Unconstrained Optimization

Taylor's Theorem

We can use this theorem to approximate functions when f is of class C^2, **(page 14)**

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T p + \frac{1}{2} p^T \nabla^2 f(\mathbf{x}) p + O(||p||^3)$$

Q/Quotient Convergence

A sequence $\{x_k\}$ which converges to x^*, is said to converge **(page 619)**

- Linearly with rate of convergence c. if $r = 1$ and $c \in (0, 1)$.
- Superlinearly if $r = 1$ and $c = 0$.
- Quadratically if $r = 2$ and c is finite.

$$c = \lim_{k\to\infty} \frac{||x_{k+1} - x^*||}{||x_k - x^*||^r}$$

Types of minima

A point x^* is is a **global** minima of f if,

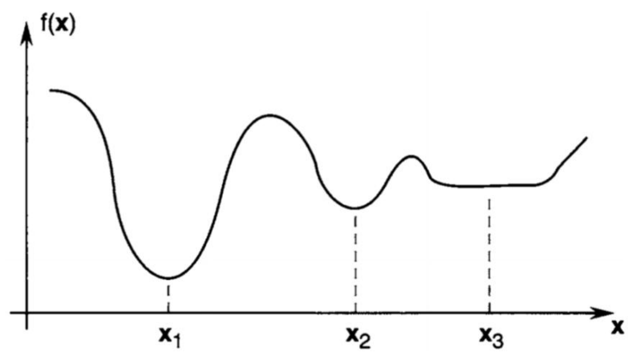
$$f(x^*) \leq f(x)\ \forall\ x \in \mathbb{R}^n$$

A point x^* is a **strict/strong local** minima of f if $\exists\ \mathbb{N}$ such that,

$$f(x^*) < f(x)\ \forall\ x \in \mathbb{N}$$

A point x^* is a **weak local** minima of f if $\exists\ \mathbb{N}$ such that,

$$f(x^*) \leq f(x)\ \forall\ x \in \mathbb{N}$$



x_1: strict global minimizer; x_2: strict local minimizer; x_3: local minimizer

In both cases, $\mathbb{N}$ is a neighbourhood of x^*, that is any open set containing x^*. **(page 12)**

Optimality conditions

If f is continuously differentiable in an open neighbourhood of x^* and if x^* is also a weak local minima then the **first order necessary** condition is just $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

The **second order necessary** conditions are $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(x^*)$ is positive **semi** definite.

The **second order sufficient** conditions for x^* to be a strong local minima is $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(x^*)$ is positive definite. **(page 15)**

Line search and Trust region Overview

These are the two broad strategies used to move from x_k to the next iterate x_{k+1},

1. Line Search : Choose a direction p_k, solve $min_{\alpha>0}\ f(x_k + \alpha p_k)$ to get the step length α.
   Now $x_{k+1} = x_k + \alpha p_k$
2. Trust Region : Choose a trust radius Δ_k. Approximate f around x_k as a quadratic function, $m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p$. Now we find $x_{k+1} = x_k + p$ by solving $min_p\ m_k(x_k + p)$ for p where $|x_{k+1} - x_k| < \Delta_k$. If this new iterate is not satisfactory we shrink Δ_k and repeat the steps.

∇f_k is the gradient of f evaluated at x_k. $\nabla^2 f_k$ is the hessian of f evaluated at x_k. B_k is an approximation of $\nabla^2 f_k$. **(page 19)**

Line search directions

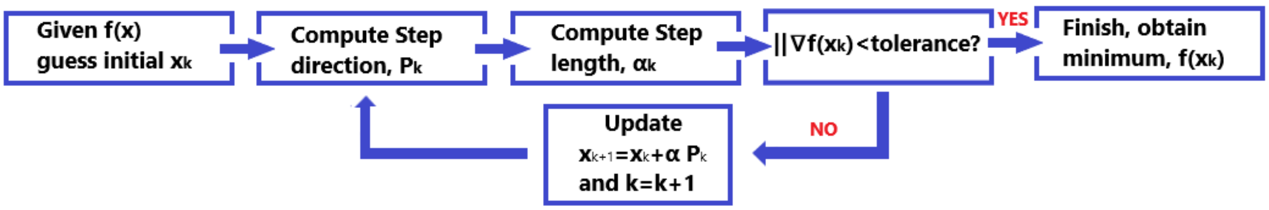
1. Steepest Direction : $p_k = \frac{-\nabla f_k}{||\nabla f_k||}$, we get the most rapid decrease in this direction from $f(x_k)$.
2. Newton Direction : $p_k = -(\nabla^2 f_k)^{-1} \nabla f_k$, the exact solution to the quadratic approximation of f. It is a descent direction only if the hessian is positive definite.
3. Quasi Newton Direction : $p_k = B_k^{-1} \nabla f_k$, where we avoid the costly hessian computation by using either Symmetric-rank-one (SR1) or BFGS iterative approximation. We impose symmetry and secant equation conditions, $B_k^T = B_k$ and $B_{k+1}(x_{k+1} - x_k) = \nabla f_{k+1} - \nabla f_k$.
4. Non Linear Conjugate Gradient directions : $p_k = -\nabla f_k + \beta_k p_{k-1}$, here β_k is a scalar that ensures p_k and p_{k-1} is conjugate. **(page 21)**

Wolfe Conditions

The armijo condition (to ensure sufficient decrease) and the curvature condition (to ensure the steps are not too small) are together known as the Wolfe conditions. **(page 33)**

