Scientific Computing WS 2018/2019

Lecture 27

Jürgen Fuhrmann

juergen.fuhrmann@wias-berlin.de

Recapitulation II: Finite Volumes

- Strong formulation of PDE
- Voronoi cells as control volumes
- Gauss theorem in control volumes
- Derivation of discrete system from fluxes between cells
- Matrix form
- Matrix element calculation
- Matrix properties
- Solution of matrix problem

Divergence theorem (Gauss' theorem)

Theorem: Let Ω be a bounded Lipschitz domain and $\mathbf{v} : \Omega \to \mathbb{R}^d$ be a continuously differentiable vector function. Let $\mathbf{n}$ be the outward normal to Ω. Then,

$$\int_\Omega \nabla \cdot \mathbf{v} \, d\mathbf{x} = \int_{\partial\Omega} \mathbf{v} \cdot \mathbf{n} \, ds$$

$\square$

Species balance over an REV

- Let $u(\mathbf{x}, t) : \Omega \times [0, T] \to \mathbb{R}$ be the local amount of some species.
- Assume *representative elementary volume (REV)* $\omega \subset \Omega$
- Subinterval in time $(t_0, t1) \subset (0, T)$
- $-\delta \nabla u \cdot \mathbf{n}$ describes the flux of these species trough $\partial\omega$, where δ is some transfer coefficient
- Let $f(\mathbf{x}, t)$ be some local source of species. Then the flux through the boundary is balanced by the change of the amount of species in ω and the source strength:

$$0 = \int_\omega (u(\mathbf{x}, t_1) - u(\mathbf{x}, t_0))\, d\mathbf{x} - \int_{t_0}^{t_1} \int_{\partial\omega} \delta \nabla u \cdot \mathbf{n}\, ds\, dt - \int_{t_0}^{t_1} \int_\omega f(\mathbf{x}, t)\, ds$$

- Using Gauss' theorem, rewrite this as

$$0 = \int_{t_0}^{t_1} \int_\omega \partial_t u(\mathbf{x}, t)\, d\mathbf{x}\, dt - \int_{t_0}^{t_1} \int_\omega \nabla \cdot (\delta \nabla u)\, d\mathbf{x}\, dt - \int_{t_0}^{t_1} \int_\omega f(\mathbf{x}, t)\, ds$$

- True for all $\omega \subset \Omega$, $(t_0, t1) \subset (0, T) \Rightarrow$ parabolic second order PDE

$$\partial_t u(x, t) - \nabla \cdot (\delta \nabla u(x, t)) = f(x, t)$$

Second order elliptic PDEs

Stationary case: $\partial_t u = 0 \Rightarrow$ second order *elliptic* PDE

$$-\nabla \cdot (\delta \nabla u(x)) = f(x)$$

- Stationary heat conduction, stationary diffusion
- Incompressible flow in saturated porous media: u: pressure
  $\delta = k$: permeability, flux$=-k\nabla u$: "Darcy's law"
- Electrical conduction: u: electric potential
  $\delta = \sigma$: electric conductivity
  flux$=-\sigma\nabla u \equiv$ current density: "Ohms's law"
- Poisson equation (electrostatics in a constant magnetic field):
  u: electrostatic potential, ∇u: electric field,
  $\delta = \varepsilon$: dielectric permittivity, f: charge density

Second order PDEs: boundary conditions

- Combine PDE in the interior with boundary conditions on variable u and/or or normal flux $\delta\nabla u \cdot \mathbf{n}$
- Assume $\partial\Omega = \cup_{i=1}^{N_\Gamma}\Gamma_i$ is the union of a finite number of non-intersecting subsets Γ_i which are locally Lipschitz.
- On each Γ_i, specify one of
  - Dirichlet ("first kind"): let $g_i : \Gamma_i \to \mathbb{R}$ (homogeneous for $g_i = 0$)

    $$u(x) = u_{\Gamma_i}(x) \quad \text{for } x \in \Gamma_i$$

  - Neumann ("second kind"): Let $g_i : \Gamma_i \to \mathbb{R}$ (homogeneus for $g_i = 0$)

    $$\delta\nabla u(x) \cdot \mathbf{n} = g_i(x) \quad \text{for } x \in \Gamma_i$$

  - Robin ("third kind"): let $\alpha_i, g_i : \Gamma_i \to \mathbb{R}$

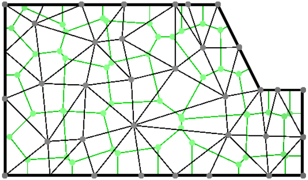
    $$\delta\nabla u(x) \cdot \mathbf{n} + \alpha_i(x)\left(u(x) - g_i(x)\right) = 0 \quad \text{for } x \in \Gamma_i$$

