4. Linear systems: iterative methods

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

numerical methods for solving this so far:

- Finite Differences
- Finite Elements

$\rightsquigarrow$ linear system $\quad Ax = b \quad , \; A \in \mathbb{R}^{n\times n}, \; b \in \mathbb{R}^n$

where A is sparse (and potentially large)

How do we solve this linear system?

- Gaußian elimination / LR-decomposition

disadvantage: (i) time: $\mathcal{O}(n^3)$ „slow"

(ii) memory: the factors $LR = A$ are not necessarily sparse again

Is there any way to solve large linear systems fast and „sparsely" ?

Observation: matrix-product Av ($v \in \mathbb{R}^n$) has $\mathcal{O}(n)$ complexity.

Idea: try to construct a method which constructs a sequence of approximate solutions $(x_n)_n$ (hopefully converging to x) and which only uses matrix-vector products Av, i.e. an iterative method for solving $Ax = b$.

4.1 Classical schemes

(4.1) Idea / basic ansatz:

$$Ax = b \iff F(x) := Ax - b = 0$$

aim here: construct a „one-step scheme" $x_{n+1} = \varphi(x_n)$

$$\begin{array}{ccc} x_{n+1} & = & \varphi(x_n) \\ \downarrow & & \downarrow \\ x & = & \varphi(x) \end{array} \qquad \varphi \text{ continuous}$$

i.e. the solution x that we are looking for is a fixed point of the „one-step scheme" φ.

Thus, we have to find a function (the „one-step scheme") φ such that the solution x of $Ax=b$ is a fixed point of φ.

One idea is:

$$A = M - N \qquad , M \text{ regular}$$

„splitting of A"

$$\text{Then } Ax = b \iff (M-N)x = b$$
$$\iff Mx = Nx + b$$
$$\iff x = \underbrace{M^{-1}N}x + \underbrace{M^{-1}}b$$
$$x = Cx + d \quad =: \varphi(x)$$

fixed point equation

(4.2) General convergence criterion

E.g. by Banach's fixed point theorem (if φ is a contraction:

$$\|\varphi(x) - \varphi(y)\| \le K \|x-y\|, \qquad 0 \le K < 1 \;)$$

Associated local criterion if φ is continously differentiable:

If $\bar{x} = \varphi(\bar{x})$ and $\rho(D\varphi(\bar{x})) < 1$, then the iteration with φ locally converges to $\bar{x}$.

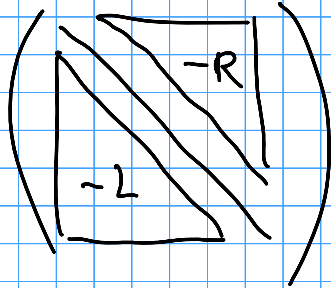
Here: $\varphi(x) = Cx + d$

$$D\varphi(x) = C$$

Thus, the fixed point iteration converges if $\rho(C) < 1$ (even globally!)

(4.3) The Jacobi method

splitting: $\underset{\substack{\| \\ M}}{A = D} - \underbrace{(L+R)}_{\substack{\| \\ N}}$, D regular

with $D = \operatorname{diag}(A)$, i.e. $A =$ 

Explicitly: $x^{(n+1)} = M^{-1}Nx^{(m)} + M^{-1}b$

$$= \underbrace{D^{-1}(L+R)}_{=C}x^{(m)} + D^{-1}b$$

Here, $M = D$ is easily invertible ($\mathcal{O}(n)$ flop).

(4.4) The Gauß-Seidel method

splitting: $A = \underbrace{(D-L)}_{=M} - \underbrace{R}_{=N}$

explicitely: $x^{(m+1)} = \underbrace{(D-L)^{-1} R}_{=C} x^{(m)} + (D-L)^{-1} b$

(4.5) Diagonally dominant matrices

A matrix $A \in \mathbb{R}^{n \times n}$ is strictly diagonally dominant if

$$\sum_{i \neq j} |a_{ij}| < |a_{ii}| \quad , \quad i = 1, \dots, n$$

A is called diagonally dominant, if

$$\sum_{i \neq j} |a_{ij}| \leq |a_{ii}| \quad , \quad i = 1, \dots, n$$

and the inequality is strict for at least one $i \in \{1, \dots, n\}$.

