Quasi-Newton methods for minimization

Lectures for PHD course on
Unconstrained Numerical Optimization

Enrico Bertolazzi

DIMS – Università di Trento

May 2008

Outline

Algorithm (General quasi-Newton algorithm)

$k \leftarrow 0$; $\boldsymbol{x}_0$ *assigned*; $\boldsymbol{g}_0 = \nabla \mathsf{f}(\boldsymbol{x}_0)^T$; $\boldsymbol{H}_0 = \nabla^2 \mathsf{f}(\boldsymbol{x}_0)^{-1}$;
while $\|\boldsymbol{g}_k\| > \epsilon$ **do**
  — compute search direction
  $\boldsymbol{d}_k = -\boldsymbol{H}_k \boldsymbol{g}_k$;
  Approximate $\arg\min_{\lambda>0} \mathsf{f}(\boldsymbol{x}_k + \lambda \boldsymbol{d}_k)$ *by linsearch;*
  — perform step
  $\boldsymbol{s}_k = \lambda_k \boldsymbol{d}_k$;
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$;
  $\boldsymbol{g}_{k+1} = \nabla \mathsf{f}(\boldsymbol{x}_{k+1})^T$;
  $\boldsymbol{y}_k = \boldsymbol{g}_{k+1} - \boldsymbol{g}_k$;
  — update $\boldsymbol{H}_{k+1}$
  $\boldsymbol{H}_{k+1} =$ *some_algorithm*$(\boldsymbol{H}_k, \boldsymbol{s}_k, \boldsymbol{y}_k)$;
  $k \leftarrow k+1$;
end while

Outline

- Let $\boldsymbol{B}_k$ and approximation of the Hessian of $\mathsf{f}(\boldsymbol{x})$. Let $\boldsymbol{x}_k$, $\boldsymbol{x}_{k+1}$, $\boldsymbol{g}_k$ and $\boldsymbol{g}_{k+1}$ points and gradients at k and $k+1$-th iterates. Using the Broyden update formula to force secant condition to $\boldsymbol{B}_{k+1}$ we obtain

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{(\boldsymbol{y}_k - \boldsymbol{B}_k\boldsymbol{s}_k)\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{s}_k},$$

where $\boldsymbol{s}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ and $\boldsymbol{y}_k = \boldsymbol{g}_{k+1} - \boldsymbol{g}_k$. By using Sherman–Morrison formula and setting $\boldsymbol{H}_k = \boldsymbol{B}_k^{-1}$ we obtain the update:

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k - \frac{(\boldsymbol{H}_k\boldsymbol{y}_k - \boldsymbol{s}_k)\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{s}_k + \boldsymbol{s}_k^T\boldsymbol{H}_k\boldsymbol{g}_{k+1}}\boldsymbol{H}_k$$

- The previous update does not maintain symmetry. In fact if $\boldsymbol{H}_k$ is symmetric then $\boldsymbol{H}_{k+1}$ not necessarily is symmetric.

- To avoid the loss of symmetry we can consider an update of the form:

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \boldsymbol{u}\boldsymbol{u}^T$$

- Imposing the secant condition (on the inverse) we obtain

$$\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \boldsymbol{s}_k \qquad \Rightarrow \qquad \boldsymbol{H}_k\boldsymbol{y}_k + \boldsymbol{u}\boldsymbol{u}^T\boldsymbol{y}_k = \boldsymbol{s}_k$$

from previous equality

$$\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k + \boldsymbol{y}_k^T\boldsymbol{u}\boldsymbol{u}^T\boldsymbol{y}_k = \boldsymbol{y}_k^T\boldsymbol{s}_k \qquad \Rightarrow$$

$$\boldsymbol{y}_k^T\boldsymbol{u} = \left(\boldsymbol{y}_k^T\boldsymbol{s}_k - \boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k\right)^{1/2}$$

we obtain

$$\boldsymbol{u} = \frac{\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k}{\boldsymbol{u}^T\boldsymbol{y}_k} = \frac{\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k}{\left(\boldsymbol{y}_k^T\boldsymbol{s}_k - \boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k\right)^{1/2}}$$

- substituting the expression of $\boldsymbol{u}$

$$\boldsymbol{u} = \frac{\boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k}{\left(\boldsymbol{y}_k^T\boldsymbol{s}_k - \boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k\right)^{1/2}}$$

in the update formula, we obtain

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{w}_k\boldsymbol{w}_k^T}{\boldsymbol{w}_k^T\boldsymbol{y}_k} \qquad \boldsymbol{w}_k = \boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k$$

- The previous update formula is the symmetric rank one formula (SR1).
- To be definite the previous formula needs $\boldsymbol{w}_k^T\boldsymbol{y}_k \neq 0$. Moreover if $\boldsymbol{w}_k^T\boldsymbol{y}_k < 0$ and $\boldsymbol{H}_k$ is positive definite then $\boldsymbol{H}_{k+1}$ may loss positive definitiveness.
- Have $\boldsymbol{H}_k$ symmetric and positive definite is important for global convergence

This lemma is used in the forward theorems

Lemma

Let be

$$\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} + c$$

with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *symmetric and positive definite. Then*

$$\begin{aligned}\boldsymbol{y}_k &= \boldsymbol{g}_{k+1} - \boldsymbol{g}_k \\ &= \boldsymbol{A}\boldsymbol{x}_{k+1} - \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{b} \\ &= \boldsymbol{A}\boldsymbol{s}_k\end{aligned}$$

where $\boldsymbol{g}_k = \nabla\mathsf{q}(\boldsymbol{x}_k)^T$.

Theorem (property of SR1 update)

Let be

$$\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} + c$$

with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ symmetric and positive definite. Let be $\boldsymbol{x}_0$ and $\boldsymbol{H}_0$ assigned. Let $\boldsymbol{x}_k$ and $\boldsymbol{H}_k$ produced by

1. $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{s}_k$*;*
2. $\boldsymbol{H}_{k+1}$ *updated by the SR1 formula*

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{w}_k\boldsymbol{w}_k^T}{\boldsymbol{w}_k^T\boldsymbol{y}_k} \qquad \boldsymbol{w}_k = \boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k$$

If $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-1}$ are linearly independent then $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$.

Proof. (1/2).

We prove by induction the hereditary property $\boldsymbol{H}_i\boldsymbol{y}_j = \boldsymbol{s}_j$.
BASE: For $i = 1$ is exactly the secant condition of the update.
INDUCTION: Suppose the relation is valid for $k > 0$ the we prove that it is valid for $k+1$. In fact, from the update formula

$$\boldsymbol{H}_{k+1}\boldsymbol{y}_j = \boldsymbol{H}_k\boldsymbol{y}_j + \frac{\boldsymbol{w}_k^T\boldsymbol{y}_j}{\boldsymbol{w}_k^T\boldsymbol{y}_k}\boldsymbol{w}_k \qquad \boldsymbol{w}_k = \boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k$$

by the induction hypothesis for $j < k$ and using lemma on slide 8 we have

$$\begin{aligned}\boldsymbol{w}_k^T\boldsymbol{y}_j &= \boldsymbol{s}_k^T\boldsymbol{y}_j - \boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_j = \boldsymbol{s}_k^T\boldsymbol{y}_j - \boldsymbol{y}_k^T\boldsymbol{s}_j \\ &= \boldsymbol{y}_k^T\boldsymbol{A}\boldsymbol{y}_j - \boldsymbol{y}_k^T\boldsymbol{A}\boldsymbol{y}_j = 0\end{aligned}$$

so that $\boldsymbol{H}_{k+1}\boldsymbol{y}_j = \boldsymbol{H}_k\boldsymbol{y}_j = \boldsymbol{s}_j$ for $j = 0, 1, \ldots, k-1$. For $j = k$ we have $\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \boldsymbol{s}_k$ trivially by construction of the SR1 formula.

