In other words: We can be sure that the Gradient norms $\|\nabla f^k\|$ converge to zero, provided that the search directions are never too close to orthogonality with the gradient.
In particular, the method of steepest descent (for which the search direction p^k is parallel to the neg. gradient) produces a gradient sequence that converges to zero, provided that it uses a line search satisfying the Wolfe conditions.

Example: (Exam 2021)

Consider the function $f: \mathbb{R}^n \to \mathbb{R}$,

$$f(x) := x^T H x - b^T x,$$

with a symm. pos. def. matrix $H \in \mathbb{R}^{n,n}$.

Let $\underbrace{x^0}_{\text{starting point}} - \underbrace{\tilde{x}}_{\text{minimizer}}$ be an eigenvector of H.

Show that the iteration $x^{k+1} = x^k + \alpha_k p^k$ with the search direction $p^k = -\nabla f^k$ finds the minimizer $\tilde{x}$ in one step.
Explain how α_k needs to be chosen for this result to hold.

We have $\nabla f = 2Hx - b$, and so,

for a minimize, we get

$$\nabla f = \underline{2H\tilde{x} - b = 0},$$

s.t. $\tilde{x} = \frac{1}{2} H^{-1} b$ (H^{-1} exists because H is a symm. pos. def. matrix).

We get for the search direction:

$$p^k = -\nabla f(x^k) = -2Hx^k + b,$$

and so, $p^0 = b - 2Hx^0$.

We want to show that

$$\tilde{x} = x^0 + \alpha p^0,$$

which is equivalent to

$$x^0 - \tilde{x} = -\alpha p^0. \qquad (1)$$

We know:

$$-p^0 = 2Hx^0 - b = 2Hx^0 - 2H\tilde{x}$$
$$= 2H(x^0 - \tilde{x}).$$

With this, our claim (1) is equivalent to

$$(I - 2\alpha_k H)(x^0 - \tilde{x}) = 0 \qquad (2)$$

Since $x^0 - \tilde{x}$ is an eigenvector of H, we have

$$(x^0 - \tilde{x}) \cdot \underline{2} = H(x^0 - \tilde{x}),$$

eigenvalue

or, equivalently,

$$(x^0 - \tilde{x})(\lambda I - H) = 0$$

$$\Leftrightarrow (x^0 - \tilde{x})\left(I - \frac{1}{\lambda} H\right) = 0$$

Comparing this with (2), we see that $\alpha_k = \frac{1}{2\lambda}$ yields the desired result.

Note that $\lambda > 0$, because H is symm. pos. def., and so, α_k is well-defined.

Sufficient Decrease and Backtracking

By using a so-called backtracking approach, we can use just the sufficient decrease (or Armijo) condition to terminate the line search procedure. This condition is

$$f(x^k + \alpha p^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T p^k .$$

Algorithm (Backtracking line search)

Choose $\bar{\alpha} > 0$, $\rho \in (0,1)$, $c \in (0,1)$;

Set $\alpha := \bar{\alpha}$

repeat until

$$f(x^k + \alpha p^k) \leq f(x^k) + c \alpha \nabla f^{kT} p^k$$

$$\alpha := \rho \cdot \alpha$$

end

Terminate with $\alpha_k := \alpha$.

An acceptable step length will be found after a finite number of trials, because α_k will eventually become small enough for the sufficient decrease condition to hold.

In this procedure, the initial step length $\bar{\alpha}$ is chosen to be 1 in Newton and Quasi-Newton methods, but it can have other values in other algorithms, such as steepest descent or conjugate gradient.

In practice, the contraction factor ρ is often allowed to vary at each iteration of the line search.

The backtracking approach ensures either that the selected step length α_k is some fixed value (the initial choice $\bar{\alpha}$), or else that it is short enough to satisfy the sufficient decrease condition, but not too short.

Now, let us comider the iteration

$$x^{k+1} = x^k + \alpha_k p^k$$

with $\boxed{B_k p^k = -\nabla f(x^k),}$

where $B_k = B_k^T$, and α_k is selected using the backtracking (Armijo) line search with parameters c and β.

We have the following convergence theorem:

Theorem: Assume that

1. $f: \mathbb{R}^n \to \mathbb{R}$ is continuosly differentiable.
2. The set $\{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$ is bounded.
3. The matrices B_k are uniformly positive definite, that is,

$$\forall k, \forall x \neq 0: \quad x^T B_k x > \alpha \|x\|^2$$

for some $\alpha > 0$.

