Math 673

Multigrid Methods: A Mostly Matrix-Based Approach

Chapter 01: Classical Iterative Methods

Abner J. Salgado and Steven M. Wise

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

Fall 2024

Chapter 01, Part 2 of 2
Classical Iterative Methods

Splitting Methods

Classical Splitting Methods

In this section we define a couple of the classical iterative methods that one meets in an elementary course on numerical analysis, namely, the Jacobi, Gauss-Seidel, and Richardson's methods. We analyze them using the spectral theory introduced in the last section.

The Jacobi and Gauss-Seidel methods are based on the canonical (diagonal) splitting of A.

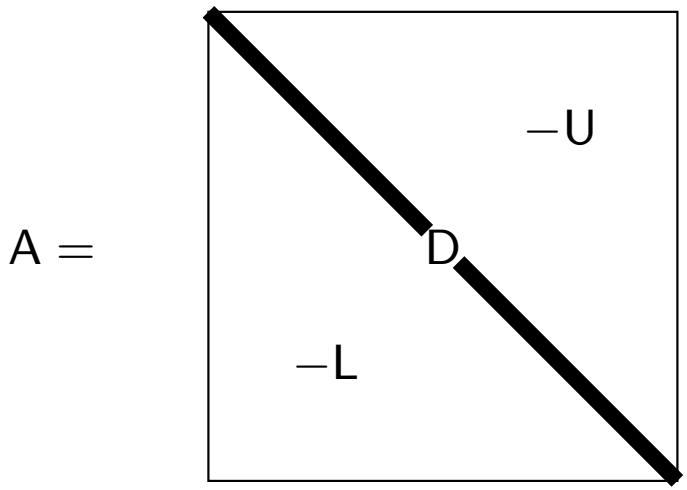

Figure: The canonical splitting of A.

Definition

Let $A \in \mathbb{R}^{n \times n}$ be given. Consider

$$A = D - L - U, \tag{1}$$

where $D \in \mathbb{R}^{n \times n}$ is the diagonal of A, $-L \in \mathbb{R}^{n \times n}$ is the strictly lower triangular part of A, and $-U \in \mathbb{R}^{n \times n}$ is the strictly upper triangular part of A, respectively, as illustrated in the figure on the previous slide. The decomposition (1) is called the **canonical splitting of** A or **diagonal splitting of** A.

Canonical/Diagonal Splitting of A

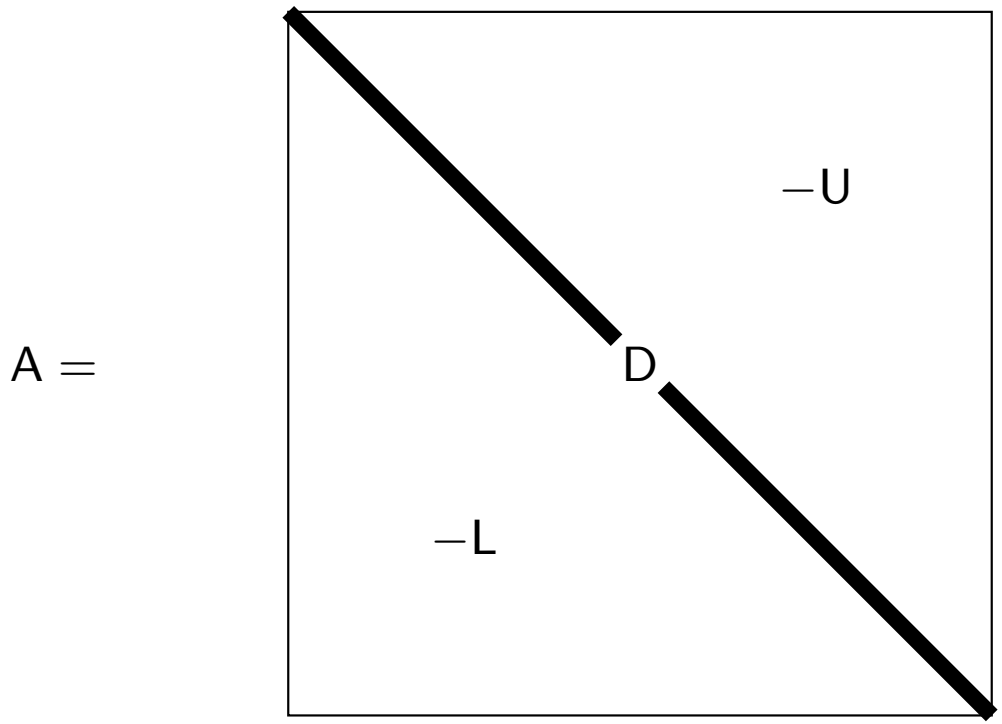

Figure: The canonical splitting of A.

Example

Suppose that

$$A = \begin{bmatrix} 10 & -2 & -10 & 5 & -10 \\ 10 & 9 & 7 & -2 & -5 \\ 0 & 6 & 9 & 3 & -10 \\ 6 & 10 & 4 & -7 & -8 \\ -8 & 3 & 5 & 4 & 7 \end{bmatrix}.$$

Then

$$D = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & -7 & 0 \\ 0 & 0 & 0 & 0 & 7 \end{bmatrix},$$

$$U = \begin{bmatrix} 0 & 2 & 10 & -5 & 10 \\ 0 & 0 & -7 & 2 & 5 \\ 0 & 0 & 0 & -3 & 10 \\ 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad L = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -10 & 0 & 0 & 0 & 0 \\ 0 & -6 & 0 & 0 & 0 \\ -6 & -10 & -4 & 0 & 0 \\ 8 & -3 & -5 & -4 & 0 \end{bmatrix}.$$

Proposition

If A *is SPD, then* D *has only positive diagonal elements, and* $\mathsf{U} = \mathsf{L}^\top$.

Proof.

Exercise. □

The Jacobi Iterative Method

Suppose that $A = D - L - U$ is the canonical splitting of A. Assume D is invertible. Then the linear system $A\boldsymbol{u} = \boldsymbol{f}$ can be rewritten as

$$D\boldsymbol{u} = (U + L)\boldsymbol{u} + \boldsymbol{f}.$$

Jacobi's method results from the iteration

$$D\boldsymbol{u}^{k+1} = (U + L)\boldsymbol{u}^{k} + \boldsymbol{f}.$$

Put another way, we have the following:

Definition (The Jacobi Method)

Suppose that $A = D - L - U$ is the canonical splitting of $A \in \mathbb{R}^{n \times n}$, where the diagonal matrix D is invertible. **The Jacobi method** is a GLIS with the iterator matrix

$$B = B_J := D^{-1}$$

and error transfer matrix

$$T = T_J := I - D^{-1}A.$$

The iteration sequence is precisely

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + D^{-1}\left(\boldsymbol{f} - A\boldsymbol{u}^k\right). \tag{2}$$

Remark

The error transfer matrix for the Jacobi method can be expressed as

$$\begin{aligned} T_J &= D^{-1}D - D^{-1}A \\ &= D^{-1}(D - D + U + L) \\ &= D^{-1}(U + L). \end{aligned} \tag{3}$$

Remark

In component form, the Jacobi method can be written as

$$u_i^{k+1} = \frac{f_i - \sum_{j=1}^{i-1} a_{i,j} u_j^k - \sum_{j=i+1}^{n} a_{i,j} u_j^k}{a_{i,i}}. \tag{4}$$

The order in which we obtain the updated components is completely immaterial.