Proof. (2/2).

To prove that $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$ notice that

$$\boldsymbol{H}_n\boldsymbol{y}_j = \boldsymbol{s}_j, \qquad \boldsymbol{A}\boldsymbol{s}_j = \boldsymbol{y}_j, \qquad j = 0, 1, \ldots, n-1$$

and combining the equality

$$\boldsymbol{H}_n\boldsymbol{A}\boldsymbol{s}_j = \boldsymbol{s}_j, \qquad j = 0, 1, \ldots, n-1$$

due to the linear independence of $\boldsymbol{s}_i$ we have $\boldsymbol{H}_n\boldsymbol{A} = \boldsymbol{I}$ i.e. $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$. □

Properties of SR1 update (1/2)

1. The SR1 update possesses the natural quadratic termination property (like CG).
2. SR1 satisfy the hereditary property $\boldsymbol{H}_k\boldsymbol{y}_j = \boldsymbol{s}_j$ for $j < k$.
3. SR1 does maintain the positive definitiveness of $\boldsymbol{H}_k$ if and only if $\boldsymbol{w}_k^T\boldsymbol{y}_k > 0$. However this condition is difficult to guarantee.
4. Sometimes $\boldsymbol{w}_k^T\boldsymbol{y}_k$ becomes very small or 0. This results in serious numerical difficulty (roundoff) or even the algorithm is broken. We can avoid this breakdown by the following strategy

Breakdown workaround for SR1 update

1. if $\left|\boldsymbol{w}_k^T\boldsymbol{y}_k\right| \geq \epsilon \left\|\boldsymbol{w}_k^T\right\| \left\|\boldsymbol{y}_k\right\|$ (i.e. the angle between $\boldsymbol{w}_k$ and $\boldsymbol{y}_k$ is far from 90 degree), then we update with the SR1 formula.
2. Otherwise we set $\boldsymbol{H}_{k+1} = \boldsymbol{H}_k$.

Properties of SR1 update (2/2)

Theorem (Convergence of nonlinear SR1 update)

Let $f(\boldsymbol{x})$ satisfying standard assumption. Let be $\{\boldsymbol{x}_k\}$ a sequence of iterates such that $\lim_{k\to\infty}\boldsymbol{x}_k = \boldsymbol{x}_\star$. Suppose we use the breakdown workaround for SR1 update and the steps $\{\boldsymbol{s}_k\}$ are uniformly linearly independent. Then we have

$$\lim_{k\to\infty} \left\| \boldsymbol{H}_k - \nabla^2 f(\boldsymbol{x}_\star)^{-1} \right\| = 0.$$

A.R.Conn, N.I.M.Gould and P.L.Toint
Convergence of quasi-Newton matrices generated by the symmetric rank one update.
Mathematic of Computation **50** 399–430, 1988.

Outline

- The SR1 update, although symmetric do not have minimum property like the Broyden update for the non symmetric case.
- The Broyden update

$$\boldsymbol{A}_{k+1} = \boldsymbol{A}_k + \frac{(\boldsymbol{y}_k - \boldsymbol{A}_k\boldsymbol{s}_k)\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{s}_k}$$

solve the minimization problem

$$\|\boldsymbol{A}_{k+1} - \boldsymbol{A}_k\|_F \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_F \qquad \text{for all } \boldsymbol{A}\boldsymbol{s}_k = \boldsymbol{y}_k$$

- If we solve a similar problem in the class of symmetric matrix we obtain the Powell-symmetric-Broyden (PSB) update

Lemma (Powell-Symmetric-Broyden update)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ symmetric and $\boldsymbol{s}, \boldsymbol{y} \in \mathbb{R}^n$ with $\boldsymbol{A}\boldsymbol{s} \neq \boldsymbol{y}$. Consider the set

$$\mathcal{B} = \left\{ \boldsymbol{B} \in \mathbb{R}^{n\times n} \,|\, \boldsymbol{B}\boldsymbol{s} = \boldsymbol{y},\, \boldsymbol{B} = \boldsymbol{B}^T \right\}$$

then there exists a unique matrix $\boldsymbol{B} \in \mathcal{B}$ such that

$$\|\boldsymbol{A} - \boldsymbol{B}\|_F \leq \|\boldsymbol{A} - \boldsymbol{C}\|_F \qquad \text{for all } \boldsymbol{C} \in \mathcal{B}$$

moreover $\boldsymbol{B}$ has the following form

$$\boldsymbol{B} = \boldsymbol{A} + \frac{\boldsymbol{\omega}\boldsymbol{s}^T + \boldsymbol{s}\boldsymbol{\omega}^T}{\boldsymbol{s}^T\boldsymbol{s}} - (\boldsymbol{\omega}^T\boldsymbol{s})\frac{\boldsymbol{s}\boldsymbol{s}^T}{(\boldsymbol{s}^T\boldsymbol{s})^2} \qquad \boldsymbol{\omega} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}$$

then $\boldsymbol{B}$ is a rank two perturbation of the matrix $\boldsymbol{A}$.

Proof. (1/9).

First of all notice that $\mathcal{B}$ is not empty, in fact $\boldsymbol{B}$ satisfy $\boldsymbol{B}\boldsymbol{s} = \boldsymbol{y}$ so that the set is not empty. Next we reformulate the problem as a constrained minimum problem:

$$\arg\min_{\boldsymbol{B}\in\mathbb{R}^{n\times n}} \frac{1}{2}\sum_{i,j=1}^{n}(A_{ij}-B_{ij})^2 \quad \text{subject to } \boldsymbol{B}\boldsymbol{s}=\boldsymbol{y} \text{ and } \boldsymbol{B}=\boldsymbol{B}^T$$

The solution is a stationary point of the Lagrangian:

$$g(\boldsymbol{B},\boldsymbol{\lambda},\boldsymbol{M}) = \frac{1}{2}\|\boldsymbol{A}-\boldsymbol{B}\|_F^2 + \boldsymbol{\lambda}^T(\boldsymbol{B}\boldsymbol{s}-\boldsymbol{y}) + \sum_{i<j}\mu_{ij}(B_{ij}-B_{ji})$$

Proof. (2/9).

taking the gradient we have

$$\frac{\partial}{\partial B_{ij}}g(\boldsymbol{B},\boldsymbol{\lambda},\boldsymbol{B}) = A_{ij} - B_{ij} + \lambda_i s_j + M_{ij} = 0$$

where

$$M_{ij} = \begin{cases} \mu_{ij} & \text{if } i<j; \\ -\mu_{ij} & \text{if } i>j; \\ 0 & \text{If } i=j. \end{cases}$$

The previous equality can be written in matrix form as

$$\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{\lambda}\boldsymbol{s}^T + \boldsymbol{M}.$$

where $\boldsymbol{M}$ is an antisymmetric matrix.

Proof. (3/9).

Imposing the symmetry for $\boldsymbol{B}$

$$\boldsymbol{A}+\boldsymbol{\lambda}\boldsymbol{s}^T+\boldsymbol{M} = \boldsymbol{A}^T+\boldsymbol{s}\boldsymbol{\lambda}^T+\boldsymbol{M}^T = \boldsymbol{A}+\boldsymbol{s}\boldsymbol{\lambda}^T-\boldsymbol{M}$$

solving for $\boldsymbol{M}$ we have

$$\boldsymbol{M} = \frac{\boldsymbol{s}\boldsymbol{\lambda}^T - \boldsymbol{\lambda}\boldsymbol{s}^T}{2}$$

substituting in $\boldsymbol{B}$ we have

$$\boldsymbol{B} = \boldsymbol{A} + \frac{\boldsymbol{s}\boldsymbol{\lambda}^T + \boldsymbol{\lambda}\boldsymbol{s}^T}{2}$$

Proof. (4/9).

