Krylov's Methods and Applications

Giovanni Barbarino[*]

March 15, 2019

[*] These notes are incomplete and not reviewed. If someone wants to help, mail me at giovanni.barbarino@gmail.com, and you'll be included in the list of authors.

Contents

5/11/18

1 Preliminaries

- Hermitian, skew-Hermitian, unitary,normal,orthogonal, diagonalizable, defective, positive definite matrices
- real/imaginary part of matrices
- characterization of normal matrices
- Schur form

Theorem 1.1 (Schur Real Form)**.** *For every matrix $A \in \mathbb{R}^{n\times n}$ there exists an orthogonal real matrix Q such that $Q^T AQ = T$ where T is block triangular with 1×1 of 2×2 blocks.*

Theorem 1.2 (Polar Form)**.** *For every matrix $A \in \mathbb{C}^{n\times n}$ there exists an unitary matrix U and an Hermitian positive semidefinite matrix S such that $A = US$.*

7/11/18

Theorem 1.3 (Gerschgorin)**.** *The G. disks of a matrix A are defined as*

$$D(a_{i,i}, R_i) = \{ z \in \mathbb{C} \mid |z - a_{i,i}| \leq R_i \}, \qquad R_i = \sum_{j=1, j\neq i}^{n} |a_{i,j}|.$$

The eigenvalues of A are contained in the union of G. disks, called $\mathscr{G}$. If $\mathscr{G}$ has a connected component consisting of p circles, then it will contain exactly p eigenvalues of A.

Notice that you can repeat the same reasonings with A^T since it has the same eigenvalues.

1.1 Stability

Definition 1.1. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be **Stable** if $\Re(\lambda_i(A)) < 0$ for every eigenvalue of A, and it is **Positive Stable** if $-A$ is stable.

Notice that a positive stable matrix is not necessarily positive definite (meaning that its hermitian part is positive definite), but the converse is true. In fact

$$A = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \qquad (A + A^H)/2 = \begin{pmatrix} 1 & a/2 \\ a/2 & 1 \end{pmatrix} \to \det = 1 - \frac{a^2}{4} > 0 \iff a^2 < 4$$

The stability comes from the dynamical system $\dot{x} = Ax$, where the solution $x(t)$ converge to zero if and only if A is stable. In the field of differential equation, we have the heat problem, where $u_t = \Delta u$ and the resulting A is negative definite.

Theorem 1.4 (Bendixson)**.** *If $A = H_1 + iH_2 \in \mathbb{C}^{n\times n}$, where H_1, H_2 are Hermitian, then every eigenvalue λ of A is bounded by*

$$\lambda_{min}(H_1) \leq \Re(\lambda) \leq \lambda_{max}(H_1), \qquad \lambda_{min}(H_2) \leq \Im(\lambda) \leq \lambda_{max}(H_2).$$

Proof. If $Ax = \lambda x$, where x is unitary, then

$$x^* Ax = \lambda \implies \Re(\lambda) = \Re(x^* Ax) = x^* H_1 x, \qquad \Im(\lambda) = \Im(x^* Ax) = x^* H_2 x$$

and then use the bounds

$$\lambda_{min}(H_1) \leq x^* H_1 x \leq \lambda_{max}(H_1), \qquad \lambda_{min}(H_2) \leq x^* H_2 x \leq \lambda_{max}(H_2).$$

□

In the real case $A \in \mathbb{R}^{n\times n}$, then A is positive definite iff $x^T Ax > 0$ for every nonzero real vector x, since if we decompose $A = H + S$ into Hermitian and skew-Hermitian part, then

$$x^T Ax = x^T Hx + x^T Sx, \qquad (x^T Sx)^T = -x^T Sx \implies x^T Ax = x^T Hx.$$

Lemma 1.1 (Kellog)**.** *Given $A = H_1 + iH_2$ the usual decomposition, if A is positive semi-definite then*

$$\|(\alpha I - A)(\alpha I + A)^{-1}\|_2 \le 1 \qquad \forall \alpha > 0,$$

and if A is positive definite, then

$$\|(\alpha I - A)(\alpha I + A)^{-1}\|_2 < 1 \qquad \forall \alpha > 0.$$

Proof.

$$\|(\alpha I - A)(\alpha I + A)^{-1}\|_2 = \sup_{x\in\mathbb{C}^n/\{0\}} \frac{\|(\alpha I - A)(\alpha I + A)^{-1}x\|_2}{\|x\|_2^2}$$

If we call $y = (\alpha I + A)^{-1}x$, then

$$\frac{\|(\alpha I - A)(\alpha I + A)^{-1}x\|_2^2}{\|x\|_2^2} = \frac{\|(\alpha I - A)y\|_2^2}{\|(\alpha I + A)y\|_2^2} = \frac{\alpha^2\|y\|_2^2 - 2\alpha\Re(y^*Ay) + \|Ay\|_2^2}{\alpha^2\|y\|_2^2 + 2\alpha\Re(y^*Ay) + \|Ay\|_2^2} \le 1$$

where the last inequality holds because $\Re(y^*Ay) \ge 0$. It is < 1 if A is positive definite, since $\Re(y^*Ay) \ge \gamma > 0$ for every y. □

Given the classic dynamical system $\dot{x} = Ax$ we can use the *Crank-Nicolson* scheme for the discretization. If $x_0 = x(0)$, $x_k = x(k\Delta t)$, then

$$x_{k+1} = (I - \frac{2}{\Delta t}A)^{-1}(I + \frac{2}{\Delta t}A)x_k.$$

We assume A a stable matrix and negative definite. In this case $\|x_k\| \to 0$ as k goes to zero, in accordance to the analytical solution $x(t) = \exp(tA)x_0$. The matrix $(I - \frac{2}{\Delta t}A)^{-1}(I + \frac{2}{\Delta t}A)$ is a rational approximation of $\exp(\Delta tA)$, and it is called $(1,1)$-*Padè Approximation.*

Suppose now that A is Skew-Hermitian $A = \mathrm{i}H$. This is positive (and negative) semidefinite, and

$$\|(I - \beta A)(I + \beta A)^{-1}\|_2 = \|(I - \beta\mathrm{i}H)(I + \beta\mathrm{i}H)^{-1}\|_2 = 1$$

since $(I - \beta A)(I + \beta A)^{-1}$ is unitary, with eigenvalues of the form $(1 - \beta\mathrm{i}\lambda)/(1 + \beta\mathrm{i}\lambda)$ with $\lambda \in \mathbb{R}$. This is called the *Cayley Transform* that brings the real line into the unitary circle.

An other example is

$$\mathrm{i}\frac{\delta\psi}{\delta t} = H\psi, \qquad \psi(0) = \psi_0 \in \mathbb{C}^n, \qquad \|\psi_0\|_2 = 1$$

that is called *Schrodinger Equation*, with H Hermitian and solution

$$\psi(t) = \exp(-\mathrm{i}tH)\psi_0$$

but the matrix is unitary, so $\|\psi(t)\|_2 = 1$. We can use Crank-Nicolson and obtain

$$\psi_{k+1} = \psi((k+1)\Delta t) = (I - \beta\mathrm{i}H)^{-1}(I + \beta\mathrm{i}H)\psi_k, \qquad \beta = 2/\Delta t.$$

Definition 1.2. A nonnegative matrix/vector is a matrix/vector with all entries nonnegative. They are also called **Digraphs**. they are denoted as

$$\mathbb{R}^n_+ = \{\, x \in \mathbb{R}^n \mid x \ge 0 \,\} \qquad \mathbb{R}^{n\times n}_+ = \{\, A \in \mathbb{R}^{n\times n} \mid A \ge 0 \,\}$$

They are used usually in probability, Markov chains, spectral graph theory, M-matrices, Economics, etc.

Definition 1.3. $A \in \mathbb{C}^{n\times n}$ is **Irreducible** if there's no permutation matrix P such that

$$PAP^T = \begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{pmatrix}$$

where $A_{1,1}$ and $A_{2,2}$ are square matrices. Otherwise A is **Reducible**.

Given a matrix A, we can build the *associated directed graph* $G(A)$ with n nodes and there's a link between nodes i and j if and only if $a_{i,j} \neq 0$. If A has a symmetric sparsity pattern, then the corresponding graph can be considered not directed.

Theorem 1.5. *A is irreducible if and only if the associated graph $G(A)$ is strongly connected.*

Theorem 1.6 (Perron-Frobenius)**.** *If $A \in \mathbb{R}_+^{n\times n}$ is irreducible, then the spectral radius $\rho(A)$ of A is a simple eigenvalue of A with a strictly positive eigenvector $x > 0$. Moreover, $\rho(A)$ increases whenever any entry of A increases.*

A counterexample for the last statement when A is not irreducible is the null matrix, when compared with the nilpotent Jordan matrix. Any nonnegative matrix can be reduced through permutation onto a p-cyclic form

$$\begin{pmatrix} & A_1 & & & \\ & & A_2 & & \\ & & & \ddots & \\ & & & & A_{p-1} \\ A_p & & & & \end{pmatrix}$$

and if $p > 1$, the spectrum of the matrix has a p-symmetry.

Theorem 1.7 (Perron-Frobenius pt.2)**.** *If $A \in \mathbb{R}_+^{n\times n}$, then the spectral radius $\rho(A)$ of A is an eigenvalue of A with a nonnegative eigenvector $x > 0$.*

In this case, we don't know the multiplicity of $\rho(A)$. For example, you an take the null matrix. In [2], you can find a proof that makes use of the Brouwer's fixed point theorem.

1.2 M-matrix

Definition 1.4. A matrix $A \in \mathbb{R}^{n\times n}$ is called **M-matrix** (Minkowski) if $A = rI - B$ with B nonnegative and $r \geq \rho(B)$.

A is also non-singular if and only if $r > \rho(B)$. Notice that in this case, if B is irreducible, then $\text{rk}(A) = n - 1$.

Theorem 1.8. *Given a matrix $A \in \mathbb{R}^{n\times n}$, TFAE:*

1. *A is a nonsingular M-matrix*
2. *$a_{i,j} \leq 0 \quad \forall i \neq j$ and $A^{-1} \geq 0$.*

Proof. Assuming 1), $A = rI - B$ with $\rho(B/r) < 1$, so $A = r(I - B/r)$ is invertible. Also,

$$A^{-1} = \frac{1}{r}\left(I + B/r + B^2/r^2 + \dots\right)$$

so $A^{-1} \geq 0$ since $B \geq 0$.

Assuming 2), we can write $A = rI - B$ for some $r \geq 0$, $B \geq 0$. Notice that $r \neq \rho(B)$, otherwise A is not invertible. If $r < \rho(B)$, then take $x \geq 0$ eigenvector of $\rho(B)$ and notice that $Ax = rx - \rho(B)x = (r - \rho(B))x = y \leq 0$. As a consequence $x = A^{-1}y \leq 0$, so $x = 0$, that is an absurd. □

12/11/18

Consider A a real nonnegative matrix and its associated graph $(V, E) = G(A)$. We can build the graph associated to A^k as

$$G(A^k) = (V, E') \qquad (i, j) \in E' \iff \exists \text{ path of length } k \text{ from } i \text{ to } j.$$

In particular, if A has no null entries on the diagonal, then

$$(i, j) \in E' \iff \exists \text{ path of length at most } k \text{ without loops from } i \text{ to } j.$$

Lemma 1.2. *If $A \in \mathbb{R}^{n\times n}$ is nonnegative and irreducible, then $(I + A)^{n-1}$ is a positive matrix.*

Corollary 1.1. *If A is an irreducible non singular M-matrix, then $A^{-1} > 0$.*

Proof. $A = rI - B$, $r > \rho(B)$, $B \geq 0$. A irreducible means B irreducible and

$$A^{-1} = \frac{1}{r}\left(I + B/r + B^2/r^2 + \dots\right)$$

and this is positive since its associated graph is complete. □

Definition 1.5. An invertible matrix A is called **Monotone** if $A^{-1} \geq 0$.

We have sen that nonsingular M matrix are monotone, but the converse is not true.

Equivalently, A is monotone iff $Ax \geq Ay$ whenever $x \geq y$.

They also satisfy a version of maximum principle. Consider $-\Delta u = f$ on $\Omega \subseteq \mathbb{R}^n$ open and bounded with Dirichlet Boundary conditions at the border $u|_{\delta\Omega} = 0$. If $f \leq 0$, then it represent a 'weight' on the surface u that pulls it down. In fact $-\Delta$ gives out a monotone operator, so that $u = (-\Delta)^{-1} f \leq 0$.

Theorem 1.9. *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix. Then $\Re(\lambda) \geq 0$ for every eigenvalue of A, and if A is non singular, then $\Re(\lambda) > 0$.*

Proof. $A = rI - B$, $r \geq \rho(B)$, $B \geq 0$. We have $\lambda(A) = r - \lambda(B)$ so

$$\Re(\lambda(A)) = r - \Re(\lambda(B)) \geq 0.$$

In the non-singular case, $r > \rho(B)$ and $\Re(\lambda(A)) > 0$. □

This result shows that non singular M matrices are positive-stable, but in general they are not positive definite. For example

$$A = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix}$$

is positive stable, but $(A + A^T)/2$ is indefinite because it has negative determinant.

Theorem 1.10. *If A is a symmetric M-matrix, then it is positive semidefinite, and it is positive definite if it is not singular.*

Notice that there exist positive definite non singular M-matrix that are not symmetric.

Definition 1.6. An SPD M-matrix is called **Stieltjes** matrix.

For example, let P be a nonnegative matrix with $0 \leq P_{i,j} \leq 1$ and row stochastic. Let also x^0 be a row probability vector. We know that $x^k = x^0 P^k$ are all probability vectors, and they describe a Discrete Markov Process, or *Markov Chain*.

The common question on Markov chains are

- Does there exist a limit steady state distribution $x = \lim_{k\to\infty} x^k$?
- Does it depends on the starting state x^0?
- How fast does it converge?

Notice that if $x^k \to x$, then

$$x^k = x^{k-1}P \implies x = xP$$

so a steady state distribution must be a right eigenvector associated to the eigenvalue 1, and it exists since $Pe = e$. Moreover, $\rho(P) = 1$, since $\rho(P) \le \|P\|_\infty = 1$.

Definition 1.7. A probability vector x that satisfies $xP = x$ is called **stationary** distribution of the chain.

If $A = I - P^T$, then it is a singular M-matrix, and the stationary distributions of the chains are the probability vectors in the right kernel of A. A is called *rate matrix*, since it represents the differential in a continuous Markov process.

We know that if P (or equivalently A) is irreducible, the eigenvalue 1 is simple, and the stationary distribution is unique. Even in this case, though, the convergence is not assured.

For example

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \to x = e/2, \qquad x^0 = e_1 \to x^1 = e_2 \to x^2 = e_1 \to \dots$$

The problem here is that P has eigenvalue -1 that has the same magnitude of 1. The cyclic behaviour happens when there are zeros on the diagonal, so one can modify P

$$P \to \widetilde{P} = (1-\alpha)I + \alpha P, \qquad \alpha \in (0,1)$$

where $\widetilde{P}$ is still row stochastic but now it is aperiodic. In this case $\widetilde{P}$ is still irreducible and it has only one stationary distribution, and every x^0 converge to it. Moreover $x\widetilde{P} = x \iff xP = x$, so we are sure to obtain the right distribution. The choice of α is determinant for the speed of the convergence.

An other example is the Laplacian matrix of a graph.

Definition 1.8. If A is the adjacency matrix of G, and $d = Ae$, then $L = diag(d) - A$ is called the **Laplacian Matrix** associated to G.

