

## Iterative Techniques for Solving $Ax = b$ - (3.8)

Consider solving linear systems of the form:  $Ax = b$  where  $A = [a_{ij}]_{n \times n}$ ,  $x = [x_i]_{n \times 1}$ ,  $b = [b_i]_{n \times 1}$ . Assume that the system has a unique solution. Let  $x^*$  be the solution. Then  $x^* = A^{-1}b$ .

### 1. Jacobi and Gauss-Seidel Methods:

Let  $A = D - L - U$  where  $D = \text{diag}\{a_{ii}\}$ ,  $L = [l_{ij}]$ , a lower triangular matrix where  $l_{ij} = a_{ij}$  for  $i > j$ , and  $U = [u_{ij}]$  an upper triangular matrix where  $u_{ij} = a_{ij}$  for  $i < j$ .

**Example**  $A = \begin{bmatrix} -1 & 2 & -3 \\ 4 & -5 & 6 \\ -7 & 8 & 10 \end{bmatrix}$ . Then

$$D = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 10 \end{bmatrix}, L = \begin{bmatrix} 0 & 0 & 0 \\ -4 & 0 & 0 \\ 7 & -8 & 0 \end{bmatrix}, U = \begin{bmatrix} 0 & -2 & 3 \\ 0 & 0 & -6 \\ 0 & 0 & 0 \end{bmatrix}$$

Assume  $a_{ii} \neq 0$  for  $i = 1, \dots, n$ .  $Ax = b \Rightarrow (D - L - U)x = b$

#### a. Jacobi's Method:

$$Dx - (L + U)x = b \Rightarrow x = D^{-1}(L + U)x + D^{-1}b$$

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

#### b. Gauss-Seidel Method:

$$(D - L)x - Ux = b \Rightarrow (D - L)^{-1}Ux + (D - L)^{-1}b$$

$$x^{(k)} = (D - L)^{-1}Ux^{(k-1)} + (D - L)^{-1}b$$

#### c. Successive Over-Relaxation (SOR) Methods:

Consider  $Ax = b$

$$\omega Ax = \omega b \Leftrightarrow \omega(D - L - U)x = \omega b \Leftrightarrow D - D + \omega D - \omega L - \omega U = \omega b$$

$$(D - \omega L)x = ((1 - \omega)D + \omega U)x + \omega b$$

$$x = (D - \omega L)^{-1}((1 - \omega)D + \omega U)x + \omega(D - \omega L)^{-1}b$$

Choose  $\omega$  so that  $(D - \omega L)$  is invertible and  $\rho(T(\omega))$  is as small as possible.

When  $\omega > 1$ , the method is called SOR method; when  $0 < \omega < 1$ , the method is called successive under-relaxation method.

**Example** Solve the system of linear equations  $Ax = b$  where  $A = \begin{bmatrix} 4 & -1 & 1 \\ -1 & 5 & -1 \\ 1 & -1 & 7 \end{bmatrix}$  and  $b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

by Jacobi, Gauss-Seidel and SOR method using  $\omega = 1.2$ .

MatLab file: ex\_Jacobi\_GS\_SOR.m

## 2. Convergence of Jacobi, Gauss-Seidel and SOR Methods:

Consider the iterative methods of the form:

$$x^{(k)} = T x^{(k-1)} + c$$

where  $T$  is an  $n \times n$  matrix and  $c$  is  $n \times 1$  vector. For Jacobi Method:

$$T_{Jac} = D^{-1}(L + U), \quad c_{Jac} = D^{-1}b;$$

for Gauss-Seidel Method:

$$T_{GS} = (D - L)^{-1}U, \quad c_{GS} = (D - L)^{-1}b$$

and for SOR methods:

$$T_{SOR} = (D - \omega L)^{-1}((1 - \omega)D + \omega U), \quad c_{SOR} = \omega(D - \omega L)^{-1}b$$

Questions:

- Under what condition(s),  $x^{(k)} \rightarrow x$ ; and
- under what condition(s),  $x^{(k)} \rightarrow x^* = A^{-1}b$ ?

