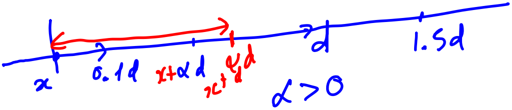
Descent method



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

$$f(x + \alpha d) < f(x) \quad \forall 0 < \alpha < \alpha_d.$$

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

$$\langle \nabla f(x), d \rangle < 0.$$

A good candidate for d is $-\nabla f(x)$

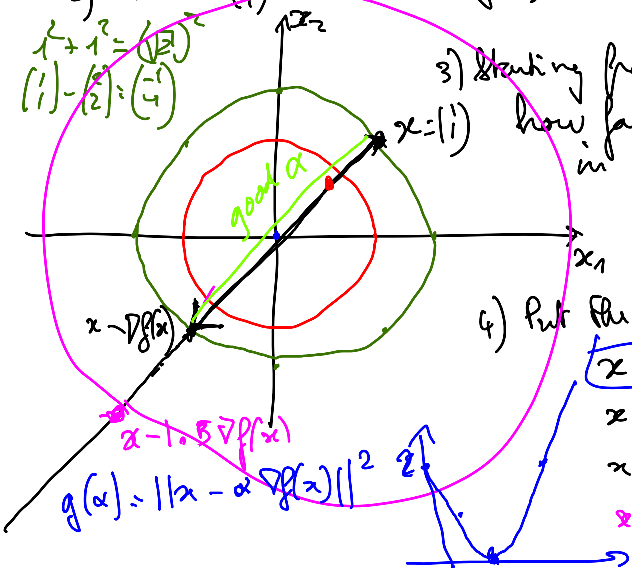
if $\nabla f(x) \neq 0$ $\quad -\|\nabla f(x)\|^2 < 0$

$$f(x) = \|x\|^2 = x_1^2 + x_2^2 \qquad f : \mathbb{R}^2 \to \mathbb{R}$$

1) Draw the level curves 0, 1, 2 of f

2) at $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ draw $-\nabla f(x)$ direction

- $f^{-1}(0)$
- $f^{-1}(1)$
- $f^{-1}(2)$



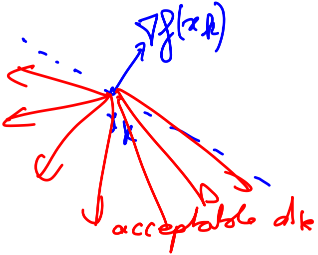
3) Starting from $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ how far can you go in $-\nabla f(x)$ direction such that

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

4) Put the points on the graph

- $x - \frac{1}{2}\nabla f(x)$
- $x - \frac{1}{4}\nabla f(x)$
- $x - \nabla f(x)$
- $x - 1.5\,\nabla f(x)$

General descent algorithm



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**

1) Choose descent direction d_k (such that $\langle \nabla f(x_k), d_k \rangle < 0$)
2) 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

$d_k = -\nabla f(x_k)$

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 $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2$
2. ∇f L-lipschitz $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|.$

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

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

The gradient method

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

converges geometrically for all initial guess, i.e.

$\|e_k\| \leq \beta^k \|e_0\|$

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

Proof: $x_k \to x^*$; $\nabla f(x^*) = 0$ $\quad e_{k+1}$ as a fxn of e_k

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

$$\|e_{k+1}\|^2 = \|\underbrace{x_k - x^*}_{a} - \underbrace{\alpha_k(\nabla f(x_k) - \nabla f(x^*))}_{b}\|^2 = \|a\|^2 - 2\langle a, b\rangle + \|b\|^2$$

$$= \|e_k\|^2 - 2\alpha_k \langle e_k, \nabla f(x_k) - \nabla f(x^*)\rangle + \|\nabla f(x_k) - \nabla f(x^*)\|^2 \alpha_k^2$$

where $e_k = x_k - x^*$, and $\langle e_k, \nabla f(x_k) - \nabla f(x^*)\rangle \geq \alpha \|x_k - x^*\|^2$; $\|\nabla f(x_k) - \nabla f(x^*)\|^2 \alpha_k^2 \leq L^2 \alpha_k^2 \|x_k - x^*\|^2$

