# CLASSICAL ITERATIVE METHODS

LONG CHEN

We discuss classic iterative methods for solving the linear operator equation

$$Au = f, \tag{1}$$

posed on a finite dimensional Hilbert space $\mathbb{V} \cong \mathbb{R}^N$ equipped with an inner product $(\cdot,\cdot)$. Here $A : \mathbb{V} \to \mathbb{V}$ is a *symmetric and positive definite (SPD)* operator, $f \in \mathbb{V}$ is given, and we seek $u \in \mathbb{V}$ satisfying (1).

The direct approach is to form $A^{-1}$ or compute the action $A^{-1}f$. Gaussian elimination or the LU factorization remains the standard method. It is a black–box procedure that applies to general matrices. For a dense matrix, one matrix–vector multiplication costs $\mathcal{O}(N^2)$ operations and a straightforward Gauss elimination costs $\mathcal{O}(N^3)$, which is too expensive when $N$ is large. State–of–the–art direct solvers achieve nearly linear complexity for certain structured sparse matrices; see, for example, [3].

When $A$ is sparse, the number of nonzero entries is $\mathcal{O}(N)$, so one matrix–vector multiplication costs only $\mathcal{O}(N)$. In this setting, it is natural to design solvers with optimal or near–optimal complexity, such as $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$. This means that evaluating $A^{-1}f$ should require only a small number of applications of $A$ to a vector.

We begin with a basic residual–correction iteration and then study several classical iterative schemes. For algorithms achieving $\mathcal{O}(N)$ complexity with uniformly bounded contraction factors, we refer the reader to *Introduction to Multigrid Methods*.

## 1. RESIDUAL-CORRECTION METHOD

We follow Xu [4, 5, 6] and introduce an iterative method in residual–correction form. Starting from an initial guess $u_0 \in \mathbb{V}$, one iteration computes $u_{k+1}$ from $u_k$ by:

(1) forming the residual $r = f - Au_k$;
(2) computing a correction $e = Br$ with a nonsingular *preconditioner* $B \approx A^{-1}$;
(3) updating the iterate $u_{k+1} = u_k + e$.

Given $B$, define the affine map

$$\Phi_B(u; f) = u + B(f - Au) = (I - BA)u + Bf.$$

Then the residual–correction method is

$$u_{k+1} = \Phi_B(u_k; f) = u_k + B(f - Au_k). \tag{2}$$

The map $\Phi_B(\cdot; 0)$ is linear in $u$ and thus (2) is called a linear iterative method.

Since the exact solution $u$ satisfies $u = \Phi_B(u; f)$, we obtain the error relation

$$u - u_{k+1} = (I - BA)(u - u_k). \tag{3}$$

The matrix $E = I - BA$ is the error amplification operator (also called the iteration matrix).

Another popular formulation of linear iterative method is based on a splitting of $A$ [2]. Let $A = M - N$ with $M$ nonsingular, where $M$ is chosen as the dominant part of $A$. Rewriting the equation as

$$Mu - Nu = f,$$

we obtain the matrix–splitting iteration

$$u_{k+1} = M^{-1}(Nu_k + f). \tag{4}$$

Comparing the residual–correction and matrix–splitting forms, we see that

$$B = M^{-1}, \qquad N = M - A = B^{-1} - A.$$

The matrix–splitting method is slightly more efficient as $Nx$ is cheaper to compute than $Ax$. The residual–correction framework highlights the step of solving the residual equation $Ae = r$, whereas the update (4) acts directly on $u$ and is often called the direct update form.

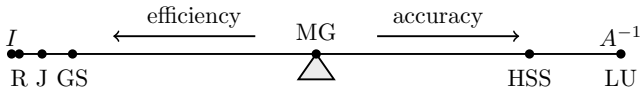


FIGURE 1. Efficiency vs. accuracy of solvers with multigrid (MG) achieving a balanced compromise.

The art of constructing *efficient* iterative methods lies in the design of $B$, which must capture the essential features of $A^{-1}$ while remaining inexpensive to apply. In this setting, the term "efficient" refers to the following two requirements:

(1) Each iteration costs only $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$ operations.
(2) The contraction rate is strictly less than 1 and remains independent of $N$.

