Several Details of Modern Numerical Algebra Methods

K. Honn

hanqch21@lzu.edu.cn

February 3, 2024

Residual of Approx Solution of $Ax = b$

Suppose A, b is not exact and about their exact values A_1, b_1 we have

$$A_1 = A + \Delta, \quad b_1 = b + \eta.$$

Problem

Find estimate of approx solution $\|x\|$ in terms of $\|\Delta\|$ and $\|\eta\|$.

Denote X and X_1 are exact solutions respectively for $Ax = b$ and $A_1 x = b_1$. Define $r = X_1 - X$, we have

$$(A + \Delta)(X + r) = b + \eta.$$

Denote X and X_1 are exact solutions respectively for $Ax = b$ and $A_1x = b_1$. Define $r = X_1 - X$, we have

$$(A + \Delta)(X + r) = b + \eta.$$

Subtract with $Ax = b$:

$$Ar + \Delta X + \Delta r = \eta \Rightarrow r = A^{-1}(\eta - \Delta X - \Delta r),$$

Denote X and X_1 are exact solutions respectively for $Ax = b$ and $A_1 x = b_1$. Define $r = X_1 - X$, we have

$$(A + \Delta)(X + r) = b + \eta.$$

Subtract with $Ax = b$:

$$Ar + \Delta X + \Delta r = \eta \Rightarrow r = A^{-1}(\eta - \Delta X - \Delta r),$$

$$\|r\| \leqslant \|A^{-1}\|\|\eta\| + \|A^{-1}\|\|\Delta\|\|X\| + \|A^{-1}\|\|\Delta\|\|r\|.$$

Denote X and X_1 are exact solutions respectively for $Ax = b$ and $A_1 x = b_1$. Define $r = X_1 - X$, we have

$$(A + \Delta)(X + r) = b + \eta.$$

Subtract with $Ax = b$:

$$Ar + \Delta X + \Delta r = \eta \Rightarrow r = A^{-1}(\eta - \Delta X - \Delta r),$$

$$\|r\| \leqslant \|A^{-1}\|\|\eta\| + \|A^{-1}\|\|\Delta\|\|X\| + \|A^{-1}\|\|\Delta\|\|r\|.$$

For $\|A^{-1}\|\|\Delta\| < 1$ we have

$$\|r\| \leqslant \frac{\|A^{-1}\|(\|\eta\| + \|\Delta\|\|X\|)}{1 - \|A^{-1}\|\|\Delta\|}. \tag{1}$$

Condition Number

Definition (Condition number of system $Ax = b$)

That is,

$$\tau = \sup_{\eta} \left(\frac{\|r\|}{\|X\|} : \frac{\|\eta\|}{\|b\|} \right) = \frac{\|b\|}{\|X\|} \sup_{\eta} \frac{\|r\|}{\|\eta\|}.$$

Definition (Condition number of matrix A)

That is,

$$\nu(A) = \sup_{b} \tau.$$

Special Case $\Delta = 0$

$$r = A^{-1}\eta, \quad \|r\| \leqslant \|A^{-1}\|\|\eta\|.$$

Special Case $\Delta = 0$

$$r = A^{-1}\eta, \quad \|r\| \leqslant \|A^{-1}\|\|\eta\|.$$

- $\sup_\eta \frac{\|r\|}{\|\eta\|} = \|A^{-1}\| \Rightarrow \tau = \frac{\|b\|}{\|X\|}\|A^{-1}\|.$

Special Case $\Delta = 0$

$$r = A^{-1}\eta, \quad \|r\| \leqslant \|A^{-1}\|\|\eta\|.$$

- $\sup_\eta \frac{\|r\|}{\|\eta\|} = \|A^{-1}\| \Rightarrow \tau = \frac{\|b\|}{\|X\|}\|A^{-1}\|.$
- $\sup_b \frac{\|b\|}{\|X\|} = \|A\| \Rightarrow \nu(A) = \|A\|\|A^{-1}\|.$

Special Case $\Delta = 0$

$$r = A^{-1}\eta, \quad \|r\| \leqslant \|A^{-1}\|\|\eta\|.$$

- $\sup_\eta \frac{\|r\|}{\|\eta\|} = \|A^{-1}\| \Rightarrow \tau = \frac{\|b\|}{\|X\|}\|A^{-1}\|$.
- $\sup_b \frac{\|b\|}{\|X\|} = \|A\| \Rightarrow \nu(A) = \|A\|\|A^{-1}\|$.

