Classical Numerical Analysis, Chapter 06

Abner J. Salgado and Steven M. Wise

asalgad1@utk.edu swise1@.utk.edu
University of Tennessee

Chapter 06, Part 1 of 3
Linear Iterative Methods

Direct Versus Iterative Methods

As an alternative to the direct methods that we studied in the previous chapters, in the present chapter we will describe so called *linear* iteration methods for constructing sequences, $\{\mathbf{x}_k\}_{k=1}^{\infty} \subset \mathbb{C}^n$, with the desire that $\mathbf{x}_k \to \mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, as $k \to \infty$. The idea is that, given some $\varepsilon > 0$, we look for a $k \in \mathbb{N}$, such that

$$\|\mathbf{x} - \mathbf{x}_k\| \leq \varepsilon$$

with respect to some norm. In this context, ε is called the *stopping tolerance*.

Usually, we do not have a direct way of approximating the error. The residual is more readily available. Suppose that $\mathbf{x}_k$ is an approximation of $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. The error is $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$ and the residual is $\mathbf{r}_k = \mathbf{f} - \mathsf{A}\mathbf{x}_k = \mathsf{A}\mathbf{e}_k$. Recall that,

$$\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}\|} \leq \kappa(\mathsf{A}) \frac{\|\mathbf{r}_k\|}{\|\mathbf{f}\|}.$$

Thus, when $\kappa(\mathsf{A})$ is large, $\frac{\|\mathbf{r}_k\|}{\|\mathbf{f}\|}$, which is easily computable, may not be a good indicator of the size of the relative error $\frac{\|\mathbf{e}_k\|}{\|\mathbf{x}\|}$, which is not directly computable.

Linear Iterative Methods

Iterative Method

Definition

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ with $\det(\mathsf{A}) \neq 0$ and $\mathbf{f} \in \mathbb{C}^n$. An **iterative method** to find an approximate solution to $\mathsf{A}\mathbf{x} = \mathbf{f}$ is a process to generate a sequence of approximations $\{\mathbf{x}_k\}_{k=1}^{\infty}$ via an iteration of the form

$$\mathbf{x}_k = \varphi(\mathsf{A}, \mathbf{f}, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_{k-r}),$$

given the starting values $\mathbf{x}_0, \cdots \mathbf{x}_{r-1} \in \mathbb{C}^n$. Here

$$\varphi(\cdot, \cdot, \ldots, \cdot) : \mathbb{C}^{n\times n} \times \mathbb{C}^n \times \cdots \times \mathbb{C}^n \to \mathbb{C}^n$$

is called the **iteration function**. If $r = 1$, we say that the process is a **two–layer** method, otherwise we say it is a multilayer method.

Consistent, Linear Iterative Methods

Definition

Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ with $\mathbf{det}(\mathsf{A}) \neq 0$ and $\mathbf{f} \in \mathbb{C}^n$. Set $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. The two–layer iterative method

$$\mathbf{x}_k = \varphi(\mathsf{A}, \mathbf{f}, \mathbf{x}_{k-1}),$$

is said to be **consistent** iff $\mathbf{x} = \varphi(\mathsf{A}, \mathbf{f}, \mathbf{x})$, i.e., $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$ is a fixed point of $\varphi(\mathsf{A}, \mathbf{f}, \cdot)$. The method is **linear** iff

$$\varphi(\mathsf{A}, \alpha\mathbf{f}_1 + \beta\mathbf{f}_2, \alpha\mathbf{x}_1 + \beta\mathbf{x}_2) = \alpha\varphi(\mathsf{A}, \mathbf{f}_1, \mathbf{x}_1) + \beta\varphi(\mathsf{A}, \mathbf{f}_2, \mathbf{x}_2),$$

for all $\alpha, \beta \in \mathbb{C}$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{C}^n$.

Proposition (general form)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *with* $\det(\mathsf{A}) \neq 0$ *and* $\mathbf{f} \in \mathbb{C}^n$*. Any two–layer, linear, and consistent method can be written in the form*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathsf{C}\mathbf{r}(\mathbf{x}_k) = \mathbf{x}_k + \mathsf{C}(\mathbf{f} - \mathsf{A}\mathbf{x}_k), \tag{1}$$

for some matrix $\mathsf{C} \in \mathbb{C}^{n\times n}$*, where* $\mathbf{r}(\mathbf{z}) = \mathbf{f} - \mathsf{A}\mathbf{z}$ *is the residual vector.*

Proof.