**Lemma** If  $\rho(T) < 1$ , then  $(I - T)^{-1}$  exists and

$$(I - T)^{-1} = I + T + T^2 + \dots = \sum_{j=0}^{\infty} T^j$$

Proof: Since  $\rho(T) < 1$ , all eigenvalues of  $T$  are less than 1. So, zero is not an eigenvalue of  $I - T$  and therefore,  $I - T$  is invertible and  $(I - T)^{-1}$  exists. Let  $S_m = \sum_{j=0}^{m-1} T^j$ . Then

$$\begin{aligned} S_m &= I + T + T^2 + \dots + T^{m-1}, \quad TS_m = T + T^2 + \dots + T^m \\ S_m - TS_m &= I - T^m, \quad (I - T)S_m = I - T^m, \quad S_m = (I - T)^{-1}(I - T^m). \end{aligned}$$

Since  $\rho(T) < 1$ ,  $\lim_{k \rightarrow \infty} T^k = 0_{n \times n}$ .

$$\lim_{m \rightarrow \infty} S_m = \lim_{m \rightarrow \infty} (I - T)^{-1}(I - T^m) = (I - T)^{-1}I.$$

**Theorem 1.** For any  $x^{(0)}$  in  $R^n$ ,  $\{x^{(k)}\}$  converges if and only if  $\rho(T) < 1$ .

Proof: Observe that

$$\begin{aligned} x^{(k)} &= T x^{(k-1)} + c = T(T x^{(k-2)} + c) + c = T^2 x^{(k-2)} + Tc + c \\ &= \dots = T^k x^{(0)} + (T^{k-1} + T^{k-2} + \dots + T + I)c \\ \lim_{k \rightarrow \infty} x^{(k)} &= \lim_{k \rightarrow \infty} (T^k x^{(0)} + (T^{k-1} + T^{k-2} + \dots + T + I)c) = (I - T)^{-1}c \end{aligned}$$

So, if  $\rho(T) < 1$ ,  $\{x^{(k)}\} \rightarrow (I - T)^{-1}c$ .

Is  $(I - T)^{-1}c = A^{-1}b$ ? Observe the following:

Jacobi Method:

$$\begin{aligned} (I - T_{Jac})^{-1}c_{Jac} &= (I - D^{-1}(L + U))^{-1}D^{-1}b = (D(I - D^{-1}(L + U)))^{-1}b \\ &= (D - L - U)^{-1}b = A^{-1}b = x^* \end{aligned}$$

Gauss-Seidel Method:

$$\begin{aligned}(I - T_{GS})^{-1}c_{GS} &= (I - (D - L)^{-1}U)^{-1}(D - L)^{-1}b = ((D - L)(I - (D - L)^{-1}U))^{-1}b \\ &= (D - L - U)^{-1}b = A^{-1}b = x^*\end{aligned}$$

SOR Methods:

$$\begin{aligned}(I - T_{SOR})^{-1}c_{SOR} &= (I - (D - \omega L)^{-1}((1 - \omega)D + \omega U))^{-1}\omega(D - \omega L)^{-1}b \\ &= (D - \omega L - ((1 - \omega)D + \omega U))^{-1}(D - \omega L)\omega(D - \omega L)^{-1}b \\ &= (\omega(D - L - U))^{-1}\omega b = \frac{1}{\omega}A^{-1}\omega b = A^{-1}b.\end{aligned}$$

Hence, if  $\{x^{(k)}\}$  converges then it converges to  $x^*$ .

Under what condition(s),  $\rho(T_{Jac}) < 1$ ,  $\rho(T_{GS}) < 1$  and  $\rho(T_{SOR}) < 1$ ?

**Definition** An  $n \times n$  matrix  $A = [a_{ij}]$  is said to be strictly diagonal dominant if  $|a_{ii}| > \sum_{i \neq j} |a_{ij}|$  for all  $i = 1, \dots, n$ .

