Lecture 1 Stationary Methods

April 14 2026

Summary

We consider splitting based iterative methods for linear systems. This allows us to repeat basic concepts from linear algebra, to give a synthesis of 60 years of "classical" convergence theory and to specialize for matrices with specific properties arising in port Hamiltonian systems.

We seek $x \in \mathbb{C}^n$ in

$$Ax = b, \;\; A \in \mathbb{C}^{n\times n} \text{ non-singular}, b \in \mathbb{C}^n.$$

If n is large and A is sparse (only $\mathcal{O}(n)$ entries a_{ij} are non-zero), the direct factorization (aka Gaussian elimination) can become unfeasible since it becomes too expensive.

Remark 1.1. The above statement is to be taken with care. There is a whole branch of applied linear algebra which considers direct methods for sparse linear systems. The catch is to permute the rows and columns of A in such a manner that Gaussian elimination becomes efficient since only few of the zeros in A become non-zeros ("fill-in") during the elimination process. Mathematical methods that enter here are discrete methods that work with the graph representing the non-zero structure of A. There is a potential conflict between permuting for numerical stability ("pivoting") and permuting for small fill-in. Sparse direct solvers often work well for discretizations of 2-dimensional problems, but less so for higher dimensions. And they usually allow only for a small degree of parallelism. High quality software packages are available. MATLAB has sparse LU solvers. Libraries are MUMPS or HSL, for example.

In this course we consider iterative solvers as an alternative to Gaussian elimination.

Keywords: Matrix splittings, stationary iteration, spectral radius and convergence, Jacobi, Gauss-Seidel, Richardson, SOR, SSOR and HSS splitting, convergence analysis

1.1 Splittings

Definition 1.2 (splitting)**.** A pair $(M, N) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n}$ is termed a *splitting* of A if M is non-singular and

$$A = M - N.$$

The associated iterative method is

$$\begin{aligned} x^{(k+1)} &= M^{-1}(Nx^{(k)} + b), \;\; k = 0, 1, \ldots, \\ & \quad x^{(0)} : \text{initial guess or "starting vector".} \end{aligned} \tag{1.1}$$

In practice, $x^{(k+1)}$ is obtained by solving a linear system with matrix M, so M must be taken such that this takes little work. If the iterates converge then the limit solves the system:

$$\begin{aligned} \lim_{k\to\infty} x^{(k)} = x^* \;\; &\Rightarrow \;\; x^* = M^{-1}(Nx^* + b) \\ &\Rightarrow \;\; Mx^* = Nx^* + b \\ &\Rightarrow \;\; (\underbrace{M - N}_{=A})x^* = b. \end{aligned}$$

Definition 1.3 (standard notation)**.** For a linear system $Ax = b$ we denote $x^* = A^{-1}b$, and for a sequence of iterates $x^{(k)}$ we denote the errors as $e^{(k)}$ and the residuals as $r^{(k)}$,

$$e^{(k)} = x^* - x^{(k)}, \;\; r^{(k)} = b - Ax^{(k)} \, (= Ae^{(k)}).$$

Moreover, for a splitting based iteration (1.1) we define the iteration matrix (aka "error propagation matrix") as

$$H = M^{-1}N = I - M^{-1}A.$$

Useful relations to remember:

$$\begin{aligned} x^{(k+1)} &= M^{-1}(Nx^{(k)} + b) \;=\; Hx^{(k)} + M^{-1}b \;=\; x^{(k)} + M^{-1}r^{(k)} \\ x^* &= M^{-1}(Nx^* + b) \;=\; Hx^* + M^{-1}b \\ e^{(k+1)} &= He^{(k)} \\ r^{(k+1)} &= (AHA^{-1})r^{(k)} \;=\; (I - AM^{-1})r^{(k)}. \end{aligned}$$

Definition 1.4 (spectrum, spectral radius)**.** The *spectrum* $\operatorname{spec}(B)$ of $B \in \mathbb{C}^{n\times n}$ is the set of all its eigenvalues, and the *spectral radius* $\rho(B)$ is

$$\rho(B) := \max\{|\lambda| : \lambda \in \operatorname{spec}(B)\}$$

We are in a finite-dimensional space: $|\operatorname{spec}(B)| \leq n$.