By $\|A\| \geqslant \max|\lambda_A|$ and $\|A^{-1}\| \geqslant \max\frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}$, we have

$$\nu(A) \geqslant \max|\lambda_A|/\min|\lambda_A| \geqslant 1,$$

Special Case $\Delta = 0$

$$r = A^{-1}\eta, \quad \|r\| \leqslant \|A^{-1}\|\|\eta\|.$$

- $\sup_\eta \frac{\|r\|}{\|\eta\|} = \|A^{-1}\| \Rightarrow \tau = \frac{\|b\|}{\|X\|}\|A^{-1}\|$.
- $\sup_b \frac{\|b\|}{\|X\|} = \|A\| \Rightarrow \nu(A) = \|A\|\|A^{-1}\|$.

By $\|A\| \geqslant \max|\lambda_A|$ and $\|A^{-1}\| \geqslant \max\frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}$, we have

$$\nu(A) \geqslant \max|\lambda_A|/\min|\lambda_A| \geqslant 1,$$

for $A^T = A$: $\nu(A) = \max|\lambda_A|/\min|\lambda_A|$.

Numerical Results: Binary word-length $= t$

In this situation,

$$\|\eta\| = O(\|b\|2^{-t}) \quad \text{and} \quad \|\eta\|/\|b\| = O(2^{-t}),$$

so

$$\|r\|/\|X\| \leqslant \nu(A)\,O(2^{-t}).$$

C.g. Method (Hestenes and Stiefel, 1952)

Problem

C.g. method is adopted to solve $Ax = b$ with $A = A^T > 0$.

C.g. Method (Hestenes and Stiefel, 1952)

Problem

C.g. method is adopted to solve $Ax = b$ with $A = A^T > 0$.

Fix an estimate x^0 of X. Denote initial residual $r^0 = A(x^0 - X)$ and n-th step residual r^n, we have to find P_n such that

$$r^n = P_n(A) r^0, \quad P_n(0) = 1$$

minimize the functional

$$\mathscr{F}(x^n) = (Ax^n, x^n) - 2(b, x^n).$$

C.g. Method (Hestenes and Stiefel, 1952)

Problem

C.g. method is adopted to solve $Ax = b$ with $A = A^T > 0$.

Fix an estimate x^0 of X. Denote initial residual $r^0 = A(x^0 - X)$ and n-th step residual r^n, we have to find P_n such that

$$r^n = P_n(A)r^0, \quad P_n(0) = 1$$

minimize the functional

$$\mathscr{F}(x^n) = (Ax^n, x^n) - 2(b, x^n).$$

Since

$$\mathscr{F}(x^n) = \|Ax^n - b\|_{A^{-1}}^2 - \|X\|_A^2 = \|r^n\|_{A^{-1}}^2 - \|X\|_A^2,$$

we have to find P_n minimize $\|r^n\|_{A^{-1}}$.

Fix

$$P_n(\lambda) = \sum_{k=0}^{n} c_k \lambda^k, \quad c_0 = 1, \quad r^0 = \sum_{i=1}^{q} r_i e_i,$$

where e_i are eigenvectors of A. W.l.o.g, each e_i corresponding to different eigenvalue λ_i. Hence r^n formed

$$r^n = \left(\sum_{k=0}^{n} c_k A^k \right) r^0 = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k A^k \right) e_i = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k \lambda_i^k \right) e_i$$

Fix

$$P_n(\lambda) = \sum_{k=0}^{n} c_k \lambda^k, \quad c_0 = 1, \quad r^0 = \sum_{i=1}^{q} r_i e_i,$$

where e_i are eigenvectors of A. W.l.o.g, each e_i corresponding to different eigenvalue λ_i. Hence r^n formed

$$r^n = \left(\sum_{k=0}^{n} c_k A^k \right) r^0 = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k A^k \right) e_i = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k \lambda_i^k \right) e_i$$

so

$$\|r^n\|_{A^{-1}}^2 = (A^{-1} r^n, r^n) = \sum_{k,j=0}^{n} c_k c_j \left(\sum_{i=1}^{q} \lambda_i^{k+j-1} r_i^2 \right).$$