## 2. CLASSIC ITERATIVE METHODS

Let us consider the case $\mathbb{V} = \mathbb{R}^N$ and let $\boldsymbol{A}$ be an SPD matrix. We derive several linear iterative methods based on the splitting

$$\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}.$$

Here $\boldsymbol{D}$, $\boldsymbol{L}$, and $\boldsymbol{U}$ are the diagonal, strictly lower triangular, and strictly upper triangular parts of $\boldsymbol{A}$, respectively. A list of corresponding preconditioner operators is given below:

- Richardson: $\boldsymbol{B}_{\mathrm{R}} = \alpha \boldsymbol{I}$
- Jacobi: $\boldsymbol{B}_{\mathrm{J}} = \boldsymbol{D}^{-1}$
- Weighted Jacobi: $\boldsymbol{B}_{\mathrm{DJ}} = \alpha \boldsymbol{D}^{-1}$
- Forward Gauss–Seidel: $\boldsymbol{B}_{\mathrm{GS}} = (\boldsymbol{D} + \boldsymbol{L})^{-1}$
- Backward Gauss–Seidel: $\boldsymbol{B}_{\mathrm{GS}} = (\boldsymbol{D} + \boldsymbol{U})^{-1}$
- Symmetric Gauss–Seidel: $\bar{\boldsymbol{B}}_{\mathrm{GS}} = (\boldsymbol{D} + \boldsymbol{U})^{-1}\, \boldsymbol{D}\, (\boldsymbol{D} + \boldsymbol{L})^{-1}$
- Successive over-relaxation (SOR): $\boldsymbol{B}_{\mathrm{SOR}} = \alpha(\boldsymbol{D} + \alpha \boldsymbol{L})^{-1}$
- Symmetric SOR: $\boldsymbol{B}_{\mathrm{SSOR}} = \alpha(2-\alpha)\, (\boldsymbol{D} + \alpha \boldsymbol{U})^{-1}\, \boldsymbol{D}\, (\boldsymbol{D} + \alpha \boldsymbol{L})^{-1}$

We use the forward Gauss–Seidel method as an example to illustrate the algorithmic form of a linear iterative scheme. Starting from the residual–correction update

$$u_{k+1} = u_k + (D + L)^{-1}(f - Au_k),$$

multiply both sides by $D + L$ to obtain

$$(D + L)u_{k+1} = (D + L)u_k + f - Au_k.$$

Since $A = D + L + U$, this gives the formal relation

$$u_{k+1} = D^{-1}\big(f - Lu_{k+1} - Uu_k\big). \tag{5}$$

This leads to the following in-place implementation of one Gauss–Seidel sweep:

```
for i=1:N
    u(i) = a_ii^{-1} (b(i)  -∑_{j=1}^{i-1} a_ij u(j) -∑_{j=i+1}^{N} a_ij u(j));
end
```

In the above algorithm, we use only one vector `u` to store both $u_{k+1}$ and $u_k$. The transition from $u_k$ to $u_{k+1}$ is built into the loop. Indeed the classical derivation of Gauss–Seidel solves the $i$-th equation

$$a_{i1}u^1 + \cdots + a_{ii}\,u^i + \cdots + a_{iN}u^N = f^i,$$

treating $u^i$ as the only unknown and moving all other terms to the right-hand side.

The form (5) is a direct update scheme. One Gauss–Seidel iteration requires essentially the same amount of work as a single matrix–vector multiplication with $A$. In contrast, the correction form

$$u_{k+1} = u_k + (D + L)^{-1}\,(f - Au_k),$$

requires both a residual evaluation and a forward substitution, and is therefore almost twice as expensive. In MATLAB, however, it is often easier and faster to implement the correction form:

```
u = u + tril(A)\(f-A*u);
```

For lower or upper triangular matrices, the inverse is applied by forward or backward substitution, and MATLAB performs an internal type check to choose the appropriate routine.

## 3. Convergence analysis of residual-correction methods

In this section we analyze the convergence of the linear residual–correction method and its variants. The $A$-inner product and $A$-symmetry play a central role in the analysis.

### 3.1. Symmetry and inner products.

Given an SPD operator $A$ on $\mathbb{V}$, we introduce the $A$-inner product

$$(u, v)_A := (Au, v) = (u, Av), \qquad u, v \in \mathbb{V}.$$

We use $(\mathbb{V}, I)$ and $(\mathbb{V}, A)$ to denote the same linear space equipped with the standard and $A$-inner products. It is the structure induced by $(\mathbb{V}, A)$ that plays a central role in the convergence analysis.