Theorem 1.5. *Let $A = M - N$ be a splitting of $A \in \mathbb{C}^{n\times n}$, non-singular. The iteration*

$$x^{(k+1)} = Hx^{(k)} + M^{-1}b, \;\; k = 0, 1, \ldots$$

converges to $x^* = A^{-1}b$ for any starting vector $x^{(0)}$ iff $\rho(H) < 1$.

Proof. Assume first that $\rho(H) \geq 1$ and let (v, λ) be an eigenpair of H with $|\lambda| \geq 1$. For given b, take $x^{(0)} = x^* - v$. Then $e^{(0)} = v$ and, inductively,

$$e^{(k)} = \lambda^k v,$$

which implies $\lim_{k\to\infty} e^{(k)} \neq 0$.

To prove the other direction we use the fact that for any operator norm $\|\cdot\|$ we have

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

see Lemma A.9. Thus, if $\rho(H) < 1$, take some $\bar{\rho} \in (\rho(H), 1)$. Then there exists k_0 s.t. for $k \geq k_0$ we have $\|H^k\| \leq \bar{\rho}^k$ and therefore, for any b and any $x^{(0)}$ we have

$$\|e^{(k)}\| = \|H^k e^{(0)}\| \leq \bar{\rho}^k \|e^{(0)}\| \to 0 \; (k \to \infty).$$

□

Remark 1.6. We can take the spectral radius $\rho(H)$ as a measure for the convergence speed, since for any norm $\|\cdot\|$ we have by Lemma A.9 that for any $\epsilon > 0$ and k sufficiently large

$$\|e^{(k)}\| \leq (\rho(H) + \epsilon)^k \|e^{(0)}\|.$$

Note that this is an *asymptotic result*, meaning that it holds for unspecified large k only.

Any iteration of the form

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

is called *linear* and *stationary*. Stationary refers to the fact that M and H do not depend on the iteration step k.

A linear stationary iteration always relies on a splitting.

Theorem 1.7. *Let $M, H \in \mathbb{C}^{n\times n}$, M non-singular be such that the iteration*

$$x^{(k+1)} = Hx^{(k)} + M^{-1}b, \;\; k = 0, 1, \ldots$$

converges to $x^* = A^{-1}b$ for any starting vector $x^{(0)}$ and for any $b \in \mathbb{C}^n$. Then $H = I - M^{-1}A = M^{-1}N$ with $A = M - N$.

Proof. Since the iteration converges to x^*, we have $x^* = Hx^* + M^{-1}b$, which gives $M(I - H)A^{-1}b = b$ for all b, i.e., $M(I - H)A^{-1} = I \Leftrightarrow H = I - M^{-1}A$. □

1.2 Standard splittings

Decompose $A = (a_{ij})$ as

$$A = D - L - U$$

$$D = \begin{pmatrix} a_{1,1} & & & & \\ & \ddots & & 0 & \\ & & \ddots & & \\ & 0 & & \ddots & \\ & & & & a_{n,n} \end{pmatrix} \quad \text{(diagonal)}$$

$$L = \begin{pmatrix} 0 & & & & \\ -a_{2,1} & \ddots & & 0 & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ -a_{n,1} & \dots & \dots & -a_{n,n-1} & 0 \end{pmatrix} \quad \begin{matrix} \text{(negative)} \\ \text{lower triangular part} \end{matrix}$$

$$U = \begin{pmatrix} 0 & -a_{1,2} & \dots & \dots & -a_{1,n} \\ & \ddots & & & \vdots \\ & & \ddots & & \vdots \\ & 0 & & \ddots & -a_{n-1,n} \\ & & & & 0 \end{pmatrix} \quad \begin{matrix} \text{(negative)} \\ \text{upper triangular part} \end{matrix}$$

Definition 1.8 (standard splitting methods)**.** With the iterates $x^{(k)}$ and the residuals $r^{(k)}$

