SWFR ENG 4TE3 (6TE3)

COMP SCI 4TE3 (6TE3)

Continuous Optimization Algorithm

Conjugate gradient

Computing and Software
McMaster University

Conjugate directions:

Generalization of orthogonality

Let A be an $n \times n$ symmetric PD matrix.
We consider the strictly convex quadratic function

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

Definition 1. *The directions (vectors) $s^1, \cdots, s^k \in R^n$ are conjugate ($A-$orthogonal) directions if $(s^i)^T A s^j = 0$ for all $1 \le i \ne j \le k$.*

(Conjugate≡orthogonal if $A = I$.)

Theorem 1. *Let $\mathcal{L}$ be a linear subspace, $\mathcal{H}_1 := y^1 + \mathcal{L}$ and $\mathcal{H}_2 := y^2 + \mathcal{L}$ be two parallel affine spaces, and let x^1 and x^2 be the minimizers of $q(x)$ over $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Then for every $s \in \mathcal{L}$, $(x^2 - x^1)$ and s are conjugate w.r.t. A.*

Theorem 2. *Let $s^1, \cdots, s^k \in R^n$ be conjugate directions w.r.t. A. Let x^1 be given and let $x^{i+1} :=$ argmin $q(x^i + \lambda s^i)$, $i = 1, \cdots, k$.*

Then x^{k+1} minimizes $q(x)$ on the affine space $\mathcal{H} = x^1 + \mathcal{L}(s^1, \cdots, s^k)$.

Proof of the Theorems

Proof of Theorem 1

$$x^1 + \lambda s \in \mathcal{H}_1 \Rightarrow \; q(x^1 + \lambda s) \geq q(x^1) \Rightarrow \; s^T \nabla q(x^1) = 0$$
$$x^2 + \lambda s \in \mathcal{H}_2 \Rightarrow \; q(x^2 + \lambda s) \geq q(x^2) \Rightarrow \; s^T \nabla q(x^2) = 0$$

This implies $s^T \left(\nabla q(x^2) - \nabla q(x^1) \right) = s^T A (x^1 - x^2) = 0$. □

Proof of Theorem 2

One has to show that $\nabla q(x^{k+1}) \perp \mathcal{L}(s^1, \cdots, s^k)$, i.e. $\nabla q(x^{k+1}) \perp s^1, \cdots, s^k$.

$$x^{i+1} := x^i + \lambda^i s^i \qquad i = 1, \cdots, k$$

where λ^i indicates the line-minimum, thus

$$x^{k+1} := x^1 + \lambda^1 s^1 + \cdots + \lambda^k s^k = x^i + \lambda^i s^i + \cdots + \lambda^k s^k.$$

Due to exact line-search we have $\nabla q(x^{i+1})^T s^i = 0$.
Using $\nabla q(x) = Ax - b$ we get

$$\nabla q(x^{k+1}) := \nabla q(x^i + \lambda^i s^i) + \sum_{j=i+1}^{k} \lambda^j A s^j.$$
$$(s^i)^T \nabla q(x^{k+1}) := (s^i)^T \nabla q(x^{i+1}) + \sum_{j=i+1}^{k} \lambda^j (s^i)^T A s^j.$$

Hence $(s^i)^T \nabla q(x^{k+1}) = 0$. □

Powell's algorithm - I

Conjugate directions without using gradient

$$\text{minimize } q(x) = \frac{1}{2}x^T Ax - b^T x.$$

Let $s^1, \cdots, s^n$ be linearly independent directions; and x^1 be an initial point, A is symmetric PD.

Cycle 1. Let $z^1 = x^1$ and

$z^{i+1} := \text{argmin}\, q(z^i + \lambda s^i) \qquad i = 1, \cdots, n.$

$x^2 = \text{argmin}\, q(z^{n+1} + \lambda t^1)$, where $t^1 = z^{n+1} - x^1$.

Let $s^i = s^{i+1}$, $i = 1, \cdots, n-1$ and $s^n = t^1$.

Cycle 2. Let $z^1 = x^2$ and

$z^{i+1} := \text{argmin}\, q(z^i + \lambda s^i) \qquad i = 1, \cdots, n.$

$x^3 = \text{argmin}\, q(z^{n+1} + \lambda t^2)$ with $t^2 = z^{n+1} - x^2$.

Then due to Thm 1. t^1 and t^2 are conjugate.

Let $s^i = s^{i+1}$, $i = 1, \cdots, n-1$ and $s^n = t^2$.

