Choose correct coefficients before summing

$$\frac{8f(x+h) - 8f(x-h) - f(x+2h) + f(x-2h)}{12h} =$$

$$f'(x) + o(h^3)$$

Numerical approximation of gradient

to approximate $\nabla f(x)$ $f : \mathbb{R}^n \to \mathbb{R}$ you need to evaluate f at $n+1$ points in $\mathbb{R}^n$

And if $f : \mathbb{R}^n \to \mathbb{R}$?

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h}$$

with e_i the canonical basis vector $(e_i)_j = \delta_{i,j} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$

Therefore $\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x + he_i) - f(x)}{h}$ or

$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x + he_i) - f(x - he_i)}{2h}$ for h small enough

Experiment numerically with notebook 1

Outline

Optimality conditions in the unconstrained case

Find extrema of a function defined on E.

Find

$$\inf_{x \in E} f(x),$$

with

$$f : E \longrightarrow \mathbb{R}$$

E normed finite dimension vector space

$E = \mathbb{R}^n$

Necessary optimality conditions

if $f(x^*) = \min_{\substack{\mathbb{R} \\ I}} f(x)$ then $f'(x) = 0$

I open interval of $\mathbb{R}$

Let $x^\star$ local minimum of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$.

1. First order optimality condition : if f is differentiable on an open neighborhood V of $x^\star$, then $\nabla f(x^\star) = 0$
2. Second order optimality condition : moreover if f is twice differentiable on V , then H$f(x^\star)$ is positive semi definite and f is locally convex in $x^\star$.

scalar case

if $f \in C^2$ $f'(x^*) = 0$ if x^* is a minimum then $f''(x^*) \geq 0$

$$f(x+th) = f(x) + Df(x)th + o(\|th\|)$$
$$= f(x) + \nabla f(x)\cdot th + o(\|th\|)$$

$$f(x^*) = \min_{x\in V(x^*)} f(x)$$

$t \in \mathbb{R}$

$$f(x+th) = f(x) + t\,\nabla f(x)\cdot h + \|h\|\,o(t)$$

$\boxed{t>0}$ for t small enough $x+th \in V(x^*)$

$$f(x+th) = f(x) + t\,\nabla f(x)\cdot h + \|h\|\,o(t)$$
$$f(x+th) - f(x) = t\,\nabla f(x)\cdot h + \|h\|\,o(t)$$
$$0 \le \frac{f(x+th)-f(x)}{t} = \nabla f(x)\cdot h + \|h\|\,\underbrace{\frac{o(t)}{t}}_{\to 0} \quad t\to 0$$

$$\nabla f(x)\cdot h \ge 0$$

$t<0 \Rightarrow \quad \nabla f(x)\cdot h \le 0$ $\Big\}\ \forall h \Rightarrow \nabla f(x) = 0$

Examples

$f(x) = x^4$ $\quad$ $0 = \min f(x)$

$f'(0) = 0$

$f''(0) = 12x^2 = 0$

Counter example : $f(x) = x^4$.
Counter example : $f(x) = x^3$.

$f(x) = x^3$

$f'(0) = 0$

$f''(0) = 6x = 0$

but 0 is not a minimum

Sufficient optimality conditions

If f is twice differentiable in $x^\star$, and if $\nabla f(x^\star) = 0$ and moreover

- either $Hf(x^\star)$ is positive definite
- either f is twice differentiable in a neighborhood V of $x^\star$ and $Hf(x)$ is positive semi definite on V $\longrightarrow$ $f(x) = x^4$

then $x^\star$ is a strict (isolated) minimizer of f on V.

$f'(0) = 0$

$f''(x) = 12x^2 \geq 0$

Uniqueness condition in the convex case

Proof by contraposée

(i) If f is convex on a convex subset $C \in \mathbb{R}^n$, any local minimum of f on C is global.
(ii) If f is strictly convex it has at most one global minimum.

Necessary and sufficient optimality condition in the convex case

If f is convex on $\mathbb{R}^n$ and C^1, $x^\star \in \mathbb{R}^n$ realizes a global minimum of f if and only if $\nabla f(x^\star) = 0$.

$$f(x) = \frac{1}{2} Ax.x + b.x + c$$

$x \in \mathbb{R}^n$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$, $A \in S^n(\mathbb{R})$

$$\nabla f(x) = Ax + b$$

$$\nabla f(x) = \frac{1}{2}(A + A^T)x + b$$

$$Hf(x) = A$$

$$Ax + b = 0_{\mathbb{R}^n}$$

1) if $A \in S^n_{++}(\mathbb{R})$ then $\exists! \, x^*$ $Ax^* = b$

$x^* = -A^{-1}b$

2) if $A \notin S^n_+(\mathbb{R})$

$\hookrightarrow f(x) \to -\infty$ for at least one $\vec{A}$

$$f(x) = \frac{1}{2} a x^2 + bx + c$$

$$f''(x) = a$$

$$f'(x) = ax + b$$

Optimality condition for quadratic problems

for $A \in S^n(\mathbb{R})$, $\lambda_1 \leq \dots \leq \lambda_n$ eigenvalues in $\mathbb{R}$

$\lambda_1 \|x\|^2 \leq Ax.x \leq \lambda_n \|x\|^2$

$$f(x) = \frac{1}{2}x^t Ax + b^t x + c = \frac{1}{2}(Ax) \cdot x + b \cdot x + c$$

with $A \in S_n^{++}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.

$$\inf_{x \in \mathbb{R}^n} f(x) \tag{1}$$

- If A is not positive semi definite then the problem (1) has no solution : no $x \in \mathbb{R}^n$ realizes a local minimum.
- If A is positive definite then $x^\star = -A^{-1}b$ is the only global minimum.

Solving systems of non linear equations

if $f(x^*) = \min_{x \in V(x^*)} f(x) \Rightarrow \nabla f(x^*) = 0$

$$\boxed{\nabla f(x) = 0_{\mathbb{R}^n}} \leftarrow$$

$\nabla f(x) \in \mathbb{R}^n$

$x \in \mathbb{R}^n$

$g(x) = 0$

$g : \mathbb{R}^n \to \mathbb{R}^n$

set of n equations
with n unknowns

Fixed point method

$\boxed{g(x) = 0}$

$\boxed{x = \varphi(x)}$, $\boxed{\varphi = Id - g}$

$Id - \varphi = g$

Definition

A fixed point $x \in \mathbb{R}^N$ of a function $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is a point such that $\boxed{x = \varphi(x)}$

Definition

A fixed point $x \in \mathbb{R}^N$ of a function $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is said to be attractive if there exists a neighborhood V of x such that for all x_0 in V, the sequence defined by $x_{n+1} = \varphi(x_n)$ converges to x. Otherwise, the point is said to be repulsive.

x_0, $x_{n+1} = \varphi(x_n)$

$\longrightarrow \quad x_n \xrightarrow[n \to +\infty]{} x$

Picard's Theorem

Theorem (Picard's Theorem)

Let F be a closed ~~space~~ subset of $\mathbb{R}^N$ and let $\varphi : F \subset \mathbb{R}^N \to \mathbb{R}^N$ be a map such that $\varphi(F) \subset F$. We assume that φ is contracting, i.e. there exists $k \in]0,1[$ such that:

$$\forall x, y \in F, \qquad \|\varphi(x) - \varphi(y)\| \le k\|x - y\|. \tag{2}$$

Then there exists a unique $x^* \in F \subset \mathbb{R}^N$ such that $\boxed{\varphi(x^*) = x^*}$ and, for all $x_0 \in F$, the sequence defined by $x_{n+1} = \varphi(x_n)$ converges to x^* (i.e. $x^\star$ is an attractive fixed point). Furthermore, there exists a constant C (depending on the choice of x_0 and the function φ) such that

$$e_n := \|x_n - x^*\| \le Ck^n. \tag{3}$$

Fixed point algorithm

$g(x) = 0$

Simplest method

```
def g(x):
    y = g(x)
    return y

def phi(x):
    return x - g(x)
```

choose x_0, $x_1 = \text{phi}(x_0)$, ε, n_{max}

while $\|x_1 - x_0\| \geq \varepsilon$ and $n \leq n_{max}$

$x_{n+1} = \varphi(x_n)$

$n = n+1$

$g(x) = 0$, scalar case

Newton method

Zeros of function $g : \mathbb{R} \longrightarrow \mathbb{R}$

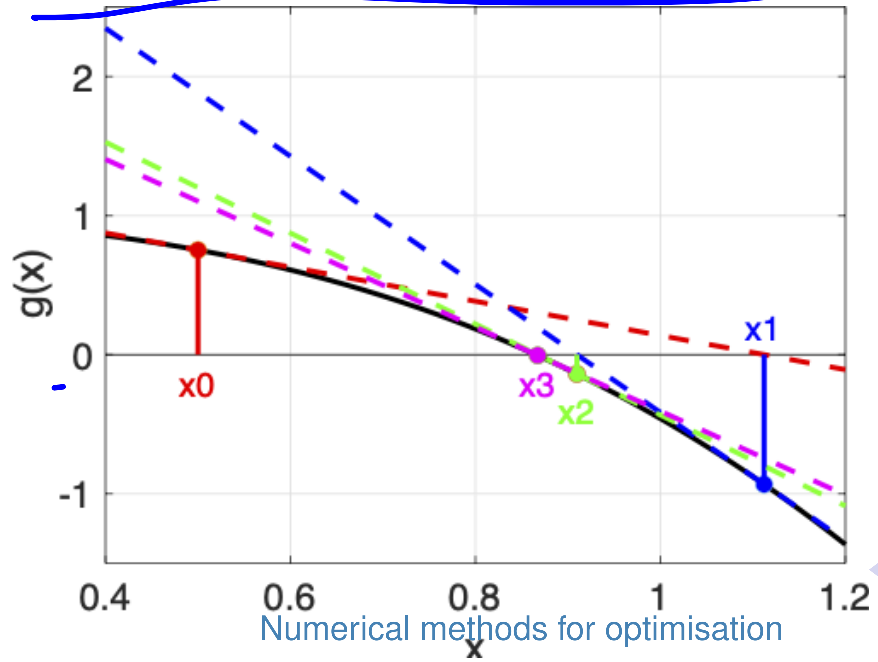
$$g(x^\star) \quad = \quad g(x) + g'(x)(x^\star - x) + o(\|x^\star - x\|).$$

Fixed point algorithm to solve a nonlinear equation

$$\psi(x) = x, \text{ with } \psi(x) = x - g(x)/g'(x).$$

Approximation by a sequence x_n

$$\boxed{x_{n+1} \quad = \quad x_n - g(x_n)/g'(x_n)}$$



$$\psi(x) = x - \frac{g(x)}{g'(x)}$$

$$\psi'(x) = 1 - \frac{g'(x)}{g'(x)} + \frac{g''(x)\,g(x)}{g'(x)^2}$$

$$= \frac{g''(x)\,g(x)}{g'(x)^2}$$

at the solution of $g(x^*) = 0$

$$\psi'(x^*) = 0$$

by continuity ψ' is contracting in a neighborhood of x^*.

