Tema 4. Optimització No Lineal sense Restriccions

Grau en Matemàtiques

Programació Matemàtica

Jordi Castro Javier Heredia Josep Homs

Departament d'Estadística i Investigació Operativa
Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya
Barcelona

Esquema del tema

1. Conceptes bàsics
   - Formulació de problemes d'optimització no lineals
   - Exemples de problemes
   - Conceptes previs
   - Convexitat: funcions, conjunts i problemes convexos
   - Breu ressenya històrica
2. Optimització no lineal sense restriccions
   - Condicions d'optimalitat
   - Mètodes basats en exploracions lineals: gradient i Newton
   - Exploració lineal. Convergència global
   - Mètode del gradient. Convergència local
   - Mètode Newton. Convergència local
   - Modificacions al mètode de Newton
3. Bibliografia

Problemes d'optimització no lineal I

- Formulació general de problemes d'ONL:

$$\begin{array}{ll} \min & f(x) \\ \text{s.a} & x \in \Omega \end{array}$$

on $f : \mathbb{R}^n \rightarrow \mathbb{R}$ és la funció objectiu (suposem "suau", $f \in \mathcal{C}^2$) i $\Omega \subseteq \mathbb{R}^n$ és el conjunt factible.

- Si $\Omega = \mathbb{R}^n$ tenim un problema sense restriccions.
- Si $\Omega = \{x : h_i(x) = 0, g_j(x) \leq 0, i = 1, \ldots, m,\ j = 1, \ldots, p\}$ ($h_i : \mathbb{R}^n \rightarrow \mathbb{R}, i = 1, \ldots, m$, $g_j : \mathbb{R}^n \rightarrow \mathbb{R}, j = 1, \ldots, p$, suposem també "suaus": $h_i, g_j \in \mathcal{C}^2$) tenim un problema amb restriccions d'igualtat i/o desigualtat.

Problemes d'optimització no lineal II

- El problema d'ONL general és:

$$\begin{array}{lll} \min & f(x) & \\ \text{s.a} & h(x) = 0 & [h_i(x) = 0 \quad i = 1, \ldots, m] \\ & g(x) \leq 0 & [g_j(x) \leq 0 \quad j = 1, \ldots, p]. \end{array}$$

- Considerarem problemes de $\min$. Notem que

$$\max f(x)\ x \in \Omega \quad \equiv \quad -\min -f(x)\ x \in \Omega$$

(Clarament: $f(x^*) \geq f(x)\ \forall x \in \Omega \Longleftrightarrow -f(x^*) \leq -f(x)\ \forall x \in \Omega$).

Problemes amb i sense restriccions

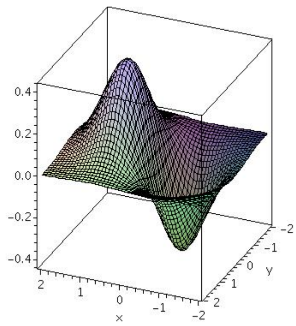
- Els mètodes de resolució també es classifiquen en mètodes per a problemes d'ONL amb i sense restriccions.
- És una de les principals diferències amb PLs:

- Els PLs sense restriccions són il.limitats:

$$-\infty = \min \quad c^T x$$

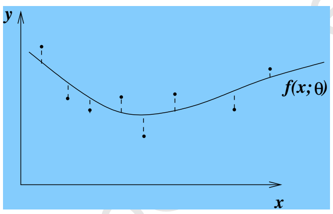
$$+\infty = \max \quad c^T x$$

- En canvi els problemes d'ONL sense restriccions poden tenir mínim i màxim. P.e., $f(x,y) = x \cdot e^{-x^2-y^2}$:



Ajust de funcions per mínims quadrats no lineals

- Ajustos de corbes: usats en Estadística, en enginyeries química, mecànica, etc
- Tenim una sèrie de m observacions $(x_i, y_i), i = 1, \ldots, m$, $x_i \in \mathbb{R}^d$ i $y_i \in \mathbb{R}^l$. Per exemple per $d = l = 1$ podríem tenir



- Volem ajustar una funció $f(x;\theta) : \mathbb{R}^d \to \mathbb{R}^l$, depenent d'un vector de paràmetres $\theta \in \mathbb{R}^n$. Per exemple,

$$d = l = 1 \quad n = 3 \qquad f(x;\theta) = \theta_1 + \sin(x^2/\theta_2) + e^{x/\theta_3}.$$

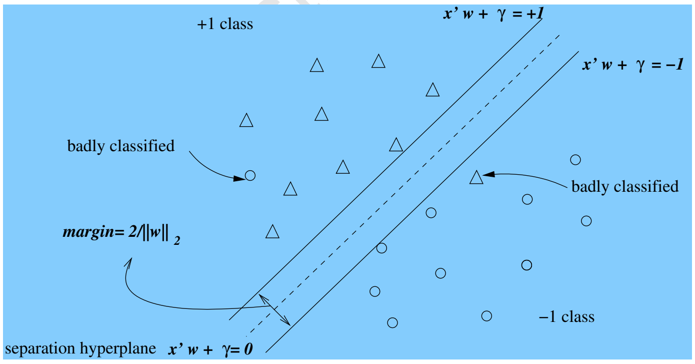
- Definim els residus $r_i = f(x_i;\theta) - y_i$, $i = 1, \ldots, m$ i formulem

$$\min_{\theta\in\mathbb{R}^n} \sum_{i=1}^m \|r_i\|_2^2 \qquad \Longleftrightarrow \qquad \min_{\theta\in\mathbb{R}^n} \sum_{i=1}^m (f(x_i;\theta) - y_i)^T (f(x_i;\theta) - y_i)$$

- Altres aplicacions dels "mínims quadrats no lineals": Google els usa per obtenir models 3D a partir de fotografies, a Street View per estimar les posicions dels cotxes i satèl·lits a partir d'observacions de sensors, etc. Google usa el paquet Ceres, visiteu http://ceres-solver.org/ per a més informació.