**Example**  $A_1 = \begin{bmatrix} 4 & -1 & -1 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{bmatrix}$ ,  $A_2 = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ .

$A_1$  is strictly diagonal dominant but  $A_2$  is not.

**Theorem 2.** If  $A$  is strictly diagonal dominant, then for any choice of  $x^{(0)}$ , both the Jacobi and Gauss-Seidel methods converge.

We will show that if  $A$  is strictly diagonal dominant, then (1)  $\rho(T_{Jac}) < 1$  and (2)  $\rho(T_{GS}) < 1$ .

(1) Show  $\rho(T_{Jac}) < 1$  by showing that  $\|T_{Jac}\| < 1$ . Since  $A$  is strictly diagonal dominant, for  $1 \leq i \leq n$ ,

$$|a_{ii}| > \sum_{j \neq i}^n |a_{ij}| \Rightarrow 1 > \frac{1}{|a_{ii}|} \sum_{j \neq i}^n |a_{ij}|$$

$$T_{Jac} = D^{-1}(L + U), \quad \|T_{Jac}\|_{\infty} = \max_{1 \leq i \leq n} \frac{1}{|a_{ii}|} \sum_{j \neq i}^n |a_{ij}| < 1.$$

$$\rho(T_{Jac}) \leq \|T_{Jac}\|_{\infty} < 1$$

(2) Show  $\rho(T_{GS}) < 1$  by showing that all eigenvalue  $\lambda_i$ 's of  $T_{GS}$  are less than 1 and then clearly,

$$\rho(T_{GS}) = \max_{1 \leq i \leq n} \{|\lambda_i|\} < 1.$$

Recall that  $T_{GS} = (D - L)^{-1}U$ . Let  $(\lambda, x)$  be an eigenpair of  $T_{GS}$  where  $x$  in  $R^n$ ,  $\|x\|_{\infty} = 1$ . Then

$$T_{GS}x = \lambda x \Leftrightarrow (D - L)^{-1}Ux = \lambda x \Leftrightarrow Ux = \lambda(D - L)x.$$

Suppose that  $|x_p| = 1$  for some  $1 \leq p \leq n$ . First let us assume that  $p = 1$ , that is,  $|x_1| = 1$  and  $|x_i| \leq 1$  for  $i = 2, \dots, n$ . Observe that the first row of the equation:  $Ux = \lambda(D - L)x$ :

$$a_{12}x_2 + \dots + a_{1n}x_n = \lambda a_{11}x_1.$$

Then

$$\begin{aligned}|\lambda a_{11}x_1| \stackrel{|x_1|=1}{=} |\lambda| |a_{11}| &= |a_{12}x_2 + \dots + a_{1n}x_n| \leq |a_{12}x_2| + \dots + |a_{1n}x_n| \\ &= |a_{12}||x_2| + \dots + |a_{1n}||x_n| \leq |a_{12}| + \dots + |a_{1n}|\end{aligned}$$

and

$$|\lambda| \leq \frac{|a_{12}| + \dots + |a_{1n}|}{|a_{11}|} < 1.$$

Now let us assume that  $p = 2$ , that is,  $|x_2| = 1$  and  $|x_i| \leq 1$  for  $i = 1, 3, \dots, n$ . Observe that the 2nd row of the equation:  $Ux = \lambda(D - L)x$ :

$$a_{23}x_3 + \dots + a_{2n}x_n = \lambda(a_{21}x_1 + a_{22}x_2).$$

Then

$$\lambda a_{22}x_2 = a_{23}x_3 + \dots + a_{2n}x_n - \lambda a_{21}x_1$$

and

$$\begin{aligned} |\lambda| |a_{22}x_2| \stackrel{|x_2|=1}{=} |\lambda| |a_{22}| &= |a_{23}x_3 + \dots + a_{2n}x_n - \lambda a_{21}x_1| \leq |a_{23}x_3| + \dots + |a_{2n}x_n| - |\lambda| |a_{21}x_1| \\ &\leq |a_{23}||x_3| + \dots + |a_{2n}||x_n| + |\lambda| |a_{21}||x_1| \leq |a_{23}| + \dots + |a_{2n}| + |\lambda| |a_{21}| \end{aligned}$$