We use $^\intercal$ for the adjoint with respect to $(\mathbb{V}, I)$ and $^*$ for the adjoint with respect to $(\mathbb{V}, A)$:

$$(B^\intercal u, v) = (u, Bv), \qquad u, v \in \mathbb{V},$$
$$(B^* u, v)_A = (u, Bv)_A, \qquad u, v \in \mathbb{V}.$$

A direct calculation shows

$$B^* = A^{-1}B^\intercal A. \tag{6}$$

If we treat $A$ as a basis transformation matrix, $B^*$ and $B^\intercal$ are different representation of the dual of $B$ in different bases.

An operator $M$ is symmetric with respect to $(\cdot,\cdot)$ if $M = M^\intercal$, and symmetric with respect to $(\cdot,\cdot)_A$ if $M = M^*$. In functional analysis, such operators are called self-adjoint, and the notion depends on the chosen inner product of the underlying Hilbert space. Throughout, when we say "symmetric", we refer to symmetry in the default inner product $(\cdot,\cdot)$, and use "$A$-symmetric" to emphasize symmetry in $(\cdot,\cdot)_A$.

For two symmetric operators $X$ and $Y$, we write $X \geq Y$ if $(Xu, u) \geq (Yu, u)$ holds for all $u \in \mathbb{V}$. For $A$-symmetric operators, we write $X \geq_A Y$ to indicate $(Xu, u)_A \geq (Yu, u)_A$ for all $u \in \mathbb{V}$. This notation introduces a partial order on the set of symmetric operators.

3.2. **General convergence analysis.** Let $e_k = u - u_k$. Recall that the error equation of the residual–correction method is

$$e_{k+1} = (I - BA)e_k = (I - BA)^{k+1}e_0.$$

The method converges if and only if $\rho(I - BA) < 1$, which is equivalent to

$$|1 - \lambda| < 1 \qquad \text{for all } \lambda \in \sigma(BA).$$

Thus the spectrum of $BA$ must lie inside the open unit disk centered at $(1, 0)$ in the complex plane. Estimating the eigenvalues of $BA$ is therefore a key part of the analysis.

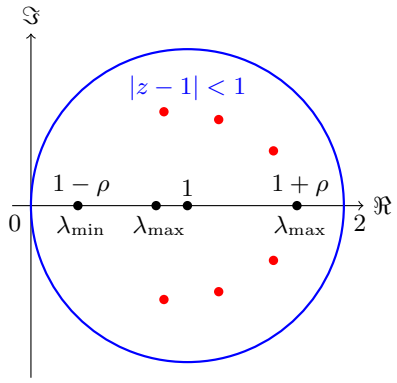


FIGURE 2. Spectrum of $BA$ contained in the open unit disk centered at 1 in the complex plane. When $B$ is symmetric, $\sigma(BA)$ is on the real axis.

For a linear operator $T \in \mathcal{L}(\mathbb{V}, \mathbb{V})$, the *spectrum* is

$$\sigma(T) = \{\, \lambda : \lambda \text{ is an eigenvalue of } T \,\},$$

and the *spectral radius* is

$$\rho(T) = \sup_{\lambda \in \sigma(T)} |\lambda|.$$

Eigenvalues depend only on the linear structure of an operator and do not involve the choice of inner product. However, selecting a suitable inner product, for example $A$-inner product, can greatly simplify the study of these eigenvalues.

3.3. **Symmetric scheme.** Eigenvalues of the operator $BA$ may be complex, which makes direct estimates difficult. When $B$ is symmetric, $BA$ need not be symmetric in $(\cdot, \cdot)$, as $(BA)^\intercal = A^\intercal B^\intercal = AB \neq BA$ in general. But $BA$ is $A$-symmetric.

**Lemma 3.1.** *When $B$ is symmetric, $BA$ is $A$-symmetric. If $B$ is also SPD, then $BA$ is $B^{-1}$-symmetric.*

*Proof.* It is straightforward to verify

$$(BA)^* = A^{-1}(BA)^\intercal A = BA.$$

The second one can be verified by symbolical change. □

The appropriate Hilbert space for the analysis is $(\mathbb{V}, A)$ rather than the default space $(\mathbb{V}, I)$. In this space the structure of $BA$ becomes transparent. In particular, $BA$ is symmetric in the $A$-inner product so all of its eigenvalues are real. Therefore

$$\rho(I - BA) = \max\bigl\{|1 - \lambda_{\min}(BA)|,\ |1 - \lambda_{\max}(BA)|\bigr\}. \tag{7}$$

From (7) we obtain a characterization of the convergence of a symmetric scheme.

