Conjugate gradients

Madhu Belur, CC, EE, IITB

15th April, 2011

Plan for conjugate gradients

- Steepest descent method: current search direction is not necessarily independent from past ones.
- One search direction per iteration: when not necessarily steepest descent
- Minimizing over multiple search directions : choose p_k and minimize over both y, α ($y \in \mathbb{R}^{k-1}$ and $\alpha \in \mathbb{R}$)
- If directions are A-orthogonal: minimization problem 'decouples'. ('Conjugate gradients')
- y found so far (upto $(k-1)^{\text{th}}$ step) is already optimal.
- There exists such a new search direction (with A-orthogonal to past search directions and $p_k^T r_{k-1} \neq 0$).

Steepest descent

Recall that $\phi(x) := \frac{1}{2}x^T Ax - x^T b$

Steepest descent

Recall that $\phi(x) := \frac{1}{2}x^T Ax - x^T b$
Choose $x_0 \in \mathbb{R}^{\mathtt{n}}$. Residue $r_0 := b - Ax_0$.

Steepest descent

Recall that $\phi(x) := \frac{1}{2}x^T Ax - x^T b$
Choose $x_0 \in \mathbb{R}^{\mathtt{n}}$. Residue $r_0 := b - Ax_0$.
Then, direction of steepest descent: $-\nabla\phi(x_{k-1}) = r_{k-1}$

Steepest descent

Recall that $\phi(x) := \frac{1}{2}x^T Ax - x^T b$
Choose $x_0 \in \mathbb{R}^{\mathtt{n}}$. Residue $r_0 := b - Ax_0$.
Then, direction of steepest descent: $-\nabla\phi(x_{k-1}) = r_{k-1}$
Can define: $p_k := -\nabla\phi(x_{k-1})$ (steepest descent method)

Steepest descent

Recall that $\phi(x) := \frac{1}{2}x^T Ax - x^T b$
Choose $x_0 \in \mathbb{R}^{\mathtt{n}}$. Residue $r_0 := b - Ax_0$.
Then, direction of steepest descent: $-\nabla\phi(x_{k-1}) = r_{k-1}$
Can define: $p_k := -\nabla\phi(x_{k-1})$ (steepest descent method)

How much along p_k

A direction p_k will help decrease of ϕ iff $p_k^T r_{k-1} \neq 0$.

How much along p_k

A direction p_k will help decrease of ϕ iff $p_k^T r_{k-1} \neq 0$.
Optimal α along p_k can then be found.

How much along p_k

A direction p_k will help decrease of ϕ iff $p_k^T r_{k-1} \neq 0$.
Optimal α along p_k can then be found.
Update x by $x_k := x_{k-1} + \alpha p_k$: best $\alpha := \frac{p_k^T r_{k-1}}{p_k^T A p_k}$ (also when p_k not steepest descent)
Can show that

$$\phi(x_k) = \phi(x_{k-1}) - \frac{(p_k^T r_{k-1})^2}{2 p_k^T A p_k}$$

Optimize over all k directions

Choose p_k, to have totally k directions now.

Optimize over all k directions

Choose p_k, to have totally k directions now.
Re-optimize over all k independent directions in k-th step.

Optimize over all k directions

Choose p_k, to have totally k directions now.
Re-optimize over all k independent directions in k-th step.

$$\min_{x \in x_0 + \text{span}\{p_1, \ldots, p_k\}} \phi(x)$$

$x = x_0 + P_{k-1}y + \alpha p_k$ (with $y \in \mathbb{R}^{k-1}$ and $\alpha \in \mathbb{R}$).

Optimize over all k directions

Choose p_k, to have totally k directions now.
Re-optimize over all k independent directions in k-th step.

$$\min_{x \in x_0 + \text{span}\{p_1, \ldots, p_k\}} \phi(x)$$

$x = x_0 + P_{k-1}y + \alpha p_k$ (with $y \in \mathbb{R}^{k-1}$ and $\alpha \in \mathbb{R}$).
Open brackets of $\phi(x_k)$ to get

Optimize over all k directions

Choose p_k, to have totally k directions now.
Re-optimize over all k independent directions in k-th step.

$$\min_{x \in x_0 + \text{span}\{p_1, \ldots, p_k\}} \phi(x)$$

$x = x_0 + P_{k-1}y + \alpha p_k$ (with $y \in \mathbb{R}^{k-1}$ and $\alpha \in \mathbb{R}$).
Open brackets of $\phi(x_k)$ to get

$$\phi(x_k) = \phi(x_0 + P_{k-1}y) + \underline{\alpha p_k^T A P_{k-1} y} + \alpha^2 p_k^T A p_k - \alpha p_k^T r_0$$

