

# Unconstrained minimization

Lectures for PHD course on  
Non-linear equations and numerical optimization

Enrico Bertolazzi

DIMS – Università di Trento

March 2005



## Outline

- 1 General iterative scheme
  - Descent direction failure
- 2 Backtracking Armijo line-search
  - Global convergence of backtracking Armijo line-search
  - Global convergence of steepest descent
- 3 Wolfe–Zoutendijk global convergence
  - The Wolfe conditions
  - The Armijo-Goldstein conditions
- 4 Algorithms for line-search
  - Armijo Parabolic-Cubic search
  - Wolfe linesearch



## The problem

(1/3)

Given  $f : \mathbb{R}^n \mapsto \mathbb{R}$ :

$$\underset{\mathbf{x} \in \mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x})$$

the following regularity about  $f(\mathbf{x})$  is assumed in the following:

### Assumption (Regularity assumption)

We assume  $f \in C^1(\mathbb{R}^n)$  with Lipschitz continuous gradient, i.e. there exists  $\gamma > 0$  such that

$$\|\nabla f(\mathbf{x})^T - \nabla f(\mathbf{y})^T\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$



## The problem

(2/3)

### Definition (Global minimum)

Given  $f : \mathbb{R}^n \mapsto \mathbb{R}$  a point  $\mathbf{x}_* \in \mathbb{R}^n$  is a **global minimum** if

$$f(\mathbf{x}_*) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

### Definition (Local minimum)

Given  $f : \mathbb{R}^n \mapsto \mathbb{R}$  a point  $\mathbf{x}_* \in \mathbb{R}^n$  is a **local minimum** if

$$f(\mathbf{x}_*) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in B(\mathbf{x}_*; \delta).$$

Obviously a global minimum is a local minimum. Find a global minimum in general is not an easy task. The algorithms presented in the sequel will approximate **local minima's**.



## Definition (Strict global minimum)

Given  $f: \mathbb{R}^n \mapsto \mathbb{R}$  a point  $\mathbf{x}_* \in \mathbb{R}^n$  is a **strict global minimum** if

$$f(\mathbf{x}_*) < f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{x}_*\}.$$

## Definition (Strict local minimum)

Given  $f: \mathbb{R}^n \mapsto \mathbb{R}$  a point  $\mathbf{x}_* \in \mathbb{R}^n$  is a **strict local minimum** if

$$f(\mathbf{x}_*) < f(\mathbf{x}), \quad \forall \mathbf{x} \in B(\mathbf{x}_*; \delta) \setminus \{\mathbf{x}_*\}.$$

Obviously a strict global minimum is a strict local minimum.



## Lemma (First order Necessary condition for local minimum)

Given  $f: \mathbb{R}^n \mapsto \mathbb{R}$  satisfying the regularity assumption. If a point  $\mathbf{x}_* \in \mathbb{R}^n$  is a **local minimum** then

$$\nabla f(\mathbf{x}_*)^T = \mathbf{0}.$$

## Proof.

Consider a generic direction  $\mathbf{d}$ , then for  $\delta$  small enough we have

$$\lambda^{-1}(f(\mathbf{x}_* + \lambda \mathbf{d}) - f(\mathbf{x}_*)) \leq 0, \quad 0 < \lambda < \delta$$

so that

$$\lim_{\lambda \rightarrow 0} \lambda^{-1}(f(\mathbf{x}_* + \lambda \mathbf{d}) - f(\mathbf{x}_*)) = \nabla f(\mathbf{x}_*)^T \mathbf{d} \leq 0,$$

because  $\mathbf{d}$  is a generic direction we have  $\nabla f(\mathbf{x}_*)^T = \mathbf{0}$ .  $\square$



- 1 The first order necessary condition do not discriminate maximum, minimum, or saddle points.
- 2 To discriminate maximum and minimum we need more information, e.g. second order derivative of  $f(\mathbf{x})$ .
- 3 With second order derivative we can build **necessary** and **sufficient** condition for a minima.
- 4 In general using only first and second order derivative at the point  $\mathbf{x}_*$  it is not possible to deduce a **necessary and sufficient** condition for a minima.



## Lemma (Second order Necessary condition for local minimum)

Given  $f \in C^2(\mathbb{R}^n)$  if a point  $\mathbf{x}_* \in \mathbb{R}^n$  is a **local minimum** then  $\nabla f(\mathbf{x}_*)^T = \mathbf{0}$  and  $\nabla^2 f(\mathbf{x}_*)$  is **semi-definite positive**, i.e.

$$\mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} \geq 0, \quad \forall \mathbf{d} \in \mathbb{R}^n$$

## Example

This condition is only, necessary, in fact consider  $f(\mathbf{x}) = x_1^2 - x_2^3$ ,

$$\nabla f(\mathbf{x}) = (2x_1, -3x_2^2), \quad \nabla^2 f(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & -6x_2 \end{pmatrix}$$

for the point  $\mathbf{x}_* = \mathbf{0}$  we have  $\nabla f(\mathbf{0}) = \mathbf{0}$  and  $\nabla^2 f(\mathbf{0})$  semi-definite positive, but  $\mathbf{0}$  is a saddle point not a minimum.