Definition

We say that a matrix $A \in \mathbb{R}^{n \times n}$ is **diagonally dominant** or **row-wise diagonally dominant** iff, for each $i \in \{1, \ldots, n\}$,

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| < |a_{i,i}|.$$

Proposition

If $A \in \mathbb{R}^{n \times n}$ is diagonally dominant, then it is invertible and none of its diagonal elements is zero.

Proof.

Exercise. □

Theorem (Convergence of the Jacobi method)

If $A \in \mathbb{R}^{n\times n}$ is diagonally dominant, then the Jacobi method is well defined and unconditionally convergent.

Proof.

We will show that $\|T_J\|_\infty < 1$. Since A is diagonally dominant, its diagonal elements are all non-zero. Thus the method is well defined because D is invertible. Now, observe that

$$T_J = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{a_{1,1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{n,n}} \end{bmatrix} \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix}.$$

Thus

$$t_{i,j} = \begin{cases} 0, & \text{if} \quad i = j, \\ -\frac{a_{i,j}}{a_{i,i}}, & \text{if} \quad i \neq j. \end{cases}.$$

Proof (Cont.)

So,

$$\|T_J\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |t_{i,j}| = \max_{1\le i\le n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{i,j}}{a_{i,i}}\right|.$$

Since A is diagonally dominant, there is some $\delta > 0$, such that

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

for each $i = 1, \ldots, n$.

Proof (Cont.)

Therefore, for each $i = 1, \ldots, n$,

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{i,j}}{a_{i,i}} \right| \leq 1 - \frac{\delta}{|a_{i,i}|} \leq 1 - \frac{\delta}{\alpha},$$

where

$$\alpha = \max_{1 \leq i \leq n} |a_{i,i}|.$$

Hence,

$$\|T_J\|_\infty \leq 1 - \frac{\delta}{\alpha} < 1.$$

Definition (Damped Jacobi Method)

Suppose that $A = D - L - U$ is the canonical splitting of $A \in \mathbb{R}^{n\times n}$ and $\omega \in (0, 1]$. Assume D is invertible. The **damped Jacobi method**, or **weighted Jacobi method**, is defined by the iteration

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{u}^k + D^{-1}(\boldsymbol{f} - A\boldsymbol{u}^k) \\ \boldsymbol{u}^{k+1} &= \omega \boldsymbol{z} + (1-\omega)\boldsymbol{u}^k. \end{aligned}$$

where the starting value $\boldsymbol{u}^0 \in \mathbb{R}^n$ is given.

Thus the damped Jacobi method can be viewed as doing one step of the Jacobi method, followed by a damping, or weighting, step. We can eliminate $\boldsymbol{z}$ in the second equation to obtain the following.

Proposition

With the same assumptions as in the last definition, the damped Jacobi method is a GLIS with iterator

$$\mathsf{B} = \mathsf{B}_{J,\omega} := \omega \mathsf{D}^{-1}$$

and error transfer matrix

$$\mathsf{T} = \mathsf{T}_{J,\omega} := \mathsf{I} - \omega \mathsf{D}^{-1}\mathsf{A}.$$

The iteration sequence can be expressed as

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^{k} + \omega \mathsf{D}^{-1}\left(\boldsymbol{f} - \mathsf{A}\boldsymbol{u}^{k}\right). \tag{5}$$

Proof.

Exercise. □

We will come back to the damped Jacobi method in a later chapter.

The Gauss-Seidel Method

Suppose that $A = D - L - U$ is the canonical splitting of $A \in \mathbb{R}^{n \times n}$. Then

$$(D - L)\boldsymbol{u} = U\boldsymbol{u} + \boldsymbol{f}.$$

The Gauss-Seidel method is simply stated as

$$(D - L)\boldsymbol{u}^{k+1} = U\boldsymbol{u}^{k} + \boldsymbol{f}.$$

Assuming $D - L$ is invertible, some manipulations give

$$\begin{aligned} \boldsymbol{u}^{k+1} &= (D - L)^{-1}U\boldsymbol{u}^{k} + (D - L)^{-1}\boldsymbol{f} \\ &= \boldsymbol{u}^{k} - (D - L)^{-1}(D - L)\boldsymbol{u}^{k} + (D - L)^{-1}U\boldsymbol{u}^{k} + (D - L)^{-1}\boldsymbol{f} \\ &= \boldsymbol{u}^{k} + (D - L)^{-1}(\boldsymbol{f} - A\boldsymbol{u}^{k}). \end{aligned}$$

Definition (Forward Gauss-Seidel Method)

Suppose that $A = D - L - U$ is the canonical splitting of $A \in \mathbb{R}^{n\times n}$ and $D - L$ is invertible. The **Gauss-Seidel method** – also called the **forward Gauss-Seidel method** – is a GLIS with iterator matrix

$$B = B_{GS} := (D - L)^{-1}$$

and error transfer matrix

$$T = T_{GS} := I - (D - L)^{-1}A.$$

The iteration sequence can be expressed as

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + (D - L)^{-1}(\boldsymbol{f} - A\boldsymbol{u}^k). \tag{6}$$

Remark

The error transfer matrix for the Gauss-Seidel method can be expressed equivalently as

$$\begin{aligned} T_{GS} &= I - (D - L)^{-1}A \\ &= (D - L)^{-1}(D - L) - (D - L)^{-1}(D - L - U) \\ &= (D - L)^{-1}U. \end{aligned} \tag{7}$$

The Component Form of Gauss-Seidel

The Gauss-Seidel method is simply stated as

$$(\mathrm{D} - \mathrm{L})\boldsymbol{u}^{k+1} - \mathrm{U}\boldsymbol{u}^{k} = \boldsymbol{f}.$$

In component form, we can express the forward Gauss-Seidel method as

$$\sum_{j=1}^{i} a_{i,j} u_j^{k+1} + \sum_{j=i+1}^{n} a_{i,j} u_j^{k} = f_i, \quad i = 1, \ldots, n.$$

The Component Form of Gauss-Seidel

To compute the update $\boldsymbol{u}^{k+1}$, we proceed from $i = 1$, in order, to $i = n$. So, the first equation we solve is

$$u_1^{k+1} = \frac{f_1 - \sum_{j=2}^{n} a_{1,j} u_j^k}{a_{1,1}}.$$

Next, we have

$$u_2^{k+1} = \frac{f_2 - a_{2,1} u_1^{k+1} - \sum_{j=3}^{n} a_{2,j} u_j^k}{a_{2,2}},$$

and

$$u_3^{k+1} = \frac{f_3 - \sum_{j=1}^{2} a_{3,j} u_j^{k+1} - \sum_{j=4}^{n} a_{3,j} u_j^k}{a_{3,3}}.$$

We continue, in the same way, with $i = 4$, in order, to $i = n - 1$, and we end with

$$u_n^{k+1} = \frac{f_n - \sum_{j=1}^{n-1} a_{n,j} u_j^{k+1}}{a_{n,n}}.$$

The Component Form of Gauss-Seidel

The idea with the method is simple. Once we obtain an updated component u_i^{k+1}, we immediately use it to obtain u_{i+1}^{k+1}. One full pass, from $i = 1$ to $i = n$ is called a **forward sweep**. The generic i^{th} step is

$$u_i^{k+1} = \frac{f_i - \sum_{j=1}^{i-1} a_{i,j} u_j^{k+1} - \sum_{j=i+1}^{n} a_{i,j} u_j^k}{a_{i,i}}. \tag{8}$$

Theorem (Convergence of the forward Gauss-Seidel method)

Suppose that $A = D - L - U$ *is the canonical splitting of* $A \in \mathbb{R}^{n \times n}$. *If* A *is diagonally dominant, then the forward Gauss-Seidel method is well defined, that is,* $D - L$ *is invertible, and the method is unconditionally convergent.*

Proof.

