Lecture-9

Conjugate Direction Algorithm
(Solution of Linear System or
Minimization of A Quadratic
Function)

Conjugate Gradient

- Linear conjugate gradient: for solving linear systems Ax=b with PD matrix, A.
  - Exact solution in *n steps* (Hestenes & Stiefel, 1950s)
  - Approximate solution in fewer than *n steps*
- Non-linear conjugate gradient: for solving large-scale non-linear optimization problems.
  - Fletcher and Reeves, 1964
  - Polk-Ribiere, 1969

Conjugate Gradient

$Ax = b$ A is symmetric PD. (1)

Or minimize the following function:

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

$\nabla \phi(x) = Ax - b = r(x)$ $r(x)$ is the residual

$S = \{p_0, p_1, \ldots, p_{n-1}\}$ The set S is conjugate wrt A if

$$p_i^T A p_j = 0 \qquad \forall i \neq j$$

Linear Independence

$S = \{p_0, p_1, \ldots, p_{n-1}\}$ S is linearly independent

if $\sigma_0 p_0 + \sigma_1 p_1 + \ldots \sigma_{n-1} p_{n-1} = 0$

then $\sigma_0 = \sigma_1 = \sigma_2 = \ldots \sigma_{n-1} = 0$

Conjugate set is also linearly independent.

$p_i^T A p_j = 0 \qquad \forall i \neq j$ Therefore, A has at most n conjugate directions.

Conjugate Direction Method

$$x_{k+1} = x_k + \alpha_k p_k$$ Line search

$$p_i^T A p_j = 0 \qquad \forall \mathrm{i} \neq \mathrm{j}$$

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

$$\alpha_k = -\frac{\nabla \phi_k^T p_k}{p_k^T A p_k}$$ 1D minimizer of a quadratic function

Convergence Rate of Steepest Descent

$$\frac{d}{d\alpha} f(x_k - \alpha g_k) = \frac{d}{d\alpha}\left(\frac{1}{2}(x_k - \alpha g_k)^T Q (x_k - \alpha g_k) - b^T (x_k - \alpha g_k)\right) = 0$$

$$= -(x_k - \alpha g_k)^T Q g_k + b^T g_k = 0$$

$$-x_k{}^T Q g_k + \alpha g_k{}^T Q g_k + b^T g_k = 0$$

$$\alpha g_k{}^T Q g_k = x_k{}^T Q g_k - b^T g_k$$

$$\alpha = \frac{x_k{}^T Q g_k - b^T g_k}{g_k{}^T Q g_k}$$

$$\alpha = \frac{(x_k{}^T Q - b^T) g_k}{g_k Q g_k} \qquad \nabla f(x) = Qx - b$$

From Lecture-5

$$\alpha = \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k}$$

$$x_{k+1} = x_k - \alpha_k \nabla f_k \qquad x_{k+1} = x_k - \frac{\nabla f_k^T \nabla f_k}{\nabla f_k^T Q \nabla f_k} \nabla f_k$$

Conjugate Direction Method

$$\alpha = \frac{x_k{}^T Q g_k - b^T g_k}{g_k{}^T Q g_k}$$

$$\alpha = \frac{(x_k{}^T A - b^T)(-p_k)}{(-p_k) A (-p_k)}$$

$$\alpha_k = -\frac{\nabla \phi_k^T p_k}{p_k^T A p_k} \qquad \nabla \phi(x) = Ax - b = r(x)$$

$$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k} \qquad p_i^T A p_j = 0 \qquad \forall \mathrm{i} \neq \mathrm{j}$$

Theorem 5.1

For any x^0 the sequence $\{x_k\}$ generated by the conjugate direction algorithm, converges to the solution x^* of the linear system in at most n steps.

- Sequence $\{x_k\}$
- Linearly independent vectors
- Conjugate vectors

Proof

$$x_{k+1} = x_k + \alpha_k p_k \qquad \alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$$

$$x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \ldots + \alpha_{k-1} p_{k-1}$$

$$x_k - x_0 = \alpha_0 p_0 + \alpha_1 p_1 + \ldots + \alpha_{k-1} p_{k-1}$$

Proof

$S = \{p_0, p_1, \ldots, p_{n-1}\}$ S is linearly independent

Therefore:

$$x^* - x_0 = \sigma_0 p_0 + \sigma_1 p_1 + \ldots \sigma_{n-1} p_{n-1}$$

$$p_k^T A(x^* - x_0) = p_k^T A(\sigma_0 p_0 + \sigma_1 p_1 + \ldots \sigma_{n-1} p_{n-1})$$

