

AMS526: Numerical Analysis I (Numerical Linear Algebra)

Lecture 24: More on Conjugate Gradient Methods

Xiangmin Jiao

SUNY Stony Brook

December 9, 2008

Conjugate Gradient Method

Algorithm: Conjugate Gradient Method

$$\mathbf{x}_0 = \mathbf{0}, \mathbf{r}_0 = \mathbf{b}, \mathbf{p}_0 = \mathbf{r}_0$$

for $n = 1$ to $1, 2, 3, \dots$

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

step length

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1}$$

approximate solution

$$\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n \mathbf{A} \mathbf{p}_{n-1}$$

residual

$$\beta_n = (\mathbf{r}_n^T \mathbf{r}_n) / (\mathbf{r}_{n-1}^T \mathbf{r}_{n-1})$$

improvement this step

$$\mathbf{p}_n = \mathbf{r}_n + \beta_n \mathbf{p}_{n-1}$$

search direction

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

An Alternative Interpretation of CG

Algorithm: CG

$$\mathbf{x}_0 = 0, \mathbf{r}_0 = \mathbf{b}, \mathbf{p}_0 = \mathbf{r}_0$$

for $n = 1$ to $1, 2, 3, \dots$

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

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1}$$

$$\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n \mathbf{A} \mathbf{p}_{n-1}$$

$$\beta_n = \mathbf{r}_n^T \mathbf{r}_n / (\mathbf{r}_{n-1}^T \mathbf{r}_{n-1})$$

$$\mathbf{p}_n = \mathbf{r}_n + \beta_n \mathbf{p}_{n-1}$$

Algorithm: A non-standard CG

$$\mathbf{x}_0 = 0, \mathbf{r}_0 = \mathbf{b}, \mathbf{p}_0 = \mathbf{r}_0$$

for $n = 1$ to $1, 2, 3, \dots$

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

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1}$$

$$\mathbf{r}_n = \mathbf{b} - \mathbf{A} \mathbf{x}_n$$

$$\beta_n = -\mathbf{r}_n^T \mathbf{A} \mathbf{p}_{n-1} / (\mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_{n-1})$$

$$\mathbf{p}_n = \mathbf{r}_n + \beta_n \mathbf{p}_{n-1}$$

- The non-standard one is less efficient but easier to understand
- It is easy to see $\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n \mathbf{A} \mathbf{p}_{n-1} = \mathbf{b} - \mathbf{A} \mathbf{x}_n$
- We need to show:
 - ▶ α_n minimizes φ along search direction \mathbf{p}_n
 - ▶ α_n and β_n are equivalent to those in standard CG
 - ▶ Minimizing φ along \mathbf{p}_n also minimizes φ within Krylov subspace

Optimality of Step Length

- Select step length α_n over vector \mathbf{p}_{n-1} to minimize

$$\varphi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} - \mathbf{x}^T \mathbf{b}$$

- Let $\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1}$,

$$\begin{aligned}\varphi(\mathbf{x}_n) &= \frac{1}{2}(\mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1})^T \mathbf{A}(\mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1}) - (\mathbf{x}_{n-1} + \alpha_n \mathbf{p}_{n-1})^T \mathbf{b} \\ &= \frac{1}{2}\alpha_n^2 \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_{n-1} + \alpha_n \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{x}_{n-1} - \alpha_n \mathbf{p}_{n-1}^T \mathbf{b} + \text{constant} \\ &= \frac{1}{2}\alpha_n^2 \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_{n-1} - \alpha_n \mathbf{p}_{n-1}^T \mathbf{r}_{n-1} + \text{constant}\end{aligned}$$

- Therefore,

$$\frac{d\varphi}{d\alpha_n} = 0 \Rightarrow \alpha_n \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_{n-1} - \mathbf{p}_{n-1}^T \mathbf{r}_{n-1} = 0 \Rightarrow \alpha_n = \frac{\mathbf{p}_{n-1}^T \mathbf{r}_{n-1}}{\mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_{n-1}}.$$

- In addition, $\mathbf{p}_{n-1}^T \mathbf{r}_{n-1} = \mathbf{r}_{n-1}^T \mathbf{r}_{n-1}$ because $\mathbf{p}_{n-1} = \mathbf{r}_{n-1} + \beta_n \mathbf{p}_{n-2}$ and $\mathbf{r}_{n-1}^T \mathbf{p}_{n-2} = 0$ due to the following theorem.

