Optimization

Màster de Fonaments de Ciència de Dades

Gerard Gómez

Lecture II. Gradient methods for unconstrained optimization

Preliminaries on optimization methods

- It should be stressed that one hardly can hope to design a single optimization method capable to solve efficiently all nonlinear optimization problems – these problems are too diverse. In fact there are numerous methods, and each of them is oriented onto certain restricted family of optimization problems.
- Methods for numerical solving nonlinear optimization problems are, in their essence, iterative routines: a method typically is unable to find exact solution in finite number of computations.
- What a method generates, is an infinite sequence $\{x_n\}$ of approximate solutions. The next iterate $\{x_{n+1}\}$ is formed, according to certain rules, on the basis of local information of the problem collected along the previous iterates.
- Optimization methods can be classified according to the type of local information they use. From this viewpoint, the methods are divided into
  - Zero-order routines using only values of the objective and the constraints and not using their derivatives;
  - First-order routines using the values and the gradients of the objective and the constraints;
  - Second-order routines using the values, the gradients and the Hessians (i.e., matrices of second-order derivatives) of the objective and the constraints.

One-dimensional unconstrained optimization

Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function with a local extremum at x^*. As we have already seen $f'(x^*) = 0$, so the local extrema are the solutions of

$$f'(x) = 0$$

Newton's method

The idea behind Newton's method is to use a guess x^k for the solution of $f'(x) = 0$, linearize f' around x^k, and solve for the point where the linear function vanishes. This point is the next guess x^{k+1}.
Accordig to this, and writting

$$f'(x) = f'(x^k) + f''(x^k)(x - x^k) + O_2,$$

we get Newton's method: given $x^0 \in \mathbb{R}$ compute

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}, \quad k = 0, 1, 2, \ldots$$

The question is to know under which conditions the resulting sequence $\{x^k\}$ formula converges to the solution x^* of our problem.

Newton's method

Lemma

Let $\phi : [a, b] \to T \subset \mathbb{R}$ with $T \subset [a, b]$ be a continuous real-valued function, and $q \in \mathbb{R}$, $q < 1$ be such that:

$$\forall x^1, x^2 \in [a, b] \quad \textit{then} \quad |\phi(x^1) - \phi(x^2)| \leq q|x^1 - x^2| \quad \textit{(contracting condition)}$$

Then, if $x^0 \in [a, b]$ and $x^{k+1} = \phi(x^k)$ it follows that:

1. *There exists a unique fixed point x^* of ϕ.*
2. *For any $k \geq 0$*
$$|x^{k+1} - x^*| \leq q^{k+1}|x^0 - x^*|.$$
3. *For any $x^0 \in [a, b]$ it follows that $\{x^k\} \to x^*$.*

Proof:

1. Since $\phi(a), \phi(b) \in [a, b]$, the function $F(x) = \phi(x) - x$ satisfies $F(a) = \phi(a) - a \geq 0$ and $F(b) = \phi(b) - b \leq 0$. Since F is continuous, there is a point x^* such that $F(x^*) = 0$, this is $\phi(x^*) = x^*$.

   To see that x^* is unique, assume that there are two distinct fixed points $x_1^* \neq x_2^*$: $\phi(x_i^*) = x_i^*$ for $i = 1, 2$, then
$$0 < |x_1^* - x_2^*| = |\phi(x_1^*) - \phi(x_2^*)| \leq q|x_1^* - x_2^*|,$$
   which is a contradiction, since $q < 1$.

Newton's method

Proof: (cont.)

2. The inequality holds for $k = 0$ since

$$|x^1 - x^*| = |\phi(x^0) - \phi(x^*)| \leq q|x^0 - x^*|.$$

   Suppose that it holds up to a certain k:

$$|x^k - x^*| \leq q^k |x^0 - x^*|.$$

   Then

$$|x^{k+1} - x^*| = |\phi(x^k) - \phi(x^*)| \leq q|x^k - x^*| \leq q^{k+1}|x^0 - x^*|.$$

3. The convergence follows from the inequality, since $q < 1$, so $q^k \to 0$.

$\square$

The next lemma deals with sufficient conditions on ϕ for being a contraction.

Lemma

Suppose that $\phi : [a, b] \to T \subset \mathbb{R}$ with $T \subset [a, b]$ has a continuous derivative on $[a, b]$, $\phi \in C^1$. If $|\phi'(x)| < 1$ for every $x \in [a, b]$ then ϕ is a contraction.

Proof: Let $x^1, x^2 \in [a, b]$. Then, by the Mean Value Theorem

$$\phi(x^1) = \phi(x^2) + \phi'(\tilde{x})(x^1 - x^2), \quad \tilde{x} \in < x^1, x^2 >$$

where $< x^1, x^2 > \equiv [\min(x^1, x^2), \max(x^1, x^2)]$.
Hence

$$|\phi(x^1) - \phi(x^2)| = |\phi'(\tilde{x})|\, |x^1 - x^2|.$$

Taking

$$q = \max_{a \le x \le b} |\phi'(x)| < 1$$

the Lemma is proved. $\square$

Newton's method

Recall Newton's method:

$$x^0 \in \mathbb{R}, \quad x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}, \quad k = 0, 1, 2, \dots$$

Theorem

Let h, γ be two real valued continuously differentiable functions on $S = [a, b] \subset \mathbb{R}$. Suppose that

1. $h(a) \cdot h(b) < 0$,
2. *For all $x \in S$ the following conditions are satisfied:*
   - $h'(x) > 0$
   - $\gamma(x) > 0$,
   - $0 \leq 1 - [\gamma(x)h(x)]' \leq q < 1$

Let

$$x^{k+1} = x^k - \gamma(x^k)h(x^k), \quad k \geq 0$$

with $x^0 \in S$, then the sequence $\{x^k\}$ converges to a solution x^* of $h(x) = 0$.

Newton's method

Proof: Define $\phi(x) = x - \gamma(x)h(x)$, so $\phi'(x) = 1 - [\gamma(x)h(x)]'$. We have that

$$0 \leq \phi'(x) \leq q < 1, \quad \forall x \in S,$$

so ϕ is monotone nondecreasing on S. The function h is monotone increasing on S and satisties $h(a) < 0$, $h(b) > 0$, hence $\phi(a) > a$ and $\phi(b) < b$, so it follows that

$$a < \phi(x) < b, \quad \forall x \in S.$$

Moreover $|\phi'(x)| < 1$ and, by the preceeding Lemma, it follows that ϕ is a contractor on S, so it has a unique fixed point $\overline{x} \in S$ and the sequence

$$x^{k+1} = x^k - \gamma(x^k)h(x^k) = \phi(x^k)$$

converges to $\overline{x}$. Finally, since $\gamma(x) > 0$, observe that x^* is a fixed point of ϕ if and only if $h(x^*) = 0$, thus $\{x^k\}$ converges to a solution of $h(x) = 0$. $\square$

Newton's method

Now we can state sufficient conditions for the convergence of Newton's method.