Choose p_k to be $A-$orthogonal to $p_1, \ldots, p_{k-1}$

Optimize over all k directions

Choose p_k, to have totally k directions now.
Re-optimize over all k independent directions in k-th step.

$$\min_{x \in x_0 + \text{span}\{p_1, \ldots, p_k\}} \phi(x)$$

$x = x_0 + P_{k-1} y + \alpha p_k$ (with $y \in \mathbb{R}^{k-1}$ and $\alpha \in \mathbb{R}$).
Open brackets of $\phi(x_k)$ to get

$$\phi(x_k) = \phi(x_0 + P_{k-1} y) + \underline{\alpha p_k^T A P_{k-1} y} + \alpha^2 p_k^T A p_k - \alpha p_k^T r_0$$

Choose p_k to be $A-$orthogonal to $p_1, \ldots, p_{k-1}$
(i.e. $p_k^T A P_{k-1} = 0$)

$$\min_{y,\alpha} \phi(x_k) = \min_{y \in \mathbb{R}^{k-1}} \phi(x_{k-1}) + \min_{\alpha} (\frac{\alpha^2}{2} p_k^T A p_k - \alpha p_k^T r_0)$$

y already solves minimization problem in previous step (and p_k is 'decoupled' from previous y).

$$\min_{y,\alpha} \phi(x_k) = \min_{y \in \mathbb{R}^{k-1}} \phi(x_{k-1}) + \min_{\alpha} \left(\frac{\alpha^2}{2} p_k^T A p_k - \alpha p_k^T r_0\right)$$

y already solves minimization problem in previous step (and p_k is 'decoupled' from previous y).
Best

$$\alpha = \frac{p_k^T r_0}{p_k^T A p_k}$$

Note: $p_k^T r_0 = p_k^T r_{k-1}$ (shown before)

Algorithm

Choose p_k such that it is $A-$orthogonal to all past p_i and such that $p_k^T r_{k-1} \neq 0$.

Algorithm

Choose p_k such that it is $A-$orthogonal to all past p_i and such that $p_k^T r_{k-1} \neq 0$.
Lemma in Golub & van Loan :
$r_{k-1} \neq 0 \Rightarrow$ there exists such a p_k.

Optimal choice of p_k

- Which new search direction (amongst various A-orthogonal directions)
- Best direction (direction that minimizes, and still is A-orthogonal) is, in fact, easy to implement.
  If best search direction didn't have to be A-orthogonal, then steepest descent ($-\nabla\phi(x_{k-1})$ is best direction to choose for p_k.)
- The $A-$orthogonal and best direction p_k happens to least norm residual for a related minization problem.
- Further, p_k is linearly dependent on only p_{k-1} and r_{k-1}.
- In fact, $r_0, \ldots, r_{n-1}$ are Lanczos vectors and they 'tridiagonalize' A. (Eigenvalues of A can also be estimated.)

Lemma 10.2.2
For $k \geqslant 2$, the vectors p_k generated by the best A-orthogonal direction algorithm satisfy

$$p_k = r_{k-1} - AP_{k-1}z_{k-1}$$

where z_{k-1} solves

$$\min_{z \in \mathbb{R}^{k-1}} \|r_{k-1} - AP_{k-1}z\|_2.$$

(We used fact that for a tall, f.c.r. matrix B, the matrices BB^+ and $(I - BB^+)$ are orthogonal projections onto range B and $(\text{range } B)^\perp$ respectively.)

Theorem 10.2.3

The sequence of r_k and p_k satisfy

1. $r_k = r_{k-1} - \alpha_k A p_k$
2. $r_k^T P_k = 0$
3. span $\{p_1, \ldots, p_k\}$ = span $\{r_0, \ldots, r_{k-1}\} = K(r_0, A, k)$
4. residuals $r_0, \ldots, r_k$ are mutually orthogonal.

Corollary 10.2.4

Residuals and search directions satisfy

$$p_k \in \operatorname{span}\{p_{k-1}, r_{k-1}\}$$

for $k \geqslant 2$.

Lanczos connection

1. Krylov subspace generated from a vector and a symmetric matrix A.

Lanczos connection

1. Krylov subspace generated from a vector and a symmetric matrix A.
2. Convergence in at most n steps.
3. If the number of distinct eigenvalues of A is n_a, then convergence in at most k steps.
   (Steepest descent for identity matrix: one step.)
4. The largest and smallest eigenvalue of A get estimated accurately much before n.
   (Use of Chebyshev polynomials.)