§1 BASIC ITERATIVE METHODS

§1.1 PRELIMINARIES

RECALL: A mapping $\|\cdot\| : X \to \mathbb{R}$ on a vector space X over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ is a norm, if the following three conditions hold:

(i) $\forall x \in X : (\|x\| = 0 \Rightarrow x = 0)$ — $\|\cdot\|$ is definit

(ii) $\forall x \in X \; \forall \lambda \in \mathbb{K} : \|\lambda x\| = |\lambda| \, \|x\|$ — $\|\cdot\|$ is homogeneous

(iii) $\forall x, y \in X : \|x + y\| \leq \|x\| + \|y\|$ — $\|\cdot\|$ satisfies the triangle inequality

---

DEF: If $\|\cdot\|$ is a norm on $\mathbb{K}^n$, we define the induced matrix norm

$$\|A\| := \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} \quad \text{for all } A \in \mathbb{K}^{n \times n},$$

for which we use the same notation $\|\cdot\|$ as for the norm on $\mathbb{K}^n$

LEMMA: If $\|\cdot\|$ is a norm on $\mathbb{K}^n$, then the induced matrix norm is indeed a norm on $\mathbb{K}^{n \times n}$. For $A, B \in \mathbb{K}^{n \times n}$ and $x \in \mathbb{K}^n$, there hold $\|Ax\| \leq \|A\| \, \|x\|$ and $\|AB\| \leq \|A\| \, \|B\|$. Moreover, for any $\varepsilon > 0$, there exists $x_\varepsilon \in \mathbb{K}^n$ with $\|x_\varepsilon\| = \varepsilon$ such that $\|A\| = \dfrac{\|Ax_\varepsilon\|}{\|x_\varepsilon\|}$.

Proof: Let $A, B \in \mathbb{K}^n$, $\lambda \in \mathbb{K}^n$.

- If $\|A\| = 0$, then $Ax = 0$ for all $x \in \mathbb{K}^n$

  Hence, $A = 0$.

- Since $\|\lambda Ax\| = |\lambda| \|Ax\|$ for all $x \in \mathbb{K}^n$, it follows that $\|\lambda A\| = |\lambda| \|A\|$.
- Since $\|(A+B)x\| \leq \|Ax\| + \|Bx\|$ for all $x \in \mathbb{K}^n$, it follows that $\|A+B\| \leq \|A\| + \|B\|$

Hence, the matrix norm is a norm on $\mathbb{K}^{n\times n}$. //

- For $x \neq 0$, it holds $\|Ax\| = \frac{\|Ax\|}{\|x\|} \cdot \|x\| \leq \|A\| \|x\|$.
  Since this holds trivially for $x = 0$ (with equality), we are led (inequality) to $\|Ax\| \leq \|A\| \|x\|$. //
- This also proves $\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\|$ and hence $\|AB\| \leq \|A\| \|B\|$. //

It remains to prove that the supremum is indeed a maximum.
Define the sphere $S := \{x \in \mathbb{K}^n \mid \|x\| = 1\}$.
Note that S is bounded and closed (as the preimage of the closed set $\{1\}$).
Hence, S is compact and $\|A(\cdot)\|$ is continous, since $|\|Ax\| - \|Ay\|| \leq \|Ax - Ay\| \leq \|A\| \|x-y\|$.
Therefore, there exists $x_1 \in \mathbb{K}^n$ with $\|x_1\| = 1$ such that

$$\|Ax_1\| = \max_{x \in S} \|Ax\| \leq \|A\| = \sup_{x \in \mathbb{K}^n \setminus \{0\}} \underbrace{\frac{\|Ax\|}{\|x\|}}_{= \|A\left(\frac{x}{\|x\|}\right)\|} \leq \|Ax_1\|.$$

$\frac{x}{\|x\|} \in S$

For any $\varepsilon > 0$, we may thus consider $x_\varepsilon := \varepsilon x_1$, since $\|x_\varepsilon\| = \varepsilon$, $\frac{\|Ax_\varepsilon\|}{\|x_\varepsilon\|} = \|Ax_1\|$. ∎

DEF: Let $A \in \mathbb{K}^{n\times n}$ be regular (or invertible), i.e., $\det A \neq 0$. Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$.
Then, $\operatorname{cond}_{\|\cdot\|}(A) := \|A\|\,\|A^{-1}\|$ is called condition number (of A with respect to $\|\cdot\|$)

LEMMA: Let $A \in \mathbb{K}^{n\times n}$ be regular. Let $x, \tilde{x}, b, \tilde{b} \in \mathbb{K}^n$ with $b \neq 0$ as well as $Ax = b$, $A\tilde{x} = \tilde{b}$. Then, it holds that

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leq \operatorname{cond}_{\|\cdot\|}(A)\,\frac{\|b-\tilde{b}\|}{\|b\|}. \qquad (*)$$

Moreover, this estimate is sharp: Then exists $b \in \mathbb{K}^n \setminus \{0\}$ such that for all $\varepsilon > 0$, then exists $\tilde{b} \in \mathbb{K}^n$ with $\|b-\tilde{b}\| \leq \varepsilon$ such that the estimate $(*)$ holds with equality.

REMARK: The lemma states that $\operatorname{cond}_{\|\cdot\|}(A)$ is the worst-case factor for error propagation if the computer uses a rounded $\tilde{b}$ instead of the exact b, under the idealistic assumption that computer arithmetics is exact. Recall that direct solvers like Gaussian elemination factorize the matrix $A = LU$ and solve sequentially with L and U. Inherently, this can lead to error propagation with $\operatorname{cond}_{\|\cdot\|}(L)\operatorname{cond}_{\|\cdot\|}(U) \geq \operatorname{cond}_{\|\cdot\|}(A)$, where $\operatorname{cond}_{\|\cdot\|}(A)$ can be (factor) much smaller.

Proof:

Note that • $\|x-\tilde{x}\| = \|A^{-1}(b-\tilde{b})\| \leq \|A^{-1}\|\,\|b-\tilde{b}\|$

• $\|b\| = \|Ax\| \leq \|A\|\,\|x\|$.

Combining this, we prove $\dfrac{\|x-\tilde{x}\|}{\|x\|} \leq \underbrace{\|A\|\|A^{-1}\|}_{=\operatorname{cond}_{\|\cdot\|}(A)} \dfrac{\|b-\tilde{b}\|}{\|b\|}$ //

To see that the estimate is sharp, we argue as follows:

• Choose $x \in \mathbb{K}^n \setminus \{0\}$ and $b := Ax$ such that $\|A\| = \dfrac{\|Ax\|}{\|x\|} = \dfrac{\|b\|}{\|x\|}$

• For $\varepsilon > 0$, choose $y_\varepsilon \in \mathbb{K}^n \setminus \{0\}$ with $\|y_\varepsilon\| = \varepsilon$ such that

$$\|A^{-1}\| = \frac{\|A^{-1}y_\varepsilon\|}{\|y_\varepsilon\|}$$

• Choose $\tilde{b} := b - y_\varepsilon$ and $\tilde{x} := A^{-1}\tilde{b}$

This leads to $x-\tilde{x} = A^{-1}(b-\tilde{b}) = A^{-1}y_\varepsilon$ and hence

$$\frac{\|x-\tilde{x}\|}{\|x\|} = \frac{\|A^{-1}y_\varepsilon\|}{\|x\|} = \underbrace{\frac{\|A^{-1}y_\varepsilon\|}{\|y_\varepsilon\|}}_{=\|A^{-1}\|} \cdot \underbrace{\frac{\|y_\varepsilon\|}{\|x\|}}_{=\frac{\|b-\tilde{b}\|}{\|x\|}} = \|A^{-1}\|\,\frac{\|b-\tilde{b}\|}{\|b\|}\,\underbrace{\frac{\|b\|}{\|x\|}}_{=\|A\|}$$

$$= \underbrace{\|A^{-1}\|\,\|A\|}_{=\operatorname{cond}_{\|\cdot\|}(A)} \frac{\|b-\tilde{b}\|}{\|b\|} \quad \text{und} \quad \|b-\tilde{b}\| = \|y_\varepsilon\| = \varepsilon$$

□

EXAMPLE ($\|\cdot\|_1$ induces the column-sum norm):

We consider $\|x\|_1 := \sum_{j=1}^{n} |x_j|$ on $\mathbb{K}^n$.

For $A \in \mathbb{K}^{n\times n}$, it follows that

$$\|Ax\|_1 = \sum_{j=1}^{n} |(Ax)_j| = \sum_{j=1}^{n} \Big| \sum_{k=1}^{n} A_{jk} x_k \Big| \leq \sum_{k} \Big(\sum_{j} |A_{jk}| \Big) |x_k|$$

$$\leq \Big(\max_{k=1,\dots,n} \sum_{j=1}^{n} |A_{jk}| \Big) \underbrace{\sum_{k=1}^{n} |x_k|}_{= \|x\|_1}.$$

This yields $\|A\|_1 \leq \max_{k} \sum_{j} |A_{jk}|$, the so-called column-sum norm of A.

To see equality, fix the index $k_0 \in \{1,\dots,n\}$ such that

$$\sum_{j=1}^{n} |A_{jk_0}| = \max_{k} \sum_{j} |A_{jk}|$$

Choose $x = e_{k_0}$ k_0-th unit vector

$\Rightarrow \|x\|_1 = 1$ and $\sum_{j} \Big| \sum_{k} A_{jk} \underbrace{x_k}_{=\delta_{kk_0}} \Big| = \sum_{j} |A_{jk_0}| = \max_{k} \sum_{j} |A_{jk}|$. ∎

EXAMPLE ($\|\cdot\|_\infty$ induces the row-sum norm):

We consider $\|x\|_\infty := \max_{j=1,\dots,n} |x_j|$ on $\mathbb{K}^n$

For $A \in \mathbb{K}^{n\times n}$, it follows that

$$\|Ax\|_\infty = \max_{j} \Big| \sum_{k} A_{jk} x_k \Big| \leq \max_{j} \sum_{k} |A_{jk}| \underbrace{|x_k|}_{\leq \|x\|_\infty}$$

$\Rightarrow \|A\|_\infty \leq \max_{j} \sum_{k} |A_{jk}|$, the so-called row-sum norm of A

To see equality, pick index $j_0 \in \{1,\dots,n\}$ with maximal value.

