Unconstrained minimization

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

Enrico Bertolazzi

DIMS – Università di Trento

March 2005

Outline

The problem (1/3)

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

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

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

Assumption (Regularity assumption)

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

$$\left\|\nabla\mathsf{f}(\boldsymbol{x})^T - \nabla\mathsf{f}(\boldsymbol{y})^T\right\| \leq \gamma \left\|\boldsymbol{x}-\boldsymbol{y}\right\|, \qquad \forall \boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$$

The problem (2/3)

Definition (Global minimum)

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

$$\mathsf{f}(\boldsymbol{x}_\star) \leq \mathsf{f}(\boldsymbol{x}), \qquad \forall \boldsymbol{x}\in\mathbb{R}^n.$$

Definition (Local minimum)

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

$$\mathsf{f}(\boldsymbol{x}_\star) \leq \mathsf{f}(\boldsymbol{x}), \qquad \forall \boldsymbol{x}\in B(\boldsymbol{x}_\star;\delta).$$

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.

The problem

Definition (Strict global minimum)

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

$$\mathsf{f}(\boldsymbol{x}_\star) < \mathsf{f}(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{x}_\star\}.$$

Definition (Strict local minimum)

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

$$\mathsf{f}(\boldsymbol{x}_\star) < \mathsf{f}(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in B(\boldsymbol{x}_\star; \delta) \setminus \{\boldsymbol{x}_\star\}.$$

Obviously a strict global minimum is a strict local minimum.

First order Necessary condition

Lemma (First order Necessary condition for local minimum)

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

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

Proof.

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

$$\lambda^{-1}\big(\mathsf{f}(\boldsymbol{x}_\star + \lambda \boldsymbol{d}) - \mathsf{f}(\boldsymbol{x}_\star)\big) \leq 0, \qquad 0 < \lambda < \delta$$

so that

$$\lim_{\lambda \to 0} \lambda^{-1}\big(\mathsf{f}(\boldsymbol{x}_\star + \lambda \boldsymbol{d}) - \mathsf{f}(\boldsymbol{x}_\star)\big) = \nabla \mathsf{f}(\boldsymbol{x}_\star)\boldsymbol{d} \leq 0,$$

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

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 $\mathsf{f}(\boldsymbol{x})$.
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 $\boldsymbol{x}_\star$ it is not possible to deduce a necessary and sufficient condition for a minima.

Second order Necessary condition

Lemma (Second order Necessary condition for local minimum)

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

$$\boldsymbol{d}^T \nabla^2 \mathsf{f}(\boldsymbol{x}_\star)\boldsymbol{d} \geq 0, \qquad \forall \boldsymbol{d} \in \mathbb{R}^n$$

Example

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

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

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

Proof.

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

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

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

$$f(\boldsymbol{x}_\star \pm \lambda\boldsymbol{d}) = f(\boldsymbol{x}_\star) \pm \nabla f(\boldsymbol{x}_\star)\lambda\boldsymbol{d} + \frac{\lambda^2}{2}\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{d} + o(\lambda^2)$$

and from the previous inequality

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

taking the limit $\lambda \to 0$ and form the arbitrariness of $\boldsymbol{d}$ we have that $\nabla^2 f(\boldsymbol{x}_\star)$ must be semi-definite positive. □

Second order sufficient condition

Lemma (Second order sufficient condition for local minimum)

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

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

$$\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x}_\star)\boldsymbol{d} > 0, \qquad \forall\boldsymbol{d} \in \mathbb{R}^n \setminus \{\boldsymbol{x}_\star\}$$

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

Remark

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

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

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

Proof.

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

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

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

$$f(\boldsymbol{x}_\star + \boldsymbol{d}) \geq f(\boldsymbol{x}_\star) + \frac{1}{4}\lambda_{min}\|\boldsymbol{d}\|^2 > f(\boldsymbol{x}_\star), \qquad \boldsymbol{d} \neq \mathbf{0},\ \|\boldsymbol{d}\| \leq \delta.$$

i.e. $\boldsymbol{x}_\star$ is a strict minimum. □

Outline

How to find a minimum

Given $f : \mathbb{R}^n \mapsto \mathbb{R}$: $\quad \text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$.