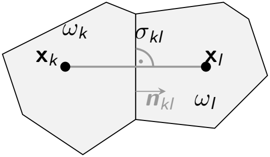
- Boundary functions may be time dependent.

Constructing control volumes I

- Assume Ω is a polygon
- Subdivide the domain Ω into a finite number of **control volumes** : $\bar{\Omega} = \bigcup_{k \in \mathcal{N}} \bar{\omega}_k$ such that
  - ω_k are open (not containing their boundary) convex domains
  - $\omega_k \cap \omega_l = \emptyset$ if $\omega_k \neq \omega_l$
  - $\sigma_{kl} = \bar{\omega}_k \cap \bar{\omega}_l$ are either empty, points or straight lines
  - we will write $|\sigma_{kl}|$ for the length
  - if $|\sigma_{kl}| > 0$ we say that ω_k, ω_l are neighbours
  - neighbours of ω_k: $\mathcal{N}_k = \{l \in \mathcal{N} : |\sigma_{kl}| > 0\}$
- To each control volume ω_k assign a **collocation point**: $\mathbf{x}_k \in \bar{\omega}_k$ such that
  - **admissibility condition**:
    if $l \in \mathcal{N}_k$ then the line $\mathbf{x}_k\mathbf{x}_l$ is orthogonal to σ_{kl}
  - **placement of boundary unknowns**:
    if ω_k is situated at the boundary, i.e. for $|\partial\omega_k \cap \partial\Omega| > 0$, then $\mathbf{x}_k \in \partial\Omega$, and $\partial\omega_k \cap \partial\Omega = \cup_{i=1}^{N_\Gamma} \gamma_{i,k}$ (where $\gamma_{i,k} = \emptyset$ is possible).

Constructing control volumes II





We know how to construct such a partitioning:

- obtain a boundary conforming Delaunay triangulation with vertices x_k
- construct restricted Voronoi cells ω_k with $x_k \in \omega_k$
- Delaunay triangulation gives connected neigborhood graph of Voronoi cells
- Admissibility condition fulfilled in a natural way
- Boundary placement of triangle nodes

Voronoi diagrams

After G. F. Voronoi, 1868-1908

Definition Let $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$. The set of points $H_{\mathbf{pq}} = \left\{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x} - \mathbf{p}|| \leq ||\mathbf{x} - \mathbf{q}||\right\}$ is the *half space* of points $\mathbf{x}$ closer to $\mathbf{p}$ than to $\mathbf{q}$.

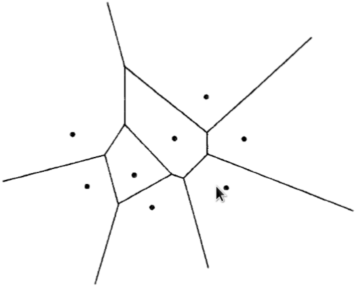
Definition Given a finite set of points $S \subset \mathbb{R}^d$, the *Voronoi region (Voronoi cell)* of a point $\mathbf{p} \in S$ is the set of points $\mathbf{x}$ closer to $\mathbf{p}$ than to any other point $\mathbf{q} \in S$:

$$V_{\mathbf{p}} = \left\{\mathbf{x} \in \mathbb{R}^d : ||\mathbf{x} - \mathbf{p}|| \leq ||\mathbf{x} - \mathbf{q}|| \,\forall \mathbf{q} \in S\right\}$$

The *Voronoi diagram* of S is the collection of the Voronoi regions of the points of S.