Then, define $x_k := \operatorname{sign}(A_{j_0 k})$, i.e., $|x_k| = 1$ and $A_{j_0 k} x_k = |A_{j_0 k}|$

With this, then holds $\|x\|_\infty = 1$ and equality. ∎

RECALL: • Each matrix $A \in \mathbb{K}^{n\times n}$ is trigonalizable over $\mathbb{C}$, i.e., there exists a regular matrix $T \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $U \in \mathbb{C}^{n\times n}$ s.t. $U = T^{-1}AT$ with $U_{ij} = 0$ for $i > j$.

• The diagonal elements $\lambda_j := U_{jj}$ are the eigenvalues of U and A, i.e., ther exist $v_j \in \mathbb{C}^n \setminus \{0\}$ such that $Av_j = \lambda_j v_j$.

• Each Hermitian matrix $A = A^H := \bar{A}^T \in \mathbb{K}^{n\times n}$ is diagonalizable over $\mathbb{K}$, i.e., ther exists a regular (and even orthogonal) matrix $Q \in \mathbb{K}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ s.t. $D = Q^HAQ$ with $D_{ij} = 0$ for $i \neq j$ and $Q^HQ = I$.

• The diagonal elements $\lambda_j := D_{jj} \in \mathbb{R}$ are the eigenvalues of D and A. Moreover, the columns v_j of Q are the eigenvectors of A, and $\{v_1, \dots, v_n\} \subseteq \mathbb{K}^n$ is an orthonormal basis of $\mathbb{K}^n$ with respect to the Euclidean scalar product

$$(x,y)_2 := x^Hy \quad \text{for } x, y \in \mathbb{K}^n.$$

DEF: For a matrix $A \in \mathbb{K}^{n\times n}$, let $\sigma(A) := \{\lambda_1, \dots, \lambda_n\}$ be the set of its eigenvalues, i.e., the so-called spectrum of A. Moreover, $\rho(A) := \max_{j=1,\dots,n} |\lambda_j|$ denotes the so-called spectral radius of A

EXAMPLE ($\|A\|_2 = \sqrt{\rho(A^H A)}$ so-called spectral norm).

Let $\{v_1, \dots, v_n\} \subseteq \mathbb{K}^n$ orthonormal basis with $A^H A v_j = \lambda_j v_j$
and $|\lambda_1| \leq \dots \leq |\lambda_n| = \rho(A^H A)$

Write $x = \sum_{j=1}^{n} \mu_j v_j$ and note that $\|x\|_2 := \left(\sum_{j=1}^{n} |x_j|^2 \right)^{1/2} \overset{!}{=} \left(\sum_{j=1}^{n} |\mu_j|^2 \right)^{1/2}$:

Indeed, $\|x\|_2^2 = (x, x)_2 = \left(\sum_j \mu_j v_j, \sum_k \mu_k v_k \right)_2 = \sum_{j,k} \bar{\mu}_j \mu_k \underbrace{(v_j, v_k)_2}_{= \delta_{jk}}$

$= \sum_j |\mu_j|^2$.

Similarly, we get

$$\|Ax\|_2^2 = (Ax, Ax)_2 = (A^H A x, x)_2 = \sum_{j,k} \bar{\mu}_j \mu_k \underbrace{(A^H A v_j, v_k)_2}_{= \lambda_j (v_j, v_k)_2 = \delta_{jk}}$$

$$= \sum_j |\mu_j|^2 \lambda_j \leq |\lambda_n| \sum_j |\mu_j|^2 = \rho(A^H A) \|x\|_2^2$$

This proves $\|Ax\|_2^2 \leq \rho(A^H A)$.

For $x = v_n$, we even see equality. ∎

EXAMPLE ($\|A\|_2 = \rho(A)$ for $A = A^H$):

Let $\{v_1, \dots, v_n\} \subseteq \mathbb{K}^n$ orthonormal basis with $A v_j = \lambda_j v_j$ and
$|\lambda_1| \leq \dots \leq |\lambda_n| = \rho(A)$.

Write $x = \sum_j \mu_j v_j$.

Then, $\|Ax\|_2^2 = (Ax, Ax)_2 = \sum_{j,k} \bar{\mu}_j \mu_k \underbrace{(A v_j, A v_k)_2}_{= \lambda_j \lambda_k (v_j, v_k)_2} = \sum_j |\mu_j|^2 \lambda_j^2$

$$\leq |\lambda_n|^2 \sum_j |\mu_j|^2 = \rho(A)^2 \|x\|_2^2.$$

This proves $\|Ax\|_2^2 \leq \rho(A)^2$.

For $x = v_n$, we again see equality. ∎

---

§ 1.2 An application of Banach's fixed point theorem

Banach fixed point theorem: Let X be a Banach space, i.e., a complete normed space. Let $0 < q < 1$ and let $\Phi: X \to X$ satisfy $\|\Phi(x) - \Phi(y)\| \leq q \|x - y\| \quad \forall x, y \in X$.

Then, there exists a unique $x^* \in X$ s.t. $\Phi(x^*) = x^*$.

Moreover, it holds that:

(i) for each $x_0 \in X$ and $x_{k+1} := \Phi(x_k)$ for all $k \in \mathbb{N}_0$, there holds $\|x^* - x_k\| \to 0$ as $k \to \infty$

(ii) $\|x^* - x_{k+1}\| \leq q \|x^* - x_k\|$

(iii) $\|x^* - x_{k+1}\| \leq \frac{q}{1-q} \|x_{k+1} - x_k\| \leq \frac{q^{k+1}}{1-q} \|x_1 - x_0\|$

(iv) $\|x_{k+1} - x_k\| \leq (1+q) \|x^* - x_k\|$

Remark:

- The first estimate in (iii) and the estimate in (iv) are so-called a-posteriori error estimates: We can control the error from above and below from the computed iterates x_k, x_{k+1}, even if we do not know the exact solution x^*.
- The estimate $\|x^* - x_{k+1}\| \leq \frac{q^{k+1}}{1-q} \|x_1 - x_0\|$ is an a-priori error estimate: Before computing x_{k+1} and only by knowledge of q, x_0, x_1, we know how many iterations it needs at most to drive the error below a prescribed tolerance.

Proof:

Step 1 (uniqueness): For $x^* = \Phi(x^*)$ and $y^* = \Phi(y^*)$, it holds that

$$\|x^* - y^*\| = \|\Phi(x^*) - \Phi(y^*)\| \leq q \|x^* - y^*\|.$$

Hence, $x^* = y^*$ since $0 < q < 1$. //

Step 2 (the limit will be a fixed point): For $x_0 \in X$, consider $x_{k+1} := \Phi(x_k)$.

Suppose that the limit $x^* := \lim_{k\to\infty} x_k$ exists.

Then, $x^* = \lim_k x_k = \lim_k x_{k+1} = \lim_k \Phi(x_k) \overset{\Phi \text{ continuous}}{=} \Phi(\lim_k x_k) = \Phi(x^*)$. //

Step 3 ($(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence): For $m \leq n$, it holds that

$$\|x_n - x_m\| \leq \sum_{k=m}^{n-1} \underbrace{\|x_{k+1} - x_k\|}_{\begin{array}{l} = \|\Phi(x_k) - \Phi(x_{k-1})\| \\ \leq q \|x_k - x_{k-1}\| \\ \vdots \\ \leq q^k \|x_1 - x_0\| \end{array}} \leq \|x_1 - x_0\| \underbrace{\sum_{k=m}^{n-1} q^k}_{= q^m \sum_{k=0}^{n-m-1} q^k \leq \frac{1}{1-q}} \leq \|x_1 - x_0\| \frac{q^m}{1-q}$$

Note that the right-hand side tends to zero as $m \to \infty$.

Hence, (x_k) is a Cauchy sequence.

Since X is a Banach space, the limit $x^* := \lim_{k\to\infty} x_k$ exists

(and is a fixed point)

Step 4 (error estimates):

$$\|x^* - x_{k+1}\| = \|\Phi(x^*) - \Phi(x_k)\| \leq q \underbrace{\|x^* - x_k\|}_{\leq \|x^* - x_{k+1}\| + \|x_{k+1} - x_k\|}$$

Hence, $\|x^* - x_{k+1}\| \leq \frac{q}{1-q} \underbrace{\|x_{k+1} - x_k\|}_{\leq q^k \|x_1 - x_0\|}$

And, $\|x_{k+1} - x_k\| \leq \|x^* - x_{k+1}\| + \|x^* - x_k\| \leq (1+q) \|x^* - x_k\|$ ▨

LEMMA (Characterizing the spectral radius): Let $A \in \mathbb{K}^{n\times n}$.

Then, $\rho(A) = \inf\{\|A\| \mid \|\cdot\| \text{ is matrix norm induced by norm on } \mathbb{K}^n\}$.

If $\sigma(A) \subseteq \mathbb{K}$, then norms on $\mathbb{K}^n$ (instead of $\mathbb{C}^n$) suffice.

Proof: „$\leq$": Let $\|\cdot\|$ be norm on $\mathbb{C}^n$. Let $\lambda \in \sigma(A)$, $x \in \mathbb{C}^n \setminus \{0\}$ with $Ax = \lambda x$.

Then, $|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \|x\|$

Hence, $|\lambda| \leq \|A\|$. //

If $\sigma(A) \subseteq \mathbb{K}$, then we can choose $x \in \mathbb{K}^n \setminus \{0\}$ and the same proof applies. Therefore, norms on $\mathbb{K}^n$ suffice in this case. //

Proof „$\geq$": According to linear Algebra, A is triangonalizable:

$$U := T^{-1}AT = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \vdots \\ 0 & & u_{nn} \end{pmatrix} \text{ with } T \in \mathbb{C}^{n\times n} \text{ regular}, U \in \mathbb{C}^{n\times n}.$$

Note that $\rho(A) = \max_{j=1,\dots,n} |u_{jj}|$, since $\sigma(A) = \sigma(U) = \{u_{11}, \dots, u_{nn}\}$.

