4.6 Iterative methods for linear systems

Idea: to solve $A\vec{x} = \vec{b}$, try to design a mapping $F: \mathbb{R}^n \to \mathbb{R}^n$

such that

$$\vec{x}^{(k)} = F(\vec{x}^{(k-1)})$$

converges to the sol'n $\vec{x}$.

(at least $\vec{x}$ has to be a fixed point of F)

$A\vec{x} = \vec{b}$ $\qquad \vec{0} = -A\vec{x} + \vec{b}$ $\qquad \vec{x} = \vec{x} - A\vec{x} + \vec{b}$

$= (I - A)\vec{x} + \vec{b}$

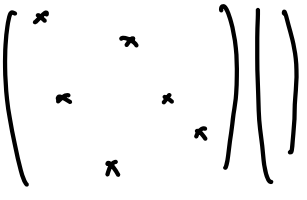
We may take $F(\vec{x}) = (I - A)\vec{x} + \vec{b}$ "Richardson method"

More generally, $Q\vec{x} = Q\vec{x} - A\vec{x} + \vec{b}$

$$\vec{x} = (I - Q^{-1}A)\vec{x} + Q^{-1}\vec{b}$$

When do we need iterative methods instead of Gaussian elimination?

- Design F s.t. the calculation of F is cheap (usually comparable w/ one calculation of $A\vec{x}$)
- If A is <u>sparse</u> (i.e., most entries are zero), then calculating $A\vec{x}$ is much cheaper than $O(n^2)$
  If iterative methods converge very fast (within much less than $O(n)$ iterations)



i	j	a_{ij}
1	1	2.5
2	3	-0.7
⋮	⋮	⋮

Jacobi method

Q = diagonal of A $\quad \text{diag}\{a_{11}, \cdots, a_{nn}\}$

$$Q\vec{x} = Q\vec{x} - A\vec{x} + \vec{b}$$

$$x_1^{(k)} = \frac{1}{a_{11}}\left(- \sum_{\substack{j=1 \\ j\neq 1}}^{n} a_{1j}\, x_j^{(k-1)} + b_1 \right)$$

$$x_2^{(k)} = \frac{1}{a_{22}}\left(- \sum_{\substack{j=1 \\ j\neq 2}}^{n} a_{2j}\, x_j^{(k-1)} + b_2 \right)$$

$$\vdots$$

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

Gauss-Seidel method

$$x_1^{(k)} = \frac{1}{a_{11}}\left(- \sum_{\substack{j=1 \\ j\neq 1}}^{n} a_{1j}\, x_j^{(k-1)} + b_1 \right)$$

$$x_2^{(k)} = \frac{1}{a_{22}}\left(- \sum_{j=1}^{1} a_{2j}\, x_j^{(k)} - \sum_{j=3}^{n} a_{2j}\, x_j^{(k-1)} + b_2 \right)$$

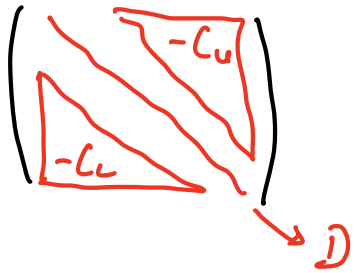
$$\vdots$$

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

(calculation cost is the same as Jacobi)

Write $A = D - C_L - C_U$

- D: diagonal
- C_L: strict lower-triangular part
- C_U: strict upper-triangular part



$$D\vec{x}^{(k)} = C_L\vec{x}^{(k)} + C_U\vec{x}^{(k-1)} + \vec{b}$$

$$(D - C_L)\vec{x}^{(k)} = \left((D - C_L) - A\right)\vec{x}^{(k-1)} + \vec{b}$$

$$Q = D - C_L \quad (\text{lower-triangular part of } A)$$

Compare Jacobi and Gauss-Seidel

- Gauss-Seidel method converges faster.
- Jacobi is more suitable for parallel computing.

