Quasi-Newton methods

Problem: Hessian is expensive to compute, solving with it even more so... need an alternative. Approximate it?

Basic idea: replace $\nabla^2 f(x_k)$ with approximation B_k. To compute search step, solve:

$$B_k p_k = -\nabla f(x_k). \qquad (*)$$

Note: B_k may not be full rank! Hence, B_k^{-1} may not exist, and need to solve $(*)$ in the least-squares sense:

$$\underset{p_k}{\text{minimize}} \; \| B_k p_k + \nabla f(x_k) \|_2^2 \quad \rightarrow \quad p_k = B_k^{\dagger} \nabla f(x_k).$$

If B_k is positive definite, equivalent to minimizing:

$$\underset{p_k}{\text{minimize}} \; f(x_k) + \nabla f(x_k)^T p_k + \frac{1}{2} p_k^T B_k p_k.$$

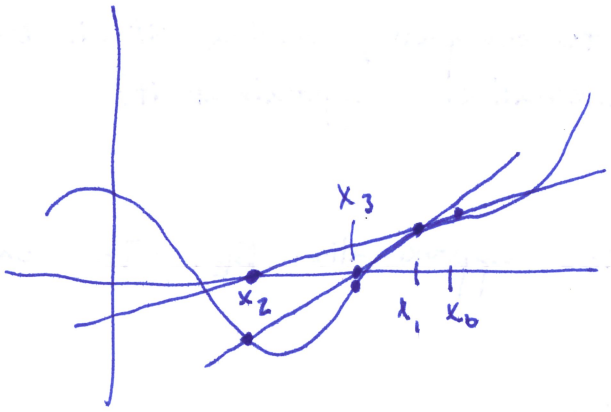
In 1D, often we use the Secant method instead of Newton's methods where we approximate $f'(x_k)$ with:

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

Giving the iteration:

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f'(x_k), \quad k \geq 1.$$

What's the idea?



The secant method converges superlinearly with order of $q = 1.618\ldots = \frac{1}{2}(1+\sqrt{5})$. ("Golden ratio")

Can easily be applied to minimize a function, in which case the approximation

$$f''(x_k) \approx \frac{f'(x_k) - f'(x_{k-1})}{x_k - x_{k-1}} \qquad (\#)$$

is used.

We want to generalize this idea to approximate the Hessian, $\nabla^2 f(x_k)$. Note that the Hessian is the natural generalization of f'' in this setting. (i.e., trying to minimize a function).

Multiplying both sides of $(\#)$ by $x_k - x_{k-1}$ gives:

$$f''(x_k)(x_k - x_{k-1}) = f'(x_k) - f'(x_{k-1}).$$

If we replace f'' with B_k and f' with ∇f we get:

$$\boxed{B_k(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1}).}$$

This is called the secant condition.

Note: this is a system which B_k must satisfy, but B_k has n^2 degrees of freedom and there are only n equations here. So need more constraints to define B_k uniquely.

How do we make sense of the secant condition? Let:

$$f(x) = \frac{1}{2} x^T Q x + c^T x, \qquad Q = Q^T.$$

Then:

$$\nabla f(x) = \frac{1}{2}(Q + Q^T)x + c = Q^T x + c = Qx + c.$$

And:

$$\nabla^2 f(x) = Q.$$

Compare w/ secant condition with $B_k = \nabla^2 f$:

$$\nabla^2 f(x_k)(x_k - x_{k-1}) = Q(x_k - x_{k-1})$$
$$= Qx_k - Qx_{k-1} = Qx_k + c - Qx_{k-1} - c$$
$$= \nabla f(x_k) - \nabla f(x_{k-1}).$$

Hence, the secant condition is a constraint on the action of a quadratic form in a particular direction: i.e, the $x_k - x_{k-1}$ direction. Of course, we will need to set the action of the Hessian in many directions. This suggests that we need to do more.

Let's take a look at some quasi-Newton updates. If B_k is our approximation of $\nabla^2 f(x_k)$, typically quasi-Newton approximations are built up over many iterations as a running sum of the form:

$$B_{k+1} = B_k + C_k,$$

where C_k is different depending on the quasi-Newton method being used.