Imposing $\boldsymbol{s}^T\boldsymbol{B}\boldsymbol{s} = \boldsymbol{s}^T\boldsymbol{y}$

$$\boldsymbol{s}^T\boldsymbol{A}\boldsymbol{s} + \frac{\boldsymbol{s}^T\boldsymbol{s}\boldsymbol{\lambda}^T\boldsymbol{s} + \boldsymbol{s}^T\boldsymbol{\lambda}\boldsymbol{s}^T\boldsymbol{s}}{2} = \boldsymbol{s}^T\boldsymbol{y} \qquad \Rightarrow$$

$$\boldsymbol{\lambda}^T\boldsymbol{s} = (\boldsymbol{s}^T\boldsymbol{\omega})/(\boldsymbol{s}^T\boldsymbol{s})$$

where $\boldsymbol{\omega} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{s}$. Imposing $\boldsymbol{B}\boldsymbol{s} = \boldsymbol{y}$

$$\boldsymbol{A}\boldsymbol{s} + \frac{\boldsymbol{s}\boldsymbol{\lambda}^T\boldsymbol{s} + \boldsymbol{\lambda}\boldsymbol{s}^T\boldsymbol{s}}{2} = \boldsymbol{y} \qquad \Rightarrow$$

$$\boldsymbol{\lambda} = \frac{2\boldsymbol{\omega}}{\boldsymbol{s}^T\boldsymbol{s}} - \frac{(\boldsymbol{s}^T\boldsymbol{\omega})\boldsymbol{s}}{(\boldsymbol{s}^T\boldsymbol{s})^2}$$

next we compute the explicit form of $\boldsymbol{B}$.

Proof. (5/9).

Substituting

$$\lambda = \frac{2\omega}{s^T s} - \frac{(s^T\omega)s}{(s^T s)^2} \qquad \text{in} \qquad B = A + \frac{s\lambda^T + \lambda s^T}{2}$$

we obtain

$$B = A + \frac{\omega s^T + s\omega^T}{s^T s} - (\omega^T s)\frac{s s^T}{(s^T s)^2} \qquad \omega = y - As$$

next we prove that B is the unique minimum.

Proof. (6/9).

The matrix B is at minimum distance, in fact consider a symmetric matrix C which satisfy $Cs = y$ so that

$$\omega = y - As = (C - A)s = Es \quad \text{where} \quad E = C - A,$$

substituting ω with Es in B of slide N.16 and noticing that $E^T = E$ we have

$$B - A = \frac{Ess^T + ss^T E}{s^T s} - (s^T E s)\frac{s s^T}{(s^T s)^2}$$

consider now the product $(B - A)s$ which result in

$$(B - A)s = Es$$

so that

$$\|(B - A)s\|_2 = \|Es\|_2$$

Proof. (7/9).

consider now the product $(B - A)z$ where z is a vector orthogonal to s (i.e $z^T s = 0$) which result in

$$(B - A)z = \frac{Ess^T z + ss^T E z}{s^T s} - (s^T E s)\frac{s s^T z}{(s^T s)^2}$$
$$= \frac{ss^T}{s^T s} E z$$

so that using Frobenius norm

$$\|(B - A)z\|_2 = \left\|\frac{ss^T}{s^T s} E z\right\|_2 \le \left\|\frac{ss^T}{s^T s}\right\|_F \|Ez\|_2$$
$$\le \|Ez\|_2 = \|(C - A)z\|_2$$

Proof. (8/9).

So that considering n othonormal vector $\{v_1, v_2, \ldots, v_n\}$ with $v_1 = s/\|s\|_2$ we have

$$\|(B - A)v_k\|_2 \le \|(C - A)v_k\|_2, \quad k = 1, 2, \ldots, n$$

and using the properties of Frobenius norm

$$\|B - A\|_F^2 = \sum_{k=1}^{n} \|(B - A)v_k\|_2^2$$
$$\le \sum_{k=1}^{n} \|(C - A)v_k\|_2^2$$
$$\le \|C - A\|_2^2$$

i.e. we have $\|B - A\|_F \le \|C - A\|_F$ for all $C \in \mathcal{B}$.

Proof. (9/9).

Let B' and B'' two different minimum. Then $\frac{1}{2}(B'+B'') \in \mathcal{B}$ moreover

$$\left\| A - \frac{1}{2}(B'+B'') \right\|_F \leq \frac{1}{2}\left\| A - B' \right\|_F + \frac{1}{2}\left\| A - B'' \right\|_F$$

If the inequality is strict we have a contradiction. From the Cauchy–Schwartz inequality we have an equality only when $A - B' = \lambda(A - B'')$ so that

$$B' - \lambda B'' = (1-\lambda)A$$

and

$$B's - \lambda B''s = (1-\lambda)As \quad \Rightarrow \quad (1-\lambda)y = (1-\lambda)As$$

cause $As \neq y$ this is true only when $\lambda = 1$, i.e. $B' = B''$. □

Algorithm (PSB quasi-Newton algorithm)

$k \leftarrow 0$;
x_0 *assigned*; $g_0 \leftarrow \nabla f(x_0)^T$; $B_0 \leftarrow \nabla^2 f(x_0)$;
while $\|g_k\| > \epsilon$ **do**
 — compute search direction
 $d_k = -B_k^{-1} g_k$; *[solve linear system $B d_k = -g_k$]*
 Approximate $\arg\min_{\alpha>0} f(x_k + \alpha d_k)$ *by linsearch*;
 — perform step
 $x_{k+1} = x_k + \alpha d_k$;
 — update B_{k+1}
 $g_{k+1} = \nabla f(x_{k+1})^T$;
 $\omega_k = g_{k+1} - g_k - \alpha B_k d_k = g_{k+1} + (\alpha-1) g_k$;
 $B_{k+1} = B_k + \dfrac{d_k \omega_k^T + \omega_k d_k^T}{\alpha d_k^T d_k} - \dfrac{d^T \omega_k}{\alpha} d_k d_k^T$;
 $k \leftarrow k+1$;
end while

Algorithm (PSB quasi-Newton algorithm)

$k \leftarrow 0$;
x *assigned*; $g \leftarrow \nabla f(x)^T$; $B \leftarrow \nabla^2 f(x)$;
while $\|g\| > \epsilon$ **do**
 — compute search direction
 $d \leftarrow -B^{-1} g$; *[solve linear system $Bd = -g$]*
 Approximate $\arg\min_{\alpha>0} f(x + \alpha d)$ *by linsearch*;
 — perform step
 $x \leftarrow x + \alpha d$;
 — update B_{k+1}
 $\omega \leftarrow \nabla f(x)^T + (\alpha-1) g$;
 $g \leftarrow \nabla f(x)^T$;
 $\beta \leftarrow (\alpha d^T d)^{-1}$;
 $\gamma \leftarrow d^T \omega / \alpha$;
 $B \leftarrow B + \beta(d\omega^T + \omega d^T) - \gamma d d^T$;
 $k \leftarrow k+1$;
end while

Outline

- The SR1 and PSB update maintains the symmetry but do not maintains the positive definitiveness of the matrix $\boldsymbol{H}_{k+1}$. To recover this further property we can try the update of the form:

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \alpha \boldsymbol{u}\boldsymbol{u}^T + \beta \boldsymbol{v}\boldsymbol{v}^T$$

- Imposing the secant condition (on the inverse)

$$\begin{aligned} \boldsymbol{H}_{k+1}\boldsymbol{y}_k &= \boldsymbol{s}_k \qquad \Rightarrow \\ \boldsymbol{H}_k\boldsymbol{y}_k + \alpha(\boldsymbol{u}^T\boldsymbol{y}_k)\boldsymbol{u} + \beta(\boldsymbol{v}^T\boldsymbol{y}_k)\boldsymbol{v} &= \boldsymbol{s}_k \qquad \Rightarrow \\ \alpha(\boldsymbol{u}^T\boldsymbol{y}_k)\boldsymbol{u} + \beta(\boldsymbol{v}^T\boldsymbol{y}_k)\boldsymbol{v} &= \boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k \end{aligned}$$

clearly this equation has not a unique solution. A natural choice for $\boldsymbol{u}$ and $\boldsymbol{v}$ is the following:

$$\boldsymbol{u} = \boldsymbol{s}_k \qquad \boldsymbol{v} = \boldsymbol{H}_k\boldsymbol{y}_k$$

- Solving for α and β the equation

$$\alpha(\boldsymbol{s}_k^T\boldsymbol{y}_k)\boldsymbol{s}_k + \beta(\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k)\boldsymbol{H}_k\boldsymbol{y}_k = \boldsymbol{s}_k - \boldsymbol{H}_k\boldsymbol{y}_k$$

we obtain

$$\alpha = \frac{1}{\boldsymbol{s}_k^T\boldsymbol{y}_k} \qquad \beta = -\frac{1}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}$$

- substituting in the updating formula we obtain the Davidon Fletcher and Powell (DFP) rank 2 update formula

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}$$

- Obviously this is only one of the possible choices and with other solutions we obtain different update formulas. Next we must prove that under suitable condition the DFP update formula maintains positive definitiveness.

Positive definitiveness of DFP update

Theorem (Positive definitiveness of DFP update)

Given $\boldsymbol{H}_k$ symmetric and positive definite, then the DFP update

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}$$

produce $\boldsymbol{H}_{k+1}$ positive definite if and only if $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$.

Remark (Wolfe ⇒ DFP update is SPD)

Expanding $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$ we have $\nabla \mathsf{f}(\boldsymbol{x}_{k+1})\boldsymbol{s}_k > \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{s}_k$. Remember that in a minimum search algorithm we have $\boldsymbol{s}_k = \alpha_k\boldsymbol{p}_k$ with $\alpha_k > 0$. But the second Wolfe condition for line-search is $\nabla \mathsf{f}(\boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k)\boldsymbol{p}_k \geq c_2 \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$ with $0 < c_2 < 1$. But this imply:

$$\nabla \mathsf{f}(\boldsymbol{x}_{k+1})\boldsymbol{s}_k \geq c_2 \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{s}_k > \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{s}_k \quad \Rightarrow \quad \boldsymbol{s}_k^T\boldsymbol{y}_k > 0.$$

Proof. (1/2).

Let be $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$: consider a $\boldsymbol{z} \neq 0$ then

$$\begin{aligned} \boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} &= \boldsymbol{z}^T\left(\boldsymbol{H}_k - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}\right)\boldsymbol{z} + \boldsymbol{z}^T\frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\boldsymbol{z} \\ &= \boldsymbol{z}^T\boldsymbol{H}_k\boldsymbol{z} - \frac{(\boldsymbol{z}^T\boldsymbol{H}_k\boldsymbol{y}_k)(\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{z})}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k} + \frac{(\boldsymbol{z}^T\boldsymbol{s}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k} \end{aligned}$$

$\boldsymbol{H}_k$ is SPD so that there exists the Cholesky decomposition $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{H}_k$. Defining $\boldsymbol{a} = \boldsymbol{L}^T\boldsymbol{z}$ and $\boldsymbol{b} = \boldsymbol{L}^T\boldsymbol{y}_k$ we can write

$$\boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} = \frac{(\boldsymbol{a}^T\boldsymbol{a})(\boldsymbol{b}^T\boldsymbol{b}) - (\boldsymbol{a}^T\boldsymbol{b})^2}{\boldsymbol{b}^T\boldsymbol{b}} + \frac{(\boldsymbol{z}^T\boldsymbol{s}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k}$$

from the Cauchy-Schwartz inequality we have $(\boldsymbol{a}^T\boldsymbol{a})(\boldsymbol{b}^T\boldsymbol{b}) \geq (\boldsymbol{a}^T\boldsymbol{b})^2$ so that $\boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} \geq 0$.

Proof. (2/2).

To prove strict inequality remember from the Cauchy-Schwartz inequality that $(\boldsymbol{a}^T\boldsymbol{a})(\boldsymbol{b}^T\boldsymbol{b}) = (\boldsymbol{a}^T\boldsymbol{b})^2$ if and only if $\boldsymbol{a} = \lambda\boldsymbol{b}$, i.e.

$$\boldsymbol{L}^T\boldsymbol{z} = \lambda\boldsymbol{L}^T\boldsymbol{y}_k \qquad \Rightarrow \qquad \boldsymbol{z} = \lambda\boldsymbol{y}_k$$

but in this case

$$\frac{(\boldsymbol{z}^T\boldsymbol{s}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k} = \lambda^2\frac{(\boldsymbol{y}^T\boldsymbol{s}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k} > 0 \qquad \Rightarrow \qquad \boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} > 0.$$

Let be $\boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} > 0$ for all $\boldsymbol{z} \neq \boldsymbol{0}$: Choosing $\boldsymbol{z} = \boldsymbol{y}_k$ we have

$$0 < \boldsymbol{y}_k^T\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \frac{(\boldsymbol{y}^T\boldsymbol{s}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k} = \boldsymbol{s}_k^T\boldsymbol{y}_k$$

□

Algorithm (DFP quasi-Newton algorithm)

$k \leftarrow 0$;
$\boldsymbol{x}$ *assigned;* $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})^T$; $\boldsymbol{H} \leftarrow \nabla^2 f(\boldsymbol{x})^{-1}$;
while $\|\boldsymbol{g}\| > \epsilon$ **do**
  — *compute search direction*
  $\boldsymbol{d} \leftarrow -\boldsymbol{H}\boldsymbol{g}$;
  Approximate $\arg\min_{\alpha>0} f(\boldsymbol{x} + \alpha\boldsymbol{d})$ *by linsearch;*
  — *perform step*
  $\boldsymbol{x} \leftarrow \boldsymbol{x} + \alpha\boldsymbol{d}$;
  — *update* $\boldsymbol{H}_{k+1}$
  $\boldsymbol{y} \leftarrow \nabla f(\boldsymbol{x})^T - \boldsymbol{g}$;
  $\boldsymbol{z} \leftarrow \boldsymbol{H}\boldsymbol{y}$;
  $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})^T$;
  $\boldsymbol{H} \leftarrow \boldsymbol{H} + \alpha\frac{\boldsymbol{d}\boldsymbol{d}^T}{\boldsymbol{d}^T\boldsymbol{y}} - \frac{\boldsymbol{z}\boldsymbol{z}^T}{\boldsymbol{y}^T\boldsymbol{z}}$;
  $k \leftarrow k + 1$;
end while

Theorem (property of DFP update)

