Quasi-Newton Methods

Background

Newton's method for finding an extreme point is

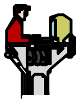
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}^{-1}(\mathbf{x}_k)\,\nabla y(\mathbf{x}_k)$$

Assumption: the evaluation of the Hessian is impractical or costly.

- Central idea underlying quasi-Newton methods is to use an approximation of the inverse Hessian.
- Form of approximation differs among methods.

- The quasi-Newton methods that build up an approximation of the inverse Hessian are often regarded as the most sophisticated for solving unconstrained problems.

Question: What is the simplest approximation?



Modified Newton Method

The Modified Newton method for finding an extreme point is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\, \mathbf{S}_k \nabla y(\mathbf{x}_k)$$

Note that:

if $\mathbf{S}_k = \mathbf{I}$, then we have the method of steepest descent

if $\mathbf{S}_k = \mathbf{H}^{-1}(\mathbf{x}_k)$ and $\alpha = 1$, then we have the "pure" Newton method

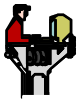
if $y(\mathbf{x}) = 0.5\, \mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T \mathbf{x}$, then $\mathbf{S}_k = \mathbf{H}^{-1}(\mathbf{x}_k) = \mathbf{Q}$ (quadratic case)

Classical Modified Newton's Method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\, \mathbf{H}^{-1}(\mathbf{x}_0) \nabla y(\mathbf{x}_k)$$

Note that the Hessian is only evaluated at the initial point $\mathbf{x}_0$.

Question: What is a measure of effectiveness for the Classical Modified Newton Method?



Quasi-Newton Methods

In quasi-Newton methods, instead of the true Hessian, an initial matrix H_0 is chosen (usually $H_0 = I$) which is subsequently updated by an update formula:

$$H_{k+1} = H_k + H_k^u$$

where H_k^u is the update matrix.

This updating can also be done with the inverse of the Hessian H^{-1} as follows:

Let $B = H^{-1}$; then the updating formula for the inverse is also of the form

$$B_{k+1} = B_k + B_k^u$$

Big question: What is the update matrix?

Hessian Matrix Updates

Given two points $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$, we define $\mathbf{g}_k = \nabla y(\mathbf{x}_k)$ and $\mathbf{g}_{k+1} = \nabla\, y(\mathbf{x}_{k+1})$. Further, let $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, then

$$\mathbf{g}_{k+1} - \mathbf{g}_k \approx \mathbf{H}(\mathbf{x}_k)\, \mathbf{p}_k$$

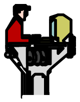
If the Hessian is constant, then

$$\mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{H}\, \mathbf{p}_k \quad \text{which can be rewritten as} \quad \mathbf{q}_k = \mathbf{H}\, \mathbf{p}_k$$

If the Hessian is constant, then the following condition would hold as well

$$\mathbf{H}^{-1}{}_{k+1}\, \mathbf{q}_i = \mathbf{p}_i \qquad\qquad 0 \le i \le k$$

This is called the <u>quasi-Newton condition</u>.



Rank One and Rank Two Updates

Let $\mathbf{B} = \mathbf{H}^{-1}$, then the quasi-Newton condition becomes $\mathbf{B}_{k+1}\, \mathbf{q}_i = \mathbf{p}_i \qquad 0 \leq i \leq k$

Substitute the updating formula $\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{B}^u_k$ and the condition becomes

$$\mathbf{p}_i = \mathbf{B}_k\, \mathbf{q}_i + \mathbf{B}^u_k\, \mathbf{q}_i \qquad\qquad (1)$$

(remember: $\mathbf{p}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ and $\mathbf{q}_i = \mathbf{g}_{i+1} - \mathbf{g}_i$)

Note: There is no unique solution to funding the update matrix $\mathbf{B}^u_k$

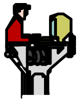
A general form is $\mathbf{B}^u_k = a\, \mathbf{u}\mathbf{u}^T + b\, \mathbf{v}\mathbf{v}^T$

where a and b are scalars and **u** and **v** are vectors satisfying condition (1).

The quantities $a\mathbf{u}\mathbf{u}^T$ and $b\mathbf{v}\mathbf{v}^T$ are symmetric matrices of (at most) rank one.

Quasi-Newton methods that take $b = 0$ are using <u>rank one</u> updates.
Quasi-Newton methods that take $b \neq 0$ are using <u>rank two</u> updates.

Note that $b \neq 0$ provides more flexibility.



Update Formulas

Rank one updates are simple, but have limitations.

Rank two updates are the most widely used schemes.

The rationale can be quite complicated (see, e.g., Luenberger).

The following two update formulas have received wide acceptance:

- Davidon -Fletcher-Powell (DFP) formula
- Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula.

Davidon-Fletcher-Powel Formula

- Earliest (and one of the most clever) schemes for constructing the inverse Hessian was originally proposed by Davidon (1959) and later developed by Fletcher and Powell (1963).
- It has the interesting property that, for a quadratic objective, it simultaneously generates the directions of the conjugate gradient method while constructing the inverse Hessian.
- The method is also referred to as the variable metric method (originally suggested by Davidon).