Newton algorithm (scalar case)

Data: Function $g(x)$ derivative $g'(x)$, tolerance ε, max number of iterations $k_{\max}$

Result: $x^\star$ such that $g(x^\star) = 0$

Initialisation : $k = 0$, x_0 initial guess for $g(x_0) = 0$.

while $|g(x_k)| > \varepsilon$ *and* $k \leq k_{\max}$ **do**

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$$

$$k \leftarrow k + 1$$

end

$x^\star \leftarrow x_k$

if x_0 is close enough to x^* it converges very fast
if not it may diverge

Convergence of Newton algorithm

Let g in C^2 on $I = [x^\star - r, x^\star + r]$ with $g(x^\star) = 0$ and $g' \neq 0$ on I. Let

$$M = \max_{x \in I} \left| \frac{g''(x)}{g'(x)} \right|, \quad \text{and } h = \min\left(r, \frac{1}{M}\right).$$

Then for any $x_0 \in]x^\star - h, x^\star + h[$ we have

$$e_k := |x_k - x^\star| \leq \frac{1}{M} \left(M|x_0 - x^\star|\right)^{2k},$$

from which we deduce $\lim_{k\to+\infty} |x_k - x^\star| = 0$.

Convergence speed

for any iterative algorithm $x_0, (x_k)_k$

Denote by $e_k = x^k - x^\star$ the error at iteration k. We say that

- the algorithm converges if $\lim_{k\to\infty} \|e_k\| = 0$
- the algorithm converges linearly if $c \in]0, 1[$ tel que $\|e_k\| \leq c\|e_{k-1}\|$ for $k > K(c)$
- the algorithm converges supra-linearly if $(c_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} c_k = 0$ such that $\|e_k\| \leq c_k\|e_{k-1}\|$
- the algorithm converges geometrically if the sequence c_k is geometric
- the algorithm~~e~~ is of order p if there exists $c \in]0, 1[$ such that $\|e_k\| \leq c\|e_{k-1}\|^p$ for $k > K(c)$

$p = 2$

The convergence can be global or local

Newton method in dimension n

$g'(x) \longleftrightarrow Jg(x)$

$$\boxed{x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}}$$

Scalar case

$G(x) = 0$ with $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$.
Let $JG(x) \in \mathbb{R}^{n\times n}$ the jacobian matrix of G in x,
$JG_{i,j}(x) = \dfrac{\partial G_i(x)}{\partial x_j}$.

$$G(x^\star) = G(x) + JG(x)(x - x^\star) + o(||x - x^\star)||.$$

Fixed point algorithm to solve a nonlinear equation

$$\Psi(X) = X, \text{ with } \Psi(\cancel{X}) = X - JG(X)^{-1}G(X).$$

Approximation by a fixed point sequence $X_{n+1} = \Psi(X_n)$

X_0 ;

$$X_{n+1} = X_n - JG(X_n)^{-1}G(X_n)$$

Except that in practice for n large, one never computes the inverse of a matrix

Newton-Ralphson algorithm

extension to $\mathbb{R}^n$

$x_{n+1} = x_n - JG(x_n)^{-1}\, G(x_n)$

Data: Function $G(x)$, jacobian matrix $JG(x)$, tolerance ε, max number of iterations $k_{\max}$

Result: $x^\star$ such that $G(x^\star) = 0$

Initialisation : $k = 0$, x_0

while $\|G(x_k)\| > \varepsilon$ *and* $k \le k_{\max}$ **do**

  Solve $JG(x_k) d_k = -G(x_k)$ do not compute $JG(x_k)^{-1}$

  $x_{k+1} = x_k + d_k$

  $k \leftarrow k + 1$

end

$x^\star \leftarrow x_k$

Convergence of Newton-Ralphson algorithm

Suppose :

- G of class C^2
- $G(x^\star) \cancel{\neq} 0$
- the tangent linear map $JG(x^\star) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ is inversible.

Jacobian matrix

Then $x^\star$ is a superattractive fixed point of

$$\Psi(x) = x - (JG(x))^{-1} G(x).$$

Local result.

Scalar case : the secant method

in the case where $g'(x)$ is not accessible

Data: Function $g(x)$, tolerance ε, max number of iterations $k_{\max}$

Result: $x^\star$ such that $g(x^\star) = 0$

Initialisation : $k = 0$, x_0 initial guess for $g(x) = 0$.

a_0 initial guess for $g'(x_0)$ (default =1)

while $|g(x_k)| > \varepsilon$ *and* $k < k_{\max}$ **do**

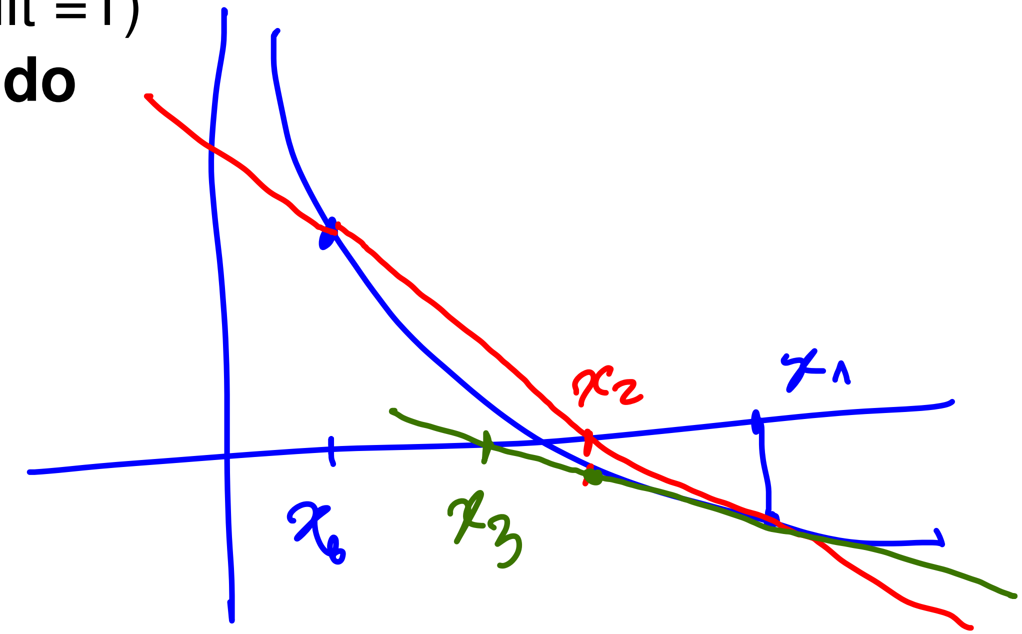
$$x_{k+1} = x_k - \frac{g(x_k)}{a_k}$$

$$a_{k+1} = \frac{g(x_k) - g(x_{k+1})}{x_k - x_{k+1}}$$

$$k \leftarrow k + 1$$

end

$x^\star \leftarrow x_k$



Vector case : the quasi-Newton method

Data: $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$
$\varepsilon > 0$.
Result: $x^\star$ such that $g(x^\star) = 0$
Initialisation : a first approximation of $x_0 \in \mathbb{R}^n$
$A_0 \approx J(x_0)$ or $W_0 \approx J(x_0)^{-1}$
$x_1 = x_0 - W_0 G(x_0)$
$d_0 = x_1 - x_0,$
$y_0 = G(x_1) - G(x_0),$
$k = 1$
while $\|G(x_k)\| > \varepsilon$ *and* $k < k_{\max}$ **do**
  Update : $W_k = W_{k-1} + B_{k-1}$
  Compute d_k solution of $d_k = -W_k G(x_k)$
  $x_{k+1} = x_k + d_k$
  $y_k = G(x_{k+1}) - G(x_k)$
  $k \leftarrow k + 1$
end
$x^\star \leftarrow x_k$