Exercise 1.1. *Verify that L is always a singular M-matrix $L = rI - B$ where $r = \max_i d_i = \rho(B)$.*

Notice that L is irreducible if and only if A is irreducible if and only if G is connected, and in this case the kernel of L is the span of the vector e. In general, the dimension of the kernel is the number of connected components of G, since we can sort the nodes so that L is block diagonal with each block irreducible. The second smallest eigenvalue λ_2 of L is called *Spectral Gap* or *Fiedler eigenvalue* of G, and indicates how connected the graph is, meaning that the largest λ_2, the hardest to disconnect G (and that's why it is also called *algebraic connectivity*). The corresponding eigenvectors are called *Fiedler vectors*. Notice that such eigenvector x_2 is orthogonal to e, so it has positive and negative entries, and we can divide the nodes into two sets V_1, V_2 depending on the corresponding sign in x_2. This partition is an approximation of an optimal partition of the graph such that the number of edges connecting V_1, V_2 are relatively small with respect to $|V_1| \cdot |V_2|$.

It is called 'Laplacian' since, if we consider a path graph $G = (V, E)$ where $(i,j) \in E \iff |i - j| = 1$, so

$$A = \begin{pmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{pmatrix} \qquad L = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}$$

and L is the discretization through central finite difference of the Laplacian operator on the path graph

$$u''(i) \sim \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2}$$

Notice that in this case 0 is a single eigenvalue of L, since the graph is connected.

Given a graph $G = (V, E)$ and a probability vector x_0, consider the problem

$$\begin{cases} \dot{x} = -Lx, \\ x(0) = x_0. \end{cases}$$

The solution is given by $x(t) = \exp(-tL)x_0$ for $t \geq 0$. It's easy to show that $\exp(-tL) \geq 0$ and it is strictly positive for $t > 0$ if G is connected. Notice that $\{\, exp(-tL) \mid t \geq 0 \,\}$ is a semigroup, since

$$S(t) = \exp(-tL) \implies S(t + t') = S(t)S(t').$$

If G is irreducible and not directed, then L is semipositive definite and $\lambda_2 > 0$, so the eigenvalues of $S(t)$ are $1, \exp(-t\lambda_2), \ldots, \exp(-t\lambda_n)$, where the eigenvectors of $S(t)$ are the same of L. If $x_1, \ldots, x_n$ is an orthonormal basis built with eigenvector of L, where $x_1 = e/\sqrt{n}$, then

$$L = \lambda_2 x_2 x_2^T + \cdots + \lambda_n x_n x_n^T \implies S(t) = x_1 x_1^T + \exp(-t\lambda_2) x_2 x_2^T + \cdots + \exp(-t\lambda_n) x_n x_n^T$$

$$\implies S(t)x_0 = (x_1^T x_0)x_1 + \sum_{i=2}^{n} \exp(-t\lambda_i)(x_i^T x_0)x_i \to \frac{1}{n}(e^t x_0)e$$

Notice that the convergence is dominated by $\exp(-t\lambda_2)$, so the spectral gap is also a measure of the speed of convergence. Cases where λ_2 is large are, for example, the Small World networks, like the social media graphs.

Suppose G is connected, so that $d > 0$. We can consider the *Normalized Laplacian Graph*, that may be defined as

$$L_1 = D^{-1}L = I - D^{-1}A, \qquad L_2 = LD^{-1} = I - AD^{-1}, \qquad L_3 = D^{-1/2}LD^{-1/2} = I - D^{-1/2}AD^{-1/2}.$$

In particular, $D^{-1}A$ is row stochastic, AD^{-1} is column stochastic, and $D^{-1/2}AD^{-1/2}$ is doubly stochastic. They all represent transition matrices for Markov chains, or *Random Walks* on G. For example

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \to D^{-1}A = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

14/11/18

2 Matrix Powers and Polynomials

Given a polynomial $p(x) \in \mathbb{C}[x]$ with degree k, we can evaluate it on a matrix A

$$p(x) = a_0 + a_1 x + \cdots + a_k x^k \to p(A) = a_0 I + a_1 A + \cdots + a_k A^k.$$

If J is the Jordan canonical form of A, with $A = XJX^{-1}$ and X invertible, then

$$p(A) = Xp(J)X^{-1}.$$

Moreover, if J_i are the Jordan blocks of J, where

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{pmatrix} \qquad J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

then

$$p(A) = X \begin{pmatrix} p(J_1) & & & \\ & p(J_2) & & \\ & & \ddots & \\ & & & p(J_s) \end{pmatrix} X^{-1}$$

If we call n_i the dimension of J_i and

$$J_0 = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$$

then

$$J_i^m = (\lambda_i I + J_0)^m = \sum_{l=0}^{m} \binom{m}{l} \lambda_i^{m-l} J_0^l = \begin{pmatrix} \lambda_i^m & m\lambda_i^{m-1} & \dots & \binom{m}{n_i-1}\lambda_i^{m-n_i+1} \\ & \lambda_i^m & \ddots & \vdots \\ & & \ddots & m\lambda_i^{m-1} \\ & & & \lambda_i^m \end{pmatrix}$$

Notice that $J_i^m \to 0$ if and only if $|\lambda_i| < 1$, so we can state the following theorem.

Theorem 2.1. *Let $A \in \mathbb{C}^{n\times n}$. Then*

$$\lim_{m\to\infty} A^m = 0 \iff \rho(A) < 1.$$

If we want the sequence A^m to be just bounded, we need $\rho(A) \leq 1$ and every eigenvalue with magnitude 1 must have only uni-sized Jordan blocks (also called semi-simple or non-defective).

Recall that for any matrix norm, $\rho(A) \leq \|A\|$, and that

$$A^m \to 0 \iff \|A^m\| \to 0.$$

If A is a normal matrix, with $A = UDU^*$, D diagonal and U unitary, then $\rho(A) = \|A\|_2$.

Exercise 2.1. *If $\|A\|_2 = \rho(A)$, then A is normal?*

If A is diagonalizable, then

$$\|A\|_2 = \|XDX^{-1}\|_2 \leq \|X\|_2\|X^{-1}\|_2\|D\|_2 = k_2(X)\rho(A).$$

Also, if $A = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix}$, then $\rho(A) = 0$, but $\|A\|_2 = |a|$ that can be arbitrarily large. This leads to the *Hump Phenomenon*: even if $\rho(A) < 1$, then A^m first rises and then converges to zero.

An other example is $A = \begin{pmatrix} 0.1 & a \\ 0 & 0.1 \end{pmatrix}$. In fact $A^k = \begin{pmatrix} 10^{-k} & ka10^{1-k} \\ 0 & 10^{-k} \end{pmatrix}$ that may converge slowly to zero if $|a|$ is large enough.

Lemma 2.1 (Varga)**.**

$$\|A^k\|_2 \sim \nu \binom{k}{p-1} [\rho(A)]^{k-(p-1)}$$

where p is the size of the largest Jordan block associated to the eigenvalues λ such that $|\lambda| = \rho(A)$, and

$$\frac{1}{k_2(X)} \leq \nu \leq k_2(X)$$

with $A = XJX^{-1}$.

Theorem 2.2 (Householder)**.** *Let $A \in \mathbb{C}^{n\times n}$. Then for every $\varepsilon > 0$ there exists a matrix norm such that*

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

It is also said as

$$\rho(A) = \inf_{\|\cdot\| \text{ matrix norm}} \|A\|.$$

Proof. Consider $\|A\|_L := \|LAL^{-1}\|_2$ that is a matrix norm for every invertible matrix L. Call

$$A' = A/\varepsilon, \qquad A' = VJV^{-1} \quad \implies \quad V^{-1}AV = \varepsilon J.$$

εJ is the Jordan form of A with ε instead of '1' above the main diagonal. This means that

$$\|A\|_V = \|\varepsilon J\|_2 \leq \|D\|_2 + \varepsilon\|E\|_2 = \rho(A) + \varepsilon\|E\|_2$$

and we can find a V for every $\varepsilon > 0$, so

$$\rho(A) \leq \inf_V \|A\|_V \leq \rho(A).u$$

□

Recall the Cayley-Hamilton theorem:

$$p_A(\lambda) = \det(\lambda I - A) \implies p_A(A) = 0.$$

It means that there always exists a degree n polynomial that vanishes on A (and its eigenvalues). This is called *characteristic polynomial* of a matrix. An other important polynomial is called *minimal polynomial* and it is the monic polynomial q_A with least degree that vanishes on A. It is unique, it vanishes on every eigenvalue of A, and any other polynomial that vanishes on A is a multiple of q_A.

Suppose $A \in \mathbb{C}^{n\times n}$ and $\deg(q_A) = m \leq n$. If A is invertible, then we can write the inverse as

$$q_A(x) = a_0 + a_1x + \cdots + a_mx^m \implies A^{-1} = -\frac{a_1}{a_0}I - \frac{a_2}{a_0}A - \cdots - \frac{1}{a_0}A^{m-1}.$$

If we want now to solve a linear system $Av = b$, then

$$v = A^{-1}b = r(A)b, \qquad \deg(r) < m.$$

Corollary 2.1. *If A is Hermitian and A has s distinct eigenvalues, then $\deg(q_A) = s$, and thus*

$$Av = b \implies v = r(A)b, \qquad \deg(r) < s.$$

Let $u \in \mathbb{R}^n$, $\|u\|_2 = 1$. The matrix $A = I + uu^T$ has minimal polynomial of degree 2, with eigenvalues 1 of multiplicity $n-1$ associated with the eigenvectors $u^\perp$ and 2 with multiplicity one associated to the eigenvector u. The minimal polynomial is thus $q_A(x) = (x-1)(x-2) = x^2-3x+2$. It means that $A^{-1} = \frac{3}{2}A - \frac{1}{2}A^2 = I - \frac{1}{2}uu^T$.

For any A, if f is a function for which $f(A)$ is defined (some way), then we can find a polynomial $p(x)$ of degree $< n$ such that $f(A) = p(A)$.

Definition 2.1. Suppose A has eigenvalues $\lambda_1, \dots, \lambda_s$, where λ_i are distinct. We say that *index*(λ_i) is the size of the largest Jordan block associated with λ_i.

Notice that A is non-defective on λ_i if and only if *index*$(\lambda_i) = 1$.

Definition 2.2. A function $f : \Omega \subseteq \mathbb{C} \to \mathbb{C}$ is defined at A if for each $\lambda_i \in \Lambda(A)$ the derivatives $f^{(k)}$ exist on λ_i for $k = 0, 1, \dots, \mathit{index}(\lambda_i) - 1$.

In this case, if $A = XJX^{-1}$, then we can define

$$f(A) = Xf(J)X^{-1}, \qquad J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{pmatrix}, \qquad f(J) = \begin{pmatrix} f(J_1) & & & \\ & f(J_2) & & \\ & & \ddots & \\ & & & f(J_r) \end{pmatrix}$$

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}, \qquad f(J_i) = \begin{pmatrix} f(\lambda_i) & f'(\lambda_i) & \dots & \frac{1}{(n_i-1)!} f^{(n_i-1)}(\lambda_i) \\ & f(\lambda_i) & \ddots & \vdots \\ & & \ddots & f'(\lambda_i) \\ & & & f(\lambda_i) \end{pmatrix}$$

Notice that if A is diagonalizable, then $f(A) = Xf(D)X^{-1}$ and

$$f(D) = \begin{pmatrix} f(\lambda_1) & & & \\ & f(\lambda_2) & & \\ & & \ddots & \\ & & & f(\lambda_n) \end{pmatrix} = \begin{pmatrix} p(\lambda_1) & & & \\ & p(\lambda_2) & & \\ & & \ddots & \\ & & & p(\lambda_n) \end{pmatrix}$$

for every polynomial p that interpolates f on λ_i.

For example, if A is nonsingular, let $f(x) = x^{-1}$. Any polynomial that interpolates $\frac{1}{x}$ on the eigenvalues will be very different from the actual function, since it has a vertical asymptote.

An other example is $f(x) = \exp(-tx)$, where the interpolation will not approximate well the function due to the horizontal asymptote. Notice that if $x(t) = exp(-tA)x_0$ is the solution to a differential problem, it can be computed as $x(t) = p(A)x_0$ for some polynomial.

Suppose A SPD Hermitian, and we are interested in the k lowest eigenvalues with $k << n$ and their eigenspaces. Call $\lambda_1, \dots, \lambda_k$ the wanted eigenvalues, and $x_1, \dots, x_k$ the corresponding eigenvectors. If $\mathscr{J} = Span\,\{\, x_1, \dots, x_k \,\}$, then it is A invariant, and we usually want to compute an orthogonal projector P into $\mathscr{J}$. Suppose $\lambda_k < \mu\lambda_{k+1}$ and define the function

$$h(x) = \begin{cases} 1 & \lambda_1 \le x < \mu, \\ \frac{1}{2} & x = \mu, \\ 0 & x > \mu. \end{cases}$$

The projector will be exactly $P = h(A)$, since

$$A = \sum_{i=1}^{n} \lambda_i x_i x_i^* \implies h(A) = \sum_{i=1}^{k} x_i x_i^*.$$

A useful approximation of the step function is given by the Fermi-Dirac function

$$h(x) = \frac{1}{1 + \exp(\beta(\lambda - \mu))}$$

that can be expanded and truncated into a Taylor polynomial or approximated with Chebychev polynomials basis.

19/11/18

3 Stationary Iterative Methods

Let $x^0, c \in \mathbb{C}^n$ and $T \in \mathbb{C}^{n\times n}$. For $k \in \mathbb{N}$, consider the recurrence

$$x^{k+1} = Tx^k + c.$$

We say that the sequence $\{x^k\}_k$ is generated by a linear first order (depends only on the previous element) stationary method; T is called the *iteration matrix* of the method. It is a short-memory method since you have to store only a matrix and two vectors for every step. It is stationary since T, c are constant in every step.

Suppose $x^k \to x^*$. In this case, by taking the limit, x^* must satisfy

$$x^* = Tx^* + c \implies (I - T)x^* = c$$

and it is a fixed point of $\phi(x) = Tx + c$. Notice that c must be in the range of $I - T$.

Consider the linear system $Ax = b$, with $A \in \mathbb{C}^{n\times n}$ and $b \in \mathbb{C}^n$.

Definition 3.1. $A = B - C$ is a **Splitting** of A if B is non-singular.

In this case, a splitting generates a stationary iteration

$$x^{k+1} = Tx^k + c, \qquad T = B^{-1}C = I - B^{-1}A, \qquad c = B^{-1}b.$$

If the sequence converges $x^k \to x^*$, then $Ax^* = b$. Conversely, given A, T, we can ask if there exists a splitting $A = B - C$ such that $T = B^{-1}C$.

Lemma 3.1. *If A is nonsingular and $1 \notin \Lambda(T)$, then there exists an unique splitting $A = B - C$ such that $T = B^{-1}C$.*

Proof.

$$T = I - B^{-1}A \iff B = A(I - T)^{-1}.$$

□

Theorem 3.1. *Given A, T, there's a splitting $A = B - C$ such that $T = B^{-1}C$ if and only if* $\ker(I - T) = \ker(A)$. *In the case A is singular, there are infinite many splittings.*

Exercise 3.1. *Prove theorem 3.1.*

For example, take

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad T = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \qquad B_1 = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, B_2 = \begin{pmatrix} 1 & -2 \\ -1 & 1 \end{pmatrix}.$$

B_1, B_2 induces two splittings for A, T.