Let $\varepsilon > 0$. We show that there exists norm $\|\cdot\|_\varepsilon$ on $\mathbb{C}^n$ with $\|A\|_\varepsilon \leq \rho(A) + C\cdot\varepsilon$ for some fix $C > 0$

Let $D_\varepsilon := \operatorname{diag}(1, \varepsilon, \dots, \varepsilon^{n-1}) = \begin{pmatrix} 1 & & 0 \\ & \varepsilon & \\ & & \ddots \\ 0 & & & \varepsilon^{n-1} \end{pmatrix}$

Define $\|x\|_\varepsilon := \|D_\varepsilon^{-1} T^{-1} x\|_\infty$

Note that $\|A\|_\varepsilon = \sup_{x\in\mathbb{C}^n\setminus\{0\}} \frac{\|Ax\|_\varepsilon}{\|x\|_\varepsilon} = \sup_{x\in\mathbb{C}^n\setminus\{0\}} \frac{\|D_\varepsilon^{-1}T^{-1}Ax\|_\infty}{\|\underbrace{D_\varepsilon^{-1}T^{-1}x}_{=:y}\|_\infty}$

($y = D_\varepsilon^{-1}T^{-1}x$, $x = TD_\varepsilon y$)

$$= \sup_{y\in\mathbb{C}^n\setminus\{0\}} \frac{\|D_\varepsilon^{-1}T^{-1}ATD_\varepsilon y\|_\infty}{\|y\|_\infty}$$

$$= \|D_\varepsilon^{-1}\overbrace{T^{-1}AT}^{=U}D_\varepsilon\|_\infty = \|D_\varepsilon^{-1}UD_\varepsilon\|_\infty$$

Note that

- multiplication with diagonal matrix from the left = scaling of rows
- multiplication with diagonal matrix from the right = scaling of columns

$$D_\varepsilon^{-1} U D_\varepsilon = \begin{pmatrix} u_{11} & \varepsilon u_{12} & \cdots & \varepsilon^{n-1} u_{1n} \\ & u_{22} & \cdots & \varepsilon^{n-2} u_{2n} \\ & & \ddots & \vdots \\ 0 & & & u_{nn} \end{pmatrix}$$

Recall that $\|A\|_\infty = \max_j \sum_k |A_{jk}|$ • is the so-called row-sum norm.

Hence, $\|A\|_\varepsilon = \|D_\varepsilon^{-1} U D_\varepsilon\|_\infty = \max_j \left(|u_{jj}| + \sum_{k=j+1}^{n} \varepsilon^{k-j} |u_{jk}| \right)$

$$\leq \underbrace{\max_j |u_{jj}|}_{= \rho(A)} + \max_j \underbrace{\sum_{k=j+1}^{n} \varepsilon^{k-j} |u_{jk}|}_{\leq \varepsilon \sum_{k=j+1}^{n} \varepsilon^{k-(j+1)} |u_{jk}| \leq C}$$

— — —

Note that the estimate „$\geq$" is independent of $\sigma(A) \subseteq \mathbb{K}$, since $\|\cdot\|_\varepsilon$ is of course a norm on $\mathbb{K}^n$ if it is (even) a norm on $\mathbb{C}^n$. ▨

THEOREM (global convergence of basic iterative methods):

Let $\Pi \in \mathbb{K}^{n\times n}$. Define $\Phi : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}^n$, $\Phi(b,x) := \Pi x + b$.

Then, the following is equivalent:

(i) $\varrho(\Pi) < 1$

(ii) $\forall b \in \mathbb{K}^n \; \exists x^* \in \mathbb{K}^n \; \forall x_0 \in \mathbb{K}^n$, $x_{k+1} := \Phi(b, x_k)$: $x^* = \lim_{k\to\infty} x_k$.

In this case $x^* = \Phi(b, x^*)$ is uniquely determined by (Π, b), and there exists $\|\cdot\|$ norm on $\mathbb{K}^n$ such that $0 < q := \|\Pi\| < 1$ and

$$\|\Phi(b,x) - \Phi(b,y)\| \le q \, \|x-y\| \qquad \forall b \in \mathbb{K}^n \; \forall x, y \in \mathbb{K}^n$$

Proof (i ⇒ ii):

According to the previous lemma, there exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ (and hence on $\mathbb{K}^n$) with $\|\Pi\| = q < 1$.

For any $b \in \mathbb{K}^n$, $x, y \in \mathbb{K}^n$, this yields

$$\|\Phi(b,x) - \Phi(b,y)\| = \|\Pi x - \Pi y\| \le \underbrace{\|\Pi\|}_{=q} \|x-y\|$$

Hence, $\Phi(b,\cdot) : \mathbb{K}^n \to \mathbb{K}^n$ satisfies Banach's fixed point theorem. ✓

— — —

Proof (ii ⇒ i) for $\mathbb{K} = \mathbb{C}$:

Let $b \in \mathbb{C}^n$, choose x^* accordingly.

Recall that limits of fixed point iterations are fixed points $x^* = \Phi(b, x^*)$.

Let $\lambda \in \sigma(\Pi)$ with $|\lambda| = \varrho(\Pi)$.

Let $v \in \mathbb{C}^n$ with $\|v\| = 1$ and $\Pi v = \lambda v$.

Choose $x_0 := v + x^* \in \mathbb{C}^n$

Note $x^* - x_1 = \Phi(b, x^*) - \Phi(b, x_0) = \Pi(\underbrace{x^* - x_0}_{=v}) = \lambda(\underbrace{x^* - x_0}_{v})$.

Hence $x^* - x_k = \lambda^k v$ by induction.

Therefore, $\varrho(\Pi)^k = |\lambda|^k = \|\lambda^k v\| = \|x^* - x_k\| \to 0$ by (ii)

Therefore, $\varrho(\Pi) < 1$. ✓

Proof "(ii) for $\mathbb{K}=\mathbb{R} \Rightarrow$ (ii) for $\mathbb{K}=\mathbb{C}$":

Note that the previous argument cannot be applied to prove (i), since $\lambda \notin \mathbb{R}$ in general.

Hence, we argue differently and suppose that (ii) holds for $\mathbb{K}=\mathbb{R}$

Note that $\mathrm{Re}(\Pi x) = \Pi(\mathrm{Re}\, x)$, $\mathrm{Im}(\Pi x) = \Pi(\mathrm{Im}\, x)$ for all $x \in \mathbb{C}^n$, since $\Pi \in \mathbb{R}^{n\times n}$!

Let $b \in \mathbb{C}^n$.

Apply (ii) for $\mathrm{Re}\, b, \mathrm{Im}\, b \in \mathbb{R}^n$.

Obtain $x^* \in \mathbb{C}^n$ s.t. (ii) holds for $(\mathrm{Re}\, b, \mathrm{Re}\, x^*)$ and $(\mathrm{Im}\, b, \mathrm{Im}\, x^*)$.

Let $x_0 \in \mathbb{C}^n$, $x_{k+1} := \Phi(b, x_k) = \Pi x_k + b$

Then, $\mathrm{Re}\, x_{k+1} = \Pi(\mathrm{Re}\, x_k) + \mathrm{Re}\, b \to \mathrm{Re}\, x^*$

$\mathrm{Im}\, x_{k+1} = \Pi(\mathrm{Im}\, x_k) + \mathrm{Im}\, b \to \mathrm{Im}\, x^*$.

Hence, $x_{k+1} \to x^*$ in $\mathbb{C}^n$, so that (ii) holds for $\mathbb{C}$.

Now, (i) follows. ∎

REMARK: For the proof for $\mathbb{K}=\mathbb{C}$, it suffices that

$$\exists\, b \in \mathbb{C}^n\ \exists\, x^* \in \mathbb{C}^n\ \forall\, x_0 \in \mathbb{C}^n,\ x_{k+1} := \Phi(b, x_k):\quad x^* = \lim_{k\to\infty} x_k,$$

i.e., "$\exists\, b \in \mathbb{C}^n$" instead of "$\forall\, b \in \mathbb{C}^n$".

EXAMPLE (DAMPED RICHARDSON ITERATION):

Let $A \in \mathbb{K}^{n\times n}$ regular, $b \in \mathbb{K}^n$

We seek $x^* \in \mathbb{K}^n$ with $Ax^* = b$.

Note that $-\lambda A x^* + \lambda b = 0$

Hence, consider $\Phi(b, x) := \underbrace{(I - \lambda A)}_{=\Pi} x + \lambda b = x + \lambda \underbrace{(b - Ax)}_{\text{so-called residual}}$

Known: Iteration converges if $\rho(\Pi) < 1$. ∎

PROPOSITION (Convergence of damped Richardson iteration):

Let $A \in \mathbb{K}^{n\times n}$ regular with $\mathrm{Re}(\mu) > 0$ for all $\mu \in \sigma(A)$.

Then, there exists $\lambda > 0$ such that $\rho(I - \lambda A) < 1$, i.e.

In this case, the damped Richardson iteration leads to convergence for any $b \in \mathbb{K}^n$ and any initial guess $x_0 \in \mathbb{K}^n$. The limit

$$x^* = \lim_{k\to\infty} x_k, \quad x_{k+1} := \Phi(b, x_k) \text{ solves } Ax^* = b.$$

Proof:

Recall that A is triagonalizable, i.e., there exist $T \in \mathbb{C}^{n\times n}$ regular and $U \in \mathbb{C}^{n\times n}$ upper triangular with $A = T^{-1}UT$.

In particular, $\sigma(A) = \{u_{jj} \mid j = 1, \dots, n\}$.

Moreover, $I - \lambda A = T^{-1}\underbrace{(I - \lambda U)}_{=:\tilde{U}}T$ with $\tilde{U}$ upper triangular.

In particular, $\sigma(I - \lambda A) = \{\tilde{u}_{jj} \mid j = 1, \dots, n\}$.

This yields

$$\begin{aligned}
|\tilde{u}_{jj}|^2 &= |1 - \lambda\,\mathrm{Re}\,u_{jj}|^2 + |\lambda\,\mathrm{Im}\,u_{jj}|^2 \\
&= 1 - 2\lambda\,\mathrm{Re}\,u_{jj} + \underbrace{\lambda^2|\mathrm{Re}\,u_{jj}|^2 + \lambda^2|\mathrm{Im}\,u_{jj}|^2}_{=\lambda^2|u_{jj}|^2} \\
&= 1 - \lambda\underbrace{\left(2\underbrace{\mathrm{Re}\,u_{jj}}_{>0} - \lambda|u_{jj}|^2\right)}_{>0 \text{ for } \lambda \text{ sufficiently small}} < 1 \quad \text{for } \lambda \text{ sufficiently small.}
\end{aligned}$$

Since $\rho(I - \lambda A) = \max_j |\tilde{u}_{jj}|$, there exists $\lambda_0 > 0$ such that

$\rho(I - \lambda A) < 1$ for all $0 < \lambda \leq \lambda_0$.