(i) *Jacobi* takes the splitting with $M = D$, i.e. for $k = 0, 1, \dots$

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j \neq i}^{n} a_{ij} x_j^{(k)} \right) = x_i^{(k)} + \frac{1}{a_{ii}} r_i^{(k)},$$
$$i = 1, \dots, n.$$

(ii) *Gauss-Seidel* takes the splitting with $M = D - L$, i.e. for $k = 0, 1, \ldots$

$$\begin{aligned} x_i^{(k+1)} &= \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \\ &= x_i^{(k)} + \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right), \ i = 1, \ldots, n. \end{aligned}$$

(iii) *Richardson* takes the splitting with $M = I$, i.e., for $k = 0, 1, \ldots$

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} + \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right) = x_i^{(k)} + r_i^{(k)}, \\ & \quad i = 1, \ldots, n. \end{aligned}$$

Relaxed variants of these methods scale the correction to the current iterate.

Definition 1.9 (relaxed methods)**.** Let $\omega \in \mathbb{C}$ be a relaxation parameter. Then

(i) *relaxed Jacobi* computes for $k = 0, 1, \ldots$

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} + \frac{\omega}{a_{ii}} r_i^{(k)}, \\ & \quad i = 1, \ldots, n. \end{aligned}$$

(ii) *relaxed Gauss-Seidel* (aka SOR "Successive OverRelaxation") computes for $k = 0, 1, \ldots$

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right), \ i = 1, \ldots, n.$$

(iii) *relaxed Richardson* computes for $k = 0, 1, \ldots$

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} + \omega \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right) = x_i^{(k)} + \omega r_i^{(k)}, \\ & \quad i = 1, \ldots, n. \end{aligned}$$

Exercise 1.10. Show that the relaxed methods belong to the following splittings (they have to belong to a splitting by Theorem 1.7):

- relaxed Jacobi: $A = \frac{1}{\omega}D - (\frac{1-\omega}{\omega}D + L + U)$
- SOR: $A = (\frac{1}{\omega}D - L) - (\frac{1-\omega}{\omega}D + U)$
- relaxed Richardson: $A = \frac{1}{\omega}I - (\frac{1}{\omega}I - A)$

1.3 Convergence

We consider two situations where substantial convergence results can be obtained: A is hpd or A is strictly diagonally dominant. See Section A.3 in the Appendix for notation and definitions, particularly for the A-norm and the Loewner ordering $\preceq$ on hpd matrices.

Theorem 1.11. *Let $A \succ 0$ and $0 < \lambda_{\min} \leq \lambda_{\max}$ be its smallest and largest eigenvalues,* $\operatorname{spec}(A) \subseteq [\lambda_{\min}, \lambda_{\max}]$. *Then relaxed Richardson converges for $\omega \in (0, \frac{2}{\lambda_{\max}})$, and the spectral radius of the iteration matrix $H_\omega = I - \omega A$ is smallest for $\omega = \omega_{opt} = \frac{2}{\lambda_{\min}+\lambda_{\max}}$.*

Proof. We have

$$\operatorname{spec}(H_\omega) = 1 - \omega \operatorname{spec}(A).$$

Discussing the function $f(z) = 1 - \omega z$ over the interval $[\lambda_{\min}, \lambda_{\max}]$; see Figure 1.1, we obtain

- $\operatorname{spec}(H_\omega) > 1$ for $\omega < 0$,
- $1 - \omega\lambda_{\max} < -1$ for $\omega > \frac{2}{\lambda_{\max}}$,
- $|\operatorname{spec}(H_\omega)| \leq \max\{1 - \omega\lambda_{\min}, |1 - \omega\lambda_{\max}|\} < 1$ for $\omega \in (0, \frac{2}{\lambda_{\max}})$,

and in the latter case $\max\{1 - \omega\lambda_{\min}, |1 - \omega\lambda_{\max}|\}$ is minimal if

