## 7.4 Relaxation Techniques for Solving Linear Systems

**Definition** Suppose $\tilde{\boldsymbol{x}} \in R^n$ is an approximation to the solution of the linear system defined by $A\boldsymbol{x} = \boldsymbol{b}$**.** The **residual vector** for $\tilde{\boldsymbol{x}}$ with respect to this system is $\boldsymbol{r} = \boldsymbol{b} - A\tilde{\boldsymbol{x}}$.

*Objective of accelerating convergence*: Let residual vector converge to **0** rapidly.

In Gauss-Seidel method, we first associate with each calculation of an approximate component

$$\boldsymbol{x}_i^{(k)} \equiv (x_1^{(k)}, x_2^{(k)}, \ldots, x_{i-1}^{(k)}, x_i^{(k-1)}, \ldots, x_n^{(k-1)})^t$$

to the solution a residual vector

$$\boldsymbol{r}_i^{(k)} = (r_{1i}^{(k)}, r_{2i}^{(k)}, \ldots, r_{ni}^{(k)})^t$$

The $i$th component of $\boldsymbol{r}_i^{(k)}$ is

$$r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1}\left(a_{ij}x_j^{(k)}\right) - \sum_{j=i+1}^{n}\left(a_{ij}x_j^{(k-1)}\right) - a_{ii}x_i^{(k-1)} \qquad \text{Eq. (1)}$$

so

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = b_i - \sum_{j=1}^{i-1}\left(a_{ij}x_j^{(k)}\right) - \sum_{j=i+1}^{n}\left(a_{ij}x_j^{(k-1)}\right).$$

Also, $x_i^{(k)}$ is computed by

$$x_i^{(k)} = \frac{1}{a_{ii}}\left[-\sum_{j=1}^{i-1}(a_{ij}x_j^{(k)}) - \sum_{j=i+1}^{n}(a_{ij}x_j^{(k-1)}) + b_i\right] \qquad \text{Eq. (2)}$$

Therefore

$$a_{ii}x_i^{(k-1)} + r_{ii}^{(k)} = a_{ii}x_i^{(k)}$$

Gauss-Seidel method is characterized by

$$x_i^{(k)} = x_i^{(k-1)} + \frac{r_{ii}^{(k)}}{a_{ii}} \qquad \text{Eq. (3)}$$

Now consider the residual vector $\boldsymbol{r}_{i+1}^{(k)}$ associated with the vector $\boldsymbol{x}_{i+1}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \dots, x_i^{(k)}, x_{i+1}^{(k-1)}, \dots, x_n^{(k-1)})^t$

The $i$th component of $\boldsymbol{r}_{i+1}^{(k)}$ is

$$r_{i,i+1}^{(k)} = b_i - \sum_{j=1}^{i-1}\left(a_{ij}x_j^{(k)}\right) - \sum_{j=i+1}^{n}\left(a_{ij}x_j^{(k-1)}\right) - a_{ii}x_i^{(k)}$$

By Eq. (2), $r_{i,i+1}^{(k)} = 0$.

***<u>Idea of Successive Over-Relaxation (SOR)</u>*** (technique to accelerate convergence)

Modify Eq. (3) to

$$x_i^{(k)} = x_i^{(k-1)} + \omega\frac{r_{ii}^{(k)}}{a_{ii}} \qquad \text{Eq. (4)}$$

so that norm of residual vector $\boldsymbol{r}_{i+1}^{(k)}$ converges to 0 rapidly. Here $\omega > 0$.

**Under-relaxation method** when $0 < \omega < 1$

**Over-relaxation method** when $\omega > 1$

Use Eq. (4) and Eq. (1),

$$x_i^{(k)} = (1-\omega)x_i^{(k-1)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1}\left(a_{ij}x_j^{(k)}\right) - \sum_{j=i+1}^{n}\left(a_{ij}x_j^{(k-1)}\right)\right] \qquad \text{for } i = 1,2,,\dots n \qquad \text{Eq. (5)}$$

***Matrix form of SOR***

Rewrite Eq. (5) as

$$a_{ii}x_i^{(k)} + \omega \sum_{j=1}^{i-1} \left(a_{ij}x_j^{(k)}\right) = (1-\omega)a_{ii}x_i^{(k-1)} - \omega \sum_{j=i+1}^{n} \left(a_{ij}x_j^{(k-1)}\right) + \omega b_i$$