The rest of the claims follows from the convergence theorem. ∎

COROLLARY: Let $A = A^H \in \mathbb{K}^{n\times n}$ be regular and Hermitian.
Suppose that A is positive definite, i.e., $(Ax, x)_2 > 0 \quad \forall x \in \mathbb{K}^n \setminus \{0\}$.
Then, any $0 < \lambda < \frac{2}{\rho(A)}$ guarantees convergence of the damped Richardson iteration in the sense of $\rho(I - \lambda A) < 1$.
In particular, $\rho(A) < 2$ even yields convergence of the (undamped) Richardson iteration, i.e., $\rho(I - A) < 1$.

Proof: Note that $\sigma(A) \subset \mathbb{R}_{>0}$.
We argue as before with

$$|\tilde{u}_{jj}| = |1 - \lambda u_{jj}|.$$

($\lambda u_{jj} > 0$)

If $\lambda u_{jj} \leq 1$, then $|\tilde{u}_{jj}| = |1 - \lambda u_{jj}| = 1 - \lambda u_{jj} < 1$

If $\lambda u_{jj} > 1$, then $|\tilde{u}_{jj}| = \lambda u_{jj} - 1 \leq \underbrace{\lambda \rho(A)}_{<2} - 1 < 1.$ ∎

REMARK: The disadvantage of the Richardson iteration and the related proofs is that $\sigma(A)$ is hardly known in general. Therefore, it is rarely possible to choose a viable (i.e., sufficiently small) λ.

REMARK: Convergence of a basic linear iteration can be slow and non-monotonically.
For instance,

$$N := \begin{pmatrix} 0 & m & & \\ & \ddots & \ddots & \\ & & \ddots & m \\ & & & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}$$

satisfies $\rho(N) = 0$, $N^n = 0$, but $\|N\|_1 = \|N\|_2 = \|N\|_\infty = m$.
One observes essentially no convergence for $k \leq n$ and $x^* = \sum_{j=0}^{n-1} N^j b$

REMARK: Let $\|\cdot\|$ be a norm on $\mathbb{K}^n$, $A \in \mathbb{K}^{n\times n}$ with $\rho(A) < 1$ and $b, x_0 \in \mathbb{K}^n$. Consider the asymptotic error contraction of $x_{k+1} := \Phi(b, x_k) = Ax_k + b$

$$q := \limsup_{k\to\infty} \left(\frac{\|x^* - x_k\|}{\|x^* - x_0\|}\right)^{1/k} = \limsup_k \left(\frac{\|A^k(x^* - x_0)\|}{\|x^* - x_0\|}\right)^{1/k} \leq \limsup_k \|A^k\|^{1/k} = \rho(A),$$

(Gelfand theorem)

i.e., $\rho(A)$ gives quite sharp bound on asymptotic error contraction.

THEOREM (Gelfand): Let $A \in \mathbb{K}^{n\times n}$ and $\|\cdot\|$ a norm on $\mathbb{K}^n$.

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

If $\|\cdot\|$ is even a norm on $\mathbb{C}^n$, then $\|A^k\|^{1/k} \geq \rho(A)$.

(or $\sigma(A) \subseteq \mathbb{K}$)

Proof:

(or $\sigma(A) \subseteq \mathbb{K}$)

① If $\|\cdot\|$ is a norm on $\mathbb{C}^n$, let $\lambda \in \sigma(A)$ with $|\lambda| = \rho(A)$.

Let $v \in \mathbb{C}^n \setminus \{0\}$ with $Av = \lambda v$ and $\|v\| = 1$.

Note that we can choose $v \in \mathbb{K}^n \setminus \{0\}$ if $\sigma(A) \subseteq \mathbb{K}$.

$$\Rightarrow \rho(A) = |\lambda| = \|\lambda^k v\|^{1/k} = \|A^k v\|^{1/k} \leq \|A^k\|^{1/k} \underbrace{\|v\|}_{=1}^{1/k}. \quad //$$

---

② Let $0 < \varepsilon < \rho(A)$.

Consider $A_\pm := \frac{1}{\rho(A) \pm \varepsilon} A$.

Then, $\rho(A_\pm) = \frac{\rho(A)}{\rho(A) \pm \varepsilon}$, i.e., $\rho(A_+) < 1$, $\rho(A_-) > 1$.

We show: (with $\lim$)

- $\exists\, k_+ \in \mathbb{N}\ \forall k \geq k_+ : \|A_+^k\| < 1$
- $\exists\, k_- \in \mathbb{N}\ \forall k \geq k_- : \|A_-^k\| > 1$.

For $k \geq \max\{k_+, k_-\}$, it follows that

- $\|\Pi^k\|^{1/k} = (\varrho(\Pi) \pm \varepsilon)\, \|\Pi_\pm^k\|^{1/k}$

and hence

$$(\varrho(\Pi) - \varepsilon) \leq (\varrho(\Pi) - \varepsilon) \underbrace{\|\Pi_-^k\|^{1/k}}_{>1} = \|\Pi^k\|^{1/k} = (\varrho(\Pi) + \varepsilon) \underbrace{\|\Pi_+^k\|^{1/k}}_{<1} \leq \varrho(\Pi) + \varepsilon$$

This proves: $\forall \varepsilon > 0\ \exists k_0 \in \mathbb{N}\ \forall k \geq k_0 : \left| \varrho(\Pi) - \|\Pi^k\|^{1/k} \right| \leq \varepsilon$

and thus $\varrho(\Pi) = \lim_{k\to\infty} \|\Pi^k\|^{1/k}$. //

---

③ Let $B \in \{M_\pm\}$. Recall the Jordan normal form $B = V J V^{-1}$,

where $V \in \mathbb{C}^{n\times n}$ is regular, J is block-diagonal $J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_s \end{pmatrix}$

with $J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_i \end{pmatrix} =: \lambda_i I_i + N_i \in \mathbb{R}^{r_i \times r_i}$

In particular, it holds $\sigma(B) = \{\lambda_1, \dots, \lambda_s\}$. and $N_i^{r_i} = 0$

Clearly, $B^k = V J^k V^{-1}$ and $J^k = \begin{pmatrix} J_1^k & & \\ & \ddots & \\ & & J_s^k \end{pmatrix}$, since J is block-diagonal.

Recall that $\|\cdot\|_\infty$ induces the row-sum norm

Hence, $\|J^k\|_\infty = \max_{i=1,\dots,s} \|J_i^k\|_\infty$

Recall the binomial formula

$$J_i^k = (\lambda_i I_i + N_i)^k = \sum_{j=0}^{k} \binom{k}{j} \lambda_i^{k-j} N_i^j \overset{N_i^{r_i}=0!}{=} \sum_{j=0}^{r_i-1} \binom{k}{j} \lambda_i^{k-j} N_i^j$$

$$\overset{\text{U.l.o.g. } \lambda_i \neq 0}{=} \lambda_i^k \underbrace{\sum_{j=0}^{r_i-1} \binom{k}{j} \lambda_i^{-j} N_i^j}_{\text{finite sum with polynomial growth in } k!}$$

If $|\lambda_i| > 1$, then $\|J_i^k\|_\infty \xrightarrow{k} \infty$.

If $|\lambda_i| < 1$, then $\|J_i^k\|_\infty \xrightarrow{k} 0$,

since $k^s q^k$ is summable!

This yields

- $\|J^k\|_\infty \longrightarrow \infty$ if $\max_i |\lambda_i| = \rho(B) > 1$,
- $\|J^k\|_\infty \to 0$ if $\max_i |\lambda_i| = \rho(B) < 1$.

Therefore

- $\|B^k\|_\infty = \|V J^k V^{-1}\|_\infty \leq \|V\|_\infty \|J^k\|_\infty \|V^{-1}\|_\infty \to 0$ if $\rho(B) < 1$

Similarly,

- $\|J^k\|_\infty = \|V^{-1} B^k V\|_\infty \leq \|V^{-1}\|_\infty \|B^k\|_\infty \|V\|_\infty$

  $\to \infty$ if $\rho(B) > 1$

yields that $\|B^k\|_\infty \to \infty$ if $\rho(B) > 1$. //

---

④ From step 2 + 3, we infer

- $\|\Pi_+^k\|_\infty \to 0$, since $\rho(\Pi_+^k) < 1$,
- $\|\Pi_-^k\|_\infty \to \infty$, since $\rho(\Pi_-^k) > 1$.

Since all norms on the finite-dimensional space $\mathbb{K}^{n\times n}$ are equivalent, we infer for the norm $\|\cdot\|$ on $\mathbb{K}^n$ and the induced matrix norm.

- $\|\Pi_+^k\| \to 0$
- $\|\Pi_-^k\| \to \infty$

▨

We aim to solve $Ax^* = b$ for given $A \in \mathbb{K}^{n\times n}$ regular, $b \in \mathbb{K}^n$ with sought $x^* \in \mathbb{K}^n$.

To this end, we consider the affine iteration function

$$\Phi : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}^n, \quad \Phi(b,x) := Mx + Nb$$

with fixed matrices $M, N \in \mathbb{K}^{n\times n}$. The matrix M is called iteration matrix.

KNOWN: We require $\rho(M) < 1$ to guarantee convergence of the linear fixed point iteration $x_{k+1} := \Phi(b; x_k)$ for arbitrary initial guesses $x_0 \in \mathbb{K}^n$

DEF: The linear fixpol point method is

- consistent, if $\forall x \in \mathbb{K}^n : x = \Phi(Ax, x)$, i.e., each solution x of $Ax =: b$ is a fixed point of the corresponding iteration.
- convergent, if $\forall b \in \mathbb{K}^n \; \exists x^* \in \mathbb{K} \; \forall x_0 \in \mathbb{K}^n, \; x_{k+1} := \Phi(b, x_k) : x^* = \lim_k x_k$

KNOWN: Convergence yields that $x^* = \Phi(b, x^*)$, i.e., each limit is a fixed point (if the limit exists)

LEMMA: Let $\Pi, N \in \mathbb{K}^{n\times n}$, let $A \in \mathbb{K}^{n\times n}$ regular, let $\Phi(b,x) := \Pi x + Nb$. Then, there holds the following:

(i) If Φ is consistent, then $I-\Pi$ is regular if and only if N is regular.