**Theorem 3.2.** *Let $B$ be a symmetric preconditioner. Then the iterative scheme $\Phi_B$ converges if and only if*

$$0 < \lambda_{\min}(BA) \le \lambda_{\max}(BA) < 2.$$

From the condition $\rho(I - BA) < 1$ we can also derive bounds on the eigenvalues.

**Corollary 3.3.** *Let $B$ be a symmetric preconditioner and set $\rho = \rho(I - BA) < 1$. Then*

$$1 - \rho \le \lambda_{\min}(BA) \le \lambda_{\max}(BA) \le 1 + \rho.$$

*Proof.* The result follows from the scalar inequality $|1 - x| \le \rho$ for all $x \in [\lambda_{\min}, \lambda_{\max}]$. □

To obtain quantitative estimates, we need bounds on $\lambda_{\min}(BA)$ and $\lambda_{\max}(BA)$ in terms of comparisons between $B^{-1}$ and $A$, or equivalently between $B$ and $A^{-1}$.

**Lemma 3.4.** *Let $B$ be symmetric and nonsingular in $(\mathbb{V}, I)$. Then*

$$\lambda_{\min}(BA) = \inf_{u \in \mathbb{V}\setminus\{0\}} \frac{(ABAu, u)}{(Au, u)} = \inf_{u \in \mathbb{V}\setminus\{0\}} \frac{(Bu, u)}{(A^{-1}u, u)} = \left[\sup_{u \in \mathbb{V}\setminus\{0\}} \frac{(B^{-1}u, u)}{(Au, u)}\right]^{-1},$$

$$\lambda_{\max}(BA) = \sup_{u \in \mathbb{V}\setminus\{0\}} \frac{(ABAu, u)}{(Au, u)} = \sup_{u \in \mathbb{V}\setminus\{0\}} \frac{(Bu, u)}{(A^{-1}u, u)} = \left[\inf_{u \in \mathbb{V}\setminus\{0\}} \frac{(B^{-1}u, u)}{(Au, u)}\right]^{-1}.$$

*Proof.* By symmetry of $BA$ in $(\mathbb{V}, A)$, the first two identities for $\lambda_{\min}(BA)$ follow from the standard Rayleigh quotient characterization in $(\cdot, \cdot)_A$. For the third one, note that

$$\lambda_{\min}^{-1}(BA) = \lambda_{\max}((BA)^{-1}) = \sup_{u \in \mathbb{V}\setminus\{0\}} \frac{((BA)^{-1}u, u)_A}{(u, u)_A} = \sup_{u \in \mathbb{V}\setminus\{0\}} \frac{(B^{-1}u, u)}{(Au, u)}.$$

The identities for $\lambda_{\max}(BA)$ are analogous. □

Using the partial ordering notation for symmetric operators, the lower bound $\lambda_{\min}(BA) \ge c_0$ is equivalent to $BA \ge_A c_0 I$, and can be obtained from any of

$$c_0 B^{-1} \le A, \qquad B \ge c_0 A^{-1}, \qquad ABA \ge c_0 A.$$

Formally, these inequalities show that symmetric operators can be manipulated in the same way as real numbers with respect to the induced orderings.

3.4. **Symmetrization of general schemes.** For a general non-symmetric preconditioner $B$, the eigenvalues of $BA$ may be complex and are therefore difficult to estimate directly. To recover symmetry, we introduce the *symmetrized scheme*

$$\Phi_{\overline{B}} := \Phi_{B^\intercal} \circ \Phi_B,$$

given by the two-step iteration

(1) $u_{k+\frac{1}{2}} = u_k + B(f - Au_k)$,
(2) $u_{k+1} = u_{k+\frac{1}{2}} + B^\intercal(f - Au_{k+\frac{1}{2}})$.