Fix

$$P_n(\lambda) = \sum_{k=0}^{n} c_k \lambda^k, \quad c_0 = 1, \quad r^0 = \sum_{i=1}^{q} r_i e_i,$$

where e_i are eigenvectors of A. W.l.o.g, each e_i corresponding to different eigenvalue λ_i. Hence r^n formed

$$r^n = \left(\sum_{k=0}^{n} c_k A^k\right) r^0 = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k A^k\right) e_i = \sum_{i=1}^{q} r_i \left(\sum_{k=0}^{n} c_k \lambda_i^k\right) e_i$$

so

$$\|r^n\|_{A^{-1}}^2 = (A^{-1} r^n, r^n) = \sum_{k,j=0}^{n} c_k c_j \left(\sum_{i=1}^{q} \lambda_i^{k+j-1} r_i^2\right).$$

Letting $\frac{\partial}{\partial c_l} \|r^n\|_{A^{-1}}^2 = 0$ for some l, we yield

$$2 \sum_{j=0}^{n} c_j \left(\sum_{i=0}^{q} \lambda_i^{j+l-1} r_i^2\right) = 2 \sum_{i=0}^{q} \left(\sum_{j=0}^{n} c_j \lambda_i^j r_i \lambda_i^{l-1} r_i\right) =$$

$$= 2(r^n, A^{l-1} r^0) = 0.$$

$$\boxed{(r^n, A^l r^0) = (r^n, r^l) = 0, \quad l = 0, \cdots, n-1.}$$

$$\boxed{(r^n, A^l r^0) = (r^n, r^l) = 0, \quad l = 0, \cdots, n-1.}$$

Denote $L_k = \mathrm{span}(r^0,\ Ar^0, \cdots, A^k r^0)$.

$$\boxed{(r^n, A^l r^0) = (r^n, r^l) = 0, \quad l = 0, \cdots, n-1.}$$

Denote $L_k = \text{span}(r^0, Ar^0, \cdots, A^k r^0)$.

- For $j < k$ we have $L_j \subset L_k$ hence $r^j \in L_k$;

$$\boxed{(r^n, A^l r^0) = (r^n, r^l) = 0, \quad l = 0, \cdots, n-1.}$$

Denote $L_k = \mathrm{span}(r^0,\ Ar^0, \cdots, A^k r^0)$.

- For $j < k$ we have $L_j \subset L_k$ hence $r^j \in L_k$;
- For $j > k$ $r^j \notin L_k$.

$$\boxed{(r^n, A^l r^0) = (r^n, r^l) = 0, \quad l = 0, \cdots, n-1.}$$

Denote $L_k = \mathrm{span}(r^0,\ Ar^0,\ \cdots,\ A^k r^0)$.

- For $j < k$ we have $L_j \subset L_k$ hence $r^j \in L_k$;
- For $j > k$ $r^j \notin L_k$.

so $L_k = \mathrm{span}(r^0,\ r^1,\ \cdots,\ r^k)$ and $r^n \perp L_{n-1}$.

$$\boxed{(r^n, A^l r^0) = (r^n, r^l) = 0, \quad l = 0, \cdots, n-1.}$$

Denote $L_k = \text{span}(r^0, Ar^0, \cdots, A^k r^0)$.

- For $j < k$ we have $L_j \subset L_k$ hence $r^j \in L_k$;
- For $j > k$ $r^j \notin L_k$.