Let be $q(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_\star) + c$ *with* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *symmetric and positive definite. Let be* $\boldsymbol{x}_0$ *and* $\boldsymbol{H}_0$ *assigned. Let* $\{\boldsymbol{x}_k\}$ *and* $\{\boldsymbol{H}_k\}$ *produced by the sequence* $\{\boldsymbol{s}_k\}$

1. $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{s}_k$;
2. $\boldsymbol{H}_{k+1} \leftarrow \boldsymbol{H}_k + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}$;

where $\boldsymbol{s}_k = \alpha_k\boldsymbol{p}_k$ *with* α_k *is obtained by exact line-search. Then for* $j < k$ *we have*

1. $\boldsymbol{g}_k^T\boldsymbol{s}_j = 0$; *[orthogonality property]*
2. $\boldsymbol{H}_k\boldsymbol{y}_j = \boldsymbol{s}_j$; *[hereditary property]*
3. $\boldsymbol{s}_k^T\boldsymbol{A}\boldsymbol{s}_j = 0$; *[conjugate direction property]*
4. *The method terminate (i.e.* $\nabla f(\boldsymbol{x}_m) = \boldsymbol{0}$*) at* $\boldsymbol{x}_m = \boldsymbol{x}_\star$ *with* $m \leq n$*. If* $n = m$ *then* $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$.

Proof. (1/4).

Points (1), (2) and (3) are proved by induction. The base of induction is obvious, let be the theorem true for $k > 0$. Due to exact line search we have:

$$\boldsymbol{g}_{k+1}^T\boldsymbol{s}_k = 0$$

moreover by induction for $j < k$ we have $\boldsymbol{g}_{k+1}^T\boldsymbol{s}_j = 0$, in fact:

$$\begin{aligned}
\boldsymbol{g}_{k+1}^T\boldsymbol{s}_j &= \boldsymbol{g}_j^T\boldsymbol{s}_j + \sum_{i=j}^{k-1}(\boldsymbol{g}_{i+1} - \boldsymbol{g}_i)^T\boldsymbol{s}_j \\
&= 0 + \sum_{i=j}^{k-1}(\boldsymbol{A}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_\star) - \boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_\star))^T\boldsymbol{s}_j \\
&= \sum_{i=j}^{k-1}(\boldsymbol{A}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i))^T\boldsymbol{s}_j \\
&= \sum_{i=j}^{k-1}\boldsymbol{s}_i^T\boldsymbol{A}\boldsymbol{s}_j = 0. \qquad \text{[induction + conjugacy prop.]}
\end{aligned}$$

Proof. (2/4).

By using $\boldsymbol{s}_{k+1} = -\alpha_{k+1}\boldsymbol{H}_{k+1}\boldsymbol{g}_{k+1}$ we have $\boldsymbol{s}_{k+1}^T\boldsymbol{A}\boldsymbol{s}_j = 0$, in fact:

$$\begin{aligned}
\boldsymbol{s}_{k+1}^T\boldsymbol{A}\boldsymbol{s}_j &= -\alpha_{k+1}\boldsymbol{g}_{k+1}^T\boldsymbol{H}_{k+1}(\boldsymbol{A}\boldsymbol{x}_{j+1} - \boldsymbol{A}\boldsymbol{x}_j) \\
&= -\alpha_{k+1}\boldsymbol{g}_{k+1}^T\boldsymbol{H}_{k+1}(\boldsymbol{A}(\boldsymbol{x}_{j+1} - \boldsymbol{x}_\star) - \boldsymbol{A}(\boldsymbol{x}_j - \boldsymbol{x}_\star)) \\
&= -\alpha_{k+1}\boldsymbol{g}_{k+1}^T\boldsymbol{H}_{k+1}(\boldsymbol{g}_{j+1} - \boldsymbol{g}_j) \\
&= -\alpha_{k+1}\boldsymbol{g}_{k+1}^T\boldsymbol{H}_{k+1}\boldsymbol{y}_j \\
&= -\alpha_{k+1}\boldsymbol{g}_{k+1}^T\boldsymbol{s}_j \qquad \text{[induction + hereditary prop.]} \\
&= 0
\end{aligned}$$

Proof. (3/4).

Due to DFP construction we have

$$\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \boldsymbol{s}_k$$

by inductive hypothesis and DFP formula for $j < k$ we have, $\boldsymbol{s}_k^T\boldsymbol{y}_j = \boldsymbol{s}_k^T\boldsymbol{A}\boldsymbol{s}_j = 0$, moreover

$$\begin{aligned}
\boldsymbol{H}_{k+1}\boldsymbol{y}_j &= \boldsymbol{H}_k\boldsymbol{y}_j + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{y}_j}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_j}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k} \\
&= \boldsymbol{s}_j + \frac{\boldsymbol{s}_k 0}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{s}_j}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k} \qquad [\boldsymbol{H}_k\boldsymbol{y}_j = \boldsymbol{s}_j] \\
&= \boldsymbol{s}_j - \frac{\boldsymbol{H}_k\boldsymbol{y}_k(\boldsymbol{g}_{k+1} - \boldsymbol{g}_k)^T\boldsymbol{s}_j}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k} \qquad [\boldsymbol{y}_j = \boldsymbol{g}_{j+1} - \boldsymbol{g}_j] \\
&= \boldsymbol{s}_j \qquad \text{[induction + ortho. prop.]}
\end{aligned}$$

Proof. (4/4).

Finally if $m = n$ we have $\boldsymbol{s}_j$ with $j = 0, 1, \ldots, n-1$ are conjugate and linearly independent. From hereditary property and lemma on slide 8

$$\boldsymbol{H}_n\boldsymbol{A}\boldsymbol{s}_k = \boldsymbol{H}_n\boldsymbol{y}_k = \boldsymbol{s}_k$$

i.e. we have

$$\boldsymbol{H}_n\boldsymbol{A}\boldsymbol{s}_k = \boldsymbol{s}_k, \qquad k = 0, 1, \ldots, n-1$$

due to linear independence of $\{\boldsymbol{s}_k\}$ follows that $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$. □

Outline

- Another update which maintain symmetry and positive definitiveness is the Broyden Fletcher Goldfarb and Shanno (BFGS,1970) rank 2 update.
- This update was independently discovered by the four authors.
- A convenient way to introduce BFGS is by the concept of duality.
- Consider an update for the Hessian, say

$$\boldsymbol{B}_{k+1} = \mathcal{U}(\boldsymbol{B}_k, \boldsymbol{s}_k, \boldsymbol{y}_k)$$

which satisfy $\boldsymbol{B}_{k+1}\boldsymbol{s}_k = \boldsymbol{y}_k$ (the secant condition on the Hessian). Then by exchanging $\boldsymbol{B}_k \rightleftharpoons \boldsymbol{H}_k$ and $\boldsymbol{s}_k \rightleftharpoons \boldsymbol{y}_k$ we obtain the dual update for the inverse of the Hessian, i.e.

$$\boldsymbol{H}_{k+1} = \mathcal{U}(\boldsymbol{H}_k, \boldsymbol{y}_k, \boldsymbol{s}_k)$$

which satisfy $\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \boldsymbol{s}_k$ (the secant condition on the inverse of the Hessian).

- Starting from the Davidon Fletcher and Powell (DFP) rank 2 update formula

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}$$

by the duality we obtain the Broyden Fletcher Goldfarb and Shanno (BFGS) update formula

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{\boldsymbol{y}_k\boldsymbol{y}_k^T}{\boldsymbol{y}_k^T\boldsymbol{s}_k} - \frac{\boldsymbol{B}_k\boldsymbol{s}_k\boldsymbol{s}_k^T\boldsymbol{B}_k}{\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{s}_k}$$

- The BFGS formula written in this way is not useful in the case of large problem. We need an equivalent formula for the inverse of the approximate Hessian. This can be done with a generalization of the Sherman-Morrison formula.