$$1 - \omega\lambda_{\min} = -1 + \omega\lambda_{\max},$$

which gives $\omega_{\text{opt}} = \frac{2}{\lambda_{\min}+\lambda_{\max}}$. □

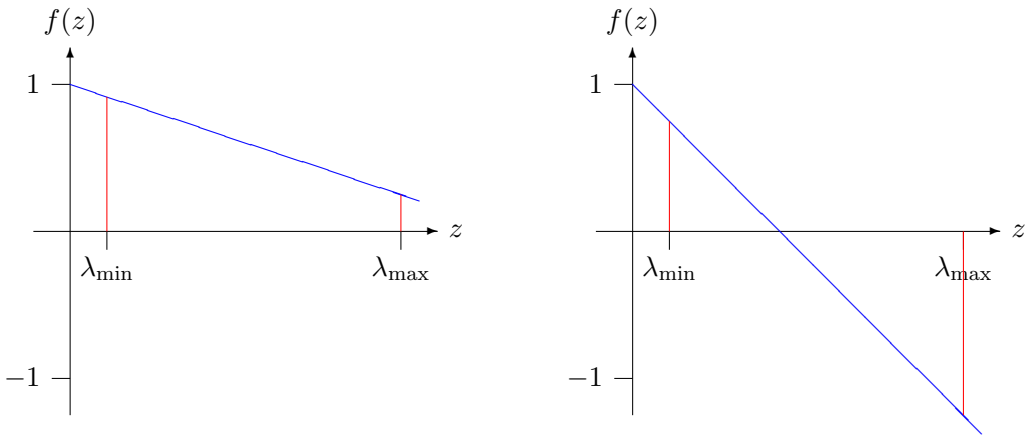

Figure 1.1: $f(z) = 1 - \omega z$ over the spectral interval of A hpd. Left $0 < \omega < \frac{1}{\lambda_{\max}}$, right: $\omega > \frac{2}{\lambda_{\max}}$

Remark 1.12. We have

$$\rho(H_{\omega_{\text{opt}}}) = 1 - \omega_{\text{opt}}\lambda_{min} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1},$$

where $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ is the *condition number* of A. So, even with the optimal choice for ω, the convergence of relaxed Richardson degrades as the condition number increases. This is a fundamental feature (and drawback) of very many iterative methods.

Definition 1.13. The splitting $A = M - N$ is called *P-regular* if $M + M^H - A \succeq 0$

Exercise 1.14. Show that the relaxed Richardson splitting is P-regular for $\omega \in (0, \frac{2}{\lambda_{\max}})$.

Theorem 1.15. *Let $A \succ 0$ and the splitting $A = M - N$ be P-regular. Then for $H = I - M^{-1}A$ we have*

$$\|H\|_A < 1.$$

Proof. Using the identity

$$H^H A H = A - AM^{-H}(M + M^H - A)M^{-1}A,$$

we see that

$$\langle Hx, Hx\rangle_A = x^H H^H AHx = \langle x, x\rangle_A - \langle M^{-1}Ax, (M + M^H - A)M^{-1}Ax\rangle.$$

For $x \neq 0$ the vector $M^{-1}Ax$ is nonzero, so that due to the positive definiteness of $M + M^H - A$ we obtain

$$x \neq 0 \Longrightarrow \langle Hx, Hx\rangle_A < \langle x, x\rangle_A.$$

Since the unit sphere in $\mathbb{C}^n$ is compact, this gives

$$\|H\|_A = \max\{\|Hx\|_A : \|x\|_A = 1\} < 1.$$

□

Note that since $\rho(H) \leq \|H\|_A$ (Lemma A.9), the theorem implies that iterations based on P-regular splittings converge.

Exercise 1.16. Show that for the relaxed Richardson splitting with $\omega \in \mathbb{R}$ we have

$$\|H_\omega\|_A = \max\{|1 - \omega\lambda_{\min}|, |1 - \omega\lambda_{\max}|\}.$$

Theorem 1.17 (convergence of SOR). *Let $A \succ 0$ and $\omega \in (0, 2)$. Then SOR converges.*

Proof. Since A is Hermitian, we have $D = L^H$. The SOR splitting

$$A = \underbrace{(\frac{1}{\omega}D - L)}_{M} - \underbrace{(\frac{1-\omega}{\omega}D + L^H)}_{N}$$

is P-regular, since

$$M + M^H - A = \frac{2-\omega}{\omega}D \succ 0 \text{ for } \omega \in (0, 2)$$

