Nonlinear Optimization

Overview of methods; the Newton method with line search

Niclas Börlin

Department of Computing Science

Umeå University
niclas.borlin@cs.umu.se

November 19, 2007

Overview

Most deterministic methods for unconstrained optimization have the following features:

- They are *iterative*, i.e. they start with an initial guess x_0 of the variables and tries to find "better" points $\{x_k\}, k = 1, \ldots$.
- They are *descent methods*, i.e. at each iteration k,

$$f(x_{k+1}) < f(x_k)$$

  is (at least) required.
- At each iteration k, the nonlinear objective function f is replaced by a simpler *model function* m_k that approximates f around x_k.
- The next iterate $x_{k+1} = x_k + p$ is sought as the minimizer of m_k.

- The model function m_k is usually defined to be a quadratic function of the form

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p,$$

  where $f_k = f(x_k)$, $\nabla f_k = \nabla f(x_k)$, and B_k is a matrix, usually a positive definite approximation of the hessian $\nabla^2 f(x_k)$.
- If B_k is positive definite, a minimizer of m_k may be found by solving

$$\nabla_p m_k(x_k + p) = 0$$

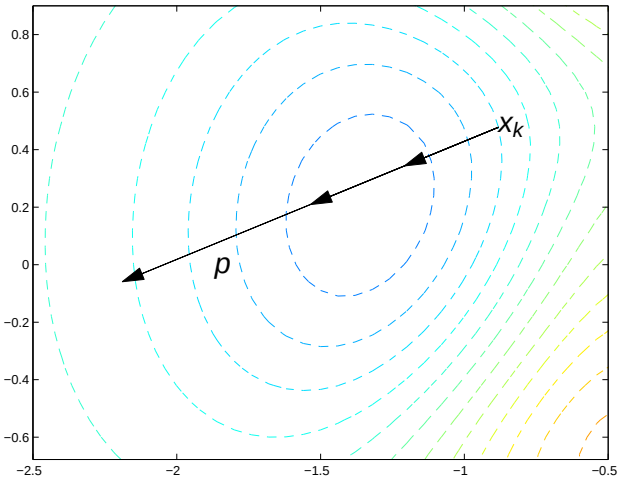
  for p.
- If the minimizer of m_k does not produce a better point, the *step* p is modified to produce a point $x_{k+1} = x_k + p$ that is better.
- The modifications come in two major flavours: *line search* and *trust-region*.

Line search

- In the line search strategy, the algorithm chooses a *search direction* p_k and tries to solve the following one-dimensional minimization problem

$$\min_{\alpha > 0} f(x_k + \alpha p_k),$$

  where the scalar α is called the *step length*.
- In theory we would like optimal step lengths, but in practice it is more efficient to test trial step lengths until we find one that gives us a good enough point.



Trust-region

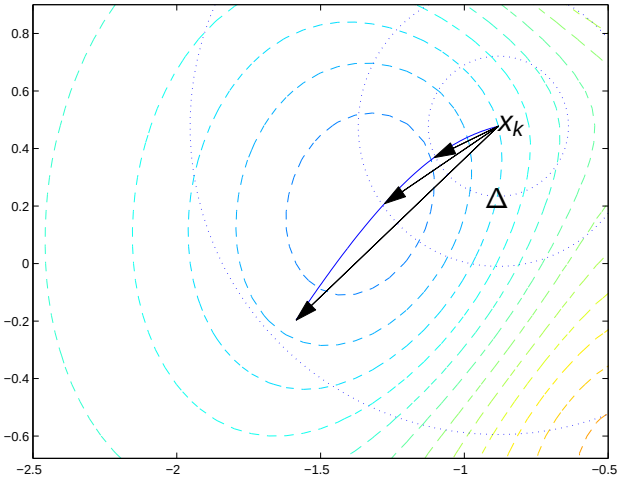
- In the trust-region strategy, the algorithm defines a *region of trust* around x_k where the current model function m_k is trusted.
- The region of trust is usually defined as