Sherman-Morrison-Woodbury formula (1/2)

Sherman-Morrison-Woodbury formula permit to explicit write the inverse of a matrix changed with a rank k perturbation

Proposition (Sherman–Morrison–Woodbury formula)

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{V}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\big(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{U}\big)^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}$$

where

$$\boldsymbol{U} = \big[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k\big] \qquad \boldsymbol{V} = \big[\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\big]$$

The Sherman–Morrison–Woodbury formula can be checked by a direct calculation.

Sherman-Morrison-Woodbury formula (2/2)

Remark

The previous formula can be written as:

$$\Big(\boldsymbol{A} + \sum_{i=1}^{k}\boldsymbol{u}_i\boldsymbol{v}_i^T\Big)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}$$

where

$$C_{ij} = \delta_{ij} + \boldsymbol{v}_i^T\boldsymbol{u}_j \qquad i,j = 1,2,\ldots,k$$

The BFGS update for $\boldsymbol{H}$

Proposition

By using the Sherman-Morrison-Woodbury formula the BFGS update for $\boldsymbol{H}$ becomes:

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{s}_k^T + \boldsymbol{s}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{s}_k^T\boldsymbol{y}_k} + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\left(1 + \frac{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\right) \qquad (A)$$

Or equivalently

$$\boldsymbol{H}_{k+1} = \left(\boldsymbol{I} - \frac{\boldsymbol{s}_k\boldsymbol{y}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\right)\boldsymbol{H}_k\left(\boldsymbol{I} - \frac{\boldsymbol{y}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\right) + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} \qquad (B)$$

Proof. (1/3).

Consider the Sherman-Morrison-Woodbury formula with $k = 2$ and

$$\boldsymbol{u}_1 = \boldsymbol{v}_1 = \frac{\boldsymbol{y}_k}{(\boldsymbol{s}_k^T\boldsymbol{y}_k)^{1/2}} \qquad \boldsymbol{u}_2 = -\boldsymbol{v}_2 = \frac{\boldsymbol{B}_k\boldsymbol{s}_k}{(\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{s}_k)^{1/2}}$$

in this way (setting $\boldsymbol{H}_k = \boldsymbol{B}_k^{-1}$) we have

$$C_{11} = 1 + \boldsymbol{v}_1^T\boldsymbol{u}_1 = 1 + \frac{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}{\boldsymbol{s}_k^T\boldsymbol{y}_k}$$

$$C_{22} = 1 + \boldsymbol{v}_2^T\boldsymbol{u}_2 = -\frac{\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{H}_k\boldsymbol{B}_k\boldsymbol{s}_k}{\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{s}_k} = 1 - 1 = 0$$

$$C_{12} = \boldsymbol{v}_1^T\boldsymbol{u}_2 = \frac{\boldsymbol{y}_k^T\boldsymbol{B}_k\boldsymbol{s}_k}{(\boldsymbol{s}_k^T\boldsymbol{y}_k)^{1/2}(\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{s}_k)^{1/2}} = \frac{(\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{s}_k)^{1/2}}{(\boldsymbol{s}_k^T\boldsymbol{y}_k)^{1/2}}$$

$$C_{21} = \boldsymbol{v}_2^T\boldsymbol{u}_1 = -C_{12}$$

Proof. (2/3).

In this way the matric $\boldsymbol{C}$ has the form

$$\boldsymbol{C} = \begin{pmatrix} \beta & \alpha \\ -\alpha & 0 \end{pmatrix} \qquad \boldsymbol{C}^{-1} = \frac{1}{\alpha^2}\begin{pmatrix} 0 & -\alpha \\ \alpha & \beta \end{pmatrix}$$

$$\beta = 1 + \frac{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}{\boldsymbol{s}_k^T\boldsymbol{y}_k} \qquad \alpha = \frac{(\boldsymbol{s}_k^T\boldsymbol{B}_k\boldsymbol{s}_k)^{1/2}}{(\boldsymbol{s}_k^T\boldsymbol{y}_k)^{1/2}}$$

where setting $\tilde{\boldsymbol{U}} = \boldsymbol{H}_k\boldsymbol{U}$ and $\tilde{\boldsymbol{V}} = \boldsymbol{H}_k\boldsymbol{V}$ where

$$\tilde{\boldsymbol{u}}_i = \boldsymbol{H}_k\boldsymbol{u}_i \quad \text{and} \quad \tilde{\boldsymbol{v}}_i = \boldsymbol{H}_k\boldsymbol{v}_i \qquad i = 1, 2$$

we have

$$\begin{aligned} \boldsymbol{H}_{k+1} &= \boldsymbol{H}_k - \boldsymbol{H}_k\boldsymbol{U}\boldsymbol{C}^{-1}\boldsymbol{V}^T\boldsymbol{H}_k = \boldsymbol{H}_k - \tilde{\boldsymbol{U}}\boldsymbol{C}^{-1}\tilde{\boldsymbol{V}}^T \\ &= \boldsymbol{H}_k + \frac{1}{\alpha}(-\tilde{\boldsymbol{u}}_1\tilde{\boldsymbol{v}}_2^T + \tilde{\boldsymbol{u}}_2\tilde{\boldsymbol{v}}_1^T) - \frac{\beta}{\alpha^2}\tilde{\boldsymbol{u}}_2\tilde{\boldsymbol{v}}_2^T \end{aligned}$$

Proof. (3/3).

Substituting the values of α, β, $\tilde{\boldsymbol{u}}$'s and $\tilde{\boldsymbol{v}}$'s we have we have

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k - \frac{\boldsymbol{H}_k\boldsymbol{y}_k\boldsymbol{s}_k^T + \boldsymbol{s}_k\boldsymbol{y}_k^T\boldsymbol{H}_k}{\boldsymbol{s}_k^T\boldsymbol{y}_k} + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\left(1 + \frac{\boldsymbol{y}_k^T\boldsymbol{H}_k\boldsymbol{y}_k}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\right)$$

At this point the update formula (B) is a straightforward calculation.

Positive definitiveness of BFGS update

Theorem (Positive definitiveness of BFGS update)

Given $\boldsymbol{H}_k$ symmetric and positive definite, then the BFGS update

$$\boldsymbol{H}_{k+1} = \Big(\boldsymbol{I} - \frac{\boldsymbol{s}_k\boldsymbol{y}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\Big)\boldsymbol{H}_k\Big(\boldsymbol{I} - \frac{\boldsymbol{y}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\Big) + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}$$

produce $\boldsymbol{H}_{k+1}$ positive definite if and only if $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$.

Remark (Wolfe ⇒ BFGS update is SPD)

Expanding $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$ we have $\nabla f(\boldsymbol{x}_{k+1})\boldsymbol{s}_k > \nabla f(\boldsymbol{x}_k)\boldsymbol{s}_k$. Remember that in a minimum search algorithm we have $\boldsymbol{s}_k = \alpha_k\boldsymbol{p}_k$ with $\alpha_k > 0$. But the second Wolfe condition for line-search is $\nabla f(\boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k)\boldsymbol{p}_k \geq c_2\,\nabla f(\boldsymbol{x}_k)\boldsymbol{p}_k$ with $0 < c_2 < 1$. But this imply:

$$\nabla f(\boldsymbol{x}_{k+1})\boldsymbol{s}_k \geq c_2\,\nabla f(\boldsymbol{x}_k)\boldsymbol{s}_k > \nabla f(\boldsymbol{x}_k)\boldsymbol{s}_k \quad \Rightarrow \quad \boldsymbol{s}_k^T\boldsymbol{y}_k > 0.$$

Proof.