(ii) If Φ is convergent, then $I-\Pi$ is regular.

(iii) If Φ is consistent and convergent, then $I-\Pi$ and N are regular.

Proof:

(i) For all $x \in \mathbb{K}^n$ holds

$$x = \Phi(Ax, x) = \Pi x + NAx$$

Hence, $I = \Pi + NA$ or equivalently $I - \Pi = NA$.

Since A is regular, it follows that: $I-\Pi$ regular $\Leftrightarrow$ N regular. //

(ii) If $I-\Pi$ is not regular, then there exists $x \in \mathbb{K}^n \setminus \{0\}$ with $(I-\Pi)x = 0$.

Hence, $\Pi x = x$, i.e., x is an eigenvector and $1 \in \sigma(\Pi)$.

In particular, it follows, $\rho(\Pi) \geq 1$.

However, convergence of a linear iteration yields $\rho(\Pi) < 1$.

Therefore, $I-\Pi$ must be regular. //

(iii) According to (ii), $I-\Pi$ is regular. (by convergence)

According to (i), this yields regularity of N (by consistency). ▨

REMARK: There are different ways to write/think the linear fixed point iteration $x_{k+1} := \Phi(b, x_k) = \Pi x_k + Nb$:

(a) With consistency, we get $I = \Pi + NA$ and hence $\Pi = I - NA$

$\Rightarrow x_{k+1} = x_k + N(b - Ax_k)$ ← "Richardson-type update formulation"

(b) $N^{-1}(x_{k+1} - x_k) = b - Ax_k$ ← solving for update as for Newton.

VO 26.03.2026

DEF: A linear fixed point iteration $\Phi(b, x) = \Pi x + Nb$ is called splitting method, if $A = G - H$ with $G \in \mathbb{R}^{n \times n}$ regular, $H \in \mathbb{R}^{n \times n}$ and $N = G^{-1}$, $\Pi = G^{-1}H = I - G^{-1}A$.

REMARK: For a splitting method $A = G - H$, it holds

$$Ax^* = b \iff Gx^* = Hx^* + b \iff x^* = \underbrace{G^{-1}H}_{=\Pi} x^* + \underbrace{G^{-1}}_{=N} b.$$

Hence, splitting methods are consistent.

The fixed point iteration reads

$$x_{k+1} = G^{-1}Hx_k + G^{-1}b = (I - G^{-1}A)x_k + G^{-1}b = \underbrace{x_k + G^{-1}(b - Ax_k)}_{\text{Richardson-type formulation.}}$$

where $G \approx A$, but is easier to solve.

Note that G^{-1} will never be computed. Instead,

- Solve $G\delta_k := b - Ax_k$
- Define $x_{k+1} := x_k + \delta_k$.

EXAMPLE: The Richardson iteration is a splitting method

$A = I - (I - A)$ with $G := I$, $H := I - A$

and hence $\Pi = G^{-1}H = I - A$, $N = G^{-1} = I$, $\Phi(b, x) = (I - A)x + b$.

EXAMPLE (JACOBI ITERATION) Let $A \in \mathbb{K}^{n\times n}$ be regular, let $D \in \mathbb{K}^{n\times n}$ be defined by $D_{jk} := A_{jk}\delta_{jk}$, i.e., the diagonal of A. Suppose that D is regular, i.e., $D_{jj} \neq 0$ for all $j = 1,\dots,n$.

Then, the Jacobi iteration is the splitting scheme $A = D + (A-D)$, i.e., $G := D$, $H := -(A-D)$; $\Pi := G^{-1}H = -D^{-1}(A-D)$, $N := G^{-1} = D^{-1}$:

- Solve $D\delta_k := b - Ax_k$
- Define $x_{k+1} := x_k + \delta_k$

Using upper indices $x^k := x_k$ for the iteration index and lower indices for the coefficients, this can explicitly be written as

$$x_i^{k+1} = x_i^k + \underbrace{\frac{1}{A_{ii}}\left(b_i - \sum_{j=1}^{n} A_{ij}x_j^k\right)}_{=\delta_i^k} \quad \text{for all } i = 1,\dots,n.$$

EXAMPLE ((forward) GAUSS-SEIDEL ITERATION) Let $A \in \mathbb{K}^{n\times n}$ be regular. Write $A = L + D + U$ with $L_{jk} = \begin{cases} A_{jk} & \text{for } j>k \\ 0 & \text{else} \end{cases}$, $U_{jk} = \begin{cases} A_{jk} & \text{for } j<k \\ 0 & \text{else} \end{cases}$, i.e., $A = \begin{smallmatrix} D & U \\ L & \end{smallmatrix}$. Suppose that D (and hence also $D+L$) is regular.

Then, the (forward) Gauss-Seidel iteration is the splitting scheme $A = (D+L) + U$, i.e., $G := D+L$, $H := -U$, $\Pi := G^{-1}H = -(D+L)^{-1}U$, $N := G^{-1} = (D+L)^{-1}$:

- Solve $(D+L)\delta_k = b - Ax_k$
- Define $x_{k+1} := x_k + \delta_k$

As above, this can be explicitly written as

$$x_i^{k+1} = x_i^k + \frac{1}{A_{ii}}\left(b_i - \sum_{j=1}^{i-1} A_{ij}x_j^{k+1} - \sum_{j=i}^{n} A_{ij}x_j^k\right) \quad \text{for all } i = 1,\dots,n.$$

Consequently, the only difference to the Jacobi method is that the newly computed coefficients x_j^{k+1} immediately replace the old x_j^k

---

EXAMPLE (BACKWARD GAUSS-SEIDEL ITERATION): $A = (D+U) + L$, $G := D+U$, $H := -L$

REMARK: Richardson and Jacobi iteration are independent of the numbering of the unknowns, whereas (forward/backward) Gauss-Seidel depends on the numbering, i.e., different numbering leads to different schemes and hence different iterates.

---

§ 1.4 DIAGONAL DOMINANT MATRICES

DEF: A matrix $A \in \mathbb{K}^{n\times n}$ is strictly diagonal dominant, if

$$\sum_{\substack{k=1\\k\neq j}}^{n} |A_{jk}| < |A_{jj}| \quad \forall j = 1,\dots,n.$$

REMARK: If $A \in \mathbb{K}^{n\times n}$ is strictly diagonal dominant, then the Jacobi and Gauss-Seidel iteration are well defined (since $A_{jj} \neq 0$).

LEMMA: Each strictly diagonal dominant matrix $A \in \mathbb{K}^{n\times n}$ is regular.

Proof:

Let $x \in \mathbb{K}^n$ with $Ax = 0$.

Then, $(A-D)x = -Dx$, i.e., $\sum_{\substack{k=1\\k\neq j}}^{n} A_{jk}x_k = A_{jj}x_j \quad \forall j = 1,\dots,n.$

Choose index j with $|x_j| = \|x\|_\infty$.

$$\Rightarrow \|x\|_\infty = |x_j| = \left|\frac{1}{A_{jj}} \sum_{\substack{k=1\\k\neq j}}^{n} A_{jk}x_k\right| \leq \underbrace{\sum_{k\neq j} \left|\frac{A_{jk}}{A_{jj}}\right|}_{<1} \underbrace{|x_k|}_{\leq \|x\|_\infty}$$

and hence $x = 0$.

This proves $\ker(A) = \{0\}$ and hence $A \in \mathbb{K}^{n\times n}$ is regular. ∎

LEMMA (NEUMANN SERIES): Let $\Pi \in \mathbb{K}^{n\times n}$ with $\rho(\Pi) < 1$.
Then, $I-\Pi$ is regular and $(I-\Pi)^{-1} = \sum_{k=0}^{\infty} \Pi^k$

Proof

Since $\rho(\Pi) < 1$, we choose a norm on $\mathbb{C}^n$ such that $\|\Pi\| < 1$.

Note that $\left\| \sum_{k=i}^{j} \Pi^k \right\| \leq \sum_{k=i}^{j} \|\Pi\|^k \leq \|\Pi\|^i \frac{1}{1-\|\Pi\|} \longrightarrow 0$ as $i,j \to \infty$.

Hence, $\sum_{k=0}^{j} \Pi^k$ is a Cauchy sequence and its limit $\sum_{k=0}^{\infty} \Pi^k$ exists.

Note that $(I-\Pi) \sum_{k=0}^{\infty} \Pi^k = \sum_{k=0}^{\infty} \Pi^k - \sum_{k=1}^{\infty} \Pi^k = I$.

Hence, $I-\Pi$ is regular and $(I-\Pi)^{-1} = \sum_{k=0}^{\infty} \Pi^k$. ∎

THEOREM (JACOBI AND GAUSS-SEIDEL FOR SDD-MATRICES):

Let $A \in \mathbb{K}^{n\times n}$ be strictly diagonal dominant and hence regular.
Let $\Pi^{(J)} := -D^{-1}(I-D)$, $\Pi^{(GS)} := -(D+L)^{-1}U$ be the iteration matrices of Jacobi and Gauss-Seidel iteration, respectively.
Then, $\|\Pi^{(GS)}\|_\infty \leq \|\Pi^{(J)}\|_\infty < 1$ and hence $\rho(\Pi^{(GS)}) < 1$ as well as $\rho(\Pi^{(J)}) < 1$, i.e., Jacobi and Gauss-Seidel iteration guarantee convergence.

REMARK: Note that the theorem proves convergence of the iterations without the need to know $\sigma(A)$, but only by knowledge of simple algebraic properties. Moreover, we can use, e.g., $\|x_{k+1} - x_k\|_\infty$ to control the iteration error / solver error.

Proof:

① We show that $\|\Pi^{(J)}\|_\infty < 1$.

Recall that $\Pi^{(J)} = -D^{-1}(A-D)$, i.e., $\Pi^{(J)}_{jk} = \begin{cases} 0 & \text{if } j=k \\ \frac{A_{jk}}{A_{jj}} & \text{if } j \neq k \end{cases}$

$$\|\Pi^{(J)}\|_\infty = \max_{j=1,\dots,n} \sum_{k=1}^{n} |\Pi^{(J)}_{jk}|$$

$$= \max_{j} \sum_{\substack{k=1 \\ k\neq j}}^{n} \left|\frac{A_{jk}}{A_{jj}}\right| < 1.$$

(by strict diagonal dominancy.)