This construction brings additional structure. From the definition,

$$I - \overline{B}A = (I - B^\intercal A)(I - BA), \tag{8}$$

which implies

$$\overline{B} = B^\intercal\big(B^{\intercal -1} + B^{-1} - A\big)B. \tag{9}$$

Since $\overline{B}$ is symmetric in $(\mathbb{V}, I)$, the operator $I - \overline{B}A$ is symmetric in $(\mathbb{V}, A)$. The next lemma shows that $I - \overline{B}A$ is also positive semidefinite.

**Lemma 3.5.** *Let $\overline{B}$ be defined by (9). Then*

$$I - \overline{B}A = (I - BA)^*(I - BA), \tag{10}$$

*where $^*$ denotes the adjoint with respect to the $A$-inner product.*

*Proof.* Using $(BA)^* = A^{-1}(BA)^\intercal A$, we obtain

$$(I - BA)^* = I - (BA)^* = I - A^{-1}(BA)^\intercal A = I - B^\intercal A,$$

which implies (10). □

Combining Lemma 3.4, 3.4, and Theorem 3.7, for the symmetrized scheme $\overline{B}$, eigenvalues of $\overline{B}A$ are real numbers and thus

$$\rho(I - \overline{B}A) = \max\{|1 - \lambda_{\min}(\overline{B}A)|, |1 - \lambda_{\max}(\overline{B}A)|\}. \tag{11}$$

By (9), $I - \overline{B}A$ is symmetric and semi-positive definite and thus $\lambda_{\min}(I - \overline{B}A) \geq 0$ which is equivalent to $\lambda_{\max}(\overline{B}A) \leq 1$. Therefore we have the following result.

**Lemma 3.6.** *For the symmetrized scheme $\Phi_{\overline{B}}$,*

$$\rho(I - \overline{B}A) = 1 - \lambda_{\min}(\overline{B}A). \tag{12}$$

We now present a criterion for the convergence of the symmetrized scheme.

**Theorem 3.7.** *The symmetrized iterative method $\Phi_{\overline{B}}$ converges if and only if*

$$B^{-1} + B^{\intercal -1} - A \text{ is SPD}. \tag{13}$$

*Proof.* By (12), the following statements are equivalent:

(1) $\Phi_{\overline{B}}$ converges;
(2) $\lambda_{\min}(\overline{B}A) > 0$;
(3) $\overline{B}A$ is SPD in $(\mathbb{V}, A)$;
(4) $\overline{B}$ is SPD in $(\mathbb{V}, I)$;
(5) $B^{-1} + B^{\intercal -1} - A$ is SPD in $(\mathbb{V}, I)$.

The equivalence of (4) and (5) follows from

$$\overline{B} = B^\intercal\big(B^{\intercal -1} + B^{-1} - A\big)B.$$

□

We summarize the convergence properties of the symmetrized scheme $\Phi_{\overline{B}}$ in the following theorem.

**Theorem 3.8.** *For the iterative scheme $\Phi_B$,*

$$\|I - BA\|_A^2 = \rho(I - \overline{B}A) = 1 - \left[\sup_{u \in \mathbb{V}\setminus\{0\}} \frac{(\overline{B}^{-1}u, u)}{(Au, u)}\right]^{-1}.$$

*Consequently, if*

$$(\overline{B}^{-1}u, u) \leq K\,(Au, u) \quad \textit{for all } u \in \mathbb{V}, \tag{14}$$

*then*

$$\|I - BA\|_A^2 \leq 1 - \frac{1}{K}.$$

3.5. **Relation of a scheme and its symmetrization.** The convergence of $\Phi_B$ and its symmetrization $\Phi_{\overline{B}}$ is related through the following inequality.

**Lemma 3.9.**

$$\rho(I - BA) \;\leq\; \sqrt{\rho(I - \overline{B}A)},$$

*and equality holds when $B$ is symmetric, i.e., $B = B^{\intercal}$.*

*Proof.* Using the relation between spectral radius and operator norms, we have

$$\rho(I - BA)^2 \;\leq\; \|\,I - BA\,\|_A^2 = \|(I - BA)^*(I - BA)\|_A = \rho(I - \overline{B}A).$$

The initial inequality becomes equality when $B$ is symmetric. □

Therefore, convergence of the symmetrized scheme $\Phi_{\overline{B}}$ implies convergence of the original scheme $\Phi_B$. However, when $B$ is non-symmetric (e.g., Gauss–Seidel), it is possible that $\Phi_B$ converges while $\Phi_{\overline{B}}$ fails to converge.

