

4TE3/6TE3
Algorithms for
Continuous Optimization
(Search directions-II CG/QN)

Tamás TERLAKY
Computing and Software
McMaster University

Hamilton, January 2004

terlaky@mcmaster.ca

Tel: 27780

Conjugate directions:

Generalization of orthogonality

Let A be an $n \times n$ symmetric PD matrix.

We consider the strictly convex quadratic function

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

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

(Conjugate \equiv orthogonal if $A = I$.)

Theorem 1 *Let \mathcal{L} be a linear subspace, $\mathcal{H}_1 := x^1 + \mathcal{L}$ and $\mathcal{H}_2 := x^2 + \mathcal{L}$ be two parallel affine spaces where x^1 and x^2 are 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 $s^1, \dots, s^k \in R^n$ be conjugate directions w.r.t. A . Let x^1 be given and let*

$$x^{i+1} := \operatorname{argmin} q(x^i + \lambda s^i) \quad i = 1, \dots, k.$$

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

Proof of the Theorems

Proof of Theorem 1

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

This imply:

$$s^T (\nabla q(x^2) - \nabla q(x^1)) = s^T A(x^1 - x^2) = 0. \quad \square$$

Proof of Theorem 2

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

$$x^{i+1} := x^i + \lambda^i s^i \quad i = 1, \dots, k$$

where λ^i indicates the line-minimum, thus

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

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

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, \dots, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

Powell's algorithm - II

Conjugate directions without using gradient

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

The directions s^1, \dots, s^n are linearly independent.

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

$$z^{i+1} := \arg \min q(z^i + \lambda s^i) \quad i = 1, \dots, n.$$
$$x^{n+1} = \operatorname{argmin} q(z^{n+1} + \lambda t^n) \text{ with } t^n = z^{n+1} - x^n.$$

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

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

Thus s^1, \dots, s^n are conjugate.

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

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

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, \dots, s_{k-1} conjugate directions be given. First we find s_k in the space of the negative gradient and the previous directions:

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

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

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

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

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

$$s_k = -g_k + \beta_k^{k-1} s_{k-1},$$

where $g_k = \nabla q(x_k)$.

Fletcher and Reeves - II

Calculating the coefficients β_{ki}

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

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

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

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 $n + 1$ line-searches are needed. The eigenvalues of A have big influence on the behavior of CG algorithm.

Other CG Methods

Polak-Ribière Method

For nonlinear problem

$$\min_{x \in \mathcal{R}^n} f(x).$$

Linear Search might be inexact. FR-CG:

$$\beta_{k+1}^{FR} = \frac{\|g_{k+1}\|^2}{\|g_k\|^2};$$

where $g_k = \nabla f(x_k)$. Note that in case that $f(x)$ is quadratic and the line search is exact, there holds $\|g_{k+1}\|^2 = g_{k+1}^T (g_{k+1} - g_k)$. 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.

Recently research on CG method focus on the global convergence of the algorithm with inexact line search for nonlinear problems. In principle, CG method is efficient for solving quadratic programming. Thus it is widely used in solving subproblems. For large-scale and specific problems, Precondition is necessary.

Variable Metric/Quasi-Newton

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} , it should be symmetric positive definite and thus the search direction is calculated by

$$s^k = -H_k \nabla q(x^k),$$

further we want to keep the Newton property $\sigma^k = H_k y^k$.

In the iterations the update

$H_{k+1} = H_k + D_k$ will be used.

Quasi-Newton - II

Desired properties of the update

$$H_{k+1} = H_k + D_k$$

The iterative sequence is: $x^1, \dots, x^k, x^{k+1}, \dots$

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

$$y^k = \nabla q(x^{k+1}) - \nabla q(x^k) \quad \text{and} \quad \sigma^k = x^{k+1} - x^k = \lambda^k s^k.$$

Desired properties:

1. Hereditary/Preservability: For all $i < k$

$$\sigma^i = H_k y^i \Rightarrow \sigma^i = H_{k+1} y^i.$$

2. Quasi-Newton (QN): Make the Newton property for k , i.e.,

$$\sigma^k = H_{k+1} y^k \Rightarrow D_k y^k = \sigma^k - H_k y^k.$$

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

Quasi-Newton Method

Approximation via local quadratic model

minimize $\hat{f}(x) = f_k + g_k^T(x - x_k) + \frac{1}{2}(x - x_k)^T B_k(x - x_k)$.

When B_k is positive definite, it has a unique solution $x_{k+1} = x_k - B_k^{-1}g_k$. Define the inverse of B_k as H_k . Thus we can use $-H_k g_k$ as a new search direction.

Algorithm: Given starting point x_0 , convergence tolerance ϵ . Choose initial matrix H_0 or B_0 .

While $\|g_k\| \geq \epsilon$;

 Compute the search direction

$$s_k = -H_k g_k,$$

 Set $x_{k+1} = x_k + \lambda_k s_k$ where λ_k is the step size determined by Wolfe-line search.

 Define $\sigma_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$, Compute H_{k+1} or B_{k+1} by QN update.