Let be $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$: consider a $\boldsymbol{z} \neq 0$ then

$$\boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} = \boldsymbol{w}^T\boldsymbol{H}_k\boldsymbol{w} + \frac{(\boldsymbol{z}^T\boldsymbol{s}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k} \quad \text{where} \quad \boldsymbol{w} = \boldsymbol{z} - \boldsymbol{y}_k\frac{\boldsymbol{s}_k^T\boldsymbol{z}}{\boldsymbol{s}_k^T\boldsymbol{y}_k}$$

In order to have $\boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} = 0$ we must have $\boldsymbol{w} = 0$ and $\boldsymbol{z}^T\boldsymbol{s}_k = 0$. But $\boldsymbol{z}^T\boldsymbol{s}_k = 0$ imply $\boldsymbol{w} = \boldsymbol{z}$ and this imply $\boldsymbol{z} = \boldsymbol{0}$.

Let be $\boldsymbol{z}^T\boldsymbol{H}_{k+1}\boldsymbol{z} > 0$ for all $\boldsymbol{z} \neq \boldsymbol{0}$: Choosing $\boldsymbol{z} = \boldsymbol{y}_k$ we have

$$0 < \boldsymbol{y}_k^T\boldsymbol{H}_{k+1}\boldsymbol{y}_k = \frac{(\boldsymbol{s}_k^T\boldsymbol{y}_k)^2}{\boldsymbol{s}_k^T\boldsymbol{y}_k} = \boldsymbol{s}_k^T\boldsymbol{y}_k$$

and thus $\boldsymbol{s}_k^T\boldsymbol{y}_k > 0$. □

Algorithm (BFGS quasi-Newton algorithm)

$k \leftarrow 0$;
$\boldsymbol{x}$ *assigned*; $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})^T$; $\boldsymbol{H} \leftarrow \nabla^2 f(\boldsymbol{x})^{-1}$;
while $\|\boldsymbol{g}\| > \epsilon$ **do**
  — compute search direction
  $\boldsymbol{d} \leftarrow -\boldsymbol{H}\boldsymbol{g}$;
  Approximate $\arg\min_{\alpha>0} f(\boldsymbol{x} + \alpha\boldsymbol{d})$ *by linsearch*;
  — perform step
  $\boldsymbol{x} \leftarrow \boldsymbol{x} + \alpha\boldsymbol{d}$;
  — update $\boldsymbol{H}_{k+1}$
  $\boldsymbol{y} \leftarrow \nabla f(\boldsymbol{x})^T - \boldsymbol{g}$;
  $\boldsymbol{z} \leftarrow \boldsymbol{H}\boldsymbol{y}/(\boldsymbol{d}^T\boldsymbol{y})$;
  $\boldsymbol{g} \leftarrow \nabla f(\boldsymbol{x})^T$;
  $\beta \leftarrow (\alpha + \boldsymbol{y}^T\boldsymbol{z})/(\boldsymbol{d}^T\boldsymbol{y})$;
  $\boldsymbol{H} \leftarrow \boldsymbol{H} - (\boldsymbol{z}\boldsymbol{d}^T + \boldsymbol{d}\boldsymbol{z}^T) + \beta\boldsymbol{d}\boldsymbol{d}^T$;
  $k \leftarrow k + 1$;
end while

Theorem (property of BFGS update)

Let be $q(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_\star)^T\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_\star) + c$ *with* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *symmetric and positive definite. Let be $\boldsymbol{x}_0$ and $\boldsymbol{H}_0$ assigned. Let $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{H}_k\}$ produced by the sequence $\{\boldsymbol{s}_k\}$*

1. $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{s}_k$;
2. $\boldsymbol{H}_{k+1} \leftarrow \Big(\boldsymbol{I} - \frac{\boldsymbol{s}_k\boldsymbol{y}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\Big)\boldsymbol{H}_k\Big(\boldsymbol{I} - \frac{\boldsymbol{y}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\Big) + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}$;

where $\boldsymbol{s}_k = \alpha_k\boldsymbol{p}_k$ with α_k is obtained by exact line-search. Then for $j < k$ we have

1. $\boldsymbol{g}_k^T\boldsymbol{s}_j = 0$; *[orthogonality property]*
2. $\boldsymbol{H}_k\boldsymbol{y}_j = \boldsymbol{s}_j$; *[hereditary property]*
3. $\boldsymbol{s}_k^T\boldsymbol{A}\boldsymbol{s}_j = 0$; *[conjugate direction property]*
4. *The method terminate (i.e. $\nabla f(\boldsymbol{x}_m) = \boldsymbol{0}$) at $\boldsymbol{x}_m = \boldsymbol{x}_\star$ with $m \leq n$. If $n = m$ then $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$.*

Proof. (1/4).

Points (1), (2) and (3) are proved by induction. The base of induction is obvious, let be the theorem true for $k > 0$. Due to exact line search we have:

$$\boldsymbol{g}_{k+1}^T \boldsymbol{s}_k = 0$$

moreover by induction for $j < k$ we have $\boldsymbol{g}_{k+1}^T \boldsymbol{s}_j = 0$, in fact:

$$\begin{aligned}
\boldsymbol{g}_{k+1}^T \boldsymbol{s}_j &= \boldsymbol{g}_j^T \boldsymbol{s}_j + \sum_{i=j}^{k-1} (\boldsymbol{g}_{i+1} - \boldsymbol{g}_i)^T \boldsymbol{s}_j \\
&= 0 + \sum_{i=j}^{k-1} (\boldsymbol{A}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_\star) - \boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_\star))^T \boldsymbol{s}_j \\
&= \sum_{i=j}^{k-1} (\boldsymbol{A}(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i))^T \boldsymbol{s}_j \\
&= \sum_{i=j}^{k-1} \boldsymbol{s}_i^T \boldsymbol{A} \boldsymbol{s}_j = 0. \qquad \text{[induction + conjugacy prop.]}
\end{aligned}$$

Proof. (2/4).

By using $\boldsymbol{s}_{k+1} = -\alpha_{k+1} \boldsymbol{H}_{k+1} \boldsymbol{g}_{k+1}$ we have $\boldsymbol{s}_{k+1}^T \boldsymbol{A} \boldsymbol{s}_j = 0$, in fact:

$$\begin{aligned}
\boldsymbol{s}_{k+1}^T \boldsymbol{A} \boldsymbol{s}_j &= -\alpha_{k+1} \boldsymbol{g}_{k+1}^T \boldsymbol{H}_{k+1} (\boldsymbol{A}\boldsymbol{x}_{j+1} - \boldsymbol{A}\boldsymbol{x}_j) \\
&= -\alpha_{k+1} \boldsymbol{g}_{k+1}^T \boldsymbol{H}_{k+1} (\boldsymbol{A}(\boldsymbol{x}_{j+1} - \boldsymbol{x}_\star) - \boldsymbol{A}(\boldsymbol{x}_j - \boldsymbol{x}_\star)) \\
&= -\alpha_{k+1} \boldsymbol{g}_{k+1}^T \boldsymbol{H}_{k+1} (\boldsymbol{g}_{j+1} - \boldsymbol{g}_j) \\
&= -\alpha_{k+1} \boldsymbol{g}_{k+1}^T \boldsymbol{H}_{k+1} \boldsymbol{y}_j \\
&= -\alpha_{k+1} \boldsymbol{g}_{k+1}^T \boldsymbol{s}_j \qquad \text{[induction + hereditary prop.]} \\
&= 0
\end{aligned}$$

notice that we have used $\boldsymbol{A}\boldsymbol{s}_j = \boldsymbol{y}_j$.