□

Remark 1.18. The proof used that the diagonal D of the hpd matrix A is hpd. With e_i denoting the ith canonical unit vector this can be seen as follows

$$a_{ii} = \langle e_i, Ae_i\rangle > 0 \text{ for } i = 1, \ldots, n,$$

and

$$\langle x, Dx\rangle = \sum_{i=1}^{n} a_{ii}|x_i|^2 > 0 \text{ for } x \neq 0.$$

The Jacobi iteration does not necessarily converge if A is hpd. We, however, can state the following non-exciting result.

Theorem 1.19. *Let A be hpd and assume that $\frac{2}{\omega}D - A = \frac{2-\omega}{\omega}D + L + L^H \succ 0$. Then relaxed Jacobi converges. In particular, standard Jacobi ($\omega = 1$) converges if $D + L + L^T$ is hpd.*

Proof. The relaxed Jacobi splitting

$$A = \underbrace{\frac{1}{\omega}D}_{M} - (\underbrace{\frac{1-\omega}{\omega}D - L - L^H}_{N})$$

satisfies

$$M + M^H - A = \frac{2}{\omega}D - A = \frac{2-\omega}{\omega}D + L + L^H.$$

Thus, the assumption just says that the splitting is P-regular. □

Definition 1.20. The matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ is *strictly diagonally dominant* (by rows), if

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|,\ i = 1, \ldots, n.$$

It will be notationally easier if we use a componentwise extension of the modulus $|\cdot|$ and $<, \leq$ from the reals to matrices (and, similarly, to vectors). Let $A = (a_{ij}), B = (b_{ij}) \in \mathbb{R}^{n\times m}$:

$$\begin{array}{rcl} |A| & = & (|a_{ij}|) \in \mathbb{R}^{n\times m}, \\ A & \leq (<) & B:\ a_{ij} \leq (<)\ b_{ij}, i = 1, \ldots, n, j = 1, \ldots, m. \end{array}$$

Positive and non-negative matrices A and vectors u are then those with $0 < A, 0 < v$ and $0 \leq A, 0 \leq u$, respectively. We also denote $e = (1, \ldots, 1)^T$ the vector of all ones.

With this notation, strict diagonal dominance can be expressed as

$$(|L| + |U|)e < |D|e.$$

The following theorem shows that Jacobi and Gauss-Seidel converge for strictly diagonally dominant matrices.

Theorem 1.21. *Let A be strictly diagonally dominant. Then*

(i) $\|D^{-1}(L+U)\|_\infty < 1$

(ii) $\|(D-L)^{-1}U\|_\infty < 1$.

Proof. First note that with our notation we have $\|B\|_\infty = \max_{i=1}^n (|B|e)_i$ for a matrix B.

(i) We multiply the vector inequality $(|L|+|U|)e < |D|e$ with the non-negative matrix $|D|^{-1} = |D^{-1}|$ and get

$$\underbrace{|D|^{-1}(|L|+|U|)}_{=|D^{-1}(L+U)|} e < e.$$

(ii) This is a bit more involved. We first show

$$|(D-L)^{-1}| \le (|D|-|L|)^{-1} \tag{1.2}$$

in three steps.

Step 1: Since $(D-L)^{-1} = (I-D^{-1}L)^{-1}D^{-1}$ and thus $|(D-L)^{-1}| = |(I-D^{-1}L)^{-1}| \cdot |D^{-1}|$, it is sufficient to consider the case $D = I$.

Step 2: Since $L^n = 0$ (L^k has its upper triangle and its first $k-1$ lower subdiagonals equal to 0), the identity

$$(I-L)\sum_{k=0}^{n-1} L^k = I - L^n = I$$

gives

$$(I-L)^{-1} = \sum_{k=0}^{n-1} L^k \text{ and } (I-|L|)^{-1} = \sum_{k=0}^{n-1} |L|^k.$$

Step 3: Since $L^k \le |L|^k$ for all k we get (1.2).