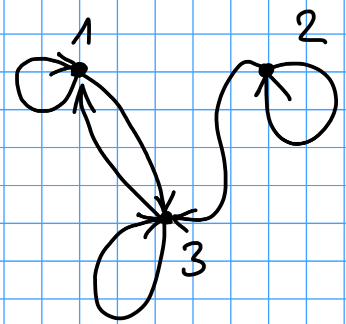
(4.6) Irreducibility

To $A \in \mathbb{R}^{n \times n}$ we associate a graph (adjacency graph, dependency graph) $G = (V, E)$ by

- $V = \{1, \dots, n\}$
- $E = \{ (i,j) \in V \times V \mid a_{ij} \neq 0 \}$

Example:

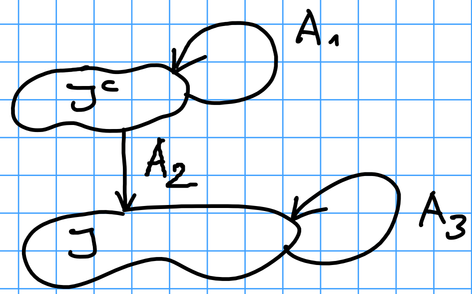
$$A = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 4 & -1 \\ -1 & 0 & 4 \end{bmatrix} \longrightarrow$$



Definition: A matrix $A \in \mathbb{R}^{n \times n}$ is called irreducible if its adjacency graph is strongly connected (i.e. for any two nodes i and j in G there is a path connecting i to j).

Lemma: If $A \in \mathbb{R}^{n\times n}$ is irreducible then there is no true subset $J \subset \{1,\dots,n\}$ such that $a_{ij} = 0$ for $i \in J,\ j \notin J$, i.e. there is no permutation of the rows and columns such that

$$A = \begin{array}{cc} \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} & \begin{array}{c} J^c \\ J \end{array} \\ \begin{array}{cc} J^c \quad & \quad J \end{array} & \end{array}$$



Proof: [ex.]

(4.7) Convergence of the Jacobi and the Gauß-Seidel method

THEOREM: (a) If A is strictly diagonally dominant then both, the Jacobi and the Gauß-Seidel method converge with

$$\rho(C) \leq \max_i \frac{1}{|a_{ii}|} \sum_{i \neq j} |a_{ij}| < 1$$

(b) If A is diagonally dominant and irreducible, then both methods converge as well.

Proof: (1) Jacobi: $\quad C = D^{-1}(L+R)$, i.e.

$$\left|(Cx)_i\right| = \left|\left(D^{-1}(L+R)x\right)_i\right| = \left|\frac{1}{a_{ii}}\sum_{i\neq j} -a_{ij}x_j\right| \leq \frac{1}{|a_{ii}|}\sum_{i\neq j}|a_{ij}||x_j|$$

$$\leq \underbrace{\left(\frac{1}{|a_{ii}|}\sum_{i\neq j}|a_{ij}|\right)}_{\max_i \, \bullet \; =: K}\|x\|_\infty$$

$$\Rightarrow \|C\|_\infty \leq K$$

- Thus $\rho(C) \leq K$ and if A is strictly diagonally dominant, then $K<1$ and so the iteration converges.

- Let A be diagonally dominant and irreducible. Assume that $\rho(C) = 1$. Then there is some eigenvalue λ of C with $|\lambda| = 1$. Let $x \in \mathbb{C}^n$ be an associated eigenvector, i.e. $Cx = \lambda x$, with $\|x\|_\infty = 1$.

 Define $J = \{1 \leq i \leq n : |x_i| = 1\}$ and we obtain for $i \in J$

$$1 = |x_i| = \left|(Cx)_i\right| \leq \frac{1}{|a_{ii}|}\sum_{i=j}|a_{ij}||x_j| \leq 1$$

$\Rightarrow$ $a_{ij} = 0$ for $j \notin J$, since otherwise the last inequality would be strict.