For this proof, let w coincide $\leq$ and $|\cdot|$ componentwise for all matrix entries.

---

② $D^{-1}(L+D) = I + D^{-1}L$ and $\rho(D^{-1}L) = 0$

The Neumann series gives $[D^{-1}(L+D)]^{-1} = I - (-D^{-1}L) = \sum_{k=0}^{\infty} (-D^{-1}L)^k$. (Neumann)

Taking $|\cdot|$, this yields

$$|[D^{-1}(L+D)]^{-1}| \leq \sum_{k=0}^{\infty} |D^{-1}L|^k = (I - |D^{-1}L|)^{-1}.$$ (Neumann)

---

③ $\Pi^{(J)} = -D^{-1}(A-D) = -D^{-1}(L+U)$

$\Rightarrow |\Pi^{(J)}| \leq |D^{-1}L| + |D^{-1}U|$, since matrix pattern L, U are disjoint and D^{-1} is diagonal (i.e., scaling of rows)

$\Rightarrow |D^{-1}U| = (|\Pi^{(J)}| - I) + (I - |D^{-1}L|)$

---

④ After these preparations, we aim to conclude the proof.

Recall that $\Pi^{(GS)} = -(D+L)^{-1}U$

$$|\Pi^{(GS)}| = |(D+L)^{-1}U| = |[D^{-1}(D+L)]^{-1}D^{-1}U| \leq |[D^{-1}(D+L)]^{-1}|\,|D^{-1}U|$$

$$\overset{②}{\leq} |(I - |D^{-1}L|)^{-1}|\,|D^{-1}U| \overset{③}{=} (I - |D^{-1}L|)^{-1}(|\Pi^{(J)}| - I) + (I - |D^{-1}L|)$$

$$= (I - |D^{-1}L|)^{-1}(|\Pi^{(J)}| - I) + I$$

Let $e = (1, \dots, 1)^T \in \mathbb{K}^n$.

$$\Rightarrow |\Pi^{(GS)}| e \leq \underbrace{(I - |D^{-1}L|)^{-1}}_{= \sum_{k=0}^{\infty} |D^{-1}L|^k \geq I} \underbrace{(|\Pi^{(J)}| e - e)}_{\leq \|\Pi^{(J)}\|_\infty e - e = (\|\Pi^{(J)}\|_\infty - 1) e, \; < 0} + e$$

$$\leq (\|\Pi^{(J)}\|_\infty - 1) e + e$$

$$= \|\Pi^{(J)}\|_\infty e$$

$$\Rightarrow \|\Pi^{(GS)}\|_\infty = \| |\Pi^{(GS)}| e \|_\infty \leq \| \|\Pi^{(J)}\|_\infty e \|_\infty = \|\Pi^{(J)}\|_\infty \underbrace{\|e\|_\infty}_{=1} \cdot \blacksquare$$

COROLLARY: Let $A \in \mathbb{K}^{n \times n}$ such that A^T is strictly diagonal dominant.

Then, A is regular, (and) the Jacobi iteration is well-defined, (for A) and convergent

---

Proof:

According to $0 \neq \det(A^T) = \det(A)$, A is regular.

Since A^T is strictly diagonal dominant, $A_{jj} \neq 0$ for all $j = 1, \dots, n$.

Therefore, $M^{(J)}$ is well-defined.

Recall that
- $\|B\|_\infty = \max_j \sum_k |B_{jk}|$ row-sum norm,
- $\|B\|_{\square} = \max_k \sum_j |B_{jk}|$ column-sum norm.

Hence $\rho(\Pi^{(J)}) \leq \|\Pi^{(J)}\|_1 = \|(\Pi^{(J)})^T\|_\infty = \|(A^T - D) D^{-1}\|_\infty < 1$ (scaling of columns; as in theorem)

$\blacksquare$

DEF: • A matrix $A \in \mathbb{K}^{n\times n}$ is reducible, if there exist $J, K \subseteq \{1,\dots,n\}$ with $J \neq \emptyset$, $K \neq \emptyset$, $J \cap K = \emptyset$, $J \cup K = \{1,\dots,n\}$ and $A_{jk} = 0 \quad \forall j \in J \; \forall k \in K$.

• A is irreducible, if it is not reducible.

EXAMPLE: The matrix $A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 4 & 0 \\ 5 & 6 & 7 \end{pmatrix}$ is reducible with $J = \{1,2\}$, $K = \{3\}$.

REMARK: If $A \in \mathbb{K}^{n\times n}$ is regular and reducible, then one can reduce the computational cost of solving $Ax^* = b$ by first solving $x_j^* \; \forall j \in J$ and then $x_k^* \; \forall k \in K$:

• For $j \in J$, note that $b_j = \sum_{\ell=1}^{n} A_{j\ell} x_\ell = \sum_{\ell \in J} A_{j\ell} x_\ell$

• For $k \in K$, there holds $b_k = \sum_{\ell \in J} A_{k\ell} x_\ell + \sum_{\ell \in K} A_{k\ell} x_\ell$

DEF: A matrix $A \in \mathbb{K}^{n\times n}$ is diagonal dominant, if

(i) $\sum_{\substack{k=1 \\ k\neq j}}^{n} |A_{jk}| \leq |A_{jj}| \quad \forall j \in 1,\dots,n$

(ii) There exist $j_0 \in \{1,\dots,n\}$ such that $\sum_{\substack{k=1 \\ k\neq j_0}}^{n} |A_{j_0 k}| < |A_{j_0 j_0}|$

REMARK: Clearly, any strictly diagonal dominant matrix $A \in \mathbb{K}^{n\times n}$ is also diagonal dominant, since (ii) holds for all $j_0 = j \in \{1,\dots,n\}$.

REMARK: Diagonal dominance is not enough to ensure that A is regular, e.g., $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

LEMMA: Let $A \in \mathbb{K}^{n\times n}$ be diagonal dominant and irreducible. Then, A is regular and $A_{jj} \neq 0 \quad \forall j = 1,\dots,n$.

Proof: ① We argue by contradiction and assume that A is not regular. Then, there exists $x \in \mathbb{K}^n \setminus \{0\}$ with $Ax = 0$.

$$\Rightarrow |A_{jj}||x_j| \leq \sum_{\substack{l=1 \\ l\neq j}}^{n} |A_{jl}||x_l| \quad \forall j = 1,\dots,n. \qquad (*)$$

Define $J := \{ j \mid |x_j| = \|x\|_\infty \}$, $K := \{ k \mid |x_k| < \|x\|_\infty \}$.

Clearly: $J \cap K = \emptyset$, $J \neq \emptyset$, $J \cup K = \{1,\dots,n\}$.

If $K = \emptyset$, then $(*)$ yields $|A_{jj}| \leq \sum_{\substack{l=1 \\ l\neq j}}^{n} |A_{jl}| \quad \forall j = 1,\dots,n$,

which contradicts diagonal dominance.

Hence, $K \neq \emptyset$.

Since A is irreducible (i.e., not reducible), there exist $j \in J$, $k \in K$ with $A_{jk} \neq 0$.

$$(*) \rightarrow |A_{jj}| \leq \sum_{\substack{l=1 \\ l\neq j}}^{n} |A_{jl}| \frac{|x_l|}{|x_j|} \leq \sum_{\substack{l=1 \\ l\neq j}}^{n} |A_{jl}|, \quad \text{since } \frac{|x_l|}{|x_j|} = \frac{|x_l|}{\|x\|_\infty} \leq 1$$

$$\text{and } A_{jk} \neq 0,\ \frac{|x_k|}{|x_j|} = \frac{|x_k|}{\|x\|_\infty} < 1.$$

This contradicts that A is diagonal dominant. ↯

Hence, A is regular. //

② If $A_{jj} = 0$ for some $j \in \{1,\dots,n\}$.

By diagonal dominance, $\sum_{k\neq j} |A_{jk}| \leq |A_{jj}| = 0$, i.e., the j-th row is zero.

This would contradicts that A is regular.

Thus $A_{jj} \neq 0$ for all $j \in \{1,\dots,n\}$. ∎

THEOREM: Let $A \in \mathbb{K}^{n\times n}$ be diagonal dominant and irreducible. Then, Jacobi iteration and Gauss-Seidel iteration are well-defined and convergent.

Proof for Jacobi:

Recall $\Pi^{(J)} = -D^{-1}(A-D)$, which is well-defined since $D_{ii} = A_{ii} \neq 0$.

We must show $\rho(\Pi^{(J)}) < 1$.

We consider $\Pi := \Pi^{(J)} - \lambda I$ for $\lambda \in \mathbb{C}$ with $|\lambda| \geq 1$.

We show that Π is regular and hence $\lambda \notin \sigma(\Pi^{(J)})$ so that $\rho(\Pi^{(J)}) < 1$.

Clearly, A irreducible $\Rightarrow$ $A-D$ irreducible $\Rightarrow$ $\Pi^{(J)}$ irreducible. $\Rightarrow$ Π irreducible.

$$\sum_{\substack{k=1\\ j\neq k}}^{n} |\Pi_{jk}| = \sum_{\substack{k=1\\ j\neq k}}^{n} |\Pi^{(J)}_{jk}| = \sum_{\substack{k=1\\ j\neq k}}^{n} \frac{|A_{jk}|}{|A_{jj}|} \overset{(*)}{\leq} 1 \leq |\lambda| = |\Pi_{jj}|$$

Def. of Π — Def. $\Pi^{(J)}$ — diagonal dominance — by choice of λ — Def. of Π

Note that $(*)$ holds strictly for at least one $j_0 \in \{1,\dots,n\}$.

Hence, Π is diagonal dominant

$\Rightarrow$ Π is regular (by being diagonal dominant + irreducible) ∎

Proof for Gauss-Seidel:

Recall $\Pi^{(GS)} = -(L+D)^{-1}U$, which is well defined, since $D_{jj} = A_{jj} \neq 0$, $L_{jj} = 0$.

We consider $\Pi := \Pi^{(GS)} - \lambda I$ for $\lambda \in \mathbb{C}$ with $|\lambda| \geq 1$.

As before, we prove that Π is regular (to see that $\rho(\Pi^{(GS)}) < 1$).