3.1 Convergence of Stationary Iterations

Given A nonsingular and a splitting $A = B - C$, take $T = B^{-1}C$ and $c = B^{-1}b$. What are conditions for convergence? If we call $r^k = b - Ax^k$, then

$$x^{k+1} = Tx^k + c \implies x^{k+1} = x^k + B^{-1}(b - Ax^k) = x^k + B^{-1}r^k.$$

This is useful in the cases when Ax^k, and thus r^k, can be computed, but we don't have direct access to A. If the limit $x^k \to x^*$ exists, then

$$x^* = Tx^* + c \implies e^{k+1} = Te^k = \cdots = T^{k+1}e^0$$

where $e^k = x^k - x^*$. In case of convergence, $e^k \to 0$, and the necessary and sufficient condition for that to happen for every e^0 is that $\rho(T) < 1$. In this case, we know there exists a matrix norm such that $\|T\|_L < 1$ and that is induced by a vector norm. In fact

$$\|x\|_L := \|\,[x, 0, \ldots, 0]\,\|_L \implies \|A\|_L = \sup \frac{\|Ax\|_L}{\|x\|_L}.$$

If A is singular, then $T = I - B^{-1}A \implies 1 \in \Lambda(T) \implies \rho(T) \geq 1$, so the convergence depends on the initial guess x^0.

Assume $x^0 = 0$ and A nonsingular.

$$x^{k+1} = (I + T + \cdots + T^k)c, \qquad \lim_{k\to\infty}(I + T + \cdots + T^k)c = (I - T)^{-1}c = (I - T)^{-1}B^{-1}b = A^{-1}b$$

If we truncate the series, we have an *approximate inverse*

$$A^{-1} \sim \sum_{l=0}^{k} T^l B^{-1.}$$

If A is sparse and B diagonal, then the approximate inverse is easy to apply. It is used as a polynomial preconditioning.

Convergence Rates Recall that $e^k = T^k e^0$. This leads to

$$\frac{\|e^k\|_2}{\|e^0\|_2} \le \|T^k\|_2$$

and this bound is *sharp*, in fact for every k we can find e^0 such that it is an equality.

Definition 3.2. Given a matrix $T \in \mathbb{C}^{n\times n}$ such that $\|T^m\|_2 < 1$ for some $m \in \mathbb{N}$, the **average rate of convergence** for m iterations associated with T is defined as

$$R(T^m) = -\ln\left[\|T^m\|_2^{1/m}\right] = -\frac{\ln\|T^m\|_2}{m}$$

Definition 3.3. The **average reduction factor** per iteration of $\{e^k\}_k$ is given by

$$\sigma = \left(\frac{\|e^k\|_2}{\|e^0\|_2}\right)^{1/k}.$$

If $\|T^k\|_2 < 1$ then

$$\sigma \le (\|T^k\|_2)^{1/k} = \exp(-R(T^k)).$$

The quantity $N_k = 1/R(T^k)$ measures the number of steps needed to reduce the initial error by a factor e, since $\sigma^{N_k} \le 1/e$.

Lemma 3.2 (Gelfand's Formula)**.**

$$\rho(T) = \lim_{k\to\infty} \|T^k\|^{1/k}.$$

Thanks to Gelfand's formula, we can define the *asymptotic rate of convergence* as

$$R_\infty(T) = \lim_{k\to\infty} R(T^k) = \lim_{k\to\infty} -\ln\left[\|T^m\|_2^{1/m}\right] = -\ln[\rho(T)]$$

Notice that

- $\rho(T^k) \le \|T^k\|_2$, so $R_\infty(T) \ge R(T^k)$.
- This is an asymptotic rate of convergence, and may happen that the convergence is very slow.

$$T = \begin{pmatrix} 0.99 & 4 \\ 0 & 0.99 \end{pmatrix} \implies R_\infty(T) = 0.01005 \implies N_\infty = 99.5.$$

In reality, this case is even worse, since $\|T^k\| \ge 1$ for $k \le 805$ and $\|T^k\|_2 < 1/e$ only when $k > 918$.

For example,

$$T = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} \to \rho(T) = 0 \implies R_\infty = \infty$$

but $\|T^k\|_\infty = 1$ for every $k < n$. This is sharp if we take $e^0 = e_n$.

- If T is normal, then $\|T^k\|_2 = \rho(T)^k$, so we know precisely the convergence.
- If T is diagonalizable $T = XDX^{-1}$, then $\|T^k\|_2 \le k_2(X)\rho(T)^k$, so for a large $k_2(X)$ we have a slow convergence.

3.2 Richardson's Iteration (1910)

$$x^{k+1} = x^k + \alpha(b - Ax^k)$$

where $\alpha > 0$ and A is invertible. It is used as a smoother in multigrid methods. We can rearrange the iteration and write

$$\frac{x^{k+1} - x^k}{\alpha} = -Ax^k + b$$

that looks like an approximation of the derivative. It is actually connected to the Forward Euler Method applied to the initial value problem

$$\begin{cases} \dot{x} = -Ax + b \\ x(0) = x^0 \end{cases}$$

and we want $\dot{x} \to 0$ so that x converges to the solution of $Ax = b$. In particular, if $x(t)$ converges for every x^0 to the solution of $Ax = b$, then A is positive stable.

The splitting corresponding to Richardson's Iteration is $A = \alpha^{-1}I - (\alpha^{-1}I - A)$ and the iteration matrix is

$$T = B^{-1}C = I - B^{-1}A = I - \alpha A.$$

If $\lambda_j = \mu_i + \mathrm{i}\nu_j$ are the eigenvalues of A, then the eigenvalues of T are

$$\lambda_j(T) = 1 - \alpha\lambda_j = (1 - \alpha\mu_j) + \mathrm{i}(\alpha\nu_j)$$

and we need

$$\rho(T) < 1 \iff |1 - \alpha\lambda_j| < 1 \iff (1 - \alpha\mu_j)^2 + (\alpha\nu_j)^2 < 1$$

so a necessary condition is

$$(1 - \alpha\mu_j)^2 < 1 \implies -1 < 1 - \alpha\mu_j < 1 \implies 0 < \alpha\mu_j < 2$$

so we need in particular that A is positive stable and α not too large. The condition of positive stability is also sufficient for an $\alpha^* > 0$ to exist s.t. the Richardson's iteration converges for every $\alpha \in (0, \alpha^*)$. In fact, we find

$$0 < \alpha < \frac{2\mu_i}{|\lambda_i|^2}$$

for every i, and we can set

$$\alpha^* = \min_i \frac{2\mu_i}{|\lambda_i|^2}.$$

If A has only real positive eigenvalues (for example Hermitian positive definite) and

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

then $\rho(T) < 1$ for every $\alpha \in (0, \alpha^*)$ where

$$\alpha^* = \min_i \frac{2\mu_i}{|\lambda_i|^2} = \min_i \frac{2}{\lambda_i} = \frac{2}{\|A\|_2} \geq \frac{2}{\|A\|_\infty}$$

so it is enough to have $\alpha \in \left(0, \frac{2}{\|A\|_\infty}\right)$.

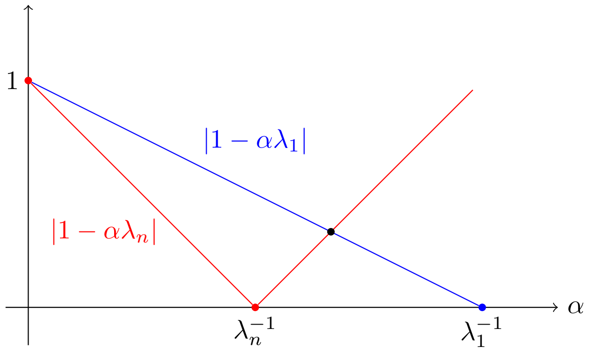
21/11/18

Under the same hypothesis, we want to minimize $\rho(T)$ with respect to α. We know that

$$\lambda_j(T) = 1 - \alpha\lambda_j \implies \rho(T) = \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\}$$

and, as a function of α, $\rho(T)$ has its minimum at

$$\alpha\lambda_n - 1 = 1 - \alpha\lambda_1 \implies \alpha = \frac{2}{\lambda_1 + \lambda_n} \implies \min_{\alpha>0} \rho(T) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$



If A is Hermitian, then the minimum is

$$\min_{\alpha>0}\rho(T)=\frac{k_2(A)-1}{k_2(A)+1}$$

since $k_2(A)=\lambda_n/\lambda_1$. When $k_2(A)$ is low, then the spectral radius is small and the convergence is fast. On the contrary, when $k_2(A)$ is large, then the convergence is slow, that is not always a bad thing.

Often we have $\lambda_1\sim 0$, so the optimal value of α is very close to the bound $2/\lambda_n$, so it could be dangerous. It is thus always better to underestimate α instead of overestimate. This is the reason we estimate

$$\alpha\sim\frac{2}{\|A\|_\infty}\leq\frac{2}{\lambda_n}$$

that is easy to compute and safe to use.

3.3 Ill-posed problems

Definition 3.4 (Hadamard)**.** A mathematical problem is **well-posed** if the solution

- exists,
- is unique,
- it depends continuously on the data.

For example, we can take *Fredholm Integral Equation of Fisrt kind.* Take $X=C[a,b]$ that is a Banach space with norm $\|u\|_\infty$, and consider a continuous function $k:[a,b]^2\to\mathbb{R}$ so that the operator T defined as

$$(Tu)(x)=\int_a^b k(x,y)u(y)dy$$

is a linear compact operator. If we want to solve $Tu=f$ for u, then it is an ill-posed problem, since f may not be in the range of T and the operator may not be injective, so the solution may not exist or be unique. Even if this is not the case, we have T compact, so T^{-1} is unbounded and

$$\widetilde{f}=f+\eta\in range(T),\quad Tu=f,T\widetilde{u}=\widetilde{f}\implies\|u-\widetilde{u}\|=\|T^{-1}\eta\|$$

that can be arbitrarily large, so the problem does not even depends continuously on the data.

This is an example of inverse problem, since we want to reconstruct u from f. Even if we discretize the problem as $Au=f$, with A invertible, then $\|A^{-1}\|$ will be huge, since it comes from the discretization of a compact operator.

Consider a rectangular system $Ax=b$, with $A\in\mathbb{R}^{m\times n}$. It may not have a solution, and it may not be unique, but we can use the normal equations

- If $m\geq n$, then $A^TAx=A^Tb$,
- If $n<m$, then $AA^Ty=b$ with $A^Ty=x$.

note that in both cases, A^TA and AA^T are symmetric and positive semidefinite. If $\text{rk}(A) = n$, then A^TA is positive definite, and if $\text{rk}(A) = m$, then AA^T is positive definite. The normal equation are used to solve the Least Square problem

$$\min_x \|Ax - b\|_2$$

that has always a solution, and it is unique if A has full rank. Even when A has not full rank, we can require an other condition (such as minimum norm) to restore the uniqueness. The solution is given by

$$\text{rk}(A) = n \implies (A^TA)^{-1}A^Tb, \qquad \text{rk}(A) = m \implies A^T(AA^T)^{-1}b.$$

The problem of this approach is that

$$k_2(A^TA) = k_2(AA^T) = k_2(A)^2$$

is often large, so the normal equations are very ill-conditioned and hard to solve numerically.

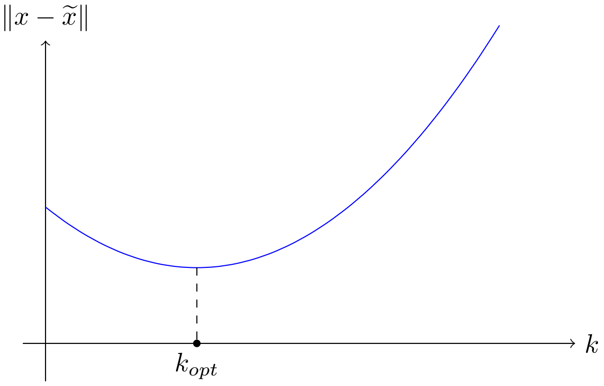
3.4 La??dweler's Iteration

$$x^{k+1} = x^k + \alpha A^T(b - Ax^k)$$

It is nothing but Richardson's method applied to $A^TAx = A^Tb$. If A has rank $n \leq m$, then A^TA is SPD and the method converge for $\alpha \in (0, \alpha^*)$ where

$$\alpha^* = \frac{2}{\lambda_n(A^TA)} = \frac{2}{\|A\|^2}, \qquad \alpha_{opt} = \frac{2}{\sigma_{\min}(A)^2 + \sigma_{\max}(A)^2}.$$

In real computations, we are solving a slightly perturbed system $Ax = \widetilde{b}$, so we do not want the exact solution $\widetilde{x} = A^{-1}\widetilde{b}$ that may be very far to the real solution because the perturbation may amplify through A^{-1}. The plot of relative error $\|x - \widetilde{x}\|$ wrt the number of iteration, we typically see a semi-convergence shaped like a parabola.



If we use the SVD decomposition, we see that

$$A = U\Sigma V = \sum \sigma_i u_i v_i^*, \qquad A^{-1} = V^*\Sigma^{-1}U^* = \sum \frac{1}{\sigma_i} v_i u_i^*$$

$$\implies x = A^{-1}b = \sum_{i=1}^n \frac{u_i^* b}{\sigma_i} v_i, \qquad \widetilde{x} = A^{-1}\widetilde{b} = \sum_{i=1}^n \frac{u_i^* \widetilde{b}}{\sigma_i} v_i = x + \sum_{i=1}^n \frac{u_i^* \eta}{\sigma_i} v_i$$

The problem is when σ_i are too small, since they boost the contamination of the perturbation η. Iteration methods usually try to approximate a k_{opt} where to stop, trying to confine themselves to a subspace relative to the highest singular values, so that the approximation is good enough.

3.5 Preconditioning

In order to solve the system $Ax = b$, instead of Richardson we may use an iteration preconditioned by a nonsingular matrix B

$$x^{k+1} = x^k + \alpha B^{-1}(b - Ax^k).$$

If $B^{-1}A$ is positive stable, then it will converge for $\alpha \in (0, \alpha^*)$ with $\alpha^* = 2/\rho(B^{-1}A)$. The aim is to enlarge $\rho(B^{-1}A)$, and if $B^{-1}A$ has real positive eigenvalues, then the convergence will be improved provided that $k_2(B^{-1}A) < k_2(A)$. In certain sense, B^{-1} is an approximate inverse of A.

If $A = B - C$ is a splitting, the stationary iteration method induced by the splitting

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

is a preconditioned Richardson with $\alpha = 1$. If $\rho(B^{-1}C) < 1$, then the eigenvalues of $B^{-1}A$ lie in the disk $D(1, \rho(B^{-1}C))$, so $\alpha^* = 2/\rho(B^{-1}A) > 1$ and $\alpha = 1 \in (0, \alpha^*)$. So if $B^{-1}A$ is positive stable, then the method converges, and if $B \sim A^{-1}$, then $\rho(B^{-1}A) \sim 1$ and $\alpha^* \sim 2$.

3.6 Classical Iterations

Given a decomposition

$$A = L + D + U$$

where L is strictly low triangular, D is diagonal and U is strictly upper triangular, where D has full rank, the *Jacobi* splitting is

$$A = D - (-L - U)$$

and the *Gauss-Seidel* splitting is

$$A = (D + L) - (-U).$$

These are simple methods that are not used anymore. Gauss-Seidel splitting often converges faster, but not always.

$$A = \begin{pmatrix} 1 & -2 & -2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{pmatrix} = I - \begin{pmatrix} 0 & 2 & 2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & -2 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 2 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$T_J = \begin{pmatrix} 0 & 2 & -2 \\ 1 & 0 & 1 \\ 2 & 2 & 0 \end{pmatrix}, \qquad T_{GS} = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{pmatrix} \implies \rho(T_J) = 0, \qquad \rho(T_{GS}) = 2 + 2\sqrt{2} > 1$$

so Jacobi converges in 3 steps maximum, but Gauss-Seidel diverges.