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

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

Define $T_\omega = (D-\omega L)^{-1}[(1-\omega)D + \omega U]$, $\boldsymbol{c}_\omega = \omega(D-\omega L)^{-1}\boldsymbol{b}$

**SOR** can be written as $\boldsymbol{x}^{(k)} = T_\omega \boldsymbol{x}^{(k-1)} + \boldsymbol{c}_\omega$.

**Example** Use SOR with $\omega = 1.25$ to solve

$$\begin{aligned} 4x_1 + 3x_2 &= 24 \\ 3x_1 + 4x_2 - x_3 &= 30 \\ -x_2 + 4x_3 &= -24 \end{aligned}$$

with $\boldsymbol{x}^{(0)} = (1,1,1)^t$.

**Theorem 7.24(Kahan)** If $a_{ii} \neq 0$, for each $i = 1,2,\ldots,n$, then $\rho(T_\omega) \geq |\omega - 1|$. This implies that the SOR method can converge only if $0 < \omega < 2$.

Recall: Theorem 7.19 $\rho(T) \leq 1$.

**Theorem 7.25(Ostrowski-Reich)** If $A$ is a positive definite matrix and $0 < \omega < 2$, then the SOR method converges for any choice of initial approximate vector $\boldsymbol{x}^{(0)}$.

**Theorem 7.26** If $A$ is a positive definite and tridiagonal, then $\rho(T_g) = [\rho(T_j)]^2 < 1$, and the optimal choice of $\omega$ for the SOR method is

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_j)]^2}}$$

With this choice of $\omega$, we have $\rho(T_\omega) = \omega - 1$.

**Example** Find the optimal choice of $\omega$ for the SOR method for the matrix

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 3 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}$$

Soln:

$$D = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & 0 & 0 \\ -3 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & -3 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$T_j = D^{-1}(L+U) = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & -3 & 0 \\ -3 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{3}{4} & 0 \\ -\frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$$

Compute eigenvalues of $T_j$.

$$\det(T_j - \lambda I) = 0$$

So $-\lambda(\lambda^2 - 0.625) = 0. \Rightarrow \lambda_1 = 0,\ \lambda_2 = \sqrt{0.625},\ \ \lambda_3 = -\sqrt{0.625}$.

Thus $\rho(T_j) = \sqrt{0.625}$. And $\omega = \frac{2}{1+\sqrt{1-[\rho(T_j)]^2}} = \frac{2}{1+\sqrt{1-0.625}} \approx 1.24$

## 7.5 Error Bounds and Iterative Refinement

Motivation. Residual vector $\boldsymbol{r} = \boldsymbol{b} - A\tilde{\boldsymbol{x}}$ can fail to provide accurate measurement on convergence

**Example** $A\boldsymbol{x} = \boldsymbol{b}$ given by

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

has the unique solution $\boldsymbol{x} = (1,1)^t$ Determine the residual vector for approximation $\tilde{\boldsymbol{x}} = (3, -0.0001)^t$

**Solution** $\boldsymbol{r} = \boldsymbol{b} - A\tilde{\boldsymbol{x}} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -0.0001 \end{bmatrix} = \begin{bmatrix} 0.0002 \\ 0 \end{bmatrix}$

**Theorem 7.27** Suppose that $\tilde{\boldsymbol{x}}$ is an approximation to the solution of $A\boldsymbol{x} = \boldsymbol{b}$, $A$ is a nonsingular matrix, and $\boldsymbol{r}$ is the residual vector for $\tilde{\boldsymbol{x}}$. Then for any natural norm,

$$||\boldsymbol{x} - \tilde{\boldsymbol{x}}|| \le ||\boldsymbol{r}|| \cdot ||A^{-1}||$$

and if $\boldsymbol{x} \neq \boldsymbol{0}$ and $\boldsymbol{b} \neq \boldsymbol{0}$

$$\frac{||\boldsymbol{x} - \tilde{\boldsymbol{x}}||}{||\boldsymbol{x}||} \le ||A|| \cdot ||A^{-1}|| \frac{||\boldsymbol{r}||}{||\boldsymbol{b}||}$$

**Condition Numbers**

**Definition** The **condition number** of the nonsingular matrix $A$ relative to the norm $||\cdot||$ is

$$K(A) = ||A|| \cdot ||A^{-1}||$$

Remark: Condition number of identity matrix $K(I) = 1$ relative to $||\cdot||_\infty$

A matrix $A$ is **well-conditioned** if $K(A)$ is close to 1, and is **ill-conditioned** if $K(A)$ is significantly greater than 1.