so $L_k = \text{span}(r^0, r^1, \cdots, r^k)$ and $r^n \perp L_{n-1}$.
For $r^n \in L_n$,

$$r^n = \sum_{k=0}^{n-1} \gamma_k r^k + \gamma_n A r^{n-1}.$$

For $j = 0, \cdots, n-3$,

$$(Ar^{n-1}, r^j) = (r^{n-1}, Ar^j) = 0.$$

Hence $\gamma_1, \gamma_2, \cdots, \gamma_{n-3} = 0$,

$$r^n = \gamma_{n-1} r^{n-1} + \gamma_{n-2} r^{n-2} + \gamma_n A r^{n-1}.$$

$$\boxed{\begin{aligned} r^n &= \gamma_{n-1} r^{n-1} + \gamma_{n-2} r^{n-2} + \gamma_n A r^{n-1}, \\ r^{n-2} &= r^0 + \sum_{k=1}^{n-2} c_k A^k r^0, \\ r^{n-1} &= r^0 + \sum_{k=1}^{n-1} c_k A^k r^0, \\ A r^{n-1} &= \sum_{k=1}^{n} p_k A^k r^0 \quad \text{(sum begins at } k=1 \text{ since } (A r^{n-1}, r^0) = 0. \end{aligned}}$$

$$
\boxed{
\begin{aligned}
r^n &= \gamma_{n-1} r^{n-1} + \gamma_{n-2} r^{n-2} + \gamma_n A r^{n-1}, \\
r^{n-2} &= r^0 + \sum_{k=1}^{n-2} c_k A^k r^0, \\
r^{n-1} &= r^0 + \sum_{k=1}^{n-1} c_k A^k r^0, \\
A r^{n-1} &= \sum_{k=1}^{n} p_k A^k r^0 \quad \text{(sum begins at } k=1 \text{ since } (A r^{n-1}, r^0) = 0.
\end{aligned}
}
$$

$$
r^n = P_n(A) r^0 = (\gamma_{n-1} + \gamma_{n-2}) r^0 + \sum_{k=1}^{n} c_k A^k r^0
$$

but $P_n(0) = 1$, hence $\gamma_{n-1} + \gamma_{n-2} = 1$. Define $\gamma_{n-1} - 1 = \alpha_{n-1}, \gamma_n = \beta_{n-1}$,

$$\boxed{\begin{aligned} r^n &= \gamma_{n-1} r^{n-1} + \gamma_{n-2} r^{n-2} + \gamma_n A r^{n-1}, \\ r^{n-2} &= r^0 + \sum_{k=1}^{n-2} c_k A^k r^0, \\ r^{n-1} &= r^0 + \sum_{k=1}^{n-1} c_k A^k r^0, \\ A r^{n-1} &= \sum_{k=1}^{n} p_k A^k r^0 \quad \text{(sum begins at } k=1 \text{ since } (A r^{n-1}, r^0) = 0. \end{aligned}}$$

$$r^n = P_n(A) r^0 = (\gamma_{n-1} + \gamma_{n-2}) r^0 + \sum_{k=1}^{n} c_k A^k r^0$$

but $P_n(0) = 1$, hence $\gamma_{n-1} + \gamma_{n-2} = 1$. Define $\gamma_{n-1} - 1 = \alpha_{n-1}, \gamma_n = \beta_{n-1}$,

$$r^n = r^{n-1} + \alpha_{n-1}(r^{n-1} - r^{n-2}) + \beta_{n-1} A r^{n-1}.$$

$$\boxed{\begin{aligned} r^n = r^{n-1} + \alpha_{n-1}(r^{n-1} - r^{n-2}) + \beta_{n-1} A r^{n-1}, \\ (r^n, r^l) = 0 \end{aligned}}$$

implies that

$$\left.\begin{aligned} (1+\alpha_{n-1})\|r^{n-1}\|^2 + \beta_{n-1}\|r_{n-1}\|_A^2 = 0, \\ -\alpha_{n-1}\|r^{n-2}\|^2 + \beta_{n-1}(Ar^{n-1}, Ar^{n-2}) = 0. \end{aligned}\right\}$$

$$\boxed{\begin{aligned} r^n &= r^{n-1} + \alpha_{n-1}(r^{n-1} - r^{n-2}) + \beta_{n-1} A r^{n-1}, \\ (r^n, r^l) &= 0 \end{aligned}}$$

implies that

$$\left.\begin{aligned} (1+\alpha_{n-1})\|r^{n-1}\|^2 + \beta_{n-1}\|r_{n-1}\|_A^2 = 0, \\ -\alpha_{n-1}\|r^{n-2}\|^2 + \beta_{n-1}(Ar^{n-1}, Ar^{n-2}) = 0. \end{aligned}\right\}$$

Note that $r^j = Ax^j = b$ and multiply by A^{-1}, we have

$$x^n = x^{n-1} + \alpha_{n-1}(x^{n-1} - x^{n-2}) + \beta_{n-1}(Ax^{n-1} - b).$$

Convergence (Prove it!)