Voronoi diagrams II

- The Voronoi diagram subdivides the whole space into "nearest neigbor" regions
- Being intersections of half planes, the Voronoi regions are convex sets



Voronoi diagram of 8 points in the plane

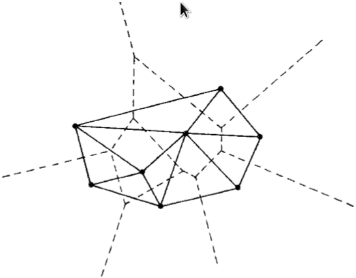
(H. Si)

Interactive example: `http://homepages.loria.fr/BLevy/GEOGRAM/geogram_demo_Delaunay2d.html`

Delaunay triangulation

After B.N. Delaunay (Delone), 1890-1980

- Assume that the points of S are in *general position*, i.e. no $d+2$ points of S are on one sphere (in 2D: no 4 points on one circle)
- Connect each pair of points whose Voronoi regions share a common edge with a line
- $\Rightarrow$ *Delaunay triangulation* of the convex hull of S



Delaunay triangulation of the convex hull of 8 points in the plane

(H. Si)

Delaunay triangulation II

- The circumsphere (circumcircle in 2D) of a d-dimensional simplex is the unique sphere containing all vertices of the simplex
- The circumball (circumdisc in 2D) of a simplex is the unique (open) ball which has the circumsphere of the simplex as boundary

Definition A triangulation of the convex hull of a point set S has the *Delaunay property* if each simplex (triangle) of the triangulation is Delaunay, i.e. its circumsphere (circumcircle) is empty wrt. S, i.e. it does not contain any points of S.

- The Delaunay triangulation of a point set S, where all points are in general position is unique
- Otherwise there is an ambiguity - if e.g. 4 points are one circle, there are two ways to connect them resulting in Delaunay triangles

Edge flips and locally Delaunay edges (2D only)

- For any two triangles **abc** and **adb** sharing a common edge **ab**, there is the *edge flip* operation which reconnects the points in such a way that two new triangles emerge: **adc** and **cdb**.
- An edge of a triangulation is locally Delaunay if it either belongs to exactly one triangle, or if it belongs to two triangles, and their respective circumdisks do not contain the points opposite wrt. the edge
- If an edge is locally Delaunay and belongs to two triangles, the sum of the angles opposite to this edge is less or equal to π.
- If all edges of a triangulation of the convex hull of S are locally Delaunay, then the triangulation is the Delaunay triangulation
- If an edge is not locally Delaunay and belongs to two triangles, the edge emerging from the corresponding edge flip will be locally Delaunay

Edge flip algorithm (Lawson)

```
Input: A stack L of edges of a given triangulation of S;
while L ≠ ∅ do
    pop an edge ab from L;
    if ab is not locally Delaunay then
        flip ab to cd;
        push edges ac, cb, db, da onto L;
    end
end
```

- This algorithm is known to terminate. After termination, all edges will be locally Delaunay, so the output is the Delaunay triangulation of S.
- Among all triangulations of a finite point set S, the Delaunay triangulation maximises the minimum angle
- All triangulations of S are connected via a flip graph

Radomized incremental flip algorithm (2D only)

- Create Delaunay triangulation of point set S by inserting points one after another, and creating the Delaunay triangulation of the emerging subset of S using the flip algorithm
- Estimated complexity: $O(n \log n)$
- In 3D, there is no simple flip algorithm, generalizations are active research subject

Triangulations of finite domains

- So far, we discussed triangulations of point sets, but in practice, we need triangulations of domains
- Create Delaunay triangulation of point set, "Intersect" with domain

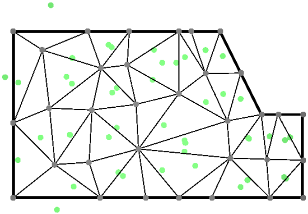
Boundary conforming Delaunay triangulations

Definition: An admissible triangulation of a polygonal Domain $\Omega \subset \mathbb{R}^d$ has the boundary conforming Delaunay property if

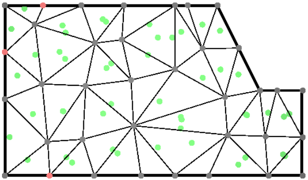
(i) All simplices are Delaunay

(ii) All boundary simplices (edges in 2D, facets in 3d) have the Gabriel property, i.e. their minimal circumdisks are empty