If A is diagonally dominant, then D has only nonzero diagonal entries. Hence, $D - L$ is invertible.

For convergence, we again want to show that, if A is diagonally dominant, then $\|T_{GS}\|_\infty < 1$. This will establish the unconditional convergence of the method. Since A is diagonally dominant, it follows that, for some $\delta > 0$,

$$|a_{i,i}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| + \delta = \sum_{j>i} |a_{i,j}| + \sum_{j<i} |a_{i,j}| + \delta.$$

Proof (Cont.)

Thus

$$|a_{i,i}| - \sum_{j<i} |a_{i,j}| \geq \sum_{j>i} |a_{i,j}| + \delta > \sum_{j>i} |a_{i,j}|,$$

which implies

$$\gamma := \max_{1 \leq i \leq n} \left\{ \frac{\sum_{j>i} |a_{i,j}|}{|a_{i,i}| - \sum_{j<i} |a_{i,j}|} \right\} < 1.$$

Now, we will show $\|\mathsf{T}_{GS}\|_\infty \leq \gamma$.

Let $\boldsymbol{x} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{y} = \mathsf{T}_{GS}\boldsymbol{x}$, that is,

$$\boldsymbol{y} = \mathsf{T}_{GS}\boldsymbol{x} = (\mathsf{D} - \mathsf{L})^{-1}\mathsf{U}\boldsymbol{x}.$$

Let k be the index such that $\|\boldsymbol{y}\|_\infty = |y_k|$. Then we have

$$|[(\mathsf{D} - \mathsf{L})\boldsymbol{y}]_k| = |[\mathsf{U}\boldsymbol{x}]_k| = |\sum_{j>k} a_{k,j}x_j| \leq \sum_{j>k} |a_{k,j}||x_j| \leq \sum_{j>k} |a_{k,j}| \, \|\boldsymbol{x}\|_\infty .$$

Proof (Cont.)

Now notice that

$$\begin{aligned}
|[(\mathsf{D}-\mathsf{L})\boldsymbol{y}]_k| &= |\sum_{j<k} a_{k,j}y_j + a_{k,k}y_j| \\
&\geq |a_{k,k}y_k| - |\sum_{j<k} a_{k,j}y_j| \\
&= |a_{k,k}|\,\|\boldsymbol{y}\|_\infty - |\sum_{j<k} a_{k,j}y_j| \\
&\geq |a_{k,k}|\,\|\boldsymbol{y}\|_\infty - \sum_{j<k} |a_{k,j}|\,\|\boldsymbol{y}\|_\infty .
\end{aligned}$$

Proof (Cont.)

Therefore, we have

$$|a_{k,k}| \, \|\boldsymbol{y}\|_\infty - \sum_{j<k} |a_{k,j}| \, \|\boldsymbol{y}\|_\infty \leq \sum_{j>k} |a_{k,j}| \, \|\boldsymbol{x}\|_\infty ,$$

which implies

$$\|\boldsymbol{y}\|_\infty \leq \frac{\sum_{j>k} |a_{k,j}|}{|a_{k,k}| - \sum_{j<k} |a_{k,j}|} \, \|\boldsymbol{x}\|_\infty .$$

So,

$$\|\mathsf{T}_{GS}\boldsymbol{x}\|_\infty \leq \gamma \, \|\boldsymbol{x}\|_\infty ,$$

which implies

$$\|\mathsf{T}_{GS}\|_\infty \leq \gamma < 1.$$

Definition (Backward Gauss-Seidel method)

Suppose that $A \in \mathbb{R}^{n\times n}$ is invertible and $A = D - L - U$ is the canonical splitting of A. Assume $D - U$ is invertible. The **backward Gauss-Seidel method** is a GLIS with iterator matrix

$$B = B_{BGS} := (D - U)^{-1}$$

and error transfer matrix

$$T = T_{BGS} := I - (D - U)^{-1}A.$$

The iteration scheme is precisely

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + (D - U)^{-1}(\boldsymbol{f} - A\boldsymbol{u}^k). \tag{9}$$

Remark

The error transfer matrix for the backward Gauss-Seidel method can be re-expressed as

$$\begin{aligned} \mathrm{T}_{BGS} &= I - (\mathrm{D} - \mathrm{U})^{-1}\mathrm{A} \\ &= \mathrm{I} - (\mathrm{D} - \mathrm{U})^{-1}(\mathrm{D} - \mathrm{L} - \mathrm{U}) \\ &= (\mathrm{D} - \mathrm{U})^{-1}\mathrm{L}. \end{aligned} \tag{10}$$

It should not be hard to modify the proof of the last theorem to obtain the following:

Theorem (Convergence of the backward Gauss-Seidel method)

Suppose that $A = D - L - U$ *is the canonical splitting of* $A \in \mathbb{R}^{n \times n}$. *If* A *is diagonally dominant, then the backward Gauss-Seidel method is well defined, that is,* $D - U$ *is invertible, and it is unconditionally convergent.*

Component Form of Backward Gauss-Seidel

The Gauss-Seidel method is simply stated as

$$-\mathrm{L}\boldsymbol{u}^k + (\mathrm{D} - \mathrm{U})\boldsymbol{u}^{k+1} = \boldsymbol{f}.$$

In component form, we can express the backward Gauss-Seidel method as

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

Component Form of Backward Gauss-Seidel

To compute the update $\boldsymbol{u}^{k+1}$, we proceed from $i = n$, in reverse order, to $i = 1$. We find the n^{th} component first:

$$u_n^{k+1} = \frac{f_n - \sum_{j=1}^{n-1} a_{n,j} u_j^k}{a_{n,n}}.$$

Next, we have

$$u_{n-1}^{k+1} = \frac{f_{n-1} - \sum_{j=1}^{n-2} a_{n-1,j} u_j^k - a_{n-1,n} u_n^{k+1}}{a_{n-1,n-1}},$$

and then

$$u_{n-2}^{k+1} = \frac{f_{n-2} - \sum_{j=1}^{n-3} a_{n-2,j} u_j^k - \sum_{j=n-1}^{n} a_{n-2,j} u_j^{k+1}}{a_{n-2,n-2}}.$$

We continue, in the same way, with $i = n - 3$, in reverse order, to $i = 2$, and we end with

$$u_1^{k+1} = \frac{f_1 - \sum_{j=1}^{n-1} a_{1,j} u_j^{k+1}}{a_{1,1}}.$$

Component Form of Backward Gauss-Seidel

One full pass, from $i = n$, proceeding in reverse order, to $i = 1$ is called a **backward sweep**. The generic i^{th} step is

$$u_i^{k+1} = \frac{f_i - \sum_{j=1}^{i-1} a_{i,j} u_j^k - \sum_{j=i+1}^{n} a_{i,j} u_j^{k+1}}{a_{i,i}}. \tag{11}$$

Richardson's Method

Suppose $\omega \in \mathbb{R}_\star := \mathbb{R} \setminus \{0\}$, and consider the splitting

$$\mathsf{A} = \omega \mathsf{I} + \mathsf{A} - \omega \mathsf{I}.$$

Suppose

$$\mathsf{A}\boldsymbol{u} = \boldsymbol{f}.$$

Then,

$$\omega \boldsymbol{u} = (\omega \mathsf{I} - \mathsf{A})\boldsymbol{u} + \boldsymbol{f}.$$

Richardson's method is, essentially,

$$\omega \boldsymbol{u}^{k+1} = (\omega \mathsf{I} - \mathsf{A})\boldsymbol{u}^k + \boldsymbol{f}.$$

Notice that this method is not based on the canonical diagonal splitting, but it is still quite useful and interesting.