We finish the proof for (ii) using (1.2) in

$$|(D-L)^{-1}U|e \le (|(D-L)^{-1}| \cdot |U|)e \le (|D|-|L|)^{-1} \cdot |U|e < e,$$

where the last inequality follows by multiplying the strictly diagonal dominance condition $(|D|-|L|)e > |U|e > 0$ with the non-negative matrix $(|D|-|L|)^{-1}$

□

1.3.1 Two-step splitting methods

It can be conceptually advantageous to express a linear stationary iteration in two steps, using two splittings $A = M_1 - N_1 = M_2 - N_2$. The two-step iteration is

$$\left.\begin{array}{rcl} x^{(k+1/2)} & = & M_1^{-1}(N_1 x^{(k)} + b), \\ x^{(k+1)} & = & M_2^{-1}(N_2 x^{(k+1/2)} + b) \end{array}\right\}, k = 0, 1, \ldots$$

Lemma 1.22. *A two-step method is equivalent to a standard stationary linear method for the splitting $A = M - N$ with*

$$M^{-1} = M_2^{-1}(I + N_2 M_1^{-1}), H = M^{-1}N = M_2^{-1}N_2 M_1^{-1}N_1$$

Proof. Plug the definition of $x^{(k+1/2)}$ into that for $x^{(k+1)}$. □

Remark 1.23. Note that we have

$$\begin{array}{rcl} H & = & I - M_2^{-1}A - M_1^{-1}A + M_2^{-1}AM_1^{-1}A \\ & = & I - M_2^{-1}(M_1 + M_2 - A)M^{-1}A \\ M^{-1} & = & M_2^{-1}(I + (M_2 - A)M_1^{-1}) = M_2^{-1}(M_1 + M_2 - A)M_1^{-1} \end{array}$$

This shows that if $\rho(H) < 1$,i.e., the iteration converges, we have that $M_2^{-1}(M_1 + M_2 - A)$ is non-singular which in turn implies that M^{-1} actually exists (as an inverse). This need not be so in the non-convergent case.

Definition 1.24 (SSOR iteration)**.** The *symmetric relaxed Gauss-Seidel method (SSOR)* is the two-step method with

$$M_1 = \frac{1}{\omega}D - L, M_2 = \frac{1}{\omega}D - U.$$

So SSOR does a "forward sweep" over the variables followed by a "backward sweep".

For SSOR, the matrix M^{-1} from Lemma 1.22 is given as

$$M^{-1} = \left(\frac{1}{\omega}D - U\right)^{-1}\left(\frac{2-\omega}{\omega}D\right)\left(\frac{1}{\omega}D - L\right)^{-1}. \tag{1.3}$$

It is hpd if A is hpd as we will see now.

Theorem 1.25. *Let $A \succeq 0$ and $\omega \in (0, 2)$. Then the matrix M^{-1} from (1.3) is indeed non-singular, and the SSOR-splitting $A = M - N$ is P-regular, implying $\|I - M^{-1}A\|_A < 1$ and thus convergence of the SSOR iteration.*

Proof. We have $U = L^H$ and thus $M = (\frac{1}{\omega}D - L)(\frac{2-\omega}{\omega}D)^{-1}(\frac{1}{\omega}D - L)^H$. This shows that $M \succ 0$. The proof now uses the equality

$$\begin{aligned} M - A &= (\frac{1}{\omega}D - L)(\frac{2-\omega}{\omega}D)^{-1}(\frac{1}{\omega}D - L)^H - D + L + L^H \\ &= \frac{1}{\omega(2-\omega)}\left((1-\omega)D + \omega L\right) D^{-1} \left((1-\omega)D + \omega L\right)^H, \end{aligned}$$

which can be verified by simply (aka stupidly) computing all individual factors and compare. This shows $M - A \succeq 0$ and thus $M + M^H - A \succ 0$, i.e., the splitting is P-regular. □

Any matrix $A \in \mathbb{C}^{n\times n}$ can be decomposed into a Hermitian and skew-Hermitian part,

$$A = \frac{1}{2}(A + A^H) + \frac{1}{2}(A - A^H) =: H + S. \tag{1.4}$$

This gives rise to 2-step splitting iteration.