1. We can solve the problem by solving the necessary condition. i.e by solving the nonlinear systems
$$\nabla f(\boldsymbol{x})^T = \mathbf{0}.$$
2. Using such an approach we looses the information about $f(\boldsymbol{x})$.
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 $\boldsymbol{x}_0$ build a sequence $\{\boldsymbol{x}_k\}$ such that $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$. In this way, at least, we avoid to converge to a strict maximum.

Iterative Methods

- in practice very rare to be able to provide explicit minimizer.
- iterative method: given starting guess $\boldsymbol{x}_0$, generate the sequence,
$$\{\boldsymbol{x}_k\}, \qquad k = 1, 2, \ldots$$
- AIM: ensure that (a subsequence) has some favorable limiting properties:
  - satisfies first-order necessary conditions
  - satisfies second-order necessary conditions

Line-search Methods

A generic iterative minimization procedure can be sketched as follows:

- calculate a search direction $\boldsymbol{p}_k$ from $\boldsymbol{x}_k$
- ensure that this direction is a descent direction, i.e.
$$\nabla f(\boldsymbol{x}_k)\boldsymbol{p}_k < 0, \qquad \text{whenever } \nabla f(\boldsymbol{x}_k)^T \neq \mathbf{0}$$
so that, at least for small steps along $\boldsymbol{p}_k$, the objective function $f(\boldsymbol{x})$ will be reduced
- use line-search to calculate a suitable step-length $\alpha_k > 0$ so that
$$f(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) < f(\boldsymbol{x}_k).$$
- Update the point:
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$$

Generic minimization algorithm

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

Generic minimization algorithm

Given an initial guess $\boldsymbol{x}_0$, let $k = 0$;
while not converged **do**
  Find a descent direction $\boldsymbol{p}_k$ at $\boldsymbol{x}_k$;
  Compute a step size α_k using a line-search along $\boldsymbol{p}_k$.
  Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$ and increase k by 1.
end while

The crucial points which differentiate the algorithms are:

1. The computation of the direction $\boldsymbol{p}_k$;
2. The computation of the step size α_k.

Practical Line-search methods

- The first developed minimization algorithms try to solve

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

  - 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 α_{init} (e.g., $\alpha_{\text{init}} = 1$);
Given $\tau \in (0, 1)$ typically $\tau = 0.5$;
Let $\alpha^{(0)} = \alpha_{\text{init}}$;
while not $f(\boldsymbol{x}_k + \alpha^{(\ell)}\boldsymbol{p}_k) < f(\boldsymbol{x}_k)$ **do**
  set $\alpha^{(\ell+1)} = \tau\alpha^{(\ell)}$;
  increase ℓ 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 $\boldsymbol{p}_k$ is a descent direction then the algorithm terminate.

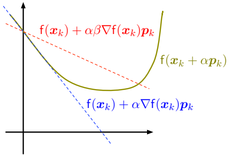
Outline

Armijo condition

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

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

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



Backtracking Armijo line-search

Given α_{init} (e.g., $\alpha_{\text{init}} = 1$);
Given $\tau \in (0,1)$ typically $\tau = 0.5$;
Let $\alpha^{(0)} = \alpha_{\text{init}}$;
while not $\mathsf{f}(\boldsymbol{x}_k + \alpha^{(\ell)}\boldsymbol{p}_k) \leq \mathsf{f}(\boldsymbol{x}_k) + \alpha^{(\ell)}\beta\nabla\mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$ **do**
  set $\alpha^{(\ell+1)} = \tau\alpha^{(\ell)}$;
  increase ℓ 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 $\mathsf{f}(\boldsymbol{x})$ *satisfy the standard assumptions and* $\beta \in (0,1)$ *and that* $\boldsymbol{p}_k$ *is a descent direction at* $\boldsymbol{x}_k$. *Then the Armijo condition*

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

is satisfied for all $\alpha_k \in [0, \omega_k]$ *where* $\omega_k = \dfrac{2(\beta-1)\nabla\mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k}{\gamma\,\|\boldsymbol{p}_k\|^2}$

Assumption (Regularity assumption)

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

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

Finite termination of Armijo line-search

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

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

Proof.