When $B$ is symmetric, condition (13) becomes both necessary and sufficient because equality holds in Lemma 3.9. In this case we may estimate either $\lambda_{\min}(BA)$, $\lambda_{\max}(BA)$, or $\lambda_{\min}(\overline{B}A)$. Note that even for symmetric $B$, its symmetrization is different:

$$\overline{B} = 2B - BAB,$$

which generally gives a better (but more expensive) preconditioner.

3.6. **Limitation of the spectral analysis.** Consider the iteration

$$x_k = Ex_{k-1} = E^k x_0.$$

Recall that for any consistent matrix norm $\|\cdot\|_X$,

$$\rho(E) \leq \|E\|_X, \qquad \rho(E) = \lim_{k\to\infty} \|E^k\|_X^{1/k}.$$

Thus $\rho(E)$ describes the *asymptotic* convergence of the sequence $\{x_k\}$. For any $\epsilon > 0$, there exists $k_0$ such that

$$\|E^k\|_X \leq (\rho(E) + \epsilon)^k, \qquad k \geq k_0.$$

However, this asymptotic information does *not* guarantee contraction at each iteration. Even when $\rho(E) < 1$, the norm $\|E^k\|$ may show *transient growth*—including exponential growth for $k \leq k_0$ followed by exponential decay for $k > k_0$. See Example D.2 in [1, Appendix D.4].

Bounding a matrix norm in terms of its spectral radius can be difficult; see, for example, [1, Appendix D.2], which uses the resolvent of a matrix to derive such bounds. For highly

non-normal matrices, the gap between the spectral radius and any induced matrix norm can be arbitrarily large. As an example, consider the rotation–shear matrix

$$R = \begin{pmatrix} 0 & 1 \\ \lambda & 0 \end{pmatrix}, \qquad \lambda > 0.$$

Its spectral radius is $\rho(R) = \sqrt{\lambda}$, while

$$\|R\| \geq \|Re_1\| = \lambda, \qquad e_1 = (1,0)^\top.$$

Hence

$$\|R\| \geq \sqrt{\lambda}\, \rho(R).$$

When $\lambda \gg 1$, the gap between the conditions $\rho(R) < 1$ and $\|R\| < 1$ becomes very large.

When the iteration matrix $E$ is diagonalizable, i.e.

$$E = T\Lambda T^{-1}, \qquad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

we obtain

$$\|E^k\| = \|T\Lambda^k T^{-1}\| \;\leq\; \|T\|\, \|\Lambda^k\|\, \|T^{-1}\| = \kappa(T)\, \rho(E)^k,$$

where the condition number $\kappa(T) = \|T\|\, \|T^{-1}\|$ may be very large.

When $E$ is unitarily diagonalizable, meaning that $T$ is unitary, then

$$\|E\| = \|\Lambda\| = \rho(E),$$

and the spectral radius fully describes the norm behavior. $E$ is unitarily diagonalizable is equivalent to $E$ is normal, i.e., for real matrices $EE^\intercal = E^\intercal E$.

## 4. Convergence analysis of classic iterative methods

We apply the convergence theory to analyze the convergence of several classic iterative methods. We begin with the Richardson method, for which $B = \alpha I$, and discuss the optimal choice of the damping parameter $\alpha$. For an SPD operator $A$, define the condition number $\kappa(A) := \lambda_{\max}(A)/\lambda_{\min}(A)$.

**Theorem 4.1.** *The Richardson iteration with $B = \alpha I$ converges if and only if*

$$0 < \alpha < \frac{2}{\lambda_{\max}(A)}.$$

*Moreover, the optimal convergence rate is obtained at*

$$\alpha^* = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)},$$

*and the corresponding optimal rate is*

$$\rho_{\alpha^*} = \frac{\kappa(A) - 1}{\kappa(A) + 1}.$$

*Proof.* Since $A$ is SPD, its eigenvalues are real and satisfy $\lambda_{\min}(A) > 0$. We have

$$\rho(I - \alpha A) = \max\left\{|1 - \alpha\lambda_{\min}(A)|,\ |1 - \alpha\lambda_{\max}(A)|\right\}.$$

The optimal $\alpha^*$ minimizes this maximum and is determined by the condition

$$\alpha^*\lambda_{\max}(A) - 1 = 1 - \alpha^*\lambda_{\min}(A).$$

The geometric interpretation is shown in Figure 4. □

We now analyze the convergence rate of Jacobi and weighted Jacobi iterations.