Corollary

Let $h(x) = f'(x)$, $\gamma(x) = 1/f''(x)$ *with* $f \in C^2$ *in* $S = [a, b]$. *Assume that* h *and* γ *fulfil the hypotheses of the preceeding Theorem*

- $h(a) \cdot h(b) < 0$
- $h'(x) > 0$
- $\gamma(x) > 0,$
- $0 \leq 1 - [\gamma(x)h(x)]' \leq q < 1$

then

$$x^{k+1} = x^k - \gamma(x^k)h(x^k) = x^k - \frac{f'(x^k)}{f''(x^k)} \longrightarrow x^*,$$

with $f'(x^*) = 0$.

Rates of convergence

- Assume that a method, as applied to a minimization problem P, generates sequence of iterates converging to the solution set X^* of the problem.
- To define the rate of convergence, we first introduce an error function $err(x)$ which measures the quality of an approximate solution x; this function should be positive outside X^* and should be zero at the latter set
- There are several choices of the error function. We can use, for instance, the distance from the approximate solution x to the solution set

  $$err(x) = inf_{x^* \in X^*} \|x - x^*\|$$

  Another choice of the error function could be the residual in terms of the objective function and the constraints

  $$err(x) = max\{f(x) - f^* \; ; \; |h_1(x)| \; ; \; ... \; ; \; |h_m(x)|\}$$

  f^* being the optimal value in P
- For properly chosen error function, convergence of the iterates to the solution set implies that the sequence

  $$r_n = err(x_n) \to 0$$

Rate of convergence

- In addition to proving convergence of a certain algorithm, it is also important to know the rate of convergence.
- We measure the quality of convergence by the rate at which $\{r_n\}$ tends to zero
- Let $\{x^k\}$, with $x^k \in \mathbb{R}^n$ be a sequence that converges to x^* with $x^k \neq x^*$ for all sufficiently large k. If there exists numbers p and $\alpha \neq 0$ such that

$$\lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^p} = \alpha,$$

then it is said that the order of convergence of $\{x^k\}$ to x^* is p, and $\|x^k - x^*\|$ is the error of the kth approximant. If $p = 1$ the rate of convergence is said to be linear, if $p = 2$ quadratic and, in general, if $p > 1$ superlinear.

Newton's method convergence

Theorem

Assume that the hypotheses of the last Theorem and Corollary hold, and that the sequence $\{x^k\}$, $x^k \in \mathbb{R}$, generated by Newton's method converges to a point x^* that satisfies $h(x^*) = 0$. Then the rate of convergence of $\{x^k\}$ towards x^* is quadratic.

Proof: The point x^* solves $h(x) = 0$ if and only if is a fixed point of

$$\phi(x) = x - \frac{h(x)}{h'(x)}.$$

By the Mean Value Theorem

$$x^{k+1} - x^* = \phi(x^k) - \phi(x^*) = \phi'(\xi^k)(x^k - x^*), \quad \xi^k \in < x^k, x^* > .$$

If we take into account that

$$\phi'(x) = 1 - \frac{(h'(x))^2 - h(x)h''(x)}{(h'(x))^2} = \frac{h(x)h''(x)}{(h'(x))^2},$$

it follows

$$|x^{k+1} - x^*| = \frac{|h(\xi^k)h''(\xi^k)|}{(h'(\xi^k))^2}|x^k - x^*|.$$

Newton's method convergence (cont.)

Since

$$|h(\xi^k)| = |h(\xi^k) - h(x^*)| = |h'(\eta^k)|\ |\xi^k - x^*| \leq |h'(\eta^k)|\ |x^k - x^*|,$$

with $\eta^k \in < \xi^k, x^* >$, hence

$$|x^{k+1} - x^*| \leq \frac{|h''(\xi^k)h'(\eta^k)|}{(h'(\xi^k))^2}|x^k - x^*|^2.$$

Taking

$$\beta = \sup_x \frac{|h''(x)h'(x)|}{(h'(x))^2},$$

we get

$$|x^{k+1} - x^*| \leq \beta|x^k - x^*|^2.$$

$\square$

The secant method

A closely related root-finding method can be obtained by approximating the second derivative $f''(x)$ by

$$f''(x^k) \simeq \frac{f'(x^k) - f'(x^{k-1})}{x^k - x^{k-1}},$$

in Newton's method formula. In this way we get secant method:

$$x^{k+1} = x^k - \frac{f'(x^k)(x^k - x^{k-1})}{f'(x^k) - f'(x^{k-1})}$$

If $f''' \neq 0$ then

$$\lim_{k\to\infty} \frac{|x^{k+1} - x^*|}{|x^{k+1} - x^*|^{\tau}} = \left| \frac{2f''(x^*)}{f'''(x^*)} \right|^{1/\tau},$$

where $\tau = (1 + \sqrt{5})/2 = 1.618...$ is a solution of the equation $t^2 - t - 1 = 0$. Thus, for large values of k, the secant method is superlinear.

One-dimensional unconstrained optimization. Line search methods

Consider numerical methods to solve the problem

$$\min_x \{f(x) \ : \ x \in \mathbb{R}\}$$

f being, at least, a continuous function.

These methods usually are called line search. Line search is a component of basically all traditional methods for multidimensional optimization.

Zero-order line search They solve the problem

$$\min_x \{f(x) \ : \ a \leq x \leq b\}, \quad -\infty < a < b < \infty$$

using only the values of f and not the derivatives.

Polynomial approximation methods

The quadratic method

Let f be the function whose minimum is sought. The basis of the quadratic method is to approximate f by

$$\phi(x) = a + bx + cx^2$$

- Suppose that we evaluate f at three points $x_1 < x_2 < x_3$.
- Letting $f(x_i) = \phi(x_i)$, $i = 1, 2, 3$ we can solve for the coefficients a, b, c.
- The minimum of the quadratic function ϕ (if it has a minimum) can be found analytically by setting $\phi'(x) = 0$, and, for a first approximation of a minimum of f we obtain
$$\tilde{x} = -\frac{b}{2c}.$$
- Assume that $c > 0$. If $c < 0$, the quadratic function is actually a parabola with a maximum and so the point $\tilde{x}$ obtained is unusable. A situation that will ensure that c is positive is
$$f(x_1) > f(x_2), \quad \text{and} \quad f(x_3) > f(x_2)$$
- If these conditions hold we can also ensure that the local minimum of f is between x_1 i x_3.