Definition (Richardson's Method)

Suppose that $\mathsf{A} \in \mathbb{R}^{n\times n}$ is invertible and $\omega \in \mathbb{R}_\star$. **Richardson's method** is a GLIS with iterator matrix

$$\mathsf{B} = \mathsf{B}_R := \omega^{-1}\mathsf{I}$$

and error transfer matrix

$$\mathsf{T} = \mathsf{T}_R := \mathsf{I} - \omega^{-1}\mathsf{A}.$$

The iteration scheme is

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \omega^{-1}(\boldsymbol{f} - \mathsf{A}\boldsymbol{u}^k). \tag{12}$$

We make a slight departure from the now familiar approach and analyze this method using the 2-norm, $\|\cdot\|_2$, under the assumption that A is SPD.

Theorem (Convergence of Richardson's Method)

Let $\mathrm{A} \in \mathbb{R}^{n\times n}$ *be SPD and* $\sigma(\mathrm{A}) = \{\lambda_1, \ldots, \lambda_n\}$, *with* $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. *Richardson's method converges unconditionally iff* $\omega \in (0, 2/\lambda_n)$. *In this case, we have the estimate*

$$\|\boldsymbol{e}^k\|_2 \leq \rho^k \|\boldsymbol{e}^0\|_2, \quad \rho = \rho(\omega) = \max\{|1-\omega\lambda_n|, |1-\omega\lambda_1|\}.$$

From this, it follows that setting

$$\omega = \omega_{\mathrm{opt}} := \frac{2}{\lambda_1 + \lambda_n},$$

one obtains the smallest possible value of ρ, ρ_{opt}, *and*

$$\rho_{\mathrm{opt}} := \rho(\omega_{\mathrm{opt}}) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\kappa_2(\mathrm{A}) - 1}{\kappa_2(\mathrm{A}) + 1},$$

where

$$\kappa_2(\mathrm{A}) = \frac{\lambda_n}{\lambda_1}.$$

Proof of Convergence of Richardson's Method.

($\Leftarrow$): Since A is SPD, we know that the eigenvalues of A are positive real numbers and

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

Notice also that $\mathsf{T}_R = \mathsf{I}_n - \omega\mathsf{A} = \mathsf{T}_R^\top$, which implies that the eigenvalues of T_R are real. Observe that $(\lambda_i, \boldsymbol{w}_i)$ is an eigenpair of A iff $(\nu_i = 1 - \omega\lambda_i, \boldsymbol{w}_i)$ is an eigenpair of T_R. Assume that $0 < \omega < 2/\lambda_n$. Then

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

which implies

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

It follows that

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

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

Proof (Cont.)

($\Rightarrow$): Conversely, if $\omega \notin (0, 2/\lambda_n)$, then $\rho(\mathsf{T}_R) \geq 1$, and the method cannot converge unconditionally.

(Error estimate): By consistency,

$$\|\boldsymbol{e}^k\|_2 = \|\mathsf{T}_R^k \boldsymbol{e}^0\|_2 \leq \rho^k \|\boldsymbol{e}^0\|_2.$$

Of course, it is easy to see that

$$\rho = \rho(\mathsf{T}_R) = \max\{|\nu_1|, |\nu_n|\} = \max\{|1 - \omega\lambda_n|, |1 - \omega\lambda_1|\}.$$

Proof (Cont.)

(Optimality): 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

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

or

$$1 - \omega\lambda_n = \omega\lambda_1 - 1,$$

which implies

$$\omega_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}.$$

Therefore

$$\rho_{\text{opt}} = 1 - \omega_{\text{opt}}\lambda_1 = \frac{\lambda_1 + \lambda_n - 2\lambda_1}{\lambda_1 + \lambda_n} = \frac{\lambda_n - \lambda_1}{\lambda_1 + \lambda_n} = \frac{\kappa_2(\mathsf{A}) - 1}{\kappa_2(\mathsf{A}) + 1}.$$