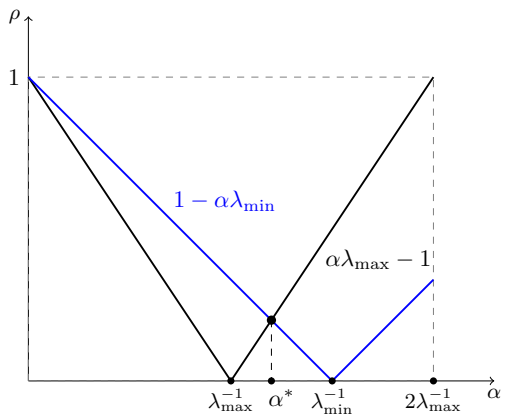


FIGURE 3. Convergence analysis of Richardson method

**Theorem 4.2.** *The Jacobi method converges if and only if*

$$2\boldsymbol{D}-\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$$

*is an SPD matrix.*

*Proof.* Since $\boldsymbol{B}_J=\boldsymbol{D}^{-1}$ is an SPD matrix, the characterization of its convergence follows from Theorem 3.7. □

A matrix $\boldsymbol{A}=(a_{ij})$ is called *diagonally dominant* if $a_{ii}\geq\sum_{j\neq i}|a_{ij}|$ for all $i$, and *strictly diagonally dominant* if it is diagonally dominant and there exists at least one $i$ such that $a_{ii}>\sum_{j\neq i}|a_{ij}|$. One can easily prove that a symmetric, strictly diagonally dominant matrix is SPD.

**Corollary 4.3.** *If $\boldsymbol{A}$ is strictly diagonally dominant, then the Jacobi iteration always converges.*

*Proof.* Note that if $\boldsymbol{A}=\boldsymbol{D}+\boldsymbol{L}+\boldsymbol{U}$ is strictly diagonally dominant, then so is

$$2\boldsymbol{D}-\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}.$$

The result then follows from the previous theorem. □

To study the weighted Jacobi iteration, we introduce the scaled matrix

$$\boldsymbol{A}_D=\boldsymbol{D}^{-1/2}\,\boldsymbol{A}\,\boldsymbol{D}^{-1/2}.$$

By the following exercise, $\sigma(\boldsymbol{A}_D)=\sigma(\boldsymbol{D}^{-1}\boldsymbol{A})$. We therefore reduce the analysis of the weighted Jacobi method to the Richardson method.

**Theorem 4.4.** *The weighted Jacobi method with $\boldsymbol{B}=\alpha\boldsymbol{D}^{-1}$ converges if and only if*

$$0<\alpha<\frac{2}{\lambda_{\max}(\boldsymbol{A}_D)}.$$

*Furthermore, the optimal convergence rate is achieved when*

$$\alpha^*=\frac{2}{\lambda_{\min}(\boldsymbol{A}_D)+\lambda_{\max}(\boldsymbol{A}_D)},$$

*and the corresponding optimal convergence factor is*

$$\rho_{\alpha^*} = \frac{\kappa(\boldsymbol{A}_D) - 1}{\kappa(\boldsymbol{A}_D) + 1},$$

*where* $\kappa(\boldsymbol{A}_D) = \lambda_{\max}(\boldsymbol{A}_D)/\lambda_{\min}(\boldsymbol{A}_D)$ *is the condition number of* $\boldsymbol{A}_D$.

The diagonal entries of the scaled matrix $\boldsymbol{A}_D$ are always equal to 1. An estimate of $\lambda_{\max}(\boldsymbol{A}_D)$ can be obtained from the Gershgorin circle theorem. For example, if $\boldsymbol{A}$ is diagonally dominated, then $\lambda_{\max}(\boldsymbol{A}_D) \leq 2$.