“Support Vector Machines” (SVM I)

- Donats els m parells $(x_i, y_i) \in \mathbb{R}^n \times \{+1, -1\}$, $i = 1, \ldots, m$ busquem l’hiperplà definit per $(w, \gamma) \in \mathbb{R}^{n+1}$ amb “marge de separació” màxim entre els hiperplans paral·lels $w^T x + \gamma \geq +1$ i $w^T x + \gamma \leq -1$.
- w és él vector normal al pla de separació, γ determina la seva posició respecte l’origen.
- Veiem que el marge de separació entre plans és $\dfrac{2}{||w||_2}$.



El marge de separació és $\frac{2}{||w||_2}$ (SVM II)

- Donats dos hiperplans paral·lels $w^T x = a$ i $w^T x = b$, i x_1 un punt del primer pla ($w^T x_1 = a$), el punt més proper a x_1 al segon hiperplà, anomenat x_2, ($w^T x_2 = b$) pot ser escrit com $x_2 = x_1 + \alpha w$.
- Llavors:

$$
\begin{aligned}
x_2 &= x_1 + \alpha w \\
w^T x_2 &= w^T x_1 + \alpha w^T w \\
b &= a + \alpha ||w||_2^2 \\
\alpha &= \frac{b-a}{||w||_2^2}
\end{aligned}
$$

- El marge de separació és la norma Euclídea de αw:

$$
||\alpha w||_2 = |\alpha| \cdot ||w||_2 = \frac{|b-a|}{||w||_2^2} ||w||_2 = \frac{|b-a|}{||w||_2}
$$

- A la SVM, $a = -\gamma - 1$ i $b = -\gamma + 1$, per tant el marge és $\dfrac{2}{||w||_2}$.

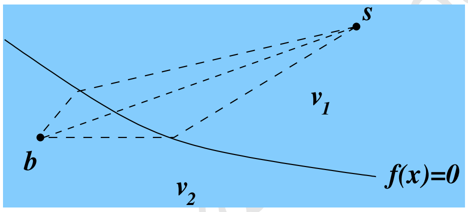
Problema d'optimització quadràtic (SVM III)

- L'objectiu és maximitzar el marge, i a la vegada minimitzar els errors de classificació $\sum_{i=1}^m s_i$. Aquest dos objectius oposats es ponderen amb el paràmetre $\nu \in \mathbb{R}$.
- Maximitzar el marge és equivalent a $\min \frac{1}{2}||w||_2$, que és equivalent a $\min \frac{1}{2}||w||_2^2 = \min \frac{1}{2} w^T w$.
- SVM és un problema d'optimització quadràtic en variables w, γ, s:

$$\begin{array}{lll} \min\limits_{(w,\gamma,s)\in\mathbb{R}^{n+1+m}} & \frac{1}{2}w^T w + \nu \sum\limits_{i=1}^m s_i & \\ \text{s. a} & y_i(w^T x_i + \gamma) + s_i \geq 1 & i = 1,\ldots,m \\ & s_i \geq 0 & i = 1,\ldots,m \end{array}$$

Problemes de distàncies: La socorrista i el banyista

- La socorrista i un banyista que s'està ofegant es troben en un pla als punts $s = (s_1, s_2)$ i $b = (b_1, b_2)$, $s, b \in \mathbb{R}^2$.



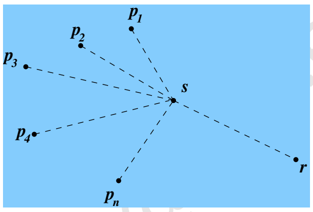
La frontera entre la platja i el mar són els punts $x : f(x) = 0$.

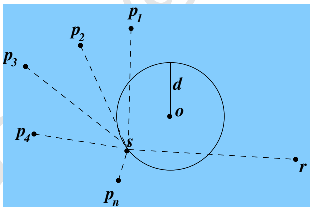
- La socorrista corre per la sorra i neda a unes velocitats $v_1 m/s$ i $v_2 m/s$, respectivament.
- Quin recorregut ha de fer per arribar al banyista en el temps mínim?
- Cal trobar $x^* \in \mathbb{R}^2$ que proporciona el temps mínim i verifica $f(x^*) = 0$:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^2} & t_1(x) + t_2(x) \\ \text{s. a} & f(x) = 0 \end{array} \quad \text{on} \quad \begin{array}{l} t_1(x) = \dfrac{||x-s||_2}{v_1} = \dfrac{\sqrt{\sum_{i=1}^2 (x_i - s_i)^2}}{v_1} \\ t_2(x) = \dfrac{||b-x||_2}{v_2} = \dfrac{\sqrt{\sum_{i=1}^2 (b_i - x_i)^2}}{v_2}. \end{array}$$

Problemes de distàncies: L'oleoducte

- Es vol construir un oleoducte per comunicar n pous de petroli de coordenades $p_i = (x_i, y_i) \in \mathbb{R}^2$, amb una refineria de coordenades $r = (x_r, y_r) \in \mathbb{R}^2$.





Determineu l'oleoducte de longitud mínima.

