Iterative methods for solving systems of linear equations.

Note Title

Let A be a $n \times n$ matrix. Let $Ax = b$ be a linear system having a unique solution.

Defn: A splitting of A is a decomposition $A = S - T$ where S is non-singular.

A splitting yields an iterative method as follows-

$$Ax = (S - T)x = b$$
$$Sx = Tx + b$$
$$x = \underbrace{S^{-1}T}_{R}x + S^{-1}b.$$
$$x = Rx + c, \quad R = S^{-1}T, \quad c = S^{-1}b.$$

This suggests the foll. iterative method: $\boxed{x_{k+1} = Rx_k + c.}$

R is called the matrix of the iterative method.

Lemma 1: Let $\|\cdot\|$ denote any induced matrix norm.

If $\|R\| < 1$, then $x_{k+1} = Rx_k + c$ converges for any x_0.

Proof: Consider

$$x_{k+1} = Rx_k + c$$
$$- \quad x = Rx + c$$
$$x_{k+1} - x = R(x_k - x).$$

Now $\|x_{k+1} - x\| = \|R(x_k - x)\| \le \|R\| \cdot \|x_k - x\|$

$\le \|R\| \cdot \|R\| \cdot \|x_{k-1} - x\|$

$= \|R\|^2 \cdot \|x_{k-1} - x\| \le \cdots \le \|R\|^{k+1} \|x_0 - x\|$

i.e. $\|x_{k+1} - x\| \le \|R\|^{k+1} \cdot \|x_0 - x\|$.

If $\|R\| < 1$, then $\|R\|^{k+1} \to 0$ as $k \to \infty$, $\therefore \|x_{k+1} - x\| \to 0$ as $k \to \infty$ i.e. $\{x_k\} \to x$ as $k \to \infty$.

Lemma 2: Let $S(R)$ denote the spectral radius of $n \times n$ matrix R.

(i.e. $S(R)$ = largest e-value of R in abs. value.)

① $S(R) \le \|R\|$, for every induced matrix norm.

② For every $\varepsilon > 0$, $\exists$ $\|\cdot\|_*$ such that $\|R\|_* \le S(R) + \varepsilon$.

($\|\cdot\|_*$ depending on R & ε; $\|R\|_*$ an induced norm.)

Proof: ① Let λ be s.t. $S(R) = |\lambda|$ & let x be the corr. evector so that $Rx = \lambda x$.

$$\|R\| = \max_{y \neq 0} \frac{\|Ry\|}{\|y\|} \geq \frac{\|Rx\|}{\|x\|} = \frac{\|\lambda x\|}{\|x\|} = |\lambda| = \rho(R).$$

② Refer to Lemma 6.5. (Demmel).

Theorem: The iterative method $x_{k+1} = Rx_k + c$ converges to the solution of $Ax = b$, $\forall x_0$ & $\forall b$

$$\iff \rho(R) < 1.$$

Proof: If $\rho(R) \geq 1$, let $|\lambda| = \rho(R)$ & let $Ry = \lambda y$.

Suppose $y + x = x_0$ i.e. $y = x_0 - x$.

then $Ry = R(x_0 - x) = \lambda(x_0 - x)$. ✓

$$x_{k+1} - x = R(x_k - x) = R^2(x_{k-1} - x) = \cdots = R^{k+1}(x_0 - x) = \lambda^{k+1}(x_0 - x)$$

$\not\to 0$ as $k \to \infty$

a contradiction to the assumption.

Conversely, if $\rho(R) < 1$, choose the norm $\|\cdot\|_A$ such that $\|R\|_A < 1$, then by lemma 1, the iterative method converges.

Questions to be considered when working with iterative methods -

1) Is the given method convergent? I.e. does there exist a matrix norm such that $\|R\| < 1$? Or is $\rho(R) < 1$?
2) Given 2 iterative methods, which converges faster? (i.e. which matrix has a smaller spectral radius?)

(Smaller spectral radius ⇒ faster convergence)

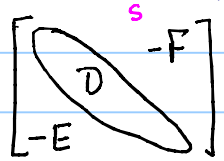
We will study 3 iterative methods for $Ax = b$:
- Jacobi
- Gauss-Seidel
- SOR

(I) **Jacobi's method:** Assume $a_{ii} \neq 0$ $\forall i$.

Split A as $A = \underbrace{D}_{S} - E - F$, where $D = \text{diag}(A)$

$-E$ = lower Δ of A

$-F$ = upper Δ of A.