We will define the vectors:

$$s_k = x_{k+1} - x_k$$

$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

Then the secant condition takes the form:

$$B_{k+1} s_k = y_k.$$

Note that if $x_{k+1} = x_k + \alpha_k p_k$, then $s_k = \alpha_k p_k$.

To come up with a Quasi-Newton method, we typically decide what properties we would like B_k to have, and then find C_k such that the secant condition is maintained. Here is one example.

Theorem: SR1 update formula (SR1 = "symmetric rank 1")

Let B_k, B_{k+1} both be symmetric. Assume $B_{k+1} = B_k + C_k$, where $C_k \neq 0$, C_k is rank 1, and such that $B_{k+1} s_k = y_k$ and $(y_k - B_k s_k)^T s_k = 0$ holds. Then:

$$C_k = \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s)^T s_k}.$$

Proof: Since C_k is symmetric and rank one, we can write it as:

$$C_k = \gamma u u^T,$$

where $\gamma \neq 0$, and $\|w\|_2 = 1$. From the secant condition, we have:

$$y_k = B_{k+1} s_k = (B_k + C_k) s_k = B_k s_k + \gamma w w^T s_k.$$

Hence:

$$y_k - B_k s_k = \gamma (w^T s_k) w. \quad (**)$$

Since $C_k \neq 0$, $\gamma \neq 0$. If $w^T s_k = 0$, then $y_k - B_k s_k = 0$. But by assumption, $(y_k - B_k s_k)^T s_k \neq 0$, so this would contradict our assumptions. So, it must be the case that $w^T s_k \neq 0$. Hence, w equals a nonzero constant times $y_k - B_k s_k$. Since $\|w\|_2 = 1$, we can conclude that:

$$w = \frac{y_k - B_k s_k}{\|y_k - B_k s_k\|_2}.$$

Now we just need to find γ. Now that we know w, write:

$$y_k - B_k s_k = \gamma \left[\left(\frac{y_k - B_k s_k}{\|y_k - B_k s_k\|_2} \right)^T s_k \right] \frac{y_k - B_k s_k}{\|y_k - B_k s_k\|_2}$$

from $(**)$. Dot both sides with $y_k - B_k s_k$ to get:

$$\|y_k - B_k s_k\|_2^2 = \gamma (y_k - B_k s_k)^T s_k.$$

Since $(y_k - B_k s_k)^T s_k \neq 0$ by assumption:

$$\gamma = \frac{\|y_k - B_k s_k\|_2^2}{(y_k - B_k s_k)^T s_k}.$$

This proves the result ∎

The general form for a quasi-Newton method is:

1. Compute warm start x_0.
2. Compute initial guess B_0 (e.g. $B_0 = I$).
3. For $k = 0, 1, \ldots$:
   a. If $\nabla f(x_k) \approx 0$, stop.
   b. Solve $B_k p_k = -\nabla f(x_k)$ for p_k.
   c. Use line search to find α_k in $x_{k+1} = x_k + \alpha_k p_k$.
   d. Update:

$$s_k = x_{k+1} - x_k$$
$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$
$$B_{k+1} = B_k + C_k(s_k, y_k).$$

Instead of trying to preserve the symmetry of B_k, we can try to preserve some other property: e.g., B_k positive definite.

Interesting fact: there is no rank 1 update which preserves both symmetry and positive definiteness.

There are infinitely many rank 2 updates which do.

An example of a family of updates which is symmetric and preserves positive definiteness of B_k is the Broyden class:

$$B_{k+1} = B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k} + \phi \cdot (s_k^T B_k s_k) v_k v_k^T$$

$$v_k = \frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k},$$

$\phi \in \mathbb{R}$.

Note that positive definiteness is ensured only if $y_k^T s_k > 0$, which is typically enforced using a line search.

Two important cases:

1) $\phi = 0 \rightsquigarrow$ "BFGS"
2) $\phi = 1 \rightsquigarrow$ "DFP"

} named after researchers who discovered

The most widely used quasi-Newton method is BFGS:

$$B_{k+1} = B_k - \frac{(B_k s_k)(B_k s_k)^T}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

See Grive for a convergence theorem for quasi-Newton methods from the Broyden family.