**Theorem 4.5.** *The Gauss–Seidel method always converges. For the forward Gauss–Seidel method with*

$$\boldsymbol{B} = (\boldsymbol{D} + \boldsymbol{L})^{-1},$$

*the convergence rate satisfies*

$$\|I - \boldsymbol{B}\boldsymbol{A}\|_{\boldsymbol{A}}^2 = \frac{c_0}{1 + c_0},$$

*where*

$$c_0 = \sup_{\|\boldsymbol{u}\|_{\boldsymbol{A}}=1} (\boldsymbol{D}^{-1}\boldsymbol{U}\boldsymbol{u},\, \boldsymbol{U}\boldsymbol{u}) = \sup_{\boldsymbol{u}\neq 0} \frac{(\boldsymbol{D}^{-1}\boldsymbol{U}\boldsymbol{u},\, \boldsymbol{U}\boldsymbol{u})}{(\boldsymbol{A}\boldsymbol{u},\, \boldsymbol{u})}.$$

*Proof.* A direct computation shows that

$$\boldsymbol{B}^{-\intercal} + \boldsymbol{B}^{-1} - \boldsymbol{A} = \boldsymbol{D},$$

which is SPD. Therefore, by Theorem 3.7, the Gauss–Seidel iteration always converges.

Moreover,

$$\overline{\boldsymbol{B}}^{-1} = \boldsymbol{A} + \boldsymbol{L}\,\boldsymbol{D}^{-1}\boldsymbol{U}.$$

Hence,

$$\lambda_{\min}^{-1}(\overline{\boldsymbol{B}}\boldsymbol{A}) = \sup_{\boldsymbol{u}\neq 0} \frac{(\overline{\boldsymbol{B}}^{-1}\boldsymbol{u},\, \boldsymbol{u})}{(\boldsymbol{A}\boldsymbol{u},\, \boldsymbol{u})} = 1 + c_0.$$

The result then follows from Theorem 3.8 and the symmetry identity $\boldsymbol{U} = \boldsymbol{L}^{\intercal}$. □

## 5. Exercise

**Exercise 5.1.** Derive the direct updated form of the Jacobi iteration and write its algorithmic description. Compare with G-S and list the main difference.

**Exercise 5.2.** Prove that if $B$ is symmetric and $B^{-1} > \frac{1}{2}A$, then $\Phi_B$ converges with a rate

$$\|I - BA\|_A^2 \leq 1 - \lambda_{\min}(B^{-1} + B^{\intercal -1} - A)\lambda_{\min}(A)\|B^{-1}\|^{-2}.$$

In view of matrix-splitting method, the condition $B^{-1} > \frac{1}{2}A$ means the matrix $M = B^{-1}$ is dominant (more than half).

**Exercise 5.3.** Consider $k$-steps of Richardson methods with different parameters $\alpha_1, \ldots, \alpha_k$. Then the error equation is

$$e_k = (I - \alpha_k A) \cdots (I - \alpha_1 A) e_0.$$

Consider the optimization problem of choosing $k$-parameters:

$$\min_{\alpha_i \in \mathbb{R}, i=1,\ldots,k} \left\{ \max_{\lambda \in [\lambda_{\min}(A), \lambda_{\max}(A)]} |(I - \alpha_k \lambda) \cdots (I - \alpha_1 \lambda)| \right\}. \tag{15}$$

Find the solution of (15) and derive the rate. This trick is known as Chebyshev acceleration.

**Exercise 5.4.** Let $A_{n\times r}$ and $B_{r\times n}$ be two matrices. Prove

$$\sigma(AB)\backslash\{0\} = \sigma(BA)\backslash\{0\}.$$

**Exercise 5.5.** Prove that the convergence rate of Richardson, weighted Jacobi method, and Gauss-Seidel method for the 5-point stencil finite difference method of the Poisson equation on a uniform mesh with size $h$, is like

$$\rho \leq 1 - Ch^2.$$

Thus when $h \to 0$, we will observe slow convergence of those classical iterative methods.

*Hint: For G-S, use the Hölder inequality of the 2-norm of a matrix $M$:*

$$\|M\|^2 \leq \|M\|_\infty \|M\|_1.$$

## REFERENCES


[1] R. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems.* Society for Industrial and Applied Mathematics, 2007.

[2] R. S. Varga. *Matrix iterative analysis*. Prentice-Hall, Englewood Cliffs, N. J., 1962.

[3] J. Xia, S. Chandrasekaran, M. Gu, and X. Li. Superfast multifrontal method for large structured linear systems of equations. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1382–1411, 2009.

[4] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34:581–613, 1992.

[5] J. Xu. An introduction to multilevel methods. 1997. published by Oxford University Press, New York.

[6] J. Xu. *Multilevel Finite Element Methods*. Lecutre Notes, 2004.