$$\leq \|e_k\|^2 - 2\alpha_k \alpha \|x_k - x^*\|^2 + L^2 \alpha_k^2 \|x_k - x^*\|^2$$

$$\leq \|e_k\|^2 \underbrace{(1 - 2\alpha_k \alpha + L^2 \alpha_k^2)}_{< 1}$$

$$\|e_{k+1}\|^2 \leq \beta^2 \|e_k\|^2$$

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

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

$$\alpha_k > 0$$

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

$$\alpha_k < \frac{2\alpha}{L^2}$$

✓

Examples of possible choices for the descent direction

- Gradient Algorithm (*steepest descent*)

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

- Newton algorithm based on direction $\langle d_k, \nabla f(x_k)\rangle < 0$

or if $Hf(x_k) \in S^n_{++}$, close to the minimum x^* where we know that $Hf(x^*) \in S^n_{+}$

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

- Quasi-Newton with

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

where $W_k \approx Hf(x_k)$ we require that $W_k \in S^n_{++}$

- Conjugate gradient method (in the quadratic case) $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$

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

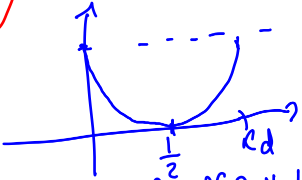
Choice of the step in a given direction

d_k is a descent direction : $\exists\, \alpha_{d_k}$; $\forall\, 0<\alpha\leq\alpha_{d_k}$ $f(x_k+\alpha d_k) < f(x_k)$

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

$$h_k : \alpha \mapsto h_k(\alpha) = f(x_k + \alpha d_k)$$

$h_k'(\alpha) = \langle d_k, \nabla f(x_k + \alpha d_k)\rangle$



for $f(x) = \|x\|^2$ and $x = \begin{pmatrix}1\\1\end{pmatrix}$ $d = -\nabla f(x)$

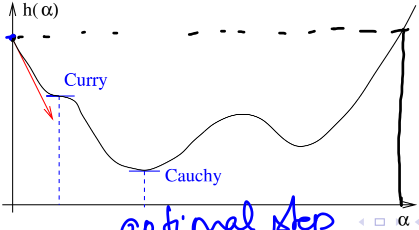
Cauchy's rule

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

Curry's rule

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

$h(0) = f(x)$



optimal step

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 → find 0 of $h'(x)$

satisfies

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

Optimal step in the quadratic case

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

$f(x^*) = \min_x f(x)$

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

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

$$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle,$$

choose d, s.t. $\langle \nabla f(x), d\rangle < 0$

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

$h : \mathbb{R} \to \mathbb{R}$

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

$\frac{1}{2}\langle A(x+\alpha d), x + \alpha d\rangle + \langle b, x + \alpha d\rangle = \frac{1}{2}\langle Ax, x\rangle + \frac{1}{2}\langle Ax, \alpha d\rangle + \frac{1}{2}\langle \alpha d, x\rangle + \frac{1}{2}\langle \alpha d, \alpha d\rangle + \langle b, x\rangle + \langle b, \alpha d\rangle$

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

$h'(\alpha) = \alpha\langle Ad, d\rangle + \langle Ax + b, d\rangle$

$$\alpha^\star = -\frac{\langle g, d\rangle}{\langle Ad, d\rangle}$$

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

$\alpha^* \langle Ad, d\rangle + \langle Ax + b, d\rangle = 0 \Rightarrow \alpha^* = \frac{-\langle Ax+b, d\rangle}{\langle Ad, d\rangle}$

Optimal step gradient method in the quadratic case

$f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$ $A \in S^n_{++}$ $b \in \mathbb{R}^n$

at each step $d_k = -\nabla f(x_k) = -(Ax_k + b) = -g_k$

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**

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

end
$x^\star \leftarrow x_k$

$-\langle g, d\rangle = \langle d, d\rangle$

$\langle v_k, d_k\rangle = \langle A d_k, d_k\rangle$

$\Leftrightarrow$ $g_{k+1} = Ax_{k+1} + b = A(x_k + \alpha_k d_k) + b = \underbrace{Ax_k + b}_{g_k} + \alpha_k \underbrace{Ad_k}_{v_k}$

