Advanced Numerical Analysis

Vladimir Kazeev

May 14, 2025

1 Linear algebra

Vectors and matrices

In this section the field $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$. m and n always denote natural numbers.

Definition 2.1. Let V be a vector space over $\mathbb{F}$. A function $\|\cdot\| : V \to \mathbb{R}$ is called a norm on V if for all $v, w \in V$ and $\alpha \in \mathbb{F}$ the following properties hold:

1. $\|v\| \geq 0$
2. $\|v\| \neq 0 \quad \forall v \neq 0$
3. $\|\alpha v\| = |\alpha|\|v\|$
4. $\|v + w\| \leq \|v\| + \|w\|$

Example 2.2. Let $V = \mathbb{F}^n$

- $\|\cdot\|_\infty : V \to \mathbb{R} : \|v\|_\infty = \max_{i=1}^n |v_i| \quad \forall v \in V$
- $\|\cdot\|_p : V \to \mathbb{R} : \|v\|_p = \sqrt[p]{\sum_{i=1}^n |v_i|^p} \quad \forall v \in V$ and $p \in [1, \infty)$

Also $\lim_{p\to\infty}\|v\|_p = \|v\|_\infty$

Example 2.3. $V = \mathbb{F}^{m\times n}$. Then we define $\|\cdot\|_{\max}, \|\cdot\|_{\mathrm{F}} : \mathbb{F}^{m\times n} \to \mathbb{R}$ as follows:

- $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ (maximum absolute value norm / Chebyshev norm)
- $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j} |a_{ij}|^2}$ (Frobenius norm)

Proposition 2.4. *Let V, U be $\mathbb{F}$-vector spaces. $\mathcal{L}$ denotes the space of continuous (w.r.t. $\|\cdot\|_V$, $\|\cdot\|_U$) linear mappings from V to U. Then $\|\cdot\|: \mathcal{L} \to \mathbb{R}$ given by*

$$\|\varphi\| = \sup_{\substack{v \in V \\ \|v\|_V = 1}} \|\varphi(v)\|_U \quad \forall \varphi \in \mathcal{L}$$

is a norm.

Definition 2.5. The norm given in Proposition 2.4 is called the *operator norm* on $\mathcal{L}$ induced by the norms $\|\cdot\|_V$ and $\|\cdot\|_U$.

Definition 2.6. $V = \mathbb{F}^n, U = \mathbb{F}^m$. $\mathcal{L}$ is identified with $W = \mathbb{F}^{m \times n}$ using the standard basis.

$$\begin{array}{ccc} \varphi \in \mathcal{L} & \longleftrightarrow & A = \mathrm{Mat}(\varphi) \in W \\ & & \varphi(v) = Av \end{array}$$

Let $\|\cdot\|$ be the operator norm on $\mathcal{L}$ induced by $\|\cdot\|_V$ and $\|\cdot\|_U$. Then $\|\cdot\| \cdot \mathrm{Mat}^{-1} : \mathbb{F}^{m \times n} \to \mathbb{R}$ is called the *matrix operator norm* induced by $\|\cdot\|_V$ and $\|\cdot\|_U$.

Example 2.7. For $p, q \in [1, \infty]$, $W = \mathbb{F}^{m \times n}$.

$$\|\cdot\|_{p,q}: \ W \to \mathbb{R} \text{ given by } \|A\|_{p,q} = \max_{\substack{v \in \mathbb{F}^n \\ \|v\|_q = 1}} \|Av\|_p \quad \forall A \in W$$

is an (matrix) operator norm induced by $\|\cdot\|_p$ and $\|\cdot\|_q$.

Definition 2.8. For $p = q \in [1, \infty]$ we write $\|\cdot\|_{p,q} = \|\cdot\|_p$ and $\|\cdot\|_p$ is called the matrix p-norm on $\mathbb{F}^{m \times n}$.

Proposition 2.9. $\mathbb{F}^{n \times 1} \simeq \mathbb{F}^n$. *The matrix p-norm on $\mathbb{F}^{n \times 1}$ coincides with the vector p-norm on $\mathbb{F}^n$.*

Proposition 2.10. *For $A \in \mathbb{F}^{m \times n}$ the following holds:*

$$\begin{aligned} \|A\|_1 &= \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}| && \textit{(column sum norm)} \\ \|A\|_\infty &= \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| && \textit{(row sum norm)} \\ \|A\|_2 &= \sqrt{\lambda_{\max}(A^*A)} = \sigma_{\max}(A) && \textit{(spectral norm)} \\ &= \max_{\substack{u \in \mathbb{F}^m \\ v \in \mathbb{F}^n \\ \|u\|_2 = \|v\|_2 = 1}} u^* A v \end{aligned}$$

where $\lambda_{\max}$ is the largest eigenvalue and $\sigma_{\max}$ is the largest singular value of A.

Definition 2.11. $U = \mathbb{F}^{k\times m}, V = \mathbb{F}^{m\times n}, W = \mathbb{F}^{k\times n}$. Let $\|\cdot\|_U, \|\cdot\|_V, \|\cdot\|_W$ be norms on U, V, W respectively. These norms are called *consistent* (or *submultiplicative*) if

$$\|AB\|_W \leq \|A\|_U \|B\|_V \quad \forall A \in U, B \in V$$

For $U = V = W$ and $\|\cdot\|_U = \|\cdot\|_V = \|\cdot\|_W$ this reduces to

$$\|AB\|_W \leq \|A\|_W \|B\|_W \quad \forall A, B \in W.$$

Proposition 2.12.

- *p-norm on* $\mathbb{F}^{n\times n}$ *is consistent for* $p \in \{1, 2, \infty\}$
- *Frobenius norm on* $\mathbb{F}^{n\times n}$ *is consistent*
- *Chebyshev norm on* $\mathbb{F}^{n\times n}$ *is* not *consistent*

  e.g. $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$: $\|A \cdot A\|_{\max} = \|\begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}\|_{\max} = 2 \not\leq 1 = \|A\|_{\max}\|A\|_{\max}$

Proposition 2.13. $U \in \mathbb{F}^{n\times n}$ *invertible and* $\|\cdot\|$ *a norm on* $\mathbb{F}^{n\times n}$. *Consider* $\|\cdot\|_*, \|\cdot\|_{**}, \|\cdot\|_{***} : \mathbb{F}^{n\times n} \to \mathbb{R}$ *given by* $\|A\|_* = \|UA\|$, $\|A\|_{**} = \|AU\|$, $\|A\|_{***} = \|U^{-1}AU\|$. *These 3 functions are norms on* $\mathbb{F}^{n\times n}$ *and they are consistent if* $\|\cdot\|$ *is consistent.*

Eigenvalues of matrices

Definition 2.14. $A \in F^{n\times n}, \lambda \in \mathbb{F}$. If $\ker(A - \lambda I) \neq \{0\}$ then λ is called an eigenvalue of A and every non-zero vector from $\ker(A - \lambda I)$ is called an eigenvector of A associated with the eigenvalue λ.

Definition 2.15. $A \in \mathbb{F}^{n\times n}$. $\chi_A : \mathbb{F} \to \mathbb{F}$ given by $\chi_A(\lambda) = \det(A - \lambda I)$ $\forall \lambda \in \mathbb{F}$ is called *characteristic polynomial*.

Proposition 2.16. $A \in \mathbb{F}^{n\times n}$. χ_A *is an algebraic polynomial of degree* n *with leading coefficient* $(-1)^n$. *For any* $\lambda \in \mathbb{F}$, λ *is an eigenvalue of* A *if and only if* $\chi_A(\lambda) = 0$.

Definition 2.17. $A \in \mathbb{F}^{n\times n}, \lambda \in \mathbb{F}$ eigenvalue of A. The *algebraic multiplicity* of λ is the multiplicity of λ as a root of χ_A.

Definition 2.18. The *geometric multiplicity* of λ is the dimension of $\ker(A - \lambda I)$. λ is called *defective* if its geometric multiplicity is less than its algebraic multiplicity. If the geometric multiplicity of λ is equal to its algebraic multiplicity then λ is called *non-defective* eigenvalue of A.

Example. $A = I \in \mathbb{F}^{n\times n}$. $\chi_A(\lambda) = \det(I - \lambda I) = (1-\lambda)^n$. So $\lambda = 1$ is the only eigenvalue of I with algebraic multiplicity n. We have that $\dim(\ker(A - I)) = \dim(\ker(0)) = n$.

If $A \in \mathbb{F}^{n\times n}$ is a Jordan block of size $n \geq 2$, then there is only one eigenvalue, $\lambda = 1$, with algebraic multiplicity n and geometric multiplicity $\dim(\ker(A - I)) = 1 < n$. So $\lambda = 1$ is a defective eigenvalue of A.

Schur canonical form

Definition 2.19. $A \in \mathbb{C}^{n\times n}$. Assume that $Q \in \mathbb{C}^{n\times n}$ is unitary and that $T = Q^*AQ$ (which is equivalent to $A = QTQ^*$) is upper triangular. Then the factorization $A = QTQ^*$ is called a Schur decomposition of A and T is called a *Schur canonical form*.

Proposition 2.20. *In the context of the previous definition, the diagonal entries of T are the eigenvalues of A repeated according to their algebraic multiplicities.*

Theorem 2.21. *Let $A \in \mathbb{C}^{n\times n}$, $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A repeated according to their algebraic multiplicities. Then there exists a unitary matrix $Q \in \mathbb{C}^{n\times n}$ such that $T = Q^*AQ$ is upper triangular with diagonal entries $\lambda_1, \ldots, \lambda_n$.*

Proof. Let x_1 be a normalized eigenvector of A associated with λ_1. Consider a matrix $X = \left[\begin{array}{c|c} x_1 & X_1 \end{array}\right] \in \mathbb{C}^{n\times n}$ unitary (with $X_1 \in \mathbb{C}^{n\times(n-1)}$). Then

$$\begin{aligned} X^*AX &= \left[\begin{array}{c} x_1^* \\ \hline X_1 \end{array}\right] A \left[\begin{array}{c|c} x_1 & X_1 \end{array}\right] = \left[\begin{array}{c|c} x_1^*Ax_1 & x_1^*AX_1 \\ \hline X_1^*Ax_1 & X_1^*AX_1 \end{array}\right] \\ &= \left[\begin{array}{c|c} X^*Ax_1 & \begin{array}{c} x_1^*AX_1 \\ X_1^*AX_1 \end{array} \end{array}\right] = \left[\begin{array}{c|c} \lambda_1 & x_1^*AX_1 \\ \hline 0 & X_1^*AX_1 \end{array}\right] = \left[\begin{array}{c|c} \lambda_1 & t_1^* \\ \hline 0 & A_1 \end{array}\right] \end{aligned}$$

where $t_1 = X_1^*A^*x_1 \in \mathbb{C}^{n-1}$ and $A_1 = X_1^*AX_1 \in \mathbb{C}^{(n-1)\times(n-1)}$. For any $\lambda \in \mathbb{C}$, we have that $\det(A - \lambda I) = \det(X^*AX - \lambda I) = (\lambda_1 - \lambda)\det(A_1 - \lambda I)$. The following are equivalent for all $\lambda \in \mathbb{C}, m \in \mathbb{N}$:

(i) λ is a root of χ_A of multiplicity m

(ii) λ is a root of χ_{A_1} of multiplicity $\begin{cases} m & \text{if } \lambda \neq \lambda_1 \\ m-1 & \text{if } \lambda = \lambda_1 \end{cases}$

Then by Proposition 2.16 and Definition 2.17 $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A repeated according to their algebraic multiplicities. Assume that there exists a unitary

matrix $Q_1 \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $T_1 = Q_1^* A_1 Q_1$ is upper triangular with diagonal entries $\lambda_2, \ldots, \lambda_n$. Then $Q = X \left[\begin{array}{c|c} 1 & \\ \hline & Q_1 \end{array}\right]$ and $T = \left[\begin{array}{c|c} \lambda_1 & t_1^* Q_1 \\ \hline & T_1 \end{array}\right]$ fullfill the claim for A:

$$Q^* A Q = \left[\begin{array}{c|c} 1 & \\ \hline & Q_1^* \end{array}\right] X^* A X \left[\begin{array}{c|c} 1 & \\ \hline & Q_1 \end{array}\right] = \left[\begin{array}{c|c} 1 & \\ \hline & Q_1^* \end{array}\right] \left[\begin{array}{c|c} \lambda_1 & t_1^* \\ \hline & A_1 \end{array}\right] \left[\begin{array}{c|c} 1 & \\ \hline & Q_1 \end{array}\right] = \left[\begin{array}{c|c} \lambda_1 & t_1^* Q_1 \\ \hline & T_1 \end{array}\right]$$

For $n = 1$: $Q = [1], T = A$ fullfill the claim. By induction the claim holds (for any $n \in \mathbb{N}$). □

Remark. For square matrices $A \in \mathbb{C}^{n\times n}$ and $S \in \mathbb{C}^{n\times n}$ invertible, the following holds:

$$\begin{aligned} \det(S^{-1}AS) &= \det(S^{-1})\det(A)\det(S) = \det(A) \\ \chi_{S^{-1}AS}(\lambda) &= \det(S^{-1}AS - \lambda I) = \det(S^{-1}(A - \lambda I)S) \\ &= \det(S^{-1})\det(A - \lambda I)\det(S) = \det(A - \lambda I) = \chi_A(\lambda) \end{aligned}$$

That is, the eigenvalues of A and $S^{-1}AS$ coincide.

Theorem (Spectral theorem)**.** *Let $n \in N$. $A \in \mathbb{F}^{n\times n}$ is diagonalizable by ...*

$\mathbb{F} = \mathbb{C}$: ... an unitary similarity transformation $\Leftrightarrow$ A is normal

$\mathbb{F} = \mathbb{R}$: ... an orthogonal similarity transformation $\Leftrightarrow$ A is symmetric

Remark. There can be non-hermitian normal matrices, e.g. $A = \begin{pmatrix} i & 1 \\ 0 & i \end{pmatrix}$

Remark 3.1. For $A \in \mathbb{C}^{n\times n}$. By theorem 2.21 A has a Schur form T. It is easy to check that A is normal if and only if T is normal.

$$(A = QTQ^*, \text{ so } A^*A - AA^* = Q(T^*T - TT^*)Q^* = 0 \Leftrightarrow T^*T = TT^*)$$

It is left as an exercise to show that

$$T \text{ is normal} \Leftrightarrow T \text{ is diagonal.}$$

So the Schur form is a generalization of diagonalization by unitary similarity transformations for normal matrices to abitrary matrices.