### Proof.

The condition  $\nabla f(\mathbf{x}_*)^T = \mathbf{0}$  comes from first order necessary conditions. Consider now a generic direction  $\mathbf{d}$ , and the finite difference:

$$\frac{f(\mathbf{x}_* + \lambda \mathbf{d}) - 2f(\mathbf{x}_*) + f(\mathbf{x}_* - \lambda \mathbf{d})}{\lambda^2} \geq 0$$

by using Taylor expansion for  $f(\mathbf{x})$

$$f(\mathbf{x}_* \pm \lambda \mathbf{d}) = f(\mathbf{x}_*) \pm \nabla f(\mathbf{x}_*)^T \lambda \mathbf{d} + \lambda^2 \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} + o(\lambda^2)$$

and from the previous inequality

$$\mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} + o(\lambda^2)/\lambda^2 \geq 0$$

taking the limit  $\lambda \rightarrow 0$  and from the arbitrariness of  $\mathbf{d}$  we have that  $\nabla^2 f(\mathbf{x}_*)$  must be semi-definite positive.  $\square$

## Second order sufficient condition

### Lemma (Second order sufficient condition for local minimum)

Given  $f \in \mathcal{C}^2(\mathbb{R}^n)$  if a point  $\mathbf{x}_* \in \mathbb{R}^n$  satisfy:

- 1  $\nabla f(\mathbf{x}_*)^T = \mathbf{0}$ ;
- 2  $\nabla^2 f(\mathbf{x}_*)$  is **definite positive**; i.e.

$$\mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} > 0, \quad \forall \mathbf{d} \in \mathbb{R}^n \setminus \{\mathbf{x}_*\}$$

then  $\mathbf{x}_* \in \mathbb{R}^n$  is a **strict local minimum**.

### Remark

Because  $\nabla^2 f(\mathbf{x}_*)$  is symmetric we can write

$$\lambda_{\min} \mathbf{d}^T \mathbf{d} \leq \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} \leq \lambda_{\max} \mathbf{d}^T \mathbf{d}$$

If  $\nabla^2 f(\mathbf{x}_*)$  is positive definite we have  $\lambda_{\min} > 0$ .

### Proof.

Consider now a generic direction  $\mathbf{d}$ , and the Taylor expansion for  $f(\mathbf{x})$

$$\begin{aligned} f(\mathbf{x}_* + \mathbf{d}) &= f(\mathbf{x}_*) + \nabla f(\mathbf{x}_*)^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \nabla^2 f(\mathbf{x}_*) \mathbf{d} + o(\|\mathbf{d}\|^2) \\ &\geq f(\mathbf{x}_*) + \lambda_{\min} \|\mathbf{d}\|^2 + o(\|\mathbf{d}\|^2) \\ &\geq f(\mathbf{x}_*) + \lambda_{\min} \|\mathbf{d}\|^2 \left(1 + o(\|\mathbf{d}\|^2)/\|\mathbf{d}\|^2\right) \end{aligned}$$

choosing  $\mathbf{d}$  small enough we can write

$$f(\mathbf{x}_* + \mathbf{d}) \geq f(\mathbf{x}_*) + \frac{\lambda_{\min}}{2} \|\mathbf{d}\|^2 > f(\mathbf{x}_*), \quad \mathbf{d} \neq \mathbf{0}, \|\mathbf{d}\| \leq \delta.$$

i.e.  $\mathbf{x}_*$  is a strict minimum.  $\square$

## Outline

- 1 General iterative scheme
  - Descent direction failure
- 2 Backtracking Armijo line-search
  - Global convergence of backtracking Armijo line-search
  - Global convergence of steepest descent
- 3 Wolfe-Zoutendijk global convergence
  - The Wolfe conditions
  - The Armijo-Goldstein conditions
- 4 Algorithms for line-search
  - Armijo Parabolic-Cubic search
  - Wolfe linesearch

## How to find a minimum

Given  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ : minimize  $x \in \mathbb{R}^n$   $f(x)$ .

- 1 We can solve the problem by solving the **necessary condition**, i.e. by solving the nonlinear systems

$$\nabla f(x)^T = 0.$$

- 2 Using such an approach we loose the information about  $f(x)$ .
- 3 Moreover such an approach can find solution corresponding to a maximum or saddle points.
- 4 A better approach is to use all the information and try to build **minimizing procedure**, i.e. procedures that, starting from a point  $x_0$  build a sequence  $\{x_k\}$  such that  $f(x_{k+1}) \leq f(x_k)$ . In this way, at least, we avoid to converge to a **strict maximum**.



## Line-search Methods

A generic iterative minimization procedure can be sketched as follows:

- calculate a **search direction**  $p_k$  from  $x_k$
- ensure that this direction is a **descent direction**, i.e.

$$\nabla f(x_k)p_k < 0, \quad \text{whenever } \nabla f(x_k)^T \neq 0$$

so that, at least for small steps along  $p_k$ , the objective function  $f(x)$  will be reduced

- use **line-search** to calculate a suitable step-length  $\alpha_k > 0$  so that

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

- Update the point:

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



## Iterative Methods

- in practice very rare to be able to provide explicit minimizer.
- iterative method: given starting **guess**  $x_0$ , generate the sequence,

$$\{x_k\}, \quad k = 1, 2, \dots$$

- **AIM**: ensure that (a subsequence) has some favorable limiting properties:
  - satisfies first-order necessary conditions
  - satisfies second-order necessary conditions



## Generic minimization algorithm

Written with a pseudo-code the minimization procedure is the following algorithm:

## Generic minimization algorithm

```

Given an initial guess  $x_0$ , let  $k = 0$ ;
while not converged do
  Find a descent direction  $p_k$  at  $x_k$ ;
  Compute a step size  $\alpha_k$  using a line-search along  $p_k$ .
  Set  $x_{k+1} = x_k + \alpha_k p_k$  and increase  $k$  by 1.
end while
  
```

The crucial points which differentiate the algorithms are:

- 1 The computation of the direction  $p_k$ ;
- 2 The computation of the step size  $\alpha_k$ .