Theorem 3.2 (Stein-Rosemberg). *[3] Let $A = L + I + U$ the above decomposition (can be always put in this form multiplying by D^{-1}) and suppose $-(L + U) \geq 0$. Then only one of the following relations is valid:*

- $\rho(T_J) = \rho(T_{GS}) = 0$,
- $0 < \rho(T_{GS}) < \rho(T_J) < 1$,
- $\rho(T_J) = \rho(T_{GS}) = 1$,
- $1 < \rho(T_J) < \rho(T_{GS})$.

It means that either both of the methods converge, or they both diverge, and in the first case Gauss-Seidel always converges faster or equally as Jacobi.

Consider the GS iteration

$$\widetilde{x}^{k+1} = T_{GS}x^k + c = -(D + L)^{-1}(Ux^k + b)$$

modified by

$$x^{k+1} = (i - \omega)x^k + \omega\widetilde{x}^{k+1}.$$

where $1 > \omega > 0$. This is called *SOR* method, and it can be written as

$$x^{k+1} = T_{SOR}x^k + c, \qquad T_{SOR} = B(\omega)^{-1}C(\omega), \qquad B(\omega) = \frac{1}{\omega}(D + L), \quad C(\omega) = \frac{1 - \omega}{\omega}(D + L) - U$$

where $A = B(\omega) + C(\omega)$. The cost is the same as GS, but the convergence is improved drastically.

We can test it on the Poisson equation $-\Delta u = f$ on $\Omega \subseteq \mathbb{R}^2$ open bounded and regular, with boundary conditions $u|_{\delta\Omega} = 0$. Using centered FD, it reduces to a linear system $Au = b$. If for example $\Omega = (0, 1)^2$, then we get

$$-\Delta u(x_i, y_i) \sim \frac{-u(x_{i+1}, y_i) + 2u(x_i, y_i) - u(x_{i-1}, y_i)}{h^2} + \frac{-u(x_i, y_{i+1}) + 2u(x_i, y_i) - u(x_i, y_{i-1})}{h^2}$$

$$\implies A = T \otimes I + I \otimes T, \qquad T = \frac{1}{h^2} trid(-1, 2, -1)$$
$$\implies A = h^{-2} trid(-I, H, -I), \qquad H = trid(-1, 4, -1).$$

A is a SPD M-matrix and

$$\rho(T_J) = \cos\left(\frac{\pi}{N}\right) = 1 - \frac{\pi^2}{2} h^2 + O(h^4), \qquad \rho(T_{GS}) = 1 - \pi^2 h^2 + O(h^4) \implies R_\infty(T_{GS}) = 2R_\infty(T_J).$$

Using SOR, we have

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(T_J)^2}} \implies \rho(T_{SOR}) = \frac{2}{1 + \sin\left(\frac{\pi}{N}\right)} - 1 \sim 1 - O(h)$$

so ρ grows to 1, but the convergence is N times slower then GS.

26/11/18

---

4 Block Variants of Stationary Methods

Some applications lead to matrices that have a natural block structure

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,p} \\ \vdots & & \vdots \\ A_{p,1} & \dots & A_{p,p} \end{pmatrix}$$

where every diagonal block $A_{i,i}$ is a square matrix. For example the 3D discrete Laplacian is

$$A = \begin{pmatrix} T & -I & & \\ -I & T & -I & \\ & \ddots & \ddots & \ddots \\ & & -I & T \end{pmatrix}, \quad T = \begin{pmatrix} B & -I & & \\ -I & B & -I & \\ & \ddots & \ddots & \ddots \\ & & -I & B \end{pmatrix}, \quad B = \begin{pmatrix} 6 & -1 & & \\ -1 & 6 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 6 \end{pmatrix}.$$

More general block tridiagonal matrix that can be founf in applications is

$$\begin{pmatrix} A_1 & B_1 & & \\ C_2 & A_2 & B_2 & \\ & \ddots & \ddots & \ddots \\ & & C_p & A_p \end{pmatrix}.$$

Using a non-overlapping domain decomposition, we can break a domain Ω into a partition of Ω_i. The discretization on Ω is split into the points of the domains Ω_i and the interface points, on the borders of Ω_i. Suppose we reorder the points according to the subdomains (useful for example in parallel computing), and we put the interface points as the last ones. We can achieve the reorder through a permutation matrix P, and in the case of the Laplacian operator in 2D, we obtain an arrow-shaped structure.

$$PAP^T = \begin{pmatrix} A_1 & & & & B_1^T \\ & A_2 & & & B_2^T \\ & & \ddots & & \vdots \\ & & & A_p & B_p^T \\ B_1 & B_2 & \dots & B_p & \tilde{A} \end{pmatrix} = \begin{pmatrix} F & G^T \\ G & \widetilde{A} \end{pmatrix}.$$

Consider the 2D Stokes problem

$$\begin{cases} -\Delta u + \nabla p = f & \Omega \subseteq \mathbb{R}^2, \text{ open, bounded} \\ \nabla \cdot u = 0 & u : \Omega \to \mathbb{R}^2 \\ u|_{\delta\Omega} = g & p : \Omega \to \mathbb{R}^2 \end{cases}$$

Upon discretization, we obtain a matrix of the form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \qquad A = \begin{pmatrix} L & 0 \\ 0 & L \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 & B_2 \end{pmatrix}.$$

where L is the discrete Laplacian, and B is a discrete divergence. This is an example of **Saddle Point** problem, and it is represented by a symmetric indefinite matrix.

4.1 Nested Iteration

Given a splitting $A = M - N$ and its iteration

$$x^{k+1} = M^{-1}(Nx^k + b)$$

we can split again $M = F - G$, and replace the exact solution of $Mx^{k+1} = \dots$ with a fixed number p of inner iterations that produce an approximated solution

$$Fy^{i+1,k} = Gy^{i,k} + Nx^k + b.$$

This is still a stationary method $x^{k+1} = Tx^k + c$ where

$$T = I - B^{-1}A, \quad B^{-1} = (I - (F^{-1}G)^p)M^{-1} = \sum_{i=0}^{p-1}(F^{-1}G)^i F^{-1}.$$

There are general convergence results for

- M-matrices and monotone matrices ($A^{-1} \geq 0$),
- Hermitian positive definite matrices,
- diagonally dominant matrices,
- indefinite matrices of saddle point type $\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$.

For diagonally dominant matrices, the algorithms used are Jacobi or Gauss-Seidel.

Definition 4.1 (regular)**.** The splitting $A = B - C$ is said to be **regular** if $B^{-1} \geq 0, C \geq 0$. It is **weak regular** if $B^{-1} \geq 0$ and $T = B^{-1}C \geq 0$.

Theorem 4.1. *Let A be a monotone matrix. If the splitting $A = B - C$ is weak regular, then $\rho(B^{-1}C) < 1$, so the iteration method is convergent.*

Proof. $T = B^{-1}C = I - B^{-1}A \geq 0$, so

$$(I + T + T^2 + \dots + T^m)(I - T) = I - T^{m+1}, \qquad B^{-1} = (I - T)A^{-1} \implies$$

$$0 \leq (I + T + T^2 + \dots + T^m)B^{-1} = (I - T^{m+1})A^{-1} \leq A^{-1}.$$

we have $B^{-1} \geq 0$, so every row of B^{-1} must contain at least one positive entry, so $(I + T + T^2 + \dots + T^m)$ has bounded elements for every m, so it is convergent for $m \to \infty$. In particular $\rho(T) < 1$. □

Corollary 4.1. *Let A be a nonsingular M-matric. If B is a nonsingular matrix obtained by setting to zero the offdiagonal entries of A, then $\rho(I - B^{-1}A) < 1$.*

Proof. The splitting $A = B - C$ is regular, since $C \geq 0$ and $B^{-1} \geq 0$. We have also $A = rI - N$, where $r > \rho(N)$ and $N \geq 0$. we have $B = rI - \widetilde{N}$, where $0 \leq \widetilde{N} \leq N$. the spectral radius is monotone on nonnegative matrices, so the rest follows. (?) □

Corollary 4.2. *Jacobi, Gauss-Seidel and their block variants are convergent when A is a nonsingular M-matrix.*

What about SOR method? We have that the SOR splitting of an M-matrix is regular only if $\omega \in (0, 1]$.

Theorem 4.2 (Kahan)**.** *Let A be a nonsingular M-matrix. Then SOR is convergent for every $\omega \in (0, \overline{\omega})$ where*

$$\overline{\omega} = \frac{2}{1 + \rho(J)}, \qquad J = -D^{-1}(L + U).$$

4.2 Convergence of Alternating Methods

Let $A = M - N = P - Q$ be two splittings, and consider the scheme

$$\begin{cases} x^{k+\frac{1}{2}} = M^{-1}Nx^k + M^{-1}b, \\ x^{k+1} = P^{-1}Qx^{k+\frac{1}{2}} + P^{-1}b. \end{cases}$$

Observe that $\rho(M^{-1}N) < 1$ and $\rho(P^{-1}Q) < 1$ are not sufficient to guarantee the convergence.

Theorem 4.3. *If A is monotone and $A = M - N = P - Q$ are weak regular, then the alternating scheme is convergent and the induced splitting is weak regular.*

Proof. We can bring to to the form $x^{k+1} = Tx^k + c$ where

$$T = (P^{-1}Q)(M^{-1}N) \geq 0, \qquad c = P^{-1}(QM^{-1} + I)b = [P^{-1} + (I - P^{-1}A)M^{-1}]b = P^{-1}(M + P - A)M^{-1}b.$$

We have

$$T = P^{-1}QM^{-1}N = (I - P^{-1}A)(I - M^{-1}A) = I - P^{-1}A - M^{-1}A + P^{-1}AM^{-1}A$$
$$\implies (I - T)A^{-1} = P^{-1} + (I - P^{-1}A)M^{-1} \geq 0$$

so

$$0 \leq (I + T + T^2 + \cdots + T^m)(I - T)A^{-1} = (I - T^{m+1})A^{-1} \leq A^{-1}$$

and we can conclude that the series converges and $\rho(T) < 1$. In this case,

$$T = I - B^{-1}A, \qquad B^{-1} = P^{-1}(M + P - A)M^{-1} = P^{-1} + (I - P^{-1}A)M^{-1} \geq 0(?)$$

so $A = B - C$ is weak regular. □

One can prove that if A is an M-matrix and $A = M - N = P - Q$ are both regular splitting, then

$$\rho(T) \leq \min\{\rho(P^{-1}Q), \rho(M^{-1}N)\}.$$

For example, we can use the symmetric Gauss-Seidel method on a symmetric matrix $A = A^T = L + D + L^T$, where the diagonal D has full rank. $A = (L + D) - (-L^T) = (L^T + D) - (-L)$ are two splittings and in this case

$$A = B - C, \qquad B = M(M + P - A)^{-1}P = (L + D)D^{-1}(L^T + D)$$

that is SPD.

$$B = LD^{-1}L^T + L + L^T + D = LD^{-1}L^T + A \implies C = LD^{-1}L^T.$$

A similar situation arise for the symmetric version of SOR.

Definition 4.2. If $A = B - C$ is a splitting, it is **P-regular** if $B^* + C$ is positive definite.

Notice that if B is hermitian, then $A = B - C$ is P-regular if $2B - A$ is positive definite.

28/11/18

Lemma 4.1. *Suppose A is Hermitian. $A = B - C$ is P-regular if and only if $A - T^*AT$ is HPD.*

Proof. $T = B^{-1}C = I - B^{-1}A$.

$$G = A - T^*AT = A - (I - B^{-1}A)^*A(I - B^{-1}A) = A - A + A^*(B^{-1})^*A - A^*(B^{-1})^*AB^{-1}A + AB^{-1}A$$
$$= (B^{-1}A)^*(B + B^* - A)(B^{-1}A).$$

The matrix $B + B^* - A$ is Hermitian, so G is positive definite if and only if $B + B^* - A = B^* + C$ is positive definite, that means the splitting id P-regular. (? we don't need A invertible?) □

Definition 4.3. Given a positive definite Hermitian matrix A,

$$\|x\|_A^2 := x^* Ax.$$

Theorem 4.4. *If A is HPD, then $A = B - C$ is a P-regular splitting if and only if $\|T\|_A < 1$.*

Proof. We need to show that $\|T\|_A < 1$ iff $A - T^*AT$ is HPD.

Assume first that $A - T^*AT$ is HPD.

$$\|Tx\|_A^2 = x^*T^*ATx = x^*Ax - x^*(A - T^*AT)x < x^*Ax - \varepsilon = \|x\|_A - \varepsilon \implies \|T\|_A < 1.$$

For the converse, assume $\|T\|_A < 1$. In this case,

$$\|x\|_A > \|Tx\|_A \quad \forall x \implies x^*Ax > x^*T^*ATx \quad \forall x \implies A - T^*AT \quad HPD.$$

□

Corollary 4.3. *If A is HPD, and $A = B - C$ is P-regular, then the splitting is convergent.*

Theorem 4.5 (Ostrowski, Reich)**.** *If A is HPD and $\omega \in (0, 2)$, then the SOR iteration for solving $Ax = b$ is convergent.*

Proof. The SOR splitting is

$$B = \frac{1}{\omega}(D + \omega L), \quad C = \frac{1}{\omega}[(1-\omega)D - \omega L^*], \quad B^* + C = \frac{2-\omega}{2}D$$

that is HPD if and only if $\omega \in (0, 2)$. □

In particular, Gauss-Seidel and the block versions converge too.

Notice that id A is Hermitian but indefinite, any splitting $A = B - C$ where B is HPD is divergent. In fact, in $B^{-1}C = I - B^{-1}A$, the matrix $B^{-1}A$ has negative and positive eigenvalues, since

$$B^{-1}A \sim B^{-1/2}AB^{-1/2}$$

that is indefinite, so $\rho(B^{-1}C) > 1$.

Returning to the alternating method,

Theorem 4.6. *If A is HPD and $A = M - N = P - Q$ are P-regular splittings, then the alternating iteration is convergent and the induced splitting $A = B - C$ is P-regular.*

Proof. We already proved that the method converges, but here it is easier to prove.

$$T = (P^{-1}Q)(M^{-1}N) \implies \|T\|_A < 1.$$

□

It proves that

$$\|T\|_A \leq \|P^{-1}Q\|_A \|M^{-1}N\|_A < \max\{\|P^{-1}Q\|_A, \|M^{-1}N\|_A\}$$

so the alternating method is faster than both original method, but it has a greater computational cost. Moreover, the resulting B is HPD, even in the block case for SSOR and symmetric Gauss-Seidel.

Exercise 4.1. *Suppose A is HPD, and $A = A_1 + A_2$ with $A_1 = A_2^*$. Let $M = A_1 + \alpha I$ and $P = A_2 + \alpha I$. Then the splittings $A = M - N = (A_1 + \alpha I) - (\alpha I - A_2)$ and $A = P - Q = (A_2 + \alpha I) - (\alpha I - A_1)$ are both P-regular for every $\alpha > 0$.*

In this case the method is convergent for every $\alpha > 0$, so it is called *unconditional convergence.* Notice that if α goes to zero or goes to infinite, then $\|T\|_A$ goes to one, so there's an optimal α that can be computed.

