Line search methods. Rate of convergence.

- Rate of convergence of steepest descent, Newton’s method, Quasi-Newton’s methods.
- Practical step length selection

**Review**

*Search directions*

**Steepest descent**

\[ p_k = -\nabla f(x_k)/\|\nabla f(x_k)\|, \]

**Newton’s method**

\[ p_k = -\left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k), \]

**Quasi-Newton’s method**

\[ p_k = -\left[B_k\right]^{-1}\nabla f(x_k), \]

where \( B_k \) approximates \( \nabla^2 f(x_k) \), for example BFGS, Broyden, Fletcher, Goldfarb, Shanno formula proposes

\[ B_{k+1} = B_k - \frac{(B_k s_k)(B_k s_k)^t}{\langle s_k, B_k s_k \rangle} + \frac{y_k y_k^t}{\langle s_k, y_k \rangle} \]

where

\[ s_k = \alpha_k p_k = (x_{k+1} - x_k), \ y_k = \nabla f(x_{k+1}) - \nabla f(x_k) \]

The question why \( B_k \approx \nabla^2 f(x_k) \) will be addressed later.

**Step-length**

**Wolfe condition**

\[ f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \langle \nabla f(x_k), p_k \rangle, \quad \langle \nabla f(x_k + \alpha p_k), p_k \rangle \geq c_2 \langle \nabla f(x_k), p_k \rangle \]

**Strong Wolfe condition**

\[ f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \langle \nabla f(x_k), p_k \rangle, \quad \langle \nabla f(x_k + \alpha p_k), p_k \rangle \leq c_2 \langle \nabla f(x_k), p_k \rangle \]

**Backtracking and sufficient decrease**

\[
\text{if } f(x_k + \alpha^i p_k) \leq f(x_k) + c\alpha^i \langle \nabla f(x_k), p_k \rangle \alpha_k = \alpha^i \text{ else } \alpha^i = \rho \alpha^i \text{ end}
\]

Our goal is to assess the rate of convergence.

*Newton’s method* has a quadratic rate of convergence, convergence of order 2, when \( \alpha_k \equiv 1 \).

**Proof** Locally

\[ ||(\nabla f(x_k) - \nabla f(x^*)) - \nabla^2 f(x_k)(x_k - x^*)|| \leq C||x_k - x^*||^2 \]

Since for the Newton’s method

\[ x_{k+1} - x^* = x_k - x^* - \left[\nabla^2 f(x_k)\right]^{-1}\nabla f(x_k) = \left[\nabla^2 f(x_k)\right]^{-1}\left[\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))\right] \]
where $\nabla f(x^*) = 0$, we have
$$||x_{k+1} - x^*|| \leq C||x_k - x^*||^2.$$  

As a byproduct we have a quadratic rate of convergence of gradients:

$$||\nabla f(x_{k+1})|| = ||\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k)|| \leq C||x_{k+1} - x_k||^2 \leq C||\nabla f(x_k)||^2$$

**Question** Why does the first equality and the last inequality in (*) hold?

**Question** What does $C$ depend on? Are the $C$’s in the proof all the same?

A Quasi-Newton's method has a superlinear rate of convergence if

$$\lim_{k \to \infty} \frac{||(B_k - \nabla^2 f(x^*))p_k||}{||p_k||} = 0.$$

**Proof:**

$$||p_k - p^N_k||/||p_k|| \leq C||(B_k - \nabla^2 f(x^*))p_k||/||p_k|| \to 0,$$

as $k \to \infty$, where $p^N_k$ is the Newton’s step. Hence the triangle inequality gives the result

$$||x_{k+1} - x^*|| \leq ||x_k + p^N_k - x^*|| + ||p_k - p^N_k||.$$

**Remark** A useful notation is $O(x)$ and $o(x)$, what does it mean?

**Steepest descent for is linearly convergent.**

**Proof** Locally

$$f(x_k + h) = f(x_k) + \langle \nabla f(x_k)h \rangle + \frac{1}{2}\langle h, \nabla^2 f(x_k)h \rangle + O(h^3).$$

We choose the step-length $\alpha_k$ of the steepest descent so that it minimizes the quadratic part of $f(x + h)$:

$$\alpha_k = \frac{||\nabla f(x_k)||^2}{\langle \nabla f(x_k), \nabla^2 f(x_k)\nabla f(x_k) \rangle} = \frac{||\nabla f(x_k)||^2}{\langle \nabla f(x_k), Q\nabla f(x_k) \rangle}(1 + O(||x_k - x^*||)), $$

where $Q = \nabla^2 f(x^*)$ is positive-definite. Hence

$$x_{k+1} = x_k - \frac{||\nabla f(x_k)||^2}{\langle \nabla f(x_k), Q\nabla f(x_k) \rangle}(1 + O(||x_k - x^*||))\nabla f(x_k).$$

Noting that

$$\nabla f(x_k) = Q(x_k - x^*)(1 + O(||x_k - x^*||))$$

we have

$$||x_k - x^*||^2_Q = (f(x_k) - f(x^*)(1 + O(||x_k - x^*||))$$

where

$$||x_k - x^*||^2_Q = \langle (x_k - x^*), Q(x_k - x^*) \rangle.$$  

Hence (see homework this week)

$$||x_{k+1} - x^*||^2_Q = C(Q)(1 + O(||x_k - x^*||)||x_k - x^*||^2_Q)$$
where
\[ C(Q) = \left[ 1 - \frac{||\nabla f(x_k)||^4}{||\nabla f(x_k)||^2 ||\nabla f(x_k)||^2 ||Q^{-1}||} \right], \]
or using the condition number
\[ \kappa(Q) = ||Q|| ||Q^{-1}|| \]
\[ C(Q) \leq 1 - \frac{1}{\kappa^2(Q)}. \]
Hence, altogether
\[ ||x_{k+1} - x^*||_Q \leq M ||x_k - x^*||_Q \]
where
\[ M \approx \sqrt{1 - 1/\kappa^2(Q)} \]

Our next goal is to discuss practical ways of step-length selection.
A procedure, that computes \( \nabla f \) as few times as possible.
*Step-length selection=bracketing phase + selection phase.*

Example 1. Interpolation
Approximate a function
\[ \phi(\alpha) = f(x + ah_k) \]
by a cubic polynomial
\[ a\alpha^3 + b\alpha^2 + c\alpha + d \]
such that values \( \phi(0), \phi'(0), \phi(\alpha_k), \phi(\alpha_{k+1}) \) agree with the corresponding values of the cubic polynomial.

Example 2. Use of Wolfe condition: Bracketing and zooming.