$$\|p\|_2 \leq \Delta,$$

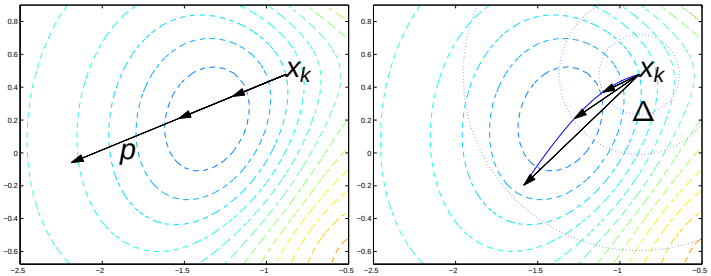
  where the scalar Δ is called the trust-region radius.
- A candidate step p is found by approximately solving the following subproblem

$$\min_p m_k(x_k + p) \text{ s.t. } \|p\|_2 \leq \Delta.$$

- If the candidate step does not produce a good enough new point, we shrink the trust-region radius and re-solve the subproblem.



- In the line search strategy, the direction is chosen first, followed by the distance.
- In the trust-region strategy, the maximum distance is chosen first, followed by the direction.



Convergence rate

- In order to compare different iterative methods, we need an efficiency measure.
- Since we do not know the number of iterations in advance, the *computational complexity* measure used by direct methods cannot be used.
- Instead the concept of a *convergence rate* is defined.

- Assume we have a series $\{x_k\}$ that converges to a solution x^*. Define the sequence of errors as

$$e_k = x_k - x^*$$

and note that

$$\lim_{k\to\infty} e_k = 0.$$

- We say that the sequence $\{x_k\}$ converges to x^* with rate r and rate constant C if

$$\lim_{k\to\infty} \frac{\|e_{k+1}\|}{\|e_k\|^r} = C$$

and $C < \infty$.

In practice there are three important rates of convergence:

- *linear convergence*, for $r = 1$ and $0 < C < 1$;
- *quadratic convergence*, for $r = 2$.
- *super-linear convergence*, for $r = 1$ and $C = 0$.

Linear convergence, examples

- For $r = 1$, $C = 0.1$ and $\|e_0\| = 1$, the norm of the error sequence becomes

$$\underbrace{1, 10^{-1}, 10^{-2}, \ldots, 10^{-7}}_{\text{7 iterations}}$$

- For $C = 0.99$ the corresponding sequence is

$$\underbrace{1, 0.9, 0.9801, \ldots, 0.997 \cdot 10^{-7}}_{\text{1604 iterations}}.$$

- Thus the constant C is of major importance for a method with linear convergence.

Quadratic convergence, examples

- For $r = 2$, $C = 0.1$ och $\|e_0\| = 1$, the sequence becomes
$$1, 10^{-1}, 10^{-3}, 10^{-7}, \ldots$$
- For $r = 2$, $C = 3$ och $\|e_0\| = 1$, the sequence diverges
$$1, 3, 27, \ldots$$
- For $r = 2$, $C = 3$ och $\|e_0\| = 0.1$, the sequence becomes
$$0.1, 0.03, 0.0027, \ldots,$$
i.e. it converges despite $C > 1$.
- For quadratic convergence, the constant C is of lesser importance. Instead it is important that the initial approximation is "close enough" to the solution, i.e. $\|e_0\|$ is small.

Local vs. global convergence

- A method is called *locally convergent* if it produces a convergent sequence toward a minimizer x^* provided a *close enough* starting approximation.
- A method is called *globally convergent* if it produces a convergent sequence toward a minimizer x^* provided *any* starting approximation.
- Note that global convergence does not imply convergence towards a global minimizer.

Globalization strategies

- The line search and trust-region methods are sometimes called globalization strategies, since they modify a "core" method (typically locally convergent) to become globally convergent.
- There are two efficiency requirements on any globalization strategy:
  - Far from the solution, they should stop the methods from going out of control.
  - Close to the solution, when the "core" method is efficient, they should interfere as little as possible.