## Practical Line-search methods

- The first developed minimization algorithms try to solve

$$\alpha_k = \arg \min_{\alpha > 0} f(\mathbf{x}_k + \alpha \mathbf{p}_k)$$

- performing **exact line-search** by univariate minimization;
  - rather expensive and certainly not cost effective.
- Modern methods implements **inexact** line-search:
  - ensure steps are neither too long nor too short
  - try to pick *useful* initial step size for fast convergence
  - best methods are based on:
    - backtracking-Armijo search;
    - Armijo-Goldstein search;
    - Franke-Wolfe search;



## backtracking line-search

To obtain a monotone decreasing sequence we can use the following algorithm:

## Backtracking line-search

```

Given  $\alpha_{\text{init}}$  (e.g.,  $\alpha_{\text{init}} = 1$ );
Given  $\tau \in (0, 1)$  typically  $\tau = 0.5$ ;
Let  $\alpha^{(0)} = \alpha_{\text{init}}$ ;
while not  $f(\mathbf{x}_k + \alpha^{(\ell)} \mathbf{p}_k) < f(\mathbf{x}_k)$  do
  set  $\alpha^{(\ell+1)} = \tau \alpha^{(\ell)}$ ;
  increase  $\ell$  by 1;
end while
Set  $\alpha_k = \alpha^{(\ell)}$ .
  
```

To be effective the previous algorithm should terminate in a finite number of steps. The next lemma assure that if  $\mathbf{p}_k$  is a descent direction then the algorithm terminate.



## Existence of a descent step

(1/3)

## Lemma (Descent Lemma)

Suppose that  $f(\mathbf{x})$  satisfy the standard assumptions and that  $\mathbf{p}_k$  is a descent direction at  $\mathbf{x}_k$ , i.e.  $\nabla f(\mathbf{x}_k) \mathbf{p}_k < 0$ . Then we have

$$f(\mathbf{x}_k + \alpha \mathbf{p}_k) \leq f(\mathbf{x}_k) + \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \frac{\gamma}{2} \alpha^2 \|\mathbf{p}_k\|^2$$

for all  $\alpha \in [0, \alpha_k^*]$  where  $\alpha_k^* = \frac{-2\nabla f(\mathbf{x}_k) \mathbf{p}_k}{\gamma \|\mathbf{p}_k\|^2} > 0$

## Assumption (Regularity assumption)

We assume  $f \in \mathcal{C}^1(\mathbb{R}^n)$  with Lipschitz continuous  $\gamma$  gradient, i.e. there exists  $\gamma > 0$  such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$



## Existence of a descent step

(2/3)

## Proof.

Let be  $g(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{p}_k)$  then we can write:

$$\begin{aligned}
 g(\alpha) - g(0) &= \int_0^\alpha g'(\xi) d\xi = \alpha g'(0) + \int_0^\alpha (g'(\xi) - g'(0)) d\xi \\
 &= \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \int_0^\alpha (\nabla f(\mathbf{x}_k + \xi \mathbf{p}_k) - \nabla f(\mathbf{x}_k)) \mathbf{p}_k d\xi \\
 &\leq \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \int_0^\alpha \|\nabla f(\mathbf{x}_k + \xi \mathbf{p}_k) - \nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\| d\xi \\
 &\leq \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \|\mathbf{p}_k\|^2 \int_0^\alpha \gamma \xi d\xi \\
 &\leq \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \frac{\gamma \alpha^2}{2} \|\mathbf{p}_k\|^2 = \alpha \left[ \nabla f(\mathbf{x}_k) \mathbf{p}_k + \frac{\gamma \alpha}{2} \|\mathbf{p}_k\|^2 \right].
 \end{aligned}$$

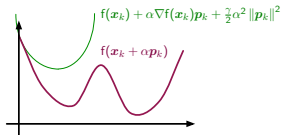
now the lemma follows trivially.



## Existence of a descent step

(3/3)

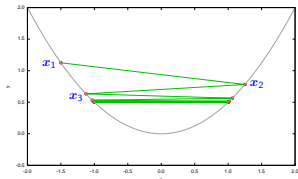
- The **descent lemma** means that there is a parabola that is entirely over the function  $f(x)$  in the direction  $p_k$  if this is a descent direction.
- The second part of the lemma permits to ensure a **minimal** reduction if the step length is chosen to be  $\alpha_k = \alpha_k^*/2$ .



## Steps may be too long

The objective function is  $f(x) = x^2$  and the iterates are generated by the descent directions  $p_k = (-1)^{k+1}$  from  $x_0 = 2$  with:

$$x_{k+1} = x_k + \alpha_k p_k, \quad \alpha_k = 2 + 3 \cdot 2^{-(k+1)}$$



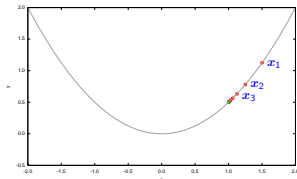
## Descent direction failure

- The simple request to have a descent direction may be not enough.
- In fact, step length may be asymptotically **too short**
- Or step length may be asymptotically **too long**

## Steps may be too short

The objective function is  $f(x) = x^2$  and the iterates are generated by the descent directions  $p_k = -1$  from  $x_0 = 2$  with:

$$x_{k+1} = x_k + \alpha_k p_k, \quad \alpha_k = 2^{-(k+1)}$$



## Outline

- 1 General iterative scheme
  - Descent direction failure
- 2 Backtracking Armijo line-search
  - Global convergence of backtracking Armijo line-search
  - Global convergence of steepest descent
- 3 Wolfe–Zoutendijk global convergence
  - The Wolfe conditions
  - The Armijo-Goldstein conditions
- 4 Algorithms for line-search
  - Armijo Parabolic-Cubic search
  - Wolfe line-search

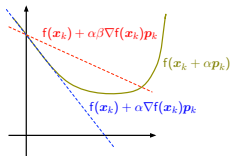


## Armijo condition

To prevent large steps relative to the decreasing of  $f(x)$  we require that

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \alpha_k \beta \nabla f(x_k) p_k$$

for some  $\beta \in (0, 1)$ . Typical values of  $\beta$  ranges from  $10^{-4}$  to 0.1.



