Today's Agenda

- Iterative methods for solving linear systems
  - (Gauss) Jacobi
  - Gauss Seidel

- Convergence analysis

Golub and Van Loan, *Matrix Computations 4th edition*, Johns Hopkins Press

Two types of approaches

- Direct methods
  - Gaussian elimination
  - Involve factorization such as LU, QR, Cholesky, *etc*
  - Impractical for large and/or sparse linear system

- Iterative methods
  - Solve Ax=b *iteratively*, yielding $x^{(1)}$, $x^{(2)}$, …
  - Some important questions:

a. How to design the iteration?
b. When to stop iterating?
c. Convergent? Does $x^{(k)} \rightarrow x$? Does $Ax^{(k)} \rightarrow b$?

A splitting framework

- Suppose $A = M - N$ for an invertible matrix M
- $Ax = b \rightarrow (M-N)x = b \rightarrow Mx = b + Nx$ (*)
- This leads to a *fixed-point* iteration

  $$x^{(k+1)} = M^{-1}b + M^{-1}Nx^{(k)}$$

- If $x^{(k)}$ converges, its limit point satisfies (*)

- Next, some choices of splitting…

UT DALLAS

Gauss-Jacobi (GJ or Jacobi)

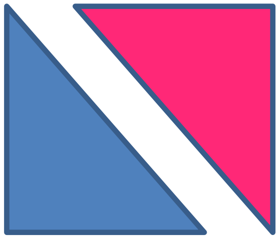
- Write A = L + D + U

  (strictly lower + diagonal + strictly upper)
- $Ax = b \rightarrow (L+D+U)x = b$

  $\rightarrow \quad Dx = b - (L+U)x$

If D is invertible

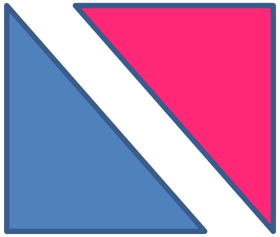
  $\rightarrow \quad x^{(k)} = D^{-1}b - D^{-1}(L+U)\, x^{(k-1)}$



- This is called the *Gauss-Jacobin* iteration.
- The GJ iteration is exhibited by the splitting A = M – N with M=D, N=-(L+U)
- It fails when diagonal has zero elements.

Gauss-Seidel (GS)

- Write A = L + D + U



- From (L+D+U)x = b, choose M = (L+D) and N = -U

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

- This is called the *Gauss-Seidel* iteration.
- It requires non-zero diagonal elements.

A toy example

- Consider a 3-by-3 system

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

- GJ $x^{(k)} = D^{-1}b - D^{-1}(L+U)\, x^{(k-1)}$ is exactly

$$\begin{aligned} x_1^{(k)} &= (b_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)})/a_{11}, \\ x_2^{(k)} &= (b_2 - a_{21}x_1^{(k-1)} - a_{23}x_3^{(k-1)})/a_{22}, \\ x_3^{(k)} &= (b_3 - a_{31}x_1^{(k-1)} - a_{32}x_2^{(k-1)})/a_{33}\,. \end{aligned}$$

GJ v.s. GS

$$
\begin{aligned}
x_1^{(k)} &= (b_1 - a_{12}x_2^{(k-1)} - a_{13}x_3^{(k-1)})/a_{11},\\
x_2^{(k)} &= (b_2 - a_{21}x_1^{(k-1)} - a_{23}x_3^{(k-1)})/a_{22},\\
x_3^{(k)} &= (b_3 - a_{31}x_1^{(k-1)} - a_{32}x_2^{(k-1)})/a_{33}.
\end{aligned}
$$

- An alternative update: use iterates k if possible.
- This becomes GS:

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

- Fixed point solution is

$$
\begin{aligned}
x_1 &= (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11},\\
x_2 &= (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22},\\
x_3 &= (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33}.
\end{aligned}
$$

In general,

- Gauss-Jacobi

for $i = 1{:}n$

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

end

- Gauss-Seidel

for $i = 1{:}n$

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

end

When to stop?