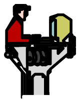
Quasi-Newton condition with rank two update substituted is

$$\mathbf{p}_i = \mathbf{B}_k\, \mathbf{q}_i + a\, \mathbf{u}\mathbf{u}^T\ \mathbf{q}_i + b\, \mathbf{v}\mathbf{v}^T\ \mathbf{q}_i$$

Set $\mathbf{u} = \mathbf{p}_k$, $\mathbf{v} = \mathbf{B}_k\mathbf{q}_k$ and let $a\mathbf{u}^T\mathbf{q}_k = 1$, $b\mathbf{v}^T\mathbf{q}_k = -1$ to determine a and b.

Resulting Davidon-Fletcher-Powell update formula is

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{p}_k\mathbf{p}_k^T}{\mathbf{p}_k^T\mathbf{q}_k} - \frac{\mathbf{B}_k\mathbf{q}_k\mathbf{q}_k^T\mathbf{B}_k}{\mathbf{q}_k^T\mathbf{B}_k\mathbf{q}_k}$$



Broyden–Fletcher–Goldfarb–Shanno Formula

Remember that $\mathbf{q}_i = \mathbf{H}_{k+1}\,\mathbf{p}_i$ and $\mathbf{H}^{-1}{}_{k+1}\,\mathbf{q}_i = \mathbf{p}_i$ (or, $\mathbf{B}_{k+1}\,\mathbf{q}_i = \mathbf{p}_i$) $\quad 0 = i = k$

Both equations have exactly the same form, except that $\mathbf{q}_i$ and $\mathbf{p}_i$ are interchanged and $\mathbf{H}$ is replaced by $\mathbf{B}$ (or vice versa).

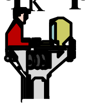
This leads to the observation that any update formula for $\mathbf{B}$ can be transformed into a corresponding complimentary formula for $\mathbf{H}$ by interchanging the roles of $\mathbf{B}$ and $\mathbf{H}$ and of $\mathbf{q}$ and $\mathbf{p}$. The reverse is also true.

Broyden–Fletcher–Goldfarb–Shanno formula update of $\mathbf{H}_k$ is obtained by taking the complimentary formula of the DFP formula, thus:

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{q}_k\mathbf{q}_k{}^T}{\mathbf{q}_k{}^T\mathbf{p}_k} - \frac{\mathbf{H}_k\mathbf{p}_k\mathbf{p}_k{}^T\mathbf{H}_k}{\mathbf{p}_k{}^T\mathbf{H}_k\mathbf{p}_k}$$

By taking the inverse, the BFGS update formula for $\mathbf{B}_{k+1}$ (i.e., $\mathbf{H}^{-1}{}_{k+1}$) is obtained:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \left(\frac{1 + \mathbf{q}_k{}^T\mathbf{B}_k\mathbf{q}_k}{\mathbf{q}_k{}^T\mathbf{p}_k}\right)\frac{\mathbf{p}_k\mathbf{p}_k{}^T}{\mathbf{p}_k{}^T\mathbf{q}_k} - \frac{\mathbf{p}_k\mathbf{q}_k{}^T\mathbf{B}_k + \mathbf{B}_k\mathbf{q}_k\mathbf{p}_k{}^T}{\mathbf{q}_k{}^T\mathbf{p}_k}$$



Some Comments on Broyden Methods

- Broyden–Fletcher–Goldfarb–Shanno formula is more complicated than DFP, but straightforward to apply
- BFGS update formula can be used exactly like DFP formula.
- Numerical experiments have shown that BFGS formula's performance is superior over DFP formula. Hence, BFGS is often preferred over DFP.

Both DFP and BFGS updates have symmetric rank two corrections that are constructed from the vectors pk and Bkqk. Weighted combinations of these formulae will therefore also have the same properties. This observation leads to a whole collection of updates, know as the Broyden family, defined by:

$$\mathbf{B}^{\phi} = (1 - \phi)\mathbf{B}^{\mathrm{DFP}} + \phi\mathbf{B}^{\mathrm{BFGS}}$$

where ϕ is a parameter that may take any real value.

Quasi-Newton Algorithm

1. Input $\mathbf{x}_0$, $\mathbf{B}_0$, termination criteria.

2. For any k, set $\mathbf{S}_k = -\,\mathbf{B}_k\mathbf{g}_k$.

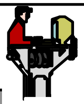
3. Compute a step size α (e.g., by line search on $y(\mathbf{x}_k + \alpha\mathbf{S}_k)$) and set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha\mathbf{S}_k$.

4. Compute the update matrix $\mathbf{B}^u{}_k$ according to a given formula (say, DFP or BFGS) using the values $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$, $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, and $\mathbf{B}_k$.

5. Set $\mathbf{B}_{k+1} = \mathbf{B}_k + \mathbf{B}^u{}_k$.

6. Continue with next k until termination criteria are satisfied.

Note: You do have to calculate the vector of first order derivatives **g** for each iteration.



Some Closing Remarks

- Both DFP and BFGS methods have theoretical properties that guarantee superlinear (fast) convergence rate and global convergence under certain conditions.

- However, both methods could fail for general nonlinear problems. Specifically,

  • DFP is highly sensitive to inaccuracies in line searches.

  • Both methods can get stuck on a saddle-point. In Newton's method, a saddle-point can be detected during modifications of the (true) Hessian. Therefore, search around the final point when using quasi-Newton methods.

  • Update of Hessian becomes "corrupted" by round-off and other inaccuracies.

- All kind of "tricks" such as scaling and preconditioning exist to boost the performance of the methods.