- Equivalent definition in 2D: sum of angles opposite to interior edges $\leq \pi$, angle opposite to boundary edge $\leq \frac{\pi}{2}$
- Creation of boundary conforming Delaunay triangulation description may involve insertion of Steiner points at the boundary



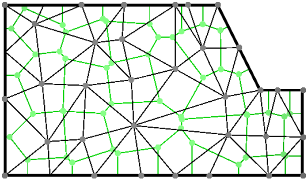
Delaunay grid of Ω



Boundary conforming Delaunay grid of Ω

Domain blendend Voronoi cells

- For Boundary conforming Delaunay triangulations, the intersection of the Voronoi diagram with the domain yields a well defined dual subdivision



Boundary conforming Delaunay triangulations II

- Weakly acute triangulations are boundary conforming Delaunay, but not vice versa!
- Working with weakly acute triangulations for general polygonal domains is unrealistic, especially in 3D
- For boundary conforming Delaunay triangulations of polygonal domains there are algoritms with mathematical termination proofs valid in many relevant cases
- Code examples:
  - 2D: Triangle by J.R.Shewchuk `https://www.cs.cmu.edu/~quake/triangle.html`
  - 3D: TetGen by H. Si `http://tetgen.org`
- Features:
  - polygonal geometry description
  - automatic insertion of points according to given mesh size criteria
  - accounting for interior boundaries
  - local mesh size control for a priori refinement
  - quality control
  - standalone executable & library

Discretization ansatz for Robin boundary value problem

Given constants $\kappa > 0$, $\alpha_i \geq 0$ $(i = 1 \dots N_\Gamma)$

$$-\nabla \cdot \kappa \nabla u = f \text{ in } \Omega$$
$$\kappa \nabla u \cdot \mathbf{n} + \alpha_i (u - g_i) = 0 \text{ on } \Gamma_i \ (i = 1 \dots N_\Gamma) \qquad (*)$$

- Given control volume ω_k, $k \in \mathcal{N}$, integrate

$$\begin{aligned}
0 &= \int_{\omega_k} (-\nabla \cdot \kappa \nabla u - f)\, d\omega \\
&= -\int_{\partial\omega_k} \kappa \nabla u \cdot \mathbf{n}_k d\gamma - \int_{\omega_k} f d\omega \qquad \text{(Gauss)} \\
&= -\sum_{l \in \mathcal{N}_k} \int_{\sigma_{kl}} \kappa \nabla u \cdot \mathbf{n}_{kl} d\gamma - \sum_{i=1}^{N_\Gamma} \int_{\gamma_{ik}} \kappa \nabla u \cdot \mathbf{n} d\gamma - \int_{\omega_k} f d\omega \\
&\approx \sum_{L \in \mathcal{N}_k} \underbrace{\kappa \frac{|\sigma_{kl}|}{h_{kl}} (u_k - u_l)}_{\nabla u \cdot \mathbf{n} \approx \frac{u_l - u_k}{h_{kl}}} + \sum_{i=1}^{N_\Gamma} \underbrace{|\gamma_{i,k}| \alpha_i (u_k - g_{i,k})}_{\text{bound. cond. } (*)} - \underbrace{|\omega_k| f_k}_{\text{quadrature}}
\end{aligned}$$

- Here, $u_k = u(\mathbf{x}_k)$, $g_{i,k} = g_i(\mathbf{x}_k)$, $f_k = f(\mathbf{x}_k)$

Properties of discretization matrix

- $N = |\mathcal{N}|$ equations (one for each control volume ω_k)
- $N = |\mathcal{N}|$ unknowns (one for each collocation point $x_k \in \omega_k$)
- weighted connected edge graph of triangulation $\equiv N \times N$ irreducible sparse discretization matrix $A = (a_{kl})$:

$$a_{kl} = \begin{cases} \sum_{l' \in \mathcal{N}_k} \kappa \frac{|\sigma_{kl'}|}{h_{kl'}} + \sum_{i=1}^{N_\Gamma} |\gamma_{i,k}| \alpha_i, & l = k \\ -\kappa \frac{\sigma_{kl}}{h_{kl}}, & l \in \mathcal{N}_k \\ 0, & else \end{cases}$$