- La millor solució passa per unir els diferents pous en una subestació $s = (x_s, y_s) \in \mathbb{R}^2$. Per trobar la millor situació de s formulem el problema sense restriccions:

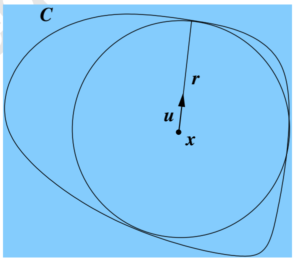
$$\min_{s\in\mathbb{R}^2} \sum_{i=1}^{n} \|s - p_i\|_2 + \|r - s\|_2 = \sum_{i=1}^{n} \sqrt{(x_s - x_i)^2 + (y_s - y_i)^2} + \sqrt{(x_r - x_s)^2 + (y_r - y_s)^2}.$$

- I si s ha d'estar a més de d Km. d'un punt $o = (x_o, y_o)$? Cal afegir la restricció

$$\|o - s\|_2^2 \geq d^2 \quad \Longleftrightarrow \quad (x_o - x_s)^2 + (y_o - y_s)^2 \geq d^2.$$

Càlcul de centres: el centre de Chebyshev (conjunt convex)

- El centre de Chebyshev d'un conjunt afitat $C \subseteq \mathbb{R}^n$ és el centre del cercle de major radi que es pot inscriure dintre de C.
- Per a un C qualsevol és un problema difícil de solucionar; considerem el cas C convex.



- C convex definit per $C = \{x : f_1(x) \leq 0, \ldots, f_m(x) \leq 0\}$, $f_i(x)$ funcions convexes.
- Formulem el següent problema, en variables $x \in \mathbb{R}^n$ i $r \in \mathbb{R}$:

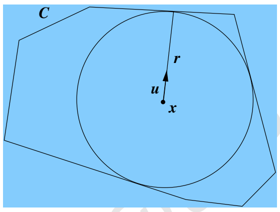
$$\begin{array}{ll} \max\limits_{x,r} & r \\ \text{s. a} & g_i(x,r) \leq 0 \quad i = 1, \ldots, m \\ & r \geq 0 \end{array} \qquad \text{on} \qquad g_i(x,r) = \max_{\|u\|_2 \leq 1} f_i(x + r\,u)$$

- Aquest problema "binivell" ("max" a restriccions) pot reformular-se com

$$\begin{array}{ll} \max\limits_{x,r,u^i\ i=1,\ldots,m} & r \\ \text{s. a} & \max f_i(x + r\,u^i) \leq 0 \quad i = 1, \ldots, m \\ & \|u^i\|_2 \leq 1 \quad i = 1, \ldots, m \\ & r \geq 0 \end{array}$$

Càlcul de centres: el centre de Chebyshev (políedre)

- C és ara el políedre definit per $C = \{x : a_i^T x \le b_i, \quad i = 1, \dots, m\}$?



- Quan C era conjunt convex teníem:

$$\begin{array}{ll} \max\limits_{x,r} & r \\ \text{s. a} & g_i(x,r) \le 0 \quad i = 1, \dots, m \\ & r \ge 0 \end{array} \qquad \text{on} \qquad g_i(x,r) = \max_{\|u\|_2 \le 1} f_i(x + r\,u)$$

- Ara $f_i(x) = a_i^T x - b_i$, i com que $\|u\|_2 \le 1$ i $a_i^T u = \|a_i\|_2 \|u\|_2 \cos(a_i, u) \le \|a_i\|_2$:

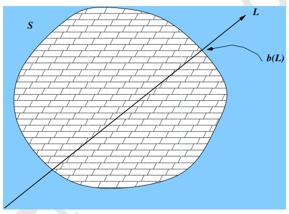
$$g_i(x,r) = \max_{\|u\|_2 \le 1} a_i^T (x + r\,u) - b_i = a_i^T x - b_i + r \max_{\|u\|_2 \le 1} a_i^T u = a_i^T x - b_i + r\|a_i\|_2.$$

- Per tant tenim un problema lineal:

$$\begin{array}{ll} \max\limits_{x,r} & r \\ \text{s. a} & a_i^T x + r\|a_i\|_2 \le b_i \quad i = 1, \dots, m \\ & r \ge 0 \end{array}$$

Tomografia computada: reconstrucció d'imatges I

- Tenim un conjunt $S \in \mathbb{R}^3$ que representa una part del cos humà. El particionem en $\{1, 2, \dots, n\}$ cel·les o "voxels". Considerem densitat homogènea dintre de cada cel·la.
- Enviem un raig L a través de S (són raigs X)



i recollim les següents dades:

  - $I(L)$: subconjunt de cel·les travessades per L.
  - $a_i(L)$: longitud del recorregut de L dintre cel·la i.
  - $b(L)$: atenuació de l'energia de L mesurada al punt de sortida.

- Objectiu: calcular les densitats x_i a cada cel·la $i = 1, \dots, n$, sabent que verifiquen

$$\begin{array}{l} \sum\limits_{i \in I(L)} a_i(L) x_i = b(L) \\ x_i \ge 0 \quad i = 1, \dots, n. \end{array}$$

Tomografia computada: reconstrucció d'imatges II

- A la pràctica s'emeten diversos raigs $L_j, j = 1, \ldots, m$, de forma que tenim:

$$\sum_{i \in I(L_j)} a_i(L_j)x_i = b(L_j) \quad j = 1, \ldots, m$$
$$x_i \geq 0 \quad i = 1, \ldots, n.$$

- Com que $n \gg m$ podem tenir moltes solucions; o fins i tot cap solució $x \geq 0$ per errors de mesura. Per tant a la pràctica se soluciona:

$$\min f(x) \triangleq \sum_{j=1}^{m} \left(\sum_{i \in I(L_j)} a_i(L_j)x_i - b(L_j) \right)^2 \quad \text{s. to } \; x \geq 0.$$

