AMS526: Numerical Analysis I
(Numerical Linear Algebra)
Lecture 22: Conjugate Gradient Method

Xiangmin Jiao

SUNY Stony Brook
Krylov Subspace Algorithms

- Create a sequence of Krylov subspaces for $Ax = b$
  
  \[ \mathcal{K}_n = \{ b, Ab, \ldots, A^{n-1}b \} \]

  and find an approximate (hopefully optimal) solutions $x_n$ in $\mathcal{K}_n$

- Only matrix-vector products involved

- For SPD matrices, most famous algorithm is Conjugate Gradient (CG) method discovered by Hestenes/Stiefel in 1952
  
  - Finds best solution $x_n \in \mathcal{K}_n$ in norm $\|x\|_A = \sqrt{x^T Ax}$
  - Only requires storing 4 vectors (instead of $n$ vectors) due to three-term recurrence
Motivation of Conjugate Gradients

- If $A$ is $m \times m$ SPD, then quadratic function
  \[ \varphi(x) = \frac{1}{2} x^T Ax - x^T b \]
  has unique minimum
- Negative gradient of this function is residual vector
  \[ -\nabla \varphi(x) = b - Ax = r \]
  so minimum is obtained precisely when $Ax = b$
- Optimization methods have form
  \[ x_{n+1} = x_n + \alpha p_n \]
  where $p_n$ is search direction and $\alpha$ is step length chosen to minimize
  $\varphi(x_n + \alpha p_n)$
- Line search parameter can be determined analytically as
  \[ \alpha = \frac{r_n^T p_n}{p_n^T Ap_n} \]
- In CG, $p_n$ is chosen to be A-conjugate (or A-orthogonal) to previous
  search directions, i.e., $p_n^T Ap_j = 0$ for $j < n$
Algorithm: Conjugate Gradient Method

\[ x_0 = 0, \; r_0 = b, \; p_0 = r_0 \]

\[ \text{for } n = 1, 2, 3, \ldots \]

\[ \alpha_n = \frac{(r_{n-1}^T r_{n-1})}{(p_{n-1}^T A p_{n-1})} \]  

step length

\[ x_n = x_{n-1} + \alpha_n p_{n-1} \]  

approximate solution

\[ r_n = r_{n-1} - \alpha_n A p_{n-1} \]  

residual

\[ \beta_n = \frac{(r_n^T r_n)}{(r_{n-1}^T r_{n-1})} \]  

improvement this step

\[ p_n = r_n + \beta_n p_{n-1} \]  

search direction

- Only one matrix-vector product \( Ap_{n-1} \) per iteration
- Apart from matrix-vector product, \#operations per iteration is \( O(m) \)
- If \( A \) is sparse with constant number of nonzeros per row, \( O(m) \) operations per iteration
- \( \text{CG can be viewed as minimization of quadratic function} \)
  \[ \varphi(x) = \frac{1}{2} x^T A x - x^T b \]  
  by modifying steepest descent
An Alternative Interpretation of CG

Algorithm: CG

\[ x_0 = 0, \quad r_0 = b, \quad p_0 = r_0 \]

\[
\text{for } n = 1 \text{ to } 1, 2, 3, \ldots \\
\alpha_n = \frac{r_{n-1}^T r_{n-1}}{(p_{n-1}^T A p_{n-1})} \\
x_n = x_{n-1} + \alpha_n p_{n-1} \\
r_n = r_{n-1} - \alpha_n A p_{n-1} \\
\beta_n = \frac{r_n^T r_n}{(r_{n-1}^T r_{n-1})} \\
p_n = r_n + \beta_n p_{n-1}
\]

Algorithm: A non-standard CG

\[ x_0 = 0, \quad r_0 = b, \quad p_0 = r_0 \]

\[
\text{for } n = 1 \text{ to } 1, 2, 3, \ldots \\
\alpha_n = \frac{r_{n-1}^T p_{n-1}}{(p_{n-1}^T A p_{n-1})} \\
x_n = x_{n-1} + \alpha_n p_{n-1} \\
r_n = b - A x_n \\
\beta_n = \frac{-r_n^T A p_{n-1}}{(p_{n-1}^T A p_{n-1})} \\
p_n = r_n + \beta_n p_{n-1}
\]