The iterative method is $x_{k+1} = \overbrace{D^{-1}(E+F)}^{J} x_k + D^{-1}b$.

i.e. $x_{k+1} = J x_k + c$, where $c = D^{-1} b$

J is called the Jacobi matrix of A.

$$J = D^{-1}(E+F) = D^{-1}E + D^{-1}F = I - I + D^{-1}E + D^{-1}F$$
$$= I - D^{-1}(D - E - F)$$
$$= I - D^{-1}A.$$

Implementing Jacobi method

$$x_0 = \begin{pmatrix} x_1^0 \\ x_2^0 \\ \vdots \\ x_n^0 \end{pmatrix}.$$

$Ax = b$:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ \vdots \quad + a_{21}x_2 + \quad &= b_2 \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

$$\left.\begin{aligned} x_1 &= \frac{1}{a_{11}}\left(b_1 - a_{12}x_2 - \cdots - a_{1n}x_n\right) \\ x_2 &= \frac{1}{a_{22}}\left(b_2 - a_{21}x_1 - \cdots - a_{2n}x_n\right) \\ &\vdots \\ x_n &= \frac{1}{a_{nn}}\left(b_n - a_{n1}x_1 - \cdots - a_{n,n-1}x_{n-1}\right) \end{aligned}\right\} \star.$$

① Choose an appropriate initial approximation $X^0 = (x_1^0, \ldots, x_n^0)$ & substitute in the RHS of $\star$, to get the first approximation $X^1 = (x_1^1, \ldots, x_n^1)$

② Substitute X^1 in the RHS of $\star$ to get the second approximation $X^2 = (x_1^2, \ldots, x_n^2)$ & so on.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$
(A, X, b)

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

$$x_1^1 = (b_1 - a_{12}x_2^0 - a_{13}x_3^0)/a_{11}$$
$$x_2^1 = (b_2 - a_{21}x_1^0 - a_{23}x_3^0)/a_{22}$$
$$x_3^1 = (b_3 - a_{31}x_1^0 - a_{32}x_2^0)/a_{33}$$

$X^0 = (x_1^0, x_2^0, x_3^0)$

To see what happens at the $(k+1)^{th}$-step, consider $J = I - D^{-1}A$

so $X^{k+1} = JX^k + c$

$= (I - D^{-1}A)X^k + c$

$= X^k - D^{-1}AX^k + D^{-1}b$

$$\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \\ \vdots \\ x_n^{k+1} \end{pmatrix} - \begin{pmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_n^k \end{pmatrix}$$

$\rightarrow X^{k+1} = X^k - D^{-1}(AX^k - b)$

$\therefore x_i^{k+1} = x_i^k - \frac{1}{a_{ii}}\left(a_{i1}x_1^k - \cdots - a_{in}x_n^k - b_i\right)$

$$= \left(a_{ii}\, x_i^k - a_{i1} x_1^k - \cdots - a_{ii} x_i^k - \cdots - a_{in} x_n^k + b_i\right) / a_{ii}$$

$$= \frac{1}{a_{ii}} \left(b_i - a_{i1} x_1^k - \cdots - \widehat{a_{ii} x_i^k} - \cdots - a_{in} x_n^k\right)$$

Algorithm: (one step of Jacobi's method)

for $j = 1$ to n

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

end.

(II) Gauss-Seidel method

The number of allocations required in Jacobi's method can be reduced using the following trick: use x_1^{k+1} to calculate x_2^{k+1} and so on.

Then $x_1^{k+1} = \frac{1}{a_{11}} \left(b_1 - a_{12} x_2^k - \cdots - a_{1n} x_n^k\right)$

$$x_2^{k+1} = \frac{1}{a_{22}} \left(b_2 - a_{21} x_1^{k+1} - a_{23} x_3^k - \cdots - a_{2n} x_n^k\right)$$

$$x_3^{k+1} = \frac{1}{a_{33}} \left(b_3 - a_{31} x_1^{k+1} - a_{32} x_2^{k+1} - a_{34} x_4^k - \cdots - a_{3n} x_n^k\right)$$

& so on.

$$E \begin{pmatrix} 0 & 0 & \cdots \\ a_{21} & 0 & \cdots \\ a_{31} & a_{32} & 0 \cdots \\ \vdots & & \end{pmatrix}$$

This corresponds to the iterative method -

$$\rightarrow X^{k+1} = D^{-1}\left(E X^{k+1} + F X^k + b\right)$$

i.e. $D X^{k+1} = E X^{k+1} + F X^k + b$

i.e. $(D - E) X^{k+1} = F X^k + b$

i.e. $X^{k+1} = \underbrace{(D-E)^{-1} F}_{\text{denote by } L_1} X^k + \underbrace{(D-E)^{-1} b}_{c}$.

$$X^{k+1} = L_1 X^k + c.$$

Splitting is $A = D - E - F$.

Let $S = D - E$, $T = -F$.

S is invertible since D is.

L_1 is the matrix of this method & is called the Gauss-Seidel matrix of A.

Algorithm (one step of G-S)

for $j = 1$ to n

$$x_j^{k+1} = \frac{1}{a_{jj}} \left(b_j - \sum_{i=1}^{j-1} a_{ji}\, x_i^{k+1} - \sum_{i=j+1}^{n} a_{ji}\, x_i^{k} \right)$$

(x_i^{k+1}: updated x_i's; x_i^k: old x_i's)

end.

(III) Successive over-relaxation:

ω : relaxation parameter $(\omega \neq 0)$.