Choice of the step in the general case (f not quadratic)

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. $\langle \nabla f(x_k), d_k \rangle < 0)$
  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
$x^\star \leftarrow x_k$

provides an acceptable α (not the best)

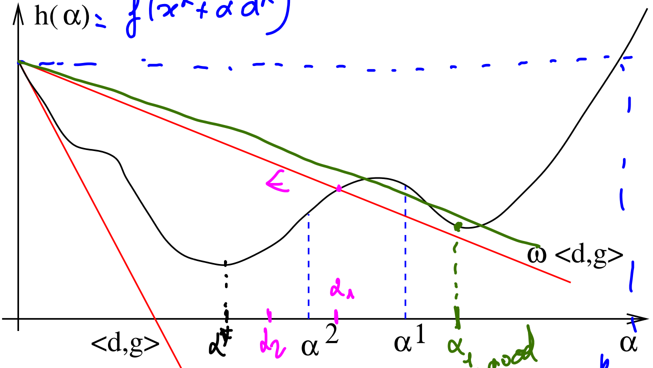
Directional minimisation -Line search

choose α_k? Backtracking Line Search BLS

Armijo's rule linearization of the constraint on α_k

$\omega \in]0,1[$

$$f(x^k + \alpha_k d^k) < f(x^k) + \omega \alpha_k \langle g^k, d^k \rangle$$



tangent to h(α) at α=0

$\alpha^k = \tau \alpha^{k-1}$ with $\tau \in]0,1[$

Armijo's rule / BLS

Red line equation (α, y)
$y = f(x) + \omega \alpha \langle d, \nabla f(x) \rangle$
you want $h(\alpha) < y$

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 $f(x + \alpha_k d) \geqslant f(x) + \omega \alpha_k \langle d, \nabla f(x) \rangle$ **do**

  Choose $\alpha_{k+1} = \tau \alpha_k$ α_{k+1} will be smaller and smaller

  $k \leftarrow k + 1$

end

$\alpha = \alpha_k$

Exercice

Consider the minimisation problem

$$x^\star = \underset{\mathbb{R}^2}{\operatorname{argmin}} f(x), \quad \text{with } f(x) = (x_1 - 1)^4 + (x_2 - 2)^4$$

1. Solve the problem exactly.
2. Write Newton algorithm for this problem. Show that it converges for any $x^0 \in \mathbb{R}^2$.

Consider now the gradient algorithm for this problem. In order to select the step α_k in the gradient direction

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

we consider the function $\varphi^k(\rho) = f(x^k - \rho \nabla f(x^k))$ and $p^k(\rho)$ polynomial of degree 2 that interpolates $\varphi(0)$, $\varphi'(0)$ and $\varphi''(0)$.

3. Define $p^k(\rho)$ from the values $\varphi(0)$, $\varphi'(0)$ and $\varphi''(0)$
4. Compute $\rho^k = \operatorname{argmin}_{\rho>0} p^k(\rho)$
5. Show that there exists $\alpha > 0$ s.t. if $\alpha_k = \alpha \rho_k$ then

$$\|x^{k+1} - x^\star\| \leq \mu \|x^k - x^\star\| \text{ with } 0 < \mu < 1$$

6. Conclude that $\lim_{k\to\infty} x^k = x^\star$ and compute the error estimate $\|x^0 - x^\star\|$.

Exercice

$f(x) = (x_1 - 1)^4 + (x_2 - 2)^4$ $\quad f : \mathbb{R}^2 \to \mathbb{R}$

$x^* = (1, 2) \Leftarrow (f(x) \geq 0$ and $f(x^*) = 0)$

find $f(x^*) = \min f(x)$

$G(x) = \nabla f(x)$ $\quad JG(x) = Hf(x)$, write Newton: $X_{k+1} = X_k - JG(X_k)^{-1} G(X_k)$

$$G(x) = \begin{pmatrix} 4(x_1 - 1)^3 \\ 4(x_2 - 2)^3 \end{pmatrix} \qquad JG(x) = \begin{pmatrix} 12(x_1 - 1)^2 & 0 \\ 0 & 12(x_2 - 2)^2 \end{pmatrix}$$

$$JG(x)^{-1} = \begin{pmatrix} \frac{1}{12(x_1 - 1)^2} & 0 \\ 0 & \frac{1}{12(x_2 - 2)^2} \end{pmatrix} \qquad JG(x) \cdot JG(x)^{-1} = Id$$

