

## Lecture 19 The Conjugate Gradients Algorithm I

MIT 18.335J / 6.337J  
Introduction to Numerical Methods

Per-Olof Persson  
November 16, 2006

1

## Krylov Subspace Algorithms

- Create a sequence of *Krylov subspaces* for  $Ax = b$ :

$$\mathcal{K}_n = \langle b, Ab, \dots, A^{n-1}b \rangle$$

and find approximate solutions  $x_n \in \mathcal{K}_n$

- Only matrix-vector products involved
- For SPD matrices, the most popular algorithm is the *Conjugate Gradients* method [Hestenes/Stiefel, 1952]
  - Finds the best solution  $x_n \in \mathcal{K}_n$  in the norm  $\|x\|_A = \sqrt{x^T A x}$
  - Only requires storage of 4 vectors (not all the  $n$  vectors in  $\mathcal{K}_n$ )
  - Remarkably simple and excellent convergence properties
  - Originally invented as a direct algorithm! (converges after  $m$  steps in exact arithmetic)

2

## The Conjugate Gradients Method

### Algorithm: Conjugate Gradients Method

$$x_0 = 0, r_0 = b, p_0 = r_0$$

for  $k = 1, 2, 3, \dots$

$\alpha_n = (r_{n-1}^T r_{n-1}) / (p_{n-1}^T A p_{n-1})$	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
$\beta_n = (r_n^T r_n) / (r_{n-1}^T r_{n-1})$	improvement this step
$p_n = r_n + \beta_n p_{n-1}$	search direction

- Only one matrix-vector product  $A p_{n-1}$  per iteration
- Operation count  $O(m)$  (excluding the matrix-vector product)

3

## Properties of Conjugate Gradients Vectors

- The spaces spanned by the solutions, the search directions, and the residuals are all equal to the Krylov subspaces:

$$\begin{aligned} \mathcal{K}_n &= \langle x_1, x_2, \dots, x_n \rangle = \langle p_0, p_1, \dots, p_{n-1} \rangle \\ &= \langle r_0, r_1, \dots, r_{n-1} \rangle = \langle b, Ab, \dots, A^{n-1}b \rangle \end{aligned}$$

- The residuals are orthogonal:

$$r_n^T r_j = 0 \quad (j < n)$$

- The search directions are A-conjugate:

$$p_n^T A p_j = 0 \quad (j < n)$$

*Proofs.* Textbook/black-board

4

## Optimality of Conjugate Gradients

- The errors  $e_n = x_* - x_n$  are minimized in the  $A$ -norm

*Proof.* For any other point  $x = x_n - \Delta x \in \mathcal{K}_n$  the error is

$$\begin{aligned} \|e\|_A^2 &= (e_n + \Delta x)^T A (e_n + \Delta x) \\ &= e_n^T A e_n + (\Delta x)^T A (\Delta x) + 2e_n^T A (\Delta x) \end{aligned}$$

But  $e_n^T A (\Delta x) = r_n^T (\Delta x) = 0$ , since  $r_n$  is orthogonal to  $\mathcal{K}_n$ , so  $\Delta x = 0$  minimizes  $\|e\|_A$

- Monotonic:  $\|e_n\|_A \leq \|e_{n-1}\|_A$ , and  $e_n = 0$  in  $n \leq m$  steps

*Proof.* Follows from  $\mathcal{K}_n \subseteq \mathcal{K}_{n+1}$ , and that  $\mathcal{K}_n \subseteq \mathbb{R}^m$  unless converged

5

## Optimization in CG

- CG can be interpreted as a minimization algorithm
- We know it minimizes  $\|e\|_A$ , but this cannot be evaluated
- CG also minimizes the quadratic function  $\varphi(x) = \frac{1}{2}x^T A x - x^T b$ :

$$\begin{aligned} \|e_n\|_A^2 &= e_n^T A e_n = (x_* - x_n)^T A (x_* - x_n) \\ &= x_n^T A x_n - 2x_n^T A x_* + x_*^T A x_* \\ &= x_n^T A x_n - 2x_n^T b + x_*^T b = 2\varphi(x_n) + \text{constant} \end{aligned}$$

- At each step  $\alpha_n$  is chosen to minimize  $x_n = x_{n-1} + \alpha_n p_{n-1}$
- The conjugated search directions  $p_n$  give minimization over all of  $\mathcal{K}_n$

6

## Polynomial Approximation by CG

- Conjugate Gradients finds an optimal polynomial  $p_n \in P_n$  of degree  $n$  with  $p(0) = 1$ , minimizing  $\|p_n(A)e_0\|$  with initial error  $e_0 = x_*$
- More specifically, with  $\Lambda(A)$  being the spectrum of  $A$ :

$$\frac{\|e_n\|_A}{\|e_0\|_A} = \inf_{p \in P_n} \frac{\|p(A)e_0\|_A}{\|e_0\|_A} \leq \inf_{p \in P_n} \max_{\lambda \in \Lambda(A)} |p(\lambda)|$$

*Proof.* It is clear that  $x_n = q_n(A)b = q_n(A)Ax_*$  with  $q_n$  degree  $n - 1$ . Then  $e_n = p_n(A)e_0$  with  $p_n \in P_n$ . The equality above then follows since CG minimizes  $\|e_n\|_A$ . For the inequality, expand in eigenvectors of  $A$ :

$$e_0 = \sum a_j v_j, \quad p(A)e_0 = \sum a_j p(\lambda_j) v_j$$

Then  $\|e_0\|_A^2 = \sum_j a_j^2 \lambda_j$  and  $\|p(A)e_0\|_A^2 = \sum_j a_j^2 \lambda_j (p(\lambda_j))^2$ , which implies the inequality.

7

## Rate of Convergence

- Important convergence results can be obtained from the polynomial approximation:
  1. If  $A$  has  $n$  distinct eigenvalues, CG converges in at most  $n$  steps  
*Proof.* The polynomial  $p(x) = \prod_{j=1}^n (1 - x/\lambda_j)$  is zero at  $\Lambda(A)$
  2. If  $A$  has 2-norm condition number  $\kappa$ , the errors are

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n \approx 2 \left( 1 - \frac{2}{\sqrt{\kappa}} \right)^n$$

*Proof.* Textbook

- In general: CG performs well with clustered eigenvalues

8