Operations Research Techniques and Algorithms (620-361)

Dr Yao-ban Chan

y.chan@ms.unimelb.edu.au
Telephone: 8344 9073
Office: Room 198, Richard Berry Building

Christina Burt
c.burt@ms.unimelb.edu.au
Telephone: 8344 1797
Office: 139 Barry St

Friday 4th April, 2008

Today's Lecture

Unimodal n-D unconstrained optimisation
- BFGS

The BFGS method is the general descent method in which d^k is chosen as the BFGS direction, that is $d^k = -H_k \nabla f(x^k)$ where H_k is constructed using the BFGS update below.

The BFGS Method

To minimise a unimodal function $f : \Re^N \to \Re$ to within tolerance ϵ.

1. Select $x^0 \in \Re^N$.
   Set $k = 0$. Set $H_0 \in \Re^{n \times n}$ to be a symmetric positive definite matrix (for example $H_0 = I$).
2. If $||\nabla f(x^k)|| < \epsilon$ then stop.
   Set $d^k = -H_k \nabla f(x^k)$.

3. Select step length t_k either
   - by solving the single-variable minimisation problem: $\min q(t) = f(x^k + td^k)$.
   - by using our procedure for finding a step length that staisfies the Armijo-Goldstein and Wolff conditions.

4. Set $k = k + 1$.
   Set $x^{k+1} = x^k + t_k d^k$.
   Update H_k as follows:

$$\begin{aligned} s^k &= x^{k+1} - x^k \\ g^k &= \nabla f(x^{k+1}) - \nabla f(x^k) \\ r^k &= H_k g^k / \langle s^k, g^k \rangle \\ H_{k+1} &= H_k + \frac{1 + \langle r^k, g^k \rangle}{\langle s^k, g^k \rangle} s^k (s^k)^T - [s^k (r^k)^T + r^k (s^k)^T] \end{aligned}$$

   Return to step 2.

Note: *Outer products* such as $s^k(s^k)^T$ and $s^k(r^k)^T$ are actually $n \times n$ matrices, because they are each the product of an $n \times 1$ matrix (column vector) with a $1 \times n$ matrix (row vector).
For example, if

$$s^k = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; r^k = \begin{bmatrix} -3 \\ 2 \end{bmatrix},$$

then

$$s^k(r^k)^T = \begin{bmatrix} 1 \\ 1 \end{bmatrix} [-3 \quad 2] = \begin{bmatrix} -3 & 2 \\ -3 & 2 \end{bmatrix}.$$

Why this weird formula?

That is a good question! To see why, we look briefly at two other quasi-Newton methods - the rank-one correction formula and the DFP algorithm.

All of these methods (rank-one, DFP, BFGS) are quasi-Newton methods and follow the same ideas. Their sole difference lies in choice of H_k. They also all ensure that

$$H_{k+1}(\nabla f(x^{k+1}) - f(x^k)) = x^{k+1} - x^k$$

as discussed above.

Furthermore, they all select H_k by updating from H_{k-1}. The difference lies in how they update it.

In the rank-one correction method, we try to update H_k as little as possible. To do this, we add a matrix of rank one to it:

$$H_{k+1} = H_k + z^k (z^k)^T.$$

After some calculation, which I will not give here, it turns out that ensuring $H_{k+1}(\nabla f(x^{k+1}) - f(x^k)) = x^{k+1} - x^k$ gives the formula

$$H_{k+1} = H_k + \frac{(s^k - H_k g^k)(s^k - H_k g^k)^T}{(g^k)^T (s^k - H_k g^k)}$$

using the notation given before.

So far so good; but this approach has problems. In particular, the updating scheme does not preserve positive definiteness. The DFP updating scheme, however, does.
The DFP scheme requires a rank 2 correction:

$$H_{k+1} = H_k + y^k(y^k)^T + z^k(z^k)^T.$$

To ensure that this works, we need

$$\begin{aligned} H_{k+1}(\nabla f(x^{k+1}) - \nabla f(x^k)) &= (H_k + y^k(y^k)^T + z^k(z^k)^T)g^k \\ &= x^{k+1} - x^k = s^k. \end{aligned}$$

Therefore

$$y^k(y^k)^T g^k + z^k(z^k)^T g^k = s^k - H_k g^k.$$

There are a few ways to make this happen; the simplest way (but not necessarily most effective!) is just to choose y^k and z^k so that the first terms on each side match, and the second terms on each side match. Then

$$y^k = \frac{1}{(y^k)^T g^k} s^k$$

$$y^k(y^k)^T = \frac{1}{((y^k)^T g^k)^2} s^k (s^k)^T$$

which is good, except for the (y^k) on the right-hand side.