## Backtracking Armijo line-search

Given  $\alpha_{\text{init}}$  (e.g.,  $\alpha_{\text{init}} = 1$ );  
 Given  $\tau \in (0, 1)$  typically  $\tau = 0.5$ ;  
 Let  $\alpha^{(0)} = \alpha_{\text{init}}$ ;  
**while** not  $f(x_k + \alpha^{(\ell)} p_k) \leq f(x_k) + \alpha^{(\ell)} \beta \nabla f(x_k) p_k$  **do**  
   set  $\alpha^{(\ell+1)} = \tau \alpha^{(\ell)}$ ;  
   increase  $\ell$  by 1;  
**end while**  
 Set  $\alpha_k = \alpha^{(\ell)}$ .

- Backtracking Armijo line-search prevents the step from getting too large.
- Now the question is: will the backtracking Armijo line-search **terminate** in a finite number of steps?



## Finite termination of Armijo line-search

## Theorem (Finite termination of Armijo linesearch)

Suppose that  $f(x)$  satisfy the standard assumptions and  $\beta \in (0, 1)$  and that  $p_k$  is a descent direction at  $x_k$ . Then the Armijo condition

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \alpha_k \beta \nabla f(x_k) p_k$$

is satisfied for all  $\alpha_k \in [0, \omega_k]$  where  $\omega_k = \frac{2(\beta - 1) \nabla f(x_k) p_k}{\gamma \|p_k\|^2}$

## Assumption (Regularity assumption)

We assume  $f \in C^1(\mathbb{R}^n)$  with Lipschitz continuous gradient, i.e. there exists  $\gamma > 0$  such that

$$\|\nabla f(x) - \nabla f(y)\| \leq \gamma \|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$



## Finite termination of Armijo line-search

To prove finite termination we need the following Taylor expansion due to the regularity assumption:

$$f(\mathbf{x} + \alpha \mathbf{p}) = f(\mathbf{x}) + \alpha \nabla f(\mathbf{x}) \mathbf{p} + E \quad \text{where} \quad |E| \leq \frac{\gamma}{2} \alpha^2 \|\mathbf{p}\|^2$$

## Proof.

If  $\alpha \leq \omega_k$  we have  $\alpha \gamma \|\mathbf{p}_k\|^2 \leq 2(\beta - 1) \nabla f(\mathbf{x}_k) \mathbf{p}_k$  and by using Taylor expansion

$$\begin{aligned} f(\mathbf{x}_k + \alpha \mathbf{p}_k) &\leq f(\mathbf{x}_k) + \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \frac{\gamma}{2} \alpha^2 \|\mathbf{p}_k\|^2 \\ &\leq f(\mathbf{x}_k) + \alpha \nabla f(\mathbf{x}_k) \mathbf{p}_k + \alpha(\beta - 1) \nabla f(\mathbf{x}_k) \mathbf{p}_k \\ &\leq f(\mathbf{x}_k) + \alpha \beta \nabla f(\mathbf{x}_k) \mathbf{p}_k \end{aligned}$$



## Finite termination of Armijo line-search

## Corollary (Finite termination of Armijo line-search)

Suppose that  $f(\mathbf{x})$  satisfy the standard assumptions and  $\beta \in (0, 1)$  and that  $\mathbf{p}_k$  is a descent direction at  $\mathbf{x}_k$ . Then the step-size generated by then backtracking-Armijo line-search terminates with

$$\alpha_k \geq \min \{ \alpha_{\text{init}}, \tau \omega_k \}, \quad \omega_k = 2(\beta - 1) \nabla f(\mathbf{x}_k) \mathbf{p}_k / (\gamma \|\mathbf{p}_k\|^2)$$

## Proof.

Line-search will terminate as soon as  $\alpha^{(\ell)} \leq \omega_k$ :

- ① May be that  $\alpha_{\text{init}}$  satisfies the Armijo condition  $\Rightarrow \alpha_k = \alpha_{\text{init}}$ .
- ② Otherwise in the last line-search iteration we have

$$\alpha^{(\ell-1)} > \omega_k, \quad \alpha_k = \alpha^{(\ell)} = \tau \alpha^{(\ell-1)} > \tau \omega_k.$$

Combining these 2 cases gives the required result. □



## Backtracking-Armijo line-search

- ① The previous analysis permit to say that Backtracking-Armijo line-search ends in a finite number of steps.
- ② The line-search produce a step length **not too long** due to the condition

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + \alpha_k \beta \nabla f(\mathbf{x}_k) \mathbf{p}_k$$

- ③ The line-search produce a step length **not too short** due to the **finite termination** theorem.
- ④ Armijo line-search can be improved by adding some further requirements on the step length acceptance criteria.



## Global convergence

## Theorem (Global convergence)

Suppose that  $f(\mathbf{x})$  satisfy the standard assumptions, then, for the iterates generated by the **Generic minimization algorithm with backtracking Armijo line-search** either:

- ①  $\nabla f(\mathbf{x}_k)^T = \mathbf{0}$  for some  $k \geq 0$ ;
- ② or  $\lim_{k \rightarrow \infty} f(\mathbf{x}_k) = -\infty$ ;
- ③ or  $\lim_{k \rightarrow \infty} |\nabla f(\mathbf{x}_k) \mathbf{p}_k| \min \{ 1, \|\mathbf{p}_k\|^{-1} \} = 0$ .