Idea: to improve on G-S by taking appropriate weighted average of x_j^{k+1} & x_i^k's:

in SOR $\quad x_j^{k+1} = (1-\omega)\, x_j^k + \omega \left(x_i^{k+1}\text{'s} \;\&\; x_i^k\text{'s} \right)$ $(i \neq j)$

For this, the splitting of A is defined as-

$$A = \left(\frac{D}{\omega} - E \right) - \left(\frac{1-\omega}{\omega} D + F \right)$$ (this simplifies to $D - E - F$)

The associated iterative method is -

$$x_{k+1} = \left(\frac{D}{\omega} - E \right)^{-1} \left(\frac{1-\omega}{\omega} D + F \right) x_k + c, \quad \text{where}$$

$(D - \omega E)\, x^{k+1} = \left((1-\omega) D + \omega F \right) x^k + b = (D - \omega D + \omega F) x^k + b$. $\quad c = \left(\frac{D}{\omega} - E \right)^{-1} b$.

The matrix of the method is - $\quad x^{k+1} = (D - \omega E)^{-1} \left[D x^k - \omega D x^k + \omega F x^k + b \right]$

$$L_\omega = \left(\frac{D}{\omega} - E \right)^{-1} \left(\frac{1-\omega}{\omega} D + F \right)$$

$$\text{OR} \quad (D - \omega E)^{-1} \left((1-\omega) D + \omega F \right).$$

If $\omega > 1$, the method is called "over-relaxation"

$\omega < 1$, " " "under-relaxation".

Implementing SOR:

$$X^{k+1} = (D-\omega E)^{-1}\left[DX^k - \omega DX^k + \omega FX^k + b\right]$$

$$X^{k+1} = L_\omega X^k + c$$

This corresponds to equations -

$$x_1^{k+1} = \frac{1}{a_{11}}\left(a_{11}x_1^k - \omega\left(a_{11}x_1^k + \cdots + a_{1n}x_n^k - b_1\right)\right)$$

$$(1-\omega)x_1^k + \frac{\omega}{a_{11}}\left(b_1 - a_{12}x_2^k - \cdots - a_{1n}x_n^k\right)$$

$$x_2^{k+1} = \frac{1}{a_{22}}\left(a_{22}x_2^k - \omega\left(a_{21}x_1^{k+1} + \cdots + a_{2n}x_n^k - b_2\right)\right)$$

$$(1-\omega)x_2^k + \frac{\omega}{a_{22}}\left(b_2 - a_{21}x_1^{k+1} - \cdots - a_{2n}x_n^k\right)$$

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(a_{ii}x_i^k - \omega\left(a_{i1}x_1^{k+1} + \cdots + a_{i,i-1}x_{i-1}^{k+1} + \cdots + a_{in}x_n^k - b_i\right)\right)$$

$$(1-\omega)x_i^k + \frac{\omega}{a_{ii}}\left(b_i - a_{i1}x_1^{k+1} - \cdots - a_{i,i-1}x_{i-1}^{k+1} - \cdots - a_{in}x_n^k\right)$$

By choosing appropriate ω, it is possible to attain a higher rate of convergence than G-S.

Algorithm (one step of SOR)

for $j = 1$ to n

$$x_j^{k+1} = (1-\omega)x_j^k + \frac{\omega}{a_{jj}}\left[b_j - \sum_{i=1}^{j-1} a_{ji}x_i^{k+1} - \sum_{i=j+1}^{n} a_{ji}x_i^k\right]$$

end.

Defn: The rate of convergence of $x_{k+1} = Rx_k + c$ is $r(R) = -\log_{10}\rho(R)$.

(So smaller the $\rho(R)$, higher the rate of convergence i.e. more number of correct decimal places computed per iteration.

See sections 6.5.3 & 6.5.4 of Demmel.

Compare $\rho(J)$, $\rho(L_1)$, $\rho(L_\omega)$ for some values ω.

Key facts about Jacobi's method -

① requires non-zero diagonal entries (can be accomplished by permuting rows/columns if not already true)

② requires $2n$ memory allocations ($\dim A = n \times n$).

③ components do not depend on one another, so they can be computed simultaneously.

④ does not always converge; converges for sure when A is SDD. [i.e. $|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$]

Key facts about G-S method.

① requires non-zero diagonal entries.

② requires n memory allocations at each step.

③ each component depends on previous ones, so must be computed successively.

④ converges if A is HPD (weaker than SDD)

⑤ When they converge together, G-S cges twice as fast as Jacobi.

Key facts/questions about SOR:

① requires non-zero diagonal entries.

② 2 questions: (i) for what ω is $\rho(L_\omega) < 1$?

Is there an interval $I \subset \mathbb{R}$ such that $\rho(L_\omega) < 1 \quad \forall \omega \in I$.

(ii) Is there an optimal ω_0 such that

$$\rho(L_{\omega_0}) = \inf_{\omega \in I} \rho(L_\omega).$$