Cycle k. Let $z^1 = x^k$ and

$z^{i+1} := \text{argmin}\, q(z^i + \lambda s^i) \qquad i = 1, \cdots, n.$

$x^{k+1} = \text{argmin}\, q(z^{n+1} + \lambda t^k)$ with $t^k = z^{n+1} - x^k$.

Then due to Thm 1. $t^1, \cdots t^k$ are conjugate.

Let $s^i = s^{i+1}$, $i = 1, \cdots, n-1$ and $s^n = t^k$.

Powell's algorithm - II

Conjugate directions without using gradient

$$\text{minimize } q(x) = \frac{1}{2}x^T Ax - b^T x.$$

Cycle n**.** Let $z^1 = x^n$ and

$$z^{i+1} := \arg\min q(z^i + \lambda s^i) \qquad i = 1, \cdots, n.$$

$$x^{n+1} = \text{argmin} q(z^{n+1} + \lambda t^n) \text{ with } t^n = z^{n+1} - x^n.$$

Then due to Thm 1. $t^1, \cdots t^n$ are conjugate.

Let $s^i = s^{i+1}$, $i = 1, \cdots, n-1$ and $s^n = t^n$.

Thus $s^1, \cdots s^n$ are conjugate.

Cycle $n+1$**.** Let $z^1 = x^n$ and

$$z^{i+1} := \arg\min q(z^i + \lambda s^i) \qquad i = 1, \cdots, n,$$

then due to Thm 2 $x^* = z^{n+1}$ is the minimizer of $q(x)$.

Observe: Without any gradient information we were able to find the exact minimum of a strictly convex quadratic function in a finite number of steps. For this at most $(n+1)^2$ line-searches are needed. We also need to store n direction vectors.

Fletcher and Reeves

Conjugate gradient method

$$\text{minimize } q(x) = \frac{1}{2}x^T Ax - b^T x.$$

Let x_1 be an initial point, A is symmetric PD.

Step 1. Let $s_1 = -\nabla q(x_1)$ and $x_2 := \arg\min q(x_1 + \lambda s_1)$.

Step k. Let x_k, $\nabla q(x_k)$ and $s_1, \cdots, s_{k-1}$ conjugate directions be given.

First we find s_k in the space of the negative gradient and the previous directions:

$$s_k := -\nabla q(x_k) + \beta_k^1 s_1 + \cdots + \beta_k^{k-1} s_{k-1}.$$

s_k should be conjugate to $s_1, \cdots, s_{k-1}$. Therefore there holds $s_i^T As_k = 0$, which implies:

$$\beta_k^i = \frac{\nabla q(x_k)^T As_i}{s_i^T As_i}$$

Then $x_{k+1} := \arg\min q(x_k + \lambda s_k)$.

With a bit of analysis we show $\beta_k^i = 0$ if $i < k - 1$, thus

$$s_k = -g_k + \beta_k^{k-1} s_{k-1}, \text{ where } g_k = \nabla q(x_k).$$

Fletcher and Reeves - II

Calculating the coefficients β_k^i

$$\beta_k^i = \frac{g_k^T A s_i}{s_i^T A s_i}.$$

Observe that

$$g_{i+1} - g_i = A(x_{i+1} - x_i) = \lambda_i A s_i$$

thus

$$\beta_k^i = \frac{g_k^T (g_{i+1} - g_i)}{s_i^T (g_{i+1} - g_i)}.$$

Note $g_k^T g_i = 0$ if $i < k$, because

$$g_i := -s_i + \beta_i^1 s_1 + \cdots + \beta_i^{i-1} s_{i-1}$$

$$g_k^T g_i := -g_k^T s_i + \beta_i^1 g_k^T s_1 + \cdots + \beta_i^{i-1} g_k^T s_{i-1} = 0$$

because $g_k \perp s_1, \cdots s_{k-1}$, by using Theorem 2.
Similarly, $g_i^T g_i = -g_i^T s_i$, thus

$$\beta_k^i = \begin{cases} 0 & \text{if} i < k-1, \\ \frac{g_k^T g_k}{-s_{k-1}^T g_{k-1}} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} & \text{if} i = k-1. \end{cases}$$

Fletcher and Reeves - III

Calculating the coefficients β_k^i

Thus the direction s^k is given by

$$s_k = -g_k + \frac{\|g_k\|^2}{\|g_{k-1}\|^2} s_{k-1},$$

Only the previous direction has to be stored and to minimize $q(x)$ at most n line-searches are needed.