Definition 1.26 (HSS iteration)**.** The splitting (1.4) is termed the *Hermitian-skew Hermitian splitting (HSS)*, and for given $\alpha \in \mathbb{R}$, the resulting two step splitting method is the *HSS iteration*

$$\left.\begin{aligned} x^{(k+1/2)} &= (\alpha I + H)^{-1}((\alpha I - S)x^{(k)} + b), \\ x^{(k+1)} &= (\alpha I + S)^{-1}((\alpha I - H)x^{(k+1/2)} + b) \end{aligned}\right\}, k = 0, 1, \ldots$$

By Lemma 1.22, the HSS iteration is induced by the single splitting $A = M - N$ with

$$\begin{aligned} M^{-1} &= (\alpha I + S)^{-1}(I + (\alpha I - H)(\alpha I + H)^{-1}), \\ K(\alpha) := I - M^{-1}A &= (\alpha I + S)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S). \end{aligned}$$

We now write $K(\alpha)$ for the iteration matrix $I - M^{-1}A$ to avoid confusion with the Hermitian part H of A.

To prove a convergence result for the HSS iteration we need an auxiliary result on the matrix *Cayley transform*

$$B \to C(\alpha) := (\alpha I + B)^{-1}(\alpha I - B). \tag{1.5}$$

Note that the two factors in (1.5) commute as can be seen by multiplying the equality

$$(\alpha I - B)(\alpha I + B) = (\alpha I + B)(\alpha I - B)$$

from the left and the right with $(\alpha I - B)^{-1}$.

Lemma 1.27 (properties of the Cayley transform). *Let $B \in \mathbb{C}^{n\times n}$, $\alpha \in \mathbb{R}$ and assume that $(\alpha I + B)$ is non-singular. Then*

(i) If $B^H = -B$, then $C(\alpha)^H C(\alpha) = I$, i.e., $C(\alpha)$ is unitary, $\|C(\alpha)\|_2 = 1$.

(ii) If $B \succ 0$ and $\alpha > 0$, then $\|C(\alpha)\|_2 < 1$, i.e., $C(\alpha)$ is a contraction wrt. the 2-norm.

Proof. For (i) we have

$$\begin{aligned} C(\alpha)^H C(\alpha) &= \big((\alpha I - S)(\alpha I + S)^{-1}\big)^H \cdot (\alpha I - S)(\alpha I + S)^{-1} \\ &= (\alpha I + S^H)^{-1}(\alpha I - S^H) \cdot (\alpha I - S)(\alpha I + S)^{-1} \\ &= (\alpha I - S)^{-1}(\alpha I + S) \cdot (\alpha I - S)(\alpha I + S)^{-1} \\ &= (\alpha I + S)(\alpha I - S)^{-1}(\alpha I - S)(\alpha I + S)^{-1} \\ &= I. \end{aligned}$$

For (ii) we first observe that now $C(\alpha)$ is Hermitian, so

$$\|C(\alpha)\|_2 = \max|\operatorname{spec}(C(\alpha))| = \max\left\{\left|\frac{\alpha-\lambda}{\alpha+\lambda}\right|, \lambda \in \operatorname{spec}(A)\right\}.$$

But for $\alpha > 0$ and $\lambda > 0$ we always have $|\alpha - \lambda| < |\alpha + \lambda| = \alpha + \lambda$. □

Theorem 1.28. *Let $A \in \mathbb{C}^{n\times n}$ be positive definite in the sense that its Hermitian part H satisfies $H \succ 0$. Then*

$$\rho(K(\alpha)) \le \sigma(\alpha) := \max\left\{\left|\frac{\alpha-\lambda}{\alpha+\lambda}\right| : \lambda \in \operatorname{spec}(H)\right\}. \tag{1.6}$$

In particular, $\rho(K(\alpha)) < 1$ if $\alpha > 0$.

Proof. The "in particular part" is what we already used in the proof of Lemma 1.27.