□

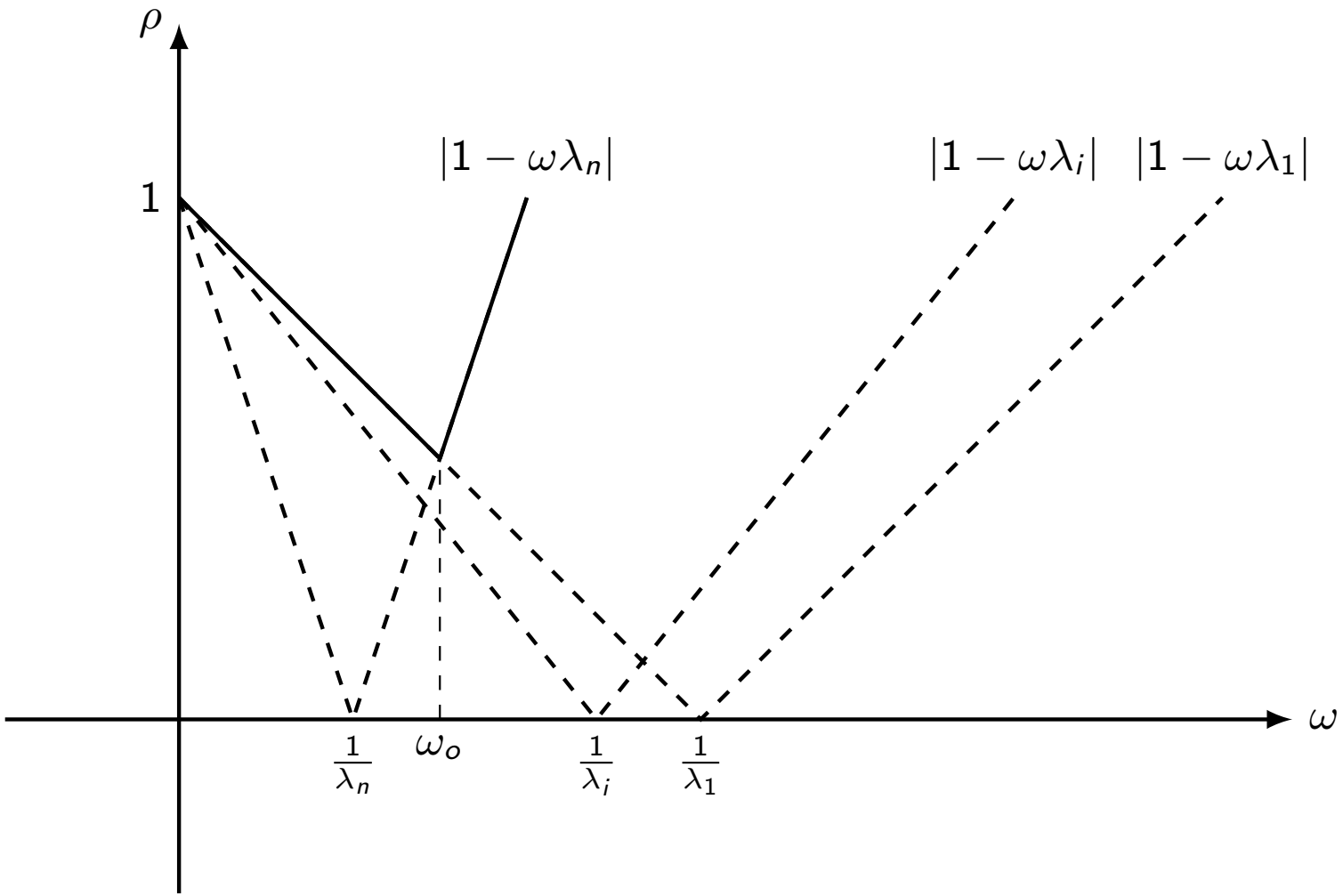

Figure: The curve $\rho(T_R)$ (in solid black) as a function of ω.

Additive, Multiplicative, and Symmetrized GLIS

Making New Methods by Combination

Recall that the iterator matrix defines the GLIS. If one has a collection of iterator matrices available, the analyst can combine them in various ways to create a new GLIS. There are three main flavors:

1. Additive combination.
2. Multiplicative combination.
3. Symmetrization.

Of course, we can make combinations of combinations...

Definition (Additive and Multiplicative GLIS)

Suppose $A \in \mathbb{R}^{n\times n}$ is invertible and $\boldsymbol{f} \in \mathbb{R}^n$ is given. Consider a family of iterator matrices, $B_i \in \mathbb{R}^{n\times n}$, for $i = 1, \ldots, r$. An iterative scheme is called **additive with respect to** $\{B_i\}_{i=1}^r$ iff

$$\begin{aligned}
\boldsymbol{z}_1 &= \boldsymbol{u}^k + B_1\left(\boldsymbol{f} - A\boldsymbol{u}^k\right),\\
\boldsymbol{z}_2 &= \boldsymbol{z}_1 + B_2\left(\boldsymbol{f} - A\boldsymbol{u}^k\right),\\
\boldsymbol{z}_3 &= \boldsymbol{z}_2 + B_3\left(\boldsymbol{f} - A\boldsymbol{u}^k\right),\\
&\vdots\\
\boldsymbol{z}_{r-1} &= \boldsymbol{z}_{r-2} + B_{r-1}\left(\boldsymbol{f} - A\boldsymbol{u}^k\right),\\
\boldsymbol{u}^{k+1} &= \boldsymbol{z}_{r-1} + B_r\left(\boldsymbol{f} - A\boldsymbol{u}^k\right).
\end{aligned} \tag{13}$$

Definition

An iterative scheme is called **multiplicative with respect to** $\{\mathrm{B}_i\}_{i=1}^{r}$ iff

$$
\begin{aligned}
\boldsymbol{z}_1 &= \boldsymbol{u}^k + \mathrm{B}_1\left(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k\right), \\
\boldsymbol{z}_2 &= \boldsymbol{z}_1 + \mathrm{B}_2\left(\boldsymbol{f} - \mathrm{A}\boldsymbol{z}_1\right), \\
\boldsymbol{z}_3 &= \boldsymbol{z}_2 + \mathrm{B}_3\left(\boldsymbol{f} - \mathrm{A}\boldsymbol{z}_2\right), \\
&\vdots \\
\boldsymbol{z}_{r-1} &= \boldsymbol{z}_{r-2} + \mathrm{B}_{r-1}\left(\boldsymbol{f} - \mathrm{A}\boldsymbol{z}_{r-2}\right), \\
\boldsymbol{u}^{k+1} &= \boldsymbol{z}_{r-1} + \mathrm{B}_r\left(\boldsymbol{f} - \mathrm{A}\boldsymbol{z}_{r-1}\right).
\end{aligned}
\tag{14}
$$

Proposition

Suppose $\mathrm{A} \in \mathbb{R}^{n \times n}$ *is invertible and* $\boldsymbol{f} \in \mathbb{R}^n$ *is given. Consider a family of iterator matrices,* $\mathrm{B}_i \in \mathbb{R}^{n \times n}$*, for* $i = 1, \ldots, r$*. An additive iterative scheme with respect to* $\{\mathrm{B}_i\}_{i=1}^r$ *is a GLIS with iterator*

$$\mathrm{B} = \sum_{i=1}^{r} \mathrm{B}_i.$$

A multiplicative iterative scheme with respect to $\{\mathrm{B}_i\}_{i=1}^r$ *is a GLIS with the following recursively-defined iterator:*

$$\mathrm{B} = \tilde{\mathrm{B}}_r,$$

where $\tilde{\mathrm{B}}_1 = \mathrm{B}_1$*, and, for* $2 \leq i \leq r$*,*

$$\tilde{\mathrm{B}}_i = \tilde{\mathrm{B}}_{i-1} + \mathrm{B}_i - \mathrm{B}_i \mathrm{A} \tilde{\mathrm{B}}_{i-1}.$$

Proof.

Exercise. □

Proposition

Suppose $A \in \mathbb{R}^{n\times n}$ *is invertible,* $\boldsymbol{f} \in \mathbb{R}^n$ *is given, and* $\{B_i\}_{i=1}^r \subset \mathbb{R}^{n\times n}$ *is a family of iterator matrices. The error transfer matrix for the additive GLIS with respect to* $\{B_i\}_{i=1}^r$ *can be expressed as*

$$T = I - \left(\sum_{i=1}^{r} B_i\right) A.$$

The error transfer matrix for the multiplicative GLIS with respect to $\{B_i\}_{i=1}^r$ *can be expressed as*

$$T = \prod_{i=1}^{r} (I - B_i A).$$

Proof.

Exercise. □

Definition (Symmetrized Multiplicative GLIS)

Suppose $A \in \mathbb{R}^{n\times n}$ is invertible and $\boldsymbol{f} \in \mathbb{R}^n$ is given. Consider a generic GLIS with iterator $B \in \mathbb{R}^{n\times n}$. The **symmetrized multiplicative GLIS (SMGLIS) with respect to** B is defined as follows: given $\boldsymbol{u}^0$, find $\boldsymbol{u}^1, \boldsymbol{u}^2, \ldots$ via

$$\boldsymbol{z} = \boldsymbol{u}^k + B(\boldsymbol{f} - A\boldsymbol{u}^k), \tag{15}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{z} + B^\top(\boldsymbol{f} - A\boldsymbol{z}). \tag{16}$$

Lemma

Suppose $\mathrm{A} \in \mathbb{R}^{n \times n}$ *is invertible and* $\boldsymbol{f} \in \mathbb{R}^n$ *is given. Assume that* B *is invertible. The SMGLIS with respect to* B *can be written as*

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \mathrm{B}_{SM}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k), \tag{17}$$

where

$$\mathrm{B}_{SM} = \mathrm{B} + \mathrm{B}^\top - \mathrm{B}^\top \mathrm{A}\mathrm{B}. \tag{18}$$

In other words, the SMGLIS with respect to B *is a GLIS with iterator* B_{SM}*. If* B *is invertible, then the iterator may be expressed as*

$$\mathrm{B}_{SM} = \mathrm{B}^\top(\mathrm{B}^{-\top} + \mathrm{B}^{-1} - \mathrm{A})\mathrm{B}. \tag{19}$$

If A *is symmetric, then* B_{SM} *is as well.*

Proof.