Compute X_{k+1} knowing X_k

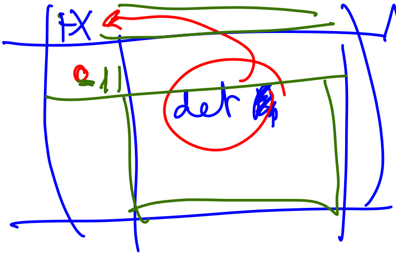
Exercice

inverse of a 2x2 matrix:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{D} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

with $D = a_{11} a_{22} - a_{12} a_{21}$

General case: use the cofactor:



M^{-1} = transpose of the cofactor matrix divided by the determinant.

Exercice

$$\nabla f(x) = 4\begin{pmatrix}(x_1-1)^3\\(x_2-2)^3\end{pmatrix} = G(x) \qquad JG(x) = Hf(x) = 12\begin{pmatrix}(x_1-1)^2 & 0\\ 0 & (x_2-1)^2\end{pmatrix}$$

$$JG^{-1}(x) = \frac{1}{12}\begin{pmatrix}\frac{1}{(x_1-1)^2} & 0\\ 0 & \frac{1}{(x_2-2)^2}\end{pmatrix}$$

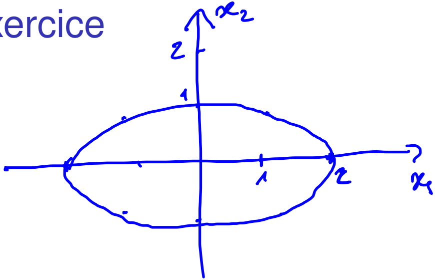
$$x - JG^{-1}(x)G(x) = \begin{pmatrix}x_1\\x_2\end{pmatrix} - \frac{4}{12}\begin{pmatrix}\frac{1}{(x_1-1)^2} & 0\\ 0 & \frac{1}{(x_2-2)^2}\end{pmatrix}\begin{pmatrix}(x_1-1)^3\\(x_2-2)^3\end{pmatrix} = \begin{pmatrix}x_1\\x_2\end{pmatrix} - \frac{1}{3}\begin{pmatrix}x_1-1\\x_2-2\end{pmatrix}$$

$$x_{k+1} = \begin{pmatrix}x_{k_1}\\x_{k_2}\end{pmatrix} - \frac{1}{3}\begin{pmatrix}x_{k_1}-1\\x_{k_2}-2\end{pmatrix} \qquad x_{k+1} - x^* = \begin{pmatrix}x_{k_1}\\x_{k_2}\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix} - \frac{1}{3}\begin{pmatrix}x_{k_1}-1\\x_{k_2}-2\end{pmatrix} = \frac{2}{3}\begin{pmatrix}x_{k_1}-1\\x_{k_2}-2\end{pmatrix}$$

$$x_{k+1} - x^* = \frac{2}{3}(x_k - x^*)$$

$$\|x_{k+1} - x^*\| = \frac{2}{3}\|x_k - x^*\| \Rightarrow \|x_k - x^*\| = \left(\frac{2}{3}\right)^k \|x_0 - x^*\| \xrightarrow[k\to\infty]{} 0 \quad \forall x_0$$

Exercice



$f(x) = x_1^2 + 2x_2^2$

1) which level curve is drawn?

2) set $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
compute $-\nabla f(x)$
set $x - \nabla f(x)$ on the graph and draw the line

3) compute α_{max} such that $f(x - \alpha \nabla f(x)) < f(x)$ for $0 < \alpha < \alpha_{max}$