- The non-standard one is less efficient but easier to understand
- It is easy to see \( r_n = r_{n-1} - \alpha_n A p_{n-1} = b - A x_n \)
- We need to show:
  - \( \alpha_n \) minimizes \( \varphi \) along search direction \( p_n \)
  - \( \alpha_n \) and \( \beta_n \) are equivalent to those in standard CG
  - Minimizing \( \varphi \) along \( p_n \) also minimizes \( \varphi \) within Krylov subspace
Optimality of Step Length

- Select step length $\alpha_n$ over vector $p_{n-1}$ to minimize
  $\varphi(x) = \frac{1}{2}x^T Ax - x^T b$
- Let $x_n = x_{n-1} + \alpha_n p_{n-1}$,

$$\varphi(x_n) = \frac{1}{2}(x_{n-1} + \alpha_n p_{n-1})^T A(x_{n-1} + \alpha_n p_{n-1}) - (x_{n-1} + \alpha_n p_{n-1})^T b$$

$$= \frac{1}{2} \alpha_n^2 p_{n-1}^T Ap_{n-1} + \alpha_n p_{n-1}^T Ax_{n-1} - \alpha_n p_{n-1}^T b + \text{constant}$$

$$= \frac{1}{2} \alpha_n^2 p_{n-1}^T Ap_{n-1} - \alpha_n p_{n-1}^T r_{n-1} + \text{constant}$$

- Therefore,

$$\frac{d\varphi}{d\alpha_n} = 0 \Rightarrow \alpha_n p_{n-1}^T Ap_{n-1} - p_{n-1}^T r_{n-1} = 0 \Rightarrow \alpha_n = \frac{p_{n-1}^T r_{n-1}}{p_{n-1}^T Ap_{n-1}}.$$ 

- In addition, $p_{n-1}^T r_{n-1} = r_{n-1}^T r_{n-1}$ because $p_{n-1} = r_{n-1} + \beta_n p_{n-2}$ and $r_{n-1}^T p_{n-2} = 0$ due to the following theorem.
Properties of Conjugate Gradients

**Theorem (38.1)**

If $r_{n-1} \neq 0$, spaces spanned by approximate solutions $x_n$, search directions $p_n$, and residuals $r_n$ are all equal to Krylov subspaces

$$K_n = \langle x_1, x_2, \ldots, x_n \rangle = \langle p_0, p_1, \ldots, p_{n-1} \rangle$$
$$= \langle r_0, r_1, \ldots, r_{n-1} \rangle = \langle b, Ab, \ldots, A^{n-1}b \rangle$$

The residuals are orthogonal (i.e., $r_T^nr_j = 0$ for $j < n$) and search directions are $A$-conjugate (i.e, $p_T^nap_j = 0$ for $j < n$).

This theorem implies that

$$\alpha_n = (r_{n-1}^Tr_{n-1})/(p_{n-1}^TAp_{n-1}) = r_{n-1}^Tp_{n-1}/(p_{n-1}^TAp_{n-1})$$

and

$$\beta_n = \frac{r_T^nr_n}{r_{n-1}^Tr_{n-1}} = \frac{r_T^n (r_{n-1} - \alpha_nAp_{n-1})}{r_{n-1}^Tr_{n-1}} = -\frac{r_T^nap_{n-1}}{p_{n-1}^TAp_{n-1}}.$$
Proof of Theorem 38.1

Prove based on notation of standard CG.

