfen ist. Bei der Einförmigkeit des Messungsgeschäfts gibt dies immer eine angenehme Unterhaltung; man sieht dann auch immer gleich, ob etwas zweifelhaftes eingeschlichen ist, was noch wünschenswerth bleibt, etc. Ich empfehle Ihnen diesen Modus zur Nachahmung. Schwerlich werden Sie je wieder direct eliminiren, wenigstens nicht, wenn Sie mehr als 2 Unbekannte

haben. Das indirecte Verfahren lässt sich halb im Schlafe ausführen, oder man kann während desselben an andere Dinge denken.

. . . . . .

GAUSS an SCHUMACHER. Göttingen, 22. December 1827.

Die Einheit in meinem Coordinatenverzeichnisse ist 443,307885 [Pariser] Linien; der Logarithm zur Reduction auf Toisen

$$= 9{,}710\,1917.$$

Inzwischen gründet sich das absolute nur auf Ihre Basis, oder vielmehr auf die von CAROC mir angegebene Entfernung zwischen Hamburg und Hohenhorn, log = 4,141 1930, wofür ich also genommen habe: 4,431 0013. Sollte nach der Definitivbestimmung Ihrer Stangen Ihre Basis, und damit die obige Angabe der Entfernung Hamburg-Hohenhorn, eine Veränderung erleiden, so werden in demselben Verhältnisse auch alle meine Coordinaten zu verändern sein.

In der Form der Behandlung ist ein wichtiges Moment, dass von jedem Beobachtungsplatz ein Tableau aufgestellt wird, worin alle Azimuthe (in meinem Sinn) geordnet enthalten sind. Man hat so zum bequemsten Gebrauch fertig alles, was man von den Beobachtungen nöthig hat, so dass man nur ausnahmsweise, um diesen oder jenen Zweifel zu lösen, zu den Originalprotocollen recurrirt. .... Ist der Standpunkt von dem Zielpunkt verschieden, so reducire ich keinesweges die Beobachtungen auf letztern (Centrirung), da sie ohne diese Reduction ebenso bequem gebraucht werden können (insofern nemlich von vielen Schnitten untergeordneter Punkte die Rede ist, die nicht wieder Standpunkte sind).

Die Bildung eines solchen Tableaus beruht nun wieder auf mehrern Momenten, wozu eine Anweisung nur auf mehrere Briefe vertheilt werden kann, daher Sie vielleicht wohl thun, dieses Tableau erst selbst gleichsam zu studiren und mit den Beobachtungen zusammenzuhalten, damit Sie mir beson-

SOR and SSOR

- SOR: Successive overrelaxation: solve $\omega A = \omega B$ and use splitting

$$\omega A = (D - \omega E) - (\omega F + (1 - \omega D))$$
$$M = \frac{1}{\omega}(D - \omega E)$$

leading to

$$(D - \omega E)u_{k+1} = (\omega F + (1 - \omega D)u_k + \omega b$$

- SSOR: Symmetric successive overrelaxation

$$(D - \omega E)u_{k+\frac{1}{2}} = (\omega F + (1 - \omega D)u_k + \omega b$$
$$(D - \omega F)u_{k+1} = (\omega E + (1 - \omega D)u_{k+\frac{1}{2}} + \omega b$$

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

- Gauss-Seidel and symmetric Gauss-Seidel are special cases for $\omega = 1$.

Block methods

- Jacobi, Gauss-Seidel, (S)SOR methods can as well be used block-wise, based on a partition of the system matrix into larger blocks,
- The blocks on the diagonal should be square matrices, and invertible
- Interesting variant for systems of partial differential equations, where multiple species interact with each other

Convergence

Let $\hat{u}$ be the solution of $Au = b$.

$$
\begin{aligned}
u_{k+1} &= u_k - M^{-1}(Au_k - b) \\
&= (I - M^{-1}A)u_k + M^{-1}b \\
u_{k+1} - \hat{u} &= u_k - \hat{u} - M^{-1}(Au_k - A\hat{u}) \\
&= (I - M^{-1}A)(u_k - \hat{u}) \\
&= (I - M^{-1}A)^k(u_0 - \hat{u})
\end{aligned}
$$

So when does $(I - M^{-1}A)^k$ converge to zero for $k \to \infty$?

