

## Math 673/4

# 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

F24/S25



# Chapter 01, Part 2 of 2 Classical Iterative Methods



# Splitting Methods

### Classical Splitting Methods

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.

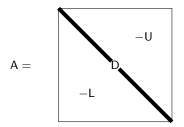


Figure: The canonical splitting of A.



#### 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



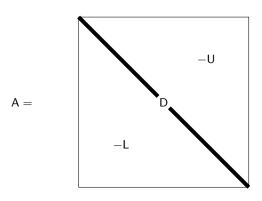


Figure: The canonical splitting of A.

#### 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

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

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

# Proposition

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

### 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  $A\pmb{u}=\pmb{f}$  can be rewritten as

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

Jacobi's method results from the iteration

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



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

$$\mathsf{B}=\mathsf{B}_J:=\mathsf{D}^{-1}$$

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

$$T_{J} = D^{-1}D - D^{-1}A$$

$$= D^{-1}(D - D + U + L)$$

$$= D^{-1}(U + L).$$
(3)

#### Remark

In component form, the Jacobi method can be written as

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

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

#### 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

$$\mathsf{T}_J = \left[ \begin{array}{ccc} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{array} \right] - \left[ \begin{array}{ccc} \frac{1}{a_{1,1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{a_{n,n}} \end{array} \right] \left[ \begin{array}{ccc} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{array} \right].$$

Thus

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

So,

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

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

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

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

Therefore, for each i = 1, ..., n,

$$\sum_{\substack{j=1\j
eq i}}^n \left| rac{\mathsf{a}_{i,j}}{\mathsf{a}_{i,i}} 
ight| \leq 1 - rac{\delta}{\left| \mathsf{a}_{i,i} 
ight|} \leq 1 - rac{\delta}{lpha},$$

where

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

Hence,

$$\left\|\mathsf{T}_{J}\right\|_{\infty} \leq 1 - \frac{\delta}{\alpha} < 1.$$



### 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

$$\mathbf{z} = \mathbf{u}^k + \mathsf{D}^{-1}(\mathbf{f} - \mathsf{A}\mathbf{u}^k)$$
  
$$\mathbf{u}^{k+1} = \omega \mathbf{z} + (1 - \omega)\mathbf{u}^k.$$

where the starting value  $u^0 \in \mathbb{R}^n$  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 z in the second equation to obtain the following.

#### **Proposition**

Splitting Methods

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

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

and error transfer matrix

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

The iteration sequence can be expressed as

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

#### 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

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

The Gauss-Seidel method is simply stated as

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

Assuming  $\mathsf{D}-\mathsf{L}$  is invertible, some manipulations give

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

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

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



### 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

$$\mathsf{B}=\mathsf{B}_{\mathit{GS}}:=\left(\mathsf{D}-\mathsf{L}\right)^{-1}$$

and error transfer matrix

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

The iteration sequence can be expressed as

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



#### Remark

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

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

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

$$= (D - L)^{-1}U.$$
(7)

# The Component Form of Gauss-Seidel



The Gauss-Seidel method is simply stated as

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

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

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

# The Component Form of Gauss-Seidel



To compute the update  $u^{k+1}$ , 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<sup>th</sup> step is

$$u_i^{k+1} = \frac{f_i - \sum_{j=1}^{i-1} a_{i,j} u_j^{k+1} - \sum_{j=i+1}^{n} a_{i,j} u_j^k}{a_{i,i}}.$$
 (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,  $\mathsf{D}-\mathsf{L}$  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  $\delta > 0$ ,

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

Thus

Splitting Methods

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

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

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

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

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



Now notice that

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



Therefore, we have

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

which implies

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

So,

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

which implies

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



### 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

$$\mathsf{B}=\mathsf{B}_{\mathit{BGS}}:=\left(\mathsf{D}-\mathsf{U}\right)^{-1}$$

and error transfer matrix

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

The iteration scheme is precisely

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



#### Remark

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

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

$$= I - (D - U)^{-1}(D - L - U)$$

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

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

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

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

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

Splitting Methods

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

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  $i^{\rm th}$  step is

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

#### Richardson's Method



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

$$A = \omega^{-1}I + A - \omega^{-1}I.$$

Suppose

$$Au = f$$
.

Then,

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

Richardson's method is, essentially,

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

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

$$\mathsf{B}=\mathsf{B}_R:=\omega\mathsf{I}$$

and error transfer matrix

$$T = T_R := I - \omega A$$
.

The iteration scheme is

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

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

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

From this, it follows that setting

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

one obtains the smallest possible value of  $\rho$ ,  $\rho_{\mathrm{opt}}$ , and

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

where

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

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

$$\lambda_n = \|A\|_2.$$

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

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

which implies

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

It follows that

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

This guarantees that  $\|T_R\|_2 = \rho(T_R) < 1$ , which implies convergence.



