Lecture 22
Conjugate Gradient Method

Songting Luo

Department of Mathematics
Iowa State University

MATH 562 Numerical Analysis II

Outline

Outline

Krylov Subspace Algorithms

- Create a sequence of Krylov subspaces for $\mathbf{Ax} = \mathbf{b}$

$$\mathcal{K}_n = \{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{n-1}\mathbf{b}\}$$

  and find an approximate (hopefully optimal) solutions $\mathbf{x}_n \in \mathcal{K}_n$
- Only matrix-vector products involved
- For SPD matrices, most famous algorithm is Conjugate Gradient (CG) method discovered by Hestenes/Stiefel in 1952
  - Find best solution $\mathbf{x}_n \in \mathcal{K}_n$ in norm $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T\mathbf{Ax}}$.
  - Only requires storing 4 vectors (instead of n vectors) due to three-term recurrence

Motivation of Conjugate Gradients

- If $\mathbf{A}$ is $m \times m$ SPD, then quadratic function

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

  has unique minimum
- Negative gradient of this function is residual vector

$$-\nabla\phi(\mathbf{x}) = \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{r}$$

  so minimum is obtained precisely when $\mathbf{A}\mathbf{x} = \mathbf{b}$
- Optimization methods have form

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

  where $\mathbf{p}_n$ is search direction and α is step length chosen to minimize $\phi(\mathbf{x}_n + \alpha\mathbf{p}_n)$
- Line search parameter can be determined analytically as $\alpha = \mathbf{r}_n^T\mathbf{p}_n / \mathbf{p}_n^T\mathbf{A}\mathbf{p}_n$
- In CG, $\mathbf{p}_n$ is chosen to be $\mathbf{A}$-conjugate (or $\mathbf{A}$-orthogonal) to previous search directions, i.e., $\mathbf{p}_n^T\mathbf{A}\mathbf{p}_j = 0$ for $j < n$.

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 $1, 2, 3, \ldots$

$\alpha_n = (\mathbf{r}_{n-1}^T \mathbf{r}_{n-1})/(\mathbf{p}_{n-1}^T \mathbf{A}\mathbf{p}_{n-1})$	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, #flops per iteration is $O(m^2)$
- 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 $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} - \mathbf{x}^T \mathbf{b}$ by modifying steepest descent

An Alternative Interpretation of CG

CG

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

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

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

A non-standard CG

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

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

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

- 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 $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$
- Let $\mathbf{x}_n = x_{n-1} + \alpha_{n-1}\mathbf{p}_{n-1}$,

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

- Therefore,

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

- 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

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*

$$\mathcal{K}_n = \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle = \langle \mathbf{p}_0, \ldots, \mathbf{p}_{n-1} \rangle = \langle \mathbf{r}_0, \ldots, \mathbf{r}_{n-1} \rangle$$

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

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 in 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 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

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

$$\|\mathbf{e}\|_{\mathbf{A}}^2 = (\mathbf{e}_n + \Delta\mathbf{x})^T \mathbf{A}(\mathbf{e}_n + \Delta\mathbf{x}) = \mathbf{e}_n^T \mathbf{A}\mathbf{e}_n + \Delta\mathbf{x}^T \mathbf{A}\Delta\mathbf{x} + 2\mathbf{e}_n^T \mathbf{A}\Delta\mathbf{x}$$

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, $\|\mathbf{e}\|_{\mathbf{A}}^2 \geqslant \|\mathbf{e}_n\|_{\mathbf{A}}^2$ 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 finds optimal polynomial $p_n \in \mathbf{P}_n$ 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

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

- In general, CG performs well with clustered eigenvalues