A two layer method is defined by an iteration function

$$\varphi(\,\cdot\,,\,\cdot\,,\,\cdot\,) : \mathbb{C}^{n\times n} \times \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}^n.$$

Given φ, define the operator

$$\mathsf{C}\mathbf{z} = \varphi(\mathsf{A}, \mathbf{z}, \mathbf{0}).$$

This is a linear operator, due to the assumed linearity of the iteration function. Consequently, C can be identified as a square matrix.

Proof, Cont.

It follows from this definition, using the consistency and linearity of φ, that

$$(\mathsf{I}_n - \mathsf{CA})\mathbf{w} = \mathbf{w} - \varphi(\mathsf{A}, \mathsf{A}\mathbf{w}, \mathbf{0}) = \varphi(\mathsf{A}, \mathsf{A}\mathbf{w}, \mathbf{w}) - \varphi(\mathsf{A}, \mathsf{A}\mathbf{w}, \mathbf{0}) = \varphi(\mathsf{A}, \mathbf{0}, \mathbf{w}).$$

Furthermore, by linearity, we can write

$$\begin{aligned} \mathbf{x}_{k+1} &= \varphi(\mathsf{A}, \mathbf{f} + \mathbf{0}, \mathbf{0} + \mathbf{x}_k) \\ &= \varphi(\mathsf{A}, \mathbf{f}, \mathbf{0}) + \varphi(\mathsf{A}, \mathbf{0}, \mathbf{x}_k) \\ &= \mathsf{C}\mathbf{f} + (\mathsf{I}_n - \mathsf{CA})\mathbf{x}_k \\ &= \mathbf{x}_k + \mathsf{C}(\mathbf{f} - \mathsf{A}\mathbf{x}_k), \end{aligned}$$

as we intended to show. □

Note: If C is invertible, we can, if we like, write

$$\mathsf{C}^{-1}(\mathbf{x}_{k+1} - \mathbf{x}_k) + \mathsf{A}\mathbf{x}_k = \mathbf{f}.$$

Two-Layer Methods

Definition

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ with $\det(\mathsf{A}) \neq 0$ and $\mathbf{f} \in \mathbb{C}^n$. A method of the form

$$\mathsf{B}_{k+1}\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + \mathsf{A}\mathbf{x}_k = \mathbf{f},$$

where $\mathsf{B}_{k+1} \in \mathbb{C}^{n\times n}$ is invertible is called an **adaptive two–layer method**. If $\mathsf{B}_{k+1} = \mathsf{B}$, where B is invertible and independent of k, then the method is called a **stationary two–layer method**. If $\mathsf{B}_{k+1} = \frac{1}{\alpha_{k+1}}\mathsf{I}_n$, where $\alpha_{k+1} \in \mathbb{C}_\star$, then we say that the adaptive two–layer method is **explicit**.

Consider a stationary two–layer method and assume that B is invertible, then

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathsf{B}^{-1}(\mathbf{f} - \mathsf{A}\mathbf{x}_k), \tag{2}$$

from this it follows that, if $\{\mathbf{x}_k\}_{k\geq 0}$ converges, then it must converge to $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$. Of course, this form is equivalent to (1) with $\mathsf{C} = \mathsf{B}^{-1}$. The matrix B in the stationary, two–layer method is called the *iterator*.

Choosing the Iterator Matrix B

Let us now consider an extreme case, namely $\mathsf{B} = \mathsf{A}$. In this case we obtain

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathsf{B}^{-1}(\mathbf{f} - \mathsf{A}\mathbf{x}_k) = \mathbf{x}_k + \mathsf{A}^{-1}(\mathbf{f} - \mathsf{A}\mathbf{x}_k) = \mathsf{A}^{-1}\mathbf{f},$$

that is, we get the *exact* solution after one step.

The previous observation shows that choice of an iterator comes with two conflicting requirements:

1. The iterator B should be easy/cheap to invert.
2. The iterator B should "approximate" the matrix A well.

In essence, the art of iterative methods is concerned with finding good iterators.