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

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

□

Finite termination of Armijo line-search

Corollary (Finite termination of Armijo linesearch)

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

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

Proof.

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

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

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

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

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

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

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

Remark

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

Proof. (1/3).

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

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

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

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

and then

$$\lim_{j\to\infty} \alpha_j |\nabla \mathsf{f}(\boldsymbol{x}_j)\boldsymbol{p}_j| = 0$$

Proof. (2/3).

Recall that

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

and consider the two index set:

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

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

$$\lim_{k\in\mathcal{K}_1\to\infty} \alpha_k |\nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k| = 0, \tag{A}$$

$$\lim_{k\in\mathcal{K}_2\to\infty} \alpha_k |\nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k| = 0, \tag{B}$$

Proof. (3/3).

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

$$\lim_{k\in\mathcal{K}_1\to\infty} \left|\nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k\right| = 0 \qquad (\star)$$

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

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

and from (B) we have

$$\lim_{k\in\mathcal{K}_1\to\infty} \frac{\left|\nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k\right|}{\left\|\boldsymbol{p}_k\right\|} = 0 \qquad (\star\star)$$

Combining $(\star)$ and $(\star\star)$ gives the required result. □

Steepest descent algorithm

Steepest descent algorithm

Given an initial guess $\boldsymbol{x}_0$, let $k = 0$;
while not converged **do**
  Compute a step-size α_k using a line-search along $-\nabla \mathsf{f}(\boldsymbol{x}_k)^T$.
  Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla \mathsf{f}(\boldsymbol{x}_k)^T$ 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 $\boldsymbol{x}_k$.
- The search direction $-\nabla \mathsf{f}(\boldsymbol{x}_k)^T$ is always a descent direction unless the point $\boldsymbol{x}_k$ is a stationary point.

Global convergence of steepest descent

Corollary (Global convergence of steepest descent)

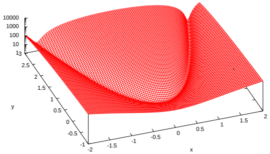
Suppose that $\mathsf{f}(\boldsymbol{x})$ satisfy the standard assumptions, then, for the iterates generated by the steepest descent algorithm with backtracking Armijo line-search either:

1. $\nabla \mathsf{f}(\boldsymbol{x}_k)^T = \mathbf{0}$ *for some* $k \geq 0$;
2. *or* $\lim_{k\to\infty} \mathsf{f}(\boldsymbol{x}_k) = -\infty$;
3. *or* $\lim_{k\to\infty} \nabla \mathsf{f}(\boldsymbol{x}_k)^T = \mathbf{0}$.

The Rosenbrock example (1/3)

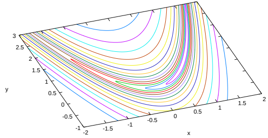
- Although the steepest descent scheme is globally convergent it can be very slow!
- A classical example is the Rosenbrock function:

$$\mathsf{f}(x,y) = 100\,(y-x^2)^2 + (x-1)^2$$



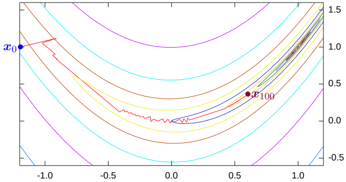
The Rosenbrock example (2/3)

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

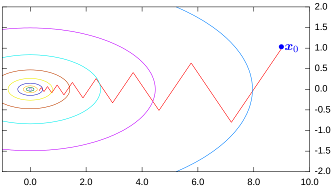


The Rosenbrock example (3/3)

- 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 $\mathsf{f}(x,y)=\dfrac{1}{2}x^2+\dfrac{9}{2}y^2$ starting from $\boldsymbol{x}_0=(9,1)^T$ we have the zig-zag pattern toward $(0,0)^T$.