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

(Error estimate): By consistency,

$$\|\mathbf{e}^{k}\|_{2} = \|\mathsf{T}_{R}^{k}\mathbf{e}^{0}\|_{2} \leq \rho^{k}\|\mathbf{e}^{0}\|_{2}.$$

Of course, it is easy to see that

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



(Optimality): Finally, showing optimality amounts to minimizing  $\rho$ . See the figure on the next slide. From this we see that the minimum of  $\rho$  is attained when

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

or

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

which implies

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

Therefore

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

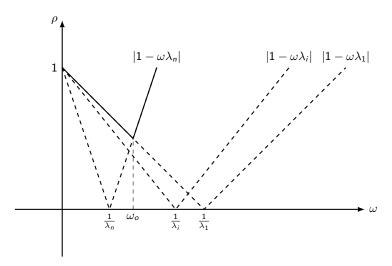


Figure: The curve  $\rho(T_R)$  (in solid black) as a function of  $\omega$ .



# 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:

- Additive combination.
- 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  $f \in \mathbb{R}^n$  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

$$z_{1} = u^{k} + B_{1} \left( f - A u^{k} \right),$$

$$z_{2} = z_{1} + B_{2} \left( f - A u^{k} \right),$$

$$z_{3} = z_{2} + B_{3} \left( f - A u^{k} \right),$$

$$\vdots$$

$$z_{r-1} = z_{r-2} + B_{r-1} \left( f - A u^{k} \right),$$

$$u^{k+1} = z_{r-1} + B_{r} \left( f - A u^{k} \right).$$
(13)

#### Definition

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

$$z_{1} = u^{k} + B_{1} (f - Au^{k}),$$

$$z_{2} = z_{1} + B_{2} (f - Az_{1}),$$

$$z_{3} = z_{2} + B_{3} (f - Az_{2}),$$

$$\vdots$$

$$z_{r-1} = z_{r-2} + B_{r-1} (f - Az_{r-2}),$$

$$u^{k+1} = z_{r-1} + B_{r} (f - Az_{r-1}).$$
(14)

## **Proposition**

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

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

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

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

where  $\tilde{B}_1 = B_1$ , and, for 2 < i < r,

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

#### Proof.

Exercise.

# Proposition

Suppose  $A \in \mathbb{R}^{n \times n}$  is invertible,  $\mathbf{f} \in \mathbb{R}^n$  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  $f \in \mathbb{R}^n$  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  $u^0$ , find  $u^1, u^2, \dots$  via

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

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

$$\mathbf{u}^{k+1} = \mathbf{z} + \mathbf{B}^{\mathsf{T}}(\mathbf{f} - \mathbf{A}\mathbf{z}). \qquad (16)$$



#### Lemma

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

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

where

$$B_{SM} = B + B^{\top} - B^{\top} AB. \tag{18}$$

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

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

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



## Proof.

Plugging (15) into (16) yields

$$\mathbf{u}^{k+1} = \mathbf{u}^{k} + \mathsf{B}(\mathbf{f} - \mathsf{A}\mathbf{u}^{k}) + \mathsf{B}^{\top} \left( \mathbf{f} - \mathsf{A} \left[ \mathbf{u}^{k} + \mathsf{B}(\mathbf{f} - \mathsf{A}\mathbf{u}^{k}) \right] \right)$$

$$= \mathbf{u}^{k} + \mathsf{B}(\mathbf{f} - \mathsf{A}\mathbf{u}^{k}) + \mathsf{B}^{\top} \left( \mathbf{f} - \mathsf{A}\mathbf{u}^{k} - \mathsf{A}\mathsf{B}(\mathbf{f} - \mathsf{A}\mathbf{u}^{k}) \right)$$

$$= \mathbf{u}^{k} + \mathsf{B}(\mathbf{f} - \mathsf{A}\mathbf{u}^{k}) + \mathsf{B}^{\top} \left( \mathbf{f} - \mathsf{A}\mathbf{u}^{k} \right) - \mathsf{B}^{\top} \mathsf{A}\mathsf{B}(\mathbf{f} - \mathsf{A}\mathbf{u}^{k})$$

$$= \mathbf{u}^{k} + \left( \mathsf{B} + \mathsf{B}^{\top} - \mathsf{B}^{\top} \mathsf{A}\mathsf{B} \right) (\mathbf{f} - \mathsf{A}\mathbf{u}^{k}),$$

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

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

If B is invertible, then

$$B_{SM} = B + B^{T} - B^{T}AB$$

$$= (I + B^{T}B^{-1} - B^{T}A)B$$

$$= (B^{T}B^{-T} + B^{T}B^{-1} - B^{T}A)B$$

$$= B^{T}(B^{-T} + B^{-1} - A)B.$$

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

$$B_{SM}^{\top} = B^{\top} + B - B^{\top} A^{\top} B$$
$$= B^{\top} + B - B^{\top} A B$$
$$= B_{SM}.$$



#### Remark

We could also consider a symmetrized additive method:

$$z = u^k + B(f - Au^k), (20)$$

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

This is a GLIS with iterator matrix

$$B = B_{SA} := B + B^{T}$$
.