The Error Transfer Matrix

Definition

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be invertible and $\mathbf{f} \in \mathbb{C}^n$. Suppose that $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$ and consider the stationary two–layer method (2) defined by the invertible matrix $\mathsf{B} \in \mathbb{C}^{n\times n}$. The matrix $\mathsf{T} = \mathsf{I}_n - \mathsf{B}^{-1}\mathsf{A}$ is called the **error transfer matrix** and satisfies

$$\mathbf{e}_{k+1} = \mathsf{T}\mathbf{e}_k,$$

where $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$ is the **error** at step k.

Here is our mission: we will seek to find conditions on T to guarantee that $\{\mathbf{e}_k\}_{k=0}^{\infty}$ converges to zero.

Spectral Convergence Theory

Theorem (convergence of linear methods)

Suppose that $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *are invertible,* $\mathbf{f}, \mathbf{x}_0 \in \mathbb{C}^n$ *are given, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*.*

1. *The sequence* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *defined by the linear, two–layer, stationary iterative method* (2) *converges to* $\mathbf{x}$ *for any starting point* $\mathbf{x}_0$ *iff* $\rho(\mathsf{T}) < 1$*, where* T *is the error transfer matrix* $\mathsf{T} = \mathsf{I}_n - \mathsf{B}^{-1}\mathsf{A}$*.*
2. *A sufficient condition for the convergence of* $\{\mathbf{x}_k\}_{k=1}^{\infty}$*, for any starting point* $\mathbf{x}_0$*, is the condition that* $\|\mathsf{T}\| < 1$*, for some induced matrix norm.*

Proof.

Before we begin the proof, observe that

$$\mathbf{e}_k = \mathsf{T}\mathbf{e}_{k-1} = \mathsf{T}^2\mathbf{e}_{k-2} = \cdots = \mathsf{T}^k\mathbf{e}_0.$$

Also, observe that $\mathbf{x}_k \to \mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, as $k \to \infty$, iff $\mathbf{e}_k \to \mathbf{0}$, as $k \to \infty$.

Suppose that $\mathbf{x}_k \to \mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, as $k \to \infty$, for any $\mathbf{x}_0$. Then $\mathbf{e}_k \to \mathbf{0}$, as $k \to \infty$, for any $\mathbf{e}_0$.

Proof, Cont.

Set $\mathbf{e}_0 = \mathbf{w}$, where $(\lambda, \mathbf{w})$ is any eigenpair of T, with $\|\mathbf{w}\|_\infty = 1$. Then

$$\mathbf{e}_k = \lambda^k \mathbf{e}_0,$$

and

$$|\lambda|^k = |\lambda|^k \|\mathbf{w}\|_\infty = \|\mathbf{e}_k\|_\infty \to 0.$$

It follows that $|\lambda| < 1$. Since λ was arbitrary, $\rho(\mathsf{T}) < 1$.

If $\rho(\mathsf{T}) < 1$, appealing to a Theorem from Chapter 3,

$$\lim_{k\to\infty} \mathbf{e}_k = \lim_{k\to\infty} \mathsf{T}^k \mathbf{e}_0 = \mathbf{0},$$

for any $\mathbf{e}_0$. Hence, $\mathbf{x}_k \to \mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$, as $k \to \infty$, for any $\mathbf{x}_0$.

Proof, Cont.

Suppose now that $\|\mathsf{T}\| < 1$ for some induced matrix norm. Since, for any induced matrix norm,

$$\rho(\mathsf{T}) \leq \|\mathsf{T}\|,$$

it follows that $\rho(\mathsf{T}) < 1$. Again, by the Theorem from Chapter 3,

$$\lim_{k\to\infty} \mathbf{e}_k = \lim_{k\to\infty} \mathsf{T}^k \mathbf{e}_0 = \mathbf{0},$$

for any $\mathbf{e}_0$. $\square$

Theorem (error estimate)

Let $\mathsf{A}, \mathsf{B} \in \mathbb{C}^{n\times n}$ *be invertible,* $\mathbf{x}_0, \mathbf{f} \in \mathbb{C}^n$ *are given, and* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$*. Let* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *be the sequence generated by the linear, two–layer, stationary method* (2)*. Assume that, for some induced norm,* $\|\mathsf{T}\| < 1$*. Then, the following estimates hold*

$$\|\mathbf{x} - \mathbf{x}_k\| \leq \|\mathsf{T}\|^k \|\mathbf{x} - \mathbf{x}_0\|,$$

$$\|\mathbf{x} - \mathbf{x}_k\| \leq \frac{\|\mathsf{T}\|^k}{1 - \|\mathsf{T}\|} \|\mathbf{x}_1 - \mathbf{x}_0\|.$$

Proof.