Since A is irreducible, this means that $J = \{1, .., n\}$.

On the other hand (by definition of „diagonally dominant") there is one row i such that

$$|a_{ii}| > \sum_{i \neq j} |a_{ij}|$$

$\Rightarrow$ $|x_i| = |(Cx)_i| < 1$ $\Longrightarrow$ $i \notin J$. $\lightning$ contradiction.

(2) Gauß-Seidel: $C = (D-L)^{-1} R$

$$|(Cx)_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j<i} |a_{ij}| |(Cx)_j| + \sum_{j>i} |a_{ij}| |(Cx)_j| \right)$$

$\downarrow$ induction on i

$$\leq \frac{1}{|a_{ii}|} \sum_{i \neq j} |a_{ij}| |x_j| \quad \rightsquigarrow \quad (1)$$

$\blacksquare$

(4.8) Relaxation / Damping: Idea

instead of using $x_{k+1} = \varphi(x_k)$ „directly", use a convex combination of $\varphi(x_k)$ and x_k :

$$x_{k+1} := \omega\varphi(x_k) + (1-\omega)x_k$$

$\omega \in [0,1]$: damping parameter

$\rightsquigarrow$ family of iteration schemes:

$$\varphi_\omega(x) = \omega\varphi(x) + (1-\omega)x \qquad \text{relaxed scheme}$$

Evidently, in our case of $\varphi(x) = Cx+d$, our goal in choosing ω is to minimize $\rho(D\varphi_\omega(x))$.

(4.9) Choice of the optimal damping parameter

With $\varphi(x) = Cx+d$ we obtain

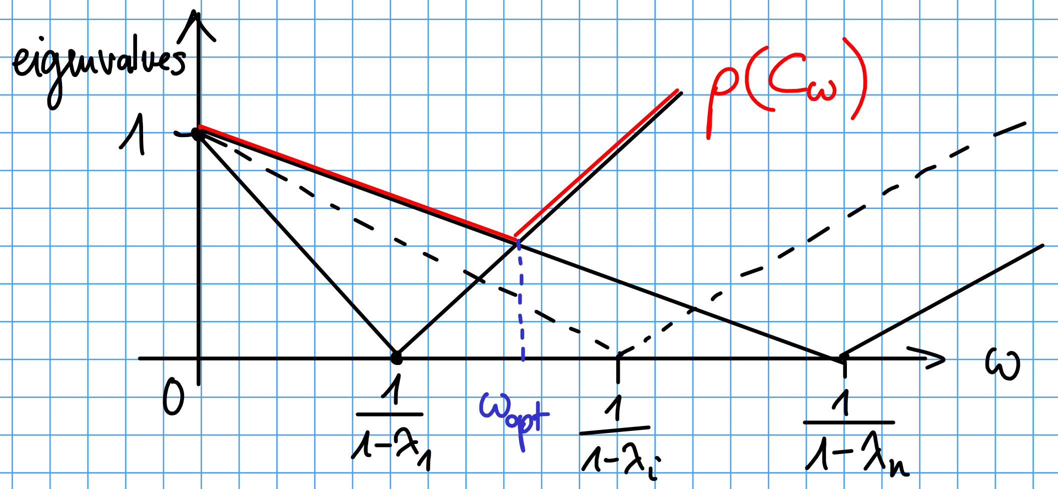
$$\varphi_\omega(x) = \omega(Cx+d) + (1-\omega)x = \underbrace{\left[\omega C + (1-\omega)I\right]}_{=C_\omega} x + \omega d$$

For simplicity assume that $\sigma(C) \subset (-\infty, 1)$:

eigenvalues of C : $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$

$\Rightarrow$ $\tilde{\lambda}_i = \omega \lambda_i + (1-\omega) = 1 - \omega(1-\lambda_i)$ eigenvalues of C_ω

$\Rightarrow$ $\rho(C_\omega) = \max_i \underline{|1-\omega(1-\lambda_i)|}$



$1-\omega(1-\lambda_i) \overset{!}{=} 0$

$\Leftrightarrow$ $\omega = \frac{1}{1-\lambda_i} \in (0, \infty)$

For the optimal $\omega = \omega_{opt}$ we have that $-(1-\omega_{opt}(1-\lambda_1)) = 1 - \omega_{opt}(1-\lambda_n)$