- En cas de solucions alternatives, busquem la de mínima densitat:

$$\min f(x) + \delta \sum_{i=1}^{n} x_i \quad \text{s. to } \; x \geq 0.$$

Gradient, Hessiana i Jacobiana

- $f : D \subseteq \mathbb{R}^n \rightarrow I \subseteq \mathbb{R}$, $f \in \mathcal{C}^2$. Vector gradient i matriu Hessiana:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \qquad \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} = \nabla^2 f(x)^T$$

- $g : D \subseteq \mathbb{R}^n \rightarrow I \subseteq \mathbb{R}^m$, $g = (g_1, \ldots, g_m)^T$, $g_i \in \mathcal{C}$. Matriu Jacobiana:

$$\nabla g(x)^T = J(x) = \begin{bmatrix} \nabla g_1(x)^T \\ \vdots \\ \nabla g_m(x)^T \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_n} \end{bmatrix}$$

Gradient, Hessiana i Jacobiana: exemple

- $f(x) = e^{x_1^2+x_2}$

$$\nabla f(x) = \begin{bmatrix} 2x_1 e^{x_1^2+x_2} \\ e^{x_1^2+x_2} \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 2e^{{x_1}^2+x_2} + 4{x_1}^2 e^{{x_1}^2+x_2} & 2x_1 e^{{x_1}^2+x_2} \\ 2x_1 e^{{x_1}^2+x_2} & e^{{x_1}^2+x_2} \end{bmatrix}$$

- $g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ 3x_1 + 4x_2 - 5 \end{bmatrix}$

$$\nabla g(x)^T = J(x) = \begin{bmatrix} \nabla g_1(x)^T \\ \nabla g_2(x)^T \end{bmatrix} = \begin{bmatrix} 2x_1 & 2x_2 \\ 3 & 4 \end{bmatrix}$$

Teorema de Taylor (amb residu, sense residu)

- $f : D \subseteq \mathbb{R} \to I \subseteq \mathbb{R}$, $f \in \mathcal{C}^k$, $\alpha \in \mathbb{R}$.

$$f(x+\alpha) = f(x) + f'(x)\alpha + \frac{f''(x)}{2!}\alpha^2 + \cdots + \frac{f^{k)}(x)}{k!}\alpha^k + R_k(\alpha) \qquad R_k(\alpha) = o(\alpha^k)$$

$$f(x+\alpha) = f(x) + f'(x)\alpha + \frac{f''(x)}{2!}\alpha^2 + \cdots + \frac{f^{k)}(z)}{k!}\alpha^k \qquad z = x + \theta\alpha, 0 \le \theta \le 1$$

- $f : D \subseteq \mathbb{R}^n \to I \subseteq \mathbb{R}$, $f \in \mathcal{C}^2$, $d \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$.
  - Polinomi d'ordre 1:

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + R_1(\alpha) \qquad R_1(\alpha) = o(\alpha)$$

$$f(x+\alpha d) = f(x) + \alpha \nabla f(z)^T d \qquad z = x + \theta\alpha d, 0 \le \theta \le 1$$

  - Polinomi d'ordre 2:

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + \frac{1}{2}\alpha^2 d^T \nabla^2 f(x) d + R_2(\alpha) \qquad R_2(\alpha) = o(\alpha^2)$$

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + \frac{1}{2}\alpha^2 d^T \nabla^2 f(z) d \qquad z = x + \theta\alpha d, 0 \le \theta \le 1$$

- Notació: $f(t) = o(g(t)$ si $\lim_{t\to 0} \frac{f(t)}{g(t)} = 0$ ($f(t)$ és infinitèssim d'ordre superior).

Existència d'òptim

Teorema (Teorema de Weierstrass d'existència d'òptim)

Si $f : \Omega \to \mathbb{R}$ és contínua i $\Omega \subseteq \mathbb{R}^n$ és compacte (tancat i afitat) $\exists x^* \in \Omega : f(x^*) \le f(x)\ \forall x \in \Omega$ (és a dir $\min_{x\in\Omega} f(x)$ té solució).

Alguns exemples "sense mínim":

- $f(x) = 1/x$, $\Omega = \{x \in \mathbb{R} : x \ge 0\}$: no té mínim ($f(x) \to 0$ si $x \to +\infty$). Falla que Ω no és afitat.
- $f(x) = -x^2$, $\Omega = \{x \in \mathbb{R} : 0 \le x < 1\}$: no té mínim (però si ínfim, que és -1). Falla que Ω no és tancat.
- En el curs ens centrarem en problemes que tenen mínim, i Ω sempre serà tancat.

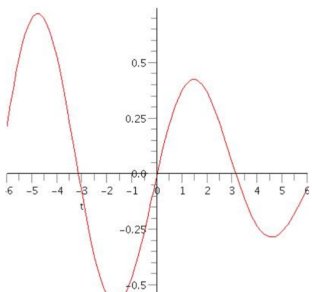
Mínim global i local

Definició

Donat el problema d'optimització
$$\begin{array}{ll} \min & f(x) \\ s.a & x \in \Omega \end{array}$$

- *diem que x^* és mínim (òptim) local si $x^* \in \Omega$ i hi ha un entorn B de x^* (conjunt obert que conté x^*) tal que $f(x^*) \le f(x)$ per a tot $x \in B \cap \Omega$ (es diu estricte si $f(x^*) < f(x)$).*
- *diem que x^* és mínim (òptim) global si $x^* \in \Omega$ i $f(x^*) \le f(x)$ $\forall x \in \Omega$.*

P.e., $f(x) = \frac{\sin x}{\left(1+\left(e^{x/10}\right)^2\right)}$, $-6 \le x \le 6$, té 3 màxims locals i 3 mínims locals, i un d'ells és respectivament màxim i mínim global



Tipus “exòtic” de mínim local: mínim local aïllat

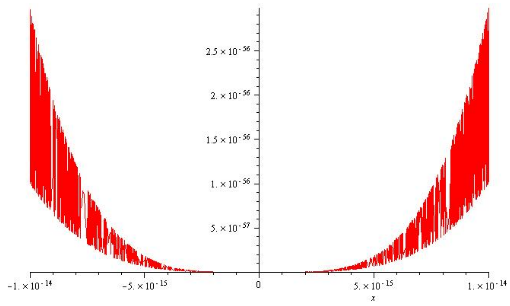
Definició

Diem que x^* és mínim local aïllat si hi ha un entorn B de x^* on x^* és l'únic mínim.