It follows, that $\mathbf{e}_k = \mathsf{T}^k \mathbf{e}_0$. By using the consistency and sub–multiplicativity of the induced matrix norm, we find

$$\|\mathbf{e}_k\| \leq \left\|\mathsf{T}^k\right\| \|\mathbf{e}_0\| \leq \|\mathsf{T}\|^k \|\mathbf{e}_0\|,$$

which proves the first estimate.

Proof, Cont.

To see the second, observe that, $\mathbf{e}_k = \mathsf{T}^{k-1}\mathbf{e}_1$, and thus $\mathsf{T}\mathbf{e}_k = \mathsf{T}^k\mathbf{e}_1$. Subtracting the last expression from $\mathbf{e}_k = \mathsf{T}^k\mathbf{e}_0$, we find $(\mathsf{I}_n - \mathsf{T})\,\mathbf{e}_k = \mathsf{T}^k\,(\mathbf{x}_1 - \mathbf{x}_0)$. Since $\|\mathsf{T}\| < 1$, a theorem guarantees that $\mathsf{I}_n - \mathsf{T}$ is invertible, and

$$\left\|(\mathsf{I}_n - \mathsf{T})^{-1}\right\| \le \frac{1}{1 - \|\mathsf{T}\|}.$$

Hence,

$$\mathbf{e}_k = (\mathsf{I}_n - \mathsf{T})^{-1}\,\mathsf{T}^k\,(\mathbf{x}_1 - \mathbf{x}_0)$$

and using the consistency and sub-multiplicativity of the norm, we get

$$\|\mathbf{e}_k\| \le \left\|(\mathsf{I}_n - \mathsf{T})^{-1}\right\|\,\|\mathsf{T}\|^k\,\|\mathbf{x}_1 - \mathbf{x}_0\| \le \frac{1}{1 - \|\mathsf{T}\|}\,\|\mathsf{T}\|^k\,\|\mathbf{x}_1 - \mathbf{x}_0\|\,.$$

The result is proven. □

Matrix Splitting Methods

The Basic Idea

Here we present some methods that are based on the idea of *matrix splitting*. Namely, we assume that we can *split* the coefficient matrix A as

$$\mathsf{A} = \mathsf{M} + \mathsf{N},$$

where M is invertible and, hopefully, easy to invert. Since $\mathsf{A}\mathbf{x} = \mathbf{f}$ iff $\mathsf{M}\mathbf{x} + \mathsf{N}\mathbf{x} = \mathbf{f}$, the strategy that we follow is then to construct a method of the form

$$\mathsf{M}\mathbf{x}_{k+1} + \mathsf{N}\mathbf{x}_k = \mathbf{f}.$$

Jacobi Method

Let $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ have non–zero diagonal elements, and consider the following splitting of A:

$$\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{U},$$

where $\mathsf{D} = \mathbf{diag}\left(a_{1,1}, \ldots, a_{n,n}\right)$ is the diagonal part of A, L is the strictly lower triangular part of A, and U, its strictly upper triangular part. Then the splitting method is

$$\mathsf{D}\mathbf{x}_{k+1} + (\mathsf{L} + \mathsf{U})\mathbf{x}_k = \mathbf{f}.$$