$\Rightarrow$ $\omega_{opt} = \frac{2}{2-\lambda_1-\lambda_n}$

(4.10) The SOR method („successive overrelaxation")

relaxed Gauß-Seidel method:

$$x_{k+1} = \omega D^{-1}(b + Lx_{k+1} + Rx_k) + (1-\omega)x_k$$

$$\rightsquigarrow \quad x_{k+1} = \underbrace{(I-\omega D^{-1}L)^{-1}((1-\omega)I + \omega D^{-1}R)}_{=C_\omega} x_k + \omega(I-\omega D^{-1}L)^{-1}D^{-1}b$$

THEOREM: (1) $\rho(C_\omega) \geq |1-\omega|$, i.e. the method can only converge for $\omega \in (0,2)$.

(2) If A is symmetric and positive definite, then this method converges for $\omega \in (0,2)$.

Remark: For $\omega \in (1,2)$ we speak of overrelaxation.

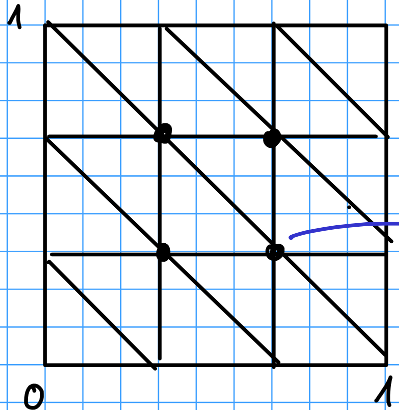
[Ex: implement SOR and compare to Jacobi / Gauß-Seidel.]

(4.11) Application to our model problem

$\Omega = (0,1)^2$

$$-\Delta u = f \quad \text{on } \Omega$$
$$u = 0 \quad \text{on } \partial\Omega$$

Linear FEM on a regular triangulation (cf. (3.15))



$\Big\}\ h = \frac{1}{m} \longrightarrow N = (m-1)^2$ unknowns

$$u_h(x) = \sum_{i,j=1}^{m-1} u_{ij}\, \varphi_{ij}(x)$$

$\rightsquigarrow$ linear system $A_h u_h = b_h$ with $A_h = \begin{bmatrix} A_m & -I_m & & \\ -I_m & \ddots & \ddots & \\ & \ddots & \ddots & -I_m \\ & & -I_m & A_m \end{bmatrix}$, $A_m = \begin{bmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix}$

Eigenvalues/-vectors $\quad A_h z^{k\ell} = \lambda^{k\ell} z^{k\ell}$

with $\lambda^{k\ell} = 2\left(2 - \cos\frac{k\pi}{m} - \cos\frac{\ell\pi}{m}\right)$

$\quad z_{ij}^{k\ell} = \sin\frac{ik\pi}{m} \sin\frac{j\ell m}{m}$, $\quad k,\ell = 1,\dots,m-1$

(4.12) Convergence of Jacobi for the model problem

Jacobi: $C = \underset{\overset{\shortparallel}{M}}{D}^{-1}(\underset{\overset{\shortparallel}{-N}}{L+R})$, $D = 4I$, $L+R = -A_h + 4I$

$= I - \frac{1}{4}A_h$

→ eigenvalues of C: $\lambda^{kl}(C) = \frac{1}{2}\cos\frac{k\pi}{m} + \frac{1}{2}\cos\frac{l\pi}{m}$, $k,l = 1,\dots,m-1$

→ $\rho(C) = \cos\frac{\pi}{m} = 1 - \frac{1}{2}\frac{\pi^2}{m^2} + \mathcal{O}(m^{-4}) = 1 - \frac{\pi^2}{2}h^2 + \mathcal{O}(h^4)$

i.e. the distance of $\rho(C)$ from 1 shrinks with increasing the number of triangles!