$JG(x_k)\, d_k = -G(x_k)$
$x_{k+1} = x_k + d_k$

$A_k d_k = -G(x_k)$
$x_{k+1} = x_k + d_k$
$A_{k+1} = f(A_k)$

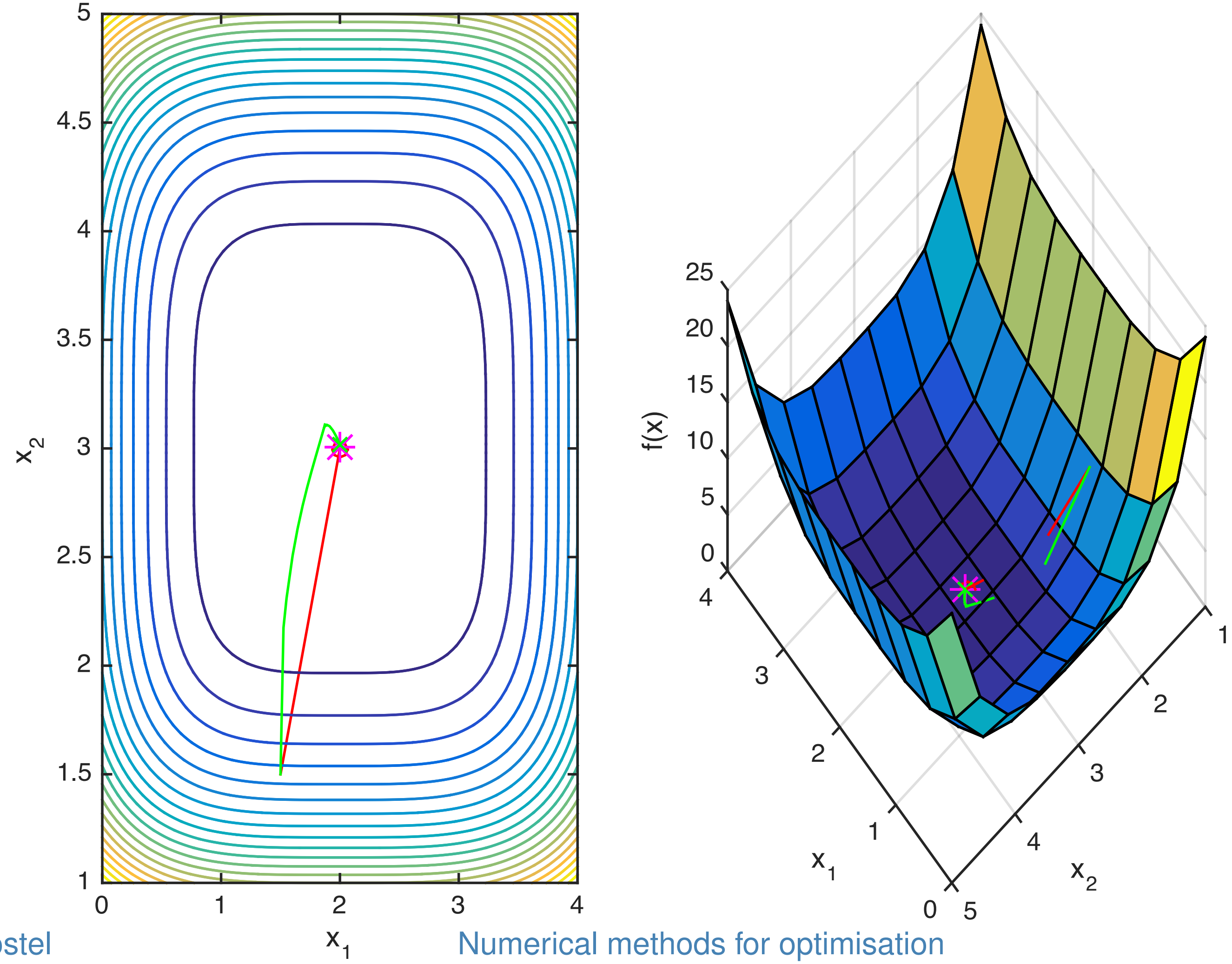
$x_{k+1} = x_k - W_k G(x_k)$
$W_{k+1} = f(W_k)$

Comparison of Newton and quasi Newton methods

Minimum of the quadratic function $f(x) = ((x_1 - 2)^4 + (x_2 - 3)^4)/2$

Newton method : 12 iterations

quasi Newton (BFGS) method : 21 iterations



Update in the quasi Newton method

$W_k \simeq JG(x_k)^{-1}$

Update :

$W_k = W_{k-1} + B_{k-1}$ $\quad W_k = \mathcal{F}(W_{k-1})$

Compute d_k solution of $d_k = -W_k G(x_k)$

$x_{k+1} = x_k + d_k$

Conditions on the W_k matrix

1. W_k should remain symetric positive definite for all k.
2. The quasi-Newton equation $W_k y_k = d_k$ is satisfied for each k
3. The difference between two consecutive approximations $W_{k+1} - W_k$ is minimum in some sense (for some norm), for example for the Frobenius norm $\|A\| = \sum_i \sum_j |A_{ij}|$

Examples of update rules satisfying the conditions

$$W_k = W_{k-1} + B_{k-1}$$
$$d_k = -W_k G(x_k)$$
$$x_{k+1} = x_k + d_k$$
$$y_k = G(x_{k+1}) - G(x_k)$$

$y_k \, d_k^T$

- The Davidon-Fletcher-Powell method

$$(DFP) \quad W_{k+1} = W_k + \frac{d_k d_k^T}{y_k \cdot d_k} - \frac{W_k y_k y_k^T W_k}{(y_k \cdot W_k y_k)}.$$

- The Broyden-Fletcher-Goldfarb-Shanno method

$$\boxed{(BFGS)} \quad W_{k+1} = W_k - \frac{d_k y_k^T W_k + W_k y_k d_k^T}{y_k \cdot d_k} + \left(1 + \frac{y_k \cdot W_k y_k}{y_k \cdot d_k}\right) \frac{d_k d_k^T}{y_k \cdot d_k}.$$

Experiment with notebook 2

$f(x^*) = \min f(x) \Rightarrow \nabla f(x) = 0$

1st set of methods

Solve $g(x) = 0_{\mathbb{R}^n}$

$f(x) = \|x\|^2$

for $\alpha > \alpha_d$ $f(x + \alpha d) > f(x)$

Descent method

build sequence $(x_n)_n$
such that $f(x_{n+1}) \leq f(x_n)$

Definition : We say that $d \in \mathbb{R}^n$ is a descent direction at x for the function f if there exists $\alpha_d > 0$ such that

$$f(x + \alpha d) < f(x) \quad \boxed{\forall 0 < \alpha \not< \alpha_d}.$$

when you cross the level curve $f(x)$ again

Property : If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable , $d \in \mathbb{R}^n$ is a descent direction at x if and only if

$$\nabla f(x) \cdot d < 0.$$

$$f(x+\alpha d) = f(x) + \underbrace{\alpha \underbrace{\nabla f(x) \cdot d}_{<0} + \|d\| o(|\alpha|)}_{<0 \text{ for } \alpha \text{ small enough}}$$

$$\underbrace{\frac{f(x+\alpha d) - f(x)}{\alpha}}_{<0} = \nabla f(x) \cdot d + o(|\alpha|)/\alpha$$

Example $f(x) = \frac{1}{2} Ax.x + b.x + c$

necessary condition to have a solution $\min f(x)$

$A \in S_n^+$; $\|\nabla f(x) - \nabla f(y)\| = \|A(x-y)\|$
$\|A\|.\|x-y\|$

sufficient: $A \in S_n^{++}$ $Ax^* = -b$ $L = \lambda_n$

eigenvalues: $0 < \lambda_1 \leq \dots \leq \lambda_n$

$$\underline{\lambda_1 \|x\|^2 \leq Ax.x}$$

f is strongly convex or λ_1-elliptic

$$Hf(x)d.d \geq \lambda_1 \|d\|^2$$

the gradient method will converge if we choose

$$0 < \alpha < \frac{2\lambda_1}{\lambda_n^2}$$

General descent algorithm

(x_k) $f(x_{k+1}) < f(x_k)$

Data: function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
required precision $\varepsilon > 0$.
Result: $x^\star$ such that $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Initial guess for the solution $x_0 \in \mathbb{R}^n$
while $\|\nabla f(x_k)\| > \varepsilon$ *and* $k < k_{\max}$ **do**

  Choose descent direction d_k (such that $\nabla f(x_k) \cdot d_k < 0$)
  Choose step α_k in direction d_k, such that
   $f(x_k + \alpha_k d_k) \leq f(x_k)$
  $x_{k+1} = x_k + \alpha_k d_k$
  $k \leftarrow k + 1$

end
$x^\star \leftarrow x_k$

Convergence of a descent algorithm

Sufficient conditions

Let f a function C^1 from $\mathbb{R}^n$ into $\mathbb{R}$ and $x^\star$ minimizer of f. If the following conditions are satisfied

1. f α-elliptic
2. ∇f L-lipschitz

$Hf(x)\, d \cdot d \geq \alpha \|d\|^2$ 2nd ord

$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha \|x-y\|^2$ 1st order

Then for all $(\alpha_k)_{k\in\mathbb{N}}$ sequence such that there exist $a, b \in \mathbb{R}$, s.t.

$\|\nabla f(x) - \nabla f(y)\|^2 \leq L \|x - y\|^2$

$$0 < a \leq \alpha_k \leq b < \frac{2\alpha}{L^2}, \quad \forall k \in \mathbb{N},$$

The gradient method

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

$d_k = -\nabla f(x_k)$

$d_k \cdot \nabla f(x_k) = -\|\nabla f(x_k)\|^2 < 0$ for $x_k \neq x^*$

converges geometrically for all initial guess, i.e.

$$\exists \beta \in]0,1[, \|x_k - x^\star\| \leq \beta^k \|x_0 - x^\star\|$$

$$e_{k+1} = x_{k+1} - x^* \quad \text{with} \quad \nabla f(x^*) = 0$$

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha_k \nabla f(x_k) - x^*\|^2$$

$$= \|x_k - x^* - \alpha_k(\nabla f(x_k) - \nabla f(x^*))\|^2$$

$$= \|x_k - x^*\|^2 - 2\alpha_k \underbrace{(x_k - x^*)\cdot(\nabla f(x_k) - \nabla f(x^*))}_{\geq \alpha \|x_k - x^*\|^2} + \alpha_k^2 \underbrace{\|\nabla f(x_k) - \nabla f(x^*)\|^2}_{\leq L^2 \|x_k - x^*\|^2}$$

$$\|e_{k+1}\|^2 \leq \|x_k - x^*\|^2 (\underbrace{1 - 2\alpha_k \alpha + L^2 \alpha_k^2}_{<1})$$

choose

$$L^2 \alpha_k^2 - 2\alpha_k \alpha < 0$$

$$\alpha_k (L^2 \alpha_k - 2\alpha) < 0$$

$$\text{if} \quad \alpha_k < \frac{2\alpha}{L^2}$$

Examples of possible choices for the descent direction

- Gradient Algorithm (*steepest descent*)

$$d_k = -\nabla f(x_k).$$

- Newton algorithm based on direction

$$d_k = \boxed{-Hf(x_k)^{-1}\nabla f(x_k).} \qquad \alpha_k = 1$$

- Quasi-Newton with

$$d_k = -W_k \nabla f(x_k),$$

where $W_k \approx Hf(x_k)$ only for $f(x) = \frac{1}{2}Ax.x + b.x + c$

- Conjugate gradient method (in the quadratic case)

$$x_{k+1} = x_k + \alpha_k d_k \qquad d_k = \begin{cases} -\nabla f(x_1) & \text{for } k = 1 \\ -\nabla f(x_k) + \beta_k d_{k-1} & \text{for } k > 1. \end{cases}$$