4.3 ADI Method

The ADI method stands for Alternating Direction Implicit method. Let A be HPD matrix with $A = H + V$ with H hermitian and V HPD. The alternating iteration is

$$\begin{cases}(H+\alpha I)x^{k+\frac{1}{2}} = (\alpha I - V)x^k + b,\\ (V+\alpha I)x^{k+1} = (\alpha I - H)x^{k+\frac{1}{2}} + b.\end{cases}$$

Theorem 4.7. *The ADI iteration converges $\forall\alpha > 0$.*

Proof. The iteration matrix is

$$T = (V+\alpha I)^{-1}(-H+\alpha I)(H+\alpha I)^{-1}(-V+\alpha I)$$

$$(V+\alpha I)T(V+\alpha I)^{-1} = (-H+\alpha I)(H+\alpha I)^{-1}(-V+\alpha I)(V+\alpha I)^{-1} = S_1S_2$$

where the eigenvalues of S_1 are $(\alpha-\lambda)/(\alpha+\lambda) < 1$ and the same holds for S_2, so $\|S_1\|_2 < 1$, $\|S_2\|_2 < 1$ and thus $\|(V+\alpha I)T(V+\alpha I)^{-1}\|_2 < 1$, so $\rho(T) < 1$. □

Notice that in this case, the unique splitting $A = B - C$ such that $T = B^{-1}C$ has

$$B = \frac{1}{2\alpha}(HV + \alpha H + \alpha V + \alpha^2 I) = \frac{1}{2}A + \frac{\alpha}{2}I + \frac{1}{2\alpha}HV$$

that is positive definite for α large enough, even though it is not Hermitian.

As an example, take A the discrete 2D Laplacian $A = T\otimes I + I\otimes T$, $T = trid(-1,2,-1)$. We take $H = T\otimes I$ and $V = I\otimes T$. In this case H is block diagonal, and also V is block diagonal after a permutation, so the computation is highly parallelizable. The two matrices V and H have the same spectrum, and the optimal choice of α is

$$\alpha_* = \sqrt{\lambda_1\lambda_n}$$

4.4 Alternating Hermitian/skew-Hermitian method

The HSS method [1] is applied to $A = H + S$ where H and S are the Hermitian and skew-Hermitian parts of A. The splittings in this case are

$$A = (H+\alpha I) - (\alpha I - S) = (S+\alpha I) - (\alpha I - H).$$

Theorem 4.8. *If A is positive definite (so that H is HPD), then HSS converges for every $\alpha > 0$.*

Proof.

$$T = (S+\alpha I)^{-1}(-H+\alpha I)(H+\alpha I)^{-1}(-S+\alpha I)$$

$$(S+\alpha I)T(S+\alpha I)^{-1} = (-H+\alpha I)(H+\alpha I)^{-1}(-S+\alpha I)(S+\alpha I)^{-1} = S_1S_2$$

As before, $\|S_1\|_2 < 1$ and it is Hermitian. S_2 is a unitary matrix, since it is the Cayley transform of a skew-Hermitian matrix, so $\|S_2\|_2 = 1$. We thus have $\rho(T) < 1$. □

In this case, the α that minimizes $\|S_1\|_2$ is again $\sqrt{\lambda_1\lambda_n}$ where the eigenvalues are the ones of H.

If we consider the Stokes problem

$$\begin{cases}-\Delta u + \nabla p = f & \Omega \subseteq \mathbb{R}^d,\ \text{open, bounded}\\ \nabla\cdot u = 0 & u : \Omega \to \mathbb{R}^d\\ u|_{\delta\Omega} = g & p : \Omega \to \mathbb{R}^d\end{cases}$$

where $d = 2$, then upon discretization, we obtain a matrix of the form

$$\begin{pmatrix} A & B^T \\ B & 0\end{pmatrix}\begin{pmatrix} u \\ p\end{pmatrix} = \begin{pmatrix} b \\ c\end{pmatrix}, \qquad A = \begin{pmatrix} L & 0 \\ 0 & L\end{pmatrix}, \qquad B = \begin{pmatrix} B_1 & B_2\end{pmatrix}.$$

where L is the discrete Laplacian, and B is a discrete divergence. This is represented by a symmetric indefinite matrix, and if we assume A HPD and B of full rank, then the discretization matrix is non singular.

Uzawa's method A popular method called *Uzawa's method* uses the splitting

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} A & 0 \\ B & -\frac{1}{\omega}I \end{pmatrix} - \begin{pmatrix} 0 & B^T \\ 0 & -\frac{1}{\omega}I \end{pmatrix},$$

where $\omega > 0$, leading to the iteration

$$\begin{cases} Au^{k+1} = b - B^T p^k, \\ p^{k+1} = p^k + \omega(Bu^{k+1} - c). \end{cases}$$

In this case,

$$\begin{aligned} u^{k+1} = A^{-1}b - A^{-1}B^T p^k &\implies p^{k+1} = p^k + \omega(BA^{-1}b - BA^{-1}B^T p^k - c) \\ &\implies p^{k+1} = (I - \omega BA^{-1}B^T)p^k + \omega BA^{-1}b - \omega c \end{aligned}$$

so it is a Richardson iteration applied to

$$BA^{-1}B^T p = BA^{-1}b - c$$

where the first matrix $S = BA^{-1}B^T$ is the Schur complement. Remember that B has full rank and S is SPD. It converges for all $\omega \in (0, \omega^*)$ where

$$\omega^* = \frac{2}{\lambda_{max}(S)}(?)$$

and

$$\omega_{opt} = \frac{2}{\lambda_{min}(S) + \lambda_{max}(S)}$$

with spectral radius

$$\rho = \frac{\lambda_{max}(S) - \lambda_{min}(S)}{\lambda_{max}(S) + \lambda_{min}(S)} = \frac{k-1}{k+1}(?).$$

3/12/18

For example, in the case of Stokes problem, discretized by *stable finite element* methods, there exist constants c_1, c_2 independent of the discretization parameter h such that

$$0 < c_1 h^2 \leq \lambda_{min}(S) < \lambda_{max}(S) \leq c_2 h^2$$

hence

$$k = \frac{\lambda_{max}(S)}{\lambda_{min}(S)} = O(1)$$

so the rate of convergence is independent from h.

Augmented Lagrangian Method (Hestenes, Powell, 1969)

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

is non singular if A is symmetric and positive semidefinite, and $\ker(A) \cap \ker(B) = 0$. But if A is not invertible, then we cannot apply Uzawa's method.

Suppose W be a SPD matrix (often diagonal) and $\gamma > 0$. Consider the augmented system

$$\begin{pmatrix} A + \gamma B^T W^{-1} B & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} \tilde{b} = b + \gamma B^T W^{-1} Bu \\ c \end{pmatrix}$$

Notice that $Bu = c$, so

$$\begin{cases} Au + B^T p = b \\ Bu = c \end{cases} \implies \begin{cases} Au + \gamma B^T W^{-1} Bu + B^T p = b + \gamma B^T W^{-1} c \\ Bu = c \end{cases}$$

hence the solution (u, p) is the same. Notice that $A + \gamma B^T W^{-1} B$ is SPD, since

$$x^T(A + \gamma B^T W^{-1} B)x = x^T Ax + \gamma (Bx)^T W^{-1}(Bx) \geq 0$$

and it is zero only when $x \in \ker(A) \cap \ker(B) \implies x = 0$. So we apply Uzawa's method to the augmented system. This is also called *method of multipliers*. It corresponds to the splitting

$$\begin{pmatrix} A + \gamma B^T W^{-1} B & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} A + \gamma B^T W^{-1} B & 0 \\ B & -\frac{1}{\omega} I \end{pmatrix} - \begin{pmatrix} 0 & -B^T \\ 0 & -\frac{1}{\omega} I \end{pmatrix}$$

so

$$0 < \omega < \frac{2}{\lambda_{max}(S_\gamma)}, \qquad S_\gamma = B(A + \gamma B^T W^{-1} B)^{-1} B^T.$$

Notice that is A is invertible, then

$$S_\gamma^{-1} = S_0^{-1} + \gamma W^{-1} \implies [B(A + \gamma B^T W^{-1} B)^{-1} B^T]^{-1} = (BA^{-1} B^T)^{-1} + \gamma W^{-1}$$

If γ diverges to $+\infty$, then the eigenvalues of S_γ go to zero, so

$$0 < \omega < \frac{2}{\lambda_{max}(S_\gamma)} \to \infty.$$

For the optimal ω, we have

$$\omega_{opt} = \frac{2}{\lambda_{min}(S_\lambda) + \lambda_{max}(S_\lambda)} \to \infty$$

and the spectral radius goes to zero, so the rate of convergence explodes. Notice moreover that if $\omega = \gamma$, then

$$\rho = \frac{1}{1 + \gamma \lambda_{min}(S)}$$

that is 1/2 when $\gamma = 1/\lambda_{min}(S)$. The catch is in the condition number, since

$$k(A + \gamma B^T W^{-1} B) \to \infty$$

so we want to achieve a large γ to improve the speed of the convergence, but not too large, otherwise the linear systems become too difficult to compute exactly, due to the big condition number.

Arrow-Hurwicz method When $Ax = b$ is too expensive to solve, then one can follow the modified iteration

$$\begin{cases} u^{k+1} = u^k + \alpha(b - Au^k - B^T p^k) \\ p^{k+1} = p^k + \omega(Bu^{k+1} - c) \end{cases}$$

where $0 < \alpha < \alpha^*$ and $0 < \omega < \omega^*$, and has the advantage to be parallelizable and there's no linear system to solve, but it is generally slow. It is induced by the splitting

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha} I & 0 \\ B & -\frac{1}{\alpha} I \end{pmatrix} - \begin{pmatrix} \frac{1}{\alpha} I - A & -B^T \\ 0 & -\frac{1}{\omega} I \end{pmatrix} = P - Q$$

One can determine the intervals for α, ω such that the method converges and estimate the convergences for optimal values, but they're generally slow.

If we consider $Q_A \sim Q$ and $Q_B \sim S = BA^{-1}B^T$ approximations that are easy to invert and SPD, we can use them as preconditioners and speed up the iteration.

$$\begin{cases} u^{k+1} = u^k + \alpha Q_A^{-1}(b - Au^k - B^T p^k) \\ p^{k+1} = p^k + \omega Q_B^{-1}(Bu^{k+1} - c) \end{cases}$$

We remark that if $Q_A = A$ and $Q_B = I$, then we recover Uzawa's method. If $Q_A = I$, then it returns to be the original Arrow-Hurwicz method. If $Q_A \sim A$ and $Q_B = I$, then it is called *Inexact* Arrow-Hurwicz method.

HSS iteration for Saddle point problems Let $M = H+S$ be the Hermitian-Skewhermitian decomposition.

$$\begin{cases} (H+\alpha I)x^{k+\frac{1}{2}} = (\alpha I - S)x^k + b, \\ (S+\alpha I)x^{k+1} = (\alpha I - H)x^{k+\frac{1}{2}} + b. \end{cases}$$

Remember that If M is positive definite (so that H is HPD), then HSS converges for every $\alpha > 0$ to the unique solution of $Mx = b$.

Notice that the system

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

can be rewritten as

$$\begin{pmatrix} A & B^T \\ -B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} b \\ -c \end{pmatrix}$$

and it can be split into

$$\begin{pmatrix} A & B^T \\ -B & 0 \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & B^T \\ -B & 0 \end{pmatrix} = H + S$$

where $H+\alpha I$ is SPD whenever A is positive semidefinite, and $S+\alpha I$ is positive definite with Schur complement

$$\alpha I + B(\alpha I)^{-1}B^T = \alpha I + \frac{1}{\alpha}BB^T.$$

Solving system with $H + \alpha I$ boils down to solve SPD system with $A + \alpha I$, that is diagonally dominant if α is large. On the other end, $S + \alpha I$ reduces to its Schur complement $\alpha I + \frac{1}{\alpha}BB^T$ that is also a SPD matrix.

5 Krylov Subspace Methods

Definition 5.1. Given a matrix $A \in \mathbb{C}^{n\times n}$ and a vector $v \in \mathbb{C}^n$, the m-th **Krylov Subspace** is

$$\mathscr{K}_m := span\,\{\, v, Av, A^2v, \dots, A^{m-1}v \,\}\,.$$

Note that for every A and v,

$$\dim(\mathscr{K}_m) \le m \le n.$$

Definition 5.2. The **minimal polynomial** of v wrt A is the monic polynomial of least degree $q_v(x)$ such that $q_v(A)v = 0$. The degree is called the **grade** of v wrt A.

Note that the grade of v is always less or equal than the degree of the minimal polynomial of A.

Lemma 5.1. *If μ is the grade of v wrt A, then the subspace $\mathscr{K}_\mu(A, v)$ is A-invariant. Moreover,*

$$\mathscr{K}_m(A, v) = \mathscr{K}_\mu(A, v) \qquad \forall\, m \ge \mu.$$

Lemma 5.2.

$$\dim(\mathscr{K}_m(A, v)) = m \iff \deg q_v(A) \ge m$$

Proof. It is equivalent to say that

$$\{\, v, Av, \dots, A^{m-1}v \,\}$$

is a basis for $\mathscr{K}_m(A, v)$ if and only if

$$\sum_{i=0}^{m-1} \alpha_i A^i v = 0 \iff \alpha_i = 0 \quad \forall\, i,$$

since in this case the grade of v is at least m. □

In other words,

$$\dim(\mathscr{K}_m(A, v)) = \min\{m, \text{grade of } v\}.$$

Suppose now that $v_0 \in \mathbb{C}^n$ is an initial guess for the solution of $Ax = b$. The *Krylov methods* are approximation methods where at the m-th iterative we find

$$x_m \in x_0 + \mathscr{K}_m(A, r_0) = \{\, u \in \mathbb{C}^n \mid u = x_0 + p_{m-1}(A)r_0 \,\}$$

where $r_0 = Ax_0 - b$ and p_{m-1} are polynomials of degree at most $m-1$. In this case, we can rewrite

$$x_m = (I - Ap_{m-1}(A))x_0 + p_{m-1}(A)b$$

and in the special case $x_0 = 0$,

$$x_m = p_{m-1}(A)b.$$

Notice that we want a polynomial such that $p_{m-1}(A) \sim A^{-1}$, and its existence is assured by the Cayley-Hamilton theorem. This is also a special case of the matrix function $f(X) = X^{-1}$.