→ estimate on the required number of iterations with the Jacobi method:

- iteration: $u^{k+1} = Cu^k + d$
- stop the iteration if $\|u^{k+1} - u_h\| < TOL\,\|u^0 - u_h\|$ (typical $TOL \approx 10^{-6}$)

  true solution $A_h u_h = b_h$
- How do we estimate this ↑ ? We can derive [ex.] an estimate

$$\|u^{k+1} - u_h\| < K \underbrace{\|u^{k+1} - u^k\|}_{\le \rho(C)^k \|u^1 - u^0\|}$$

- using the estimate on $\rho(C)$ from above we get that

$$\left(1-\frac{\pi^2}{2}h^2\right)^k > TOL$$

i.e.

$$k > \frac{2}{\pi^2} h^{-2} \left|\log TOL\right| = \mathcal{O}(N)$$

→ estimate for the computational effort

$$\#\text{flop} = \mathcal{O}\left(\#\text{flop}(A_h v) \cdot \#\text{iterations}\right)$$
$$= \mathcal{O}(N \cdot N) = \mathcal{O}(N^2)$$

4.2 The method of conjugate gradients (cg)

(4.13) Idea

given: $A \in \mathbb{R}^{n \times n}$ symmetric, positive definite, $b \in \mathbb{R}^n$

$\rightarrow$ scalar product: $\langle x,y \rangle_A := x^T A y$

associated norm: $\|x\|_A = \sqrt{\langle x,x \rangle_A}$ energy norm

we still look for $x^* \in \mathbb{R}^n$ such that $Ax^* = b$

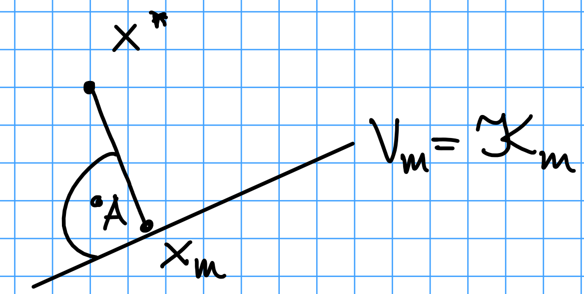
$$\Leftrightarrow \quad \langle Ax^*, y \rangle = \langle b, y \rangle, \quad y \in \mathbb{R}^n$$

Idea (Ritz-Galerkin): choose subspace $V_m \subset \mathbb{R}^n$ and determine $x_m \in V_m$ s.t.

$$\langle Ax_m, y \rangle = \langle b, y \rangle \qquad y \in V_m$$

Here: $V_m = \mathcal{K}_m = \mathcal{K}_m(A,b) = \operatorname{span}\{A^0 b, A^1 b, A^2 b, \dots, A^{m-1} b\}$ Krylov space

(4.14) Derivation



cf. Lemma of Céa: $\quad x^* - x_m \perp_A \mathcal{K}_m$

i.e. x_m is the A-orthogonal projection of x^* onto $\mathcal{K}_m$.

Let $p_1, \dots, p_m$ be an A-orthogonal basis of $\mathcal{K}_m$, then

$$x_m = \sum_{k=1}^{m} \frac{\langle p_k, x^* \rangle_A}{\langle p_k, p_k \rangle_A} p_k \overset{\lightning}{=} \underbrace{\sum_{k=1}^{m} \frac{p_k^T b}{p_k^T A p_k} p_k}_{x^* \text{ does not appear here any more}} \qquad \langle p_k, x^* \rangle_A = p_k^T A x^*$$

We can phrase this in term of an iteration:

$$x_m = x_{m-1} + \alpha_m p_m \qquad\qquad (cg.2)$$

where
$$\alpha_m = \frac{p_m^T b}{p_m^T A p_m} \qquad\qquad (cg.1)$$

The residual is
$$r_m = b - Ax_m = A(x^* - x_m) = A(\underbrace{x^* - x_{m-1}}) - \alpha_m p_m)$$
$$= r_{m-1} - \alpha_m A p_m \qquad\qquad (cg.3)$$

We start with $x_0 = 0$, i.e. $r_0 = b$.