Choice of the step in a given direction

at iteration k

$$h_k : \alpha \mapsto h_k(\alpha) = f_{\cancel{k}}(x_k + \alpha d_k)$$

Cauchy's rule

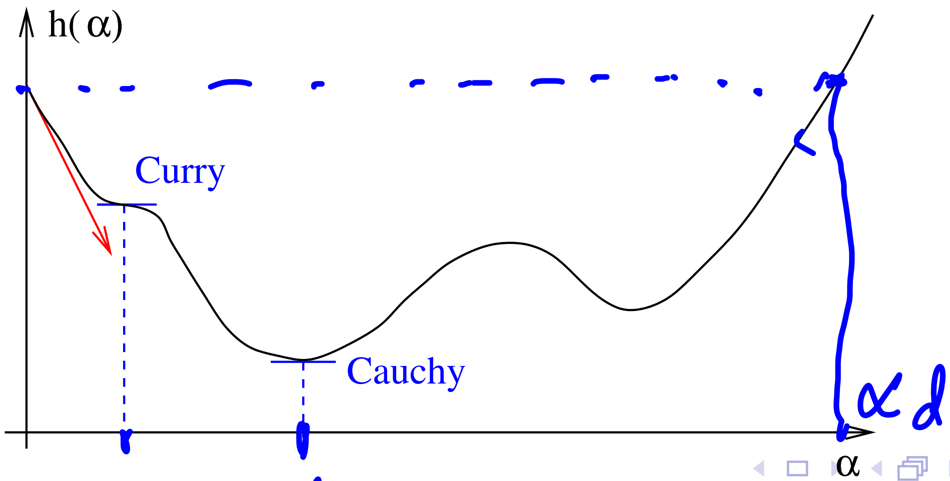
$$h_k'(\alpha) = d_k \cdot \nabla f(x_k + \alpha d_k) = 0 \;!!!$$

$$\alpha_k = \underset{\alpha > 0}{\operatorname{argmin}}\, h_k(\alpha)$$

difficult.

Curry's rule

$$\alpha_k = \inf\left\{\alpha > 0;\, h_k'(\alpha) = 0,\, h_k(\alpha) < h_k(0)\right\}$$



Convergence of optimal step gradient

Suppose $\nabla f(x)$ L-Lipschitz on $\{x, f(x) \leq f(x^0)\}$.
Then the gradient algorithm with:

- $d_k = -\nabla f(x^k)$
- α_k fixed by Curry's rule

satisfies

- either $f(x^k)$ non-bounded below
- either $\nabla f(x^k) \to 0$ when $k \to \infty$.

$$f(x) = \frac{1}{2} Ax \cdot x + b \cdot x + c \qquad A \in S_n^{++}$$

$$h(\alpha) = f(x + \alpha d)$$

$$h'(\alpha) = d \cdot \nabla f(x + \alpha d)$$

$$= d \cdot \left(A(x + \alpha d) + b\right)$$

$$= d \cdot (Ax + b) + \alpha\, d \cdot Ad$$

$$h'(\alpha) = 0 \iff \alpha = \frac{d \cdot \nabla f(x)}{d \cdot Ad} \quad \text{better optimal}$$

optimal step in the quadratic case

$$0 < \alpha < \frac{2\lambda_1}{\lambda_n^2} \quad \text{sufficient}$$

Optimal step in the quadratic case

$$f(x) = \frac{1}{2} Ax \cdot x + b \cdot x$$

with $A \in S^n$ and $b \in \mathbb{R}^n$,

$$h(\alpha) = f(x + \alpha d) = f(x) + \frac{\alpha^2}{2} Ad \cdot d + \alpha (Ax + b) \cdot d$$

$\alpha^\star = \operatorname{argmin} h(\alpha)$

$$\alpha^\star = -\frac{g \cdot d}{(Ad) \cdot d}$$

with $g = \nabla f(x) = Ax + b$.

Optimal step gradient method in the quadratic case

Data: A, b, ε
Result: $x^\star$ s.t. $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$, $x_0 \in \mathbb{R}^n$
$g_0 = Ax_0 + b$
while $\|g_k\| > \varepsilon$ *and* $k < k_{\max}$ **do**

$$d_k = -g_k$$
$$v_k = Ad_k$$
$$\alpha_k = \frac{d_k \cdot d_k}{v_k \cdot d_k}$$
$$x_{k+1} = x_k + \alpha_k d_k$$
$$g_{k+1} = g_k + \alpha_k v_k$$
$$k \leftarrow k + 1$$

end
$x^\star \leftarrow x_k$

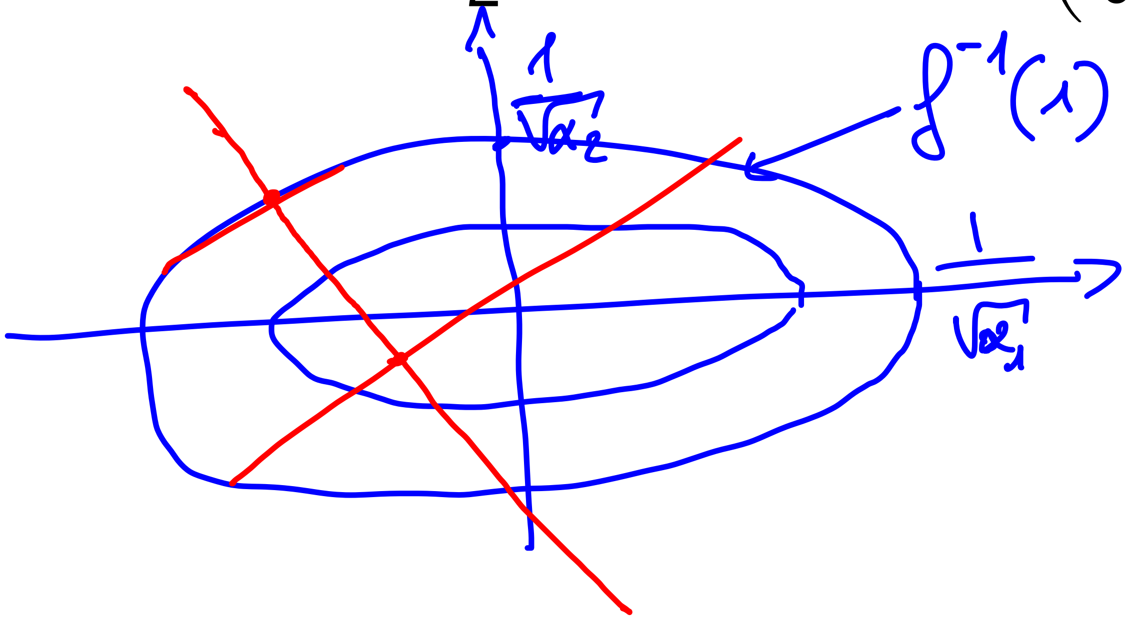
Conjugate gradient method : motivation

$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

with α_k optimal

will converge only asymptotically

$$
\begin{aligned}
f(x^*) &= \inf_{x \in \mathbb{R}^n} f(x) \\
f(x) &= \frac{1}{2}(\alpha_1 x_1^2 + \alpha_2 x^2), \quad \text{with } 0 < \alpha_1 < \alpha_2 \\
&= \frac{1}{2}(Ax)\cdot x \quad \text{with } A = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix}
\end{aligned}
$$



A-conjugate directions

if $f(x) = \frac{1}{2} Ax.x + b.x + c$

$Ax.x$ is a scalar product in the eigen vector basis

Definition : Let $A \in S^n_{++}$.

- 2 non zero vectors v, w are called $A-$conjugate iff $Av \cdot w = 0$.
- A family of non zero vectors $(v_i)_{i=1,\dots m}$, is called $A-$conjugate iff $Av_i \cdot v_j = 0$ for all $i = 1, \dots, m$, $j = 1, \dots, m$, $i \neq j$.

Property : $A-$conjugate vectors are linearly independent. If $m = n$ a $A-$conjugate family is a basis of $\mathbb{R}^n$.
Definition : a conjugate descent method is a method where the successive descent directions form a $A-$conjugate family

Expression of the minimum of f in a $A-$conjugate basis

$$f(x) = \frac{1}{2}Ax \cdot x + b \cdot x$$

Suppose we have a basis $(d_i)_{i=1,\dots n}$, such that $Ad_i \cdot d_j = 0$ for $j \neq i$

$$x^\star = \sum_{i=1}^{n} \alpha_i d_i, \text{ and } Ax^\star + b = 0,$$

$$\text{therefore } Ax^\star = -b = \sum_{i=1}^{n} \alpha_i A d_i, \text{ then for any } j = 1, \dots, n$$

$$-b \cdot d_j = \sum_{i=1}^{n} \alpha_i A d_i \cdot d_j = \alpha_j A d_j \cdot d_j$$

$$\alpha_j = \frac{-b \cdot d_j}{Ad_j \cdot d_j}$$

Construction of the $A-$conjugate basis

Let $g_k = \nabla f(x_k) = Ax_k + b$ be the gradient at step k
Choose $d_0 = -g_0$ (The first step is a standard gradient descent step)
Then $d_k = -g_{k-1} + \beta_{k-1} d_{k-1}$ satisfying:
(CG1) $\quad Ad_k \cdot d_j = 0$ for $j = 0, \ldots, k-1$ and
(CG2) $\quad g_k \cdot d_j = 0$ for $j = 0, \ldots, k-1$
Update at step k : $x_{k+1} = x_k + \alpha_k d_k$
Next gradient $g_{k+1} = Ax_{k+1} + b = g_k + \alpha_k Ad_k$
Property : For all initial guess x_0 there exists $(\alpha_k)_k$ and $(\beta_k)_k$ such that (CG1) and (CG2) are satisfied.
Property : (CG1) and (CG2) $\Rightarrow g_k \cdot g_j = 0$ for $j \neq k$

Convergence of a conjugate method

Property : A conjugate descent method using directions satisfying conditions (CG1) and (CG2) converges in at most *n* steps.