Jordan canonical form of a matrix A

- λ_i $(i = 1 \ldots p)$: eigenvalues of A
- $\sigma(A) = \{\lambda_1 \ldots \lambda_p\}$: spectrum of A
- μ_i: algebraic multiplicity of λ_i:
  multiplicity as zero of the characteristic polynomial $\det(A - \lambda I)$
- γ_i geometric multiplicity of λ_i: dimension of $\mathrm{K}er(A - \lambda I)$
- l_i: index of the eigenvalue: the smallest integer for which
  $\mathrm{K}er(A - \lambda I)^{l_i+1} = \mathrm{K}er(A - \lambda I)^{l_i})$
- $l_i \leq \mu_i$

Theorem (Saad, Th. 1.8) Matrix A can be transformed to a block diagonal matrix consisting of p diagonal blocks, each associated with a distinct eigenvalue λ_i.

- Each of these diagonal blocks has itself a block diagonal structure consisting of γ_i *Jordan blocks*
- Each of the Jordan blocks is an upper bidiagonal matrix of size not exceeding l_i with λ_i on the diagonal and 1 on the first upper diagonal.

Jordan canonical form of a matrix II

$$X^{-1}AX = J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_p \end{pmatrix}$$

$$J_i = \begin{pmatrix} J_{i,1} & & & \\ & J_{i,2} & & \\ & & \ddots & \\ & & & J_{i,\gamma_i} \end{pmatrix}$$

$$J_{i,k} = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

Each $J_{i,k}$ is of size l_i and corresponds to a different eigenvector of A.

Spectral radius and convergence

- $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$: spectral radius

Theorem (Saad, Th. 1.10) $\lim_{k \to \infty} A^k = 0 \Leftrightarrow \rho(A) < 1$.

Proof, $\Rightarrow$: Let u_i be a unit eigenvector associated with an eigenvalue λ_i. Then

$$
\begin{aligned}
A u_i &= \lambda_i u_i \\
A^2 u_i &= \lambda_i A_i u_i = \lambda^2 u_i \\
&\vdots \\
A^k u_i &= \lambda^k u_i \\
\text{therefore} \quad ||A^k u_i||_2 &= |\lambda^k| \\
\text{and} \quad \lim_{k \to \infty} |\lambda^k| &= 0
\end{aligned}
$$

so we must have $\rho(A) < 1$

Spectral radius and convergence II

Proof, $\Leftarrow$: Jordan form $X^{-1}AX = J$. Then $X^{-1}A^kX = J^k$.
Sufficient to regard Jordan block $J_i = \lambda_i I + E_i$ where $|\lambda_i| < 1$ and $E_i^{l_i} = 0$.
Let $k \geq l_i$. Then

$$J_i^k = \sum_{j=0}^{l_i-1} \binom{k}{j} \lambda^{k-j} E_i^j$$

$$||J_i||^k \leq \sum_{j=0}^{l_i-1} \binom{k}{j} |\lambda|^{k-j} ||E_i||^j$$

One has $\binom{k}{j} = \frac{k!}{j!(k-j)!} = \sum_{i=0}^{j} \begin{bmatrix} j \\ i \end{bmatrix} \frac{k^i}{j!}$ is a polynomial
where for $k > 0$, the Stirling numbers of the first kind are given by
$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = 1, \quad \begin{bmatrix} j \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ j \end{bmatrix} = 0, \quad \begin{bmatrix} j+1 \\ i \end{bmatrix} = j \begin{bmatrix} j \\ i \end{bmatrix} + \begin{bmatrix} j \\ i-1 \end{bmatrix}$.

Thus, $\binom{k}{j} |\lambda|^{k-j} \to 0$ $(k \to \infty)$.

Corollary from proof

Theorem (Saad, Th. 1.12)

$$\lim_{k \to \infty} ||A^k||^{\frac{1}{k}} = \rho(A)$$

Back to iterative methods

Sufficient condition for convergence: $\rho(I - M^{-1}A) < 1$.