(4.15) Construction of A-orthogonal basis $p_1, \dots, p_m$

Lemma: The residuals $r_0, \dots, r_m$ form an orthogonal basis of $\mathcal{K}_{m+1}$ if $r_m \neq 0$.

Proof: (i) $\mathcal{K}_1 = \text{span}\{r_0\} = \text{span}\{b\}$

$= \text{span}\{p_1, \dots, p_{m+1}\}$
$= \text{span}\{b, Ab, \dots, A^{m-1}b\}$

(ii) Assume the statement holds for $m-1$ then $r_m \in \mathcal{K}_{m+1}$ (by (cg.3), since $r_{m-1} \in \mathcal{K}_m$, $Ap_m \in \mathcal{K}_{m+1}$).

In addition, for any $y \in \mathcal{K}_m$

$$r_m^T y = \left(A(x^* - x_m)\right)^T y = (x^* - x_m)^T A^T y = (x^* - x_m) A y$$
$$= \langle x^* - x_m, y \rangle_A = 0$$

i.e. r_m is orthogonal to $\mathcal{K}_m$. So if $r_m \neq 0$ then

$$\mathcal{K}_{m+1} = \text{span}\{r_0, \dots, r_m\}.$$

∎

From the basis $r_0, \dots, r_{m-1}$ of $\mathcal{K}_m$ we construct an A-orthogonal basis $p_1, \dots, p_m$ by Gram-Schmidt :

(i) $p_1 = r_0 = b$

(ii) $$p_{m+1} = r_m - \sum_{k=1}^{m} \frac{\langle p_k, r_m \rangle_A}{\langle p_k, p_k \rangle_A} p_k = r_m - \frac{\langle p_m, r_m \rangle_A}{\langle p_m, p_m \rangle_A} p_m$$

since $\langle p_k, r_m \rangle_A = p_k^T A r_m = \underbrace{(A p_k)}_{\in \mathcal{K}_m}{}^T r_m \overset{\text{Lemma}}{=} 0$

for $k = 1, \dots, m-1$

We thus also get a recursion formula for the p_m's :

$$p_{m+1} = r_m + \beta_m p_m \qquad (cg.4)$$

$$\beta_m = - \frac{p_m^T A r_m}{p_m^T A p_m} \qquad (cg.5)$$

(4.16) The cg-Algorithm

(Hestenes, Stiefel, 1952):

$x_0 = 0; \quad p_1 = b;$

for $m = 1, 2, 3, \ldots$

$$\alpha_m = \frac{r_{m-1}^T r_{m-1}}{p_m^T A p_m} \qquad (cg.1) \qquad \text{„steplength"}$$

$$x_m = x_{m-1} + \alpha_m p_m \qquad (cg.2) \qquad \text{„descent"}$$

$$r_m = r_{m-1} - \alpha_m A p_m \qquad (cg.3) \qquad \text{residual}$$

$$\beta_m = \frac{r_m^T r_m}{r_{m-1}^T r_{m-1}} \qquad (cg.4)$$

$$p_{m+1} = r_m + \beta_m p_m \qquad (cg.5) \qquad \text{„search direction"}$$

interpretation of cg as an optimization method [ex.]

These variants of (cg.1) and (cg.4) can be derived as follows:

(i) $$p_m^T b = p_m^T A x^* = p_m^T A (x^* - x_0) = \langle p_m, x^* - x_0 \rangle \underset{\uparrow\,(cg.2)}{=} \langle p_m, x^* - x_{m-1} \rangle_A$$

$$= p_m^T A (x^* - x_{m-1}) = p_m^T r_{m-1} \underset{\leftarrow\,(cg.5)}{=} r_{m-1}^T r_{m-1}$$

(ii) $$-\alpha_m p_m^T A r_m = (-\alpha_m A p_m)^T r_m \underset{\text{(cg.3)}}{=} (r_m - r_{m-1})^T r_m = r_m^T r_m$$

$$\Rightarrow \beta_m = \frac{1}{\alpha_m} \frac{r_m^T r_m}{p_m^T A p_m} \underset{\text{[ex.]}}{=} \frac{r_m^T r_m}{r_{m-1}^T r_{m-1}}$$