- Els mínims local estrictes no són sempre aïllats. P.e.,

$$f(x) = \begin{cases} x^4 \cos(\frac{1}{x}) + 2x^4 & x \neq 0 \\ 0 & x = 0 \end{cases}$$

  té un mínim local estricte a $x^* = 0$, però no és mínim aïllat: hi ha molts mínims locals estrictes tan a prop com vulguem de $x^* = 0$.



- Però tots els mínims locals aïllats són mínims locals estrictes (es deixa com exercici).

Cerca d'òptims globals

- Sempre es desitja obtenir l'òptim global, però a la pràctica això no és possible
- Els mètodes numèrics de ONL es basen en condicions d'optimalitat locals (són “miops”)
- Optimització global: branca de la ONL que cerca òptims globals. Actualment no hi ha cap mètode efectiu que garanteixi òptims globals per a qualsevol problema
  - Si existís, s'hauria acabat la dificultat de la PE: sabeu formular $x \in \{0, 1\}$ com a restricció no lineal?:
  - $x \in \{0, 1\} \Leftrightarrow x(x-1) = 0$
- Però per a un tipus de problemes sí podem garantir òptim globals: problemes convexos.

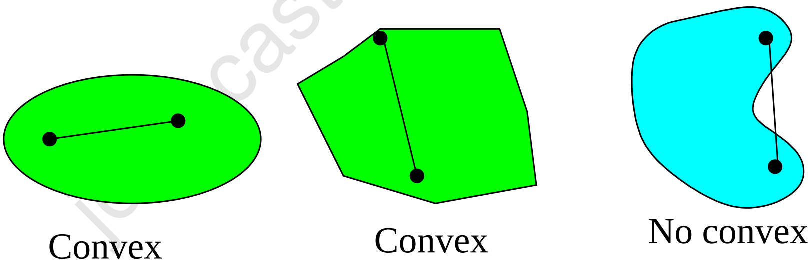
Conjunt convex

Definició

Ω és conjunt convex si per a tot $x, y \in \Omega$

$$\alpha x + (1-\alpha) y \in \Omega \quad 0 \le \alpha \le 1$$

- Gràficament, això vol dir que el segment $\overline{xy}$ pertany a Ω:



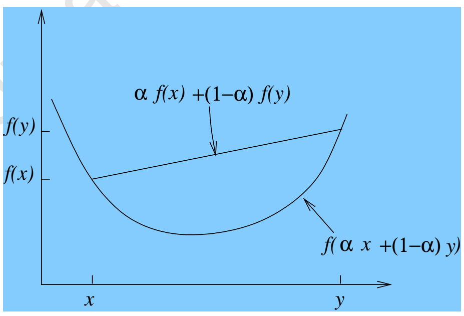
Funció convexa

Definició

f és convexa en $\Omega \subseteq \mathbb{R}^n$ si per a tot $x, y \in \Omega$

$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y) \quad 0 \le \alpha \le 1$$

- Si la desigualtat anterior és estricta (per $x \neq y$) f és estrictament convexa.
- Gràficament:



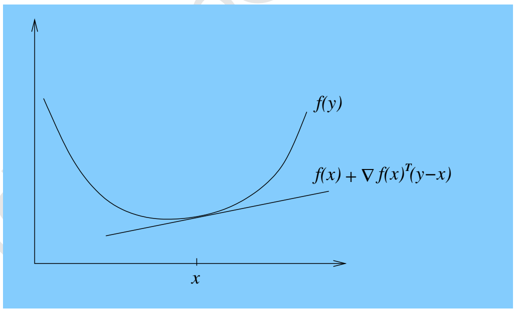
Caracterització de funcions convexes usant ∇f

Proposició

$f \in \mathcal{C}^1$ és convexa en $\Omega \subseteq \mathbb{R}^n$ convex si i només si

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall\, x, y \in \Omega$$

- Gràficament:



- Hi ha convexitat estricta si i només si la desigualtat és $>$ per $x \neq y$.

Recordatori: matrius (semi)definides positives

Definició

Una matriu simètrica $H \in \mathbb{R}^{n\times n}$ és semidefinida positiva (definida positiva) si per a tot $x \in \mathbb{R}^n$, $x \neq 0$, $x^T Hx \geq 0$ ($x^T Hx > 0$).