10/12/18

12/12/18

Theorem 5.1. *Let $[\alpha, \beta] \subset \mathbb{R}$ with $-\infty < \alpha < \beta < \infty$ and $\gamma \in \mathbb{R} \setminus [\alpha, \beta]$. Then the problem*

$$\min_{p \in \mathbb{P}_k, p(\gamma)=1} \max_{t \in [\alpha,\beta]} |p(t)| = \min_{p \in \mathbb{P}_k, p(\gamma)=1} \|p\|_{\infty,[\alpha,\beta]}$$

is solved by taking

$$p(t) = \widehat{C}_k(t) = \frac{C_k\left(1 + 2\frac{t-\beta}{\beta-\alpha}\right)}{C_k\left(1 + 2\frac{\gamma-\beta}{\beta-\alpha}\right)}.$$

Notice that $\|C_k\|_{\infty,[-1,1]} = 1$ for every $k \geq 0$, hence

$$\min_{p \in \mathbb{P}_k, p(\gamma)=1} \max_{t \in [\alpha,\beta]} |p(t)| = \frac{1}{\left|C_k\left(1 + 2\frac{\gamma-\beta}{\beta-\alpha}\right)\right|} = \frac{1}{\left|C_k\left(2\frac{\gamma-\mu}{\beta-\alpha}\right)\right|}, \qquad \mu = \frac{\alpha+\beta}{2}.$$

Moreover, $C_k\left(2\frac{\gamma-\mu}{\beta-\alpha}\right) < 0$ only if $\gamma < \alpha$. In this case the best approximation polynomial is

$$p(t) = \widehat{C}_k(t) = \frac{C_k\left(1 + 2\frac{\alpha-t}{\beta-\alpha}\right)}{C_k\left(1 + 2\frac{\alpha-\gamma}{\beta-\alpha}\right)}.$$

If $|t| > 1$, then the definition of Chebyshev polynomial becomes

$$C_k(t) := \cosh[k \cosh^{-1}(t)] = \frac{1}{2}\left[(t + \sqrt{t^2-1})^k + (t + \sqrt{t^2-1})^{-k}\right].$$

When $k >> 1$ and $t > 1$, the first term dominates

$$C_k(t) \sim \frac{1}{2}(t + \sqrt{t^2-1})^k.$$

Let $\eta = \lambda_{\min}(A)/(\lambda_{\min}(A) + \lambda_{\max}(A))$. It is positive since A is a HPD. For $m \geq 1$, we have

$$C_m(t) = \frac{1}{2}\left[(t + \sqrt{t^2-1})^m + (t + \sqrt{t^2-1})^{-m}\right] \geq \frac{1}{2}(t + \sqrt{t^2-1})^m$$

$$C_m(1+2\eta) \geq \frac{1}{2}(1 + 2\eta + 2\sqrt{\eta(\eta+1)})^m,$$

$$1 + 2\eta + 2\sqrt{\eta(\eta+1)} = (\sqrt{\eta} + \sqrt{\eta+1})^2 = \frac{(\sqrt{\lambda_{\min}(A)} + \sqrt{\lambda_{\max}(A)})^2}{\lambda_{\max}(A) - \lambda_{\min}(A)} = (?)$$

$$\implies \frac{1}{C_m(1+2\eta)} \leq 2\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^m, \qquad \forall\, m \geq 1$$

Theorem 5.2. *Let A be HPD, $b, x_0 \in \mathbb{C}^n$ and $e_0 = A^{-1}b - x_0$. Denote with $x_m \in x_0 + \mathscr{K}_m(A, b)$ the unique minimizer of $\|x - A^{-1}b\|_A$. Then*

$$\|e_m\|_A = \|A^{-1}b - x_m\|_A \leq 2\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^m \|e_0\|_A.$$

Proof. We already know that

$$\|e_m\|_A = \min_{p\in\Pi_m} \|p(A)e_0\|_A.$$

Let λ_i be the eigenvalues of A and let $u_1, u_2, \dots, u_n$ the corresponding eigenvectors. Expand e_0 wrt the orthonormal basis of u_i

$$e_0 = \sum_{i=1}^n \xi_i u_i$$

so that

$$p(A)e_0 = \sum_{i=1}^n \xi_i p(\lambda_i)u_i \implies \|p(A)e_0\|_A^2 = \sum_{i=1}^n |\xi_i|^2 \lambda_i p(\lambda_i) \leq \max_{1\leq i\leq n} (p(\lambda_i))^2 \|e_0\|_A^2 \leq \max_{x\in[\lambda_{\min},\lambda_{\max}]} (p(x))^2 \|e_0\|_A^2.$$

Therefore,

$$\|e_m\|_A \leq \min_{p\in\Pi_m} \max_{x\in[\lambda_{\min},\lambda_{\max}]} |p(x)| \|e_0\|_A \leq \frac{1}{C_m(1+2\eta)} \|e_0\|_A \leq 2\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^m \|e_0\|_A$$

□

Notice that we have a fast convergence if $k_2(A) \sim 1$, so the common technique is to look for a good preconditioner. On the other hand, if it is large, the predicted convergence may be a lot slower than the actual convergence (since it always finishes at step n), so the estimation is not very sharp.

Corollary 5.1.

$$\|e_m\|_2 \leq 2\sqrt{k_2(A)}\left(\frac{\sqrt{k_2(A)}-1}{\sqrt{k_2(A)}+1}\right)^m \|e_0\|_2.$$

Proof.

$$\lambda_{\min}^{1/2}(A) I \leq A^{1/2} \leq \lambda_{\max}^{1/2}(A) I$$

and

$$\|x\|_A = \|A^{1/2}x\|_2$$

lead to the wanted relation. □

The error $\|e_m\|_A$ decays monotonously, but the same cannot be said for $\|e_m\|_2$ that usually oscillates.

Let now A be diagonalizable $A = XDX^{-1}$ where $D = \operatorname{diag}(\lambda_i)$ and $\lambda_i \in \mathbb{C}$. In this case

$$\|r_m\|_2 = \|b - Ax\|_2 = \min_{p\in\Pi_m} \|p(A)r_0\|_2 = \min_{p\in\Pi_m} \|Xp(D)X^{-1}r_0\|_2$$

$$\leq k_2(X)\|r_0\|_2 \min_{p\in\Pi_m} \|p(D)\|_2 = k_2(X)\|r_0\|_2 \min_{p\in\Pi_m} \max_{1\leq i\leq n} |p(\lambda_i)| \leq k_2(X)\|r_0\|_2 \min_{p\in\Pi_m} \max_{x\in S} |p(x)|$$

where S is a set in $\mathbb{C}$ that contains all the eigenvalues λ_i. For a unitary matrix, $k_2(X) = 1$ and we find again a familiar relation. The residual is now monotonic in 2-norm, but the bound is

$$\frac{\|r_m\|_2}{\|r_0\|_2} \leq k_2(X) \min_{p\in\Pi_m} \max_{x\in S} |p(x)|.$$

Notice that the RHS may be ≥ 1 and in this case, the bound is useless. Therefore, if $k_2(X)$ is big, (for example when A is almost singular) then the bound is often not informative, and $k_2(X)$ is not easy to compute. On the other hand, if A is almost normal, meaning that $k_2(X) \sim 1$, then the eigenvalues of A alone are almost sufficient to predict the convergence behaviour, so it boils down again to find good preconditioners. In fact if the eigenvalues are clustered away from zero, then the convergence will be fast.

A result of Greenbaum, Strakos and Ptak [5] states that "Any non-increasing convergence curve is possible for GMRES", where GMRES stands for "generalized minimum residual method". In fact, given any set of complex numbers $\lambda_1, \dots, \lambda_n$ and any non-increasing convergence profile $(m, \|r_m\|_2)$ there exists $A \in \mathbb{C}^{n\times n}$ with eigenvalues λ_i and residual r_m at step m. It means that the eigenvalues don't tell us everything we need about convergence, even in the case they are clustered.

5.1 Hermitian Indefinite Case

Suppose A is an indefinite Hermitian matrix, with positive and negative eigenvalues, so that the convex hull of its eigenvalues always comprehends the point 0. For example, the Saddle point problems fall in this category

$$A = \begin{pmatrix} M & B^T \\ B & -C \end{pmatrix}$$

where M, C are symmetric real (semi)positive definite $M > 0$, $C \geq 0$. Also the eigenvalue problem $A - \mu B$ with $A, B > 0$ for certain μ the matrix $A - \mu B$ is indefinite. Also in linear differential problem, there may happen to have an indefinite discretization system, for example associated to

$$-\Delta u - ku = f \quad \Omega, \qquad u|_{\partial\Omega} = \gamma.$$

In these cases, the problem

$$\min_{p \in \Pi_m} \max_{1 \leq i \leq n} |p(\lambda_i)|$$

cannot be replaced by

$$\min_{p \in \Pi_m} \max_{x \in [\lambda_{min}, \lambda_{\max}]} |p(x)|$$

because $p(0) = 1$, so the minimum never goes below 1, so we have to work on two separated intervals $I_- = [\lambda_1, \lambda_s]$ and $I_+ = [\lambda_{s+1}, \lambda_n]$, where $\lambda_s < 0 < \lambda_{s+1}$, so that

$$\min_{p \in \Pi_m} \max_{1 \leq i \leq n} |p(\lambda_i)| \leq \min_{p \in \Pi_m} \max_{x \in I_- \cup I_+} |p(x)|.$$

There's no known close form solution, except in special cases. For example, we have a solution (De Boor, Rice, 1982) when $I_- = -I_+$ that leads to the bound

$$\min_{p \in \Pi_m} \max_{1 \leq i \leq n} |p(\lambda_i)| \leq \min_{p \in \Pi_m} \max_{x \in I_- \cup I_+} |p(x)| = 2 \left(\frac{\sqrt{|\lambda_1 \lambda_n|} - \sqrt{|\lambda_s \lambda_{s+1}|}}{\sqrt{|\lambda_1 \lambda_n|} + \sqrt{|\lambda_s \lambda_{s+1}|}} \right)^{[m/2]}$$

and gives an analogous bound on $\|r_m\|_2$. Notice that it can be adapted to all cases by enlarging I_- and I_+ in order to make them specular sets. If $\lambda_n = 1 = -\lambda_1$ and $\lambda_s = -\lambda_{s+1}$, then

$$\frac{\|r_m\|_2}{\|r_0\|_2} \leq 2 \left(\frac{\sqrt{|\lambda_1 \lambda_n|} - \sqrt{|\lambda_s \lambda_{s+1}|}}{\sqrt{|\lambda_1 \lambda_n|} + \sqrt{|\lambda_s \lambda_{s+1}|}} \right)^{[m/2]} = 2 \left(\frac{k_2(A) - 1}{k_2(A) + 1} \right)^{[m/2]}.$$

16/01/19

5.2 Steepest Descent

For $Ax = b$ with A SPD,

- Compute $r_0 = b - Ax_0$, $p_0 = Ar$, and set $k = 0$.
- Until convergence, do
  - $\alpha_k = \frac{(r_k, r_k)}{(p_k, r_k)}$
  - $x_{k+1} = x_k + \alpha_k r_k$
  - $r_{k+1} = r_k - \alpha_k p_k (= b - Ax_{k+1})$
  - $p_{k+1} = Ar_k$
  - $k = k + 1$

In case of A SPD, we know that the Steepest DEscent (SD) method minimizes at each step the function

$$f(x) = \|x - x^*\|_A^2 = (x - x^*)^T A (x - x*)$$

over all vectors of the form $x_k + \alpha r_k$, $r_k = -\Delta f(x_k)$. Here $Ax^* = b$. the coefficient α_k is chosen by putting $\phi'(\alpha) = 0$ where $\phi(\alpha) = f(x_k + \alpha r_k)$.

It can be shown that

Theorem 5.3.

$$\|e_m\|_A \leq \left(\frac{k-1}{k+1} \right)^m \|e_0\|_A, \qquad k_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

The method is quite slow.

MINRES Remember that for MINRES in the Hermitian Indefinite case we have

$$\min_{p\in\Pi_m}\max_{1\leq i\leq n}|p(\lambda_i)|\leq 2\left(\frac{k-1}{k+1}\right)^{[m/2]}.$$

which is about as bad as SD.

CG on Normal Equation Given A full column rank, then minimize the quantity $\|Ax-b\|_2$ is equivalent to solve

$$A^*Ax=A^*b.$$

Here A^*A is HPD, and the error bound for CG is

$$\|e_m\|_{A^*A}\leq\left(\frac{\sqrt{k(A^*A)}-1}{\sqrt{k(A^*A)}+1}\right)^m\|e_0\|_{A^*A}=\left(\frac{k(A)-1}{k(A)+1}\right)^m\|e_0\|_{A^*A}$$

5.3 Asymptotic Converge Factor

Suppose our eigenvalues are clustered on two intervals I^-, I^+. In this case we do not have explicit bounds, so it is useful to introduce the *asymptotic converge factor*

$$\rho(I^-\cup I^+):=\lim_{m\to\infty}\left(\min_{p\in\Pi_m}\max_{\lambda\in I^-\cup I^+}|p(\lambda)|\right)^{1/m}$$

which can be estimated in some cases.

Consider a family of problems $\{A_n\}_n$ where $A\in\mathbb{C}^{n\times n}$ that naturally arises from discretization of linear PDE through FE, FD, Isogeometric Analysis, etc. etc. We would like to estimate the rate of convergence of Krylov methods for $n\to\infty$. The condition number of A_n goes as $O(n^2)$, so the CG deteriorates and the number of iterations grows. For this reason, we usually use a preconditioner P such that $k(P_n^{-1}A_n)$ is bounded uniformly, and the convergence of CG becomes independent from n.

Stationary Stokes Problem

$$\begin{cases}-\nabla u+\Delta p=f & \Omega\subseteq\mathbb{R}^d\\ div(u)=0 & \Omega\\ B.C.\end{cases}$$

where Ω is bounded with Lipschitz border, and $u:\Omega\to\mathbb{R}^d$ is a velocity field and $p:\Omega\to\mathbb{R}$ is a pressure field.

The discretization is an other example of saddle point problem

$$\mathscr{A}\begin{pmatrix}u\\p\end{pmatrix}=\begin{pmatrix}A & B^T\\ B & -\beta C\end{pmatrix}\begin{pmatrix}u\\p\end{pmatrix}=\begin{pmatrix}f\\0\end{pmatrix}$$

Usually A is a block diagonal matrix with Laplacian blocks on the diagonal, and B^T is a discrete gradient.

If the boundary conditions are $u|_{\partial\Omega}=0$, then u is a function in $(H_0^1(\Omega))^d$ and $p\in L_0^2(\Omega)$ that is $L^2(\Omega)$ quotiented by the constant functions. Using piecewise linear finite elements for both spaces, it leads to an instability on the pressure term (since $\beta=0$), called "Checkerboard instability". To solve the question, we need to consider higher degree elements for the velocity, or to introduce a small stabilization term $\beta>0$ with C positive semidefinite.

Consider the preconditioner $P=D_A\otimes D_C$ where $D_A=\operatorname{diag}(A)$ is SPD and $D_C=\beta h^dI$ if $C=0$ or $D_C=\beta\operatorname{diag}(C)$ otherwise. In this case

$$\Lambda(P^{-1}\mathscr{A})\subseteq(-a,-bh)\cup(ch^2,d)$$

where a,b,c,d are positive constants. the condition number is still $O(h^{-2})$, so it is not as much of preconditioner. If we call $I^-=(-a,-bh)$ and $I^+=(ch^2,d)$, then they have different lengths, and if we symmetrize them as $\widetilde{I}^-=(-d,-ch^2)$, we obtain an estimate that goes as $O(h^{-2})$. For the asymptotic convergence factor, it has been shown that [Wathen, Fischer, Silvester (1995)]

$$\rho(I^-\cup I^+)=O(1-\sqrt{\frac{bc}{ad}}h^{3/2}).$$

In the symmetrized version, we get

$$\rho(\widetilde{I^-} \cup I^+) = O(1 - \frac{c}{d}h^2).$$

An independent rate of convergence is obtained by using as preconditioner the matrix

$$\widetilde{P} = \begin{pmatrix} A & 0 \\ 0 & M_p \end{pmatrix}, \qquad (M_p)_{i,j} = \int_\Omega \varphi_i(x)\varphi_j(x)dx$$

where M_p is the pressure mass matrix and φ_i are the basis functions for the piecewise polynomial functions (of some degree) in L_0^2. Actually, we can replace M_p with its diagonal, and A with any good SPD preconditioner for A (such that $k(\widetilde{A}^{-1}A)$ is uniformly bounded, and in this case, $\widetilde{A}$ and A are said to be *spectral equivalent*).

5.4 Minimum Residual Methods for non-Normal systems

Recall the definition of *Field of Values*

$$F(A) := \{ x^*Ax \mid x \in \mathbb{C}^n, \|x\|_2 = 1 \}.$$

It is always compact and convex (Haussdorf-Toeplitz Theorem) and contains the spectrum of A. If A is normal, then $F(A)$ is the convex hull of $\Lambda(A)$. Otherwise, they are quite different. If P is an invertible matrix, notice also that $F(P^{-1}A)$ and $F(AP^{-1})$ can be arbitrarily different, even though they have the same eigenvalues (they are similar).