Convergence analysis

$$\vec{x}^{(k)} = (I - Q^{-1}A)\vec{x}^{(k-1)} + Q^{-1}\vec{b}$$

Write $G = I - Q^{-1}A$, $\vec{c} = Q^{-1}\vec{b}$, then we get a general form

$$\vec{x}^{(k)} = G\vec{x}^{(k-1)} + \vec{c} := F(\vec{x}^{(k-1)})$$

Thm If $\|G\| < 1$ for some subordinate matrix norm, then $\{\vec{x}^{(k)}\}$ converges to the unique sol'n to $\vec{x} = G\vec{x} + \vec{c}$.

(In case $G = I - Q^{-1}A$, $\vec{c} = Q^{-1}\vec{b}$, $\{\vec{x}^{(k)}\}$ converges to the sol'n to $A\vec{x} = \vec{b}$).

Proof $\|F(\vec{x}) - F(\vec{y})\| = \|(G\vec{x} + \vec{c}) - (G\vec{y} + \vec{c})\|$

$$= \|G(\vec{x} - \vec{y})\| \leq \|G\| \cdot \|\vec{x} - \vec{y}\|$$

$\|G\| < 1$

$\Rightarrow F$ is contractive on $\mathbb{R}^n$.

- This also shows the convergence rate is at least linear.

Thm If A is row or column diagonally dominant, then Jacobi method converges.

Proof Suppose A is row diag. dom.

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \qquad i = 1, \cdots, n$$

$$G = I - Q^{-1}A$$

$$G_{ij} = \begin{cases} 0 & j = i \\ -\dfrac{a_{ij}}{a_{ii}} & j \neq i \end{cases}$$

$\|G\|_\infty$ = largest row sum of G (in abs. value)

$$\sum_{j=1}^{n} |G_{ij}| = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} = \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < 1$$

$\Rightarrow \|G\|_\infty < 1$.

For column diag. dom. matrix, use column sum, $\|\cdot\|_1$.

Thm For any $n \times n$ matrix A, its <u>spectral radius</u> $\rho(A)$ (i.e., largest possibly complex eigenvalue in abs. value), satisfies

$$\rho(A) = \inf_{\|\cdot\|} \|A\|$$

where inf is taken over all subordinate matrix norms.

Corollary The iteration $\vec{x}^{(k)} = G\,\vec{x}^{(k-1)} + \vec{c}$ converges to $(I-G)^{-1}\vec{c}$ for any $\vec{x}^{(0)}$ and any $\vec{c}$

$$\Leftrightarrow \rho(G) < 1$$

Proof "$\Leftarrow$" follows from thm

"$\Rightarrow$" If $\rho(G) \geq 1$, then let λ be the largest eigenvalue of G

(in abs. value), $|\lambda| \geq 1$

Take $\vec{c} = \vec{0}$, $\vec{x}^{(0)} = \vec{v}$ (an eigenvector corresponding to λ)

then $\vec{x}^{(k)} = G^k \vec{v} = \lambda^k \vec{v} \not\to \vec{0}$

Proof of Thm (sketch)

① Prove $\rho(A) \leq \|A\|$ for any sub. matrix norm $\|\cdot\|$.

Let λ be an eigenvalue of A. Then $A\vec{v} = \lambda\vec{v}$ for some $\vec{v} \neq \vec{0}$

Then $|\lambda| \cdot \|\vec{v}\| = \|\lambda\vec{v}\| = \|A\vec{v}\| \leq \|A\| \cdot \|\vec{v}\|$

$\Rightarrow |\lambda| \leq \|A\|$.

② Prove $\forall \varepsilon > 0$, $\exists \|\cdot\|$ s.t. $\rho(A) \geq \|A\| - \varepsilon$

Take suitable nonsingular matrix S to make $S^{-1}AS$ having small strict upper-triangular part

Then define $\|\vec{v}\| = \|S^{-1}\vec{v}\|_\infty$