$$p_k^T A(x^* - x_0) = (0 \ + 0 + \ldots + \sigma_k p_k^T A p_k + \ldots + 0)$$ conjugate

$$p_k^T A(x^* - x_0) = p_k^T A p_k \qquad \text{(A)}$$

$$\sigma_k = \frac{p_k^T A(x^* - x_0)}{p_k^T A p_k} \qquad \text{(B)}$$

Proof

$$x_{k+1} = x_k + \alpha_k p_k \qquad \alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$$

$$x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \ldots + \alpha_{k-1} p_{k-1}$$

$$x_k - x_0 = \alpha_0 p_0 + \alpha_1 p_1 + \ldots + \alpha_{k-1} p_{k-1}$$

$$p_k^T A(x_k - x_0) = 0$$

$$p_k^T A x_k = p_k^T A x_0$$

(A)

$$p_k^T A(x^* - x_0) = p_k^T A(x^* - x_k) = p_k^T (b - A x_k) = -p_k^T r_k$$

$$\nabla \phi(x) = Ax - b = r(x)$$

$$p_k^T A(x^* - x_0) = -p_k^T r_k$$

Proof

$$p_k^T A(x^* - x_0) = -p_k^T r_k$$

$$\sigma_k = \frac{p_k^T A(x^* - x_0)}{p_k^T A p_k}$$

(B)

$$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$$

Therefore:

$$\sigma_k = \alpha_k$$

QED

Interpretation of Theorem 5.1

If A is a diagonal matrix, then we can minimize the 1-D function along coordinate axes in n iterations.

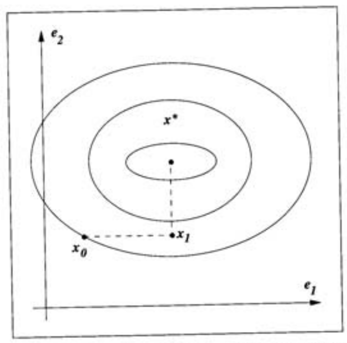
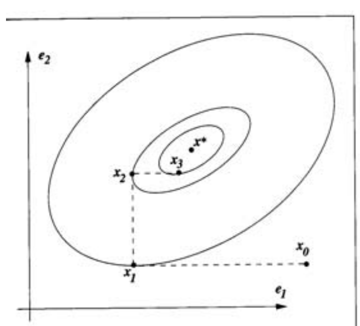


Figure 5.1 Successive minimizations along the coordinate directions find the minimizer of a quadratic with a diagonal Hessian in n iterations.

Interpretation of Theorem 5.1

If A is not a diagonal matrix, then we can not minimize the function along coordinate axes in n iterations.



igure 5.2 Successive minimization along coordinate axes does not find the solution n iterations, for a general convex quadratic.

Transformed Problem

Let

$$\hat{x} = S^{-1}x \qquad \text{where} \qquad S = [p_0, p_1, \ldots, p_{n-1}]$$

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

$$\hat{\varphi}(\hat{x}) = \varphi(x) = \frac{1}{2}\hat{x}^T (S^T AS)\hat{x} - (S^T b)^T \hat{x}$$

$$\hat{\varphi}(\hat{x}) = \varphi(x) = \frac{1}{2}\hat{x}^T D\hat{x} - (c)^T \hat{x}$$

By conjugacy S^TAS is a diagonal matrix.

Now we can minimize along coordinate directions in transformed space.

However, each coordinate direction in transformed space correspond to the conjugate direction in the original space due to

Therefore, we conclude the conjugate direction algorithm converges in n steps.

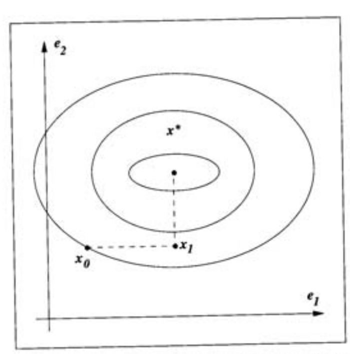


Figure 5.1 Successive minimizations along the coordinate directions
minimizer of a quadratic with a diagonal Hessian in n iterations.

When Hessian is diagonal, each coordinate minimization correctly determines one of the components of the solution x^*. Therefore, after k 1-D minimizations, the quadratic has been minimized on the subspace spanned by $e_1, e_2, \ldots, e_k$.