In other words,

$$\begin{aligned} \mathbf{f} &= \mathsf{D}\mathbf{x}_{k+1} + (\mathsf{A} - \mathsf{D})\mathbf{x}_k \\ &= \mathsf{D}(\mathbf{x}_{k+1} - \mathbf{x}_k) + \mathsf{A}\mathbf{x}_k \end{aligned}$$

so that

$$\mathbf{x}_{k+1} - \mathbf{x}_k = \mathsf{D}^{-1}\mathbf{r}(\mathbf{x}_k).$$

In other words, the *Jacobi Method* is a stationary two-layer method with the iterator

$$\mathsf{B} = \mathsf{B}_{\mathrm{J}} = \mathsf{D}.$$

The Error Transfer Matrix for the Jacobi Method

In this case, the error transfer matrix is

$$\mathsf{T} = \mathsf{T}_{\mathrm{J}} = \mathsf{I}_n - \mathsf{B}_{\mathrm{J}}^{-1}\mathsf{A} = \mathsf{I}_n - \mathsf{D}^{-1}\mathsf{A}, \tag{3}$$

so that

$$\mathsf{T}_{\mathrm{J}} = -\begin{bmatrix} 0 & \frac{a_{1,2}}{a_{1,1}} & \cdots & \frac{a_{1,n}}{a_{1,1}} \\ \frac{a_{2,1}}{a_{2,2}} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{a_{n-1,n}}{a_{n-1,n-1}} \\ \frac{a_{n,1}}{a_{n,n}} & \cdots & \frac{a_{n,n-1}}{a_{n,n}} & 0 \end{bmatrix}.$$

Alternatively, we may write the Jacobi method in component form via

$$x_{i,k+1} = [\mathbf{x}_{k+1}]_i = \frac{f_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j} x_{j,k}}{a_{i,i}},$$

where $x_{i,k} = [\mathbf{x}_k]_i$.

Theorem (convergence)

Let $\mathrm{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ *be strictly diagonally dominant (SDD) of magnitude* $\delta > 0$*, and* $\mathbf{f} \in \mathbb{C}^n$*. Then, the Jacobi iteration method for approximating the solution to* $\mathrm{A}\mathbf{x} = \mathbf{f}$ *is convergent.*

Proof.

Since A is SDD of magnitude $\delta > 0$, it follows that D is invertible, and T_{J}, given in (3), is well–defined. Then

$$\begin{aligned}\|\mathrm{T}_{\mathrm{J}}\|_\infty &= \max_{1\le i\le n} \sum_{j=1}^{n} \left| \delta_{i,j} - \frac{1}{a_{i,i}} a_{i,j} \right| \\ &= \max_{1\le i\le n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left| \frac{a_{i,j}}{a_{i,i}} \right| \\ &= \max_{1\le i\le n} \frac{1}{|a_{i,i}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{i,j}| .\end{aligned}$$

Hence $\|\mathrm{T}_{\mathrm{J}}\|_\infty < 1$. By a previous theorem the method converges. □

Theorem (convergence)

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is column–wise strictly diagonally dominant (SDD) of magnitude* $\delta > 0$, *i.e.,*

$$|a_{j,j}| - \delta \ge \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{i,j}|, \quad j = 1, \ldots, n,$$

and $\mathbf{f} \in \mathbb{C}^n$ *is given. Then the sequence* $\{\mathbf{x}_k\}_{k=1}^{\infty}$ *generated by the Jacobi iteration method converges, for any starting value* $\mathbf{x}_0$, *to the vector* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$.

Proof.

It will suffice to prove that

$$\rho(\mathsf{T}_{\mathrm{J}}) < 1,$$

where T_{J} is given by (3). Since $\mathsf{A} \in \mathbb{C}^{n\times n}$ is column–wise strictly diagonally dominant of magnitude $\delta > 0$, A^{H} is row–wise SDD of magnitude $\delta > 0$. Therefore, by the last theorem,

$$\rho\left(\mathsf{I}_n - \mathsf{D}^{-\mathsf{H}}\mathsf{A}^{\mathsf{H}}\right) \le \left\|\mathsf{I}_n - \mathsf{D}^{-\mathsf{H}}\mathsf{A}^{\mathsf{H}}\right\|_{\infty} < 1.$$

Proof, Cont.