To get rid of this we notice that

$$(g^k)^T y^k (y^k)^T g^k = [(y^k)^T g^k]^T (y^k)^T g^k = (g^k)^T s^k$$

and since $(y^k)^T g^k$ is a scalar, the left-hand side is $((y^k)^T g^k))^2$. So

$$y^k (y^k)^T = \frac{1}{\langle g^k, s^k \rangle} s^k (s^k)^T.$$

We can do the same thing for z^k.

$$z^k = -\frac{1}{(z^k)^T g^k} H_k g^k$$

$$z^k(z^k)^T = \frac{1}{((z^k)^T g^k)^2}(H_k g^k)(H_k g^k)^T$$

and again we have to get rid of z^k on the right-hand side.

$$(g^k)^T z^k (z^k)^T g^k = -(g^k)^T H_k g^k$$

and again, the left-hand side is $((z^k)^T g^k)^2$. So

$$z^k(z^k)^T = -\frac{1}{\langle g^k, H_k g^k \rangle}(H_k g^k)(H_k g^k)^T.$$

Putting it all together, we arrive at the DFP update scheme:

$$\begin{aligned} H_{k+1} &= H_k + y^k (y^k)^T + z^k (z^k)^T \\ &= H_k + \frac{1}{\langle g^k, s^k \rangle} s^k (s^k)^T - \frac{1}{\langle g^k, H_k g^k \rangle} (H_k g^k)(H_k g^k)^T. \end{aligned}$$

By inserting this formula into the general quasi-Newton framework, we get the DFP algorithm.

In fact not only does this satisfy the property

$$H_{k+1}(\nabla f(x^{k+1}) - \nabla f(x^k)) = x^{k+1} - x^k,$$

but it turns out that

$$H_{k+1}(\nabla f(x^{i+1}) - \nabla f(x^i)) = x^{i+1} - x^i$$

for any $0 \leq i \leq k$!

As if this wasn't complicated enough, some bright fellows decided to make it even more complicated!
Broyden, Fletcher, Goldfarb and Shanno noticed that the DFP algorithm updates the Hessian to preserve the relation

$$H_{k+1}g^k = s^k.$$

But, if we turn it around and have a sequence of matrices B_k which satisfy

$$B_{k+1}s^k = g^k,$$

then we can invert B_k to get H_k. Furthermore, we can actually use the formula we use to update H_k to update B_k if we simply swap g^k and s^k around!

This leads to the BFGS update system:

$$B_k = H_k^{-1}$$

$$B_{k+1} = B_k + \frac{1}{\langle s^k, g^k \rangle} g^k (g^k)^T - \frac{1}{\langle s^k, B_k s^k \rangle} (B_k s^k)(B_k s^k)^T$$

$$H_{k+1} = B_{k+1}^{-1}$$

This is a fairly horrendous mess by now, but it can be simplified somewhat by means of the following lemma.

Lemma 8: *Let A be a nonsingular matrix, and let u and v be column vectors such that $1 + v^T A^{-1} u \neq 0$. Then $A + uv^T$ is nonsingular, and*

$$(A + uv^T)^{-1} = A^{-1} - \frac{(A^{-1}u)(v^T A^{-1})}{1 + v^T A^{-1} u}.$$

Proof: Assignment question! (Maybe.)

By using the lemma twice on B_{k+1}, we obtain the formula for H_{k+1} that we specified above.
It turns out that the BFGS direction also satisfies the property

$$H_{k+1}(\nabla f(x^{i+1}) - \nabla f(x^i)) = x^{i+1} - x^i$$

for any $0 \leq i \leq k$.

Example: (same as before) Minimise

$$f(x) = x_1^2 + x_2^2 - x_1 x_2 - 3x_1 + 3x_2 + 3$$

starting from the point $x^0 = (0, 0)$.