Note:

$$\Pi \text{ regular} \iff -(L+D)\Pi = -(L+D)\left[-(L+D)^{-1}U - \lambda I\right] = U + \lambda L + \lambda D$$

regular.

Note $A = U+L+D$ irreducible $\Leftrightarrow$ $\tilde{\Pi} := U + \lambda L + \lambda D$ irreducible

Note that

$$\sum_{\substack{k=1\\k\neq j}}^{n} |\tilde{\Pi}_{jk}| = \sum_{k=j+1}^{n} \underbrace{|A_{jk}|}_{=|U_{jk}|} + |\lambda| \sum_{k=1}^{j-1} \underbrace{|A_{jk}|}_{=|L_{jk}|} \leq |\lambda| \sum_{\substack{k=1\\k\neq j}}^{n} |A_{jk}| \leq |\lambda| \underbrace{|A_{jj}|}_{=|\tilde{\Pi}_{jj}|}$$

($|\lambda| \geq 1$; A is diagonal dominant.)

Hence, $\tilde{\Pi}$ is diagonal dominant.

$\Rightarrow$ $\tilde{\Pi}$ is regular (by being diagonal dominant and irreducible)

$\Rightarrow$ Π is regular. ∎

In the framework of the finite element method, a particularly important class of matrices are SPD matrices.

DEF: A matrix $A \in \mathbb{K}^{n\times n}$ is SPD, if

- A is selfadjoint, i.e., $A = A^H$ (or: symmetric for $\mathbb{K} = \mathbb{R}$)
- A is positive definite, i.e., $\forall x \in \mathbb{K}^n \setminus \{0\}: \langle Ax, x\rangle_2 > 0$.

LEMMA: If $A \in \mathbb{K}^{n\times n}$ is SPD, then A is regular and

$$A_{jj} = \langle Ae_j, e_j\rangle_2 > 0 \quad \forall j = 1,\dots,n$$

so that Jacobi and Gauss-Seidel iteration are well-defined.

Proof: Since A is positive definite, then holds $Ax \neq 0$ $\forall x \in \mathbb{K}^n \setminus \{0\}$. Hence, A has trivial kernel and is hence regular. ∎

LEMMA: Let $A \in \mathbb{K}^{n\times n}$ be SPD. Then, $\langle x, y\rangle_A := \langle Ax, y\rangle_2$ is a scalar product on $\mathbb{K}^n$ with induced norm $\|x\|_A := \langle Ax, x\rangle_2^{1/2}$.

Proof: We need to show that $\langle \cdot, \cdot\rangle_A$ is linear in the second argument, antisymmetric (i.e., $\langle x, y\rangle_A = \overline{\langle y, x\rangle_A}$) and definit (i.e., $\langle x, x\rangle_A > 0$).

- linearity is clear, definitness follows since A is positive definite
- To see antisymmetry, note

$$\langle x, y\rangle_A = \langle Ax, y\rangle_2 = (Ax)^H y = x^H A^H y = \overline{x^H A y} = \overline{\sum_{j,k} \bar{x}_j A_{jk} y_k}$$

$$\langle y, x\rangle_A = \sum_{j,k} \bar{y}_j \underbrace{A_{jk}}_{=\overline{A_{kj}}} x_k = \overline{\sum_{j,k} \bar{x}_k A_{kj} y_j} = \overline{\langle x, y\rangle_A}$$

The Cauchy-Schwarz inequality shows that $\|\cdot\|_A$ is a norm on $\mathbb{K}^n$. ∎

Recall that an SPD matrix is diagonalizable, i.e., there exists a diagonal matrix $\Lambda = \mathrm{diag}(d_1, \dots, d_n)$ and an orthogonal matrix $Q \in \mathbb{K}^{n \times n}$ such that $A = Q \Lambda Q^H$ and $Q^H Q = \mathrm{Id}$.
In particular, $\sigma(A) = \sigma(\Lambda)$, the j-th column of v_j of Q is an eigenvector of A, and $d_j = v_j^H A v_j = \langle v_j, A v_j \rangle_2 > 0$.
(with $A v_j = d_j v_j$)

DEF: This allows to define the root $A^{1/2} := Q \Lambda^{1/2} Q^H$ with $\Lambda^{1/2} := \mathrm{diag}(d_1^{1/2}, \dots, d_n^{1/2})$ and $A^{-1/2} := Q \Lambda^{-1/2} Q^H$ with $\Lambda^{-1/2} := \mathrm{diag}(d_1^{-1/2}, \dots, d_n^{-1/2})$.

Clearly, $A^{1/2} A^{1/2} = A$, $A^{1/2} A^{-1/2} = \mathrm{Id}$, $A^{-1/2} A^{-1/2} = A^{-1}$.

NOTATION: Let $A, B \in \mathbb{K}^{n \times n}$ be selfadjoint. In the following, we write $A > 0$ if A is SPD and $A > B$ if $A - B > 0$.

LEMMA: Let $A, B \in \mathbb{K}^{n \times n}$ be selfadjoint. Then, there holds:

(i) $A > 0 \iff \forall C \in \mathbb{K}^{n \times n}$ regular : $C^H A C > 0$

(ii) $A > B \iff \forall C \in \mathbb{K}^{n \times n}$ regular : $C^H A C > C^H B C$

(iii) $A, B > 0 \implies A + B > 0$

(iv) $d_{\min} \mathrm{Id} < A < d_{\max} \mathrm{Id} \iff \sigma(A) \subseteq (d_{\min}, d_{\max})$

(v) $A > B > 0 \iff 0 < A^{-1} < B^{-1}$

Proof: (i) It suffices to prove "$\Rightarrow$" (since $(C^H)^{-1} = (C^{-1})^H =: C^{-H}$)
Let $C \in \mathbb{K}^{n \times n}$ be regular.
Clearly, $C^H A C$ is selfadjoint (as $(C^H A C)^H = C^H \underbrace{A^H}_{=A} \underbrace{C^{HH}}_{=C}$).
For $x \in \mathbb{K}^n \setminus \{0\}$, it holds $Cx \neq 0$ and hence

$\langle C^H A C x, x \rangle_2 = \langle A C x, C x \rangle_2 > 0$, i.e., $C^H A C$ is positive definite. //

Proof: (ii) follows from (i) since $C^H A C - C^H B C = C^H (A-B) C$. //

(iii) is obvious

(iv) Since A is selfadjoint, there exists orthonormal basis

$\{v_1, \dots, v_n\} \subseteq \mathbb{K}^n$ such that $Av_j = \lambda_j v_j$, $\langle v_j, v_k \rangle_2 = \delta_{jk}$, $\lambda_1 \le \dots \le \lambda_n$.

Recall that $\|x\|_2^2 = \sum_{j=1}^n |\mu_j|^2$, $\langle Ax, x \rangle_2 = \sum_{j=1}^n \lambda_j |\mu_j|^2$

for all $x = \sum_{j=1}^n \mu_j v_j \in \mathbb{K}^n$.

Hence, $\langle (A - \lambda_{min} \mathrm{Id}) x, x \rangle_2 = \langle Ax, x \rangle_2 - \lambda_{min} \|x\|_2^2$

$= \sum_{j=1}^n (\lambda_j - \lambda_{min}) |\mu_j|^2 \quad \forall x = \sum_j \mu_j v_j$

This proves that: $\left(\lambda_{min} \mathrm{Id} < A \iff \lambda_{min} < \lambda_j \ \forall j = 1, \dots, n \right)$

Similarly, it follows that: $\left(A < \lambda_{max} \mathrm{Id} \iff \forall j = 1, \dots, n : \lambda_j < \lambda_{max} \right)$. //

(v) It suffices to prove "$\Rightarrow$" (since $(C^{-1})^{-1} = C$).

Note that: $Kv = \lambda v \iff K^{-1} v = \frac{1}{\lambda} v$ if K is SPD.

Hence, $A > 0$ yields $A^{-1} > 0$.

Moreover, $A > B$ yields $B^{-1/2} A B^{-1/2} > B^{-1/2} B B^{-1/2} = \mathrm{Id}$.

$\Rightarrow \sigma(\underbrace{B^{-1/2} A B^{-1/2}}_{=: X \text{ is SPD}}) \subseteq (1, \infty)$

As noted above X^{-1} is SPD and $\sigma(X^{-1}) \subseteq (0, 1)$.

$\Rightarrow B^{1/2} A^{-1} B^{1/2} = X^{-1} < \mathrm{Id}$

$\Rightarrow \underbrace{B^{-1/2} X^{-1} B^{-1/2}}_{= A^{-1}} < B^{-1/2} B^{-1/2} = B^{-1}$.

∎

THEOREM (Convergence of damped Jacobi): Let $A \in \mathbb{K}^{n\times n}$ be SPD.
Let $D \in \mathbb{K}^{n\times n}$, $D_{jk} := A_{jj}\delta_{jk} > 0$ be the diagonal of A.
For $\omega > 0$, let $\Pi_\omega^J := \mathrm{Id} - \omega D^{-1}A$ be the iteration matrix of damped Jacobi. Then, $\rho(\Pi_\omega^J) = \| A^{1/2}\Pi_\omega^J A^{-1/2}\|_2 = \|\Pi_\omega^J\|_A$.
Moreover, then holds: $\|\Pi_\omega^J\|_A < 1 \iff A < \frac{2}{\omega} D$.
In this case, $\|\Pi_\omega^J\|_A < 1$ is thus the contraction factor of the damped Jacobi iteration w.r.t. energy norm.