Suppose we have a family of linear systems $\{A_n\}_n$. If there exists a compact set $K \subseteq \mathbb{C}$ independent of n with $0 \notin K$ and $F(A_n) \subseteq K$ definitively in n, then the (generalized) Minimum Residual Method(MRM) converges with rate independent of n. In fact, in this case, we can find a polynomial that approximates $1/z$ with error ε independent from n.

Remember that, for the Bendixson theorem, we can bound the spectrum of A with the eigenvalues of H_1, H_2 that are Hermitian and Skew-Hermitian part of A. Actually, we can say more:

$$F(A) \subseteq [\lambda_{\min}(H_1), \lambda_{\max}(H_1)] \times [\lambda_{\min}(H_2), \lambda_{\max}(H_2)].$$

If $\lambda_{\min}(H_1) > c > 0$ and $\|A\|_2 \leq C$ for all n, then $F(A)$ is contained in a compact $K \subseteq \mathbb{R}^+ \times \mathbb{R}$ not containing zero.

21/01/19

Lemma 5.3 (Elmon, ~1983)**.** *If A is positive definite, (meaning that $\Re(A)$ is HPD) then at each step m of a minimal residual method, the residuals $r_m = b - Ax_m$ satisfy*

$$\|r_{m+1}\|_2 \leq (1 - \frac{\mu^2}{\sigma^2})^{1/2}\|r_m\|_2$$

where $\mu = \lambda_{\min}(\Re(A))$ and $\sigma = \|A\|_2 = \sigma_{\max}(A)$.

Crouzeix's Conjecture "For any $A \in \mathbb{C}^{n\times n}$ and any analytic function $g : \Omega \to \mathbb{C}$ with $F(A) \subseteq \Omega$, and Ω being an open set, it holds that

$$\|g(A)\|_2 \leq 2\|g\|_{\infty,F(A)}$$

where $\|g\|_{\infty,F(A)} = \max_{z\in F(A)} |g(z)|$."

This holds for normal matrices with

$$\|g(A)\|_2 \leq \|g\|_{\infty,F(A)}$$

and in 2005 Crouzeix proved that in full generality

$$\|g(A)\|_2 \leq 11.08\|g\|_{\infty,F(A)}.$$

In ~2017, Crouzeix and Palencia proved that the constant is at most $1 + \sqrt{2}$. Given an approximation of $A^{-1}b$, we have

$$\|p_m(A)b - A^{-1}b\|_2 \leq \|p_m(A) - A^{-1}\|_2\|b\|_2 \leq (1 + \sqrt{2})\|p_m(z) - z^{-1}\|_{\infty,F(A)}\|b\|_2$$

and from approximation theory we know that

$$\min_{p_m \in \Pi_m} \|p_m(z) - z^{-1}\|_{\infty,F(A)} = O(\exp(-\alpha m))$$

if $F(A)$ does not contain the origin, for some $\alpha > 0$.

6 Arnoldi Iterations

It is a projection method onto the Krylov subspaces. It builds a orthonormal basis.

```
• choose a vector $v_1$ s.t. $\|v_1\| = 1$
• for $j = 1 : m$ do
    – compute $h_{i,j} = (Av_i, v_i)$ for every $i = 1 : j$
    – compute $w_j = Av_j - \sum_{i=1}^{j} h_{i,j} v_i$
    – $h_{j+1,j} = \|w_j\|_2$
    – if $h_{j+1,j} = 0$ then stop
    – $v_{j+1} = w_j / h_{j+1,j}$.
```

This is not very stable, since it loses orthogonality between vectors quickly.

Lemma 6.1. *If the Arnoldi process does not stop before m steps, then the Arnoldi vectors $v_1, \dots, v_j$ form an orthonormal basis for the Krylov subspace $\mathscr{K}_m(A, v_1)$.*

Proof. They are all orthonormal by construction, and it is clear by induction that $v_j \in \mathscr{K}_j(A, v_1)$, so we conclude that they are a basis thanks to the dimensions. In fact, $v_1 \in \mathscr{K}_1(A, v_1)$, and by induction $w_j \in A\mathscr{K}_j(A, v_1) + \mathscr{K}_j(A, v_1) \subseteq K_{j+1}(A, v_1)$, so the same holds for v_{j+1}. □

If we denote

$$V_m = [v_1, v_2, \dots, v_m]$$

then

Lemma 6.2. *Let $\widehat{H}_m$ be the $(m+1) \times m$ Hessemberg matrix whose entries are the quantities $h_{i,j}$ of the Arnoldi process. Also let H_m be the $m \times m$ Hessemberg matrix obtained from $\widehat{H}_m$ by deleting the last row. Then the Arnoldi identities hold:*

$$V_{m+1}\widehat{H}_m = V_m H_m + w_m e_m^T = AV_m, \qquad V_m^* A V_m = H_m.$$

Proof. $V_{m+1}\widehat{H}_m = V_m H_m + w_m e_m^T = AV_m$ descend naturally from the iterations.

$$V_m^* A V_m = V_m^* V_m H_m + V_m^* w_m e_m^T = H_m.$$

□

The spectral properties of H_m reflect the ones of A in some sense. In fact the eigenvalues of H_m, called *Ritz values*, are good approximations of the eigenvalues of A. If Arnodi process reaches the n-th step, we obtain that H_n is similar to A through an orthogonal transformation, so it has the same eigenvalues and singular values of A.

Lemma 6.3 (Breakdown)**.** *Arnoldi process breaks down at step j iff*

$$grade_A(v_1) = j$$

. Under these assumption, the space $\mathscr{K}_j(A, v_1)$ is A-invariant.

Proof. If the grade is j, then $w_j = 0$, since $\mathscr{K}_{j+1} = \mathscr{K}_j$, so Arnoldi breaks down. On the other side, if $w_j = 0$, then $w_j = Ap(A)v_1 = 0$ so the degree of v_1 is less or equal than j. To show it is exactly j, just refer to the first part of the proof and show that in that case the Arnoldi process would have stopped before.

The rest of the proof was already an exercise before.

□

Notice that if a breakdown occurs at step m, we have found an invariant subspace of A and the projection on this subspace is exact.

$$A = V_j H_j V_j^*.$$

In fact, if $K \subseteq \mathbb{C}^n$ is a subspace, consider the linear system $Ax = b$, and assume that $\dim(K) = m << n$. Let P_K be the orthogonal projection onto K and let $C \subseteq \mathbb{C}^n$ be another subspace (possibly the same) and let Q_K^C the orthogonal projector onto K orthogonally to C. (?) these projectors can be defined by $P_k x \in K$,

$x - P_K x \in K^\perp$ and $Q_K^C x \in K$, $x - Q_K^C x \in C^\perp$. Let $A_m = Q_K^C A P_K$, so that, if $K = \mathscr{K}_m$ the Krylov subspace, if $x_0 = 0$, consider the projected system

$$Q_K^C(b - Ax) = 0$$

for $x \in K$, is equivalent to $A_m x = \hat{b}$ where $\hat{b} = Q_K^C b$. Thus we are looking for an approximate solution $\widehat{x}$ in K, the projected problem has effectively dimension m. If K is A-invariant, then $\widehat{x}$ is actually the exact solution of $Ax = b$.

Theorem 6.1. *For a linear system $Ax = b$ assume $x_0 = 0$ and $b \in K$, where K has dimension m. (incomplete statement)*

Proof. (missing due to low battery computer) □

23/01/19

6.1 Arnoldi based methods for $Ax = b$

Consider the case $C = K = K_m = K_m(A, r_0)$. We seek an approximation $x_m \sim A^{-1}b$ with $x_m \in x_0 + K_m$. By Galerkin condition,

$$b - Ax_m \perp K_m.$$

Let $v_1 = r_0/\|r_0\|_2$ in Arnoldi's method; set $\beta_0 = \|r_0\|_2$. Then

$$V_m^T A V_m = H_m$$

is an upper Hessemberg matrix where

$$V_0^T r_0 = \beta_0 e_1$$

by orthogonality. Hence, the resulting approximation x_m is of the form

$$x_m = x_0 + V_m y_m, \qquad y_m = \beta_0 H_m^{-1} e_1$$

since

$$H_m y_m = V_m^T A V_m y_m = V_m^T A(x_m - x_0) = V_m^T(r_0 - b + Ax_m) = V_m^T r_0 = \beta_0 e_1.$$

The inverse of H_m is a dense matrix, but for example the entries of the first column become smaller with exponential decay in m, that proves the convergence of the method.

The resulting Full Orthogonalization Method(FOM) is as follows

```
• choose $x_0 \in \mathbb{R}^m$
• compute $r_0 = b - Ax_0$, $\beta = \|r_0\|_2$, $v_1 = r_0/\beta$
• For $j = 1 : m$, do
    – compute $w_j = Av_j$ (bottleneck)
    – For $i = 1 : j$ do
        ∗ compute $h_{i,j} = (w_i, w_j)$
        ∗ compute $w_j = w_j - h_{i,j} v_i$
    – compute $h_{j+1,j} = \|w_j\|_2$
    – If $h_{j+1,j} = 0$, set $m = j$ and break
    – set $v_{j+1} = w_j / h_{j+1,j}$
• solve $H_m y_m = \beta_0 e_1$
• set $x_m = x_0 + V_m y$
```

It is a modified GS. We can monitor $\|r_m\|_2$ without computing it at every step, since

Lemma 6.4. *The FOM residual at step m satisfies*

$$r_m = -h_{m+1,m} v_{m+1} e_m^T y_m, \qquad \|r_m\|_2 = |h_{m+1,m}(y_m)_m|.$$

Proof.

$$\begin{aligned} r_m = r_0 - AV_m y_m &= \beta v_1 - V_m H_m y_m - h_{m+1,m} v_{m+1} e_m^T y_m \\ = \beta v_1 - \beta V_m e_1 - h_{m+1,m} v_{m+1} e_m^T y_m &= -h_{m+1,m} v_{m+1} e_m^T y_m. \end{aligned}$$

□

Actually, the true residual may differ from $|h_{m+1,m}(y_m)_m|$ so we may recompute it every five to ten step.

Each step of FOM costs approximately

$$2\text{nz}(A) + 2mn$$

where $\text{nz}(A)$ is the number of nonzero entries in A, so that sparse matrices lead to faster methods. The storage cost grows as $(m+3)n + \frac{m^2}{2}$.

A remedy to the lose of orthogonality and the growing costs is Restarting: fix an index m to stop the iterations and restart the algorithm with $x_0 = x_m$.

6.2 GMRES

Implemented by Saad and Schultz in 1986. We take $K = K_m$ again, but $C = AK_m$. If we let v_1 be the first residual normalized, imposing $r_m \perp C$ results in a method minimizes the norm of the residual over x_0+K_m.Recall that

$$x_m = x_0 + V_m y_m$$

so to derive the algorithm we define a "cost function"

$$J(y_m) = \|r_m\|_2 = \|b - A(x_0 + V_m y_m)\|_2 = \|\beta e_1 - \widehat{H}_m y_m\|_2.$$

Minimize J is a least square problem over x_0+K_m. This is inexpensive since the matrix $\widehat{H}_m$ is upper Hessemberg, so a QR factorization can be very cheap. Moreover as m increases we can upgrade the factorization instead of recomputing it.

The GMRES algorithm can be described as

```
• compute $r_0 = b - Ax_0$, $\beta = \|r_0\|$, $v_1 = r_0/\beta$
• For $j = 1 : m$
    – compute $w_j = Av_j$
    – For $i = 1 : j$
        * compute $h_{i,j} = (w_i, w_j)$
        * compute $w_j = w_j - h_{i,j} v_i$
    – compute $h_{j+1,j} = \|w_j\|_2$
    – If $h_{j+1,j} = 0$, set $m = j$ and break
    – set $v_{j+1} = w_j / h_{j+1,j}$
• $\widehat{H}_m = (h_{i,j})_{i=1:m+1}^{j=1:m}$
• solve $y_m = \min_y \|\beta e_1 - \widehat{H}_m y\|_2$
• set $x_m = x_0 + V_m y_m$
```

Notice that different variants exist corresponding to different ways to orthogonalize the vectors. Moreover, the cost is similar to FOM: linear increasing in operations and quadratic increase in storage. It means that also restarted GMRES is widely used.

The LS problem inside GMRES is typically solved by QR factorization via Givens rotations. Starting from the Hessemberg matrix, the factorization cost is linear at each step, and it is a very stable algorithm. If $\widehat{H}_m = Q_m \widehat{R}_m$, where Q_m is a product of m plane rotations, then we can call $\widetilde{g}_m = \beta Q e_1$, and prune the last row of $\widehat{R}_m$ and $\widehat{g}_m$ to obtain a square upper triangular matrix R_m and a vector g_m so that the solution of the LS problem is given by

$$y_m = R_m^{-1} g_m.$$

One can do it gradually at each step, so that the cost remains sub-quadratic. The residual will be

$$\|r_m\|_2 = \|V_{m+1}(\beta e_1 - \widehat{H}_m y_m)\|_2 = |(\widehat{g}_m)_{m+1}|.$$

Theorem 6.2. *Let A be a nonsingular matrix. Then GMRES breaks down at step j iff $x_j = A^{-1}b$.*

This is called a 'happy breakdown' and one can stop whenever $h_{j+1,j}$ is smaller than a fixed tolerance. In practice, GMRES is preferred to FOM due to "faster" convergence on real problem (optimality in 2-norm). Notice that the method with restarting at m may not converge for every m, but we have the following theorem

Theorem 6.3. *If A is positive definite, then GMRES with restarting at m converges to the solution of $Ax = b$ for every $m \geq 1$.*

Proof. Let us prove this for $m = 1$. The case $m > 1$ follows from the fact that we minimize the residual norm on a larger subspace, so the convergence cannot be ruined. The proof for $m = 1$ descends from the next section analysis. □

Minimum Residual Method If we generalize the steepest descent method to A not Hermitian, we get

- $r_0 = b - Ax_0$, $p_0 = Ar_0$, $k = 0$
- until convergence do
  - $\alpha_k = \frac{(r_k,r_k)}{(p_k,p_k)}$
  - $x_{k+1} = x_k + \alpha_k r_k$
  - $r_{k+1} = r_k - \alpha_k p_k$
  - $p_{k+1} = Ar_{k+1}$

This convergence for A positive definite thanks to Lemma 5.3. Notice that the method at each step minimizes $\|b - Ax_{k+1}\|_2$ over $x_0 + Span(r_k)$.

Lemma 6.5 (Elmon, ∼1983)**.** *If A is positive definite, (meaning that $\Re(A)$ is HPD) then at each step m of a minimal residual method, the residuals $r_m = b - Ax_m$ satisfy*

$$\|r_{m+1}\|_2 \leq (1 - \frac{\mu^2}{\sigma^2})^{1/2} \|r_m\|_2$$

where $\mu = \lambda_{\min}(\Re(A))$ and $\sigma = \|A\|_2 = \sigma_{\max}(A)$.

Proof.

$$\|r_{k+1}\|_2^2 = (r_k - \alpha_k Ar_k, r_k - \alpha_k Ar_k) = (r_k - \alpha_k Ar_k, r_k) - \alpha_k (r_k - \alpha_k Ar_k, Ar_k) = (r_k - \alpha_k Ar_k, r_k)$$

$$= \|r_k\|^2 - \alpha_k (Ar_k, r_k) = \|r_k\|^2 \left(1 - \alpha_k \frac{(Ar_k, r_k)}{(r_k, r_k)}\right) = \|r_k\|^2 \left(1 - \frac{(r_k, r_k)}{(Ar_k, Ar_k)} \frac{(Ar_k, r_k)}{(r_k, r_k)}\right)$$

Now we use

$$\frac{(Ar_k, r_k)}{(r_k, r_k)} \geq \lambda_{\min}, \qquad \frac{(r_k, r_k)}{(Ar_k, Ar_k)} \leq \|A\|_2$$

to finish the proof. □

27/02/19

---

7 Reordering

$$\begin{cases} -\nabla u + 100(u_x + u_y) = f & (0,1)^2 \\ u|_{\partial\Omega} = 0 \end{cases}$$

When we use central finite differences with order 2 of precision, we get an oscillation behaviour, especially if the diffusion coefficient is large. With upwind approximation we lose the oscillations but we only get a first order precision.