## Remark

If the theorem, point 1 means that we found a stationary point in a finite number of steps. Point 2 means that function  $f(\mathbf{x})$  is unbounded below, so that a minimum does not exist. Point 3 alone do not imply convergence, but if  $\nabla f(\mathbf{x}_k)$  and  $\mathbf{p}_k$  do not become orthogonal and  $\|\mathbf{p}_k\| \not\rightarrow 0$  then  $\|\nabla f(\mathbf{x}_k)\| \rightarrow 0$ .





Proof.

(1/3).

Assume points 1 and 2 are not satisfied, then we prove point 3.  
Consider

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \alpha_k \beta \nabla f(\mathbf{x}_k) \mathbf{p}_k \leq f(\mathbf{x}_0) + \sum_{j=0}^k \alpha_j \beta \nabla f(\mathbf{x}_j) \mathbf{p}_j$$

by the fact that  $\mathbf{p}_k$  is a descent direction we have that the series:

$$\sum_{j=0}^{\infty} \alpha_j |\nabla f(\mathbf{x}_j) \mathbf{p}_j| \leq \beta^{-1} \lim_{k \rightarrow \infty} [f(\mathbf{x}_0) - f(\mathbf{x}_{k+1})] < \infty$$

and then

$$\lim_{j \rightarrow \infty} \alpha_j |\nabla f(\mathbf{x}_j) \mathbf{p}_j| = 0$$



Proof.

(3/3).

For  $k \in \mathcal{K}_1$  we have  $\alpha_k = \alpha_{\text{init}}$  and  
 $\alpha_k |\nabla f(\mathbf{x}_k) \mathbf{p}_k| = \alpha_{\text{init}} |\nabla f(\mathbf{x}_k) \mathbf{p}_k|$  and from (A) we have

$$\lim_{k \in \mathcal{K}_1 \rightarrow \infty} |\nabla f(\mathbf{x}_k) \mathbf{p}_k| = 0 \quad (*)$$

For  $k \in \mathcal{K}_2$  we have  $\tau \omega_k \leq \alpha_k \leq \omega_k$  so

$$\alpha_k |\nabla f(\mathbf{x}_k) \mathbf{p}_k| \geq \tau \omega_k |\nabla f(\mathbf{x}_k) \mathbf{p}_k| \geq 2\tau(1-\beta) \frac{|\nabla f(\mathbf{x}_k) \mathbf{p}_k|^2}{\gamma \|\mathbf{p}_k\|^2}$$

and from (B) we have

$$\lim_{k \in \mathcal{K}_2 \rightarrow \infty} \frac{|\nabla f(\mathbf{x}_k) \mathbf{p}_k|}{\|\mathbf{p}_k\|} = 0 \quad (**)$$

Combining (\*) and (\*\*) gives the required result.  $\square$



Proof.

(2/3).

Recall that

$$\alpha_k \geq \min \{ \alpha_{\text{init}}, \tau \omega_k \}, \quad \omega_k = 2(\beta - 1) \nabla f(\mathbf{x}_k) \mathbf{p}_k / (\gamma \|\mathbf{p}_k\|^2)$$

and consider the two index set:

$$\mathcal{K}_1 = \{k \mid \alpha_k = \alpha_{\text{init}}\}, \quad \mathcal{K}_2 = \{k \mid \alpha_k < \alpha_{\text{init}}\},$$

Obviously  $\mathbb{N} = \mathcal{K}_1 \cup \mathcal{K}_2$  and from  $\lim_{k \rightarrow \infty} \alpha_k |\nabla f(\mathbf{x}_k) \mathbf{p}_k| = 0$  we have

$$\lim_{k \in \mathcal{K}_1 \rightarrow \infty} \alpha_k |\nabla f(\mathbf{x}_k) \mathbf{p}_k| = 0, \quad (A)$$

$$\lim_{k \in \mathcal{K}_2 \rightarrow \infty} \alpha_k |\nabla f(\mathbf{x}_k) \mathbf{p}_k| = 0, \quad (B)$$



## Steepest descent algorithm

### Steepest descent algorithm

Given an initial guess  $\mathbf{x}_0$ , let  $k = 0$ ;

**while not converged do**

    Compute a step-size  $\alpha_k$  using a line-search along  $-\nabla f(\mathbf{x}_k)^T$ .

    Set  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)^T$  and increase  $k$  by 1.

**end while**

- The steepest descent algorithm is simply the **generic minimization algorithm** with search direction the opposite of the gradient in  $\mathbf{x}_k$ .
- The search direction  $-\nabla f(\mathbf{x}_k)^T$  is always a **descent direction** unless the point  $\mathbf{x}_k$  is a stationary point.



## Global convergence of steepest descent

## Corollary (Global convergence of steepest descent)

Suppose that  $f(x)$  satisfy the standard assumptions, then, for the iterates generated by the **steepest descent algorithm** with **backtracking Armijo line-search** either:

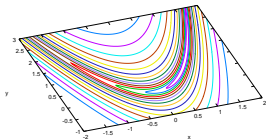
- 1  $\nabla f(x_k)^T = 0$  for some  $k \geq 0$ ;
- 2  $\lim_{k \rightarrow \infty} f(x_k) = -\infty$ ;
- 3  $\lim_{k \rightarrow \infty} \nabla f(x_k)^T = 0$ .



## The Rosenbrock example

(2/3)

- This function has a unique minimum at  $(1, 1)^T$  inside a **banana shaped** valley.