Property : $\beta_k = -\dfrac{Ad_{k-1} \cdot g_k}{Ad_{k-1} \cdot d_{k-1}} = \dfrac{\|g_k\|^2}{\|g_{k-1}\|^2}.$ ← to update the direction

Property : $\alpha_k = -\dfrac{g^k \cdot d^k}{Ad^k \cdot d^k}$ ← same optimal step as before

$d_{k+1} = -\nabla f(x_k) + \beta_k d_k$

Conjugate gradient algorithm

optimal step
Gradient method

Data: Matrix A, vector b, tolerance ε
Result: $x^\star$ such that $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Initial guess for solution $x^0 \in \mathbb{R}^n$
$g^0 = Ax^0 + b$
$d^0 = -g^0$
while $\|g^k\| > \varepsilon$ **do**

- Compute directionnal minimum :
  $v^k = Ad^k$
  $\alpha_k = -\dfrac{g^k \cdot d^k}{v^k \cdot d^k}$
  $x^{k+1} = x^k + \alpha_k d^k$
- Update gradient :
  $g^{k+1} = g^k + \alpha_k v^k$
- Compute new direction :
  $\beta_{k+1} = \dfrac{g^{k+1} \cdot g^{k+1}}{g^k \cdot g^k}$
  $d^{k+1} = -g^{k+1} + \beta_{k+1} d^k$

$k \leftarrow k + 1$
end
$x^\star \leftarrow x^k$

$v^k = Ad_k$
$\alpha_k = -\dfrac{g^k \cdot d^k}{v^k \cdot d^k}$
$x^{k+1} = x^k + \alpha_k d^k$
$g^{k+1} = g^k + \alpha_k v^k$

Monotonicity of the conjugate gradient algorithm

Property : If $d_k \neq 0$ and $\alpha_{k+1} \neq 0$ then $f(x_{k+1}) < f(x_k)$.
If $\alpha_{k+1} = 0$, x_k is the minimizer of f and $Ax_k + b = 0$

Polak-Ribière method

for f non quadratic

Data: Function f, gradient ∇f, tolerance ε
Result: $x^\star$ such that $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Initial guess for $x^0 \in \mathbb{R}^n$
$g^0 = \nabla f(x^0)$
$d^0 = -g^0$
while $\|g^k\| > \varepsilon$ *and* $k < k_{\max}$ **do**

- Compute the step in direction d_k:
  $f(x^k + \alpha_k d^k) \leq f(x^k + \alpha d^k) < f(x^k)$ for all $0 < \alpha \leq \alpha_k$
- Compute new position :
  $x^{k+1} = x^k + \alpha_k d^k$
- Compute new direction :
  $g^{k+1} = \nabla f(x^{k+1})$
  $c_{k+1} = \dfrac{(g^{k+1} - g^k) \cdot g^{k+1}}{g^k \cdot g^k}$
  $d^{k+1} = -g^{k+1} + c_{k+1} d^k$

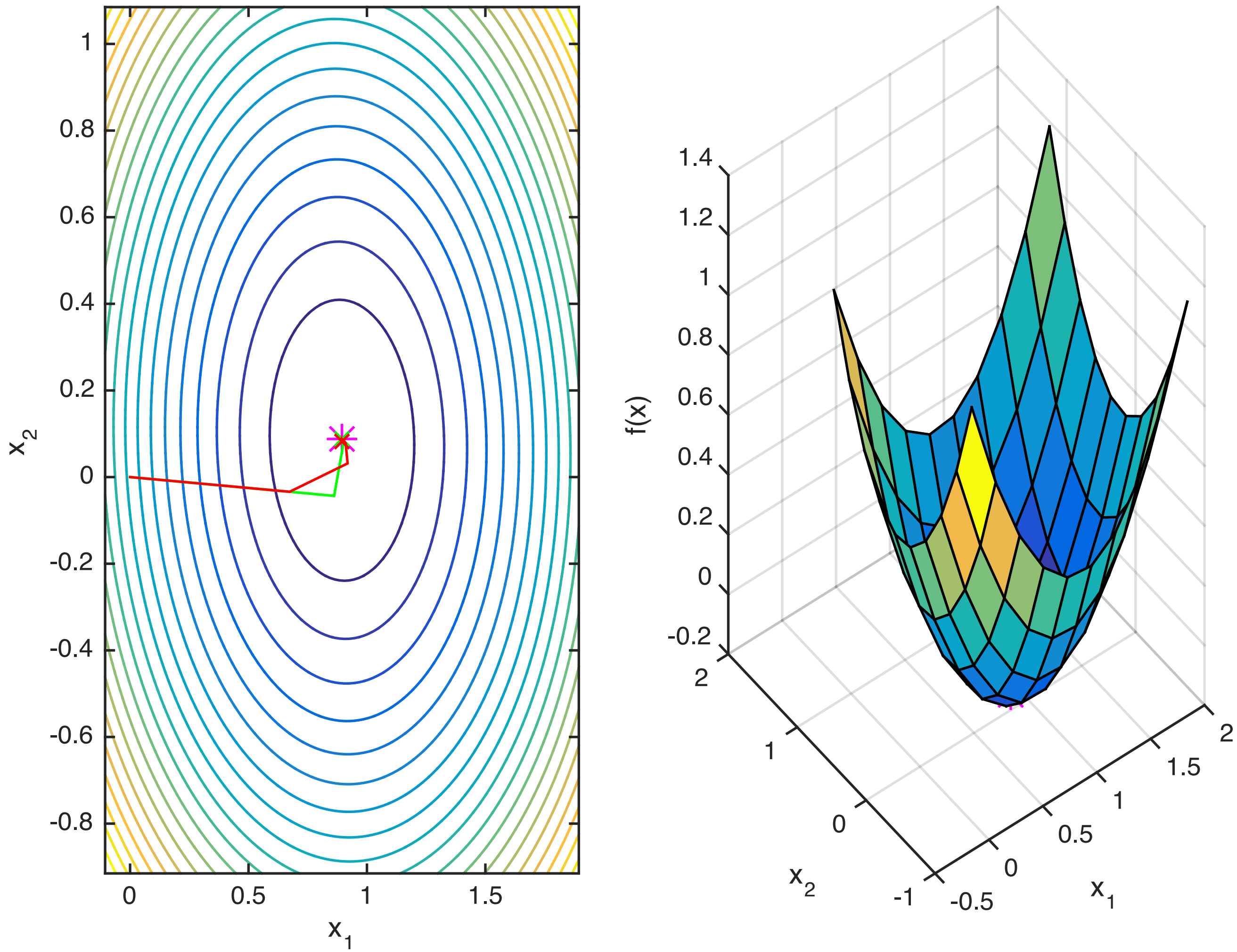
$k \leftarrow k + 1$
end
$x^\star \leftarrow x^k$

1) will not converge in a finite number of iteration

2) no explicit formula for α optimal

Comparaison of conjugate Gradient (green, 4 steps)and Polak-Ribière (red, 8 steps) methods.

f quadratic function in $\mathbb{R}^5$. Projection on $(0, x_1, x_2)$.



Choice of the step in the general case

Data: Function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
Required precision $\varepsilon > 0$.
Result: $x^\star$ s.t. $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Intial guess $x_0 \in \mathbb{R}^n$
while $\|\nabla f(x_k)\| > \varepsilon$ *and* $k < k_{\max}$ **do**
    Choose d_k, s.t. $\nabla f(x_k) \cdot d_k < 0)$
    Choose step α_k in direction d_k, s.t. $f(x_k + \alpha_k d_k) \leq f(x_k)$
    $x_{k+1} = x_k + \alpha_k d_k$
    $k \leftarrow k + 1$
end
$x^\star \leftarrow x_k$