Plugging (15) into (16) yields

$$\begin{aligned}
\boldsymbol{u}^{k+1} &= \boldsymbol{u}^k + \mathrm{B}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k) + \mathrm{B}^\top \left(\boldsymbol{f} - \mathrm{A}\left[\boldsymbol{u}^k + \mathrm{B}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k)\right]\right) \\
&= \boldsymbol{u}^k + \mathrm{B}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k) + \mathrm{B}^\top \left(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k - \mathrm{AB}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k)\right) \\
&= \boldsymbol{u}^k + \mathrm{B}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k) + \mathrm{B}^\top \left(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k\right) - \mathrm{B}^\top \mathrm{AB}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k) \\
&= \boldsymbol{u}^k + \left(\mathrm{B} + \mathrm{B}^\top - \mathrm{B}^\top \mathrm{AB}\right)(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k),
\end{aligned}$$

which shows the SMGLIS with respect to B is a GLIS with iterator

$$\mathrm{B} = \mathrm{B}_{SM} = \mathrm{B} + \mathrm{B}^\top - \mathrm{B}^\top \mathrm{AB}.$$

Proof (Cont.)

If B is invertible, then

$$\begin{aligned} \mathrm{B}_{SM} &= \mathrm{B} + \mathrm{B}^\top - \mathrm{B}^\top \mathrm{A}\mathrm{B} \\ &= \left(I + \mathrm{B}^\top \mathrm{B}^{-1} - \mathrm{B}^\top \mathrm{A} \right) \mathrm{B} \\ &= \left(\mathrm{B}^\top B^{-\top} + \mathrm{B}^\top \mathrm{B}^{-1} - \mathrm{B}^\top \mathrm{A} \right) \mathrm{B} \\ &= \mathrm{B}^\top (\mathrm{B}^{-\top} + \mathrm{B}^{-1} - \mathrm{A}) \mathrm{B}. \end{aligned}$$

If A is symmetric, then B_{SM} is symmetric, since

$$\begin{aligned} \mathrm{B}_{SM}^\top &= \mathrm{B}^\top + \mathrm{B} - \mathrm{B}^\top \mathrm{A}^\top \mathrm{B} \\ &= \mathrm{B}^\top + \mathrm{B} - \mathrm{B}^\top \mathrm{A}\mathrm{B} \\ &= \mathrm{B}_{SM}. \end{aligned}$$

Remark

We could also consider a symmetrized additive method:

$$\boldsymbol{z} = \boldsymbol{u}^k + \mathrm{B}(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k), \tag{20}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{z} + \mathrm{B}^\top(\boldsymbol{f} - \mathrm{A}\boldsymbol{u}^k). \tag{21}$$

This is a GLIS with iterator matrix

$$\mathrm{B} = \mathrm{B}_{SA} := \mathrm{B} + \mathrm{B}^\top.$$

But this, it turns out, is not as useful to us in the multigrid setting.

The forward Gauss-Seidel method is the best example of a GLIS where the iterator is non-symmetric, even when A is symmetric. However, we can use the methodology presented in the last section to symmetrize it.

Definition (The Symmetric Gauss-Seidel Method)

Suppose that $A \in \mathbb{R}^{n\times n}$ is invertible and $A = D - L - U$ is the canonical splitting of A. Assume $D - L$ is invertible, and consider the Gauss-Seidel method, which is characterized by the iterator

$$B_{GS} = (D - L)^{-1}.$$

The **symmetric Gauss Seidel method** is the SMGLIS with respect to B_{GS} and is, therefore, a GLIS with iterator matrix

$$B_{SGS} = B_{GS} + B_{GS}^{\top} - B_{GS}^{\top} A B_{GS}. \tag{22}$$

Proposition

Suppose that $A \in \mathbb{R}^{n\times n}$ *is symmetric and invertible and* $A = D - L - U$ *is the canonical splitting of* A. *Assume that* $D - L$ *is invertible. The iterator for the symmetric Gauss Seidel method can be written as*

$$B_{SGS} = (D - U)^{-1} D (D - L)^{-1}, \tag{23}$$

and it is symmetric.

Proof.

Using Equation (18), we find

$$\begin{aligned} B_{SGS} &= B_{GS}^{\top}(B_{GS}^{-\top} + B_{GS}^{-1} - A)B_{GS} \\ &= (D^{\top} - L^{\top})^{-1}\left(D^{\top} - L^{\top} + D - L - A\right)(D - L)^{-1} \\ &= (D - U)^{-1}(D - U + D - L - A)(D - L)^{-1} \\ &= (D - U)^{-1} D (D - L)^{-1}. \end{aligned}$$

□

Convergence in the Energy Norm

If the coefficient matrix A is SPD, we can use it to construct a useful norm, one which will be the basis of measuring convergence in a new way.

Definition (Energy Norm)

Suppose that A is SPD. The **energy norm** associated with A is

$$\|\boldsymbol{u}\|_{\mathrm{A}} = \sqrt{(\boldsymbol{u}, \boldsymbol{u})_{\mathrm{A}}}.$$

The energy norm is a natural metric for the convergence of the GLIS, when A is SPD, as the next few results show.

Theorem (Convergence in Energy Norm)

Suppose $\mathsf{A} \in \mathbb{R}^{n\times n}$ *is SPD and* $\mathsf{B} \in \mathbb{R}^{n\times n}$ *is symmetric and invertible. Define*

$$\mathsf{Q} := \mathsf{B}^{-1} - \frac{1}{2}\mathsf{A}.$$

If Q *is SPD, then the GLIS*

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^{k} + \mathsf{B}(\boldsymbol{f} - \mathsf{A}\boldsymbol{u}^{k}) \tag{24}$$

is unconditionally convergent with respect to the A*-norm, that is,*

$$\left\|\boldsymbol{e}^{k}\right\|_{\mathsf{A}} \to 0, \quad \text{as} \quad k \to \infty.$$

Moreover, the convergence is always monotonic in the sense that

$$\left\|\boldsymbol{e}^{k+1}\right\|_{\mathsf{A}}^{2} \leq \left\|\boldsymbol{e}^{k}\right\|_{\mathsf{A}}^{2}, \quad \forall\, k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}.$$

Proof.

Let $\boldsymbol{u}$ be the exact solution and $\boldsymbol{u}^k$ be the GLIS approximation at the k^{th} step. Then

$$\boldsymbol{e}^{k+1} = \mathrm{T}\boldsymbol{e}^k = \boldsymbol{e}^k - \mathrm{BA}\boldsymbol{e}^k.$$

Set $\boldsymbol{v}^{k+1} := \boldsymbol{e}^{k+1} - \boldsymbol{e}^k$. Then

$$\mathrm{B}^{-1}\boldsymbol{v}^{k+1} + \mathrm{A}\boldsymbol{e}^k = \boldsymbol{0}. \tag{25}$$

Taking the Euclidean inner product of the last equation with $\boldsymbol{v}^{k+1}$ gives

$$\left(\mathrm{B}^{-1}\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}\right) + \left(\mathrm{A}\boldsymbol{e}^k, \boldsymbol{v}^{k+1}\right) = 0.$$

Now, observe that

$$\boldsymbol{e}^k = \frac{1}{2}(\boldsymbol{e}^{k+1} + \boldsymbol{e}^k) - \frac{1}{2}(\boldsymbol{e}^{k+1} - \boldsymbol{e}^k) = \frac{1}{2}(\boldsymbol{e}^{k+1} + \boldsymbol{e}^k) - \frac{1}{2}\boldsymbol{v}^{k+1}.$$

Proof (Cont.)