The quadratic method

- Under the above conditions, the minimum of ϕ so found will also satisfy

$$f(x_1) > \phi(\tilde{x}) \quad \text{and} \quad f(x_3) > \phi(\tilde{x})$$

- Now consider the four points $(x_1, f(x_1)$, $(x_2, f(x_2)$, $(x_3, f(x_3)$, $(\tilde{x}, f(\tilde{x}))$.
- Choose as the new x_2 one of the four points at which f has been computed and which yielded the lowest value of f and let the new x_1 and x_3 be the two points adjacent to the new x_2 from the left and right, respectively, and repeat the iteration.
- This algorithm can be terminated if either

$$|f(\tilde{x}) - \phi(\tilde{x})| < \epsilon$$

  for some tolerance $\epsilon > 0$, or if estimates of the minimum point in two or more succesive iterations are closer than some predetermined distance.
- If $\tilde{x} = x_2$ the algorithm will not evaluate new points, although x_2 may not be a local minimum of f. In such a degenerate case, some perturbations on $\tilde{x}$ are needed in order to proceed with the computations.

Exercises

Exercise 3. Show that the inequalities $f(x_1) > f(x_2)$ and $f(x_3) > f(x_2)$ imply that the coefficient c of the quadratic approximation $\phi(x) = a + bx + cx^2$ is positive and that the predicted stationary point of ϕ is indeed a minimum.

Exercise 4. Let f be a real function on $\mathbb{R}^n$. Also let $x_0 \in \mathbb{R}^n$, $z \in \mathbb{R}^n$,and $\theta \in \mathbb{R}$. Define

$$F(\theta) = f(x_0 + \theta z)$$

and suppose that we are looking for the minimum of F (that is, for the minimum of f in the direction z through the point x_0). Let $x_0 + \theta_1 z$, $x_0 + \theta_2 z$ and $x_0 + \theta_3 z$ be three points where f is evaluated. Show that the minimum predicetd by applying the quadratic approximation method is $x_0 + \theta^* z$, where

$$\theta^* = \frac{[\theta_2^2 - \theta_3^2]F(\theta_1) + [\theta_3^2 - \theta_1^2]F(\theta_2) + [\theta_1^2 - \theta_2^2]F(\theta_3)}{2[(\theta_2 - \theta_3)F(\theta_1) + (\theta_3 - \theta_1)F(\theta_2) + (\theta_1 - \theta_2)F(\theta_3)]}$$

and it is indeed the minimum of the parabola passing through the above three points if

$$\frac{(\theta_2 - \theta_3)F(\theta_1) + (\theta_3 - \theta_1)F(\theta_2) + (\theta_1 - \theta_2)F(\theta_3)}{(\theta_2 - \theta_3)(\theta_3 - \theta_1)(\theta_1 - \theta_2)} < 0$$

The cubic method

The function f is approximated by

$$\phi(x) = a + bx + cx^2 + dx^3$$

We will assume that the first derivatives of f can be evaluated.
We start at an arbitrary point x_1 and compute $f(x_1)$ and $f'(x_1)$. Assume that $f'(x_1) < 0$. Then we compute $x_2 > x_1$ such that

$$f'(x_2) \geq 0, \quad \text{or} \quad f(x_2) > f(x_1).$$

The coefficients a, b, c and d can be now computed solving the system

$$\begin{aligned} f(x_1) &= a + bx_1 + cx_1^2 + dx_1^3, \\ f'(x_1) &= b + 2cx_1 + 3dx_1^2, \\ f(x_2) &= a + bx_2 + cx_2^2 + dx_2^3, \\ f'(x_2) &= b + 2cx_2 + 3dx_2^2. \end{aligned}$$

The solution of these equations can be found by a simple change of variables. Define

$$z = x - x_1,$$

and, instead of f and ϕ, use the functions

$$g(z) = f(x_1 + z), \quad \psi(z) = \phi(x_1 + z)$$

The cubic method

It can be easily seen that

$$\psi'(z) = g'(0) - \frac{2z}{\lambda}(g'(0) + \alpha) + \frac{z^2}{\lambda^2}(g'(0) + g'(\lambda) + 2\alpha),$$

where $\lambda = x_2 - x_1$ and

$$\alpha = \frac{3(g(0) - g(\lambda))}{\lambda} + g'(0) + g'(\lambda).$$

The point that satisfies $\psi'(z) = 0$ is

$$\tilde{z} = \lambda(1 - \beta),$$

where

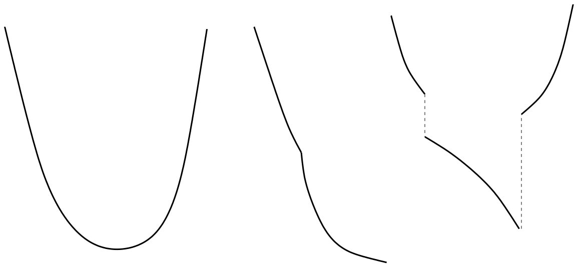
$$\beta = \frac{g'(\lambda) + (\alpha^2 + g'(0)g'(\lambda))^{1/2} - \alpha}{g'(\lambda) - g'(0) + 2(\alpha^2 + g'(0)g'(\lambda))^{1/2}}.$$

If $|g'(\tilde{z})| < \epsilon$ the procedure is terminated; otherwise the algorithm must be restarted by a procedure similar tothe one of the quadratic method

Unimodal functions

Let $L \subset \mathbb{R}$ be a closed interval. A real-valued function f is said to be unimodal on L if there exist $x^* \in L$ such that x^* minimizes f on L and for any two points $x_1, x_2 \in L$, such that $x_1 < x_2$ we have

$$\begin{aligned} x_2 \leq x^* \quad &\Rightarrow \quad f(x_1) > f(x_2), \\ x^* \leq x_1 \quad &\Rightarrow \quad f(x_2) > f(x_1). \end{aligned}$$



In another words, f is unimodal on $[a, b]$ if it possesses a unique local minimum x^* on $[a, b]$, which implies that that f is strictly decreasing in $[a, b]$ to the left of x^* and strictly increasing in $[a, b]$ to the right of x^*.

Let f be an unimodal function.

The startegy of the zero-order line search method is based in the following. Choose somehow two points x_1 and x_2 such that $a < x_1 < x_2 < b$ and compute the values of f at these points. The basic observation is that:

- Case A: *if* $f(x_1) \leq f(x_2)$*, then* x^* *is to the left of* x_2
- Case B: *if* $f(x_1) \geq f(x_2)$*, then* x^* *is to the right of* x_1

Algorithm
Let $L = \{x \mid l_1 \leq x \leq r_1\} = [l_1, r_1]$ and $x_1, x_2 \in L$ two points such that $x_1 < x_2$. We evaluate the unimodal function f at both points: $f(x_1)$ and $f(x_2)$.

- If $f(x_1) < f(x_2)$. Since f is unimodal, it follows that either $x^* \leq x_1 < x_2$ or $x_1 \leq x^* \leq x_2$. In both cases $x^* \in [l_1, x_2]$.
- If $f(x_1) > f(x_2)$. Since f is unimodal, it follows that $x^* \in [x_1, r_1]$.
- If $f(x_1) = f(x_2)$. Since f is unimodal, it follows that $x^* \in [x_1, x_2]$.