Define $\tilde{\mathsf{T}} = \mathsf{I}_n - \mathsf{D}^{-\mathsf{H}}\mathsf{A}^{\mathsf{H}}$. Then

$$\tilde{\mathsf{T}}^{\mathsf{H}} = \mathsf{I}_n - \mathsf{A}\mathsf{D}^{-1},$$

and

$$\mathsf{D}^{-1}\tilde{\mathsf{T}}^{\mathsf{H}}\mathsf{D} = \mathsf{I}_n - \mathsf{D}^{-1}\mathsf{A} = \mathsf{T}_{\mathrm{J}}.$$

Therefore

$$\sigma(\mathsf{T}_{\mathrm{J}}) = \sigma(\tilde{\mathsf{T}}^{\mathsf{H}}) = \overline{\sigma(\tilde{\mathsf{T}})},$$

and

$$\rho(\mathsf{T}_{\mathrm{J}}) = \rho(\tilde{\mathsf{T}}) < 1. \qquad \square$$

The Gauss-Seidel Method

Recall that Jacobi method can be written in the form

$$\sum_{j=1}^{i-1} a_{i,j} x_{j,k} + a_{i,i} x_{i,k+1} + \sum_{j=i+1}^{n} a_{i,j} x_{j,k} = f_i.$$

However, at this stage, we have already computed new approximations for components x_m, $m = 1, \ldots, i-1$, which we are not using. The Gauss–Seidel method uses these newly computed approximations to obtain the method

$$\sum_{j=1}^{i} a_{i,j} x_{j,k+1} + \sum_{j=i+1}^{n} a_{i,j} x_{j,k} = f_i,$$

As before, we require $a_{i,i} \neq 0$, for all $i = 1, \ldots, n$, so that the method is well–defined.

The Error Transfer Matrix for the Gauss-Seidel Method

Recall the splitting $\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{U}$. Choosing the iterator matrix as

$$\mathsf{B} = \mathsf{B}_{\mathrm{GS}} = \mathsf{L} + \mathsf{D}$$

results in the so-called *Gauss–Seidel method*. The linear iteration process for the Gauss-Seidel method may be expressed as

$$(\mathsf{L} + \mathsf{D})\mathbf{x}_{k+1} + \mathsf{U}\mathbf{x}_k = \mathbf{f}.$$

Therefore, assuming that D is invertible,

$$\mathbf{x}_{k+1} = -(\mathsf{L} + \mathsf{D})^{-1}\mathsf{U}\mathbf{x}_k + (\mathsf{L} + \mathsf{D})^{-1}\mathbf{f}.$$

The error transfer matrix may be expressed as

$$\mathsf{T}_{\mathrm{GS}} = \mathsf{I}_n - \mathsf{B}_{\mathrm{GS}}\mathsf{A} = -(\mathsf{L} + \mathsf{D})^{-1}\mathsf{U} = -(\mathsf{A} - \mathsf{U})^{-1}\mathsf{U}. \qquad (4)$$

Theorem (convergence)

Suppose that $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ *is strictly diagonally dominant (SDD) and* $\mathbf{f} \in \mathbb{C}^n$. *Then for any starting value* $\mathbf{x}_0$, *the sequence generated by the Gauss–Seidel method converges to* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$.

Proof.

Let us define

$$\gamma = \max_{i=1}^{n} \left\{ \frac{\sum_{j=i+1}^{n} |a_{i,j}|}{|a_{i,i}| - \sum_{j=1}^{i-1} |a_{i,j}|} \right\}.$$

Owing to the fact that A is SDD,

$$|a_{i,i}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{i,j}| = \sum_{j=1}^{i-1} |a_{i,j}| + \sum_{j=i+1}^{n} |a_{i,j}| \quad \Longrightarrow \quad |a_{i,i}| - \sum_{j=1}^{i-1} |a_{i,j}| > \sum_{j=i+1}^{n} |a_{i,j}|,$$

and thus $\gamma \in [0,1)$. We will show convergence of the Gauss–Seidel method by proving that $\|\mathsf{T}_{\mathrm{GS}}\|_\infty \leq \gamma$.

Proof, Cont.