Descent directions

- Consider the Taylor expansion of the objective function along a search direction p
$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + \tau p) p,$$
for some $\tau \in (0, \alpha)$
- Any direction p such that $p^T \nabla f_k < 0$ will produce a reduction of the objective function for a short enough step.
- A direction p such that
$$p^T \nabla f_k < 0$$
is called a *descent direction*.

- If the search direction has the form

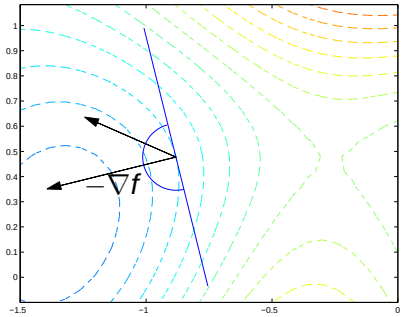
$$p_k = -B_k^{-1}\nabla f_k,$$

the descent condition

$$p_k^T \nabla f_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0$$

is satisfied whenever B_k is positive definite.

- Since $\cos\theta = \frac{-p^\nabla f_k}{\|p\|\|\nabla f_k\|}$ is the angle between the search direction and the negative gradient, descent directions are in the same half-plane as the negative gradient.
- The search direction corresponding to the negative gradient $p = -\nabla f_k$ is called the direction of *steepest descent*.



Line search

- Each iteration of a line search method computes a search direction p_k and then decides how far to move along that direction.
- The next iteration is given by

$$x_{k+1} = x_k + \alpha_k p_k$$

- We will require p_k to be a descent direction. This assures that the objective function will decrease

$$f(x_k + \alpha_k p_k) < f(x_k)$$

for some small $\alpha_k > 0$.

Exact and inexact line searches

- Consider the function

$$\phi(\alpha) = f(x_k + \alpha p_k), \quad \alpha > 0.$$

- Ideally we would like to find the global minimizer of ϕ for every iteration. This is called an *exact* line search.
- However, it is possible to construct *inexact* line search methods that produce an adequate reduction of f at a minimal cost.
- Inexact line search methods construct a number of candidate values for α and stop when certain conditions are satisfied.

The Sufficient Decrease Condition

- Mathematically, the descent condition $f(x_k + \alpha p_k) < f(x_k)$ is not enough to guarantee convergence.
- Instead, the *sufficient decrease* condition is formulated from the linear Taylor approximation of $\phi(\alpha)$

$$\phi(\alpha) \approx \phi(0) + \alpha\phi'(0)$$

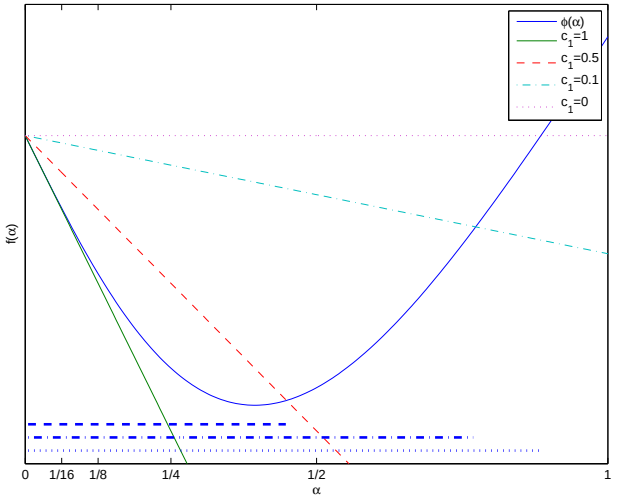
or

$$f(x_k + \alpha p_k) \approx f(x_k) + \alpha \nabla f_k^T p_k.$$

- The sufficient decrease condition states that the new point must at least produce a fraction $0 < c_1 < 1$ of the decrease predicted by the Taylor approximation, i.e.

$$f(x_k + \alpha p_k) < f(x_k) + c_1 \alpha \nabla f_k^T p_k.$$

- This condition is sometimes called the *Armijo* condition.