The line search method

- In all the cases, after the first two function evaluations, a portion of L to the right of x_2 or the left of x_1 can be eliminated from further search. So we have found a new interval $[l_2, r_2]$ such that $x^* \in [l_2, r_2]$. Then we repeat the procedure iteratively.
- It is immediately seen that we may ensure linear convergence of the lengths of subsequent uncertainty segments to 0. If x_1, x_2 are chosen to split $[l_n, r_n]$ into three equal parts, we ensure $|r_{n+1} - l_{n+1}| = (2/3)|r_n - l_n|$so

$$|x_n - x^*| \leq \left(\frac{2}{3}\right)^n |b - a|$$

One-dimensional unconstrained optimization. The Fibonacci method

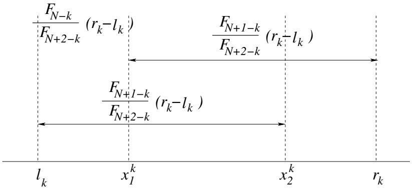
- The Fibonacci numbers, F_k are defined by

$$F_0 = 0, \quad F_1 = 1, \quad F_k = F_{k-1} + F_{k-2}, \;\; k = 2, 3, \ldots$$

  The first Fibonacci numbers are: 0, 1, 1, 2, 3, 5, 8, 13, 21, 34,...

- Let N the total number of points at which the unimodal function f will be evaluated. For N function evaluations, we will do $N-1$ interval reductions (iterations)
- At iteration number k the interval containing x^* is $[l_k, r_k]$.
- For $k = 1, 2, \ldots, N-1$ the function values are compared at the two points

$$x_1^k = l_k + \frac{F_{N-k}}{F_{N+2-k}}(r_k - l_k), \quad x_2^k = l_k + \frac{F_{N+1-k}}{F_{N+2-k}}(r_k - l_k). \tag{1}$$



The Fibonacci method

Note that, except for $k = 1$ the function f has already been evaluated in a previous iteration at one of the two points.
Note also that the points x_1^k and x_2^k are placed symmetrically in the interval $[l_k, r_k]$, since

$$\begin{aligned} x_2^k - l_k &= \frac{F_{N+1-k}}{F_{N+2-k}}(r_k - l_k) = \frac{F_{N+2-k} - F_{N-k}}{F_{N+2-k}}(r_k - l_k) \\ &= r_k - l_k - \frac{F_{N-k}}{F_{N+2-k}}(r_k - l_k) = r_k - x_1^k. \end{aligned}$$

At the last iteration ($k = N - 1$) formulas (1) give

$$x_1^{N-1} = x_2^{N-2} = l_{N-1} + \frac{1}{2}(r_{N-1} - l_{N-1}),$$

and no further interval reduction is possible.
After N function evaluations, the length of the interval containing x^* is

$$r_N - l_N = \frac{r_1 - l_1}{F_{N+1}}$$

In this way, we can bracket the minimum of any unimodal function within 1% of the starting interval by 11 function evaluations ($F_{12} = 144$), and within 0.1% by 16 evaluations

The Fibonacci method. Example

Consider the function $f(x) = (x-3)^2$. Set $N = 4$, $L = [0, 10]$, then

$$r_4 - l_4 = \frac{10-0}{5} = 2$$

According to (1)

$$x_1^1 = 0+\frac{F_3}{F_5}(10-4) = 0+\frac{2}{5}(10-4) = 4, \quad x_2^1 = 0+\frac{F_4}{F_5}(10-4) = 0+\frac{3}{5}(10-4) = 6.$$

Computing $f(x_1^1) = 1$, $f(x_2^1) = 9$, we get $l_2 = 0$ i $r_2 = 6$. From these values it follows

$$x_1^2 = 0+\frac{F_2}{F_4}(6-4) = 0+\frac{1}{3}(6-4) = 2, \quad x_2^2 = 0+\frac{F_3}{F_4}(6-4) = 0+\frac{2}{3}(6-4) = 4.$$

Note that $x_2^2 = x_1^1$, and we can computer $l_3 = 2$ and $r_3 = 6$ together with

$$x_1^3 = 2+\frac{F_1}{F_3}(6-2) = 2+\frac{1}{2}(6-2) = 4, \quad x_2^3 = 2+\frac{F_2}{F_3}(6-2) = 2+\frac{1}{2}(6-2) = 4.$$

The final interval is $[2, 4]$.

Among all the search procedures with N function evaluations, the Fibonacci method minimizes the length of the maximum possible interval remaining after N function evaluations and containing the sought minimum.

One-dimensional unconstrained optimization. The golden search method

One of the disadvantages of the Fibonacci method is that the number of function evaluations N must be known in advance, prior to starting the search. This requirement is not necessary in a related technique, called the golden section method, which is a good approximation of the Fibonacci search.

It can be shown that

$$\lim_{n\to\infty} \frac{F_{N-1}}{F_N} = \frac{1}{\tau} = \frac{\sqrt{5}-1}{2} = 0.618...$$

In this way

$$x_2^k = l_k + \frac{F_{N+1-k}}{F_{N+2-k}}(r_k - l_k) \simeq l_k + \frac{1}{\tau}(r_k - l_k),$$

The golden section method then places the points at which the function is to be evaluated at

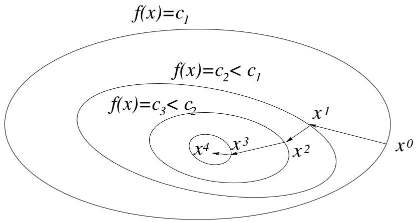
$$\begin{aligned} x_1^{kG} &= l_k + \frac{\tau - 1}{\tau}(r_k - l_k), \\ x_2^{kG} &= l_k + \frac{1}{\tau}(r_k - l_k). \end{aligned}$$

The golden section method reduces the initial interval containing the minimum by a factor $1/\tau^{N-1}$ in front of the factor of the Fibonacci method that is $1/F_{N+1}$. It can also be shown that

$$\lim_{n\to\infty} \frac{F_{N+1}}{\tau^{N-1}} = \frac{\tau^2}{\sqrt{5}} = 1.17...$$

n-dimensional unconstrained optimization. Descent methods

- We consider methods for unconstrained optimization, although the fundamental concepts apply also to constrained optimization.
- Most of the interesting algorithms for this problem rely on an important idea: the iterative descent.
- The iterative descent method
  - Let
    $$f : \mathbb{R}^n \longrightarrow \mathbb{R}$$
    be a continuosly differentiable function.
  - Let $x^0 \in \mathbb{R}^n$ be an initial guess
  - Generate a sequence of points x^1, x^2,... such that the value of f is decreased at each iteration, this is
    $$f(x^{k+1}) < f(x^k), \quad k = 0, 1, 2, ...$$



Iterative descent for minimizing f

The gradient

- Recall that the gardient of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is the vectorfield