If we use an uniform 400×400 grid and a 5 points FDM, with an ILUT$(10^{-3}, 10)$ preconditioner and a Krylov method (BiCGStab), we notice that the ordering of the grid points modify the density of the matrices and the number of iterations needed to achieve convergence.

For example, the *Nested Dissection* sorting makes a binary subdivision of the domain. When we divide Ω into 4 squares, we can put first the points inside the squares in lexicographic order and then the border points. The resulting matrix will be block diagonal with arrow-type blocks, except for the final rows/columns.

In order to estimate an incomplete factorization we have two measures

$$N_1 = \|A - LU\| \qquad N_2 = \|I - A(LU)^{-1}\|$$

where the second estimates the stability. With ND we have good N_1 but bad N_2. Incomplete factorizations usually are strongly sequential so hardly parallelizable.

8 Sparse Approximate Inverse

The idea is to explicitly build a sparse matrix M such that $M \sim A^{-1}$, that can be used as a preconditioner. Usually we find a product $M = M_1 M_2$ without actually computing explicitly the product. Notice that the preconditioning operation reduces to a matrix-vector product that is easily parallelizable.

Generally, A^{-1} is dense, so a lot of research has gone to prove that most elements are actually small. In general we consider A an irreducible matrix (so that it cannot be reduced to smaller problems).

Remember that the *Structural Inverse* of a matrix A is the union of all sparsity patterns of A^{-1} as the nonzero entries if A spans all possible values. It turns out that the structural inverse of an irreducible matrix is full. In fact, if $\|A\| < 1$ then

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

but A is irreducible, so the associated graph is strongly connected and thus every entry i, j will be non-zero in some power A^k, since there exists a path from i to j. Therefore, the sparsity pattern of the inverse is full, since it is the transitive closure of the graph. Indeed, if A comes from the discretization of an Elliptic PDE on a connected domain Ω, the resulting matrix will have a full inverse.

Notice that $\|A\|^k \to 0$ so we may think to truncate the series, but it is not a very good way to approximate the inverse. It hints nonetheless that if A is banded, the biggest elements are concentrated around the main diagonal.

For example, take $A = trid(-1, 3, -1)$ and consider A^{-1}. It will be a dense matrix with very fast decay off-diagonal, so it is usually approximated by band matrices.

Definition 8.1. Given an even number $m \in \mathbb{N}$, we say that A is m-banded if $a_{i,j} = 0$ if $|i - j| > m/2$.

For instance, we say that a tridiagonal matrix is 2-banded. We can assume that the matrices are *structurally symmetric*, meaning that $a_{i,j} \neq 0 \iff a_{j,i} \neq 0$. Moreover, given a compact subset K of $\mathbb{C}$, we will use $\|f\|$ for the sup of $|f|$ over k. Moreover, let $E_N(f, K)$ be the best polynomial approximation of f on K of degree at most N with $\|\cdot\|$ norm.

Theorem 8.1 (Chebyshev)**.** *Let $f(x) = x^{-1}$ on $[a, b]$ with $0 < a < b < \infty$. if $k = b/a$ and $q = \frac{\sqrt{k}-1}{\sqrt{k}+1}$, then*

$$E_N(f, [a, b]) = \frac{(1 + \sqrt{k})^2}{2b} q^{N+1}.$$

Theorem 8.2 (Demko, Moss, Smith, '83)**.** *Let $A = A^T$ an SPD and m-banded matrix with $a = \lambda_1(A)$, $b = \lambda_n(A)$. Let $C_0 = \frac{(1+\sqrt{k})^2}{2b}$ and let $\lambda = q^{2/m}$. Then*

$$|(A^{-1})_{i,j}| \leq \max\{C_0, 1/a\} \lambda^{|i-j|}$$

Proof. Note that A^k is km-banded and so is $p(A)$ if the degree of p is less or equal to k. If $A = QDQ^T$ with Q orthogonal and D diagonal, then $p(A) = Qp(D)Q^T$ and $A^{-1} = QD^{-1}Q^T$ so

$$\|A^{-1} - p_N(A)\|_2 = \left\| \frac{1}{x} - p_N(x) \right\|_{\Lambda(A)}$$

and if p_N is the best approximant, then

$$\left\| \frac{1}{x} - p_N(x) \right\|_{\Lambda(A)} = E_N(\frac{1}{x}, [a,b]) = C_0 q^{N+1}.$$

Now write $|i-j| = N\frac{m}{2} + k$ with positive k so that $|i-j|\frac{2}{m} \leq N+1$, therefore

$$|(A^{-1})_{i,j}| = |(A^{-1})_{i,j} - p_N(A)_{i,j}| \leq \|A^{-1} - p_N(A)\|_2 \leq C_0 \lambda^{|i-j|}.$$

Eventually, if $i = j$ then $|A^{-1}_{i,j}| \leq \|A^{-1}\|_2 = \frac{1}{a}$. □

With infinite matrices, the bound is sharp.

Corollary 8.1. *If C and λ are independent of n, then fo every $\varepsilon > 0$ we can find an index p independent from n and a p banded matrix B such that*

$$\|A^{-1} - B\| \leq \varepsilon.$$

Moreover B can be computed in $O(n)$ time.

If A is not symmetric, but bounded and invertible, we can use $A^{-1} = (A^TA)^{-1}A^T$ where A^TA is SPD and banded, so we can truncate its inverse and obtain a product of banded matrices.
If A is not banded but sparse and SPD, then there exists $C > 0$ and $\rho \in (0,1)$ s.t. $|(A^{-1})_{i,j}| \leq C\rho^{d(i,j)}$ where $d(i,j)$ is the distance between i,j in the associated graph. If the diameter of the graph is small, we in fact cannot expect decay.

Using Schur complement,

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix} \sim \begin{pmatrix} A & 0 \\ C & \widetilde{S} \end{pmatrix}$$

is a good preconditioner where $\widetilde{S} = D - CFB$ where F is a good approximation of A^{-1}.

In eigenvalue problem, we usually get $Ax = \lambda Sx$ where S is the *mass matrix* corresponding to the scalar product of a basis $\{ \varphi_i \}$ of the space used to discretize a continuous space of functions. In this case, S has a low condition number, so we can use a Cholesky factorization to obtain a standard eigenvalue problem

$$L^{-1}AL^{-T}y = \lambda y, \qquad y = L^{-T}x.$$

13/03/19

If A is an M-matrix then the AINV algorithm (incomplete form of Gram-Schmidt algorithm in A-inner product applied to the unit basis vectors e_i, that produces an approximation of A^{-1}) converge.
The *comparison matrix* $\hat{A}$ referred to the matrix A is

$$\hat{A}_{i,j} = \begin{cases} |a_{i,j}| & i = j \\ -|a_{i,j}| & i \neq j \end{cases}$$

Definition 8.2. A matrix $A = (a_{i,j})$ such that the comparison matrix $\hat{A}$ is an M-matrix, is called **H-matrix**.

An M-matrix or a diagonally dominant matrix is always an H-matrix, and, as a rule of thumb, if a result holds for an M-matrix, then it holds for an H-matrix too.

9 *AINV

The Stabilized AINV algorithm trades speed for robustness. If A is an SPD matrix, breakdown are possible within AINV.

$$\overline{P}_i = a_i^T z_i^{(i-1)} = e_i^T A z_i^{(i-1)}$$

and when it is zero or negative, a breakdown occurs. In absence of dropping, $Z^T A Z = D = diag(P_1, \dots, P_n)$, so

$$P_i = z_i^T A z_i > 0$$

since A is SPD and $z_i = Ze_i \neq 0$. Therefore we can use

$$\overline{P}_i = (z_i^{(i-1)})^T A z_i^{(i-1)}$$

in the algorithm instead of the previous relation. When Z, A are sparse, then it can be compared with the cost of an inner product, but it requires careful programming.

When dealing with non-symmetric matrices we can use the *Nonsymmetric AINV* or also called *incomplete biconjugation algorithm.* Running AINV without stopping, we obtain an unit upper triangular Z and a diagonal D such that $AZ = LD$ with L unit lower triangular matrix (incomplete LU factorization). By uniqueness, $Z = U^{-1}$ where $A = LDU$ and it is guaranteed whenever the determinant of all the leading principle submatrices are not zero (same for LU).

If we run AINV on A^T, we obtain $A^T W = U^T D$ where D, U are the same used before and $W = L^{-T}$ thank to the uniqueness of LU factorization. as a consequence

$$A^{-1} = ZD^{-1}W^T$$

so $w_i^T A z_j = 0$ whenever $i \neq j$ and W, Z are called A-bi-conjugated. Notice that $h(x, y) = x^T Ay$ is not an inner product and in general it is not positive definite.

If we take an M-matrix or diagonally dominant the AINV algorithm does not drop. If A is positive definite ($A + A^T$ SPD) then we can stabilize AINV in a similar way as in the symmetric case, by replacing

$$\overline{P}_i = a_i^T z_i^{(i-1)} \to \overline{P}_i = (z_i^{(i-1)})^T A z_i^{(i-1)} > 0$$

since $2x^T Ax = x^T (A + A^T)x$. Similarly we do for the process on A^T.

When A is very indefinite, the process fails miserably.

10 FSAI

Given $A \in \mathbb{R}^{n\times n}$ an SPD matrix, fix a lower triangular sparsity pattern $\mathscr{S}_L$ (usually the diagonal positions are included). WE can compute a lower triangular $\hat{G}$ with sparsity pattern $\mathscr{S}_L$ such that $(\hat{G}A)_{i,j} = \delta_{i,j}$ for every $(i, j) \in \mathscr{S}_L$. Set

$$D = (diag(\hat{G}))^{-1}, \qquad G = D^{1/2}\hat{G}.$$

As a consequence, we have that

$$\mathrm{diag}(GAG^T) = I_n.$$

Clearly, G is an approximation for L in the Cholesky factorization of A. G can be also characterized of the unique solution of the problem

$$\min_G \|I - GL\|_F \qquad s.t. \qquad G \text{ has sparsity pattern } \mathscr{S}_L, \quad \mathrm{diag}(GAG^T) = I_n.$$

The equations $(\hat{G}A)_{i,j} = \delta_{i,j}$ can be written as

$$\sum_k \hat{g}_{i,k} a_{k,j} = \delta_{i,j}.$$

If we fix i, then let $x_k := \hat{g}_{i,k}$ so that

$$\sum_k a_{k,j} x_k = \delta_{i,j}$$

but $\hat{G}$ and A are sparse, so we can solve a large collection of sparse linear systems independently.

10.1 Application

Consider a "one particle Hamiltonian" setting with $H = -\frac{1}{2}\Delta + V$ in 3D. We want to find the eigenvalues

$$H\psi_n = \varepsilon_n \psi_n.$$

Numerically, we introduce a set of basis functions $\{\varphi_n\}_{n=1}^N$ and use the weak form of the problem. we approximate ψ_n as a combination of φ_k so that we obtain a mass or overlap matrix $S_{i,j} = (\varphi_i, \varphi_j)$ that is usually sparse. The final form will be $Hx = \lambda Sx$.

A method (Lowdin) prescribes to transform it into

$$S^{-1/2}HS^{-1/2}y = \lambda y, \qquad S^{1/2}x = y$$

where we can keep $S^{-1/2}$ sparse or approximately sparse by dropping small entries.

An other method (Inverse Cholesky) prescribe to take $S = LL^T$ and

$$L^{-1}HL^{-T}y = \lambda y$$

where usually we take $Z \sim L^{-1}$ through the AINV process.

11 Preconditioned GMRES (on the right)

To solve $Ax = b$ through a preconditioner M, the method prescribe to

- $r_0 = b - Ax$, $\beta = \|r_0\|_2$, $v_1 = r_0/\beta$
- For $j = 1 : m$ do
  - $w = AM^{-1}v_j$
  - For $i = 1 : j$ do
    - $h_{i,j} = w^t v_i$
    - $w = w - h_{i,j}v_j$
  - $h_{j+1,j} = \|w\|_2$, $v_{j+1} = w/h_{j+1,j}$
  - $V_m = [v_1, \dots, v_m]$, $\hat{H}_m = [h_{i,j}]_{1\leq i\leq m+1,\ \ 1\leq j\leq m}$
- $y_m = \arg\min_y \|\beta e_1 - \hat{H}_m y\|_2$, $x_m = x_0 + M^{-1}V_m y_m$
- If satisfied, then stop. Otherwise, let $x_0 = x_m$ and restart

We are minimizing

$$\|b - AM^{-1}y\|_2 \text{ over } x_0 + K_m(AM^{-1}, r_0).$$

If the preconditioning would be applied on the left, we'd have

$$\|M^{-1}(b - Ax)\|_2 \leq \|M^{-1}\|_2\|b - Ax\|_2$$

11.1 FGMRES

The Flexible GMRES (Saad, 1991) can accommodate a variable preconditioner. The algorithm is identical to preconditioned GMRES except for the computation $w = AM^{-1}v_j$ that is substituted with $z_j = M_j^{-1}v_j$ and $w = Az_j$, and the matrix V_m is substituted with

$$Z_m = [z_1, \dots, z_m]$$

and finally

$$x_m = x_0 + Z_m y_m.$$

We have auxiliary storage cost (almost double), and it is not a Krylov method any more, since it minimizes $\|r_m\|_2$ over

$$x_0 + Span\{z_1, \dots, z_m\}.$$

There are other Flexible methods like Flexible CG (Notay), Flexible QMR (Szyld, Vogel, 2000).

As an example, take the Oseen Problem

$$\begin{cases} -\nu\Delta u + (\overline{u}\cdot\nabla)u + \nabla p = f & \Omega \subset \mathbb{R}^3 \\ \nabla \cdot u = 0 & \Omega \subset \mathbb{R}^3 \\ Bu = g & \partial\Omega \end{cases}$$

where Δ is the vector Laplacian and $\overline{u}$ is a constant vector. It comes out from an iteration method that produces a contraction with factor $1 - O(\nu)$. A discretization of this problem produces a matrix $A = L + N$ with L associated to the Laplacian operator (usually symmetric) and N usually skew-symmetric, and a matrix B associated to the gradient and divergence operators.

$$\begin{pmatrix} L_1 + N_1 & & & B_1^T \\ & L_2 + N_2 & & B_2^T \\ & & L_3 + N_3 & B_3^T \\ B_1 & B_2 & B_3 & \end{pmatrix}$$

It is an indefinite non-symmetric matrix, but we can decompose it into

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} I & \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix} = \begin{pmatrix} I & \\ BA^{-1} & I \end{pmatrix} P$$

where $S = -BA^{-1}B^T$. we can use P as preconditioner and obtain

$$\begin{pmatrix} I & \\ BA^{-1} & I \end{pmatrix}$$

as preconditioned matrix, that has only eigenvalues 1.

References

[1] Bai, Golub, SIAM J. Matrix Anal. Appl. (2003)

[2] Alexandroff-Hopf *Topologie.*

[3] Richard S. Varga *Matrix Iterative Analysis*

[4] Young *Iterative Solution of Large Linear System*

[5] Greenbaum, A. and Pták, V. and Strakoš, Z. *Any Nonincreasing Convergence Curve is Possible for GMRES,* SIAM Journal on Matrix Analysis and Applications, 17, 3, 465–469, (1996)