1. Armijo condition : $f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k$: this enforces sufficient decrease in f_k every iteration. $c_1 \in (0, 1)$, usually 10^{-4}
2. Curvature condition : $\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k$: here $c_2 \in (c_1, 1)$ (usually 0.9) and this ensures the new slope $\phi'(\alpha)$ is less negative than $c_2 \times \phi'(0)$. Where $\phi(\alpha) = f(x_k + \alpha p_k)$

Algorithm flow chart of line search methods



Convergence of line search methods

To analyze convergence, we use the angle between our descent direction and the steepest descent direction. **(Page-37)**

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{||\nabla f_k|| ||p_k||}$$

Zoutendijk Theorem:

(Theorem-3.2)

Consider any iteration of the form $x_{k+1} = x_k + \alpha_k p_k$,
• where p_k is a descent direction
• α_k satisfies the Wolfe conditions
• f is bounded below in $\mathbb{R}^n$
• f is continuously differentiable in an open set N containing the level set $L = \{x : f(x) \leq f(x_0)\}$,
• Assume also that the gradient ∇f is Lipschitz continuous on N.
Then,

$$\sum_{k\geq 0} \cos^2\theta_k ||\nabla f_k||^2 < \infty \tag{1}$$

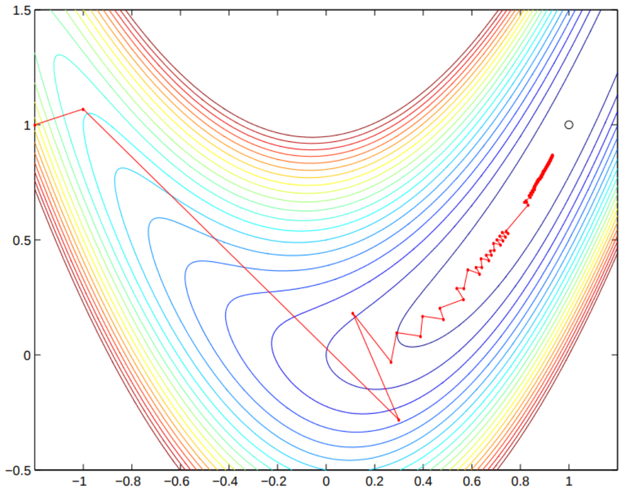
(1) is called **Zoutendijk condition** which implies that

$$\cos^2\theta_k ||\nabla f_k||^2 \to 0 \tag{2}$$

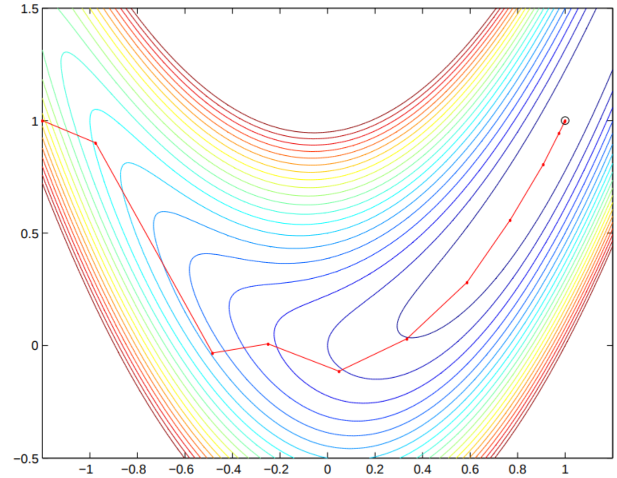
when $-90^o < \theta_k < 90^o$ for all k. It follows immediately from (2) that

$$\lim_{k\to\infty} ||\nabla f|| = 0 \tag{3}$$

The algorithms which satisfy (3) are called **globally convergent**. Zoutendijk condition shows that the steepest descent method is globally convergent. For other algorithms it describes how far p_k can deviate from the steepest descent direction and still give rise to a globally convergent iteration.



(a) iterates generated by the generic line search **steepest-descent method**



(b) iterates generated by the Generic Line search **Newton method**

Contours for the objective function $f(x, y) = 10(y - x^2)^2 + (x - 1)^2$ **(Rosenbrock function)**

Rate of Convergence

Convergence of Steepest Descent

• $p_k = \frac{-\nabla f_k}{||\nabla f_k||}$
• Globally convergent (converges to a local minimiser from any starting point x_0).
• many other methods resort to steepest descent in bad cases
• not scale invariant (changing the inner product on IR^n changes the notion of gradient!).
• convergence is usually very (very!) slow (linear)
• numerically often not convergent at all

Convergence of Newton's method

• $p_k = -\nabla^2 f_k^{-1} \nabla f_k$
• convergence is often faster than steepest descent
• may be viewed as "scaled" steepest descent
• If Hessian matrix $\nabla^2 f_k^{-1}$ is not Positive definite then p_k is not a descent direction. Two ways for obtaining globally convergent iteration:

- **line search approach**, in which the Hessian is modified, to make it positive definite.
- **trust region approach**, in which Hessian is used to form a quadratic model that is minimized in a ball.

Step length Selection algorithm

The line search is done in two stages: A **bracketing phase** finds an interval containing desirable step lengths, and a **bisection or interpolation phase** computes a good step length within this interval.

$$\phi(\alpha) = f(x_k + \alpha p_k)$$

If f is a convex quadratic, $f(x) = \frac{1}{2}x^T Qx - b^T x$, its one-dimensional minimizer along the ray $x_k + \alpha p_k$ can be computed analytically and is given by

$$\alpha_k = \frac{-\nabla f_k^T p_k}{p_k^T Q p_k}$$

For general nonlinear functions, it is necessary to use an iterative procedure. The line search procedure deserves particular attention because it has a major impact on the robustness and efficiency of all nonlinear optimization methods.