- Una matriu és semidefinida positiva (definida positiva) si i només si:
  - Tots els seus valors propis són ≥ 0 (> 0).
  - Tots els *menors principals* són ≥ 0 (els *n menors principals dominants* són > 0).
- Exemple:

$$H = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

  - $x = [x_1\ x_2]^T$, $x^T Hx = 2x_1^2 + 2x_1x_2 + x_2^2 = x_1^2 + (x_1 + x_2)^2 > 0$.
  - Primer menor: $\Delta_1 = |2| > 0$. Segon menor: $\Delta_2 = |H| = 2 - 1 = 1 > 0$.
  - Calculem valors propis $|H - \lambda I| = 0$.

$$|H-\lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = (2-\lambda)(1-\lambda)-1 = \lambda^2-3\lambda+1 = 0 \qquad \lambda = \frac{3 \pm \sqrt{5}}{2} > 0.$$

Caracterització de funcions convexes usant $\nabla^2 f$

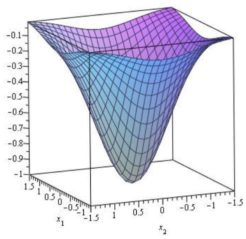
Proposició

$f \in \mathcal{C}^2$ és convexa en $\Omega \subseteq \mathbb{R}^n$ convex i que conté com a mínim un punt interior si i només si $\nabla^2 f$ és semidefinida positiva en Ω.

- Si $\nabla^2 f \succ 0$ la convexitat és estricta (condició suficient, no necessària).
- Exemple: $f(x) = -e^{-({x_1}^2+{x_2}^2)}$

$$\nabla f(x) = \begin{bmatrix} 2\,x_1 e^{-({x_1}^2+{x_2}^2)} \\ 2\,x_2 e^{-({x_1}^2+{x_2}^2)} \end{bmatrix} \quad \nabla^2 f(x) = e^{-({x_1}^2+{x_2}^2)} \begin{bmatrix} 2-4{x_1}^2 & -4\,x_1 x_2 \\ -4\,x_1 x_2 & 2-4\,{x_2}^2 \end{bmatrix}$$

- $\Delta_1 = e^{-({x_1}^2-{x_2}^2)}(2-4x_1^2)$.
- $\Delta_2 = e^{-({x_1}^2-{x_2}^2)}(4-8(x_1^2+x_2^2))$.
- f convexa a $\Omega^1 = \{x \in \mathbb{R}^2 : |x_1| \le \frac{1}{2}, |x_2| \le \frac{1}{2}\}$.
- f convexa a $\Omega^2 = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \le \frac{1}{2}\}$
- f no convexa a $x \in \mathbb{R}^2 \setminus \Omega^2$.



Caracterització de conjunts convexos I

- En general no és fàcil determinar si un conjunt $\Omega = \{x : h(x) = 0, g(x) \le 0\}$ és convex.
- Però en alguns casos és possible:

Proposició

Sigui $g : D \to \mathbb{R}$ (D convex) una funció convexa. Llavors el conjunt $S_c = \{x : x \in D, g(x) \le c\}$ és convex per a tot $c \in \mathbb{R}$.

Demostració.

(Es deixa com exercici) □

Caracterització de conjunts convexos II

- Com a conseqüència del resultat anterior el conjunt $\Omega = \{x : g_j(x) \leq 0, j = 1, \ldots, p\}$ és convex si g_j són funcions convexes, donat que la intersecció de conjunt convexos és un conjunt convex (fàcil de provar, es deixa com exercici).
- El conjunt $\Omega = \{x : h_i(x) = 0, i = 1, \ldots, m\}$ és convex si h_i son funcions afins, és a dir, $h_i(x) = a^T x - b$ (fàcil de provar, es deixa com exercici).
- La condició de la proposició anterior és suficient per garantir una regió factible convexa, però no necessària. Se us acut un exemple de regió factible convexa que no garanteix aquesta condició?
  - Per exemple $\Omega = \{x \in \mathbb{R} : x^2 - 4 \geq 0, x + 2 \leq 0\} = ((-\infty, -2] \cup [2, +\infty)) \cap (-\infty, -2] = (-\infty, -2]$.

Problema convex i òptims globals I

Definició

El problema d'optimització

$$\begin{array}{ll} \min & f(x) \\ s.a & x \in \Omega \end{array}$$

és convex si f és funció convexa i Ω és un conjunt convex

Proposició

Sigui $\Omega \subseteq \mathbb{R}^n$ conjunt convex, i $f : \Omega \to \mathbb{R}$ funció convexa. Aleshores un mínim local de f a Ω és també mínim global de f a Ω. A més si f és estrictament convexa, només hi ha un únic mínim global de f a Ω.

Problema convex i òptims globals II

Demostració.

- Part 1. Suposem que x és mínim local de f i hi ha un altre mínim global $y \neq x$ tal que $f(y) < f(x)$. Com que Ω és convex $\alpha x + (1-\alpha)y \in \Omega, 0 \leq \alpha \leq 1$. Usant que f és convexa i $f(y) < f(x)$ tenim

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) < f(x)$$

  contradint que x és mínim local.
- Part 2. Suposem que x i y ($x \neq y$) són dos mínims globals, $f(x) = f(y)$. El punt $\alpha x + (1-\alpha)y \in \Omega, 0 < \alpha <$ perquè Ω és convex, i com que f és estrictament convexa

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y) = f(x) = f(y)$$

  contradint que x i y siguin mínims globals.

$\square$

- Els PLs són problemes convexos? Podem garantir que un mínim local d'un PL és mínim global? Què en penseu?

Convexitat del conjunt de solucions

Proposició

Sigui $\min f(x)\ x \in \Omega$ *problema convex (f convexa i* Ω *convex). Llavors el conjunt de solucions* $\Omega^* \subseteq \Omega$ *és convex.*

Demostració.