Backtracking

- The sufficient decrease condition alone is not enough to guarantee convergence, since it is satisfied for arbitrarily small values of α.
- The sufficient decrease condition has to be combined with a strategy that favours large step lengths over small.
- A simple such strategy is called *backtracking*: Accept the first element of the sequence

$$1, \frac{1}{2}, \frac{1}{4}, \ldots, 2^{-i}, \ldots$$

that satisfies the sufficient decrease condition. Such a step length always exist.

- Large step lengths are tested before small ones. Thus, the step length will not be too small.
- This technique works well for Newton-type algorithms.

The Curvature Condition

- Another approximation to the solution of

$$\min_{\alpha>0} \phi(\alpha) \equiv f(x_k + \alpha p_k)$$

is to solve for $\phi'(\alpha) = 0$, which is approximated to the condition

$$|\phi'(\alpha_k)| \le c_2 |\phi'(0)|,$$

where c_2 is a constant $c_1 < c_2 < 1$.

- Since $\phi'(\alpha) = p_k^T \nabla f(x_k + \alpha p_k)$, we get

$$|p_k^T \nabla f(x_k + \alpha_k p_k)| \le c_2 |p_k^T \nabla f(x_k)|.$$

This condition is called the *curvature condition*.

The Wolfe Condition

- The sufficient decrease condition and the curvature condition

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k,$$
$$|p_k^T \nabla f(x_k + \alpha_k p_k)| \leq c_2 |p_k^T \nabla f(x_k)|,$$

  where $0 < c_1 < c_2 < 1$, are collectively called the *strong Wolfe conditions*.
- Step length methods that use the Wolfe conditions are more complicated than backtracking.
- Several popular implementations of nonlinear optimization routines are based on the Wolfe conditions, notably the BFGS quasi-Newton method.

The Newton-Raphson method in $\Re^1$

- Consider the non-linear problem $f(x) = 0$, where $f, x \in \Re$.
- The Newton-Raphson method for solving this problem is based on the linear Taylor approximation of f around x_k

$$f(x_k + p) \approx f(x_k) + pf'(x_k).$$

- If $f'(x_k) \neq 0$ we solve the linear equation

$$f(x_k) + pf'(x_k) = 0$$

  for p and get

$$p = -f(x_k)/f'(x_k).$$

- The new iterate is given by

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

The Classical Newton minimization method in $\Re^n$

- In order to use Newton's method to find a minimizer we apply the first-order necessary conditions on a function f

$$\nabla f(x) = 0 \qquad (f'(x) = 0)$$

- This results in the Newton sequence

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \qquad (x_{k+1} = x_k - f'(x)/f''(x))$$

- This is often written as $x_{k+1} = x_k + p_k$, where p_k is the solution of the *Newton equation*:

$$\nabla^2 f(x_k) p_k = -\nabla f(x_k).$$

  This formulation emphasizes that a linear equation system is solved in each step, usually by other means than calculating an inverse.

Geometrical interpretation; the model function

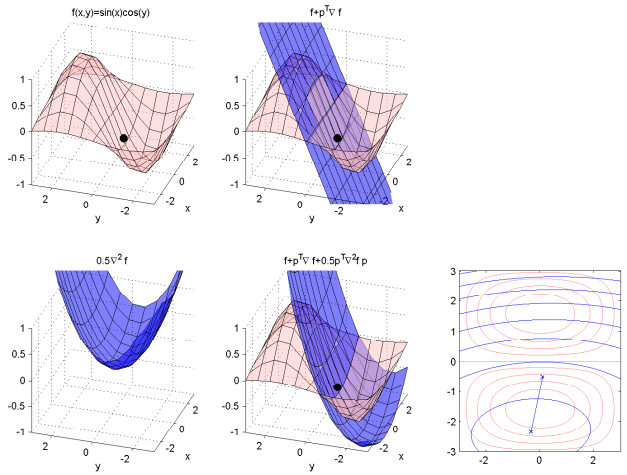
- The approximation of the non-linear function $\nabla f(x)$ with the linear (in p) polynomial

$$\nabla f(x_k + p) \approx \nabla f(x_k) + \nabla^2 f(x_k) p$$

  corresponds to approximating the non-linear function $f(x)$ with the quadratic (in p) Taylor expansion

$$m_k(x_k + p) \equiv f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p,$$

  i.e. $B_k = \nabla^2 f(x_k)$.
- Newton's method can be interpreted as that at each iteration k, f is approximated by the quadratic Taylor expansion m_k around x_k and x_{k+1} is calculated as the minimizer of m_k.