Let $\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{U}$ be the usual decomposition into lower triangular, diagonal, and upper triangular parts, respectively, and set $\mathbf{y} = \mathsf{T}_{\mathrm{GS}}\mathbf{x}$, i.e., $(\mathsf{L} + \mathsf{D})\mathbf{y} = -\mathsf{U}\mathbf{x}$. We have

$$\sum_{j=1}^{i-1} a_{i,j} y_j + a_{i,i} y_i = -\sum_{j=i+1}^{n} a_{i,j} x_j, \quad 1 \leq i \leq n.$$

Let i be such that $|y_i| = \|\mathbf{y}\|_\infty$. By the triangle inequality,

$$\left|\sum_{j=1}^{i-1} a_{i,j} y_j + a_{i,i} y_i\right| \geq |a_{i,i}||y_i| - \sum_{j=1}^{i-1} |a_{i,j}||y_j| \geq \left(|a_{i,i}| - \sum_{j=1}^{i-1} |a_{i,j}|\right) \|\mathbf{y}\|_\infty.$$

Also, we have

$$\left|\sum_{j=i+1}^{n} a_{i,j} x_j\right| \leq \sum_{j=i+1}^{n} |a_{i,j}| \|\mathbf{x}\|_\infty.$$

Proof, Cont.

Consequently,

$$\|T_{\mathrm{GS}}\mathbf{x}\|_\infty = \|\mathbf{y}\|_\infty \le \frac{\sum_{j=i+1}^{n} |a_{i,j}|}{|a_{i,i}| - \sum_{j=1}^{i-1} |a_{i,j}|}\|\mathbf{x}\|_\infty \le \gamma\|\mathbf{x}\|_\infty.$$

This implies that, for all $\mathbf{x} \in \mathbb{C}^n_\star$,

$$\frac{\|T_{\mathrm{GS}}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le \gamma,$$

which shows that

$$\|T_{\mathrm{GS}}\|_\infty \le \gamma < 1.$$

□

The Faster Convergence of Gauss-Seidel

Theorem (SDD matrices)

Suppose that $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ *is strictly diagonally dominant (SDD) of magnitude* $\delta > 0$ *and* $\mathbf{f} \in \mathbb{C}^n$. *Then*

$$\|\mathsf{T}_{\mathrm{GS}}\|_\infty \leq \|\mathsf{T}_{\mathrm{J}}\|_\infty < 1,$$

where T_{GS} *and* T_{J} *denote, respectively, the error transfer matrices of the Gauss–Seidel and Jacobi methods. In particular, for any starting value* $\mathbf{x}_0$, *the sequences generated by the Jacobi and the Gauss–Seidel methods both converge to* $\mathbf{x} = \mathsf{A}^{-1}\mathbf{f}$.

Theorem (tridiagonal matrices)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be tridiagonal with non–zero diagonal elements. Denote by* T_{J} *and* T_{GS} *the error transfer matrices of the Jacobi and Gauss–Seidel methods, respectively. Then we have*

$$\rho(\mathsf{T}_{\mathrm{GS}}) = \rho(\mathsf{T}_{\mathrm{J}})^2.$$

In particular, one method converges iff the other method converges.

Richardson's Method

Richardson's Method

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be invertible and $\mathbf{f} \in \mathbb{C}^n$ be given. Let us consider now what is known as *Richardson's method*:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha\left(\mathbf{f} - \mathsf{A}\mathbf{x}_k\right).$$

Clearly, this is a stationary two–layer method that results from choosing

$$\mathsf{B} = \mathsf{B}_{\mathrm{R}} = \frac{1}{\alpha}\mathsf{I}_n,$$

where $\alpha \in \mathbb{C}_\star$. In this case, $\mathsf{T}_{\mathrm{R}} = \mathsf{I}_n - \alpha\mathsf{A}$.

Theorem (convergence)

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *be HPD,* $\sigma(\mathsf{A}) = \{\lambda_1, \ldots, \lambda_n\}$*, with* $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$*. Assume that* $\alpha \in \mathbb{R}_\star$*. Then, Richardson's method converges iff* $\alpha \in (0, 2/\lambda_n)$*. In this case we have the estimate*

$$\|\mathbf{e}_k\|_2 \leq \rho^k \|\mathbf{e}_0\|_2, \quad \rho = \rho(\alpha) = \max\{|1-\alpha\lambda_n|, |1-\alpha\lambda_1|\}.$$

From this, it follows that setting

$$\alpha = \alpha_{\mathrm{opt}} := \frac{2}{\lambda_1 + \lambda_n},$$

one obtains the smallest possible value of ρ*, and*

$$\rho_{\mathrm{opt}} = \rho(\alpha_{\mathrm{opt}}) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\kappa_2(\mathsf{A}) - 1}{\kappa_2(\mathsf{A}) + 1}.$$

Proof of Convergence of Richardson's Method.