Moreover, B_k are bounded, that is, there is some $m > 0$, $M > 0$ with

$$m \leq \lambda_{min}(B_k) \leq \lambda_{max}(B_k) \leq M,$$

where λ_{min} and λ_{max} are the smallest and largest eigenvalues of B_k, respectively.

Then the sequence $\{x^k\}$ is bounded, and every limit point is a stationary point for f.

Exact line search methods

Steepest Descent - Zick-Zack-Theorem

Theorem ("Zick-Zack-Theorem").
Let x^k be the sequence generated by the steepest descent algorithm with exact line search. Then, for all k, $x^{k+1} - x^k$ is orthogonal to $x^{k+2} - x^{k+1}$.

Proof: Our iterations read

$$x^{k+1} = x^k - \alpha_k \nabla f^k$$
$$x^{k+2} = x^{k+1} - \alpha_{k+1} \nabla f^{k+1}.$$

Hence,

$$\langle x^{k+1} - x^k, x^{k+2} - x^{x+1} \rangle$$
$$= \alpha_k \alpha_{k+1} \langle \nabla f^k, \nabla f^{k+1} \rangle.$$

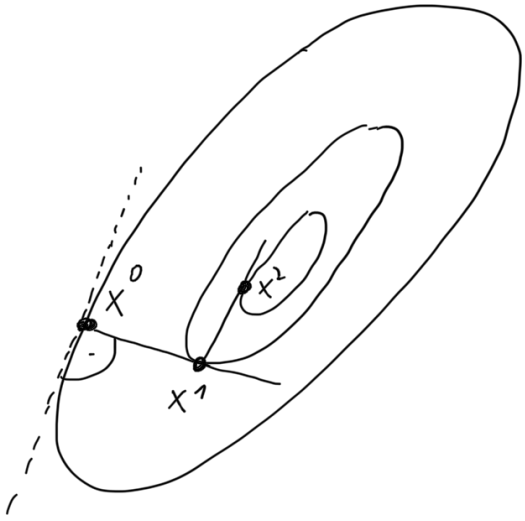
Recall that if α_k minimizes $\phi(\alpha_k) = f(x^k - \alpha_k \nabla f^k)$, then

$$\phi'(\alpha_k) = 0 = -\nabla f^{kT} \nabla f(x^k - \alpha_k \nabla f^k)$$
$$= \nabla f^{kT} \nabla f^{k+1}.$$

Therefore,

$$\langle x^{k+1} - x^k, x^{k+2} - x^{k+1} \rangle$$
$$= \alpha_k \alpha_{k+1} \langle \nabla f^k, \nabla f^{k+1} \rangle = 0.$$

∎



Steepest Descent Method - Convergence Rate -

The goal is to design optimization algorithms with good convergence properties: We need to ensure that the search directions p^k do not tend to become orthogonal to ∇f^k, according to Zoutendijk's result and its consequences.
One could compute $\cos\theta_k$ at every iteration and turn p^k towards the steepest descent direction in case that $\cos\theta_k < \delta$ for some preselected $\delta > 0$.
However, such angle tests are undesirable, because: 1. They may impede a fast rate of convergence
2. They may destroy the invariance of Quasi-Newton-Methods.

Let us recall some basic terms:

Let us see

Definition: Let $\{x^k\}$ converge to x^*. Then we say that the convergence is of order p ($p \geq 1$) with factor γ ($\gamma > 0$) if there exists some k_0 s.t. for all $k \geq k_0$:

$$\|x^{k+1} - x^*\| \leq \gamma \|x^k - x^*\|^p.$$

Remark:
- The large the power p, the faste the convergence.
- For the same p, the smalle γ, the faste the convergence.
- If $\{x^k\}$ converges with orde p, it also converges with orde p' for any $p' \geq p$.
- If $\{x^k\}$ converges with orde p and factor γ, then it also converges with orde p and factor γ' for any $\gamma' \geq \gamma$.

Therefore, we typically look for the largest p and the smallest γ for which the inequality holds.

Terminology:
- If $p = 1$ and $\gamma < 1$, we say that the convergence is linear.
- If $p = 1$ and $\gamma = 1$, we say that it is sublinear

the convergence is linear.

- If $p=2$, the convergence is called quadratic.

Example: Let $x^k = 1 + \frac{1}{2^k}$ with $x^* = 1$.

We have

$$\left| 1 + \frac{1}{2^{k+1}} - 1 \right| = \frac{1}{2^{k+1}}$$

$$= \frac{1}{2} \cdot \frac{1}{2^k}$$

$$= \frac{1}{2} \cdot \left| 1 + \frac{1}{2^k} - 1 \right|$$

Thus, the iterations converge linearly to x^* with factor $\rho = \frac{1}{2}$.