(Es deixa com exercici. Podeu usar que $g(x) \leq c$ és conjunt convex si $g(x)$ convexa.) $\square$

Evolució de la ONL I

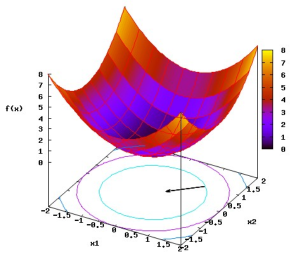
- Els orígens de la ONL es troben al càlcul de mínims de problemes d'una variable primer, i després al "càlcul de variacions" (CV) dels S. XVIII i XIX. Al CV les variables a optimitzar són funcions (optimització en dimensió infinita), i resol problemes com:
  - Quina és la forma òptima d'un automòbil que minimitza la resistència a l'aire?
  - Quina trajectòria segueix un raig de llum en un medi irregular?
  - El problema de la braquistòcrona (Johann Bernoulli, 1696): quina trajectòria ha de seguir una bola per anar d'un punt origen a un destí, si només intervé la força de gravetat i no hi ha fricció? Bernoulli, Newton i Leibniz (entre d'altres) van proposar solucions a aquest problema. Euler i Lagrange van estudiar problemes de CV.
- Es va reprendre més tard l'estudi de l'optimització (en dimensió finita).
- Lagrange va crear el concepte de multiplicador de Lagrange (1778).

Evolució de la ONL II

- Euler i Lagrange només consideraven restriccions d'igualtat.
- Weierstrass va estudiar ONL i transformava $g(x) \leq 0$ en igualtats fent $g(x) + s^2 = 0$. Ara no tenim desigualtat, però tenim més variables, la restricció és ara no lineal si $g(x)$ no ho era, i el conjunt factible pot esdevenir no convex. Exemple: $\Omega = \{x \in \mathbb{R} : -x \leq 0\}$ és convex, mentre que $\Omega' = \{(x, s) \in \mathbb{R}^2 : -x + s^2 = 0\}$ no és convex.
- La gran aportació a les condicions d'optimalitat de problemes amb desigualtats $g(x) \leq 0$ són les anomenades condicions necessàries de primer ordre de Karush-Kuhn-Tucker, anomenades KKT. Harold Kuhn i Albert Tucker les publiquen el 1951. Després se sap que William Karush les havia publicat a la seva tesi de Master el 1939 (Univ. of Chicago). A. Tucker i H. Kuhn van ser respectivament director de tesi i company de John Nash a Princeton.
- El desenvolupament de l'Optimització Lineal (simplex, mètodes de punt interior), cas particular de la ONL, és posterior.

Exemple introductori

- $\min f(x_1, x_2) = x_1^2 + x_2^2$: el mínim local i global estricte és $x^* = [0\ 0]^T$.
- Un altre punt x no és mínim perquè hi ha direccions $d \in \mathbb{R}^2 : f(x + \alpha d) < f(x)$, $\alpha > 0$.



- Per exemple, al punt $x = [1\ 1]^T$ al llarg de la direcció $d = [-1\ \ -1]^T$ tenim
  $f(x + \alpha d) = (1 - \alpha)^2 + (1 - \alpha)^2 = 2 - 4\alpha + 2\alpha^2 = f(x) - 4\alpha + 2\alpha^2 < 0$ si $\alpha \in (0, 2)$.
- Pel Teorema de Taylor
  $$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^T d + o(\alpha)$$
  veiem que tindrem una direcció $d : f(x + \alpha d) < f(x)$ si $\nabla f(x)^T d < 0$.
- I sempre podem fer $\nabla f(x)^T d < 0$ escollint per exemple $d = -\nabla f(x)$:
  $$\nabla f(x)^T d = -\nabla f(x)^T \nabla f(x) = -\|\nabla f(x)\|_2^2 \leq 0$$
- La única forma d'evitar $\nabla f(x)^T d < 0$ és que $\nabla f(x) = 0$. Aquesta és la condició necessària d'optimalitat de primer ordre.

Condicions necessàries de primer i segon ordre

Teorema

Si x^* és un mínim local de f, i $f \in \mathcal{C}^2$ en un entorn obert de x^*, llavors

i) $\nabla f(x^*) = 0$ *(condició de primer ordre).*

ii) $\nabla^2 f(x^*)$ *és semidefinida positiva (condició de segon ordre).*

(Demostració a la pissarra.)

- Els punts $x^* : \nabla f(x^*) = 0$ s'anomenen punts estacionaris.

Condicions suficients de segon ordre

Teorema

Si $f \in \mathcal{C}^2$ en un entorn obert de x^*, $\nabla f(x^*) = 0$ i $\nabla^2 f(x^*)$ és definida positiva llavors x^* és mínim local estricte de f.

(Demostració a la pissarra.)

- Les condicions suficients ens garanteix que x^* és mínim local estricte, mentre que les condicions necessàries només consideren mínims locals.
- Són condicions suficients, no necessàries: hi ha mínims locals estrictes que no verifiquen les condicions suficients (com veurem als exemples).

Condicions d'optimalitat en problemes convexos

- Si $f \in \mathcal{C}^1$ és convexa només cal comprovar si el punt és estacionari: si això es verifica el punt és mínim local.
- I no només és mínim local: és també mínim global.
- A diferència de les condicions suficients no cal considerar que $\nabla^2 f(x^*)$ sigui definida positiva: en ser f convexa n'hi ha prou amb que $\nabla^2 f(x^*)$ sigui semidefinida positiva.

Teorema

Si f és convexa i diferenciable, llavors un punt estacionari x^* ($\nabla f(x^*) = 0$) és mínim global de f.

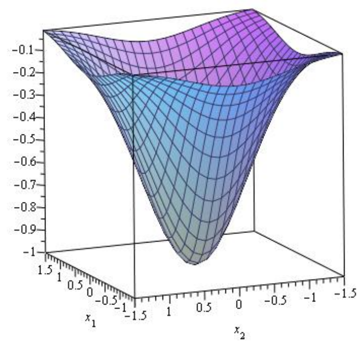
(Demostració a la pissarra.)