Proof. (3/4).

Due to BFGS construction we have

$$\boldsymbol{H}_{k+1} \boldsymbol{y}_k = \boldsymbol{s}_k$$

by inductive hypothesis and BFGS formula for $j < k$ we have, $\boldsymbol{s}_k^T \boldsymbol{y}_j = \boldsymbol{s}_k^T \boldsymbol{A} \boldsymbol{s}_j = 0$,

$$\begin{aligned}
\boldsymbol{H}_{k+1} \boldsymbol{y}_j &= \left(\boldsymbol{I} - \frac{\boldsymbol{s}_k \boldsymbol{y}_k^T}{\boldsymbol{s}_k^T \boldsymbol{y}_k}\right) \boldsymbol{H}_k \left(\boldsymbol{y}_j - \frac{\boldsymbol{s}_k^T \boldsymbol{y}_j}{\boldsymbol{s}_k^T \boldsymbol{y}_k} \boldsymbol{y}_k\right) + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T \boldsymbol{y}_j}{\boldsymbol{s}_k^T \boldsymbol{y}_k} \\
&= \left(\boldsymbol{I} - \frac{\boldsymbol{s}_k \boldsymbol{y}_k^T}{\boldsymbol{s}_k^T \boldsymbol{y}_k}\right) \boldsymbol{H}_k \boldsymbol{y}_j + \frac{\boldsymbol{s}_k 0}{\boldsymbol{s}_k^T \boldsymbol{y}_k} \qquad [\boldsymbol{H}_k \boldsymbol{y}_j = \boldsymbol{s}_j] \\
&= \boldsymbol{s}_j - \frac{\boldsymbol{y}_k^T \boldsymbol{s}_j}{\boldsymbol{s}_k^T \boldsymbol{y}_k} \boldsymbol{s}_k \\
&= \boldsymbol{s}_j
\end{aligned}$$

Proof. (4/4).

Finally if $m = n$ we have $\boldsymbol{s}_j$ with $j = 0, 1, \ldots, n-1$ are conjugate and linearly independent. From hereditary property and lemma on slide 8

$$\boldsymbol{H}_n \boldsymbol{A} \boldsymbol{s}_k = \boldsymbol{H}_n \boldsymbol{y}_k = \boldsymbol{s}_k$$

i.e. we have

$$\boldsymbol{H}_n \boldsymbol{A} \boldsymbol{s}_k = \boldsymbol{s}_k, \qquad k = 0, 1, \ldots, n-1$$

due to linear independence of $\{\boldsymbol{s}_k\}$ follows that $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$. □

Outline

- The BFGS update

$$\boldsymbol{H}_{k+1}^{BFGS} \leftarrow \boldsymbol{H}_k - \frac{\boldsymbol{H}_k \boldsymbol{y}_k \boldsymbol{s}_k^T + \boldsymbol{s}_k \boldsymbol{y}_k^T \boldsymbol{H}_k}{\boldsymbol{s}_k^T \boldsymbol{y}_k} + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T}{\boldsymbol{s}_k^T \boldsymbol{y}_k}\left(1 + \frac{\boldsymbol{y}_k^T \boldsymbol{H}_k \boldsymbol{y}_k}{\boldsymbol{s}_k^T \boldsymbol{y}_k}\right)$$

and DFP update

$$\boldsymbol{H}_{k+1}^{DFP} \leftarrow \boldsymbol{H}_k + \frac{\boldsymbol{s}_k \boldsymbol{s}_k^T}{\boldsymbol{s}_k^T \boldsymbol{y}_k} - \frac{\boldsymbol{H}_k \boldsymbol{y}_k \boldsymbol{y}_k^T \boldsymbol{H}_k}{\boldsymbol{y}_k^T \boldsymbol{H}_k \boldsymbol{y}_k}$$

maintains the symmetry and positive definitiveness.

- The following update

$$\boldsymbol{H}_{k+1}^{\theta} \leftarrow (1-\theta)\boldsymbol{H}_{k+1}^{DFP} + \theta \boldsymbol{H}_{k+1}^{BFGS}$$

maintain for any θ the symmetry, and for $\theta \in [0,1]$ also the positive definitiveness.

Positive definitiveness of Broyden Class update

Theorem (Positive definitiveness of Broyden Class update)

Given $\boldsymbol{H}_k$ symmetric and positive definite, then the Broyden Class update

$$\boldsymbol{H}_{k+1}^{\theta} \leftarrow (1-\theta)\boldsymbol{H}_{k+1}^{DFP} + \theta \boldsymbol{H}_{k+1}^{BFGS}$$

produce $\boldsymbol{H}_{k+1}^{\theta}$ positive definite for any $\theta \in [0,1]$ if and only if $\boldsymbol{s}_k^T \boldsymbol{y}_k > 0$.

Theorem (property of Broyden Class update)

Let be $\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_\star)^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_\star) + c$ *with* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *symmetric and positive definite. Let be $\boldsymbol{x}_0$ and $\boldsymbol{H}_0$ assigned. Let $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{H}_k\}$ produced by the sequence $\{\boldsymbol{s}_k\}$*

1. $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{s}_k$;
2. $\boldsymbol{H}_{k+1}^{\theta} \leftarrow (1-\theta)\boldsymbol{H}_{k+1}^{DFP} + \theta \boldsymbol{H}_{k+1}^{BFGS}$;

where $\boldsymbol{s}_k = \alpha_k \boldsymbol{p}_k$ with α_k is obtained by exact line-search. Then for $j < k$ we have

1. $\boldsymbol{g}_k^T \boldsymbol{s}_j = 0$; [orthogonality property]
2. $\boldsymbol{H}_k \boldsymbol{y}_j = \boldsymbol{s}_j$; [hereditary property]
3. $\boldsymbol{s}_k^T \boldsymbol{A} \boldsymbol{s}_j = 0$; [conjugate direction property]
4. *The method terminate (i.e. $\nabla \mathsf{f}(\boldsymbol{x}_m) = \boldsymbol{0}$) at $\boldsymbol{x}_m = \boldsymbol{x}_\star$ with $m \le n$. If $n = m$ then $\boldsymbol{H}_n = \boldsymbol{A}^{-1}$.*

- The Broyden Class update can be written as

$$H_{k+1}^{\theta} = H_{k+1}^{DFP} + \theta w_k w_k^T$$
$$= H_{k+1}^{BFGS} + (\theta - 1) w_k w_k^T$$

where

$$w_k = \left(y_k^T H_k y_k\right)^{1/2} \left[\frac{s_k}{s_k^T y_k} - \frac{H_k y_k}{y_k^T H_k y_k}\right]$$

- For particular values of θ we obtain
  1. $\theta = 0$, the DFP update
  2. $\theta = 1$, the BFGS update
  3. $\theta = s_k^T y_k / (s_k - H_k y_k)^T y_k$ the SR1 update
  4. $\theta = (1 \pm (y_k^T H_k y_k / s_k^T y_k))^{-1}$ the Hoshino update

References

J. Stoer and R. Bulirsch
Introduction to numerical analysis
Springer-Verlag, Texts in Applied Mathematics, **12**, 2002.

J. E. Dennis, Jr. and Robert B. Schnabel
Numerical Methods for Unconstrained Optimization and Nonlinear Equations
SIAM, Classics in Applied Mathematics, **16**, 1996.

Robert B. Schnabel
Minimum Norm Symmetric Quasi-Newton Updates Restricted to Subspaces
Mathematics of Computation, **32**. 1978