Spectral radius of a matrix: The behavior of matrix powers

Definition 3.2. For $A \in \mathbb{F}^{n\times n}$ the set of eigenvalues of A is called the *spectrum* of A. We will denote the spectrum of A by $\boldsymbol{\lambda}(A)$. (i.e., $\boldsymbol{\lambda}(A)$ is the zero set of χ_A).

Definition 3.3. For $A \in \mathbb{F}^{n\times n}$, $\rho(A) = \max_{\lambda\in\boldsymbol{\lambda}(A)} |\lambda|$ is called the *spectral radius* of A. Any $\lambda \in \boldsymbol{\lambda}(A)$ with $|\lambda| = \rho(A)$ is called a *dominant eigenvalue* of A.

Lemma 3.4. *Let $\|\cdot\|$ be a consistent norm on $\mathbb{F}^{n\times n}$. Then $\rho(A) \leq \|A\|$ for all $A \in \mathbb{F}^{n\times n}$.*

Proof. (Auxiliary result: Consider $y \in \mathbb{C}^n$ non-zero. Let $\|\cdot\|_* : \mathbb{F}^n \to \mathbb{R}$ be given by $\|x\|_* = \|xy^*\| \quad \forall x \in \mathbb{F}^n$. Then $\|Ax\|_* = \|(Ax)y^*\| = \|A(xy^*)\| \leq \|A\|\|xy^*\| = \|A\|\|x\|_*$. So the norms $\|\cdot\|_*, \|\cdot\|, \|\cdot\|_*$ are consistent norms.)

Let $\|\cdot\|_*$ be a norm on $\mathbb{F}^{n\times n}$ with which $\|\cdot\|$ is consistent (i.e. $\|\cdot\|_*, \|\cdot\|, \|\cdot\|_*$ are consistent). Let $\lambda \in \mathbb{F}$ be an eigenvalue of A and $x \in \mathbb{F}^n$ an associated eigenvector of unit length w.r.t. $\|\cdot\|_*$. Then $\|A\| = \|A\|\|x\|_* \geq \|Ax\|_* = \|\lambda x\|_* = |\lambda|\|x\|_* = |\lambda|$. $\square$

Lemma 3.5. *Let $A \in \mathbb{F}^{n\times n}$ and $\varepsilon > 0$. Then there exists a consistent norm $\|\cdot\|_{A,\varepsilon}$ on $\mathbb{F}^{n\times n}$ such that $\|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon$.*

If the dominant eigenvalues of A are non-defective, then there exists a consistent norm $\|\cdot\|_A$ on $\mathbb{F}^{n\times n}$ such that $\|A\|_A = \rho(A)$.

Proof. By theorem 2.21, A has a Schur decomposition $A = QTQ^*$ with $Q \in \mathbb{C}^{n\times n}$ unitary and $T \in \mathbb{C}^{n\times n}$ upper triangular. Let $\Lambda = \operatorname{diag}(T)$ and $U = T - \Lambda = \operatorname{offdiag}(T)$ (so $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ in the context of theorem 2.21, U is strictly upper triangular).

For $\eta > 0$ consider $D_\eta = \operatorname{diag}(\eta^0, \eta^1, \ldots, \eta^{n-1}) \in \mathbb{C}^{n\times n}$ then, for all $i, j \in \{1, \ldots, n\}$, we have

$$(D_\eta^{-1} U D_\eta)_{ij} = \eta^{1-i} U_{ij} \eta^{j-1} = \begin{cases} 0 & \text{if } i \geq j \\ \eta^{j-i} U_{ij} & \text{if } i < j \end{cases}$$

So there exists $\eta_* > 0$ such that $\|D_{\eta_*}^{-1} U D_{\eta_*}\|_\infty < \varepsilon$. Let $D = D_{\eta_*}$. Then

$$\begin{aligned} \|D^{-1}Q^*AQD\|_\infty &= \|D^{-1}\Lambda D + D^{-1}UD\|_\infty = \|\Lambda + D^{-1}UD\|_\infty \\ &\leq \|\Lambda\|_\infty + \|D^{-1}UD\|_\infty < \rho(A) + \varepsilon \end{aligned}$$

Let us define $\|\cdot\|_{A,\varepsilon} : \mathbb{C}^{n\times n} \to \mathbb{R}$ by $\|B\|_{A,\varepsilon} = \|D^{-1}Q^*BQD\|_\infty$. By proposition 2.13 $\|\cdot\|_{A,\varepsilon}$ is a consistent norm on $\mathbb{C}^{n\times n}$. On the other hand, $\|\cdot\|_A < \rho(A) + \varepsilon$.

For the second claim, let us assume that $\lambda_1, \ldots, \lambda_k$ with $k \in \{1, \ldots, n\}$ are the dominant eigenvalues of A (i.e. $|\lambda_1| = \ldots = |\lambda_k| = \rho(A) > |\lambda_{k+1}|, \ldots, |\lambda_n|$) and that they are non-defective.

If $\rho(A) = 0$, then $\lambda_1 = \ldots = \lambda_n = 0$, so 0 is a non-defective eigenvalue of A with algebraic multiplicity = geometric multiplicity = n, so $A = 0$. Then any consistent norm fullfills the claim.

If $k = n$, all eigenvalues are non-defective. Then A is diagonalizable, i.e. there is $S \in \mathbb{C}^{n\times n}$ invertible such that $A = S\Lambda S^{-1}$ with $\Lambda = S^{-1}AS$ diagonal. Let $\|\cdot\|_A : \mathbb{C}^{n\times n} \to \mathbb{R}$ be given by $\|B\|_A = \|S^{-1}BS\|_\infty$ for all $B \in \mathbb{C}^{n\times n}$. As discussed earlier, $\|\cdot\|_A$ is a consistent norm and $\|A\|_A = \|\Lambda\|_\infty = \rho(A)$.

For the remainder of the proof, assume that $k < n$. Let $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_k) \in \mathbb{C}^{k\times k}$ and $\Lambda_2 = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n) \in \mathbb{C}^{(n-k)\times(n-k)}$. Then $\Lambda = \left[\begin{array}{c|c} \Lambda_1 & \\ \hline & \Lambda_2 \end{array}\right] \in \mathbb{C}^{n\times n}$. We consider a Schur decomposition $A = QTQ^*$ with Q unitary and T upper triangular with $\mathrm{diag}(T) = \Lambda$. Partition T as $T = \left[\begin{array}{c|c} T_1 & T_{12} \\ \hline & T_2 \end{array}\right]$ with $T_{11} \in \mathbb{C}^{k\times k}$. We have:

(i) Every dominant eigenvalue of A is not an eigenvalue of T_2 but is an eigenvalue of T_1.

(ii) For all $\lambda \in \{\lambda_1, \ldots, \lambda_k\}$, $T_2 - \lambda I$ is invertible, so we have $\dim(\ker(A - \lambda I)) = \dim(\ker(T - \lambda I)) = \dim(\ker(T_1 - \lambda I))$.

So T_1 is diagonalizable: $\exists S_1 \in \mathbb{C}^{k\times k}$ invertible such that $S_1^{-1}T_1S_1 = \Lambda_1$. Let us consider the matrix $S = \left[\begin{array}{c|c} S_1 & \\ \hline & I_2 \end{array}\right]$ with $I_2 \in \mathbb{C}^{(n-k)\times(n-k)}$ the identity matrix. We have

$$S^{-1}Q^*AQS = S^{-1}TS = \left[\begin{array}{c|c} \Lambda_1 & \\ \hline & \Lambda_2 \end{array}\right] + \left[\begin{array}{c|c} 0 & T_{12} \\ \hline 0 & U_2 \end{array}\right]$$

where $U_2 = T_2 - \Lambda_2 = \mathrm{offdiag}(T_2)$ is strictly upper triangular.

Consider $\eta > 0, D = \mathrm{diag}(\eta^0, \ldots, \eta^{n-1}) = \left[\begin{array}{c|c} D_1 & \\ \hline & D_2 \end{array}\right]$ with $D_1 \in \mathbb{C}^{k\times k}$.

$$D^{-1}S^{-1}Q^*AQSD = \left[\begin{array}{c|c} \Lambda_1 & \\ \hline & \Lambda_2 \end{array}\right] + \left[\begin{array}{c|c} 0 & Z_{12} \\ \hline 0 & Z_2 \end{array}\right]$$

where $Z_{12} = D_1^{-1}T_{12}D_2$ and $Z_2 = D_2^{-1}U_2D_2$. We will consider $\|\cdot\|_A : \mathbb{C}^{n\times n} \to \mathbb{R}$ given by $\|B\|_A = \|D^{-1}S^{-1}Q^*BQSD\|_1$ for all $B \in \mathbb{C}^{n\times n}$. Again, $\|\cdot\|_A$ is a consistent norm.

Block Z_2: U_2 is strictly upper triangular, so Z_2 is strictly upper triangular. For all $i, j \in \{1, \dots, n-k\}$ such that $i < j$ we have

$$(Z_2)_{ij} = \eta^{j-i}(U_2)_{ij} \xrightarrow{\eta \to 0} 0$$

Block Z_{12}: For all $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, n-k\}$ we have

$$(Z_{12})_{ij} = \frac{\eta^{k+j}}{\eta^i}(T_{12})_{ij} = \eta^{k-i}\eta^j(T_{12})_{ij} \xrightarrow{\eta \to 0} 0$$

So $\left\| \left[\frac{Z_{12}}{Z_2} \right] \right\|_1 \xrightarrow{\eta \to 0} 0$. So there exists $\eta_* > 0$ such that $\left\| \left[\frac{Z_{12}}{Z_2} \right] \right\|_1 < \frac{1}{2}(\|\Lambda_1\|_1 - \|\Lambda_2\|_1)$. For D defined with $\eta = \eta_*$ we have

$$\|A\|_A = \|D^{-1}S^{-1}Q^*AQSD\|_1 = \left\| \left[\begin{array}{c|c} \Lambda_1 & \\ \hline & \Lambda_2 \end{array} \right] + \left[\begin{array}{c|c} 0 & Z_{12} \\ \hline 0 & Z_2 \end{array} \right] \right\|_1$$

$$= \max\left\{ \|\Lambda_1\|_1, \left\| \left[\frac{Z_{12}}{\Lambda_2 + Z_2} \right] \right\|_1 \right\} = \|\Lambda_1\|_1 = \rho(A)$$

where we used

$$\left\| \left[\frac{Z_{12}}{\Lambda_2 + Z_2} \right] \right\|_1 \leq \|\Lambda_2\|_1 + \left\| \left[\frac{Z_{12}}{Z_2} \right] \right\|_1 < \|\Lambda_2\|_1 + \frac{1}{2}(\|\Lambda_1\|_1 - \|\Lambda_2\|_1)$$

$$= \frac{1}{2}(\|\Lambda_1\|_1 + \|\Lambda_2\|_1) < \|\Lambda_1\|_1 = \|\Lambda\|_1 = \rho(A)$$

□

Lemma 3.6. *Let $A \in \mathbb{C}^{n \times n}$ and $\|\cdot\|$ be a norm on $\mathbb{C}^{n \times n}$, $\varepsilon > 0$, then there exists $c > 0$ (depending on n, $\|\cdot\|$, but not on A or ε) and $C > 0$ (depending on n, $\|\cdot\|$, A and ε) such that*

$$c\rho^k \leq \|A^k\| \leq C(\rho + \varepsilon)^k \quad \forall k \in \mathbb{N} \qquad \text{where } \rho = \rho(A).$$

If the dominant eigenvalues of A are non-defective, the same holds with $\varepsilon = 0$.

Proof. (i) For the lowerbound, consider a dominant eigenvalue $\lambda \in \mathbb{C}$ of A and a corresponding eigenvector, s.t. $\|x\|_2 = 1$. Then $\|A^k x\|_2 = \|\lambda^k x\|_2 = |\lambda|^k \|x\|_2 = \rho^k$. By the equivalence of norms, there exists $c > 0$ such that $c\|B\|_2 \leq \|B\|$ for all $B \in \mathbb{C}^{n\times n}$. So $\|A^k\| \geq c\|A^k x\|_2 = c\rho^k$ for all $k \in \mathbb{N}$.

(ii) For the upperbound: Let $\|\cdot\|_{A,\varepsilon}$ be a consistent norm on $\mathbb{C}^{n\times n}$ such that $\|A\|_{A,\varepsilon} \leq \rho(A) + \varepsilon$ (Lemma 3.5). Consistency yields $\|A^k\|_{A,\varepsilon} \leq \|A\|_{A,\varepsilon}^k$, so $\|A^k\|_{A,\varepsilon} \leq (\rho + \varepsilon)^k$ for all $k \in \mathbb{N}$.

By the equivalence of norms, there exists $C > 0$ (depending on n, $\|\cdot\|$, A and ε) such that $\|B\| \leq C\|B\|_{A,\varepsilon}$ for all $B \in \mathbb{C}^{n\times n}$. So $\|A^k\| \leq C\|A^k\|_{A,\varepsilon} \leq C(\rho+\varepsilon)^k$ for all $k \in \mathbb{N}$.

(iii) If the dominant eigenvalues of A are non-defective, the same holds with $\varepsilon = 0$.

□

Remark. Taking k-th root and limit $k \to \infty$ in the previous lemma yields

$$\rho(A) \leq \lim_{k\to\infty} \|A^k\|^{1/k} \leq \rho(A) + \varepsilon$$

if the middle limit exists.

Definition 3.7. $A \in \mathbb{C}^{n\times n}$ is called

- row-wise (non-)strictly diagonally dominant if

$$|a_{ii}| > (\geq) \sum_{j\in\{1,\dots,n\}\setminus\{i\}} |a_{ij}| \quad \forall i \in \{1,\dots,n\}$$

- column-wise (non-)strictly diagonally dominant if

$$|a_{jj}| > (\geq) \sum_{i\in\{1,\dots,n\}\setminus\{j\}} |a_{ij}| \quad \forall j \in \{1,\dots,n\}$$

Theorem 3.8 (Levy-Desplanques)**.** *For any $A \in \mathbb{C}^{n\times n}$, if A is row-wise or column-wise strictly diagonally dominant, then it is invertible.*

TODO: Proof

Definition 3.9 (Gerschgorin dishes)**.** Let $A \in \mathbb{C}^{n\times n}$. For each $k \in \{1, \ldots, n\}$

$$\mathcal{R}_k = \left\{ z \in \mathbb{C} : \ |z - a_{kk}| \leq \sum_{j\in\{1,\ldots,n\}\setminus\{k\}} |a_{kj}| \subset \mathbb{C} \right\}$$
$$\mathcal{C}_k = \left\{ z \in \mathbb{C} : \ |z - a_{kk}| \leq \sum_{i\in\{1,\ldots,n\}\setminus\{k\}} |a_{ik}| \subset \mathbb{C} \right\}$$

are called the k-th row-wise and column-wise *Gerschgorin dishes* of A.