Theorem (Estimate of residual of c.g.)

$$\|x^n - X\|_A \leqslant \frac{2}{\lambda_0^n + \lambda_0^{-n}} \|x^0 - X\|_A,$$

where $\lambda_0 = \frac{\sqrt{M}+\sqrt{\mu}}{\sqrt{M}-\sqrt{\mu}}$, M *and* μ *are max and min eigenvalue.*

Algorithm 1: c.g. method

Input: x^0

Output: x^n

Compute $s_1 = r^0 = Ax^0 - b$;
for $k = 1,\, 2,\, \cdots,\, n$ **do**
    Compute $\alpha_k = (r^{k-1}, r^{k-1})/(As_k, s_k)$;
    Compute $r^k = r^{k-1} - \alpha_k A s_k$;
    Compute $x^k = x^{k-1} - \alpha_k s_k$;
    Compute $\beta_k = (r^k, r^k)/(r^{k-1}, r^{k-1})$;
    Compute $s_{k+1} = r^k + \beta_k s_k$
end

R.p.c.g. Method (Bai and Li, 2002)

Problem

Solve the system $Ax = b$ with the situation that $A = PHQ$, H is s.p.d and P, Q are nonsingular. Suppose $M = PGQ$ is a pre-conditioner to A, where G is s.p.d.

Since G is s.p.d, it can be represented as $G = S^T S$ where S is nonsingular.

R.p.c.g. Method (Bai and Li, 2002)

Problem

Solve the system $Ax = b$ with the situation that $A = PHQ$, H is s.p.d and P, Q are nonsingular. Suppose $M = PGQ$ is a pre-conditioner to A, where G is s.p.d.

Since G is s.p.d, it can be represented as $G = S^T S$ where S is nonsingular.

$$Ax = b \sim Mx = b \sim \underbrace{PS^T}_{b\mathbf{b}^{-1}} \underbrace{SQx}_{\mathbf{x}} = b.$$

R.p.c.g. Method (Bai and Li, 2002)

Problem

Solve the system $Ax = b$ with the situation that $A = PHQ$, H is s.p.d and P, Q are nonsingular. Suppose $M = PGQ$ is a pre-conditioner to A, where G is s.p.d.

Since G is s.p.d, it can be represented as $G = S^T S$ where S is nonsingular.

$$Ax = b \sim Mx = b \sim \underbrace{PS^T}_{b\mathbf{b}^{-1}} \underbrace{SQx}_{\mathbf{x}} = b.$$

Hence that's equivalent to

$$\begin{aligned} R\mathbf{x} &= \mathbf{b}, \\ R &= (PS^T)^{-1} A (SQ)^{-1} = S^{-T} H S^{-1}, \\ \mathbf{x} &= SQx, \\ \mathbf{b} &= (PS^T)^{-1} b. \end{aligned}$$

Algorithm 2: r.p.c.g. method

Input: x^0

Output: x^n

```
Solve Mz_0 = r_0, set p_0 := z_0;
Solve Wv_0 = z_0, set q_0 := v_0;
for k = 1, 2, ..., n do
    Compute α_k = (r^{k-1}, r^{k-1})/(Ap_k, q_k);
    Compute r^k = r^{k-1} - α_k Ap_k;
    Compute x^k = x^{k-1} - α_k p_k;
    Solve Mz_k = r_k;
    Solve Wv_k = z_k;
    Compute β_k = (v^k, r^k)/(v^{k-1}, r^{k-1});
    Compute p_{k+1} = z^k + β_k p_k;
    Compute q_{k+1} = v_k + β_k q_k
end
```

where $W = Q^{-1}P^T$.

Convergence

Theorem (Estimate of residual of r.p.c.g)

$$\|x^k - X\|_{W^{-T}A} \leqslant 2\left(\frac{\sqrt{\nu(M^{-1}A)}-1}{\sqrt{\nu(M^{-1}A)}+1}\right)^k \|x^0 - X\|_{W^{-T}A}.$$