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)^T$$

- If $s \in \mathbb{R}^n$ is a unitary vector, the directional derivative of $f : \mathbb{R}^n \to \mathbb{R}$ at a point $x \in \mathbb{R}^n$ in the direction of s, which measures the rate of change of the function along s is equal to

$$Df(x, s) = \lim_{\lambda \to 0} \frac{f(x + \lambda s) - f(x)}{\lambda} = (\nabla f(x))^T s \in \mathbb{R}$$

- Since the directional derivative is

$$(\nabla f(x))^T s = \|\nabla f(x)\| \|s\| \cos\theta = \|\nabla f(x)\| \cos\theta$$

  the maximum rate of change of f at the point x occurs when $\cos\theta$ is maximized, this is when $\theta = 0$. Thus, the greatest increase occurs in the direction of $\nabla f(x)$, and the greatest decrease occurs in the direction of $-\nabla f(x)$

Gradient methods. Basic principle

Givent $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$, consider the half line

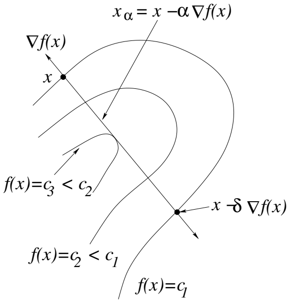
$$x_\alpha = x - \alpha \nabla f(x), \quad \alpha \geq 0.$$

According to Taylor's formula, and since $\nabla f(x)^T \nabla f(x) = \|\nabla f(x)\|^2$, we have

$$\begin{aligned} f(x_\alpha) &= f(x) + \nabla f(x)^T (x_\alpha - x) + o(\|x_\alpha - x\|) \\ &= f(x) - \alpha \|\nabla f(x)\|^2 + o(\alpha \|\nabla f(x)\|). \end{aligned}$$

Since $\nabla f(x) \neq 0$, then for α within a certain (small enough) positive interval $0 \leq \alpha \leq \delta$, we have

$$f(x_\alpha) < f(x)$$



Gradient methods. Basic principle

The above procedure can be generalised. Consider the half line

$$x_\alpha = x + \alpha d, \quad \alpha \geq 0,$$

where the direction $d \in \mathbb{R}^n$ makes an angle with $\nabla f(x)$ that is greater than 90^o, this is

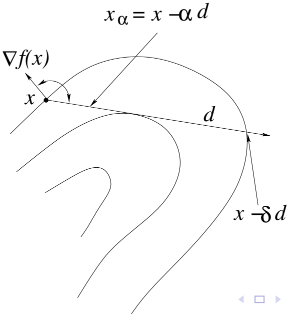
$$\nabla f(x)^T d < 0.$$

According to Taylor's formula

$$f(x_\alpha) = f(x) + \alpha \nabla f(x)^T d + o(\alpha),$$

For positive and small enough values of α $(0 \leq \alpha \leq \delta)$, we also have

$$f(x + \alpha d) < f(x)$$



General gradient methods

- The general expression of a gradient method is

$$x^{k+1} = x^k + \alpha^k d^k, \quad k = 0, 1, \ldots$$

  where, if $\nabla f(x^k) \neq 0$, the direction d^k is chosen so that

$$\nabla f(x^k)^T d^k < 0,$$

  and the stepsize is $\alpha^k > 0$. The name "gradient methods" is due to the relation between d^k and $\nabla f(x^k)$
- When $\nabla f(x^k) = 0$ the method stops.
- Most of the gradients methods that will be considered are also descent methods, this is, the step size α^k is such that

$$f(x^k + \alpha^k d^k) < f(x^k), \quad k = 0, 1, \ldots$$

Selecting the descent direction d^k

- There are many possibilities for choosing the direction d^k (and also the step size α^k)
- Consider gradient methods, $x^{k+1} = x^k + \alpha^k d^k$, with the following direction $d^k = -D^k \nabla f(x^k)$, this is, with the following general pattern

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

  where D^k is a positive definite symmetric matrix $(z^T D^k z > 0, \forall z \neq 0)$

- Since

$$d^k = -D^k \nabla f(x^k)$$

  the descent condition $\nabla f(x^k)^T d^k < 0$ becomes

$$-\nabla f(x^k)^T D^k \nabla f(x^k) < 0 \quad \Leftrightarrow \quad \nabla f(x^k)^T D^k \nabla f(x^k) > 0$$

  which holds, since D^k is positive definite

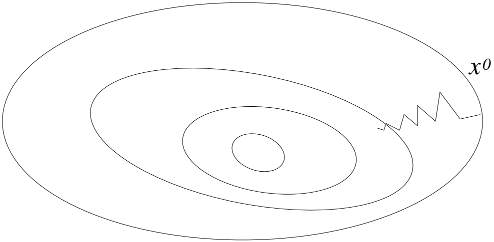
- Let us see some choices of the matrix D^k that define different methods

Some choices of the matrix D^k. The steepest descent method

The simplest choice for D^k is

$$D^k = I, \quad k = 0, 1, \ldots \quad \Rightarrow \quad x^{k+1} = x^k - \alpha^k \nabla f(x^k), \quad k = 0, 1, \ldots$$

where I is the identity matrix. In this case the method is known as the steepest descent method



This choice often leads to slow convergence

The steepest descent method

The name of the above method "steepest descent" is due to the following. Recall that if

$$x^{k+1} = x^k + \alpha d^k, \quad \alpha \geq 0,$$

then

$$f(x^{k+1}) = f(x^k) + \alpha \nabla f(x^k)^T d + o(\alpha),$$

so the rate of chage of f is $\nabla f(x^k)^T d$

Consider any unitary direction $d \in \mathbb{R}^n$, $(\|d\| = 1)$. According to Schwartz inequality[1], the rate of change of f verifies

$$\nabla f(x^k)^T d \leq \|\nabla f(x^k)\| \, \|d\| = \|\nabla f(x^k)\|$$

If we set

$$d = \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$$

then

$$\nabla f(x^k)^T d = \|\nabla f(x^k)\|$$

therefore, $-\nabla f(x^k)$ is the max-rate descending direction of f

---

[1]

$$|x^T y| \leq \|x\| \|y\|, \quad \text{and} \quad |x^T y| = \|x\| \|y\| \;\Leftrightarrow\; x = \alpha y$$

Some choices of the matrix D^k. Newton's method

- Take
$$D^k = (\nabla^2 f(x^k))^{-1}, \quad k = 0, 1, \dots$$
so
$$x^{k+1} = x^k - \alpha^k (\nabla^2 f(x^k))^{-1} \nabla f(x^k), \quad k = 0, 1, \dots$$
provided $\nabla^2 f(x^k)$ is positive definite (if not some modification must be done).
- The idea of Newton's method is to minimize, at each iteration, the quadratic approximation of f around the current point x^k. This approximation is given by
$$G(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k).$$
By setting the derivative of $G(x)$ (with respect to x) equal to zero, we get
$$G'(x) = \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0,$$
from which, isolating x and setting $x^{k+1} = x$, we have
$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$
This is the "pure" Newton iteration ($\alpha^k = 1$), the general procedure is the one already written.
- Usually the convergence of the method is fast and has not the zig-zagging behavior of the steepest descent method.