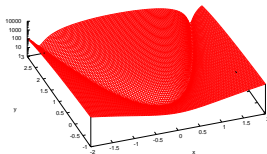


## The Rosenbrock example

(1/3)

- Although the **steepest descent** scheme is globally convergent it can be very slow!
- A classical example is the Rosenbrock function:

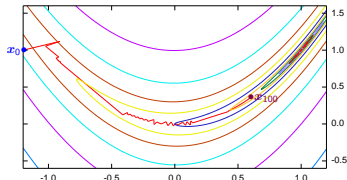
$$f(x, y) = 100(y - x^2)^2 + (x - 1)^2$$



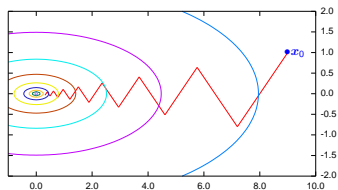
## The Rosenbrock example

(3/3)

- After 100 iteration starting from  $(-1.2, 1)^T$  the approximate minimum is **far** from the solution.



- The steepest descent is a slow method, not only on a difficult test case like the Rosenbrock example.
- Given the function  $f(x, y) = \frac{1}{2}x^2 + \frac{9}{2}y^2$  starting from  $x_0 = (9, 1)^T$  we have the zig-zag pattern toward  $(0, 0)^T$ .



## Outline

- General iterative scheme
  - Descent direction failure
- Backtracking Armijo line-search
  - Global convergence of backtracking Armijo line-search
  - Global convergence of steepest descent
- Wolfe-Zoutendijk global convergence
  - The Wolfe conditions
  - The Armijo-Goldstein conditions
- Algorithms for line-search
  - Armijo Parabolic-Cubic search
  - Wolfe linesearch

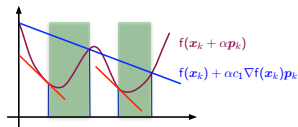
## The Wolfe and Armijo Goldstein conditions

- The simple condition of **descent step** is in general not enough for the convergence of an iterative minimization scheme.
- The condition of **sufficient decrease** of backtracking Armijo line-search may be insufficient on general inexact line-search algorithm.
- Adding another condition to the **sufficient decrease** condition such that we avoid **too short** step length we obtain **globally convergent** numerical procedure.
- Depending on which additional condition is added we obtain the:
  - Wolfe conditions;
  - Armijo Goldstein conditions.

## The Wolfe conditions

Let  $c_1$  and  $c_2$  two constant such that  $0 < c_1 < c_2 < 1$ . We say that the step length  $\alpha_k$  satisfy the Wolfe conditions if  $\alpha_k$  satisfy:

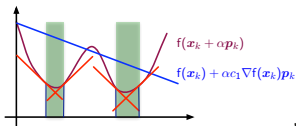
- sufficient decrease:**  $f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k) p_k$ ;
- curvature condition:**  $\nabla f(x_k + \alpha_k p_k) p_k \geq c_2 \nabla f(x_k) p_k$ .



### The strong Wolfe conditions

Let  $c_1$  and  $c_2$  two constant such that  $0 < c_1 < c_2 < 1$ . We say that the step length  $\alpha_k$  satisfy the strong Wolfe conditions if  $\alpha_k$  satisfy:

- 1 **sufficient decrease**:  $f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + c_1 \alpha_k \nabla f(\mathbf{x}_k) \mathbf{p}_k$ ;
- 2 **curvature condition**:  $|\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \mathbf{p}_k| \leq c_2 |\nabla f(\mathbf{x}_k) \mathbf{p}_k|$ .



### Existence of "Wolfe" step length

- The Wolfe condition seems quite restrictive.
- The next lemma answer to the question if a step length satisfying Wolfe conditions does exists.

#### Lemma (strong Wolfe step length)

Let  $f : \mathbb{R}^n \mapsto \mathbb{R}$  satisfying the regularity assumption. If the following condition are satisfied:

- 1  $\mathbf{p}_k$  is a descent direction for the point  $\mathbf{x}_k$ , i.e.  $\nabla f(\mathbf{x}_k) \mathbf{p}_k < 0$ ;
- 2  $f(\mathbf{x}_k + \alpha \mathbf{p}_k)$  is bounded from below, i.e.  $\lim_{\alpha \rightarrow \infty} f(\mathbf{x}_k + \alpha \mathbf{p}_k) > -\infty$ .

then for any  $0 < c_1 < c_2 < 1$  there exists an interval  $[a, b]$  such that all  $\alpha_k \in [a, b]$  satisfy the strong Wolfe conditions.



### Proof.

Define  $\ell(\alpha) = f(\mathbf{x}_k) + \alpha c_1 \nabla f(\mathbf{x}_k) \mathbf{p}_k$  and  $g(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{p}_k)$ . From  $\lim_{\alpha \rightarrow \infty} \ell(\alpha) = -\infty$  and from condition 1 it follows that there exists  $\alpha_* > 0$  such that

$$\ell(\alpha_*) = g(\alpha_*) \quad \text{and} \quad \ell(\alpha) > g(\alpha), \quad \forall \alpha \in (0, \alpha_*)$$

so that all step length  $\alpha \in (0, \alpha_*)$  satisfy strong Wolfe condition 1. Because  $\ell(0) = g(0)$  from Cauchy-Rolle theorem there exists  $\alpha_{**} \in (0, \alpha_*)$  such that

$$g'(\alpha_{**}) = \ell'(\alpha_{**}) \Rightarrow$$

$$0 > \nabla f(\mathbf{x}_k + \alpha_{**} \mathbf{p}_k) \mathbf{p}_k = c_1 \nabla f(\mathbf{x}_k) \mathbf{p}_k > c_2 \nabla f(\mathbf{x}_k) \mathbf{p}_k$$

by continuity we find an interval around  $\alpha_{**}$  with step lengths satisfying strong Wolfe conditions.  $\square$



### The Zoutendijk condition

#### Theorem (Zoutendijk)

Let  $f : \mathbb{R}^n \mapsto \mathbb{R}$  satisfying the regularity assumption and bounded from below, i.e.

$$\inf_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) > -\infty$$

Let  $\{\mathbf{x}_k\}$ ,  $k = 0, 1, \dots, \infty$  generated by a **generic minimization algorithm** where line-search satisfy **Wolfe conditions**, then

$$\sum_{k=1}^{\infty} (\cos \theta_k)^2 \|\nabla f(\mathbf{x}_k)\|^2 < +\infty$$

where

$$\cos \theta_k = \frac{-\nabla f(\mathbf{x}_k) \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\|}$$



Proof.