To prove (1.6), we use the similarity invariance of the spectrum (Lemma A.10) to see that $K(\alpha)$ and

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

have the same spectrum. Therefore,

$$\begin{aligned}\rho(K(\alpha)) &= \rho((\alpha I - S)(\alpha I + S)^{-1}(\alpha I - H)(\alpha I + H)^{-1}) \\ &\leq \|(\alpha I - S)(\alpha I + S)^{-1}\|_2 \cdot \|(\alpha I - H)(\alpha I + H)^{-1}\|_2 \\ &= \sigma(\alpha),\end{aligned}$$

where the last equality is due to Lemma 1.27. □

Remark 1.29. An important question is how to choose the parameter α. One possibility is to minimize the bound $\sigma(\alpha)$ for the spectral radius, which assumes knowledge about $\mathrm{spec}(H)$. If we know

$$\mathrm{spec}(H) \subseteq [a, b] \text{ with } a > 0,$$

then the quantity

$$\max\{\left|\frac{\alpha - \lambda}{\alpha + \lambda}\right| : \lambda \in [a, b]\}$$

is minimal for $\alpha = \alpha_{\mathrm{opt}} = \sqrt{ab}$. Then

$$\sigma(\alpha_{\mathrm{opt}}) = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \text{ where } \kappa = \frac{b}{a}.$$

Note that κ is the condition number of H if $[a, b] = [\lambda_{\min}, \lambda_{\max}]$, the best possible spectral interval for H.

Exercise 1.30. Prove the remark. Hint: Show first that for given $\lambda \in [a, b]$ we have

$$\begin{aligned}\text{if } \alpha \geq \lambda: &\quad \left|\frac{\alpha - \lambda}{\alpha + \lambda}\right| = 1 - \frac{2\lambda}{\alpha + \lambda} \leq 1 - \frac{2a}{a + \alpha} \\ \text{if } \alpha \leq \lambda: &\quad \left|\frac{\alpha - \lambda}{\alpha + \lambda}\right| = 1 - \frac{2\alpha}{\alpha + \lambda} \leq 1 - \frac{2\alpha}{b + \alpha}\end{aligned}$$

Then discuss $\max\{1 - \frac{2a}{a+\alpha}, 1 - \frac{2\alpha}{b+\alpha}\}$ as a function of α.

Remark 1.31. In all splittings considered before, systems with M (or M_1, M_2 in the two-step case) were easy to solve since the matrices were diagonal or lower triangular. This is not the case for the HSS-splitting. So its practical use is questionable, unless one turns to *inexact* HSS methods where an additional, "inner" iteration is used to solve the systems with matrices $\alpha I + S$, $\alpha I + H$ arising in each "outer" HSS iteration.

Programming exercises

Exercise 1.32 (spectral radius vs norm)**.** For $n \in \mathbb{N}$ let A be the tridigonal matrix

$$A = \begin{pmatrix} 2 & -2 & & & \\ -0.5 & 2 & -2 & & \\ & \ddots & \ddots & \ddots & \\ & & -0.5 & 2 & -2 \\ & & & -0.5 & 2 \end{pmatrix}$$

Implement Jacobi for $Ax = b$.

- Take a random right hand side b and $x^{(0)} = 0$.
- Plot the relative residual 2-norm as a function of the iteration index m. Stop after 1 000 iterations.
- Test $n = 8, 16, 32, 64$

Interpret this in the light of the fact that $\rho(H) = \cos(\frac{\pi}{n+1})$ for the Jacobi iteration matrix H.

Exercise 1.33 (HSS iteration for model problem)**.**
The Matlab function `diff_advec(N)` produces matrices arising from a finite difference discretization of an advection-diffusion equation on a grid with $N \times N$ interior grid points on the unit square. It returns several different matrices in a cell array. Work with B, the `{1}{4}` cell entry. If the equation evolves in time, the implicit Euler scheme has to solve linear systems with $A = \frac{1}{(N-1)^2} I + B$.

Implement the HSS scheme and apply it to A.

- Plot the relative norm of the residual as a function of the iteration number.
- Try $N = 10, 20, 40$.
- Discuss what you observe.