Since A is HPD, we know that the eigenvalues of A are positive real numbers and

$$\lambda_n = \|\mathsf{A}\|_2.$$

Notice also that $\mathsf{T}_{\mathrm{R}} = \mathsf{I}_n - \alpha\mathsf{A} = \mathsf{T}_{\mathrm{R}}^{\mathsf{H}}$, which implies that the eigenvalues of T_{R} are real. Observe that $(\lambda_i, \mathbf{w}_i)$ is an eigenpair of A iff $(\nu_i = 1 - \alpha\lambda_i, \mathbf{w}_i)$ is an eigenpair of T_{R}. Assume that $0 < \alpha < 2/\lambda_n$. Then

$$0 < \lambda_i\alpha < 2\frac{\lambda_i}{\lambda_n}, \quad i = 1, \ldots, n,$$

which implies

$$1 > 1 - \lambda_i\alpha > 1 - 2\frac{\lambda_i}{\lambda_n} \geq -1, \quad i = 1, \ldots, n.$$

It follows that

$$1 > \nu_1 \geq \cdots \geq \nu_n > -1, \qquad \nu_i = 1 - \alpha\lambda_i.$$

This guarantees that $\|\mathsf{T}_{\mathrm{R}}\|_2 = \rho(\mathsf{T}_{\mathrm{R}}) < 1$, which implies convergence.

Proof, Cont.

Conversely, if $\alpha \notin (0, 2/\lambda_n)$, then $\rho(\mathsf{T}_{\mathrm{R}}) \geq 1$, and the method does not converge.

By consistency,

$$\|\mathbf{e}_k\|_2 = \|\mathsf{T}_{\mathrm{R}}^k \mathbf{e}_0\|_2 \leq \rho^k \|\mathbf{e}_0\|_2.$$

Of course, it is easy to see that

$$\rho = \rho(\mathsf{T}_{\mathrm{R}}) = \max\{|\nu_1|, |\nu_n|\} = \max\{|1 - \alpha\lambda_n|, |1 - \alpha\lambda_1|\}.$$

Finally, showing optimality amounts to minimizing ρ. See the figure on the next slide. From this we see that the minimum of ρ is attained when

$$|1 - \alpha\lambda_1| = |1 - \alpha\lambda_n|$$

or

$$1 - \alpha\lambda_n = \alpha\lambda_1 - 1$$

which implies

$$\alpha_{\mathrm{opt}} = \frac{2}{\lambda_1 + \lambda_n}.$$

$\square$

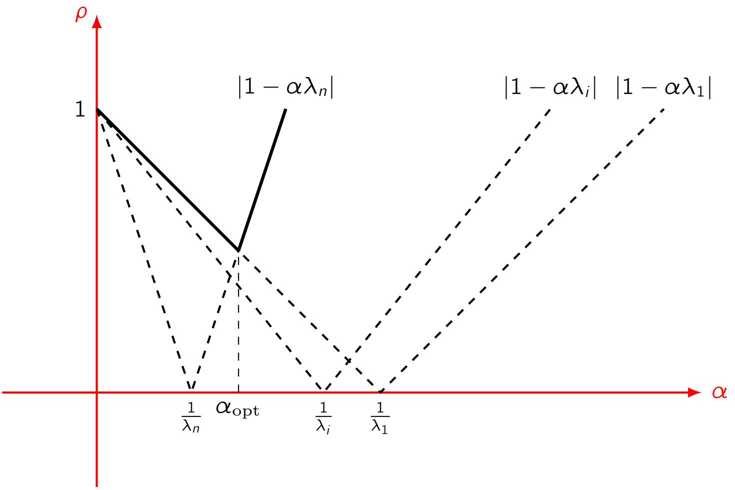

Figure: The curve $\rho(\mathsf{T}_{\mathrm{R}})$ (in solid black) as a function of α.