$k \leftarrow k + 1$;

End(while)

Wolfe-condition and QN Update

Wolfe-condition

$$\begin{aligned} f(x_k + \lambda_k s_k) &\leq f(x_k) + c_1 \lambda_k g_k^T s_k; \\ c_2 g_k^T s_k &\leq \nabla f(x_k + \lambda_k s_k)^T s_k, \end{aligned}$$

with $0 < c_1 < c_2 < 1$.

Suppose B_k, H_k are known, how to update $H_{k+1} = H_k + \Delta H_k$? Similarly how to update B_k ?

Desired properties of update:

1. **Quasi-Newton (QN):** Satisfy the secant equation, i.e.,

$$\sigma_k = H_{k+1} y_k \iff B_{k+1} \sigma_k = y_k.$$

that is $\Delta H_k y_k = \sigma_k - H_k y_k$.

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

Quasi-Newton - III

Choices for ΔH_k

From the QN property we see that a matrix D is to be found that for given y and z satisfies:

$$Dy = z.$$

Among others, for any $a \in R^n$ the rank-one matrix $D = \frac{za^T}{a^T y}$ always satisfies the equation if $a^T y \neq 0$.

Popular choices:

Symmetric rank-one (SR1) update:

$$\Delta H_k = \frac{(\sigma_k - H_k y_k)(\sigma_k - H_k y_k)^T}{(\sigma_k - H_k y_k)^T y_k},$$

here we have chosen $a = \sigma_k - H_k y_k$ with $y = y_k$ and $z = \sigma_k - H_k y_k$.

No guarantee to keep positive definiteness.

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

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

If $y_k^T \sigma_k > 0$, then H_{k+1} is positive definite!!

Quasi-Newton - IV

Proof of the positive definiteness of DFP

First, for any positive definite H and nonzero y , we can prove that

$$H - \frac{Hy y^T H}{y^T H y} = H^{\frac{1}{2}} \left(I - \frac{H^{\frac{1}{2}} y y^T H^{\frac{1}{2}}}{y^T H y} \right) H^{\frac{1}{2}},$$

is positive semidefinite. Denote $d = H^{\frac{1}{2}} y$, one can see that the matrix

$$I - \frac{d d^T}{d^T d}$$

is positive semidefinite. So is the matrix $H - \frac{Hy y^T H}{y^T H y}$ and y is the only eigenvector corresponding to its eigenvalue 0.

Second, we prove that if \bar{H} is P.SD., and d is the only nonzero eigenvector satisfying $\bar{H}d = 0$. Then for any $\theta > 0, z \in \mathfrak{R}^n$ with $z^T d \neq 0$, the matrix $\bar{H} + \theta z z^T$ is positive definite.

Replacing \bar{H} by $H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}$, θ by $\frac{1}{y_k^T \sigma_k}$, and z by σ_k , we see that the DFP update is P.D.

BFGS, Broyden's Family

Broyden-Fletcher-Goldfarb-Shanno update:

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

Similarly when B_k is P.D. and $\sigma_k^T y_k > 0$, then B_{k+1} is P.D.

Broyden's family:

$$B_{k+1}(\phi) = B_{k+1}^{BFGS} + \phi \sigma_k^T B_k \sigma_k w_k w_k^T, \quad \phi \geq 0,$$

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

$$H_{k+1}(\theta) = H_{k+1}^{DFP} + \theta y_k^T H_k y_k v_k v_k^T, \quad \theta \geq 0$$

where

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

Properties of Broyden's Family

Since $H_k B_k = I$, one can show that

$$H_{k+1}(1)B_{k+1}(0) = I,$$

and

$$H_{k+1}(0)B_{k+1}(1) = I.$$

Self-Dual QN Method, $\theta = \phi = \frac{1}{2}$.

How to keep $\sigma_k^T y_k > 0$? Note that $s_k = \lambda_k s_k$ and $y_k = g_{k+1} - g_k$. Thus we need to keep

$$g_{k+1}^T s_k > g_k^T s_k.$$

This can be guaranteed by using line-search, such as Wolfe-search.

Local convergence had been proven in early 1970s.

The global convergence of QN method was first proven by Powell (1976) for convex optimization by using BFGS update. In 1987, Byrd, Nocedal and Yuan showed the whole Broyden family is convergent except the DFP update.

QN Method:VII

In practice, BFGS is the most efficient update in Broyden's family. This coincides with the theoretical conclusion.

The first QN method is DFP, however, no one can prove it is convergent except for very special problems. The convergence of DFP is an open problem.

In practice for large-scale problems, limited memory QN method (or restart QN update in finite steps) is also a good choice. Since many problems are not globally well-posed, but locally convex. In such cases, restart can help us to get rid of the wrong information obtained from the early iterations. Precondition is helpful in choosing the initial matrices H_0 and B_0 .

It is known that when the initial matrix is sufficiently good and the sequence of the iterates converges, then the matrix sequence with SR1 update will converge to the real Hessian of the objective at the solution point! This is a very important property that might help to deal with non-convex problem.