(1/3).

Using the second condition of Wolfe

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \mathbf{p}_k \geq c_2 \nabla f(\mathbf{x}_k) \mathbf{p}_k$$

$$(\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) - \nabla f(\mathbf{x}_k)) \mathbf{p}_k \geq (c_2 - 1) \nabla f(\mathbf{x}_k) \mathbf{p}_k$$

by using Lipschitz regularity

$$\begin{aligned} \|\nabla f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) - \nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\| &\leq \gamma \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \|\mathbf{p}_k\| \\ &= \alpha_k \gamma \|\mathbf{p}_k\|^2 \end{aligned}$$

and using both inequality we obtain the estimate for  $\alpha_k$ :

$$\alpha_k \geq \frac{c_2 - 1}{\gamma \|\mathbf{p}_k\|^2} \nabla f(\mathbf{x}_k) \mathbf{p}_k$$



Proof.

(2/3).

Using the first condition of Wolfe and estimate of  $\alpha_k$ 

$$\begin{aligned} f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) &\leq f(\mathbf{x}_k) + \alpha_k c_1 \nabla f(\mathbf{x}_k) \mathbf{p}_k \\ &\leq f(\mathbf{x}_k) - \frac{c_1(1-c_2)}{\gamma \|\mathbf{p}_k\|^2} (\nabla f(\mathbf{x}_k) \mathbf{p}_k)^2 \end{aligned}$$

setting  $A = c_1(1-c_2)/\gamma$  and using the definition of  $\cos \theta_k$ 

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) - A (\cos \theta_k)^2 \|\nabla f(\mathbf{x}_k)\|^2$$

and by induction

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_1) - A \sum_{j=1}^k (\cos \theta_j)^2 \|\nabla f(\mathbf{x}_j)\|^2$$



Proof.

(3/3).

The function  $f(\mathbf{x})$  is bounded from below, i.e.

$$\inf_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) > -\infty$$

so that

$$A \sum_{j=1}^k (\cos \theta_j)^2 \|\nabla f(\mathbf{x}_j)\|^2 \leq f(\mathbf{x}_1) - f(\mathbf{x}_{k+1})$$

and

$$A \sum_{j=1}^{\infty} (\cos \theta_j)^2 \|\nabla f(\mathbf{x}_j)\|^2 \leq f(\mathbf{x}_1) - \lim_{k \rightarrow \infty} f(\mathbf{x}_{k+1}) < +\infty$$

□



Corollary (Zoutendijk condition)

Let  $f: \mathbb{R}^n \mapsto \mathbb{R}$  satisfying the regularity assumption and bounded from below. Let  $\{\mathbf{x}_k\}$ ,  $k = 0, 1, \dots, \infty$  generated by a generic minimization algorithm where line-search satisfy **Wolfe conditions**, then

$$\cos \theta_k \|\nabla f(\mathbf{x}_k)\|^T \rightarrow 0 \quad \text{where} \quad \cos \theta_k = \frac{-\nabla f(\mathbf{x}_k) \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\|^T \|\mathbf{p}_k\|}$$

Remark

If  $\cos \theta_k \geq \delta > 0$  for all  $k$  from the Zoutendijk condition we have:

$$\|\nabla f(\mathbf{x}_k)\|^T \rightarrow 0$$

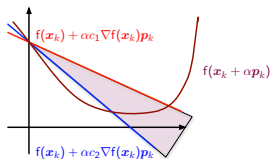
i.e. the **generic minimization algorithm** where line-search satisfy **Wolfe conditions** converge to a stationary point.



### The Armijo-Goldstein conditions

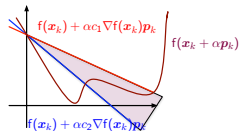
Let  $c_1$  and  $c_2$  two constant such that  $0 < c_1 < c_2 < 1$ . We say that the step length  $\alpha_k$  satisfy the Wolfe conditions if  $\alpha_k$  satisfy:

- 1  $f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f(x_k) p_k$ ;
- 2  $f(x_k + \alpha_k p_k) \geq f(x_k) + c_2 \alpha_k \nabla f(x_k) p_k$ ;



### The Armijo-Goldstein conditions

- 1 Armijo-Goldstein conditions has very similar theoretical properties like the Wolfe conditions.
- 2 Global convergence theorems can be established.
- 3 The weakness of Armijo-Goldstein conditions respect to Wolfe conditions is that the former can exclude **local minima's** from the step length as you can see in the figure below.



### Outline

- 1 General iterative scheme
  - Descent direction failure
- 2 Backtracking Armijo line-search
  - Global convergence of backtracking Armijo line-search
  - Global convergence of steepest descent
- 3 Wolfe-Zoutendijk global convergence
  - The Wolfe conditions
  - The Armijo-Goldstein conditions
- 4 Algorithms for line-search
  - Armijo Parabolic-Cubic search
  - Wolfe linesearch

### Armijo Parabolic-Cubic search

- 1 Backtracking-Armijo line-search can be slow if a large number of reduction must be performed to satisfy Armijo condition.
- 2 A better performance is obtained if instead of reducing by a fixed factor we use polynomial interpolation to estimate the location of the minimum.
- 3 Assuming that that  $f(x_k)$  and  $\nabla f(x_k) p_k$  are known at the first step we know also  $f(x_k + \lambda p_k)$  if  $\lambda$  is the first trial step.
- 4 In this case a parabolic interpolation can be used to estimate the minimum.
- 5 If we store the last trial step length, in the successive iteration we can use cubic interpolation to estimate the minima's.
- 6 The resulting algorithm is in the following slides.