Then,

$$
\begin{aligned}
0 &= (\mathrm{B}^{-1}\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}) + (\mathrm{A}\boldsymbol{e}^{k}, \boldsymbol{v}^{k+1}) \\
&= (\mathrm{B}^{-1}\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}) + \frac{1}{2}(\mathrm{A}(\boldsymbol{e}^{k+1} + \boldsymbol{e}^{k}), \boldsymbol{v}^{k+1}) - \frac{1}{2}(\mathrm{A}\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}) \\
&= \left(\left(\mathrm{B}^{-1} - \frac{1}{2}\mathrm{A}\right)\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}\right) + \frac{1}{2}(A(\boldsymbol{e}^{k+1} + \boldsymbol{e}^{k}), \boldsymbol{v}^{k+1}) \\
&= \left(\left(\mathrm{B}^{-1} - \frac{1}{2}\mathrm{A}\right)\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}\right) + \frac{1}{2}(\mathrm{A}(\boldsymbol{e}^{k+1} + \boldsymbol{e}^{k}), \boldsymbol{e}^{k+1} - \boldsymbol{e}^{k}) \\
&= \left(\left(\mathrm{B}^{-1} - \frac{1}{2}\mathrm{A}\right)\boldsymbol{v}^{k+1}, \boldsymbol{v}^{k+1}\right) + \frac{1}{2}\left(\left\|\boldsymbol{e}^{k+1}\right\|_{\mathrm{A}}^{2} - \left\|\boldsymbol{e}^{k}\right\|_{\mathrm{A}}^{2}\right) \\
&= \left\|\boldsymbol{v}^{k+1}\right\|_{\mathrm{Q}}^{2} + \frac{1}{2}\left(\left\|\boldsymbol{e}^{k+1}\right\|_{\mathrm{A}}^{2} - \left\|\boldsymbol{e}^{k}\right\|_{\mathrm{A}}^{2}\right). \qquad (26)
\end{aligned}
$$

Proof (Cont.)

By assumption, $Q := B^{-1} - \frac{1}{2}A$ is SPD. Hence

$$\left\| \boldsymbol{e}^{k+1} \right\|_A^2 \leq \left\| \boldsymbol{e}^{k} \right\|_A^2 .$$

and

$$\left\| \boldsymbol{e}^{k+1} \right\|_A \leq \left\| \boldsymbol{e}^{k} \right\|_A .$$

Thus $\left\{ \left\| \boldsymbol{e}^{k} \right\|_A \right\}$ is a decreasing sequence of nonnegative numbers. By the Monotone Convergence Theorem, there is some $\gamma \geq 0$, such that

$$\left\| \boldsymbol{e}^{k} \right\|_A \to \gamma \geq 0, \quad \text{as} \quad k \to \infty.$$

Passing to the limit $k \to \infty$ in (26), we get

$$\lim_{k \to \infty} \left\| \boldsymbol{v}^{k+1} \right\|_Q^2 = 0.$$

Using (25),

$$\boldsymbol{e}^{k} = -A^{-1} B^{-1} \boldsymbol{v}^{k+1},$$

we must have $\left\| \boldsymbol{e}^{k} \right\|_A \to 0$. □

We can generalize the last result. But first, we need the following:

Lemma

Suppose that $\mathsf{Q} \in \mathbb{R}^{n\times n}$ *is positive definite in the sense that*

$$(\mathsf{Q}\boldsymbol{y}, \boldsymbol{y}) > 0, \quad \forall\, \boldsymbol{y} \in \mathbb{R}^n_\star,$$

but Q *is not necessarily symmetric. Then*

$$\|\boldsymbol{w}\|_{\mathsf{Q}} = \sqrt{(\mathsf{Q}\boldsymbol{w}, \boldsymbol{w})}, \quad \forall\, \boldsymbol{w} \in \mathbb{R}^n,$$

defines a norm.

Proof.

Suppose that Q is not symmetric, to avoid the simple case. Then

$$\mathsf{Q} = \mathsf{Q}_S + \mathsf{Q}_A,$$

where

$$\mathsf{Q}_S := \frac{1}{2}\left(\mathsf{Q} + \mathsf{Q}^\top\right), \qquad \mathsf{Q}_A := \frac{1}{2}\left(\mathsf{Q} - \mathsf{Q}^\top\right),$$

are the symmetric and anti-symmetric parts, respectively. Observe that $\mathsf{Q}_S^\top = \mathsf{Q}_S$ and $\mathsf{Q}_A^\top = -\mathsf{Q}_A$. It follows that $(\mathsf{Q}_A \boldsymbol{y}, \boldsymbol{y})$ is zero, for any $\boldsymbol{y} \in \mathbb{R}^n$, because

$$(\mathsf{Q}_A \boldsymbol{y}, \boldsymbol{y}) = \boldsymbol{y}^\top \mathsf{Q}_A \boldsymbol{y} = \boldsymbol{y}^\top \mathsf{Q}_A^\top \boldsymbol{y} = -\boldsymbol{y}^\top \mathsf{Q}_A \boldsymbol{y} = -(\mathsf{Q}_A \boldsymbol{y}, \boldsymbol{y}).$$

Therefore, for all $\boldsymbol{y} \in \mathbb{R}^n_\star$,

$$0 < (\mathsf{Q}\boldsymbol{y}, \boldsymbol{y}) = (\mathsf{Q}_S \boldsymbol{y}, \boldsymbol{y}) + (\mathsf{Q}_A \boldsymbol{y}, \boldsymbol{y}) = (\mathsf{Q}_S \boldsymbol{y}, \boldsymbol{y}).$$

Thus, Q_S is SPD. Since

$$\|\boldsymbol{w}\|_{\mathsf{Q}_S} = \sqrt{(\mathsf{Q}_S \boldsymbol{w}, \boldsymbol{w})}, \quad \forall\, \boldsymbol{w} \in \mathbb{R}^n,$$

defines a norm, the result follows. □

Corollary

Suppose A *is SPD and* B *is invertible. Define*

$$\mathsf{Q} := \mathsf{B}^{-1} - \frac{1}{2}\mathsf{A},$$

and assume that Q *is positive definite in the sense that*

$$(\mathsf{Q}\boldsymbol{y}, \boldsymbol{y}) > 0, \quad \forall\, \boldsymbol{y} \in \mathbb{R}^n_\star,$$

but is not necessarily symmetric. Then the GLIS is convergent with respect to the A*-norm, that is,*

$$\left\| \boldsymbol{e}^k \right\|_{\mathsf{A}} \to 0, \quad \text{as} \quad k \to \infty.$$

Moreover, the convergence is always monotonic:

$$\left\| \boldsymbol{e}^{k+1} \right\|^2_{\mathsf{A}} \leq \left\| \boldsymbol{e}^{k} \right\|^2_{\mathsf{A}}, \quad \forall\, k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}.$$

Proof.