- Proof of equality of subspaces by simple induction.
- To prove $r_n^T r_j = 0$, note that $r_n = r_{n-1} - \alpha_n A p_{n-1}$ and
  $(A p_{n-1})^T = p_{n-1}^T A$, so

  $$r_n^T r_j = (r_{n-1} - \alpha_n A p_{n-1})^T r_j = r_{n-1}^T r_j - \alpha_n p_{n-1}^T A r_j.$$  

  ▶ If $j < n - 1$, then both terms on right are zero by induction.
  ▶ If $j = n - 1$, plug in $\alpha_n = (r_{n-1}^T r_{n-1})/(p_{n-1}^T A p_{n-1})$

  $$r_{n-1}^T r_j - \alpha_n p_{n-1}^T A r_j = r_{n-1}^T r_{n-1} - r_{n-1}^T r_{n-1} \frac{p_{n-1}^T A r_{n-1}}{p_{n-1}^T A p_{n-1}},$$

  which is zero because

  $$p_{n-1}^T A p_{n-1} = p_{n-1}^T A(r_{n-1} + \beta_n p_{n-2}) = p_{n-1}^T A r_{n-1}$$

  by induction hypothesis.
Proof of Theorem 38.1 Cont’d

To prove $\mathbf{p}_n^T \mathbf{A} \mathbf{p}_j = 0$, note that $\mathbf{p}_n = \mathbf{r}_n + \beta_n \mathbf{p}_{n-1}$, so

$$
\mathbf{p}_n^T \mathbf{A} \mathbf{p}_j = \mathbf{r}_n^T \mathbf{A} \mathbf{p}_j + \beta_n \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_j.
$$

▶ If $j < n - 1$, then both terms on right are zero by induction.
▶ If $j = n - 1$, plug in $\beta_n = (\mathbf{r}_n^T \mathbf{r}_n)/(\mathbf{r}_{n-1}^T \mathbf{r}_{n-1})$,

$$
\mathbf{r}_n^T \mathbf{A} \mathbf{p}_j + \beta_n \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_j = \mathbf{r}_n^T \mathbf{A} \mathbf{p}_{n-1} + \frac{1}{\alpha_n} \mathbf{r}_n^T \mathbf{r}_n
$$

$$
= \frac{1}{\alpha_n} \mathbf{r}_n^T (\mathbf{r}_n + \alpha_n \mathbf{A} \mathbf{p}_{n-1})
$$

$$
= \frac{1}{\alpha_n} \mathbf{r}_n^T \mathbf{r}_{n-1}
$$

$$
= 0.
$$
Optimality of Conjugate Gradients

Theorem (38.2)

If \( r_{n-1} \neq 0 \), then error \( e_n = x_* - x_n \) are minimized in A-norm in \( \mathcal{K}_n \).

Proof.

Consider arbitrary point \( x = x_n - \Delta x \in \mathcal{K}_n \) with error \( e = x_* - x = e_n + \Delta x \). So

\[
\|e\|_A^2 = (e_n + \Delta x)^T A (e_n + \Delta x) \\
= e_n^T A e_n + \Delta x^T A \Delta x + 2 e_n^T A \Delta x,
\]

where \( e_n^T A \Delta x = r_n^T \Delta x = 0 \) because \( r_n \perp \mathcal{K}_n \). Since \( A \) is SPD, \( \|e\|_A^2 \geq \|e_n\|_A^2 \) and equality holds iff \( \Delta x = 0 \).

- Because \( \mathcal{K}_n \) grows monotonically, error decreases monotonically.
Rate of Convergence

- In addition, CG can be studied in terms of polynomial approximation
  - It finds optimal polynomial \( p_n \in P_n \) of degree \( n \) with \( p(0) = 1 \), minimizing \( \|p_n(A)e_0\|_A \) with initial error \( e_0 = x_* \)
  - Convergence results can be obtained from this polynomial approximation

- Some important convergence results
  - If \( A \) has \( n \) distinct eigenvalues, CG converges in at most \( n \) steps
  - If \( A \) has 2-norm condition number \( \kappa \), the errors are
    \[
    \frac{\|e_n\|_A}{\|e_0\|_A} \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n
    \]
    which is \( \approx 2 \left( 1 - \frac{2}{\sqrt{\kappa}} \right)^n \) as \( \kappa \to \infty \). So convergence is expected in \( O(\sqrt{\kappa}) \) iterations.

- In general, CG performs well with clustered eigenvalues