Theorem 3.10 (1st Gerchgorin theorem)**.** *Let $A \in \mathbb{C}^{n\times n}$. Then*

$$\boldsymbol{\lambda}(A) \subseteq \bigcup_{k=1}^{n} \mathcal{R}_k, \ \boldsymbol{\lambda}(A) \subseteq \bigcup_{k=1}^{n} \mathcal{C}_k.$$

TODO: Proof

2 Iterative methods for linear systems

We consider the problem of finding a solution $x \in \mathbb{F}^n$ of the linear system $Ax = b$ with $A \in \mathbb{F}^{n\times n}$ the matrix of the linear system and $b \in \mathbb{F}^n$ the right-hand side. A and b is the data of the problem. We assume that A is invertible, so the system has a unique solution $x = A^{-1}b$.

Iterative methods (in contrast to direct methods) perform a sequence of computation steps (iterations) that produce an *approximation* (an approximate solution, an iterate) to the the exact solution x. The main question is:

> How close is the approximate solution to the exact solution?

$k \in \mathbb{N}$ will denote the iteration index. An iterative method produces iterates $(x_1, x_2, \ldots) = (x_k)_{k\in\mathbb{N}}$ from an *initial guess* (initial approximation) $x_0 \in \mathbb{F}^n$. We are interested in the errors $e_k = x_k - x \in \mathbb{F}^n$. So the above question is how $\|e_k\|$ behaves w.r.t. $k \in \mathbb{N}$, where $\|\cdot\|$ is a norm on $\mathbb{F}^n$.

Linear iterative methods

$e_k = T_k \cdot e_{k-1}$ for all $k \in \mathbb{N}$, where $T_k \in \mathbb{F}^{n\times n}$ is the *iteration matrix* at iteration k. For some methods we have $T_k = T$ for all $k \in \mathbb{N}$, where $T \in \mathbb{F}^{n\times n}$ – those are *stationary methods*.

Let $A_1, A_2 \in \mathbb{F}^{n\times n}$ be such that $A = A_1 + A_2$. Then the linear system $Ax = b$ can be rewritten as

$$\begin{aligned} A_1x + A_2x &= b \\ \Leftrightarrow A_1x &= b - A_2x \\ \Leftrightarrow A_1x_k &= b - A_2x_{k-1} = b - Ax_{k-1} + A_1x_{k-1} \\ \Leftrightarrow x_k &= A_1^{-1}(b - A_2x_{k-1}) \end{aligned}$$

For each $k \in \mathbb{N}$, let x_k be given by the above equation. Also we can see:

$$\begin{aligned} x_k &= x_{k-1} + A_1^{-1}(b - Ax_{k-1}) \\ &= (I - A_1^{-1}A)x_{k-1} + A_1^{-1}b \\ &= (I - A_1^{-1}A)(x + e_{k-1}) + A_1^{-1}b = x + (I - A_1^{-1}A)e_{k-1} \end{aligned}$$

$$\Rightarrow e_k = (I - A_1^{-1}A)e_{k-1} = -A_1^{-1}A_2e_{k-1}$$

Consider $A \in \mathbb{F}^{n\times n}$ invertible. Let $D, L, U \in$ be the diagonal, strictly lower triangular and upper triangular part of A respectively. I.e.

$$D_{ij} = \begin{cases} A_{ij} & \text{if } i = j \\ 0 & \text{else} \end{cases}, \quad L_{ij} = \begin{cases} A_{ij} & \text{if } i > j \\ 0 & \text{else} \end{cases}, \quad U_{ij} = \begin{cases} A_{ij} & \text{if } i < j \\ 0 & \text{else} \end{cases}$$

Then $A = D + L + U$.

Jacobi iteration	Gauss-Seidel iteration
$A_1 = D, A_2 = L + U$	$A_1 = D + L, A_2 = U$
A_1 is invertible $\Leftrightarrow$ the diagonal entries of A are all non-zero	
$x_k = D^{-1}(b - (L + U)x_{k-1})$	$x_k = (D + L)^{-1}(b - Ux_{k-1})$
$e_k = x_k - x = Te_{k-1}$ with	$e_k = x_k - x = Te_{k-1}$ with
$T = -D^{-1}(L + U)$	$T = -(D + L)^{-1}U$

So $\|e_k\| = \|T^k e_0\| \leq \|T\|^k \|e_0\|$ for a consistent norm.

Let us denote the iteration matrices as follows

$$J = -D^{-1}(L + U), \quad G = -(D + L)^{-1}U.$$

We assume that A (and therefore D and $D + L$) has no zeroes on the diagonal.

Theorem 4.1. *Let $A \in \mathbb{C}^{n\times n}$ be row-wise and column-wise strictly diagonally dominant. Then D and $D + L$ are invertible and $\rho(J) < 1$ and $\rho(G) < 1$.*

TODO: Proof

Corollary 4.2. *Let $A \in \mathbb{C}^{n\times n}$ be row-wise and column-wise strictly diagonally dominant. Then the linear system $Ax = b$ has a unique solution $x \in \mathbb{C}^n$ and the Jacobi and Gauss-Seidel iterations converge to x for any initial guess $x_0 \in \mathbb{C}^n$. Furthermore, for any norm $\|\cdot\|$ on $\mathbb{C}^n$ and for either method, the iterates $(x_k)_{k\in\mathbb{N}}$ satisfy with any $\varepsilon \in (0, 1 - \rho)$, where $\rho = \rho(T) < 1$ and C positive constant:*

$$\|x_k - x\| \leq C(\rho + \varepsilon)^k \|x_0 - x\|$$

with $x_{k-1} - x = T^k(x_0 - x)$.

Example 4.3. **TODO**

Towards generalization: consider a splitting $A = P_k - N_k$ with $P_k \in \mathbb{F}^{n\times n}$ invertible and the associated iteration

$$\boxed{x_k} = P_k^{-1}(N_k x_{k-1} + b) = P_k^{-1}N_k x_{k-1} + P_k^{-1}b = \boxed{B_k x_{k-1} + P_k^{-1}b}$$

with $B_k = P_k^{-1}N_k = I - P_k^{-1}A$, the kth *iteration matrix* for all $k \in \mathbb{N}$. Also note

$$\boxed{x_k} = (I - P_k^{-1}A)x_{k-1} + P_k^{-1}b = P_k^{-1}(P_k x_{k-1} - Ax_{k-1} + b) = \boxed{x_{k-1} + P_k^{-1}r_{k-1}}$$

where $r_{k-1} = b - Ax_{k-1} = Ae_{k-1}$ is the *residual vector* at step k. Also

$$\boxed{e_k} = x_k - x = x_{k-1} - x + P_k^{-1}A(x_{k-1} - x) = (I - P_k^{-1}A)(x_{k-1} - x) = \boxed{B_k e_{k-1}}$$

If $\|B_k\| \leq \rho$ for all $k \in \mathbb{N}$ and for some $\rho \in (0,1)$ and a norm $\|\cdot\|$ on $\mathbb{F}^n$, then this yields the exponential convergence of the iterates $(x_k)_{k\in\mathbb{N}}$ to the exact solution x.

Stationary linear iterative schemes

P_k is the same for all $k \in \mathbb{N}$ (so are N_k and B_k).

When is such a method efficient?

- $U \mapsto P^{-1}U$ should be easy to evaluate
- P should approximate A in the sense that $P^{-1}A \approx I$ (precisely, $\rho(B)$, should be as small as possible)

P is often called a preconditioner for A.

Example 5.1.

- Jacobi iteration: $U \mapsto D^{-1}U$ takes $\mathcal{O}(n)$ operations
- Gauss-Seidel iteration: $U \mapsto (D+L)^{-1}U$ takes $\mathcal{O}(n^2)$ operations

After K iterations $\sim Kn$ operations for Jacobi and $\sim Kn^2$ operations for Gauss-Seidel, which is $\ll n^3$ if $K \ll n$

Example 5.2 (related stationary linear iterative schemes)**.**

- Backwards Gauss-Seidel method: $P = D + U$. The analysis and behavior are analogous to the Gauss-Seidel method.
- Jacobi over-relaxation (JOR)

  $P = \frac{1}{\omega}D,\ \omega > 0$ is a *relaxation parameter* ("learning rate")

$$x_k = x_{k-1} + \omega D^{-1}r_{k-1}$$

- Successive over-relaxation (SOR)

  $P = \frac{1}{\omega}D + L, \quad x_k = x_{k-1} + \omega(D + \omega L)^{-1} r_{k-1} \; \forall k \in N$

  (i) does not change for any $\omega \in (0, 2]$

  (ii) If A is symmetric positive definite, the SOR iteration converges for any $\omega \in (0, 2)$.

Remark 5.3 (Any b? Any x_0?). The behaviour of the $(x_k)_{k\in\mathbb{N}}$ is determined by $(P_k)_{k\in\mathbb{N}}$ and

- the initial residual or
- the initial error

For any RHS $b \in \mathbb{F}^n$ and for all $k \in \mathbb{N}$ we have

$$\begin{aligned} x_k &= x_{k-1} + P_k^{-1} r_{k-1} \\ r_k &= b - Ax_k = b - Ax_{k-1} - AP_k^{-1} r_{k-1} = (I - AP_k^{-1}) r_{k-1} \\ e_k &= A^{-1} r_k = (I - P_k^{-1} A) e_{k-1} \end{aligned}$$

$b \leftarrow b - Ax_0$ and $x_0 \leftarrow 0$ (does not effect initial residual and the initial error)

Theorem 5.4. *$A \in \mathbb{C}^{n\times n}$. A* stationary *linear iterative scheme associated with $P \in \mathbb{C}^{n\times n}$ invertible converges for a zero initial guess to the solution of $Ax = b$ with any $b \in \mathbb{C}^n$ if and only if $\rho(I - P^{-1}A) < 1$.*

Proof. Let $\rho = \rho(I - P^{-1}A)$. Consider a norm $\|\cdot\|$ on $\mathbb{C}^n$ and the corresponding operator norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$. Then $\|B^k u\| \le \|B\|^k \|u\|$ for all $u \in \mathbb{C}^n$. If $\rho < 1$, then the upper bound given by 3.6 with $\varepsilon = \frac{1}{2}(1 - \rho)$ (s.t. $\rho + \varepsilon < 1$) yields that the method converges exponentially to any RHS $b \in \mathbb{C}^n$.

$\rightarrow$ if $\rho \ge 1$, consider an eigenvector $u \in \mathbb{C}^n$ of the iteration matrix B corresponding to a dominant eigenvalue λ. Then $\|B^k u\|_k = \rho^k \|u\|$ for all $k \in \mathbb{N}$. So $B^k u \nrightarrow 0$ as $k \to \infty$. So the method does not converge when $e_0 = u$ (i.e. for $x_0 = 0$ and $b = Au$). □

Remark 5.5. For stationary methods we can first precondition the problem and then consider an iterative scheme with preconditioned I.

Define $\tilde{A} = P^{-1}A$ and $\tilde{b} = P^{-1}b$. Then for all $x \in \mathbb{C}^n$ we have $Ax = b \Leftrightarrow \tilde{A}x = \tilde{b}$. The residuals, for any $x \in \mathbb{C}^n$ are $r = b - Ax$, $\tilde{r} = \tilde{b} - \tilde{A}x = P^{-1}r$. A linear iterative

scheme for the original system with preconditioner P

$$x_k = x_{k-1} + P^{-1} r_{k-1}$$

is equivalent to a linear iterative scheme for the preconditioned system with preconditioner I:

$$\tilde{x}_k = \tilde{x}_{k-1} + \tilde{r}_{k-1}$$

that is, $\tilde{x}_k = x_k \ \forall k \in \mathbb{N}$ if $\tilde{x}_0 = x_0$.

Richardson iteration

Definition 5.6. $A \in \mathbb{F}^{n\times n}, b \in \mathbb{F}^n$. The stationary Richardson method for $Ax = b$ with an initial guess $x_0 \in \mathbb{F}^n$ is given by

$$x_k = x_{k-1} + \alpha r_{k-1} \quad \forall k \in \mathbb{N}$$

where $r_{k-1} = b - Ax_{k-1}$ is the residual vector and $\alpha \in \mathbb{R} \setminus \{0\}$ is a relaxation parameter.

The non-stationary Richardson method is given by

$$x_k = x_{k-1} + \alpha_k r_{k-1} \quad \forall k \in \mathbb{N}$$

with $\alpha_k \in \mathbb{R} \setminus \{0\}$.

The iteration matrix is given by $T_k = I - \alpha_k A$.

Theorem 5.7. *Let $A \in \mathbb{C}^{n\times n}$. The stationary Richardson method with zero initial guess converges (for any linear system $Ax = b$) if and only if*

$$\frac{2\mathrm{Re}\lambda}{\alpha|\lambda|^2} > 1 \quad \forall \lambda \in \boldsymbol{\lambda}(A)$$

Proof. **TODO**
First, the iteration converges $\forall b \in \mathbb{C}^n$ (after Theorem 5.7) if and only if

$$\rho(B) < 1 \ \Leftrightarrow \ \begin{cases} \alpha > 0 \\ \alpha < \frac{2}{\lambda_1} \end{cases} \ \Leftrightarrow \ \alpha \in (0, \frac{2}{\alpha_1})$$

Consider $\alpha \in (0, \frac{2}{\alpha_1})$ and $\Phi_\alpha : \ \mathbb{R} \to \mathbb{R}$ given by $\Phi_\alpha(\lambda) = |1 - \alpha\lambda| \forall \lambda \in \mathbb{R}$ □

Theorem 5.8. *Let $A \in \mathbb{C}^{n\times n}$, $\boldsymbol{\lambda}(A) = \{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{R}$ with $\lambda_1 \geq \ldots \geq \lambda_n > 0$. Then the stationary Richardson scheme for $Ax = b$ converges with any $b \in \mathbb{C}^n$ if and only if $\alpha \in (0, \frac{2}{\lambda_1})$.*

Furthermore, $\alpha_* = \frac{2}{\lambda_1+\lambda_n}$ is the unique minimizer of $\rho(B)$ with respect to $\alpha \in \mathbb{R}\setminus\{0\}$, and it yields $\rho(B) = \frac{\kappa-1}{\kappa+1} = \frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}$.