- A is irreducibly diagonally dominant if at least for one i, $|\gamma_{i,k}|\alpha_i > 0$
- Main diagonal entries are positive, off diagonal entries are non-positive
- $\Rightarrow A$ has the M-property.
- A is symmetric $\Rightarrow A$ is positive definite

Matrix assembly – main part

- Keep list of global node numbers per triangle τ mapping local node numbers of the triangle to the global node numbers: $\{0,1,2\} \to \{k_{\tau,0}, k_{\tau,1}, k_{\tau,2}\}$
- Loop over all triangles $\tau \in \mathcal{T}$, add up contributions

for $k, l = 1 \dots N$ **do**
  set $a_{kl} = 0$
end
for $\tau \in \mathcal{T}$ **do**
  for $n, m = 0 \dots 2, n \neq m$ **do**

$$\sigma = \sigma_{k_{\tau,m},k_{\tau,n}} \cap \tau$$

$$\sigma_h = \kappa \frac{|\sigma|}{h_{k_{\tau,m},k_{\tau,n}}}$$

$$a_{k_{\tau,m},k_{\tau,m}} += \sigma_h$$

$$a_{k_{\tau,m},k_{\tau,n}} -= \sigma_h$$

$$a_{k_{\tau,n},k_{\tau,m}} -= \sigma_h$$

$$a_{k_{\tau,n},k_{\tau,n}} += \sigma_h$$

  end
end

Matrix assembly – boundary part

- Keep list of global node numbers per boundary element γ mapping local node element to the global node numbers: $\{0,1\} \to \{k_{\gamma,0}, k_{\gamma,1}\}$
- Keep list of boundary part numbers per boundary element i_γ
- Loop over all boundary elements $\gamma \in \mathcal{G}$ of the discretization, add up contributions

for $\gamma \in \mathcal{G}$ **do**
  for $n = 0, 1$ **do**
    $a_{k_{\gamma n}, k_{\gamma n}} += \alpha_{i_\gamma} |\gamma \cap \omega_{k_{\gamma n}}|$
  end
end

RHS assembly: calculate control volumes

- Denote $w_k = |\omega_k|$
- Loop over triangles, add up contributions

for $k \dots N$ **do**
  set $w_k = 0$
end
for $\tau \in \mathcal{T}$ **do**
  for $n = \dots 3$ **do**
    $w_k + = |\omega_{k_{\tau,m}} \cap \tau|$
  end
end

Matrix assembly: summary

- Sufficient to keep list of triangles, boundary segments – they typically come out of the mesh generator
- Be able to calculate triangular contributions to form factors: $|\omega_k \cap \tau|$, $|\sigma_{kl} \cap \tau|$ – we need only the numbers, and not the construction of the geometrical objects
- $O(N)$ operation, one loop over triangles, one loop over boundary elements

Variations of the discretization ansatz

- 3D: tetrahedron based
- $\kappa = \kappa(x) \Rightarrow \kappa(x)\nabla u \approx \kappa_{kl}\frac{u_l - u_k}{h_{kl}}$
- Non-constant α_i, g
- Nonlinear dependencies ...

Interpretation of results

- One solution value per control volume ω_k allocated to the collocation point $x_k \Rightarrow$ piecewise constant function on collection of control volumes
- But: x_k are at the same time nodes of the corresponding Delaunay mesh $\Rightarrow$ representation as piecewise linear function on triangles

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
- Preconditioner: $M = D$, where D is the main diagonal of $A \Rightarrow$

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

- Equivalent to the succesive (row by row) solution of

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

- Already calculated results not taken into account
- Alternative formulation with $A = M - N$:

$$\begin{aligned} u_{k+1} &= D^{-1}(E+F)u_k + D^{-1}b \\ &= M^{-1}Nu_k + M^{-1}b \end{aligned}$$

- Variable ordering does not matter

The Gauss-Seidel method

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

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

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

- May be it is faster
- Variable order probably matters
- Preconditioners: forward $M = D - E$, backward: $M = D - F$
- Splitting formulation: $A = M - N$
  forward: $N = F$, backward: $M = E$
- Forward case:

$$\begin{aligned} u_{k+1} &= (D - E)^{-1}Fu_k + (D - E)^{-1}b \\ &= M^{-1}Nu_k + M^{-1}b \end{aligned}$$

Convergence

- Let $\hat{u}$ be the solution of $Au = b$.
- Let $e_k = u_j - \hat{u}$ be the error of the k-th iteration step

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

resulting in

$$e_{k+1} = (I - M^{-1}A)^k e_0$$

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

Spectral radius and convergence

Definition The spectral radius $\rho(A)$ is the largest absolute value of any eigenvalue of A: $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$.