Some choices of the matrix D^k. Newton's method

- Newton's method determines the minimum of a quadratic positive definite function in ONE iteration. Let

$$f(x) = x^T Qx + b^T x + a$$

with Q positive definite. Note that $\nabla^2 f(x) = Q$ is constant.

Let x^0 be an arbitrary point in $\mathbb{R}^n$ and x^* the minimum of f. Then

$$\nabla f(x^0) = Qx^0 + b, \quad \text{and} \quad \nabla f(x^*) = 0 = Qx^* + b$$

From these two equations we get

$$x^0 = Q^{-1}\nabla f(x^0) - Q^{-1}b, \qquad x^* = -Q^{-1}b$$

and

$$x^* = x^0 - Q^{-1}\nabla f(x^0) = x^0 - (\nabla^2 f(x^0))^{-1}\nabla f(x^0)$$

which is the first iteration of Newton's method starting at x^0

Some choices of the matrix D^k. Newton's method

Example Consider the quadratic function

$$f(\vec{x}) = (x - y + z)^2 + (-x + y + z)^2 + (x + y - z)^2$$

that can be written as

$$f(\vec{x}) = \frac{1}{2} x^T Q x, \quad \text{with} \quad Q = \begin{pmatrix} 6 & -2 & -2 \\ -2 & 6 & -2 \\ -2 & -2 & 6 \end{pmatrix}$$

Let $x^0 = (1/2, 1, 1/2)^T$, then

$$\nabla f(x^0) = Q x^0 = (0, 4, 0)^T$$

and

$$x^* = x^0 - Q^{-1} \nabla f(x^0) = \begin{pmatrix} 1/2 \\ 1 \\ 1/2 \end{pmatrix} - \begin{pmatrix} 1/4 & 1/8 & 1/8 \\ 1/8 & 1/4 & 1/8 \\ 1/8 & 1/8 & 1/4 \end{pmatrix} \begin{pmatrix} 0 \\ 4 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

So, f has a local (and global) minimum at $(0, 0, 0)^T$

Some choices of the matrix D^k. Diagonally scaled steepest descent

- In the general gradient method

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

the diagonally scaled steepest descent method uses

$$D^k = \begin{pmatrix} d_1^k & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & d_2^k & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & d_{n-1}^k & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & d_n^k \end{pmatrix}, \quad k = 0, 1, \dots$$

where $d_i^k \in \mathbb{R}$ are all positive, thus ensuring that D^k is positive definite.

- A popular choice, resulting in a method known as a diagonal approximation to Newton's method is to take d_i^k to be an approximation of the inverted second partial derivative of f with respect to x_i, this is

$$d_i^k \approx \left(\frac{\partial^2 f(x^k)}{\partial x_i^2} \right)^{-1}.$$

Some choices of the matrix D^k. Modified and Discretized Newton's methods

- Modified Newton's method
  In the general gradient method
  $$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$$
  take
  $$D^k = \left(\nabla^2 f(x^0)\right)^{-1}, \quad k = 0, 1, \ldots$$
  provided $\nabla^2 f(x^0))$ is positive definte.
  This method is the same as Newton's method except that the Hessian matrix is not computed at each step. A related method recomputes the Hessian matrix every $p > 1$ steps (p not necessarily fixed)
- Discretized newton's method
  In the general gradient method
  $$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$$
  take
  $$D^k = \left(H(x^k)\right)^{-1}, \quad k = 0, 1, \ldots$$
  where $H(x^k)$ is a positive definite symmetric approximation of $\nabla^2 f(x^k)$ computed using finite difference approximations of the second derivatives of f (eventually using the values of f').

Some choices of the matrix D^k. Gauss-Newton method

This method is applicable to the problem of minimizing the sum of squares of real valued functions $g_1,\dots,g_m$. By denoting $g = (g_1, \dots, g_m)$ the problem can be written as

$$\begin{array}{ll} \text{minimize} & f(x) = \frac{1}{2}\|g(x)\|^2 = \frac{1}{2}\sum_{i=1}^{m} g_i^2(x), \\ \text{subject to} & x \in \mathbb{R} \end{array}$$

To solve this problem we use the linealization of $g(x)$ around x^k:

$$g(x) \approx g(x^k) + \nabla g(x^k)^T (x - x^k)$$

and compute the minimum of $\frac{1}{2}\|g(x)\|^2$ using this approximation, this is, the minimum of

$$\frac{1}{2}\left(g(x^k) + \nabla g(x^k)^T (x - x^k)\right)^T \left(g(x^k) + \nabla g(x^k)^T (x - x^k)\right) =$$

$$\frac{1}{2}\left(\|g(x^k)\|^2 + 2(x - x^k)^T \nabla g(x^k) g(x^k) + (x - x^k)^T \nabla g(x^k) \nabla g(x^k)^T (x - x^k)\right)$$

Equating to zero the derivative of this expression, we get

$$\nabla g(x^k)\ g(x^k) + \nabla g(x^k) \nabla g(x^k)^T (x - x^k) = 0$$

Gauss-Newton method (cont.)

If the matrix $\nabla g(x^k)\nabla g(x^k)^T$ is non-singular, then

$$\nabla g(x^k)\ g(x^k) + \nabla g(x^k)\nabla g(x^k)^T(x - x^k) = 0 \quad \Rightarrow$$

$$x^{k+1} = x^k - \left(\nabla g(x^k)\nabla g(x^k)^T\right)^{-1}\nabla g(x^k)\ g(x^k),$$

Note that, since $f(x) = (1/2)g(x)^T g(x)$, then

$$\nabla f(x^k) = \nabla g(x^k)\ g(x^k).$$

According to the general pattern of gradient methods, we can write Gauss-Newton method as

$$\begin{aligned} x^{k+1} &= x^k - \alpha^k\left(\nabla g(x^k)\nabla g(x^k)^T\right)^{-1}\nabla g(x^k)\ g(x^k) \\ &= x^k - \alpha^k\left(\nabla g(x^k)\nabla g(x^k)^T\right)^{-1}\nabla f(x^k), \end{aligned}$$

so

$$D^k = \left(\nabla g(x^k)\nabla g(x^k)^T\right)^{-1}, \quad k = 0, 1, \ldots$$

We have assumed that $\nabla g(x^k)\nabla g(x^k)^T$ is non-singular and it will be always positive semidefinite. It will be positive definite if the matrix $\nabla g(x^k)$ has rang n

Gauss-Newton method (cont.)