($\kappa = \frac{\lambda_1}{\lambda_n}$ is the spectral condition number of A)

Proof. **TODO**

□

Iteration methods for systems with symmetric positive matrices

$\boldsymbol{\lambda}(A) = \{\lambda_k\}_{k=1}^n \subseteq \mathbb{R}$ with $\lambda_1 \geq \ldots \geq \lambda_n$. Then A is positive definite if and only if

$$\min_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^*Ax}{x^*x} = \lambda_n > 0$$

$Ax = b$ for $x \in \mathbb{R}^n$ is the (necessary and !sufficient!) *1st order optimality condition* for minimizing $J : \mathbb{R}^n \to \mathbb{R}$ given by

$$J(x) = \frac{1}{2}x^TAx - b^Tx \quad \forall x \in \mathbb{R}^n$$

Note:

$$\begin{aligned}\nabla J(x) &= Ax - b = -r(x) \quad \forall x \in \mathbb{R}^n\\ \nabla^2 J(x) &= A \quad \forall x \in \mathbb{R}^n\end{aligned}$$

But A is positive definite $\Leftrightarrow$ J is strictly convex $\Leftrightarrow$ J has a unique minimum (sufficient condition).

For any $e \in \mathbb{R}^n$ and the exact solution x, we have

$$\begin{aligned}J(x+e) - J(x) &= \frac{1}{2}(x+e)^TA(x+e) - b^T(x+e) - \frac{1}{2}x^TAx + b^Tx\\ &= \frac{1}{2}e^TAe + \underbrace{x^TAe - b^Te}_{=-e^Tr(x)} \underbrace{=}_{r(x)=0} \frac{1}{2}e^TAe\end{aligned}$$

Since A is SPD the function $\|\cdot\|_A:\ \mathbb{R}^n \to \mathbb{R}$ given by $\|u\|_A = \sqrt{u^T A u}$ is a norm on $\mathbb{R}^n$. Then

$$J(x+e) - J(x) = \frac{1}{2}\|e\|_A^2$$

Gradient-type methods for systems with SPD matrices.

Solve $Ax = b$, $b \in \mathbb{R}^n$ with $A \in \mathbb{R}^{n\times n}$ SPD and $x_0 \in \mathbb{R}^n$ an initial guess. The gradient descent method is given by

$$x_k = x_{k-1} - \alpha_k \nabla J(x_{k-1}) \quad \forall k \in \mathbb{N},$$

where $\nabla J(x_{k-1}) = -r(x_{k-1})$.

For the steepest gradient method, choose α_k so as to minimize J along $\nabla J(x_{k-1})$:

$$J(x_{k-1} + \alpha_k r_k) = \frac{1}{2}\alpha_k^2 r_{k-1}^T A r_{k-1} + \alpha_k r_{k-1}^T A x_{k-1} + \frac{1}{2} x_{k-1}^T A x_{k-1} - \alpha_k b^T r_{k-1} - b^T x_{k-1} \quad \forall \alpha_k \in \mathbb{N}$$

The optimal value of α_k is given by

$$\alpha_k = \frac{(b - Ax_{k-1})^T r_{k-1}}{r_{k-1}^T A r_{k-1}} = \frac{r_{k-1}^T r_{k-1}}{r_{k-1}^T A r_{k-1}} = \frac{\|r_{k-1}\|_2^2}{\|r_{k-1}\|_A^2}$$

Lemma 6.1. *$A \in \mathbb{R}^{n\times n}$ is SPD. Then there is a unique SPD matrix $B \in \mathbb{R}^{n\times n}$ such that $B^2 = A$.*

Proof. A is SPD $\Rightarrow\ \exists \Lambda \in \mathbb{R}^{n\times n}$ diagonal, $Q \in \mathbb{R}^{n\times n}$ orthogonal s.t. $A = Q^*\Lambda Q$ with positive diagonal entries. W.L.O.G. let $\Lambda = \operatorname{diag}(\lambda_1 I_{s_1}, \ldots, \lambda_r I_{s_r})$ with $\lambda_1, \ldots, \lambda_r$ distinct positive and $s_1, \ldots, s_r$ in $\mathbb{N}$ such that $s_1 + \ldots + s_r = n$.

For $B = Q\Lambda^{\frac{1}{2}}Q^T$ where $\Lambda^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1} I_{s_1}, \ldots, \sqrt{\lambda_r} I_{s_r})$ we have $B^2 = Q\Lambda Q^T = A$.

Let $\tilde{B} \in \mathbb{R}^{n\times n}$ be an SPD matrix such that $\tilde{B}^2 = A$. Since $\tilde{B}$ is SPD there exists a spectral decomposition $\tilde{B} = \tilde{Q} D \tilde{Q}^T$.

$\tilde{B}^2 = A \Rightarrow \tilde{Q} D^2 \tilde{Q}^T = A$, D^2 is similar to A and hence to Λ. D^2 and Λ are diagonal, so the diagonal entries coincide up to a permutation.

W.L.O.G assume $D^2 = \Lambda$. Then $A = Q\Lambda Q^T = \tilde{Q}\Lambda\tilde{Q}^T$.

Partition Q and $\tilde{Q}$: $Q = [Q_1, \dots, Q_r]$ and $\tilde{Q} = [\tilde{Q}_1, \dots, \tilde{Q}_r]$ with $Q_k,\ \tilde{Q}_k \in \mathbb{R}^{n \times s_k} \quad \forall k \in \{1, \dots, r\}$.

For each $k \in \{1, \dots, r\}$, the columns of Q_k and the columns of $\tilde{Q}_k$ form an orthogonal basis for the same subspace, the eigenspace corresponding to λ_k.

$$\Rightarrow \exists V_k \in \mathbb{R}^{s_k \times s_k} \text{ orthogonal s.t. } \tilde{Q}_k = Q_k V_k \forall k \in \{1, \dots, r\}. \text{ So } \tilde{Q} = Q \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_r \end{bmatrix}$$

$$\tilde{B} = \tilde{Q}\Lambda^{\frac{1}{2}}\tilde{Q}^T = QV\Lambda^{\frac{1}{2}}V^TQ^T = Q\Lambda^{\frac{1}{2}}Q^T = B$$

because

$$V\Lambda^{\frac{1}{2}}V^T = \begin{pmatrix} V_1\sqrt{\lambda_1}I_{s_1}V_1^T & & \\ & \ddots & \\ & & V_r\sqrt{\lambda_r}I_{s_r}V_r^T \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1}I_{s_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_r}I_{s_r} \end{pmatrix}$$

B is called the principle square root of A or the SPD square root of A. □

Theorem 6.2. *Let $A, M \in \mathbb{R}^{n \times n}$ be commuting SPD matrices. Then the Richardson iteration for $Ax = b$ with $b \in \mathbb{R}^n$ satisfies*

$$\|e_k\|_M \leq \|I - \alpha_n A\|_2 \|e_{k-1}\|_M \text{ and}$$
$$\|e_k\|_M \leq \|(I - \alpha_k A) \cdots (I - \alpha_1 A)\|_2 \|e_0\|_M \quad \forall k \in \mathbb{N}$$

Proof. **TODO**

□

Krylov subspaces

Let $\mathcal{K}_k = \mathcal{K}_k(A, r_0) = \text{span}\{A_j r_0\}_{j=0}^{k-1} = \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}$. Obviously $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}\ \forall k \in \mathbb{N}$.

Remark 6.3. $r_{k-1} \in \mathcal{K}_k$ and $x_k - x_0 \in \mathcal{K}_k \quad \forall k \in \mathbb{N}$.

Proof. **TODO**

□

Remark 6.4. At step k, we, update the current iterate along the current residual, $x_{k-1} - x_0 \in \mathcal{K}_{k-1}$, $x_k - x_{k-1} \in span\{r_{k-1}\}$. So $\mathcal{K}_k = \mathcal{K}_{k-1} + \text{span}\{r_{k-1}\} \quad \forall k \in \mathbb{N}$.

Remark 6.5. For all $k \in \mathbb{N}$,

$$e_k = (I - \alpha_k A) \dots (I - \alpha_1 A) e_0 = q_k(A) e_0$$

where $q_k(A) \in P_k$, an algebraic polynomials of degree k given by

$$q_k = (1 - \alpha_k t) \cdots (1 - \alpha_1 t) = \sum_{j=0}^{k} c_j t^j \ \forall t \in \mathbb{F}$$

We have

$$q_k(A) = \sum_{j=0}^{k} c_j A^j = (I - \alpha_k A) \cdots (I - \alpha_1 A) \ \forall A \in \mathbb{F}^{n \times n}$$

Denote $Q_k = \{q \in P_k : \ deg(q) = k, q(0) = 1\} \subset P_k$. The iterative method implies

$$e_k = q_k(A) q(A) e_0 \text{ with } q_k \in Q_k$$

On the other hand: For $\mathbb{F} = \mathbb{C}$, if $\tilde{q}_k \in Q_k$, then 0 is not a root of $\tilde{q}_k$. So

$$\tilde{q}_k(t) = \prod_{j=1}^{k} (1 - \tilde{\alpha}_j t) \ \forall t \in \mathbb{C} \text{ with some } \tilde{\alpha}_1, \dots, \tilde{\alpha}_k \in \mathbb{C} \setminus \{0\}$$

and the iteration $x_k = x_{k-1} + \tilde{\alpha}_k r_{k-1}$ satisifies $e_k = \tilde{q}_k(A) e_0$.

Remark 6.6. Let $q_k \in Q_k$ be such that $e_k = q_k(A) e_0 \ \forall k \in \mathbb{N}$. Then

$$\begin{aligned} x_k - x_0 = e_k - e_0 &= -(I - q_k(A)) e_0 \\ &= (I - q_k(A)) A^{-1} r_0 = \pi_k(A) r_0 \quad \forall k \in \mathbb{N}_0 \end{aligned}$$

where $\pi_0 = 0 \in P_0$ and $\pi_k \in P_{k-1}$ (due to $q_k \in Q_k$) is given by

$$\pi_k(t) = \frac{1}{t}(1 - q_k(t)) \quad \forall t \in \mathbb{F}$$

so $x_k = x_0 + \underbrace{\pi_k(A) r_0}_{\in \mathcal{K}_k}$.

For $\mathbb{F} = \mathbb{R}$, consider $M \in \mathbb{R}^{n \times n}$ commuting with A. By Theorem 6.2 $\|e_k\|_M \leq \|q_k(A)\|_2 \|e_0\|_M$. If A is SPD, it has a spectral decomposition $A = Q \Lambda Q^T$ with Q orthogonal and Λ diagonal. Then

$$q_k(A) = Q q_k(\Lambda) Q^T \text{ and } \|q_k(A)\|_2 = \|q_k(\Lambda)\|_2 = \max_{t \in \boldsymbol{\lambda}(A)} |q_k(t)|$$

Definition 7.1. Let $A \in \mathbb{F}^{n\times n},\ b \in \mathbb{F}^n$. For each $k \in \mathbb{N}_0$,

$$\mathcal{K}_k(A,b) = \operatorname{span}\{A^j b\}_{j=0}^{k-1} \subseteq \mathbb{F}^n$$

is the kth *Krylov subspace* of A generated by b. In particular,

$$\mathcal{K}_k(A,b) = \mathcal{K}_{k-1}(A,b) + \operatorname{span}\{A^{k-1}b\}$$

with $\mathcal{K}_0(A,b) = \operatorname{span}\{0\}$ and $\dim \mathcal{K}_0(A,b) = 0$, so we have $\mathcal{K}_{k-1} \subseteq K_k$ and $\dim \mathcal{K}_k \leq k$ for all $k \in \mathbb{N}$.

Remark.

- Jacobi: $x_k - x_0 \in \mathcal{K}(D^{-1}A, D^{-1}(b - Ax_0))$
- Gauß-Seidel: $x_k - x_0 \in \mathcal{K}((D+U)^{-1}A, (D+U)^{-1}(b - Ax_0))$

Proposition 7.2. *$A \in \mathbb{F}^{n\times n}, b \in \mathbb{F}^n, k \in \mathbb{N}$. Then*

$$\mathcal{K}_k(A,b) = \{\pi(A) : \pi \in P_{k-1}\}$$

Remark 7.3. $A \in \mathbb{F}^{n\times n}, b \in \mathbb{F}^n, x_0 \in \mathbb{F}^n, r_0 = b - Ax_0$ Consider $k \in \mathbb{N}_0$.

if $k = 0$, let $\pi_k = 0 \in P_0$

if $k \in \mathbb{N}$, assume $\pi_k \in P_{k-1}$ is of degree $k-1$. Consider $x_k = x_0 + \pi_k(A)r_0$, we have

$$\begin{aligned} e_k &= x_k - x = \pi_k(A)r_0 - (x - x_0) \\ &= \pi_k(A)A(x - x_0) - (x - x_0) = -(I - \pi_k(A)A)(x - x_0) \\ &= q_k(A)e_0 \end{aligned}$$

with $q_k \in P_k$ given by

$$\boxed{q_k(t) = 1 - \pi_k(t)t} \quad \forall t \in \mathbb{F} \tag{*}$$

(*) implies that $q_k(0) = 1$ and $\deg(q_k) = k$, i.e. $q_k \in Q_k$.

Choose $\pi_k \in P_{k-1} \setminus P_{k-2} \implies$ get $q_k \in Q_k$.

If $q_k \in Q_k$, then $\pi_k \in P_{k-1}$ given by $\pi_k(t) = \frac{1}{t}(1 - q_k(t))$ satisfies $\deg(\pi_k) = k-1$ and $e_k = q_k(A)e_0$ implies $x_k - x_0 = \pi_k(A)r_0$.