Properties of the Newton method

Advantages:

- It converges quadratically toward a stationary point.

Disadvantages:

- It does not necessarily converge toward a minimizer.
- It may diverge if the starting approximation is too far from the solution.
- It will fail if $\nabla^2 f(x_k)$ is not invertible for some k.
- It requires second-order information $\nabla^2 f(x_k)$.

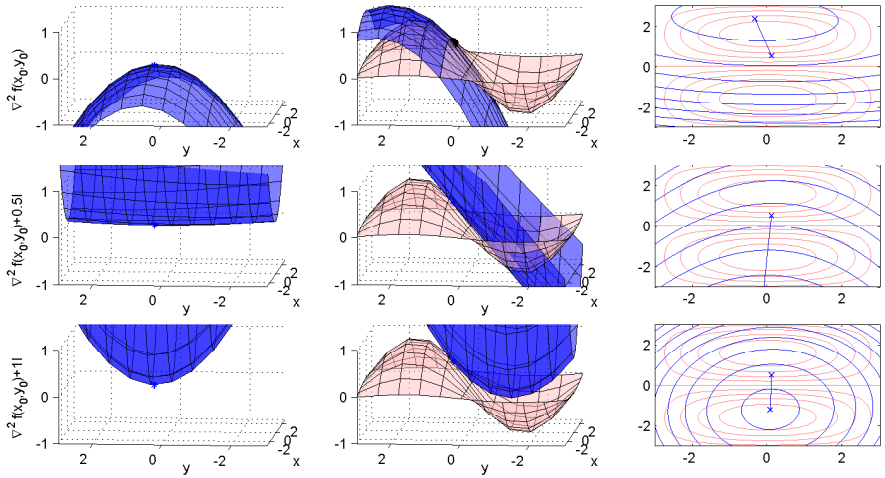
Newton's method is rarely used in its classical formulation. However, many methods may be seen as approximations of Newton's method.

Ensuring a descent direction

- Since the Newton search direction p^N is written as

$$p^N = -B_k^{-1}\nabla f_k,$$

  with $B_k = \nabla^2 f_k$, p^N will be a descent direction if $\nabla^2 f_k$ is positive definite.
- If $\nabla^2 f_k$ is *not* positive definite, the Newton direction p^N may not a descent direction.
- In that case we will choose B_k as a positive definite approximation of $\nabla^2 f_k$.
- Performed in a proper way, this modified algorithm will converge toward a minimizer. Furthermore, close to the solution the Hessian is usually positive definite and the modification will only be performed "far" from the solution.



- The positive definite approximation B_k of the Hessian may be found with minimal extra effort: The search direction p is calculated as the solution of

$$\nabla^2 f(x)p = -\nabla f(x).$$

- If $\nabla^2 f(x)$ is positive definite, the matrix factorization

$$\nabla^2 f(x) = LDL^T$$

may be used, where the diagonal elements of D are positive.
- If $\nabla^2 f(x)$ is not positive definite, at some point during the factorization, a diagonal element will be $d_{ii} \leq 0$. In this case, the element may be replaced with a suitable positive entry.
- Finally, the factorization is used to calculate the search direction

$$(LDL^T)p = -\nabla f(x).$$

The modified Newton algorithm with line search

Specify a starting approximation x_0 and a convergence tolerance ε.
Repeat for $k = 0, 1, \ldots$

- If $\|\nabla f(x_k)\| < \varepsilon$, stop.
- Compute the modified LDL^T factorization of the Hessian.
- Solve

$$(LDL^T)p_k^N = -\nabla f(x_k)$$

for the search direction p_k^N.
- Perform a line search to determine the new approximation $x_{k+1} = x_k + \alpha_k p_k^N$.