Convergence rate

Assume λ with $|\lambda| = \rho(I - M^{-1}A)$ is the largest eigenvalue and has a single Jordan block. Then the convergence rate is dominated by this Jordan block, and therein by the term

$$\lambda^{k-p+1} \binom{k}{p-1} E^{p-1}$$

$$||(I - M^{-1}A)^k(u_0 - \hat{u})|| = O\left(|\lambda^{k-p+1}| \binom{k}{p-1}\right)$$

and the "worst case" convergence factor ρ equals the spectral radius:

$$\begin{aligned}
\rho &= \lim_{k\to\infty} \left(\max_{u_0} \frac{||(I - M^{-1}A)^k(u_0 - \hat{u})||}{||u_0 - \hat{u}||} \right)^{\frac{1}{k}} \\
&= \lim_{k\to\infty} ||(I - M^{-1}A)^k||^{\frac{1}{k}} \\
&= \rho(I - M^{-1}A)
\end{aligned}$$

Depending on u_0, the rate may be faster, though

Richardson iteration

$M = \frac{1}{\alpha}$, $I - M^{-1}A = I - \alpha A$. Assume for the eigenvalues of A:
$\lambda_{min} \leq \lambda_i \leq \lambda_{max}$.

Then for the eigenvalues μ_i of $I - \alpha A$ one has $1 - \alpha\lambda_{max} \leq \lambda_i \leq 1 - \alpha\lambda_{min}$.

If $\lambda_{min} < 0$ and $\lambda_{max} < 0$, at least one $\mu_i > 1$.

So, assume $\lambda_{min} > 0$. Then we must have

$1 - \alpha\lambda_{max} > -1, 1 - \alpha\lambda_{min} < 1 \Rightarrow$
$0 < \alpha < \frac{2}{\lambda_{max}}$.

$\rho = \max(|1 - \alpha\lambda_{max}|, |1 - \alpha\lambda_{min}|)$

$\alpha_{opt} = \frac{2}{\lambda_{min}+\lambda_{max}}$

$\rho_{opt} = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}$

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 $u_{k+1} = M^{-1}Nu_k + M^{-1}b$.

When does it converge ?

1.10 Nonnegative Matrices, M-Matrices

Nonnegative matrices play a crucial role in the theory of matrices. They are important in the study of convergence of iterative methods and arise in many applications including economics, queuing theory, and chemical engineering.

A *nonnegative matrix* is simply a matrix whose entries are nonnegative. More generally, a partial order relation can be defined on the set of matrices.

Definition 1.23 *Let A and B be two $n \times m$ matrices. Then*

$$A \leq B$$

if by definition, $a_{ij} \leq b_{ij}$ for $1 \leq i \leq n$, $1 \leq j \leq m$. If O denotes the $n \times m$ zero matrix, then A is nonnegative if $A \geq O$, and positive if $A > O$. Similar definitions hold in which "positive" is replaced by "negative".

The binary relation "$\leq$" imposes only a *partial* order on $\mathbb{R}^{n \times m}$ since two arbitrary matrices in $\mathbb{R}^{n \times m}$ are not necessarily comparable by this relation. For the remainder of this section, we now assume that only square matrices are involved. The next proposition lists a number of rather trivial properties regarding the partial order relation just defined.

Properties of $\leq$ for matrices

Proposition 1.24 *The following properties hold.*

1. *The relation $\leq$ for matrices is reflexive ($A \leq A$), antisymmetric (if $A \leq B$ and $B \leq A$, then $A = B$), and transitive (if $A \leq B$ and $B \leq C$, then $A \leq C$).*
2. *If A and B are nonnegative, then so is their product AB and their sum $A + B$.*
3. *If A is nonnegative, then so is A^k.*
4. *If $A \leq B$, then $A^T \leq B^T$.*
5. *If $O \leq A \leq B$, then $\|A\|_1 \leq \|B\|_1$ and similarly $\|A\|_\infty \leq \|B\|_\infty$.*

Irreducible matrices

A is *irreducible* if there is a permutation matrix P such that PAP^T is upper block triangular.

Perron-Frobenius Theorem

Theorem (Saad Th.1.25) Let A be a real $n \times n$ nonnegative irreducible martrix. Then:

- The spectral radius $\rho(A)$ is a simple eigenvalue of A.
- There exists an eigenvector u associated wit $\rho(A)$ which has positive elements