- **Advantage** of Gauss-Newton method over Newton's method: no second derivatives of g are needed
- **Disadvantage** of Gauss-Newton method over Newton's method: convergence is slower

Selecting the stepsize

There are a number of rules for choosing the stepsize α^k in a gradient method. Some of the most usual are:

- **Constant stepsize.** A fixed stepsize $s > 0$ is selected and

  $$\alpha^k = s, \quad k = 0, 1, \ldots$$

  In this simple rule, if the stepsize is too large, probably divergence will occur, while if the stepsize is too small, the rate of convergence may be very slow.

- **Minimization rule.** Take α^k such that the cost function is minimized along the direction d^k, that is α^k satisfies

  $$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k).$$

- **Limited minimization rule.** Fix a certain $s > 0$ and choose α^k such that

  $$f(x^k + \alpha^k d^k) = \min_{0 \leq \alpha \leq s} f(x^k + \alpha d^k).$$

Remark: The last two rules must be implemented together with an efficient one-dimensional minimization procedure.

Selecting the stepsize

- **Successive stepsize reduction.**
  In the simplest rule of this type an initial stepsize s is chosen. If

  $$f(x^k + sd^k) < f(x^k),$$

  we take $x^{k+1} = x^k + sd^k$ and continue the iterative procedure. If the above condition is not fulfilled the stepsize is reduced, perhaps repeatedly, by a certain factor, until the value of f is improved.

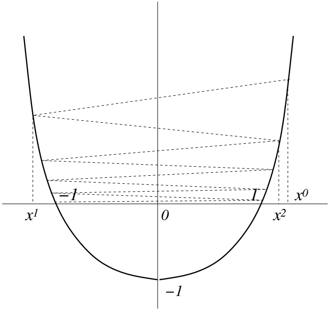
  Remark: It may happen that the cost improvement obtained at each iteration may not be substantial enough to guarantee convergence as is shown in the following example.

Successive stepsize reduction

Example.

Consider the function

$$f(x) = \begin{cases} \dfrac{3(1-x)^2}{4} - 2(1-x), & \text{if} \quad x > 1, \\ \dfrac{3(1+x)^2}{4} - 2(1+x), & \text{if} \quad x < -1, \\ x^2 - 1, & \text{if} \quad -1 \le x \le 1. \end{cases}$$



Clearly f is convex, continuously differentiable, is minimized at $x^* = 0$, and

$$f(x) < f(y) \quad \text{if and only if} \quad |x| < |y|.$$

Example (cont.)

The gradient of f is given by

$$\nabla f(x) = \begin{cases} \dfrac{3x}{2} + \frac{1}{2}, & \text{if} \quad x > 1, \\ \dfrac{3x}{2} - \frac{1}{2}, & \text{if} \quad x < -1, \\ 2x, & \text{if} \quad -1 \leq x \leq 1. \end{cases}$$

If we take $x > 1$, then

$$x - \nabla f(x) = x - \frac{3x}{2} - \frac{1}{2} = -\left(\frac{x}{2} + \frac{1}{2}\right),$$

from which it can be verified that

$$|x - \nabla f(x)| < |x| \quad \Rightarrow \quad f(x - \nabla f(x)) < f(x)$$

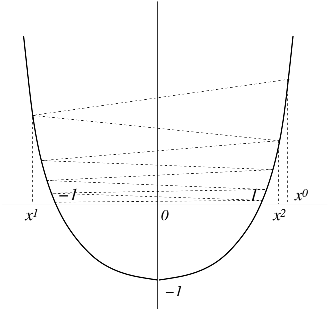
and also

$$x - \nabla f(x) < -1$$

Similarly, if $x < -1$, then $f(x - \nabla f(x)) < f(x)$, and $x - \nabla f(x) > 1$.

Example (cont.)

Consider the steepest descent iteration where the stepsize is successively reduced from an initial stepsize $s = 1$ until descent is obtained.



As in the figure, take $x^0 > 1$ (or $|x^0| > 1$), then $|x^1| > 1$, $|x^2| > 1$,.., $|x^k| > 1$ so it cannot converge to the unique minimum $x^* = 0$

Limit points of gradient methods

We want to analize when each limit point x^* of a sequence $\{x^k\}$ generated by a gradient method is a stationary point: $\nabla f(x^*) = 0$

- From Taylor's formula

$$f(x^{k+1}) = f(x^k) + \alpha^k (\nabla f(x^k))^T d^k + o(\alpha^k)$$

  we see that if the slope of f at x^k along the direction d^k, which is $(\nabla f(x^k))^T d^k$ is large, the rate of progress of the method will be, in principle, also large.

- On the other hand, if the directions d^k tend to become asymptotically orthogonal to the gradien direction

$$\frac{(\nabla f(x^k))^T d^k}{\|\nabla f(x^k)\| \|d^k\|} \to 0$$

  as x^k approaches a nonstationary point, there is a chance that the method will get "stuk" near that point.

- To ensure that this does not happen, we consider some non-orthogonality condition on the directions d^k, the so called gradient related condition.

The gradient related condition

Assume that the direction d^k is uniquely determined by the corresponding iterate x^k, that is, d^k is obtained as a given function of x^k

Definition

We say that the direction sequence $\{d^k\}$ is gradient related to $\{x^k\}$ if the following property holds: For any subsequence $\{x^k\}_{k\in\mathcal{K}}$ of $\{x^k\}$ convergent towards a non-stationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\lim_{k\to\infty} \sup_{k\in\mathcal{K}} \nabla f(x^k)^T d^k < 0. \tag{2}$$

- If $\{d^k\}$ is gradient related, it follows that if a subsequence $\{\nabla f(x^k)\}_{k\in\mathcal{K}}$tends to a nonzero vector, the corresponding sequence of directions d^k is bounded and does not tend to be orthogonal to $\nabla f(x^k)$.
- Roughly, this means that d^k does not becom "too small" or "too large" relative to $\nabla f(x^k)$, and that the angle between $\nabla f(x^k)$ and d^k does not get "too close" to 90 degrees

Successive stepsize reduction. Armijo rule

- The Armijo rule is essentially the succesive reduction rule suitably modified to eliminate the convergence difficulty shown in the above example.
- Fix scalars s, β i σ such that $0 < \beta < 1$ i $0 < \sigma < 1$.
- In $x^{k+1} = x^k + \alpha^k d$ take
$$\alpha^k = \beta^{m_k} s,$$
where m_k is the first non-negative integer m for which
$$f(x^k) - f(x^k + \beta^{m_k} s d^k) \geq -\sigma \beta^{m_k} s \nabla f(x^k)^T d^k$$

- The above rule means that the stepsizes $\beta^m s$, $m = 0, 1, \ldots$ are tried until the above inequality is satisfied (that guarantees that the cost improvement is large enough) and then we set $m_k = m$.
- Usually σ is chosen close to zero, for instance $\sigma \in [10^{-5}, 10^{-1}]$. The reductionfactor β is usually chosen between $1/2$ and $1/10$, depending on the confidence we have on the quality on the initial stepsize s.