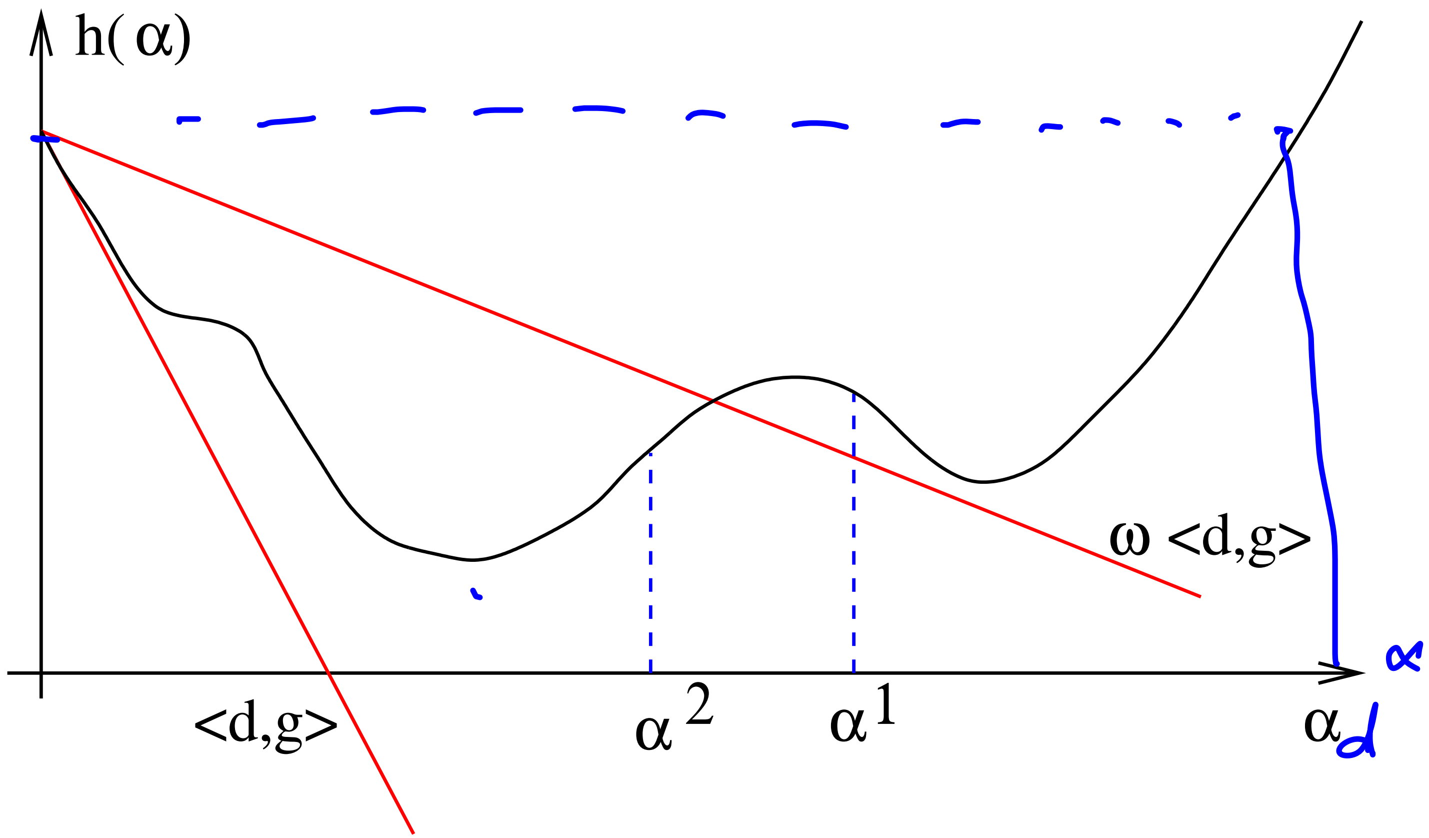
Directional minimisation -Line search

BLS
Backtracking
Line search

Armijo's rule linearization of the constraint on α_k

$$f(x^k + \alpha_k d^k) < f(x^k) + \omega \alpha_k g^k \cdot d^k$$

$g = \nabla f(x)$



Armijo's rule

BLS

this converges unconditionnaly for any $d.\nabla f(x) < 0$

Data: Function f, current position x, descent direction d,
coefficients $\tau \in]0,1[$ and $\omega \in]0,1[$

Result: α s.t. $f(x+\alpha d) < f(x)$

Initialisation : $k = 0$, initial guess α_0

while $f(x+\alpha_k d) > f(x) + \omega \alpha_k d \cdot \nabla f(x)$ **do**

  Choose $\alpha_{k+1} = \tau \alpha_k$

  $k \leftarrow k+1$

end

$\alpha = \alpha_k$

red line in the previous slide

Choice of the first value α_k^1

Assume a quadratic model for $\varphi(\alpha) = f(x^k + \alpha d^k)$

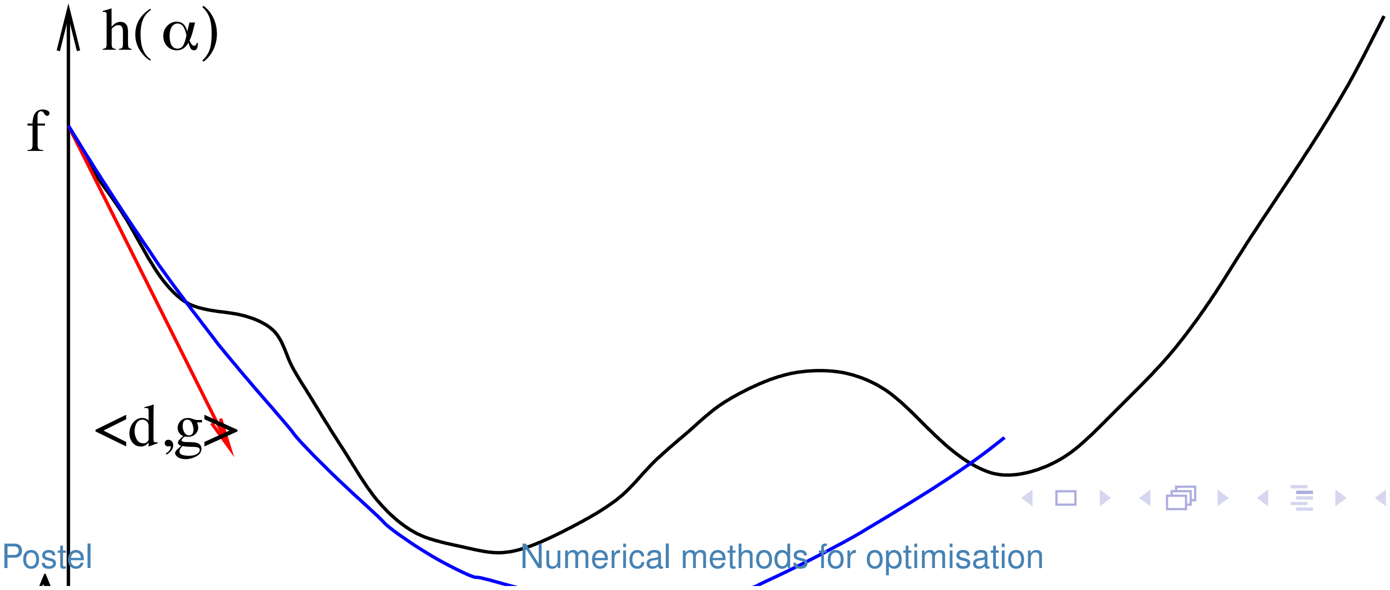
$$h(\alpha) = a_0 + a_1\alpha + a_2\alpha^2/2$$

with

$$\begin{cases} a_0 = f(x^k) \\ a_1 = d^k \cdot \nabla f(x^k) \end{cases}$$

a_2 is fixed by setting Δ, the maximal decrease of φ

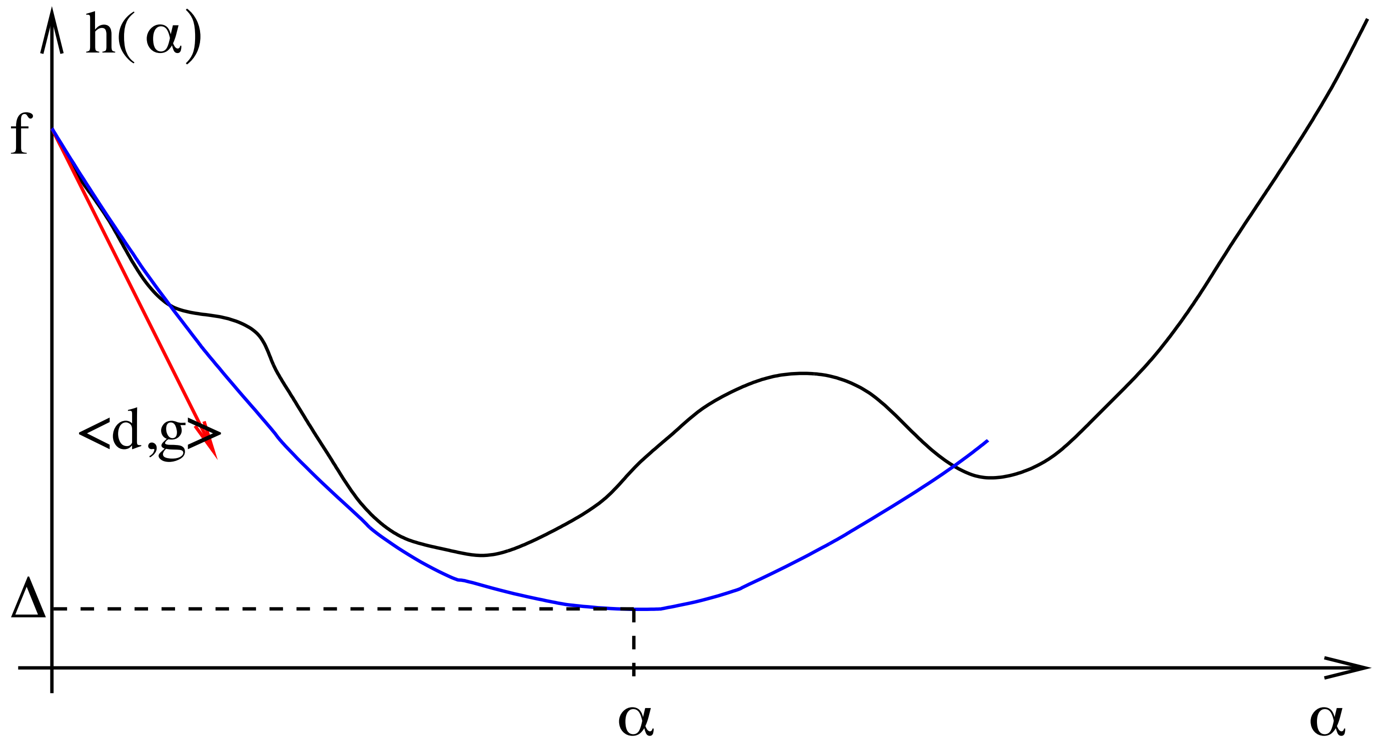
$$\Delta = \varphi(0) - \varphi_{\min} = a_1^2/(2a_2).$$