Exercise. □

Theorem (Householder–John)

Suppose that $A \in \mathbb{R}^{n\times n}$ *is nonsingular and symmetric and* $B \in \mathbb{R}^{n\times n}$ *is invertible. Assume that*

$$Q = B^{-1} + B^{-\top} - A$$

is SPD. Then the GLIS, with error transfer matrix $T = I_n - BA$*, converges unconditionally iff* A *is SPD. Moreover, if* A *is SPD, the convergence is always monotonic in the energy norm:*

$$\left\|\boldsymbol{e}^{k+1}\right\|_A^2 \leq \left\|\boldsymbol{e}^{k}\right\|_A^2, \quad \forall\, k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}.$$

Proof.

($\Leftarrow$): Suppose that A is SPD. Recall that $\|\cdot\|_A$ defines a norm on $\mathbb{R}^n$. The error equation is precisely

$$\boldsymbol{e}^{k+1} = (I - BA)\,\boldsymbol{e}^{k}.$$

Proof (Cont.)

Then,

$$
\begin{aligned}
\left\|\boldsymbol{e}^{k+1}\right\|_{\mathrm{A}}^{2} &= \left(\left(\mathrm{I}-\mathrm{BA}\right)\boldsymbol{e}^{k}\right)^{\top}\mathrm{A}\left(\left(\mathrm{I}-\mathrm{BA}\right)\boldsymbol{e}^{k}\right) \\
&= \left(\boldsymbol{e}^{k\top}\left(\mathrm{I}-\mathrm{AB}^{\top}\right)\right)\mathrm{A}\left(\left(\mathrm{I}-\mathrm{BA}\right)\boldsymbol{e}^{k}\right) \\
&= \left(\boldsymbol{e}^{k\top}-\boldsymbol{e}^{k\top}\mathrm{AB}^{\top}\right)\left(\mathrm{A}\boldsymbol{e}^{k}-\mathrm{ABA}\boldsymbol{e}^{k}\right) \\
&= \boldsymbol{e}^{k\top}\mathrm{A}\boldsymbol{e}^{k}-\boldsymbol{e}^{k\top}\mathrm{AB}^{\top}\mathrm{A}\boldsymbol{e}^{k}-\boldsymbol{e}^{k\top}\mathrm{ABA}\boldsymbol{e}^{k}+\boldsymbol{e}^{k\top}\mathrm{AB}^{\top}\mathrm{ABA}\boldsymbol{e}^{k} \\
&= \left\|\boldsymbol{e}^{k}\right\|_{\mathrm{A}}^{2}-\boldsymbol{e}^{k\top}\mathrm{A}\left(\mathrm{B}^{\top}+\mathrm{B}-\mathrm{B}^{\top}\mathrm{AB}\right)\mathrm{A}\boldsymbol{e}^{k} \\
&= \left\|\boldsymbol{e}^{k}\right\|_{\mathrm{A}}^{2}-\boldsymbol{e}^{k\top}\mathrm{AB}^{\top}\left(\mathrm{B}^{-1}+\mathrm{B}^{-\top}-\mathrm{A}\right)\mathrm{BA}\boldsymbol{e}^{k} \\
&= \left\|\boldsymbol{e}^{k}\right\|_{\mathrm{A}}^{2}-\left(\mathrm{BA}\boldsymbol{e}^{k}\right)^{\top}\left(\mathrm{B}^{-1}+\mathrm{B}^{-\top}-\mathrm{A}\right)\mathrm{BA}\boldsymbol{e}^{k}.
\end{aligned}
$$

Proof (Cont.)

Since $\mathrm{B}^{-1} + \mathrm{B}^{-\top} - \mathrm{A}$ is SPD and $\mathrm{BA}\boldsymbol{e}^k \neq \boldsymbol{0}$, in general, it follows that

$$\left\|\boldsymbol{e}^{k+1}\right\|_{\mathrm{A}}^2 + \left\|\mathrm{BA}\boldsymbol{e}^k\right\|_{\mathrm{Q}}^2 = \left\|\boldsymbol{e}^k\right\|_{\mathrm{A}}^2,$$

and $\left\|\boldsymbol{e}^k\right\|_{\mathrm{A}}$ is a decreasing sequence. Therefore, by the Monotone Convergence Theorem, $\left\|\boldsymbol{e}^k\right\|_{\mathrm{A}}$ converges, that is, there is some $\alpha \in [0, \infty)$, such that

$$\lim_{k\to\infty} \left\|\boldsymbol{e}^k\right\|_{\mathrm{A}} = \alpha = \lim_{k\to\infty} \left\|\boldsymbol{e}^{k+1}\right\|_{\mathrm{A}}.$$

This implies

$$\lim_{k\to\infty} \left\|\mathrm{BA}\boldsymbol{e}^k\right\|_{\mathrm{Q}} = 0,$$

which, in turn, implies that $\boldsymbol{e}^k \to \boldsymbol{0}$ as $k \to \infty$.

($\Rightarrow$): This direction is left as an exercise. □

The energy methods described earlier prove to be a more useful tool for establishing the convergence of the Gauss-Seidel Methods than the spectral approach when the coefficient matrix is SPD.

Theorem (Energy Convergence of Gauss-Seidel)

Suppose that $A \in \mathbb{R}^{n\times n}$ is SPD. Then the forward, backward and symmetric Gauss-Seidel methods are well defined, and the methods converge unconditionally.

Proof.

(Forward): For the forward Gauss-Seidel method, consider

$$\begin{aligned} Q_1 &= B_{GS}^{-1} + B_{GS}^{-\top} - A \\ &= D - L + D - U - D + L + U \\ &= D. \end{aligned}$$

Since A is SPD, its diagonal elements are positive. Thus, D is SPD. Now we can apply the Householder-John Theorem to see that the Gauss-Seidel method converges unconditionally.

Proof (Cont.)

We can also use our previous corollary. For that, set

$$\begin{aligned} Q_2 &= B_{GS}^{-1} - \frac{1}{2}A. \\ &= D - L - \frac{1}{2}D + \frac{1}{2}L + \frac{1}{2}U \\ &= \frac{1}{2}D - \frac{1}{2}L + \frac{1}{2}U. \end{aligned}$$

Now, let us consider the symmetric part of Q_2:

$$\begin{aligned} Q_{2,S} &= \frac{1}{2}\left(Q_2 + Q_2^\top\right) \\ &= \frac{1}{2}\left(\frac{1}{2}D - \frac{1}{2}L + \frac{1}{2}U + \frac{1}{2}D - \frac{1}{2}U + \frac{1}{2}L\right) \\ &= \frac{1}{2}D. \end{aligned}$$

Since $Q_{2,S}$ is SPD, the corollary implies that the Gauss-Seidel method converges.

Proof (Cont.)

(Backward): The same techniques show that the backward Gauss-Seidel method converges.

(Symmetric): For the symmetric Gauss-Seidel method, we have

$$B_{SGS} = (D - U)^{-1} D (D - L)^{-1}.$$

Observe that that the matrix

$$Q = B_{SGS}^{-1} + B_{SGS}^{-\top} - A = A + 2U^{\top} D^{-1} U$$

is SPD, since A is SPD. Now one can apply the Householder-John Theorem to see that the symmetric Gauss-Seidel method converges. □