- Residual: $\frac{||Ax^{(k)}-b||_2}{||b||_2} < \tau_1$
- Relative errors of two consecutive iterates:

$$\frac{||x^{(k)} - x^{(k-1)}||_2}{||x^{(k)}||_2} < \tau_2$$

- Maximum iteration: $k < k_{\max}$

Pre-set $\tau_1, \tau_2, k_{\max}$

Convergence analysis

Convergence in general

- Analyze via the splitting framework, i.e.,

$$Mx^{(k)} = Nx^{(k-1)} + b \quad \text{with} \quad A = M - N \; (*)$$

- Define $G = M^{-1}N$ as the *iteration matrix.*

- *Whether* (*) converges depends on the eigenvalues of G.

Proof sketch

$$Mx^{(k)} = Nx^{(k-1)} + b \qquad Mx = Nx + b$$

$$M(x^{(k)} - x) = N(x^{(k-1)} - x) \qquad e^{(k)} = x^{(k)} - x$$

$$e^{(k)} = M^{-1}Ne^{(k-1)} = Ge^{(k-1)} = G^k e^{(0)}$$

$$\| e^{(k)} \| = \| G^k e^{(0)} \| \leq \| G^k \| \| e^{(0)} \| \leq \| G \|^k \| e^{(0)} \|.$$

Proof (cont'd)

It is the largest eigenvalue of G that matters.

For example,

$$G = \begin{bmatrix} \lambda & \alpha \\ 0 & \lambda \end{bmatrix},$$

Then

$$G^k = \begin{bmatrix} \lambda^k & \alpha\lambda^{k-1} \\ 0 & \lambda^k \end{bmatrix}.$$

Convergence Theorem

- Define the spectral radius of any matrix C

$$\rho(C) = \max\{ |\lambda| : \lambda \in \lambda(C) \}.$$

- Theorem statement

Theorem 11.2.1. *Suppose $A = M - N$ is a splitting of a nonsingular matrix $A \in \mathbb{R}^{n \times n}$. Assuming that M is nonsingular, the iteration (11.2.6) converges to $x = A^{-1}b$ for all starting n-vectors $x^{(0)}$ if and only if $\rho(G) < 1$ where $G = M^{-1}N$.*

- Spectral radius is different to the matrix spectral norm. They are same for sym matrix.

Recall

- Any vector norm induces a matrix norm

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

- Spectral radius

$$\rho(C) = \max\{\, |\lambda| : \lambda \in \lambda(C) \,\}.$$

- Spectral radius is the lower bound of any vector-induced matrix norms: $\rho(C) \leq ||C||$
- Matrix infinity norm (max row sum)

$$\|A\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|A\mathbf{x}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|.$$

GJ convergence

Recall that

$$\| A \|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$$

strictly diagonally dominant

$$\sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}| < |a_{ii}|, \qquad i = 1{:}n.$$

$$M_J \, x^{(k)} = N_J \, x^{(k-1)} + b$$

where $M_J = D_A$ and $N_J = -(L_A + U_A)$

Since $G_J = -D_A^{-1}(L_A + U_A)$ it follows that

$$\| G_J \|_\infty = \| D_A^{-1}(L_A + U_A) \|_\infty = \max_{1 \le i \le n} \sum_{\substack{j=1 \\ j \ne i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1.$$

Jacobi and Gauss-Seidel Convergence Theorem

If A is diagonally dominant, then the Jacobi and Gauss-Seidel methods converge for any starting vector $x^{(0)}$.

GJ vs GS

Consider four systems, with matrix A_i given below.
Consider their iteration matrices G_J and G_S.

$$A_1 = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & 1 & 2 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \qquad A_4 = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}.$$

Matlab:

```
D = diag(diag(A));
L = tril(A,-1);
U = triu(A,1);

GJ = -inv(D)*(L+U);
GS = -inv(L+D)*U;

max(abs(eig(GJ)))
max(abs(eig(GS)))
```

GJ vs GS

- For the system given by A_1, $\rho(G_J)>1$ but $\rho(G_S)<1$ → GJ diverges but GS converges
- For the system given by A_2, $\rho(G_J)<1$ but $\rho(G_S)>1$ → GJ converges but GS diverges
- For the system given by A_3, $\rho(G_J)=0.44$ and $\rho(G_S)=0.018$ → both converge but GJ typically converges slower than GS
- For the system given by A_4, $\rho(G_J)=0.64$ and $\rho(G_S)=0.77$ → both converge but GJ typically converges faster than GS

→ Take-away: there isn't a "one size fits all" answer for algorithm selection.