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

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

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

Back to iterative methods

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

Convergence rate

Assume λ with $|\lambda| = \rho(I - M^{-1}A) < 1$ is the largest eigenvalue and has a single Jordan block of size l. Then the convergence rate is dominated by this Jordan block, and therein by the term with the lowest possible power in λ which due to $E^l = 0$ is

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

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

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

The Gershgorin Circle Theorem (Semyon Gershgorin,1931)

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

Theorem (Varga, Th. 1.11) Let A be an $n \times n$ (real or 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 exists r, $1 \leq r \leq n$ such that

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

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

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

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, $\sigma(A) = \sigma(A^T)$. □

Gershgorin circles: heat example I

$$A = \begin{pmatrix} \frac{2}{h} & -\frac{1}{h} & & & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & & & -\frac{1}{h} & \frac{2}{h} & \end{pmatrix}$$

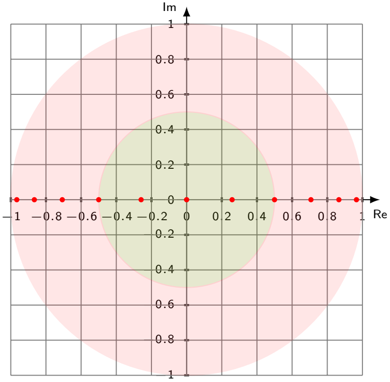
$$B = (I - D^{-1}A) = \begin{pmatrix} 0 & \frac{1}{2} & & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & & \frac{1}{2} & 0 \end{pmatrix}$$

We have $b_{ii} = 0$, $\Lambda_i = \begin{cases} \frac{1}{2}, & i = 1, n \\ 1 & i = 2 \dots n-1 \end{cases} \Rightarrow$ estimate $|\lambda_i| \leq 1$

Gershgorin circles: heat example II

Let n=11, h=0.1:

$$\lambda_i = \cos\left(\frac{ih\pi}{1+2h}\right) \quad (i = 1 \ldots n)$$



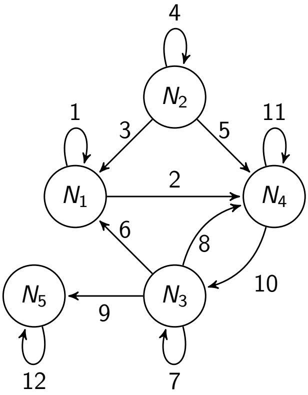
⇒ the Gershgorin circle theorem is too pessimistic...

Weighted directed graph representation of matrices

Define a directed graph from the nonzero entries of a matrix $A = (a_{ik})$:

- Nodes: $\mathcal{N} = \{N_i\}_{i=1\dots n}$
- Directed edges: $\mathcal{E} = \{\overrightarrow{N_k N_l} | a_{kl} \neq 0\}$
- Matrix entries $\equiv$ weights of directed edges

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



- 1:1 equivalence between matrices and weighted directed graphs
- Convenient e.g. for sparse matrices

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.

Theorem (Varga, Th. 1.17): 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 consecutive nonzero matrix entries $a_{ik_1}, a_{k_1 k_2}, a_{k_2 k_3} \dots, a_{k_{r-1} k_r} a_{k_r j}$.