Polak-Ribière Method

For the nonlinear problem $\min_{x \in \mathbb{R}^n} f(x)$,

Linear Search might be inexact. FR-CG: $\beta_{k+1}^{FR} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2}$;

Note that in case that $f(x)$ is quadratic and the line search is exact, it holds $\left\|g_{k+1}\right\|^2 = g_{k+1}^T (g_{k+1} - g_k)$. Another choice is PR-CG where:

$$\beta_{k+1}^{PR} = \frac{g_{k+1}^T (g_{k+1} - g_k)}{\|g_k\|^2};$$

Numerical experience shows PR-CG is more robust and efficient.

Quasi-Newton Methods

Approximate the inverse Hessian

$$\text{minimize } q(x) = \frac{1}{2}x^T Ax - b^T x.$$

Let x_1 be an initial point, A is symmetric PD.

For any two points x^k, x^{k+1} we have

$$\nabla q(x^{k+1}) - \nabla q(x^k) = Ax^{k+1} - b - (Ax^k - b) = A(x^{k+1} - x^k).$$

Let $y^k = \nabla q(x^{k+1}) - \nabla q(x^k)$ and $\sigma^k = x^{k+1} - x^k = \lambda^k s^k$, so we get:

$$\sigma^k = A^{-1} y^k$$

We are going to approximate A^{-1} by a matrix H_k. The matrix H_k should behave like the inverse Hessian A^{-1}. The search direction is calculated by

$$s^k = -H_k \nabla q(x^k) \quad \text{and} \quad x^{k+1} := \arg\min q(x^k + \lambda s^k)$$

In the iterations the update $H_{k+1} = H_k + D_k$ will be used.

Quasi-Newton - II

Desired properties of the update

1. **Symmetric and PD:** To guarantee a decreasing direction we need H_{k+1} to be symmetric and positive definite.

2. **Quasi-Newton (QN):** Maintain the Newton property $\sigma^k = H_{k+1} y^k$.

3. **Hereditary:** For all $1 \leq i \leq k$ $\sigma^i = H_{k+1} y^i$.

Quasi-Newton - III

Choices for D_k

Symmetric rank-one (SR1) update:

$$D_k = \frac{(\sigma_k - H_k y_k)(\sigma_k - H_k y_k)^T}{(\sigma_k - H_k y_k)^T y_k}.$$

No guarantee to keep positive definiteness, need $(\sigma_k - H_k y_k)^T y_k > 0$.

Davidon-Fletcher-Powell (DFP) rank-2:

$$D_k = \frac{\sigma_k \sigma_k^T}{\sigma_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}.$$

If H_k is positive definite, then so is H_{k+1}.

Quasi-Newton - IV

Consider now using approximations of A itself, denoted B_k, with

$$B_{k+1} = B_k + \Delta B_k.$$

Broyden-Fletcher-Goldfarb-Shanno update:

$$\Delta B_k = \frac{y_k y_k^T}{\sigma_k^T y_k} - \frac{B_k \sigma_k \sigma_k^T B_k}{\sigma_k^T B_k \sigma_k}.$$

Taking its inverse,

$$D_k = \left(1 + \frac{y_k^T H_k y_k}{\sigma_k^T y_k}\right) \frac{\sigma_k \sigma_k^T}{\sigma_k^T y_k} - \frac{\sigma_k y_k^T H_k + H_k y_k \sigma_k^T}{\sigma_k^T y_k}$$

If B_k is positive definite, then so is B_{k+1} (same proof as for DFP).

QN Method - V

Broyden's family:

$$
\begin{aligned}
B_{k+1}(\phi) &= (1-\phi)B_{k+1}^{BFGS} + \phi B_{k+1}^{DFP} \\
&= B_{k+1}^{BFGS} + \phi \sigma_k^T B_k \sigma_k w_k w_k^T
\end{aligned}
$$

where

$$
w_k = \frac{y_k}{\sigma_k^T y_k} - \frac{B_k \sigma_k}{\sigma_k^T B_k \sigma_k}.
$$

and its inverse form

$$
\begin{aligned}
H_{k+1}(\theta) &= (1-\theta)H_{k+1}^{DFP} + \theta H_{k+1}^{BFGS} \\
&= H_{k+1}^{DFP} + \theta y_k^T H_k y_k v_k v_k^T
\end{aligned}
$$

where

$$
v_k = \frac{\sigma_k}{\sigma_k^T y_k} - \frac{H_k y_k}{y_k^T H_k y_k}.
$$

If $\phi, \theta \geq 0$ then B_{k+1}, H_{k+1} remain positive definite.

In practice, BFGS has been found to be the most efficient update in Broyden's family.