Outline

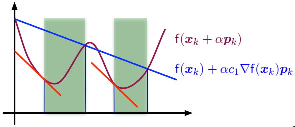
The Wolfe and Armijo Goldstein conditions

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

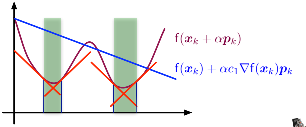
1. sufficient decrease: $\mathsf{f}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \leq \mathsf{f}(\boldsymbol{x}_k) + c_1\, \alpha_k \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$;
2. curvature condition: $\nabla \mathsf{f}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k)\boldsymbol{p}_k \geq c_2\, \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$.



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 α_k satisfy the strong Wolfe conditions if α_k satisfy:

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



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

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

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

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

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

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

$$0 > \nabla \mathsf{f}(\boldsymbol{x}_k + \alpha_{\star\star}\boldsymbol{p}_k)\boldsymbol{p}_k = c_1 \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k > c_2 \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$$

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

The Zoutendijk condition

Theorem (Zoutendijk)

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

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

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

$$\sum_{k=1}^{\infty} (\cos\theta_k)^2 \left\|\nabla \mathsf{f}(\boldsymbol{x}_k)^T\right\|^2 < +\infty$$

where

$$\cos\theta_k = \frac{-\nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k}{\|\nabla \mathsf{f}(\boldsymbol{x}_k)^T\| \, \|\boldsymbol{p}_k\|}$$

Proof. (1/3).

Using the second condition of Wolfe

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

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

by using Lipschitz regularity

$$\left\|\big(\nabla \mathsf{f}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) - \nabla \mathsf{f}(\boldsymbol{x}_k)\big)\boldsymbol{p}_k\right\| \leq \gamma \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \, \|\boldsymbol{p}_k\|$$

$$= \alpha_k \gamma \|\boldsymbol{p}_k\|^2$$

and using both inequality we obtain the estimate for α_k:

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

Proof. (2/3).

Using the first condition of Wolfe and estimate of α_k

$$\mathsf{f}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \leq \mathsf{f}(\boldsymbol{x}_k) + \alpha_k c_1 \nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$$

$$\leq \mathsf{f}(\boldsymbol{x}_k) - \frac{c_1(1 - c_2)}{\gamma \|\boldsymbol{p}_k\|^2} \big(\nabla \mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k\big)^2$$

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

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

and by induction

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

Proof. (3/3).

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

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

so that

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

and

$$A\sum_{j=1}^{\infty}(\cos\theta_j)^2\left\|\nabla\mathsf{f}(\boldsymbol{x}_j)^T\right\|^2 \leq \mathsf{f}(\boldsymbol{x}_1) - \lim_{k\to\infty}\mathsf{f}(\boldsymbol{x}_{k+1}) < +\infty$$

□

Corollary (Zoutendijk condition)

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

$$\cos\theta_k\left\|\nabla\mathsf{f}(\boldsymbol{x}_k)^T\right\| \to 0 \qquad \textit{where} \qquad \cos\theta_k = \frac{-\nabla\mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k}{\left\|\nabla\mathsf{f}(\boldsymbol{x}_k)^T\right\|\left\|\boldsymbol{p}_k\right\|}$$

Remark

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

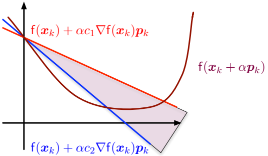
$$\left\|\nabla\mathsf{f}(\boldsymbol{x}_k)^T\right\| \to 0$$

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

The Armijo-Goldstein conditions

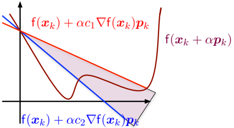
Let c_1 and c_2 two constant such that $0 < c_1 < c_2 < 1$. We say that the step length α_k satisfy the Wolfe conditions if α_k satisfy:

1. $\mathsf{f}(\boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k) \leq \mathsf{f}(\boldsymbol{x}_k) + c_1\,\alpha_k\nabla\mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$;
2. $\mathsf{f}(\boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k) \geq \mathsf{f}(\boldsymbol{x}_k) + c_2\,\alpha_k\nabla\mathsf{f}(\boldsymbol{x}_k)\boldsymbol{p}_k$;



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

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 $f(\boldsymbol{x}_k)$ and $\nabla f(\boldsymbol{x}_k)\boldsymbol{p}_k$ are known at the first step we know also $f(\boldsymbol{x}_k + \lambda \boldsymbol{p}_k)$ if λ 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$(f, \boldsymbol{x}, \boldsymbol{p}, \tau)$
$f_0 \leftarrow f(\boldsymbol{x})$; $\nabla f_0 \leftarrow \nabla f(\boldsymbol{x})\boldsymbol{p}$; $\lambda \leftarrow 1$;
while $\lambda \geq \lambda_{\min}$ **do**
  $f_\lambda \leftarrow f(\boldsymbol{x} + \lambda \boldsymbol{p})$;
  if $f_\lambda \leq f_0 + \lambda\tau\nabla f_0$ **then**
    return λ ; *successful search*
  else
    if $\lambda = 1$ **then**
      $\lambda_{tmp} \leftarrow \nabla f_0 / \big[2(f_0 + \nabla f_0 - f_\lambda)\big]$;
    else
      $\lambda_{tmp} \leftarrow$ *cubic*$(f_0, \nabla f_0, f_\lambda, \lambda, f_p, \lambda_p)$;
    end if
    $\lambda_p \leftarrow \lambda$; $f_p \leftarrow f_\lambda$; $\lambda \leftarrow$ *range*$(\lambda_{tmp}, \lambda/10, \lambda/2)$;
  end if
end while
return $\lambda_{\min}$; *failed search*

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

range(λ, a, b)
if $\lambda < a$ **then**
  return a;
else if $\lambda > b$ **then**
  return b;
else
  return λ ;
end if

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

cubic$(\mathsf{f}_0, \nabla\mathsf{f}_0, \mathsf{f}_\lambda, \lambda, \mathsf{f}_p, \lambda_p)$
Evaluate:

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{\lambda^2\lambda_p^2(\lambda-\lambda_p)}\begin{pmatrix} \lambda_p^2 & -\lambda^2 \\ -\lambda_p^3 & \lambda^3 \end{pmatrix}\begin{pmatrix} \mathsf{f}_\lambda - \mathsf{f}_0 - \lambda\nabla\mathsf{f}_0 \\ \mathsf{f}_p - \mathsf{f}_0 - \lambda_p\nabla\mathsf{f}_0 \end{pmatrix}$$

if $a = 0$ **then**
  return $-\nabla\mathsf{f}_0/(2b)$; *cubic is a quadratic*
else
  $d \leftarrow b^2 - 3\,a\,\nabla\mathsf{f}_0$; *discriminant*
  return $(-b+\sqrt{d})/(3a)$; *legitimate cubic*
end if

Wolfe linesearch

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

Algorithm (Wolfe linesearch (1/3))

wolfe_linesearch$(\mathsf{f}, \boldsymbol{x}, \boldsymbol{p}, c_1, c_2)$
$\mathsf{f}_0 \leftarrow \mathsf{f}(\boldsymbol{x})$; $\nabla\mathsf{f}_0 \leftarrow \nabla\mathsf{f}(\boldsymbol{x})\boldsymbol{p}$; $\lambda \leftarrow 1$;
while $\lambda \geq \lambda_{\min}$ **do**
  $\mathsf{f}_\lambda \leftarrow \mathsf{f}(\boldsymbol{x}+\lambda\boldsymbol{p})$;
  if $\mathsf{f}_\lambda \leq \mathsf{f}_0 + \lambda c_1 \nabla\mathsf{f}_0$ **then**
    go to *ZOOM*; *found a λ satisfying condition 1*
  else
    if $\lambda = 1$ **then**
      $\lambda_{tmp} \leftarrow \nabla\mathsf{f}_0 / \big[2(\mathsf{f}_0 + \nabla\mathsf{f}_0 - \mathsf{f}_\lambda)\big]$;
    else
      $\lambda_{tmp} \leftarrow$ *cubic*$(\mathsf{f}_0, \nabla\mathsf{f}_0, \mathsf{f}_\lambda, \lambda, \mathsf{f}_p, \lambda_p)$;
    end if
    $\lambda_p \leftarrow \lambda$; $\mathsf{f}_p \leftarrow \mathsf{f}_\lambda$; $\lambda \leftarrow$ *range*$(\lambda_{tmp}, \lambda/10, \lambda/2)$;
  end if
end while
return $\lambda_{\min}$; *failed search*