that implies,

$$|\lambda| (|a_{22}| - |a_{21}|) \leq |a_{23}| + \dots + |a_{2n}|.$$

From this inequality, we have

$$|\lambda| \leq \frac{|a_{23}| + \dots + |a_{2n}|}{|a_{22}| - |a_{21}|} < 1.$$

For more general, assume  $|x_p| = 1$  and  $|x_i| \leq 1$  for  $i = 1, \dots, p-1, p+1, \dots, n$ . Observe that the  $p$ th row of the equation:  $Ux = \lambda(D - L)x$ :

$$\begin{aligned} -(a_{p,p+1}x_{p+1} + a_{p,p+2}x_{p+2} + \dots + a_{p,n}x_n) &= \lambda(a_{pp}x_p + a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pn}x_n) \\ \lambda a_{pp}x_p &= -(a_{p,p+1}x_{p+1} + a_{p,p+2}x_{p+2} + \dots + a_{p,n}x_n) - \lambda a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pn}x_n \\ |\lambda| |a_{pp}x_p| = |\lambda| |a_{pp}| &= |-(a_{p,p+1}x_{p+1} + a_{p,p+2}x_{p+2} + \dots + a_{p,n}x_n) - \lambda a_{p1}x_1 + a_{p2}x_2 + \dots + a_{p,p-1}x_{p-1}| \\ &\leq \sum_{i=p+1}^n |a_{p,i}| + |\lambda| \sum_{i=1}^{p-1} |a_{pi}| \\ |\lambda| |a_{pp}x_p| - |\lambda| \sum_{i=1}^{p-1} |a_{pi}| &= |\lambda| \left( |a_{pp}x_p| - \sum_{i=1}^{p-1} |a_{pi}| \right) \leq \sum_{i=p+1}^n |a_{p,i}| \\ |\lambda| &\leq \frac{\sum_{i=p+1}^n |a_{p,i}|}{|a_{pp}x_p| - \sum_{i=1}^{p-1} |a_{pi}|} < 1. \end{aligned}$$

**Example**  $A = \begin{bmatrix} 4 & -1 & -1 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{bmatrix}$  and  $b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$  and  $x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ .

Eigenvalues of  $T_{Jac}$  are:  $-\frac{1}{4}, -\frac{1}{4}$  and  $\frac{1}{2}$ ; and eigenvalues of  $T_{GS}$  are 0,  $-0.05949631350069$  and  $0.26262131350069$ .

Hence,  $\rho(T_{Jac}) = \frac{1}{2} < 1$  and  $\rho(T_{GS}) = 0.26262131350069 < 1$ . Both methods converge. Since

$$\rho(T_{GS}) < \rho(T_{Jac})$$

we expect Gauss-Seidel Method converges faster. Check with the MatLab programs Jacobi.m and Gauss\_Seidel.m.

>> [xsol,xn,flag,k]=Jacobi(A,[1;2;3],3,zeros(3,1),10^(-8),100);

Jacobi Method converges

k = 26

```
>> [xsol,xn,flag,k]=Gauss_Seidel(A,[1;2;3],3,zeros(3,1),10^(-8),100);
```

Gauss-Seidel Method converges

k = 15

**Example**  $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$  and  $b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$  and  $x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ .

Note that the condition in Theorem 3 is a sufficient condition, that is, Jacobi or Gauss-Seidel Method could still converge even if  $A$  is not positive definite. For this example, clearly  $A$  is not strictly positive definite. However, eigenvalues of  $T_{Jac}$  are:  $-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}$  and 0; and eigenvalues of  $T_{GS}$  are 0, 0, and  $\frac{1}{2}$ .