## Algorithm (Armijo Parabolic-Cubic search) (1/3)

```

armijo_linesearch(f, x, p, τ)
f0 ← f(x); ∇f0 ← ∇f(x)p; λ ← 1;
while λ ≥ λmin do
  fλ ← f(x + λp);
  if fλ ≤ f0 + λτ∇f0 then
    return λ ; successful search
  else
    if λ = 1 then
      λtmp ← ∇f0 / [2(f0 + ∇f0 - fλ)];
    else
      λtmp ← cubic(f0, ∇f0, fλ, λ, fp, λp);
    end if
    λp ← λ; fp ← fλ; λ ← range(λtmp, λ/10, λ/2);
  end if
end while
return λmin ; failed search

```



## Algorithm (Armijo Parabolic-Cubic search) (2/3)

```

range(λ, a, b)
if λ < a then
  return a;
else if λ > b then
  return b;
else
  return λ ;
end if

```



## Algorithm (Armijo Parabolic-Cubic search) (3/3)

*cubic*(f<sub>0</sub>, ∇f<sub>0</sub>, f<sub>λ</sub>, λ, f<sub>p</sub>, λ<sub>p</sub>)

Evaluate:

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{\lambda^2 \lambda_p^2 (\lambda - \lambda_p)} \begin{pmatrix} \lambda_p^2 & -\lambda^2 \\ -\lambda_p^3 & \lambda^3 \end{pmatrix} \begin{pmatrix} f_\lambda - f_0 - \lambda \nabla f_0 \\ f_p - f_0 - \lambda_p \nabla f_0 \end{pmatrix}$$

if a = 0 then

return -∇f<sub>0</sub>/(2b);

*cubic is a quadratic*

else

d ← b<sup>2</sup> - 3a ∇f<sub>0</sub>;

*discriminant*

return (-b + √d)/(3a);

*legitimate cubic*

end if



## Wolfe linesearch

- Wolfe linesearch is identical to the Armijo Parabolic-Cubic search, until a point satisfying the first condition is found.
- At this point the Armijo algorithm stop while Wolfe search try to refine the search until the second condition is satisfied.
- If the step estimated is too short then is is enlarged until it contains a minimum.
- If the step estimated is too long it is reduced until the second condition is satisfied.



## Algorithm (Wolfe line-search) (1/3)

```

wolfe_line_search(f, x, p, c1, c2)
f0 ← f(x); ∇f0 ← ∇f(x)p; λ ← 1;
while λ ≥ λ_min do
  f_λ ← f(x + λp);
  if f_λ ≤ f0 + λc1∇f0 then
    go to ZOOM; found a λ satisfying condition 1
  else
    if λ = 1 then
      λ_tmp ← ∇f0 / [2(f0 + ∇f0 - f_λ)];
    else
      λ_tmp ← cubic(f0, ∇f0, f_λ, λ, f_p, λ_p);
    end if
    λ_p ← λ; f_p ← f_λ; λ ← range(λ_tmp, λ/10, λ/2);
  end if
end while
return λ_min ; failed search

```



## Algorithm (Wolfe line-search) (2/3)

```

ZOOM:
∇f_λ ← ∇f(x + λp)p;
if ∇f_λ ≥ c2∇f0 then return λ;          found Wolfe point!
if λ = 1 then
  forward search of an interval bracketing a minimum
  while λ ≤ λ_max do
    {λ_p, f_p} ← {λ, f_λ};              save values
    λ ← 2λ; f_λ ← f(x + λp);
    if not f_λ ≤ f0 + λc1∇f0 then
      {λ_p, f_p} = {λ, f_λ}; go to REFINE;  swap values
    end if
    ∇f_λ ← ∇f(x + λp)p;
    if ∇f_λ ≥ c2∇f0 then return λ;      found Wolfe point!
  end while
  return λ_max ; failed search
end if

```



## Algorithm (Wolfe line-search) (3/3)



```

REFINE:
{λ_lo, f_lo, ∇f_lo} ← {λ, f_λ, ∇f_λ}; Δ ← λ_p - λ_lo;
while Δ > ε do
  δλ ← Δ^2 ∇f_lo / [2(f_lo + ∇f_lo Δ - f_p)];
  δλ ← range(δλ, 0.2Δ, 0.8Δ);
  λ ← λ_lo + δλ; f_λ ← f(x + λp);
  if f_λ ≤ f0 + λc1∇f0 then
    ∇f_λ ← ∇f(x + λp)p;
    if ∇f_λ ≥ c2∇f0 then return λ;      found Wolfe point!
    {λ_lo, f_lo, ∇f_lo} ← {λ, f_λ, ∇f_λ}; Δ ← Δ - δλ;
  else
    {λ_p, f_p} ← {λ, f_λ}; Δ ← δλ;
  end if
end while
return λ; failed search

```



## References

-  J. Stoer and R. Bulirsch  
Introduction to numerical analysis  
Springer-Verlag, Texts in Applied Mathematics, 12, 2002.
-  J. E. Dennis, Jr. and Robert B. Schnabel  
Numerical Methods for Unconstrained Optimization and  
Nonlinear Equations  
SIAM, Classics in Applied Mathematics, 16, 1996.