Fletcher's rule

α_k^1 is chosen to minimize the quadratic model

$$\alpha_k^1 = \frac{2\Delta}{d^k \cdot \nabla f(x^k)}$$



Convergence of gradient + Armijo methods

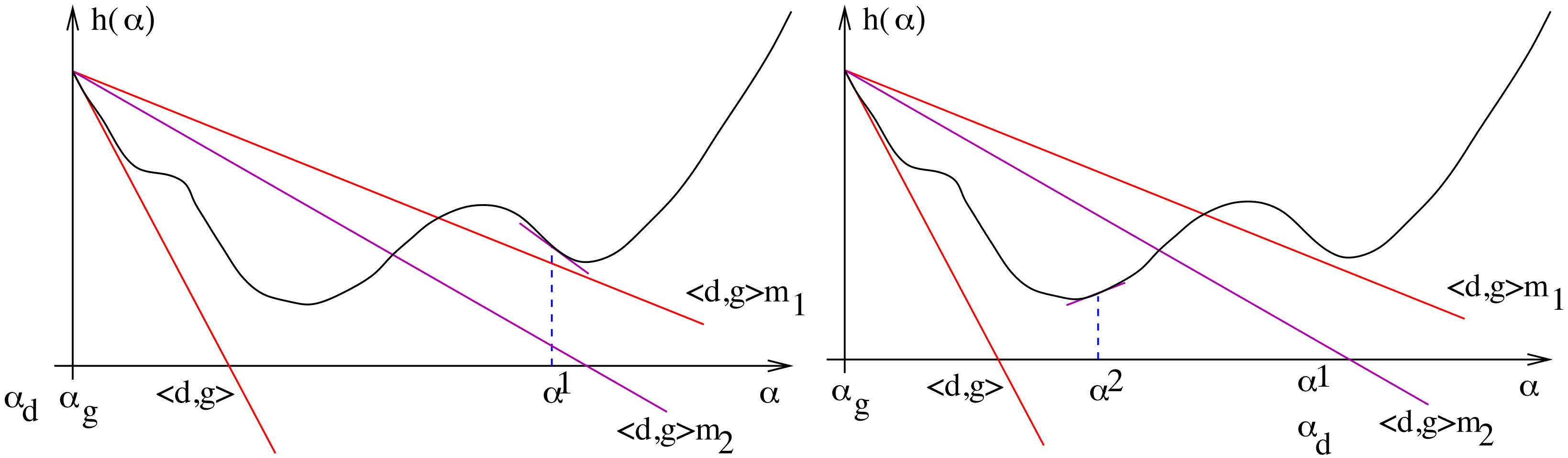
If $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 and $\nabla f(x)$ is γ-lipschitz then Armijo's rule is satisfied for all

$$\alpha \in [0, \omega], \quad \text{with } \omega = \frac{(\omega_1 - 1)\nabla f(x) \cdot d}{\gamma \|d\|^2}.$$

Drawback of Armijo strategy: $\alpha_k^{i+1} < \alpha_k^i$, slow convergence.

Wolfe's method

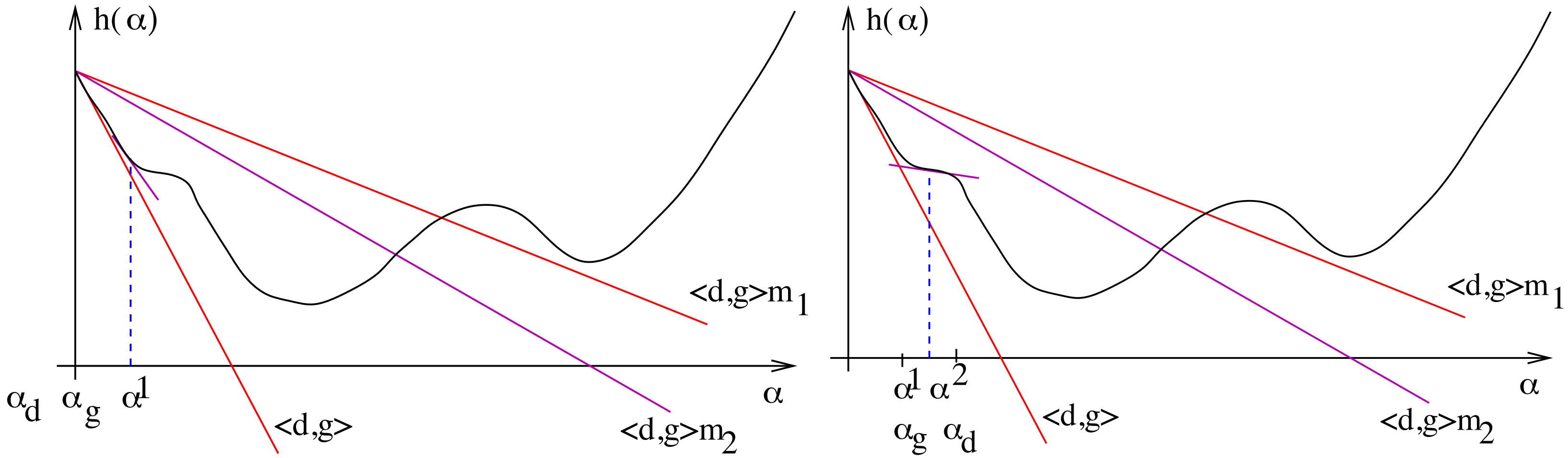
$$
\begin{array}{llcl}
(A) & f(x^k + \alpha_k d^k) & < & f(x^k) + \omega_1 \alpha_k g^k \cdot d^k \\
(W) & \nabla f((x^k + \alpha_k d^k) \cdot d^k & > & \omega_2 g^k \cdot d^k, \text{ with } 0 < \omega_1 < \omega_2 < 1.
\end{array}
$$



Wolfe algorithm

Data: Function f, gradient ∇f, current point x^k, descente direction d^k, coefficients $0 < \omega_1 < \omega_2 < 1$
Result: α_k s.t. (A) and (W) are satisfied
Initialisation : Fix $\alpha_D = -1$ and $\alpha_G = 0$.
$p = 0$, Fix initial guess α_k^p (Fletcher's rule)
while $f(x^k + \alpha_k^p d^k) > f(x^k) + \omega_1 \alpha_k^p d^k \cdot \nabla f(x^k)$ *or* $\nabla f((x^k + \alpha_k^p d^k) \cdot d^k < \omega_2 g^k \cdot d^k$ **do**
  if $f(x^k + \alpha_k^p d^k) > f(x^k) + \omega_1 g^k \cdot d^k \alpha_k^p$ **then**
    $\alpha_D = \alpha_k^p$
  end
  else
    $\alpha_G = \alpha_k^p$
  end
  if $\alpha_D < 0$ *(not yet updated)* **then**
    $\alpha_k^{p+1} = 2\alpha_G$
  end
  else
    $\alpha_k^{p+1} = (\alpha_G + \alpha_D)/2$
  end
  $p \leftarrow p + 1$
end
$\alpha_k = \alpha_k^p$

Wolfe example



Convergence of Wolfe method

Let $f : \mathbb{R}^n \to \mathbb{R}$, bounded below, C^1 différentiable on

$$\mathcal{N} = \{f(x) \leq f(x_0)\}$$

and gradient $\nabla f(x)$ L-lipschitz. Then, if coefficients $(\alpha_k)_k$ satisfy conditions (A) and (W)

$$\sum_k \cos\theta_k^2 \|\nabla f(x^k)\|^2 < \infty, \quad \text{with } \cos\theta_k = \frac{-d^k \cdot \nabla f(x^k)}{\|d^k\|, \|\nabla f(x^k)\|}.$$

Experiment with notebook 5

X is rectangular

$$E(\theta) = \frac{1}{m} \| X\theta - y \|^2$$

$$= \frac{1}{m} (X\theta - y) \cdot (X\theta - y)$$

$$\nabla E(\theta) = \frac{2}{m} X^T (X\theta - y)$$

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}$$

m lines $\times$ n columns

$$\nabla E(\theta^*) = 0_{\mathbb{R}^n}$$

$$X^T (X\theta^* - y) = 0$$

No simplification $\rightarrow$

$$X^T X \, \theta^* = X^T y$$

$$\theta^* = (X^T X)^{-1} X^T y$$

$X^T X$ invertible if n points at least are different among the m $\Leftrightarrow$ rank $X = n$

Linear regression

$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$ $\theta . x = \sum \theta_i x_i$

Find θ defining a linear model

$$\hat{y} = h_\theta(x) = \theta^{\cancel{T}}.x$$

Let m measurements (x_i, y_i), $i = 1, \ldots, m$, where explaining variables are in $\mathbb{R}^n$ $(x_i = (x_i^j)_{j=1,\ldots,n}$. θ is found by minimizing the least squared error

$$E(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta^{\cancel{T}}.x_i - y_i\right)^2 = \frac{1}{m} \|X\theta - y\|^2$$

$\nabla E(\theta) =$

The "normal equation" gives the best solution

$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}$

$$\hat{\theta} = (X^T.X)^{-1}.X^T.y \in$$

complexity in $O(n^3)$ and $O(m)$.