Hence,  $\rho(T_{Jac}) = \frac{1}{\sqrt{2}} < 1$  and  $\rho(T_{GS}) = \frac{1}{2} < 1$ . Both methods converge. Since

$$\rho(T_{GS}) < \rho(T_{Jac})$$

we expect Gauss-Seidel Method converges faster. Check with the MatLab programs Jacobi.m and Gauss\_Seidel.m.

```
>> [xsol,xn,flag,k]=Jacobi(A,[1;2;3],3,zeros(3,1),10^(-8),100);
```

Jacobi Method converges

k = 54

```
>> [xsol,xn,flag,k]=Gauss_Seidel(A,[1;2;3],3,zeros(3,1),10^(-8),100);
```

Gauss-Seidel Method converges

k = 29

**Theorem 3.** If  $a_{ij} \leq 0$ , for each  $i \neq j$ , and  $a_{ii} > 0$  for each  $i = 1, 2, \dots, n$ , then one and only one of the following statements holds:

- $0 \leq \rho(T_{GS}) < \rho(T_J) < 1$ ;
- $1 < \rho(T_J) < \rho(T_{GS})$ ;
- $\rho(T_J) = \rho(T_{GS}) = 0$ ;
- $\rho(T_J) = \rho(T_{GS}) = 1$ .

Therefore, if both methods converge, then Gauss-Seidel Method is better. The previous example has illustrated the results given in Theorem 3.

**Definition** An  $n \times n$  symmetric matrix  $A$  is **positive definite** if  $x^T A x > 0$  for all nonzero vector in  $R^n$ .

Note: A symmetric matrix is positive definite if and only if all its eigenvalues are positive.

**Example**  $A = \begin{bmatrix} 4 & -1 & -1 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{bmatrix}$  is symmetric. Its eigenvalues are 5, 5, 2 so it is positive definite.

**Definition** An  $n \times n$  matrix  $A = [a_{ij}]$  is a **tridiagonal** matrix if  $a_{ij} = 0$  whenever  $i > j + 1$  or  $j > i + 1$ .

**Example**  $A = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -2 & 0 \\ 0 & -2 & 4 & -3 \\ 0 & 0 & -3 & 4 \end{bmatrix}$  is a tridiagonal matrix.

**Theorem 4.** Let  $T(\omega) = T_{SOR}$  for a given  $\omega$ . Then

(1) If  $a_{ii} \neq 0$ , for each  $i = 1, 2, \dots, n$ , then  $\rho(T(\omega)) \geq |\omega - 1|$ .

(2) If  $A$  is a **positive definite** matrix and  $0 < \omega < 2$ , then the SOR method converges for any choice of initial approximation vector  $x^{(0)}$ .

(3) If  $A$  is a **positive definite** and **tridiagonal** matrix, then  $\rho(T_{GS}) = [\rho(T_{Jac})]^2 < 1$ , and the optimal choice of  $\omega$  for the SOR method is

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - (\rho(T_{Jac}))^2}}.$$

With this choice of  $\omega_{opt}$ ,  $\rho(T(\omega_{opt})) = |\omega_{opt} - 1|$ .

(1) Let  $\lambda_i$  be eigenvalue of  $T(\omega)$ .

$$\begin{aligned} \det(T(\omega)) &= \det((D - \omega L)^{-1}((1 - \omega)D + \omega U)) \\ &= \det((D - \omega L)^{-1}) \det((1 - \omega)D + \omega U) \\ &= \frac{1}{a_{11} \dots a_{nn}} (1 - \omega)^n (a_{11} \dots a_{nn}) = (1 - \omega)^n \end{aligned}$$

$$\det(T(\omega)) = \prod_{k=1}^n \lambda_k = (1 - \omega)^n$$

$|\lambda_k| \geq |1 - \omega|$  at least for one  $k$ .

$$\rho(T(\omega)) = \max_{1 \leq k \leq n} \{|\lambda_k|\} \geq |1 - \omega|.$$

This implies that the SOR method can converge only if  $0 < \omega < 2$ .