**Example** Determine the condition number for $A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}$.

**Solution** $A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix}$.

$$||A^{-1}||_\infty = 20000$$

$$K(A) = \left|\left|A\ \ \right|\right|_\infty \cdot \left|\left|A^{-1}\right|\right|_\infty = 3.0001 \cdot 20000 = 60002.$$

**Significance of condition number** Well-conditioned $A\boldsymbol{x} = \boldsymbol{b}$ implies a small residual error corresponds to accurate approximate solution.

**Estimate condition number**

Assume that $t$-digit arithmetic and Gaussian elimination are used to solve $A\boldsymbol{x} = \boldsymbol{b}$, the residual vector $\boldsymbol{r}$ for the approximation $\widetilde{\boldsymbol{x}}$ has

$$||\boldsymbol{r}|| \approx 10^{-t}||A|| \cdot ||\widetilde{\boldsymbol{x}}||$$

Consider to solve $A\boldsymbol{y} = \boldsymbol{r}$ with $t$-digit arithmetic. Let $\widetilde{\boldsymbol{y}}$ be approximation to $A\boldsymbol{y} = \boldsymbol{r}$

$$\widetilde{\boldsymbol{y}} \approx A^{-1}\boldsymbol{r} = A^{-1}(\boldsymbol{b} - A\widetilde{\boldsymbol{x}}) = A^{-1}\boldsymbol{b} - A^{-1}A\widetilde{\boldsymbol{x}} = \boldsymbol{x} - \widetilde{\boldsymbol{x}}$$

This implies $\boldsymbol{x} \approx \widetilde{\boldsymbol{x}} + \widetilde{\boldsymbol{y}}$.

$$\left|\left|\widetilde{\boldsymbol{y}}\right|\right| \approx \left|\left|A^{-1}\boldsymbol{r}\right|\right| \leq \left|\left|A^{-1}\right|\right| \cdot \left|\left|\boldsymbol{r}\right|\right| \approx \left|\left|A^{-1}\right|\right|\left(10^{-t}\left|\left|A\right|\right| \cdot \left|\left|\widetilde{\boldsymbol{x}}\right|\right|\right) = 10^{-t}\left|\left|\widetilde{\boldsymbol{x}}\right|\right|K(A)$$

Therefore

$$K(A) \approx \frac{\left|\left|\widetilde{\boldsymbol{y}}\right|\right|}{\left|\left|\widetilde{\boldsymbol{x}}\right|\right|}10^t.$$

**Example** Estimate condition number for system $\begin{bmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.710 & 9.6120 \\ 1.5611 & 5.1791 & 1.6852 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15913 \\ 28.544 \\ 8.4254 \end{bmatrix}$ solved by 5-digit rounding arithmetic. The exact solution is $\boldsymbol{x} = (1,1,1)^t$

**Solution** Use Gaussian elimination to solve with 5-digit rounding arithmetic gives

$\widetilde{\boldsymbol{x}} = (1.2001, 0.99991, 0.92538)^t$

The corresponding residual vector $\boldsymbol{r} = (-0.00518, 0.27412914, -0.186160367)^t$

Solving $A\boldsymbol{y} = \boldsymbol{r}$ by Gaussian elimination gives $\widetilde{\boldsymbol{y}} = (-0.20008, 8.9987 \times 10^{-5}, 0.074607)^t$

$$K(A) \approx \frac{\left|\left|\widetilde{\boldsymbol{y}}\right|\right|_\infty}{\left|\left|\widetilde{\boldsymbol{x}}\right|\right|_\infty}10^t = \frac{0.20008}{1.2001}10^5 = 16672$$

How does the round-off errors affect a system like $\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$?

Let the $(A + \delta A)\boldsymbol{x} = \boldsymbol{b} + \delta\boldsymbol{b}$ be the perturbed system associated with $A\boldsymbol{x} = \boldsymbol{b}$.

**Theorem 7.29**. Suppose $A$ is nonsingular and $||\delta A|| < \frac{1}{||A^{-1}||}$. The solution $\tilde{\boldsymbol{x}}$ to $(A + \delta A)\boldsymbol{x} = \boldsymbol{b} + \delta\boldsymbol{b}$ approximates the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b}$ with the error estimate $\frac{||x - \tilde{x}||}{||x||} \leq \frac{K(A)||A||}{||A|| - K(A)||\delta A||}\left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||}\right)$.