Conjugate-gradient method

$A \in \mathbb{F}^{n\times n}$ Hermitian positive definite ($\mathbb{F} = \mathbb{R}$: symmetric) and $b, x_0 \in \mathbb{F}^n$. As A is Hermitian positive definite, it induces an inner product

$$\langle u, v\rangle_A : \ \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F} \quad \text{given by}$$

$$\langle u, v\rangle_A = u^* A v \quad \forall u, v \in \mathbb{F}^n$$

($\|\cdot\|_A = \sqrt{\langle \cdot, \cdot\rangle_A}$ is the induced norm)

The conjugate gradient method for $Ax = b$ starting at x_0 generates $(x_k)_{k\in\mathbb{N}}$ such that

$$y_k = x_k - x_0 = \operatorname{argmin}_{y\in\mathcal{K}_k(A,r_0)} \|\underbrace{y - (x - x_0)}_{(x_0+y)-x}\|_A$$

$$= \underbrace{\prod_{A,\mathcal{K}_k(A,r_0)}}_{\textbf{orth. proj.}} (x - x_0)$$

using an A-orthogonal basis for $\mathcal{K}_k(A, r_0)$.

Lemma 7.4. *Let $A \in \mathbb{F}^{n\times n}$ be Hermitian positive definite, $r_0 \in \mathbb{F}^n$, $k \in \mathbb{N}$, $y_k = \operatorname{argmin}_{y\in\mathcal{K}_k(A,r_0)}\|y - A^{-1}r_0\|_A$ and $r_k = r_0 - Ay_k$. Then*

$$r_k \perp \mathcal{K}_k(A, r_0).$$

Proof. The optimality of y_k is characterized by the $y_k \perp_A \mathcal{K}_k(A, r_0)$, i.e.

$$A(y_k - A^{-1}r_0) \perp \mathcal{K}_k(A, r_0), \text{ i.e.}$$
$$r_k \perp \mathcal{K}_k(A, r_0)$$

□

Lemma 7.5. *Let $n \in \mathbb{N}$, $A \in \mathbb{F}^{n\times n}$ be Hermitian positive definite, $r_0 \in \mathbb{F}^n, k \in \mathbb{N}_0, r_k \in \mathcal{K}_{k+1}(A, r_0)$ be non-zero and orthogonal to $\mathcal{K}_n(A, r_0)$. Let $p_{k+1} \in \mathcal{K}_{k+1}$ be non-zero and A-orthogonal to $\mathcal{K}_k(A, r_0)$. When $k \in \mathbb{N}$, assume additionaly that $p_k \in \mathcal{K}_k(A, r_0)$ is a non-zero vector A-orthogonal to $\mathcal{K}_{k-1}(A, r_0)$.*

Let $\gamma_k = \frac{r_k^* p_{k+1}}{r_k^* r_k}$. Then $p_{k+1} = \gamma_k r_k$ if $k = 0$ and $p_{k+1} = \gamma_k(r_k + \beta_k p_k)$ with $\beta_k = -\frac{p_k^* A r_k}{p_k^* A p_k}$ if $k \in \mathbb{N}$.

Proof. Since $\dim \mathcal{K}_{k+1}(A, r_0) \leq \dim \mathcal{K}_k(A, r_0) + 1$ and $r_k \in \mathcal{K}_{k+1}(A, r_0) \setminus \mathcal{K}_k(A, r_0)$, we have $\mathcal{K}_{k+1}(A, r_0) = \mathcal{K}_k(A, r_0) + \operatorname{span}\{r_k\}$. When $k = 0$, this gives $\mathcal{K}_1(A, r_0) = \operatorname{span}\{r_0\}$, so $p_1 \in \operatorname{span}\{r_0\}$. Then the coefficient of p_1 along r_0 is γ_0, so $p_1 = \gamma_0 r_0$.

When $k \in \mathbb{N}$, we have $p_k \in \mathcal{K}_k(A, r_0) \setminus \mathcal{K}_{k-1}(A, r_0)$, so, since $\dim \mathcal{K}_k(A, r_0) \leq \dim \mathcal{K}_{k-1}(A, r_0)+1$, we have $\mathcal{K}_k(A, r_0) = \mathcal{K}_{k-1}(A, r_0)+\operatorname{span}\{p_k\}$. Then $\mathcal{K}_{k+1}(A, r_0) = \mathcal{K}_{k-1}(A, r_0) + \operatorname{span}\{r_k, p_k\}$.

Due to $p_{k+1} \in \mathcal{K}_{k+1}(A, r_0)$, there exists $u_k \in \mathcal{K}_{k-1}(A, r_0)$, $\mu_k, \nu_k \in \mathbb{F}$ such that $p_{k+1} = u_k + \mu_k r_k + \nu_k p_k$

Since $A\mathcal{K}_{k-1}(A, r_0) \subseteq \mathcal{K}_k(A, r_0)$, we have $r_k \perp_A \mathcal{K}_{k-1}(A, r_0)$ since $r_k \perp \mathcal{K}_k(A, r_0)$. Further, $p_k \perp_A \mathcal{K}_{k-1}(A, r_0)$. Finally, recall that $p_{k+1} \perp_A \mathcal{K}_k(A, r_0)$ and hence $p_{k+1} \perp_A \mathcal{K}_{k-1}(A, r_0)$.

This yields $u_k = p_{k+1} - \mu_k r_k - \nu_k p_k \perp_A \mathcal{K}_{k-1}(A, r_0)$, i.e., $u_k = 0$.

Project p_{k+1} onto p_k w.r.t. the A-inner product: using the A-orthogonality of p_{k+1} to $\mathcal{K}_k(A, r_0)$, we optain $0 = p_k^* A p_{k+1} = \mu_k p_k^* A r_k + \nu_k p_k^* A p_k$.

Project p_{k+1} onto r_k w.r.t. the standard inner product: $r_k^* p_{k+1} = \mu_k r_k^* r_k$ because $r_k \perp \mathcal{K}_k(A, r_0)$, so $\mu_k = \gamma_k$ and $\nu_k = -\mu_k \frac{p_k^* A r_k}{p_k^* A p_k} = \gamma_k \beta_k$. □

Lemma 7.6. *Let $n \in \mathbb{N}$, $A \in \mathbb{F}^{n \times n}$ be Hermitian positive definite, $r_0 \in \mathbb{F}^n, k \in \mathbb{N}$*

$$y_i = \operatorname{argmin}_{y \in \mathcal{K}_i(A, r_0)} \|y - A^{-1} r_0\|_A \text{ and } r_i = r_0 - A y_i \text{ for } i \in \{k-1, k\}$$

Let $p_k \in \mathcal{K}_k(A, r_0)$ be a non-zero vector A-orthogonal to $\mathcal{K}_{k-1}(A, r_0)$. Then $y_k = y_{k-1} + \alpha_k p_k$ with $\alpha_k = \frac{p_k^* r_{k-1}}{p_k^* A p_k}$

Proof. Since $\dim \mathcal{K}_k(A, r_0) \leq \dim \mathcal{K}_{k-1}(A, r_0)+1$ and $p_k \in \mathcal{K}_k(A, r_0) \setminus \mathcal{K}_{k-1}(A, r_0)$, so $\dim \mathcal{K}_k(A, r_0) = \dim \mathcal{K}_{k-1}(A, r_0)+1$ and hence $\mathcal{K}_k(A, r_0) = \mathcal{K}_{k-1}(A, r_0) \oplus_{\perp_A} \operatorname{span}\{p_k\}$. Since y_k, y_{k-1} are A-orthogonal projections of $A^{-1} r_0$ onto $\mathcal{K}_k$ and $\mathcal{K}_{k-1}$, we have $y_k = y_{k-1} + \alpha_k p_k$ with $\alpha_k = \frac{p_k^* A (A^{-1} r_0)}{p_k^* A p_k} = \frac{p_k^* r_0}{p_k^* A p_k}$.

Since $r_k \perp \mathcal{K}_k(A, r_0)$ (Lemma 7.4) and $r_{k-1} = r_0 - A y_{k-1}$ and $A y_{k-1} \in A\mathcal{K}_{k-1}(A, r_0) \subseteq \mathcal{K}_k(A, r_0)$, we have that $p_k^* r_0 = p_k^* r_{k-1}$. So $\alpha_k = \frac{p_k^* r_{k-1}}{p_k^* A p_k}$. □

Theorem 7.7. *Let $A \in \mathbb{F}^{n \times n}$ be Hermitian positive definite, $r_0 \in \mathbb{F}^n$, $m \in \mathbb{N}$ and set $r_k = r_0 - A y_k$ for any $k \in \{1, \ldots, m\}$ and $r_{k-1} \neq 0$. We also assume that*

$p_1, \ldots, p_m \in \mathbb{F}^n$ *be A-orthogonal and such that* $p_k^* r_{k-1} = r_{k-1}^* r_{k-1}$ *(*$\gamma_{k-1} = 1$*) and* $p_1, \ldots, p_k$ *is a basis for* $\mathcal{K}_k(A, r_0)$*. And* $y_k = \operatorname{argmin}_{y \in \mathcal{K}_k(A,r_0)} \|y - A^{-1} r_0\|_A$*. Then*

$$y_k = y_{k-1} + \alpha_k p_k \text{ and } r_k = r_{k-1} - \alpha_k A p_k \text{ with } \alpha_k = \frac{r_{k-1}^* r_{k-1}}{p_k^* A p_k} \ \forall k \in \{1, \ldots, m\}$$

$$\text{and } p_{k+1} = r_k + \beta_k p_k \text{ with } \beta_k = \frac{r_k^* r_k}{r_{k-1}^* r_{k-1}} \ \forall k \in \{1, \ldots, m-1\}$$

Proof. By Lemma 7.6, we have $y_k = y_{k-1} + \alpha_k p_k$ with $\alpha_k = \frac{p_k^* r_{k-1}}{p_k^* A p_k}$, so $\alpha_k = \frac{r_{k-1}^* r_{k-1}}{p_k^* A p_k}$ for any $k \in \{1, \ldots, m\}$. This implies $r_k = r_{k-1} - A(x_k - x_{k-1}) = r_{k-1} - \alpha_k A p_k \ \forall k \in \{1, \ldots, m\}$. Finally, for each $k \in \{1, \ldots, m-1\}$, by Lemma 7.5, we have $p_{k+1} = r_k + \beta_k p_k$ with $\beta_k = -\frac{p_k^* A r_k}{p_k^* A p_k}$. Since $r_{k-1} \neq 0$, we have $\alpha_k \neq 0$ and hence $A_{p_k} = \frac{1}{\alpha_k}(r_{k-1} - r_k)$. Then $r_k^* A p_k = \frac{1}{\alpha_k}(\underbrace{r_k^* r_{k-1}}_{=0 \text{ since } r_k \perp \mathcal{K}_k} - r_k^* r_k) = -\frac{1}{\alpha_k} r_k^* r_k$. So $\beta_k = \frac{r_k^* r_k}{p_k^* A p_k} \cdot \frac{1}{\overline{\alpha}_k} = \frac{r_k^* r_k}{p_k^* A p_k} \cdot \frac{1}{\alpha_k} = \frac{r_k^* r_k}{r_{k-1}^* r_{k-1}}$. □

Algorithm 7.8 (The conjugate gradient method)**.**

Given: $A \in \mathbb{F}^{n \times n}$ Hermitian positive definite, $b \in \mathbb{F}^n$ and $x_0 \in \mathbb{F}^n$ s.t. $b - A x_0 \neq 0$.

Initialize: set $r_0 = b - A x_0$ and $p_1 = r_0$.

Iterate for $k = 1, 2, \ldots$:

set $\alpha_k = \frac{r_{k-1}^* r_{k-1}}{p_k^* A p_k}$, set $x_k = x_{k-1} + \alpha_k p_k$

set $r_k = r_{k-1} - \alpha_k A p_k$

if r_k is zero (or "small"), then terminate

set $\beta_k = \frac{r_k^* r_k}{r_{k-1}^* r_{k-1}}$, set $p_{k+1} = r_k + \beta_k p_k$.

Remark. Only need to store x_k, r_k, p_k (and $A p_k$ maybe) and compute only one matrix-vector product $A p_k$ per iteration.

Remark 8.1. $A \in \mathbb{F}^{n \times n}$ Hermitian positive definite, $b, x_0 \in \mathbb{F}^n$ and $r_0 = b - A x_0$. Let $k \in \mathbb{N}$ be such that $\dim \mathcal{K}_k(A, r_0) = k$. Then $\dim \mathcal{K}_j(A, r_0) = j$ for all $j \in \{1, \ldots, k\}$ The CG method produces x_j with $j \in \{1, \ldots, k\}$. By Proposition 7.2 these are generated by polynomials $\pi_j \in P_{j-1} \setminus P_{j-2}$ ($P_0 = \{0\}, P_{-1} = \emptyset$) with $j \in \{1, \ldots, k\}$:

$$x_j = x_0 + \pi_j(A)r_0 \quad \forall j \in \{1, \dots, k\}$$

Let $\pi_0 = 0$, so that $x_j = x_0 + \pi_j(A)r_0$ holds also for $j = 0$. Let $\sigma_j = \pi_j - \pi_{j-1}$. Then $\sigma_j \in P_{j-1} \setminus P_{j-2}$ and $x_j - x_{j-1} = \sigma_j(A)r_0$ for all $j \in \{1, \dots, k\}$.

The A-orthogonality of $p_1, \dots, p_k$, due to

$$x_j - x_{j-1} = \alpha_j p_j \quad \forall j \in \{1, \dots, k\},$$

is equivavalent to the orthogonality of of $\sigma_1, \dots, \sigma_k$ w.r.t. to a suitable inner product.

Indeed, let $\langle \cdot, \cdot \rangle : P_{k-1} \times P_{k-1} \to \mathbb{F}$ be given by $\langle u, v \rangle = (u(x)r_0)^* A(v(x)r_0)$. This function is an inner product on P_{k-1} (A is Hermitian positive definite and ?). Then $p_i^* A p_j = \langle \sigma_i, \sigma_j \rangle \frac{1}{\alpha_i \alpha_j}$ $\forall i, j \in \{1, \dots, k\}$ and hence $\langle \sigma_i, \sigma_j \rangle = 0$ $\forall i, j \in \{1, \dots, k\}$ such that $i \neq j$.

The inner product $\langle \cdot, \cdot \rangle$ is an L^2 inner product on P_{k-1} w.r.t. to a suitable Stieltjes measure.