Convergence

Let us study the convergence behavior of the gradient methods. The following theorem is the main convergence result.

Theorem

Let $\{x^k\}$ be a sequence generated by a gradient method

$$x^{k+1} = x^k + \alpha^k d^k,$$

and assume that $\{d^k\}$ is gradient related to $\{x^k\}$, and that α^k is chosen by the Armijo rule. Then, every limit point of $\{x^k\}$ is a stationary point ($\nabla f(x^*) = 0$).

Proof of the convergence Theorem

Proof

Consider the Armijo rule and, to arrive to a contradiction, assume that x^* is a limit point of $\{x^k\}$ such that $\nabla f(x^*) \neq 0$.

- Since $\{f(x^k)\}$ is monotonically non-increasing, then $\{f(x^k)\}$ either converges to a finite value or diverges to $-\infty$.
- Since f is continuous, then

  $$\lim_{k\to\infty} f(x^k) = f(x^*),$$

  so, it follows that

  $$f(x^k) - f(x^{k+1}) \to 0.$$
- By the definition of the Armijo rule, we have

  $$f(x^k) - f(x^{k+1}) \geq -\sigma\alpha^k \nabla f(x^k)^T d^k, \tag{3}$$

  hence $\alpha^k \nabla f(x^k)^T d^k \to 0$.
- Let $\{x^k\}_{k\in\mathcal{K}}$ be a subsequence converging to x^*. Since $\{d^k\}$ is gradient related and $\nabla f(\overline{x}) \neq 0$, we have that

  $$\lim_{k\to\infty} \sup_{k\in\mathcal{K}} \nabla f(x^k)^T d^k < 0 \quad \Rightarrow \quad \{\alpha^k\}_{\mathcal{K}} \to 0.$$

Proof of the convergence Theorem (cont.)

By the definition of the Armijo rule, we must have for some index $\bar{k} \geq 0$ that

$$f(x^k) - f\left(x^k + \frac{\alpha^k}{\beta} d^k\right) < -\sigma \frac{\alpha^k}{\beta} \nabla f(x^k)^T d^k, \quad \forall k \in \mathcal{K}, k \geq \bar{k}, \tag{4}$$

that is, the initial stepsize s will be reduced at least once for all $k \in \mathcal{K}$, $k \geq \bar{k}$. Denote

$$p^k = \frac{d^k}{\|d^k\|}, \quad \bar{\alpha}^k = \frac{\alpha^k \|d^k\|}{\beta}.$$

since $\{d^k\}$ is gradient related, the sequence $\{\|d^k\|\}_{\mathcal{K}}$ is bounded, and it follows that

$$\{\bar{\alpha}^k\}_{\mathcal{K}} \to 0.$$

SInce $\|p^k\| = 1$ for all $k \in \mathcal{K}$, there exist a subsequence $\{p^k\}_{\overline{\mathcal{K}}}$ of $\{p^k\}_{\mathcal{K}}$ such that

$$\{p^k\}_{\mathcal{K}} \to \bar{p},$$

where $\bar{p}$ is some vector with $\|\bar{p}\| = 1$. From equation(4), we have

$$\frac{f(x^k) - f(x^k + \bar{\alpha}^k p^k)}{\bar{\alpha}^k} < -\sigma \nabla f(x^k)^T p^k, \quad \forall k \in \mathcal{K}, k \geq \bar{k}. \tag{5}$$

Proof of the convergence Theorem (cont.)

Using the mean value Theorem, the above relation is written as

$$-\nabla f(x^k + \tilde{\alpha}^k p^k)^T p^k < -\sigma \nabla f(x^k)^T p^k, \quad \forall k \in \mathcal{K}, k \geq \overline{k},$$

where $\tilde{\alpha}^k \in [0, \overline{\alpha}^k]$. Taking limits in the above equation one gets

$$-\nabla f(\overline{x})^T \overline{p} \leq -\sigma \nabla f(\overline{x})^T \overline{p},$$

this is

$$0 \leq (1 - \sigma) \nabla f(\overline{x})^T \overline{p}.$$

Since $\sigma < 1$, it follows that

$$0 \leq \nabla f(\overline{x})^T \overline{p}. \tag{6}$$

On the other hand we have

$$\nabla f(x^k)^T p^k = \frac{\nabla f(x^k)^T d^k}{\|d^k\|}.$$

By taking the limit as $k \in \mathcal{K}$, $k \to \infty$

$$\nabla f(\overline{x})^T \overline{p} \leq \frac{\limsup_{k \to \infty, k \in \mathcal{K}} \nabla f(x^k)^T d^k}{\limsup_{k \to \infty, k \in \mathcal{K}} \|d^k\|} < 0,$$

which contradicts (6). This proves the result.

Second convergence Theorem

Theorem
Let $\{x^k\}$ *be a sequence generated by a gradient method*

$$x^{k+1} = x^k + \alpha^k d^k,$$

and assume that $\{d^k\}$ *is gradient related to* $\{x^k\}$*, and that* α^k *is chosen by the minimization rule, or the limited minimization rule. Then, every limit point of* $\{x^k\}$ *is a stationary point* ($\nabla f(x^*) = 0$)*.*

Proof
Consider the minimization rule, and let $\{x^k\}_\mathcal{K}$ converge to $\overline{x}$ with $\nabla f(\overline{x}) \neq 0$. Again we have that $\{f(x^k)\}$ decreases monotonically to $f(\overline{x})$. Let $\tilde{x}^{k+1}$ be the point generated from x^k using the Armijo rule, and let $\tilde{\alpha}^k$ be the corresponding stepsize. We have

$$f(x^k) - f(x^{k+1}) \geq f(x^k) - f(\tilde{x}^{k+1}) \geq -\sigma \tilde{\alpha}^k \nabla f(x^k)^T d^k.$$

By repeating the argument of the earlier proof following equation (2) replacing α^k by $\tilde{\alpha}^k$ we can obtain a contradiction. In particular we have

$$\{\tilde{\alpha}^k\}_\mathcal{K} \to 0,$$

and, by the definition of the Armijo rule, we have for some index $\overline{k} \geq 0$

$$f(x^k) - f\left(x^k + \frac{\alpha^k}{\beta} d^k\right) < -\sigma \frac{\alpha^k}{\beta} \nabla f(x^k)^T d^k, \quad \forall k \in \mathcal{K}, k \geq \overline{k},$$

Proof of the second convergence Theorem (cont.)

Proceeding as earlier, we obtain (4) and (5) with $\overline{\alpha}^k = \tilde{\alpha}^k \|d^k\|/\beta$, and a contradiction.

The argument just used establishes that any stepsize rule that gives a larger reduction in cost at each iteration than the Armijo rule inherits its convergence properties. This also proves the proposition for the limited minimization rule