Theorem 5.2

Let x_0 be any starting point and suppose that the sequence $\{x_k\}$ is generated by the conjugate direction algorithm. Then

$$r_k^T p_i = 0 \qquad \text{for } i = 0,\ldots,k-1$$

and x_k is minimizer of $\phi(x) = \frac{1}{2}x^T Ax - b^T x$ over the set

$$\{x \mid x = x_0 + span\{p_0,\ldots,p_{k-1}\}\} \qquad (3)$$

Proof

First show that a point minimizes over the set (3) if and only if

$$r(\tilde{x})^T p_i = 0 \qquad \text{for } i = 0,\ldots,k-1$$

$$\{x \mid x = x_0 + span\{p_0,\ldots,p_{k-1}\}\}$$

Let $h(\sigma) = \phi(x_0 + \sigma_0 p_0 + \ldots + \sigma_{k-1} p_{k-1})$

Where $\sigma = (\sigma_0, \sigma_1, \ldots \sigma_{k-1})$

Since is strictly convex quadratic, it has a unique minimizer:

$$\frac{\partial h(x^*)}{\partial \sigma_i} = 0, \qquad i = 0,\ldots,k-1$$

$$\nabla \phi(x_0 + \sigma_0^* p_0 + \ldots + \sigma^*_{k-1} p_{k-1})^T p_i = 0 \qquad i = 0,\ldots,k-1 \qquad \text{Chain rule}$$

$r(x)$ is the residual

$$r(\tilde{x})^T p_i = 0 \qquad i = 0,\ldots,k-1$$

Proof

$r_k^T p_i = 0 \qquad \text{for } i = 0, \ldots, k-1$

$\nabla \phi(x) = Ax - b = r(x) \qquad x_{k+1} = x_k + \alpha_k p_k$

$r_{k+1} = r_k + \alpha_k A p_k$

$r_k = r_{k-1} + \alpha_{k-1} A p_{k-1} \qquad \text{(A)}$

From (A)

$r_1 = r_0 + \alpha_0 A p_0$

$r_1^T p_0 = (r_0 + \alpha_0 A p_0)^T p_0$

$r_1^T p_0 = r_0^T p_0 + \alpha_0 p_0^T A p_0$

$r_1^T p_0 = 0$

Because

$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$

Proof

$r_k = r_{k-1} + \alpha_{k-1} A p_{k-1} \qquad \text{(A)}$

True $\quad r_{k-1}^T p_i = 0 \text{ for } i = 0, \ldots, k-2$

From (A)

$p_{k-1}^T r_k = p_{k-1}^T r_{k-1} + \alpha_{k-1} p_{k-1}^T A p_{k-1} = 0$

$\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$

Definition

And

$p_i^T r_k = p_i^T r_{k-1} + \alpha_{k-1} p_i^T A p_{k-1} = 0 \qquad i = 0, \ldots, k-2$

Conjugacy

$r_k^T p_i = 0 \qquad \text{for } i = 0, \ldots, k-2$

induction

Therefore $\quad r_k^T p_i = 0 \qquad \text{for } i = 0, \ldots, k-1 \qquad$ QED

How do we select conjugate directions

- Eigenvalues of *A* are mutually orthogonal and conjugate wrt to *A*.
- Gram-Schmidt process to produce conjugate directions instead of orthogonal vectors.

Basic Properties of the CG

Each direction is chosen to be a linear combination of the steepest descent direction and the previous direction.

$$p_k = -\nabla \phi_k + \beta_k p_{k-1}$$

$$p_k = -r_k + \beta_k p_{k-1}$$

$$p_{k-1}^T A p_k = -r_k p_{k-1}^T A + \beta_k p_{k-1}^T A p_{k-1}$$

$$\beta_k = \frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}$$

Algorithm 5.1

Given x_0;

set $r_0 \leftarrow Ax_0 - b,\ p_0 \leftarrow -r_0, k \leftarrow 0$

While $r_k \neq 0$

$$\alpha_k \leftarrow -\frac{r_k^T p_k}{p_k^T A p_k};$$

$$x_{k+1} \leftarrow x_k + \alpha_k p_k;$$

$$r_{k+1} \leftarrow Ax_{k+1} - b;$$

$$\beta_{k+1} \leftarrow \frac{r_{k+1}^T A p_k}{p_k^T A p_k};$$

$$p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k;$$

$$k \leftarrow k+1;$$

end(*while*)

p_0 is steepest descent