Consider a spectral decomposition of A: $A = Q\Lambda Q^T$ with $Q \in \mathbb{F}^{n \times n}$ unitary and $\Lambda \in \mathbb{F}^{n \times n}$ diagonal. Let $W = Q^* r_0$. Then

$$\begin{aligned}
\langle u, v \rangle &= (u(A)r_0)^* A(v(A)r_0) = (Qu(\Lambda)Q^* r_0)^* Q\Lambda Q^* (Qv(\Lambda)Q^* r_0) \\
&= (u(\Lambda)W)^* \Lambda (v(\Lambda)W) = \sum_{i=1}^{n} |w_i|^2 \lambda_i u(\lambda_i) \\
&= \int_{\mathbb{R}} u(t)v(t)d\Theta(t) = \langle u, v \rangle_{L^2_\Theta(\mathbb{R})}
\end{aligned}$$

where

$$\Theta = \sum_{i=1}^{n} \lambda_i |w_i|^2 \Theta_{\lambda_i}$$

Here, for any $\lambda \in \mathbb{R}$, $\Theta_\lambda : \mathbb{R} \to \mathbb{R}$ is the Heaviside function jumping at λ:

$$\Theta_\lambda(t) = \begin{cases} 1, & t \geq \lambda \\ 0, & t < \lambda \end{cases} \quad \forall t \in \mathbb{R}$$

In terms of generalized functions: $\Theta'_\lambda = \delta_\lambda$ $\forall \lambda \in \mathbb{R}$, so that

$$d\Theta(t) = \sum_{i=1}^{n} \lambda_i |w_i|^2 \delta_{\lambda_i}(t)dt$$

For a system $\{\sigma_j\}_{j=0}^{\infty}$ of polynomials ($\sigma_j \in P_j$ of degree j $\forall j \in \mathbb{N}_0$), orthogonality w.r.t. a Stieltjes measure is a equivalent to a theree-term recurrence relation: $\exists$ $\{\xi_j\}_{j\in\mathbb{N}}$, $\{\eta_j\}_{j\in\mathbb{N}}$, $\{\zeta_j\}_{j\in\mathbb{N}}$ such that

$$\sigma_{j+1}(t) = (\xi_{j+1} + \eta_{j+1}t)\sigma_j(t) + \zeta_{j+1}\sigma_{j-1}(t) \quad \forall t \in \mathbb{R},\ j \in \mathbb{N}$$

The coefficients correspond to the inner product.

Example.

- Chebyshev polynomials:

$$T_{j+1}(t) = 2tT_j(t) - T_{j-1}(t) \quad \forall t \in \mathbb{R},\ j \in \mathbb{N}$$

  $(T_j)_{j=0}^{k-1}$ are orthogonal w.r.t. $\int_{-1}^{1} u(t)v(t)\frac{dt}{\sqrt{1-t^2}} \quad \forall u, v \in P_{k-1}$.

- Legendre polynomials:

$$(j+1)P_{j+1}(t) = (2j+1)tP_j(t) - jP_{j-1}(t) \quad \forall t \in \mathbb{R},\ j \in \mathbb{N}$$

  $(P_j)_{j=0}^{k-1}$ are orthogonal w.r.t. $\int_{-1}^{1} u(t)v(t)dt \quad \forall u, v \in P_{k-1}$.

Lemma 8.1 (A)**.** *Let $A \in \mathbb{F}^{n\times n}$ be invertible, $r_0 \in \mathbb{F}^n$ be non-zero.*

Let $m = \max_{k\in\mathbb{N}} \dim \mathcal{K}_k(A, r_0) \in \mathbb{N}$. Then

(i) $\dim \mathcal{K}_k(A, r_0) = \min\{k, m\}$ $\forall k \in \mathbb{N}$

(ii) $A^{-1}r_0 \in \mathcal{K}_m(A, r_0) \setminus \mathcal{K}_{m-1}(A, r_0)$.

Proof. (i) Since $r_0 \neq 0$, we have $\dim \mathcal{K}_1(A, r_0) = \dim \operatorname{span}\{r_0\} = 1$, so that $\dim \mathcal{K}_1(A, r_0) = \dim \mathcal{K}_0(A, r_0) + 1$. Let $\mathcal{M} = \{k \in \mathbb{N} : \dim \mathcal{K}_k(A, r_0) = \dim \mathcal{K}_{k-1}(A, r_0) + 1\}$.

Note: $1 \in \mathcal{M}$ and $\mathcal{M}$ is bounded because $\dim \mathcal{K}_k(A, r_0) \leq n$.

Let $\tilde{m} = \max \mathcal{M}$. Then $\mathcal{M} = \{1, \ldots, \tilde{m}\}$! Indeed, for $k \in \{1, \ldots, \tilde{m}\}$, we had $k \notin \mathcal{M}$, we would have $\mathcal{K}_j(A, r_0) = \mathcal{K}_k(A, r_0)$ for all $j \in \mathbb{N}$ such that $j \geq k$, hence $\tilde{m} \notin \mathcal{M}$. So $\dim \mathcal{K}_k(A, r_0) = k$ $\forall k \in \{1, \ldots, \tilde{m}\}$.

On the other hand, $\dim \mathcal{K}_k(A, r_0) = \dim \mathcal{K}_{\tilde{m}}(A, r_0)$ $\forall k \in \mathbb{N}$ such that $k \geq \tilde{m}$. Then $m = \tilde{m}$ and $\dim \mathcal{K}_k(A, r_0) = \min\{k, \tilde{m}\}$ $\forall k \in \mathbb{N}$.

(ii) Since $\mathcal{K}_{m+1}(A, r_0) = \mathcal{K}_m(A, r_0)$, we have $A^m r_0 \in \mathcal{K}_m(A, r_0)$. Due to $r_0 \neq 0$, this means $A^m r_0 = \sum_{k=\ell}^{m} c_k A^{k-1} r_0$ for some $\ell \in \{0, \ldots, m\}$ and $c_\ell, \ldots, c_m \in \mathbb{F}$

s.t. $c_\ell \neq 0$. Then

$$\begin{aligned} A^{-1}r_0 = A^{-\ell}(A^{\ell-1}r_0) &= A^{-\ell}\frac{1}{c_\ell}(A^m r_0 - \sum_{k=\ell+1}^{m} c_k A^{k-1} r_0) \\ &= \frac{1}{c_\ell}A^{m-\ell}r_0 - \frac{1}{c_\ell}\sum_{k=\ell+1}^{m} c_k A^{k-\ell-1} r_0 \\ \Rightarrow A^{-1}r_0 &\in \mathcal{K}_m(A, r_0) \end{aligned}$$

proven: $A^{-1}r_0 \in \mathcal{K}_m(A, r_0)$. Further, let us consider $\tilde{m} = \min\{k \in \mathbb{N} : A^{-1}r_0 \in \mathcal{K}_k(A, r_0)\}$ ($A^{-1}r_0 \in \mathcal{K}_{\tilde{m}}(A, r_0) \setminus \mathcal{K}_{\tilde{m}-1}(A, r_0)$). The set is nonempty ($m$ is in the set), so $\tilde{m} \in \{1, \ldots, m\}$.

It remains to show that $\tilde{m} = m$: $\dim \mathcal{K}_{\tilde{m}}(A, r_0) \leq \dim \mathcal{K}_{\tilde{m}-1}(A, r_0) + 1$, so $\dim \mathcal{K}_{\tilde{m}}(A, r_0) = \operatorname{span}\{A^{-1}r_0\} + \mathcal{K}_{\tilde{m}-1}(A, r_0)$.

$$\begin{aligned} \Rightarrow A\mathcal{K}_{\tilde{m}}(A, r_0) &= A\mathcal{K}_{\tilde{m}-1}(A, r_0) + \operatorname{span}\{r_0\} \\ &\subseteq \mathcal{K}_{\tilde{m}}(A, r_0) + \mathcal{K}_{\tilde{m}}(A, r_0) = \mathcal{K}_{\tilde{m}}(A, r_0) \end{aligned}$$

So $\mathcal{K}_{\tilde{m}+1}(A, r_0) = \mathcal{K}_{\tilde{m}}(A, r_0)$ and hence $\mathcal{K}_m(A, r_0) = \mathcal{K}_{\tilde{m}}(A, r_0)$ $\forall k \in \mathbb{N}$ such that $k \geq \tilde{m}$. This means $\tilde{m} = m$.

□

Lemma 8.1 (B)**.** *Let $A \in \mathbb{F}^{n\times n}$ be diagonalizable, $A = S\Lambda S^{-1}$ be an eigenvalue decomposition of A ($\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$). Let $r_0 \in \mathbb{F}^n$, $\omega \in S^{-1}r_0$. Assume that $\dim \mathcal{K}_k(A, r_0) = k$ for some $k \in \mathbb{N}$. Then*

$$\#\{\lambda_i \mid i \in \{1, \ldots, n\}, \omega_i \neq 0\} \geq k$$

(Note: $S^{-1}e_0 = S^{-1}(x_0 - x) = -S^{-1}A^{-1}r_0 = -\Lambda^{-1}\omega$ (if A is invertible))

Proof. Note that $\mathcal{K}_k(A, r_0) = S\mathcal{K}_k(\Lambda, \omega)$ and hence $\dim \mathcal{K}_k(\Lambda, \omega) = k$ since S is invertible. So

$$\dim \mathcal{K}_k(\Lambda, \omega) = \operatorname{rank} \begin{pmatrix} \omega_1 & \lambda_1\omega_1 & \lambda_1^2\omega_1 & \cdots & \lambda_1^{k-1}\omega_1 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \omega_n & \lambda_n\omega_n & \lambda_n^2\omega_n & \cdots & \lambda_n^{k-1}\omega_n \end{pmatrix} = k$$

$\Rightarrow$ $\exists i_1, \ldots, i_k \in \{1, \ldots, n\}$ distinct such that the matrix

$$\begin{pmatrix} \omega_{i_1} & \lambda_{i_1}\omega_{i_1} & \lambda_{i_1}^2\omega_{i_1} & \cdots & \lambda_{i_1}^{k-1}\omega_{i_1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \omega_{i_k} & \lambda_{i_k}\omega_{i_k} & \lambda_{i_k}^2\omega_{i_k} & \cdots & \lambda_{i_k}^{k-1}\omega_{i_k} \end{pmatrix}$$

is non-singular. Then $\omega_{i_1}, \ldots, \omega_{i_k}$ are all non-zero and $\lambda_{i_1}, \ldots, \lambda_{i_k}$ are distinct. $\square$

Remark 8.2 (in the notations of Remark 8.1). Consider $j \in \{2, \ldots, k-1\}$, then $p_{j+1} = r_j + \beta_j p_j$ and $p_j = r_{j-1} + \beta_{j-1}p_{j-1}$, where $r_j = r_{j-1} - \alpha_j A p_j$. Expressing $r_j = p_{j+1} - \beta_j p_j$, $r_{j-1} = p_j - \beta_{j-1}p_{j-1}$ and substituting those expressions into the recurrence for residuals, we obtain:

$$p_{j+1} - \beta_j p_j = p_j - \beta_{j-1}p_{j-1} - \alpha_j A p_j. \qquad (*)$$

Since $p_i = \frac{1}{\alpha_i}\sigma_i(A)r_0$ for all $i \in \{0, \ldots, k\}$, where we set $\alpha = 0$ for convenience, $(*)$ gives us $\frac{1}{\alpha_{j+1}}\sigma_{j+1}(A)r_j$ **TODO: There is some error, find correct remark**

Lemma 8.2 (A)**.** *Let $A \in \mathbb{F}^{n\times n}$ be Hermitian positive definite. Assume that $\boldsymbol{\lambda}(A) \subseteq [\lambda, \Lambda]$ for some $\lambda, \Lambda \in \mathbb{R}$ with $0 < \lambda < \Lambda$. Consider*

$$\phi : \ \mathbb{F} \to \mathbb{F}, \ \text{given by } \phi(t) = -\frac{2t - (\Lambda + \lambda)}{\Lambda - \lambda} \quad \forall t \in \mathbb{F}$$

and set $\tau = \frac{\Lambda+\lambda}{\Lambda-\lambda}$. Then for each $k \in \mathbb{N}$, $q_k = \frac{T_k \circ \phi}{T_k(\tau)}$, where T_k is the degree k Chebyshev polynomial of the first kind, satisfies $q_k \in Q_k$ ($q_k \in P_k$ and $q_k(0) = 1$) and $\|q_k(A)\|_2 \leq 2(\frac{\sqrt{a}-1}{\sqrt{a}+1})^k$, where $a = \frac{\Lambda}{\lambda} \geq \operatorname{cond}_2(A)$.

TODO: Proof

Theorem 8.3. *Let $A \in \mathbb{F}^{n\times n}$ be Hermitian positive definite, $b, x_0 \in \mathbb{F}^n$ and $r_0 = b - Ax_0$. Then for each $k \in \mathbb{N}$, the k-th CG-iteration for $Ax = b$ and initial guess x_0, $x_k = x_0 + \operatorname{argmin}_{y \in \mathcal{K}_k(A, r_0)} \|y - A^{-1}r_0\|_A$, satisfies*

$$\|x_k - x\|_A \leq 2\left(\frac{\sqrt{a}-1}{\sqrt{a}+1}\right)^k \|x_0 - x\|_A,$$

where $a = \frac{\Lambda}{\lambda}$ and $\boldsymbol{\lambda}(A) \subseteq [\lambda, \Lambda]$ for $0 < \lambda < \Lambda$.

Proof. Due to the optimality of x_k, we have

$$\|x_k - x\|_A = \|(x_k - x_0) + (x_0 - x)\|_A \leq \|y_k - A^{-1}r_0\|_A \quad \forall y_k \in \mathcal{K}_k(A, r_0)$$

q_k be given as in Lemma 8.2 and $\pi_k \in P_{k-1}$ be given by

$$\pi_k(t) = \frac{1}{t}(1 - q_k(t)) \quad \forall t \in \mathbb{F} \quad \text{(see Remark 7.3)}$$

Then $\pi_k(A)r_0 \in \mathcal{K}_k(A, r_0)$ by Proposition 7.2. So

$$\begin{aligned}\|x_k - x\|_A &\leq \|\pi_k(A)r_0 - A^{-1}r_0\|_A = \|A^{-1}r_0 - A^{-1}q_k(A)r_0\|_A \\ &\leq \|q_k(A)\|_2\|x - x_0\|_A \leq 2\left(\frac{\sqrt{a}-1}{\sqrt{a}+1}\right)^k \|x - x_0\|_A \quad \forall k \in \mathbb{N}\end{aligned}$$

□

Remark. Improvement over gradient descent:

Instead of $(\frac{a-1}{a+1})^k$, we have $(\frac{\sqrt{a}-1}{\sqrt{a}+1})^k$.

Preconditioned CG-iteration

Our assumptions: $k \in \mathbb{N}$, $A \in \mathbb{F}^{n\times n}$ Hermitian positive definite, λ, Λ - spectral bounds, maybe unfavorable

Example.