Properties of Conjugate Gradients

Theorem (38.1)

If $\mathbf{r}_{n-1} \neq \mathbf{0}$, spaces spanned by approximate solutions \mathbf{x}_n , search directions \mathbf{p}_n , and residuals \mathbf{r}_n are all equal to Krylov subspaces

$$\begin{aligned}\mathcal{K}_n &= \langle \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \rangle = \langle \mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_{n-1} \rangle \\ &= \langle \mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{n-1} \rangle = \langle \mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{n-1}\mathbf{b} \rangle\end{aligned}$$

The residuals are orthogonal (i.e., $\mathbf{r}_n^T \mathbf{r}_j = 0$ for $j < n$) and search directions are \mathbf{A} -conjugate (i.e., $\mathbf{p}_n^T \mathbf{A}\mathbf{p}_j = 0$ for $j < n$).

This theorem implies that

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

and

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

Proof of Theorem 38.1

Prove based on notation of standard CG.

- Proof of equality of subspaces by simple induction.
- To prove $\mathbf{r}_n^T \mathbf{r}_j = 0$, note that $\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_n \mathbf{A} \mathbf{p}_{n-1}$ and $(\mathbf{A} \mathbf{p}_{n-1})^T = \mathbf{p}_{n-1}^T \mathbf{A}$, so

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

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

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

which is zero because

$$\mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_{n-1} = \mathbf{p}_{n-1}^T \mathbf{A} (\mathbf{r}_{n-1} + \beta_n \mathbf{p}_{n-2}) = \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{r}_{n-1}$$

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 $\beta_n = (\mathbf{r}_n^T \mathbf{r}_n) / (\mathbf{r}_{n-1}^T \mathbf{r}_{n-1})$,

$$\begin{aligned} \mathbf{r}_n^T \mathbf{A} \mathbf{p}_j + \beta_n \mathbf{p}_{n-1}^T \mathbf{A} \mathbf{p}_j &= \mathbf{r}_n^T \mathbf{A} \mathbf{p}_{n-1} + \frac{1}{\alpha_n} \mathbf{r}_n^T \mathbf{r}_n \\ &= \frac{1}{\alpha_n} \mathbf{r}_n^T (\mathbf{r}_n + \alpha_n \mathbf{A} \mathbf{p}_{n-1}) \\ &= \frac{1}{\alpha_n} \mathbf{r}_n^T \mathbf{r}_{n-1} \\ &= 0. \end{aligned}$$

Optimality of Conjugate Gradients

Theorem (38.2)

If $\mathbf{r}_{n-1} \neq \mathbf{0}$, then error $\mathbf{e}_n = \mathbf{x}_* - \mathbf{x}_n$ are minimized in \mathbf{A} -norm in \mathcal{K}_n .

Proof.

Consider arbitrary point $\mathbf{x} = \mathbf{x}_n - \Delta\mathbf{x} \in \mathcal{K}_n$ with error $\mathbf{e} = \mathbf{x}_* - \mathbf{x} = \mathbf{e}_n + \Delta\mathbf{x}$. So

$$\begin{aligned}\|\mathbf{e}\|_{\mathbf{A}}^2 &= (\mathbf{e}_n + \Delta\mathbf{x})^T \mathbf{A} (\mathbf{e}_n + \Delta\mathbf{x}) \\ &= \mathbf{e}_n^T \mathbf{A} \mathbf{e}_n + \Delta\mathbf{x}^T \mathbf{A} \Delta\mathbf{x} + 2\mathbf{e}_n^T \mathbf{A} \Delta\mathbf{x},\end{aligned}$$

where $\mathbf{e}_n^T \mathbf{A} \Delta\mathbf{x} = \mathbf{r}_n^T \Delta\mathbf{x} = 0$ because $\mathbf{r}_n \perp \mathcal{K}_n$. Since \mathbf{A} is SPD, $\|\mathbf{e}\|_{\mathbf{A}}^2 \geq \|\mathbf{e}_n\|_{\mathbf{A}}^2$ and equality holds iff $\Delta\mathbf{x} = \mathbf{0}$. □

- Because \mathcal{K}_n grows monotonically, error decreases monotonically.

Rate of Convergence

- In addition, CG can be studied in terms of polynomial approximation
 - ▶ It find optimal polynomial $p_n \in P_n$ of degree n with $p(0) = 1$, minimizing $\|p_n(\mathbf{A})\mathbf{e}_0\|_{\mathbf{A}}$ with initial error $\mathbf{e}_0 = \mathbf{x}_*$
 - ▶ Convergence results can be obtained from this polynomial approximation
- Some important convergence results
 - ▶ If \mathbf{A} has n distinct eigenvalues, CG converges in at most n steps
 - ▶ If \mathbf{A} has 2-norm condition number κ , the errors are

$$\frac{\|\mathbf{e}_n\|_{\mathbf{A}}}{\|\mathbf{e}_0\|_{\mathbf{A}}} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n$$

which is $\approx 2 \left(1 - \frac{2}{\sqrt{\kappa}}\right)^n$ as $\kappa \rightarrow \infty$. So convergence is expected in $O(\sqrt{\kappa})$ iterations.

- In general, CG performs well with clustered eigenvalues