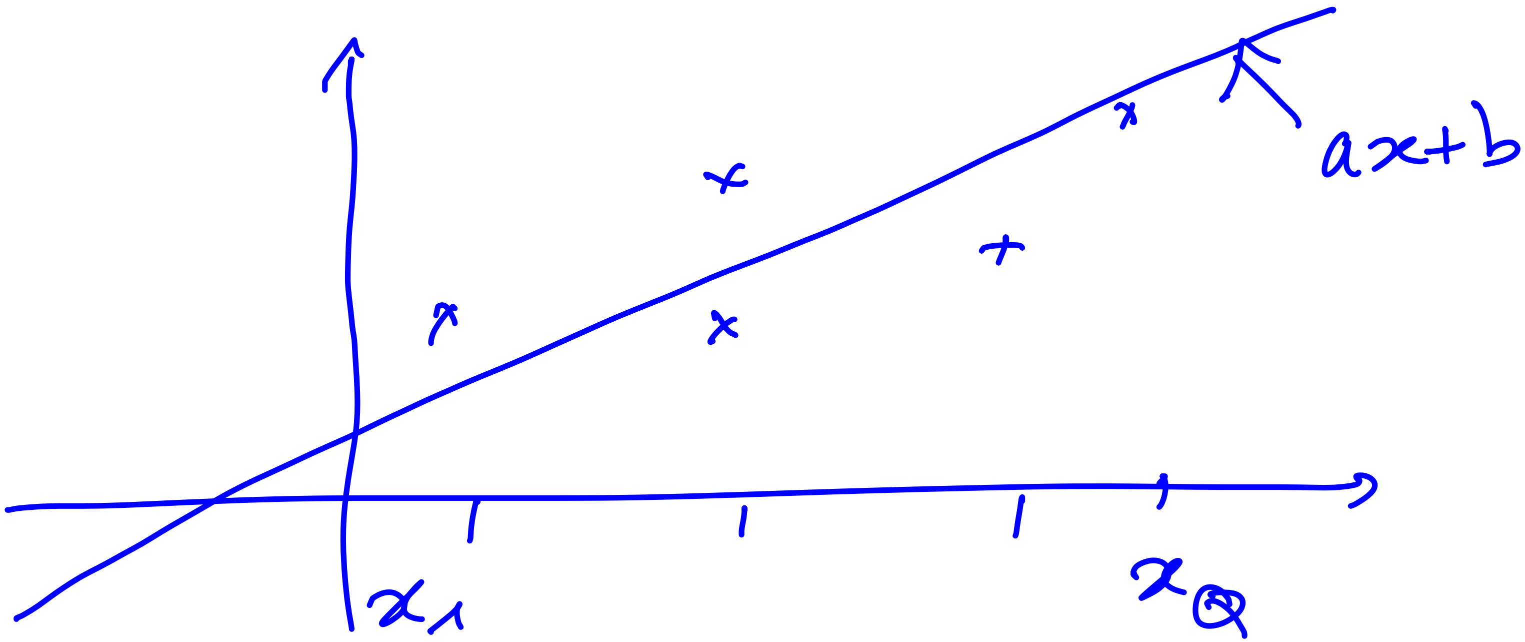
Nonlinear least squares

$$f : \left\{ \begin{array}{rcl} \mathbb{R}^P & \rightarrow & \mathbb{R}^Q \\ x = (x_1, \ldots, x_P)^t & \mapsto & (f_1(x), \ldots, f_Q(x))^t \end{array} \right.$$

for $Q > P$ we seek a solution to the problem $f(x) = 0$.

Examples

- Find a line that passes through Q points with $Q > 2$



$$f : \mathbb{R}^{2=P} \longrightarrow \mathbb{R}^Q$$

$$z = \begin{pmatrix} a \\ b \end{pmatrix} \longmapsto \begin{pmatrix} a x_1 + b \\ \vdots \\ a x_Q + b \end{pmatrix} = \begin{pmatrix} f_1(z) \\ \vdots \\ f_Q(z) \end{pmatrix}$$

Examples

- Find the parameters N_0 and λ of a radioactive material whose emissions are monitored over time $N(t) = N_0 e^{-\lambda t}$

$\tilde{N}_1 = N(t_1)$ measurements $\tilde{N}_i$ at t_i

$\vdots \qquad \vdots$

$\tilde{N}_q - t_q$

Toy example

(N_0, λ)

$f : \mathbb{R}^2 \to R^Q$ with Q large
$(N_i)_{i=1,\ldots,Q}$ radioactivity measurements at times $(t_i)_{i=1,\ldots,Q}$

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_Q(x) \end{pmatrix} = \begin{pmatrix} x_1 e^{-x_2 t_1} - N_1 \\ x_1 e^{-x_2 t_2} - N_2 \\ \vdots \\ x_1 e^{-x_2 t_Q} - N_Q \end{pmatrix} . = 0_{\mathbb{R}^q}$$

Calculate the Jacobian matrix $Jf(x)$

Toy example

$f : \mathbb{R}^2 \to R^Q$ with Q large

$(N_i)_{i=1,\ldots,Q}$ radioactivity measurements at times $(t_i)_{i=1,\ldots,Q}$

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_Q(x) \end{pmatrix} = \begin{pmatrix} x_1 e^{-x_2 t_1} - N_1 \\ x_1 e^{-x_2 t_2} - N_2 \\ \vdots \\ x_1 e^{-x_2 t_Q} - N_Q \end{pmatrix}.$$

Calculate the Jacobian matrix $Jf(x)$

$$Jf(x) = \begin{pmatrix} e^{-x_2 t_1} & -x_1 t_1 e^{-x_2 t_1} \\ e^{-x_2 t_2} & -x_1 t_2 e^{-x_2 t_2} \\ \vdots & \\ e^{-x_2 t_Q} & -x_1 t_Q e^{-x_2 t_Q} \end{pmatrix}.$$

Reminders: linear least squares

$Ax = b$ for $b \in \mathbb{R}^Q$ and $A \in \mathcal{M}_{Q,P}(\mathbb{R})$ with $Q > P$ and $\text{rg}(A) = P$.
The problem: find $x \in \mathbb{R}^P$ such that

$$\|Ax - b\|^2 = \min_{y \in \mathbb{R}^P} \|Ay - b\|^2$$

admits a unique solution given by the normal equation

$$A^t Ax = A^t b.$$

Nonlinear case

$\nabla \|f(x)\|^2 =$

$$\left\{\begin{array}{l}\text{Find } x^* \in \mathbb{R}^P \text{ such that} \\ \|f(x^*)\|^2 = \min_{x \in \mathbb{R}^P} \|f(x)\|^2 \quad \left(\|f(x)\|^2 = \sum_{k=1}^{Q} (f_k(x))^2\right),\end{array}\right.$$

We suppose that :

$$\forall x \in \mathbb{R}^P, \qquad J_f(x) \in \mathcal{M}_{Q,P}(\mathbb{R}) \text{ has rank } P.$$

In particular, we will have $\underbrace{(J_f(x))^t J_f(x)}_{\text{can be inverted}}$ symmetric defined positive.

Nonlinear case (continued)

We note

$$g : \begin{cases} \mathbb{R}^P & \to & \mathbb{R} \\ x & \mapsto & \|f(x)\|^2 \end{cases}$$

If g is strictly convex and coercive then the problem $g(x^*) = \min_x g(x)$ admits a unique solution x^*

$$\nabla g(x^*) = 0.$$

Calculating the gradient of g

$Jf(x) = \underset{q}{}\left(\quad \right)$ 2 columns

$g = \|f(x)\|^2 = N o f$, composition of
$N : \mathbb{R}^Q \to \mathbb{R}$, $N(y) = ||y||^2$ and $f : \mathbb{R}^P \to \mathbb{R}^Q$.
The rule for differentiating a composite function gives
$Dg(x) = DN(f(x))Df(x)$
For $y, \delta \in \mathbb{R}^Q$, $DN(y)\delta = 2y \cdot \delta$
For $x, h \in \mathbb{R}^P$, $Df(x)h = Jf(x)h \in \mathbb{R}^Q$

$$h, x \in \mathbb{R}^P, \quad Dg(x)h = 2f(x) \cdot Jf(x)h = 2Jf(x)^T f(x) \cdot h$$

$$\nabla g(x) = 2Jf(x)^T f(x).$$

Pay attention to the dimensions of the different terms. The order of the operations is important

Find the zeros of ∇g or the zeros of $f(x)$

- Zeroing $\nabla g(x) = \underbrace{2Jf(x)^T f(x)}$ with Newton method requires $Hf(x)$

$\nabla g(x) \in \mathbb{R}^p$ (toy problem $p=2$)

Find the zeros of ∇g or the zeros of $f(x)$

- Zeroing $\nabla g(x) = 2Jf(x)^T f(x)$ with Newton method requires $Hf(x)$
- If $f(x)$ is a function of $\mathbb{R}^P$ in $\mathbb{R}^P$ we find the zeros by Newton's algorithm

$$x_{k+1} = x_k + d_k$$

with $Jf(x_k)d_k = -f(x_k)$.

Here it is not possible because

$$Jf(x) \in \mathcal{M}_{Q \times P}$$

Find the zeros of ∇g or the zeros of $f(x)$

- Zeroing $\nabla g(x) = 2Jf(x)^T f(x)$ with Newton method requires $Hf(x)$
- If $f(x)$ is a function of $\mathbb{R}^P$ in $\mathbb{R}^P$ we find the zeros by Newton's algorithm

$$x_{k+1} = x_k + d_k$$

$$\text{with } Jf(x_k)d_k = -f(x_k).$$

- Here $f(x)$ is a function of $\mathbb{R}^P$ in $\mathbb{R}^Q$ so the system $Jf(x_k)d_k = -f(x_k)$ of size $Q \times P$ is solved in the least squares sense

$$Jf(x_k)^T Jf(x_k)d_k = -Jf(x_k)^T f(x_k)$$

$$\Leftrightarrow \quad d_k = -(Jf(x_k)^T Jf(x_k))^{-1} Jf(x_k)^T f(x_k).$$

Gauss Newton method

to minimize $\|f(x_k)\|^2$

$d_k = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \in \mathbb{R}^P$

$\|d_k\| > \varepsilon$

- Initialize $x_0 \in \mathbb{R}^P$
- While $\cancel{\|f(x_k)\|} > \varepsilon$ and $k < k_{\max}$
  - Solve $(Jf(x_k)^T Jf(x_k))d_k = -Jf(x_k)^T f(x_k)$ ← normal equation in the linear case
  - Update $x_{k+1} = x_k + d_k$
  - Update $k \to k+1$

Same pros and cons as Newton method

Convergence of the Gauss Newton method

We recall that $Jf(x)$ of rank P and $g(x)$ is strictly convex coercive

- Let $x_k \in \mathbb{R}^P$, then the direction $d_k = -(Jf(x_k)^T Jf(x_k))^{-1} Jf(x_k)^T f(x_k)$ satisfies

  $$\nabla g(x_k) \cdot d_k \leq 0.$$

  If $x_k \neq x^*$ then

  $$\nabla g(x_k) \cdot d_k < 0.$$

  So d_k is a descent direction for g at x_k.
- If the sequence $(x_k)_k$ converges, then its limit is x^*.

Experiment with notebook 4

toy problem
with radioactive
materials