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \quad \kappa = \frac{\Lambda}{\lambda} \sim n^2$$

$Ax = b$, P (invertible) as preconditioner $\rightsquigarrow P^{-1}Ax = P^{-1}b$.

Problem: $P^{-1}A$ might not be Hermitian positive definite.

Let us assume P is Hermitian positive definite (P^{-1} is so as well, so it has a Cholesky decomposition) and $C \in \mathbb{F}^{n\times n}$ is non-singular such that $P^{-1} = CC^*$.

Or: Take $(P^{-1})^{1/2}$, the HPD square root

Then $Ax = b \Rightarrow \underbrace{C^*AC(C^{-1}x) = C^*b}_{\text{symmetric two-sided preconditioning}} \rightsquigarrow \tilde{A}\tilde{x} = \tilde{b}$

If $\tilde{x}$ solves the preconditioned system, $x = C\tilde{x}$ solves the original system.

$$\|C\tilde{x}_k - x\| = \|C(\tilde{x}_k - \tilde{x})\| \leq \|C\|\|\tilde{x}_k - \tilde{x}\|$$

Easy to check: $\tilde{A}$ is Hermitian positive definite (check $x^*\tilde{A}x > 0$ and $\tilde{A}^* = \tilde{A}$), so we can apply the CG method to $\tilde{A}\tilde{x} = \tilde{b}$.

(i) accuracy: If $\tilde{x}_k$ is an approximation of $\tilde{x}$, then $x_k = C\tilde{x}_k$ satisfies

$$\|x_k - x\|_A = \sqrt{(\tilde{x}_k - \tilde{x})^* C^* AC(\tilde{x}_k - \tilde{x})} = \|\tilde{x}_k - \tilde{x}\|_{\tilde{A}}$$

So the CG method for the preconditioned system minimizes also the A-norm of the error of the initial system.

(ii) conditioning: $\operatorname{cond}_2 \tilde{A} = \operatorname{cond}_2 P^{-1}A$ by the following result

Proposition 9.1. *Let $A \in \mathbb{F}^{n\times n}$ be Hermitian non-singular and $C \in \mathbb{F}^{n\times n}$ be non-singular. Then*

$$\operatorname{cond}_2 C^*AC \leq \operatorname{cond}_2 CC^*A$$

Proof. Let $P = (CC^*)^{-1}$, $\tilde{A} = C^*AC$ and $B = P^{-1}A$. For each $k \in \mathbb{N}$, we have

$$\tilde{A}^k = C^{-1}(CC^*A)^k C = C^{-1}B^kC \text{ and } \|\tilde{A}^{-1}\|_2^k = \|\tilde{A}^{-k}\|_2 \quad \forall k \in \mathbb{N}$$

Then

$$\|\tilde{A}\|_2^k = \|\tilde{A}^k\|_2 \leq \|C^{-1}\|_2\|C\|_2\|B\|_2^k \text{ and } \|\tilde{A}^{-1}\|_2^k = \|\tilde{A}^{-k}\|_2 \leq \|C^{-1}\|_2\|C\|_2\|B^{-1}\|_2^k.$$

Taking the k-th root and passing to $k \to \infty$, we get $\|\tilde{A}\|_2 \leq \|B\|_2$ and $\|\tilde{A}^{-1}\|_2 \leq \|B^{-1}\|_2$. So

$$\operatorname{cond}_2 \tilde{A} \leq \operatorname{cond}_2 B$$

□

Proposition 9.2. *Let $A \in \mathbb{F}^{n\times n}$ be HPD, $R \in \mathbb{F}^{n\times n}$ be Hermitian such that $\rho(I - RA) < 1$. Then R is positive definite and, if $R = CC^*$ for some $C \in \mathbb{F}^{n\times n}$, then $\tilde{A} = C^*AC$ satisfies*

$$\lambda_{\max}(\tilde{A}) \leq 1 + \rho,\ \lambda_{\min}(\tilde{A}) \geq 1 - \rho \quad (\Rightarrow \operatorname{cond}_2 \tilde{A} = \frac{\lambda_{\max}}{\lambda_{\min}} \leq \frac{1+\rho}{1-\rho})$$

Proof. Let $U \in \mathbb{F}^{n\times n}$ be non-singular such that $A = UU^*$ (e.g. the Cholesky factor of A). Then $I - U^*RU = U^*(I - RA)U^{-*}$. So $\rho(I - U^*RU) = \rho(I - RA) < 1$

U^*RU is Hermitian $\rightarrow$ $\boldsymbol{\lambda}(U^*RU) \subset \mathbb{R}$ and hence $\boldsymbol{\lambda}(U^*RU) \subset [1-\rho, 1+\rho] \subset (0, 2)$, where $\rho = \rho(I - RA)$. So U^*RU is positive definitem so R is as well. Let $C \in \mathbb{F}^{n\times n}$ be such that $R - CC^*$ then $\rho(I - \tilde{A}) = \rho(I - C^*AC) = \rho(I - RA) = \rho$. Since $\tilde{A}$ is Hermitian, this implies $\boldsymbol{\lambda}(\tilde{A}) \subset [1-\rho, 1+\rho]$. □

Remark (Reformulation of CG algorithm for $\tilde{A}\tilde{x} = \tilde{b}$). We consider $P \in \mathbb{F}^{n\times n}$ Hermitian positive definite such that $P^{-1} = CC^*$. For x_0 (Initial guess), we set $r_0 = b - Ax_0$, $\tilde{x_0} = C^{-1}x_0$. Algorithm 7.8 for $\tilde{A}\tilde{x} = \tilde{b}$ starting at $\tilde{x_0}$:

$$\tilde{r}_0 = \tilde{b} - \tilde{A}\tilde{x_0},\ \tilde{p}_1 = \tilde{r}_0$$

Note that $\tilde{r}_0 = C^*b - C^*ACC^{-1}x_0 = C^*r_0$. For $k \in \mathbb{N}$, the k-th iteration takes the form:

$$\begin{aligned}
\tilde{\alpha}_k &= \frac{\tilde{r}_{k-1}^*\tilde{r}_{k-1}}{\tilde{p}_k^*\tilde{A}\tilde{p}_k} \\
\tilde{x_k} &= \tilde{x}_{k-1} + \tilde{\alpha}_k\tilde{p}_k \\
\tilde{r_k} &= \tilde{r}_{k-1} - \tilde{\alpha}_k\tilde{A}\tilde{p}_k \quad (\tilde{r}_k = 0 \rightarrow \text{ terminate}) \\
\tilde{\beta}_k &= \frac{\tilde{r}_k^*\tilde{r}_k}{\tilde{r}_{k-1}^*\tilde{r}_{k-1}} \\
\tilde{p}_{k+1} &= \tilde{r}_k + \tilde{\beta}_k\tilde{p}_k
\end{aligned}$$

First,

$$\tilde{\alpha}_k = \frac{\tilde{r}_{k-1}^*C^{-1}\overbrace{CC^*}^{P^{-1}}C^{-*}\tilde{r}_{k-1}}{\tilde{p}_k^*C^*AC\tilde{p}_k} = \frac{(C^{-*}\tilde{r}_{k-1})^*P^{-1}(C^{-*}\tilde{r}_{k-1})}{(C\tilde{p}_k)^*A(C\tilde{p}_k)}$$

$$r_k = C^{-*}\tilde{r}_k = C^{-*}\tilde{r}_{k-1} - \tilde{\alpha}_kC^{-*}C^*AC\tilde{p}_k = C^{-*}\tilde{r}_{k-1} - \tilde{\alpha}_kA\underbrace{C\tilde{p}_k}_{v_k}$$

$$v_{k+1} = C\tilde{p}_{k+1} = C\tilde{r}_k + \tilde{\beta}_kC\tilde{p}_k = CC^*C^{-*}\tilde{r}_k + \tilde{\beta}_kC\tilde{p}_k = P^{-1}\underbrace{C^{-*}\tilde{r}_k}_{r_k} + \tilde{\beta}_k\underbrace{C\tilde{p}_k}_{v_k}$$

$$x_k = C\tilde{x}_k = \underbrace{C\tilde{x}_{k-1}}_{x_{k-1}} + \tilde{\alpha}_kC\tilde{p}_k$$

($\rightarrow$ $r_k = b - Ax_k$ holds)

TODO: Ask about the $\tilde{x_k}$ definition. Kazeev did write $\tilde{p}_{k-1}$ instead of $\tilde{p}_k$

$$\tilde{\alpha}_k = \frac{r_{k-1}^* P^{-1} r_{k-1}}{p_k^* A p_k} \quad \text{and} \quad \tilde{\beta}_k = \frac{r_k^* P^{-1} r_k}{r_{k-1}^* P^{-1} r_{k-1}}$$

We can evaluate

$$z_0 = P^{-1} r_0, \quad v_1 = C\tilde{p}_1 = C\tilde{r}_0 = P^{-1} r_0 = z_0.$$

For $k \in \mathbb{N}$, assuming $z_{k-1} = P^{-1} r_{k-1}$ and $v_k = C\tilde{p}_k$ have been evaluated, we define

$$z_k = P^{-1} r_k \quad \text{and} \quad v_{k+1} = C\tilde{p}_{k+1}$$

Algorithm 9.3 (The preconditioned CG method, PCG)**.**

Given: $A, P \in \mathbb{F}^{n\times n}$ Hermitian positive definite, $b, x_0 \in \mathbb{F}^n$ such that $b - Ax_0 \neq 0$

Initialize: set $r_0 = b - Ax_0,\ z_0 = P^{-1} r_0,\ v_1 = z_0$

Iterate for $k = 1, 2, \ldots$:

set $\tilde{\alpha}_k = \frac{r_{k-1}^* z_{k-1}}{v_k^* A v_k}$

set $x_k = x_{k-1} + \tilde{\alpha}_k v_k$

set $r_k = r_{k-1} - \tilde{\alpha}_k A v_k$

if $r_k = 0$, then terminate

set $z_k = P^{-1} r_k$

set $\tilde{\beta}_k = \frac{r_k^* z_k}{r_{k-1}^* z_{k-1}}$

set $v_{k+1} = z_k + \tilde{\beta}_k v_k$

Proposition 9.4. *Let $A, P \in \mathbb{F}^{n\times n}$ be Hermitian positive definite, $b, x_0 \in \mathbb{F}^n$ such that $b - Ax_0 \neq 0$. $C \in \mathbb{F}^{n\times n}$ be such that $P^{-1} = CC^*$. Let $\tilde{x}_0 = C^{-1} x_0$. Let $\tilde{x}_1, \ldots, \tilde{x}_k \in \mathbb{F}^{n\times n}$ and $x_1 = C\tilde{x}_1, \ldots, x_k = C\tilde{x}_k$.*

Then the following statements are equivalent:

(i) Algorithm 7.8 applied to $(C^*AC)x = C^*b$ produces iterates $x_1, \ldots, x_k$.

(ii) Algorithm 9.3 applied to $Ax = b$ with preconditioner P produces iterates $\tilde{x}_1, \ldots, \tilde{x}_k$.

Proof. Given above. □

The above proposition allows to apply the error bound of Theorem 8.3 to the iterates produced by Algorithm 9.3. Since, as we noted at the beginning of the lecture, $\|x_k - x\|_A = \|\tilde{x_k} - \tilde{x}\|_A \ \forall k \in \mathbb{N}$. By Proposition 9.2, $\rho(I - P^{-1}A) = \rho < 1$ implies $\mathrm{cond}_2(C^*AC) \leq \frac{1+\rho}{1-\rho}$, and the further ρ is away from 1, the smaller $\mathrm{cond}_2(C^*AC)$ is and faster the bound of Theorem 8.3 converges to zero with respect to the iteration index $k \in \mathbb{N}$.

Chebyshev iterative methods

Lemma 8.2A above considers $q_k \in Q_k$ given by $q_k = \frac{T_k \circ \phi}{T_k(\tau)}$, where $k \in \mathbb{N}$, T_k is the Chebyshev polynomial of the first kind and degree k, $\tau = \frac{\kappa+1}{\kappa-1}$ with $\kappa = \frac{\Lambda}{\lambda}$ and $\phi : \mathbb{R} \to \mathbb{R}$ is given by:

$$\phi(t) = -\frac{2t - (\Lambda + \lambda)}{\Lambda - \lambda} \quad \forall t \in \mathbb{R}$$

The parameters $\lambda, \Lambda > 0$ are choses so that $\lambda < \Lambda$ and $\boldsymbol{\lambda}(A) \subseteq [\lambda, \Lambda]$ for an Hermitian matrix $A \in \mathbb{F}^{n \times n}$. Such a choice of q_k leads to the so-called *Chebyshev iterative methods* of two types.

(i) **Chebyshev method of Richardson type.**

For $m \in \mathbb{N}$, let $q_m = \frac{T_m \circ \phi}{T_m(\tau)}$ be realized by m Richardson iterations: for each $k \in \{1, \ldots, m\}$, choose $q_k \in Q_k$ so that

$$q_k(t) = (1 - \alpha_1 t) \cdots (1 - \alpha_m t) \quad \forall t \in \mathbb{R}$$

where $\alpha_1, \ldots, \alpha_m$ are the roots of $T_m \circ \phi$, i.e., (form a permutation of)

$$\phi^{-1}(\cos(\frac{\pi}{m}(j + \frac{1}{2})) \text{ with } j \in \{0, \ldots, m-1\}$$

After m iterations, a stopping criterion may be checked and the process may be restarted, so that

$$q_{mi+k} = q_k(q_m)^i \quad \forall i \in \mathbb{N},\ k \in \{1, \ldots, m\}.$$

Note that while q_m is a rescaled Chebyshev polynomial, $q_1, \ldots, q_m$ are not such. Lemma 8.2A gives a convergence bound for the any integral number of such blocks at m Richardson iterations. The recurrence for x_k involves only two consecutive terms (as in any Richardson iteration).