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

$$\mathsf{B}_{GS}=(\mathsf{D}-\mathsf{L})^{-1}.$$

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

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





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

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

and it is symmetric.

#### Proof.

Using Equation (18), we find

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



# 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

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

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

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

If Q is SPD, then the GLIS

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

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

$$\|e^k\|_{\Lambda} \to 0$$
, as  $k \to \infty$ .

Moreover, the convergence is always monotonic in the sense that

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



#### Proof.

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

$$e^{k+1} = Te^k = e^k - BAe^k$$
.

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

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

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

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

Now, observe that

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



Then,

$$0 = (B^{-1}v^{k+1}, v^{k+1}) + (Ae^{k}, v^{k+1})$$

$$= (B^{-1}v^{k+1}, v^{k+1}) + \frac{1}{2}(A(e^{k+1} + e^{k}), v^{k+1}) - \frac{1}{2}(Av^{k+1}, v^{k+1})$$

$$= ((B^{-1} - \frac{1}{2}A)v^{k+1}, v^{k+1}) + \frac{1}{2}(A(e^{k+1} + e^{k}), v^{k+1})$$

$$= ((B^{-1} - \frac{1}{2}A)v^{k+1}, v^{k+1}) + \frac{1}{2}(A(e^{k+1} + e^{k}), e^{k+1} - e^{k})$$

$$= ((B^{-1} - \frac{1}{2}A)v^{k+1}, v^{k+1}) + \frac{1}{2}(||e^{k+1}||_{A}^{2} - ||e^{k}||_{A}^{2})$$

$$= ||v^{k+1}||_{Q}^{2} + \frac{1}{2}(||e^{k+1}||_{A}^{2} - ||e^{k}||_{A}^{2}).$$
(26)

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

$$\left\| \boldsymbol{e}^{k+1} \right\|_{\mathsf{A}} \leq \left\| \boldsymbol{e}^{k} \right\|_{\mathsf{A}}.$$

Thus  $\left\{\left\|\mathbf{e}^{k}\right\|_{\mathsf{A}}\right\}$  is a decreasing sequence of nonnegative numbers. By the Monotone Convergence Theorem, there is some  $\gamma\geq0$ , such that

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

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

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

Using (25),

$$\mathbf{e}^k = -\mathsf{A}^{-1}\mathsf{B}^{-1}\mathbf{v}^{k+1}.$$

we must have  $\|\boldsymbol{e}^k\|_{\Lambda} \to 0$ .



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

#### Lemma

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

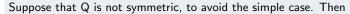
$$(Qy, y) > 0, \quad \forall y \in \mathbb{R}^n_{\star},$$

but Q is not necessarily symmetric. Then

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

defines a norm.

#### Proof.



$$Q = Q_S + Q_A$$

where

$$\mathsf{Q}_{\mathcal{S}} := \frac{1}{2} \left( \mathsf{Q} + \mathsf{Q}^\top \right), \qquad \mathsf{Q}_{\mathcal{A}} := \frac{1}{2} \left( \mathsf{Q} - \mathsf{Q}^\top \right),$$

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

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

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

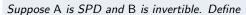
$$0 < (Qy, y) = (Q_Sy, y) + (Q_Ay, y) = (Q_Sy, y).$$

Thus,  $Q_S$  is SPD. Since

$$\|\mathbf{w}\|_{Q_{S}} = \sqrt{(Q_{S}\mathbf{w}, \mathbf{w})}, \quad \forall \mathbf{w} \in \mathbb{R}^{n},$$

defines a norm, the result follows.

#### Corollary



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

and assume that Q is positive definite in the sense that

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

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

$$\|e^k\|_{\Lambda} \to 0$$
, as  $k \to \infty$ .

Moreover, the convergence is always monotonic:

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

#### Proof.

Exercise.

# T

#### 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

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

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:

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

#### Proof.

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

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



Then.

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

Convergence in the Energy Norm

# Proof (Cont.)

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

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

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

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

This implies

$$\lim_{k\to\infty}\left\|\mathsf{B}\mathsf{A}\boldsymbol{e}^k\right\|_{\mathsf{Q}}=0,$$

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

(⇒): 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

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

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.

Convergence in the Energy Norm 0000000000000000

# Proof (Cont.)

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

$$Q_{2} = B_{GS}^{-1} - \frac{1}{2}A.$$

$$= D - L - \frac{1}{2}D + \frac{1}{2}L + \frac{1}{2}U$$

$$= \frac{1}{2}D - \frac{1}{2}L + \frac{1}{2}U.$$

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<sub>2,5</sub> is SPD, the corollary implies that the Gauss-Seidel method converges.



(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.  $\hfill\Box$