Algorithm (Wolfe linesearch (2/3))

ZOOM:
$\nabla\mathsf{f}_\lambda \leftarrow \nabla\mathsf{f}(\boldsymbol{x}+\lambda\boldsymbol{p})\boldsymbol{p}$;
if $\nabla\mathsf{f}_\lambda \geq c_2\nabla\mathsf{f}_0$ **then return** λ; *found Wolfe point!*
if $\lambda = 1$ **then**
  forward search of an interval bracketing a minimum
  while $\lambda \leq \lambda_{\max}$ **do**
    $\{\lambda_p, \mathsf{f}_p\} \leftarrow \{\lambda, \mathsf{f}_\lambda\}$; *save values*
    $\lambda \leftarrow 2\lambda$; $\mathsf{f}_\lambda \leftarrow \mathsf{f}(\boldsymbol{x}+\lambda\boldsymbol{p})$;
    if *not* $\mathsf{f}_\lambda \leq \mathsf{f}_0 + \lambda c_1\nabla\mathsf{f}_0$ **then**
      $\{\lambda_p, \mathsf{f}_p\} \rightleftharpoons \{\lambda, \mathsf{f}_\lambda\}$; **go to** *REFINE*; *swap values*
    end if
    $\nabla\mathsf{f}_\lambda \leftarrow \nabla\mathsf{f}(\boldsymbol{x}+\lambda\boldsymbol{p})\boldsymbol{p}$;
    if $\nabla\mathsf{f}_\lambda \geq c_2\nabla\mathsf{f}_0$ **then return** λ; *found Wolfe point!*
  end while
  return $\lambda_{\max}$; *failed search*
end if

Algorithm (Wolfe linesearch (3/3))

REFINE:
$\{\lambda_{lo}, f_{lo}, \nabla f_{lo}\} \leftarrow \{\lambda, f_\lambda, \nabla f_\lambda\}$; $\Delta \leftarrow \lambda_p - \lambda_{lo}$;
while $\Delta > \epsilon$ **do**
  $\delta\lambda \leftarrow \Delta^2 \nabla f_{lo} / \big[2(f_{lo} + \nabla f_{lo}\Delta - f_p)\big]$;
  $\delta\lambda \leftarrow$ *range*$(\delta\lambda, 0.2\Delta, 0.8\Delta)$;
  $\lambda \leftarrow \lambda_{lo} + \delta\lambda$; $f_\lambda \leftarrow f(\boldsymbol{x} + \lambda\boldsymbol{p})$;
  if $f_\lambda \leq f_0 + \lambda c_1 \nabla f_0$ **then**
    $\nabla f_\lambda \leftarrow \nabla f(\boldsymbol{x} + \lambda\boldsymbol{p})\boldsymbol{p}$;
    if $\nabla f_\lambda \geq c_2 \nabla f_0$ **then return** λ; *found Wolfe point!*
    $\{\lambda_{lo}, f_{lo}, \nabla f_{lo}\} \leftarrow \{\lambda, f_\lambda, \nabla f_\lambda\}$; $\Delta \leftarrow \Delta - \delta\lambda$;
  else
    $\{\lambda_p, f_p\} \leftarrow \{\lambda, f_\lambda\}$; $\Delta \leftarrow \delta\lambda$;
  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.