(ii) **Chebychev method of Krylov type**

For $k \in \mathbb{N},\ q_k \in Q_k$ is chosen as follows:

$$q_k = \frac{T_k \circ \phi}{T_k(\tau)}$$

This is not a Richardson iteration:

$$\left\{\cos\frac{\pi}{k}(j+\frac{1}{2})\right\}_{j=0}^{k-1} \bigcap \left\{\cos\frac{\pi}{k+1}(j+\frac{1}{2})\right\}_{j=0}^{k}$$

are disjoint sets, so that $\frac{q_{k+1}}{q_k}$ is not a polynomial $\forall k \in \mathbb{N}$. So this method cannot be realized as a Richardson iteration (computing each iterate x_k with $k \in \mathbb{N}$ by an independent sequence of k Richardson iterations is very expensive and therefore not smart). Such a method, however, can be efficiently realized using the three-term recurrence relation for $\{T_k\}_{k\in\mathbb{N}}$, which translates into such for $\{q_k\}_{k\in\mathbb{N}}$, which, thanks to Remark 7.3, yields such for $\{\pi_k\}_{k\in\mathbb{N}}$ and hence for $\{x_k = x_0 + \pi_k(A)r_0\}_{k\in\mathbb{N}}$. Indeed, we have

$$T_{k+1}(t) = (\xi_{k+1} + \eta_{k+1}t)T_k(t) + \zeta_{k+1}T_{k-1}(t) \quad \forall t \in \mathbb{R}\ \forall k \in \mathbb{N}$$

with $\xi_{k+1} = 0,\ \eta_{k+1} = 2$ and $\zeta_{k+1} = -1$ for $k \in \mathbb{N}$.

(We considered such a general form of a three-term recurrence relation for orthogonal polynomials in lecture 8, see right below Remark 8.1)

Let us set $c_k = T_k(\tau)\ \forall k \in \mathbb{N}_0$. As we showed in the proof of Lemma 8.2A, we have $c_k = \frac{1}{2}(\mu^k + \mu^{-k})\ \forall k \in \mathbb{N}$, where $\mu = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$. Let $\omega = \frac{1}{\Lambda+\lambda}$. Then

$$\phi(t) = \tau - \frac{2t}{\Lambda - \lambda} = \tau(1 - \omega t) \quad \forall t \in \mathbb{R}$$

TODO: Something doesn't add up here, i.e.

$$\kappa = \frac{\Lambda}{\lambda},\ \tau = \frac{\kappa+1}{\kappa-1} = \frac{\Lambda+\lambda}{\Lambda-\lambda}$$
$$\phi(t) = \tau - \frac{2t}{\Lambda-\lambda} = \tau(1 - \frac{2t}{\tau(\Lambda-\lambda)}) = \tau(1 - \frac{2t}{\Lambda+\lambda})$$

Then first, $q_0(t) = 1$ and $q_1(t) = \frac{1}{\tau}\phi(t) = 1 - \omega t \ \forall t \in \mathbb{R}$, **TODO: V.K. wrote** α_0
so that $\pi_0(t) = 0$ and $\pi_1(t) = \frac{1}{t}(1 - q_1(t)) = \omega \ \forall t \in \mathbb{R}$. So we have

$$x_1 = x_0 + \omega r_0 \quad \text{(c.f. Theorem 5.8)}$$

Further, for each $k \in \mathbb{N}$, the three-term recurrence relation between T_{k+1}, T_k and T_{k-1} gives

$$c_{k+1} = (\xi_{k+1} + \eta_{k+1}\tau)c_k + \zeta_{k+1}c_{k-1} \qquad (\star)$$

$$q_{k+1}(t) = \frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\phi(t))q_k(t) + \frac{c_{k-1}}{c_{k+1}}\zeta_{k+1}q_{k-1}(t) =$$
$$= \frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\phi(t))q_k(t) + (1 - \frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\tau))q_{k-1}(t) \quad \forall t \in \mathbb{R} \quad (*)$$

Let us look at the incremental iteration updates: $x_k - x_{k-1} = \sigma_k(A)r_0 \ \forall k \in \mathbb{N}$, where, as in Remark 8.1,

$$\sigma_k = \pi_k - \pi_{k-1} \ \in P_{k-1} \setminus P_{k-2} \quad \forall k \in \mathbb{N} \quad (P_{-1} = \emptyset)$$

For these incremental polynomials, we have

$$\sigma_k(t) = \frac{1}{t}(1 - q_k(t) - 1 + q_{k-1}(t)) = \frac{1}{t}(q_{k-1}(t) - q_k(t)) \quad \forall t \in \mathbb{R} \ \forall k \in \mathbb{N}$$

and obtain the following from $(*)$:

$$\begin{aligned}\sigma_{k+1}(t) = & -\frac{c_k}{c_{k+1}}\frac{1}{t}(\xi_{k+1} + \eta_{k+1}\tau(1 - \omega t))q_k(t) + \frac{1}{t}q_k(t) \\ & + \left(\frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\tau) - 1\right)\frac{1}{t}q_{k-1}(t) \\ = & \frac{c_k}{c_{k+1}}\eta_{k+1}\omega\tau q_k(t) + \left(\frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\tau) - 1\right)\sigma_k(t) \quad \forall t \in \mathbb{R} \ \forall k \in \mathbb{N},\end{aligned}$$

i.e.

$$\sigma_{k+1} = \frac{c_k}{c_{k+1}}\eta_{k+1}\omega\tau q_k + \left(\frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\tau) - 1\right)\sigma_k \quad \forall k \in \mathbb{N}$$

Setting

$$\alpha_{k+1} = \frac{c_k}{c_{k+1}}\eta_{k+1}\tau \quad \text{and}$$
$$\beta_{k+1} = \frac{c_k}{c_{k+1}}(\xi_{k+1} + \eta_{k+1}\tau) - 1 \quad \forall k \in \mathbb{N},$$

we arrive at

$$\sigma_{k+1} = \alpha_{k+1}\omega q_k + \beta_{k+1}\sigma_k \quad \forall k \in \mathbb{N},$$

which means

$$\boxed{x_{k+1} = x_k + \alpha_{k+1}\omega r_k + \beta_{k+1}(x_k - x_{k-1}) \quad \forall k \in \mathbb{N}} \tag{$*$}$$

Introducing $p_k = x_k - x_{k-1}$, we rewrite the above three-term recurrence as follows:

$$\begin{cases} x_k = x_{k-1} + p_k \\ p_{k+1} = \alpha_{k+1}\omega r_k + \beta_{k+1} p_k \end{cases} \qquad \forall k \in \mathbb{N} \tag{$**$}$$

Let us now specify the expressions for the coefficients. For the Chebyshev polynomials of the first kind, we have $\xi_{k+1} = 0$ and $\eta_{k+1} = 2\ \forall k \in \mathbb{N}$, so that $\alpha = 2\frac{c_k}{c_{k+1}}\tau$ and $\beta = \alpha_{k+1} - 1\ \forall k \in \mathbb{N}$. The recursion $(**)$ is initialized with $p_1 = \omega r_0$. Finally,

$$\begin{aligned} \alpha_{k+1} = 2\,\frac{c_k}{c_{k+1}}\tau = 2\,\frac{\mu^k + \mu^{-k}}{\mu^{k+1} + \mu^{-(k+1)}}\tau &= 2\,\frac{\tau}{\mu}\,\frac{1+\mu^{-2k}}{1+\mu^{-2(k+1)}} \\ &= 2\left(1 - \frac{2\sqrt{\kappa}}{(\sqrt{\kappa}+1)^2}\right)\left(1 + \frac{1-\mu^{-2}}{\mu^{2k}+\mu^{-2}}\right) \quad \forall k \in \mathbb{N} \end{aligned}$$

with $\mu = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} > 1$ (as above), so that $\lim_{k\to\infty} \alpha_{k+1} = 2\left(1 - \frac{2\sqrt{\kappa}}{(\sqrt{\kappa}+1)^2}\right) \in (1,2)$. From $(\star)$ above, we obtain

$$\frac{c_{k+1}}{c_k} = 2\tau - \frac{c_{k-1}}{c_k} \quad \forall k \in \mathbb{N}$$

so that

$$\alpha_{k+1} = \frac{2\tau}{2\tau - \frac{c_{k-1}}{c_k}} = \frac{1}{1 - \frac{\alpha_k}{4\tau^2}} \quad \forall k \in \mathbb{N}$$

The (Krylov-type) Chebyshev method then takes the following form.

Algorithm 9.5 (Krylov-type Chebyshev method)**.**

Given: $A \in \mathbb{F}^{n\times n}$ Hermitian positive definite, $b, x_0 \in \mathbb{F}^n, \Lambda > 0, \lambda \in (0, \Lambda)$.

Initialize: set $\kappa = \frac{\Lambda}{\lambda},\ \tau = \frac{\kappa+1}{\kappa-1},\ \omega = \frac{2}{\lambda+\Lambda},\ r_0 = b - Ax_0,\ p_1 = \omega r_0, \alpha_0 = 2, \alpha_1 = \frac{1}{1-\frac{\alpha_0}{4\tau^2}}$.

Iterate: for $k = 1, 2, \ldots$ do:

set $x_k = x_{k-1} + p_k$

set $r_k = r_{k-1} - Ap_k$

if $r_k = 0$ (or in practice, small), then terminate

set $\alpha_{k+1} = \frac{1}{1-\frac{\alpha_k}{4\tau^2}}$

set $p_{k+1} = \alpha_{k+1}\omega r_k + (\alpha_{k+1} - 1)p_k$

Algorithm 9.5 is very similar to Algorithm 7.8. The only essential difference is how the coefficients of the recurrence are determined. For the CG method, they are adapted to the matrix A and initial residual r_0 – specifically, they correspond to A, r_0-dependent measure Θ, which was discussed in Remark 8.1.

In the Chebyshev method, the recurrence coefficients corresponding to the measure of orthogonality of the Chebyshev polynomials mapped to $[\lambda, \Lambda]$, whose only relation to A is through λ and Λ and which is independent of r_0.

Again Lemma 8.2A gives a convergence bound for any number of steps.

Arnoldi decomposition

The CG method is based on constructing a sequence $\{p_k\}_{k\in\mathbb{N}}$ such that $p_1, \ldots, p_k$ is an A-orthogonal basis for $\mathcal{K}_k(A, r_0)$ $\forall k \in \mathbb{N}$. This is clearly not possible when A is not Hermitian positive definite. In such casesm we need to give up the A-orthogonality, but still can construct a sequence $\{v_k\}_{k\in\mathbb{N}}$ such that $v_1, \ldots, v_k$ form a basis for $\mathcal{K}_k(A, r_0)$ that is orthogonal w.r.t. the standard inner product $\forall k \in \mathbb{N}$. This can be done by Gram-Schmidt orthogonalization, resulting in what is called *Arnoldi decomposition*.

Algorithm 9.6 (Arnoldi decomposition)**.**

Given: $A \in \mathbb{F}^{n\times n}$, $v \in \mathbb{F}^n$

Initialize: set $v_1 = \frac{v}{\|v\|_2}$

Orthogonalize: for $k = 1, 2, \ldots, n-1$ do:

set $w_{k+1} = Av_k$

set $h_{ik} = v_i^* w_{k+1}$ for $i = 1, \ldots, k$

set $\tilde{v}_{k+1} = w_{k+1} - \sum_{i=1}^{k} h_{ik} v_i$

set $h_{k+1,k} = \|\tilde{v}_{k+1}\|_2$

if $h_{k+1,k} = 0$, then terminate

set $v_{k+1} = \frac{\tilde{v}_{k+1}}{k+1}$ **TODO: typo**

Output: *Arnoldi vectors* $v_1, \ldots, v_k$ and h_{ij} with $j \in \{1, \ldots, m\}$ and $i \in \{1, \ldots, j+1\}$, where m is the number of steps performed.

Algorithm 9.6 terminates at step $m = \max_{k \in \mathbb{N}} \dim \mathcal{K}_k(A, v)$ (see Lemma 8.1A) and produces a matrix $V_m = [v_1, \ldots, v_m] \in \mathbb{F}^{n \times m}$ with orthonormal columns and an *upper Hessenberg matrix*

$$H_m = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1m} \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2m} \\ 0 & h_{32} & h_{33} & \cdots & h_{3m} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{mm-1} & h_{mm} \end{pmatrix} \in \mathbb{F}^{m \times m}.$$

($A \in \mathbb{F}^{m \times m}$ is upper Hessenberg if $a_{ij} = 0$ for $i > j + 1$)

For any $k \in \mathbb{N}$, let $V_k = [v_1, \ldots, v_k] \in \mathbb{F}^{n \times k}$, $H_k \in \mathbb{F}^{k \times k}$ as above and $\hat{H}_k = \left[\begin{array}{c} H_k \\ \hline 0 \ldots 0 \; h_{k+1,k} \end{array}\right] \in \mathbb{F}^{(k+1) \times k}$.

Then we have the following relations:

$$V_k^* A V_k = H_k \quad \text{and} \quad V_{k+1}^* A V_k = \hat{H}_k$$

For each $k \in \{1, \ldots, m\}$, $v_1, \ldots, v_k$ form an orthonormal basis for $\mathcal{K}_k(A, v)$

Remark 9.7. If the Arnoldi decomposition is constructed for an Hermitian matrix A, then $H_m = V_m^* A V_m$ is also Hermitian and hence tridiagonal. In this case, the Arnoldi decomposition is called *Lanczos decomposition* and the Arnoldi vectors are called *Lanczos vectors*.

We will consider, for $A \in \mathbb{F}^{n \times n}$ invertible, $b, x_0 \in \mathbb{F}^n$ and $r_0 = b - Ax_0$, two Krylov methods for the approximate solution of $Ax = b$.

(i) The *Arnoldi method* defines for each $k \in \mathbb{N}$, $x_k \in x_0 + \mathcal{K}_k(A, r_0)$ by

$$r_k = b - Ax_k \perp \mathcal{K}_k(A, r_0).$$

When A is Hermitian positive definite, the method is called *Lanczos method*. In this case, the iterates coincide (in exact arithmetic) with the iterates of the CG method (see Lemma 7.4) The two methods, however, use different bases to construct the same iterates.

(ii) The *method of generalized minimal residual* (GMRES) defines, for each $k \in \mathbb{N}$, $x_k \in x_0 + \mathcal{K}_k(A, r_0)$ by minimizing $\|r_k\|_2$.

Note that $\dim \mathcal{K}_{k+1}(A, r_0) = k$ holds for $k \in \mathbb{N}_0$ if and only if $A^{-1}r_0 \in \mathcal{K}_k(A, r_0) \setminus K_{k-1}(A, r_0)$, as we showed in Lemma 8.1A assuming only invertibility of A, in which case, for $r_0 \neq 0$, these two methods, and any other Krylov method doing something reasonable on Krylov subspaces, find the exact solution if and only if $k = \max_{\ell \in \mathbb{N}} \dim \mathcal{K}_\ell(A, r_0)$.