**Example**  $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$  and  $b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$  and  $x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ .

Previously, we have  $\rho(T_{Jac}) = \frac{1}{\sqrt{2}}$  and  $\rho(T_{GS}) = \frac{1}{2} = (\rho(T_{Jac}))^2$ .  $A$  is a symmetric positive definite tridiagonal matrix. Then

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \left(\frac{1}{\sqrt{2}}\right)^2}} = 1.171572875$$

and  $\rho(T(\omega_{opt})) = |1.171572875 - 1| = 0.171572875$ . Check with the MatLab programs:

>> [xsol,xn,flag,k]=SOR(A,[1;2;3],3,zeros(3,1),10^(-8),100,omega);
SOR Method converges
>> k

$\omega$	$k$
1.10	21
1.15	17
1.171572875	14
1.2	14
1.25	15

### 3. Rate of Convergence of Jacobi and Gauss-Seidel Methods:

**Theorem 5.:** If  $\|T\| < 1$  for all natural matrix norm and  $c$  is a given vector, then  $\{x^{(k)}\}$  defined by  $x^{(k)} = Tx^{(k-1)} + c$  converges for any  $x^{(0)}$  to a vector  $x$  where  $x^{(0)}$  and  $x$  are in  $R^n$ , and

$$a. \quad \left\| x^* - x^{(k)} \right\| \leq \|T\|^k \left\| x^* - x^{(0)} \right\|; \quad \text{and } b. \quad \|x^* - x^{(k)}\| \leq \frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\|$$

where  $x^* = A^{-1}b$ .

Proof:

a. Observe that  $x^* = Tx^* + c$ , and

$$\begin{aligned} \left\| x^* - x^{(k)} \right\| &= \left\| (Tx^* + c) - (Tx^{(k-1)} + c) \right\| = \|T(x^* - x^{(k-1)})\| \\ &\leq \|T\| \|x^* - x^{(k-1)}\| \dots \\ &\leq \|T\|^k \left\| x^* - x^{(0)} \right\|. \\ \|x^* - x^{(k)}\| &\approx \rho(T) \|x^* - x^{(0)}\| \end{aligned}$$

Since

$$\frac{\|x^* - x^{(k)}\|}{\|x^* - x^{(k-1)}\|} \leq \|T\|,$$

by the definition of rate of convergence, we know  $x^{(k)} \rightarrow x^*$  linearly, and the asymptotic error constant is less than  $\|T\|$ . So, the smaller  $\|T\|$  is, the faster  $x^{(k)} \rightarrow x^*$ .

b. Because we don't know  $x^*$  at advance,

$$\begin{aligned} x^{(1)} - x^{(0)} &= Tx^{(0)} + c - x^{(0)} - x^* + x^* = Tx^{(0)} + c - x^{(0)} - Tx^* - c + x^* \\ &= T(x^{(0)} - x^*) - (x^{(0)} - x^*) = (I - T)(x^* - x^{(0)}) \\ x^* - x^{(0)} &= (I - T)^{-1}(x^{(1)} - x^{(0)}) \\ \|x^* - x^{(k)}\| &\leq \|T\|^k \|x^* - x^{(0)}\| = \|T\|^k \left\| (I - T)^{-1}(x^{(1)} - x^{(0)}) \right\| \\ &= \|T\|^k \left\| \left( I + T + T^2 + \dots \right) (x^{(1)} - x^{(0)}) \right\| \\ &\leq \|T\|^k (1 + \|T\| + \|T\|^2 + \dots) \|x^{(1)} - x^{(0)}\| \\ &= \frac{\|T\|^k}{1 - \|T\|} \|x^{(1)} - x^{(0)}\| \end{aligned}$$