□

Taussky theorem (Olga Taussky, 1948)

Theorem (Varga, Th. 1.18) Let A be irreducible. Assume that the eigenvalue λ is a boundary point of the union of all the disks

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

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

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

Taussky theorem proof

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

$$(\lambda - a_{rr})x_r = \sum_{\substack{j=1\dots n\\ j\neq r}} a_{rj}x_j \qquad (1)$$

$$|\lambda - a_{rr}| \leq \sum_{\substack{j=1\dots n\\ j\neq r}} |a_{rj}|\cdot|x_j| \leq \sum_{\substack{j=1\dots n\\ j\neq r}} |a_{rj}| = \Lambda_r \qquad (2)$$

λ is boundary point $\Rightarrow |\lambda - a_{rr}| = \sum_{\substack{j=1\dots n\\ j\neq r}} |a_{rj}|\cdot|x_j| = \Lambda_r$

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

Due to irreducibility there is at least one p with $a_{rp} \neq 0$. For this p, $|x_p| = 1$ and equation (2) is valid (with p in place of r) $\Rightarrow |\lambda - a_{pp}| = \Lambda_p$

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

Consequences for heat example from Taussky theorem

- $B = I - D^{-1}A$
- We had $b_{ii} = 0$, $\Lambda_i = \begin{cases} \frac{1}{2}, & i = 1, n \\ 1 & i = 2 \dots n-1 \end{cases}$ $\Rightarrow$ estimate $|\lambda_i| \leq 1$
- Assume $|\lambda_i| = 1$. Then λ_i lies on the boundary of the union of the Gershgorin circles. But then it must lie on the boundary of both circles with radius $\frac{1}{2}$ and 1 around 0.
- Contradiction $\Rightarrow |\lambda_i| < 1$, $\rho(B) < 1$!

Diagonally dominant matrices

Definition Let $A = (a_{ij})$ be an $n \times n$ matrix.

- A is *diagonally dominant* if
  (i) 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
  (i) 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
  (i) A is irreducible
  (ii) A is diagonally dominant –
  for $i = 1 \dots n$, $|a_{ii}| \geq \sum_{\substack{j=1\dots n \\ j \neq i}} |a_{ij}|$
  (iii) 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 (Varga, Th. 1.21): Let A be strictly diagonally dominant or irreducibly diagonally dominant. Then A is nonsingular.

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

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

Corollary

Theorem: If A is complex hermitian or real 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.

□

Perron-Frobenius Theorem (1912/1907)

Definition: A real n-vector $\mathbf{x}$ is

- positive ($\mathbf{x} > 0$) if all entries of $\mathbf{x}$ are positive
- nonnegative ($\mathbf{x} \geq 0$) if all entries of $\mathbf{x}$ are nonnegative

Definition: A real $n \times n$ matrix A is

- positive ($A > 0$) if all entries of A are positive
- nonnegative ($A \geq 0$) if all entries of A are nonnegative

Theorem(Varga, Th. 2.7) Let $A \geq 0$ be an irreducible $n \times n$ matrix. Then

(i) A has a positive real eigenvalue equal to its spectral radius $\rho(A)$.
(ii) To $\rho(A)$ there corresponds a positive eigenvector $\mathbf{x} > 0$.
(iii) $\rho(A)$ increases when any entry of A increases.
(iv) $\rho(A)$ is a simple eigenvalue of A.

Proof: See Varga. □

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 (for general matrices), $\rho(G)$ is an eigenvalue with a nonnegative eigenvector $\mathbf{x}$. Thus,

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

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

Convergence rate comparison

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: $\tau_2 = \rho(A^{-1}N_2) \geq \rho(A^{-1}N_1) = \tau_1$

But $\frac{\tau}{1+\tau}$ is strictly increasing. $\square$

M-Matrix definition

Definition Let A be an $n \times n$ real matrix. A is called M-Matrix if

(i) $a_{ij} \leq 0$ for $i \neq j$

(ii) A is nonsingular

(iii) $A^{-1} \geq 0$

Corollary: If A is an M-Matrix, then $A^{-1} > 0 \Leftrightarrow A$ is irreducible.

Proof: See Varga. □

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.

Proof: We know that A is nonsingular, but we have to show $A^{-1} \geq 0$.

- Let $B = I - D^{-1}A$. Then $\rho(B) < 1$, therefore $I - B$ is nonsingular.
- We have for $k > 0$:

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

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 $A^{-1} \geq 0$. □

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.
- More general: Block Jacobi, Block Gauss Seidel etc.

Examinations

Tue Feb 26.
Wed Feb 27.
Wed Mar 14.
Thu Mar 15.
Tue Mar 26.
Wed Mar 27.
Thu Mar 28.
Wed May 8. 14:00-17:00

- Please give your yellow sheets before the examinations to Frau Gillmeister (MA370)