Proof:

① For any $\Pi \in \mathbb{K}^{n\times n}$, $x \in \mathbb{K}^n$, $y := A^{1/2}x$, it holds

$$\|x\|_A^2 = \langle Ax, x\rangle_2 = \langle A^{1/2}x, A^{1/2}x\rangle_2 = \|A^{1/2}x\|_2^2 = \|y\|_2^2,$$

$$\|\Pi x\|_A = \|A^{1/2}\Pi x\|_2 = \|A^{1/2}\Pi A^{-1/2}y\|_2.$$

$$\Rightarrow \|\Pi\|_A = \sup_{x\neq 0} \frac{\|\Pi x\|_A}{\|x\|_A} = \sup_{y\neq 0} \frac{\|A^{1/2}\Pi A^{-1/2}y\|_2}{\|y\|_2} = \|A^{1/2}\Pi A^{-1/2}\|_2$$

— — —

② $A^{1/2}\Pi A^{-1/2} \overset{(\Pi = \Pi_\omega^J)}{=} \mathrm{Id} - \omega A^{1/2}D^{-1}A^{1/2}$ is self-adjoint.

$$\Rightarrow \underbrace{\rho(A^{1/2}\Pi A^{-1/2})}_{=\rho(\Pi) \text{ as } \sigma(A^{1/2}\Pi A^{-1/2}) = \sigma(\Pi)} \overset{!}{=} \|A^{1/2}\Pi A^{-1/2}\|_2 = \|\Pi\|_A \quad \text{by index examples.}$$

— — —

③ $\rho(\Pi) < 1 \iff \sigma(A^{1/2}\Pi A^{-1/2}) \subsetneq (-1,1) \iff -\mathrm{Id} < \mathrm{Id} - \omega A^{1/2}D^{-1}A^{1/2} < \mathrm{Id}$

$\iff -2\,\mathrm{Id} < -\omega A^{1/2}D^{-1}A^{1/2} < 0$

$\iff 0 < \omega A^{1/2}D^{-1}A^{1/2} < 2\,\mathrm{Id}$

$\iff \omega D^{-1} < 2A^{-1}$

$\iff 0 < A < \frac{2}{\omega}D$ ∎

The convergence of the damped Jacobi iteration depends on $\sigma(A)$ and holds, if $0<\omega$ is sufficiently small. In contrast, the following (only) SOR iteration ("successive over relaxation") is convergent for all $0<\omega<2$.

— — —

RECALL: The Gauss-Seidel method reads

$$x_i^{k+1} := x_i^k + \frac{1}{A_{ii}}\left(b_i - \sum_{j=1}^{i-1} A_{ij}\, x_j^{k+1} - \sum_{j=i}^{n} A_{ij}\, x_j^k\right) \quad \forall i=1,\dots,n$$

With the disjoint splitting $A = L+D+U$, this can be written

$$\begin{aligned} Dx_{k+1} &= Dx_k + \left[b - Lx_{k+1} - (D+U)x_k\right] \\ &= (D+L)x_k + \left[b - Lx_{k+1} - Ax_k\right] \end{aligned}$$

$$\Leftrightarrow (D+L)x_{k+1} = (D+L)x_k + (b - Ax_k)$$

$$\Leftrightarrow x_{k+1} = x_k + (D+L)^{-1}(b - Ax_k).$$

$\Pi^{GS} = I - (D+L)^{-1}A$

The SOR iteration reads, with fixed $\omega>0$,

$$x_i^{k+1} := x_i^k + \frac{\omega}{A_{ii}}\left(b_i - \sum_{j=1}^{i-1} A_{ij}\, x_j^{k+1} - \sum_{j=i}^{n} A_{ij}\, x_j^k\right) \quad \forall i=1,\dots,n$$

Arguing as before, this can be written

$$\begin{aligned} Dx_{k+1} &= Dx_k + \omega\left[b - Lx_{k+1} - (D+U)x_k\right] \\ &= (D+\omega L)x_k + \omega\left[b - Lx_{k+1} - Ax_k\right] \end{aligned}$$

$$\Leftrightarrow (D+\omega L)x_{k+1} = (D+\omega L)x_k + \omega(b - Ax_k)$$

$$\Leftrightarrow x_{k+1} = x_k + (D+\omega L)^{-1}(b - Ax_k).$$

$\Pi_\omega^{SOR} := I + \omega(D+\omega L)^{-1}A$ ∎

THEOREM (Convergence of SOR): Let $A \in \mathbb{K}^{n\times n}$ be SPD.
Let $0 < \omega < 2$. Let $\Pi_\omega^{SOR} := I - \omega(D+\omega L)^{-1}A$, where
$D, L \in \mathbb{K}^{n\times n}$ read $D_{jj} = A_{jj}\delta_{jk}$, $L_{jk} = \begin{cases} A_{jk} & \text{for } j>k \\ 0 & \text{else} \end{cases}$. (always)
Then, then holds $\rho(\Pi_\omega^{SOR}) \le \|\Pi_\omega^{SOR}\|_A < 1$, i.e., SOR leads to convergence. (Note that $\Pi_1^{SOR} = \Pi^{GS}$, i.e., Gauss-Seidel is covered)

Proof: ① Denote $\Pi := I - U^{-1}A$, $U = \frac{1}{\omega}(D+\omega L) = \frac{1}{\omega}D + L$

Note that $\|\Pi x\|_A^2 = \langle A\Pi x, \Pi x\rangle_2 = \langle \Pi^H A \Pi x, x\rangle_2$

Note $\Pi^H A \Pi = (I - AU^{-H})A(I - U^{-1}A)$

$= A - AU^{-H}A - AU^{-1}A + AU^{-H}AU^{-1}A$

$= A - A(\underbrace{U^{-H} + U^{-1} + U^{-H}AU^{-1}}_{= U^{-H}(U+U^H)U^{-1} =: S})A$

$(-A)$

Suppose that $S > 0$.
Then, it follows

$$\|\Pi x\|_A^2 = \underbrace{\langle Ax, x\rangle_2}_{=\|x\|_A^2} - \underbrace{\langle ASAx, x\rangle_2}_{=\langle SAx, Ax\rangle_2} \overset{*}{<} \|x\|_A^2$$

Hence, $\|\Pi\|_A = \sup_{x\neq 0} \frac{\|\Pi x\|_A}{\|x\|_A} < 1$. //

---

② Note that

$$U + U^H - A = \left(2\tfrac{1}{\omega}D + L + L^H\right) - \underbrace{(L + D + L^H)}_{=A} = \left(\frac{2}{\omega} - 1\right)D$$

Since D is positive definite, then holds equivalence

$$0 < \omega < 2 \iff \frac{2}{\omega} - 1 > 0 \iff U + U^H - A > 0$$

$$\iff S = U^{-H}(U + U^H - A)U^{-1} > 0$$

(SPD)

A potential drawback of SOR is that, despite A being SPD, the iteration matrix

$$\Pi_\omega^{SOR} := I - \omega (D + \omega L)^{-1} A$$

is not ~~symmetric~~ selfadjoint. This can be cured by symmetrization:

- First, do one step of forward SOR

$$\tilde{x}_{k+1} = x_k + \omega (D+\omega L)^{-1} (b - A x_k), \quad \Pi_\omega^{SOR} = Id - \omega (D+\omega L)^{-1} A$$

- Second, do one step of backward SOR

$$x_{k+1} = \tilde{x}_{k+1} + \omega (D+\omega U)^{-1} (b - A \tilde{x}_{k+1}), \quad \overline{\Pi}_\omega^{SOR} := Id - \omega (D+\omega U)^{-1} A$$

This leads to symmetric SOR (SSOR) with iteration matrix $\Pi_\omega^{SSOR} = \overline{\Pi}_\omega^{SOR} \Pi_\omega^{SOR}$.

THEOREM: Let $A \in \mathbb{K}^{n \times n}$ be SPD and $0 < \omega < 2$.

Then, $\rho(\Pi_\omega^{SSOR}) = \| \Pi_\omega^{SSOR} \|_A = \| \Pi_\omega^{SOR} \|_A^2 < 1$.

In particular, SSOR is always converging.

Proof:

(1) For a matrix $\Pi \in \mathbb{K}^{n \times n}$, define $\Pi^A := A^{-1} \Pi^H A$

Then, $\langle \Pi^A x, y \rangle_A = \langle A (A^{-1} \Pi^H A) x, y \rangle_2 = \langle \Pi^H A x, y \rangle_2 = \langle A x, \Pi y \rangle_2 = \langle x, \Pi y \rangle_A$

i.e., Π^A is the adjoint of Π w.r.t. $\langle \cdot, \cdot \rangle_A$. //

(2) Consider $\Pi := \Pi_\omega^{SOR} = Id - \omega (D+\omega L)^{-1} A$

Then, $\Pi^H = I - \omega \underbrace{A^H}_{=A} \underbrace{(D+\omega L)^{-H}}_{= (D+\omega L^H)^{-1}}$

$$\Rightarrow \Pi^A = A^{-1} \Pi^H A = A^{-1} \left(Id - \omega A \underbrace{(D+\omega L^H)^{-1}}_{=U} \right) A = \underset{\overset{!}{=} \overline{\Pi}_\omega^{SOR}}{Id - \omega (D+\omega U)^{-1} A}$$

Hence, $\Pi_\omega^{SSOR} = \Pi^A \Pi$ is selfadjoint w.r.t. $\langle \cdot, \cdot \rangle_A$

since $(\Pi^A \Pi)^A = A^{-1} (\Pi^A \Pi)^H A = A^{-1} \Pi^H (A^{-1} \Pi^H A)^H A = A^{-1} \Pi^H \underbrace{A^H}_{=A} \underbrace{A^{-H} A}_{= Id}$

$\overset{!}{=} \Pi^A \Pi.$

Moreover, $\Pi^A \Pi$ is positive semidefinite

since $\langle \Pi^A \Pi x, x \rangle_A = \langle \Pi x, \Pi x \rangle_A = \|\Pi x\|_A \geq 0.$

From the introductory examples, we learn

$$\sigma(\Pi^A \Pi) \subseteq \mathbb{R}_{\geq 0} \quad \underline{\text{and}} \quad \rho(\Pi^A \Pi) = \|\Pi^A \Pi\|_A .$$

Finally, $\|\Pi\|_A^2 = \sup_{x \neq 0} \frac{\|\Pi x\|_A^2}{\|x\|_A^2} = \sup_{x \neq 0} \frac{\langle \Pi^A \Pi x, x \rangle_A}{\|x\|_A^2} \overset{!}{=} \rho(\Pi^A \Pi).$
$= \rho(\Pi_\omega^{SSOR}).$

From the previous theorem on SOR, we know that

$$\|\Pi\|_A < 1 \iff 0 < \omega < 2$$

for $\Pi = \Pi_\omega^{SOR}$. ∎

REMARK: A laborious but elementary calculation shows that

$$\Pi_\omega^{SSOR} = \overline{\Pi}_\omega^{SOR} \Pi_\omega^{SOR} = Id - \omega(2-\omega)(D+\omega U)^{-1} D (D+\omega L)^{-1} A$$

which is selfadjoint wrt. $\langle \cdot, \cdot \rangle_A$, provided that A is SPD
(and hence $U = L^H$).