Exemple 1: mínim local estricte

- $f(x) = -e^{-({x_1}^2+{x_2}^2)}$

$$\nabla f(x) = \begin{bmatrix} 2\,x_1 e^{-({x_1}^2+{x_2}^2)} \\ 2\,x_2 e^{-({x_1}^2+{x_2}^2)} \end{bmatrix}$$

$$\nabla^2 f(x) = e^{-({x_1}^2+{x_2}^2)} \begin{bmatrix} 2-4{x_1}^2 & -4\,x_1 x_2 \\ -4\,x_1 x_2 & 2-4\,{x_2}^2 \end{bmatrix}$$



- Punts estacionaris: $\nabla f(x^*) = 0 \Rightarrow x^* = [0 \;\; 0]^T$.
- $\nabla^2 f(x^*)$ és definida positiva: x^* és mínim local estricte.

$$\nabla^2 f(x^*) = e^0 \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

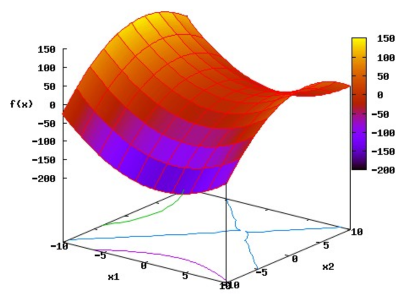
- En aquest cas és mínim global també, tot i que f no és convexa.

Exemple 2: punt de sella

- $f(x) = x_1^2 - x_2^2 - 4x_1 + 6x_2 - 5$

$$\nabla f(x) = \begin{bmatrix} 2x_1 - 4 \\ -2x_2 + 6 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$



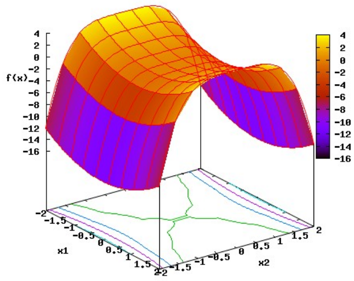
- Punts estacionaris: $\nabla f(x^*) = 0 \Rightarrow x^* = [2 \;\; 3]^T$.
- $\nabla^2 f(x^*)$ és indefinida: té valors propis positius i negatius.
- En aquest cas x^* no és mínim: és un punt de sella. Hi ha direccions d a partir de x^* on f augmenta o decreix (com s'observa a la gràfica).

Exemple 3: punt que satisfà condicions necessàries i no és mínim

- $f(x) = x_1^2 - x_2^4$

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ -4x_2^3 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -12x_2^2 \end{bmatrix}$$



- Punts estacionaris:
  $\nabla f(x^*) = 0 \Rightarrow x^* = [0 \;\; 0]^T$.
- $\nabla^2 f(x^*)$ és semidefinida positiva, no definida positiva: no podem assegurar que sigui mínim.
- De fet no és mínim (veure gràfica): al llarg de qualsevol direcció $d = [0 \;\; d_2]^T$ f disminueix:

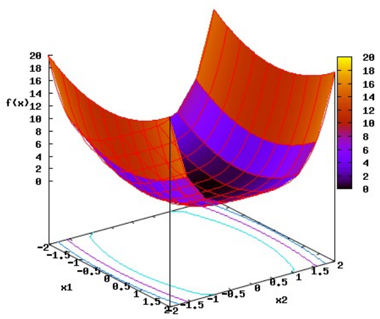
$$f(x^* + \alpha d) = f([0, \alpha d_2]^T) = -(\alpha d_2)^4 < 0 = f(x^*).$$

Exemple 4: mínim local estricte que no satisfà condicions suficients

- $f(x) = x_1^2 + x_2^4$

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 4x_2^3 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 12x_2^2 \end{bmatrix}$$



- Punts estacionaris:
  $\nabla f(x^*) = 0 \Rightarrow x^* = [0 \;\; 0]^T$.
- $\nabla^2 f(x^*)$ és semidefinida positiva, no definida positiva: no podem assegurar que sigui mínim.
- Però en aquest cas sí és mínim local (i estricte).
- Per tant hi ha mínims locals estrictes que no satisfan les condicions suficients.

Per què calen mètodes d'optimització? Tipus de mètodes.

- Per què no solucionar $\nabla f(x) = 0$ directament?
  - $\nabla f(x) = 0$ és sovint un sistema d'equacions no lineal: se soluciona numèricament de totes formes.
  - Els punts estacionari no tenen per què ser mínims (màxims, punts de sella...). Els mètodes d'optimització estan dissenyats per anar al mínim.
  - Poden ser més ràpids que solucionar el sistema $\nabla f(x) = 0$ (fins i tot si és un sistema lineal! per exemple usant gradients conjugats).
- Tipus de mètodes
  - Mètodes d'exploració lineal: primer calculem $d \in \mathbb{R}^n$, després $\alpha \in \mathbb{R}$, i obtenim $x^{k+1} = x^k + \alpha d$.
  - Mètodes de regió de garantia: primer fixem $\Delta \in \mathbb{R}$, després calculem $d \in \mathbb{R}^n$ solucionant

$$\begin{array}{ll} \min_{d \in \mathbb{R}^n} & m_k(d) \\ & \|d\|_2 \leq \Delta \end{array}$$

    on m_k és una aproximació de f a x^k, i obtenim $x^{k+1} = x^k + d$.

Procediment general dels mètodes basats en exploracions lineals

- Els mètodes generen una seqüència de punts $\{x^k\}_0^\infty$ que (sota certes condicions) convergeix a l'òptim.
- L'algorisme general per a $\min_{x \in \mathbb{R}^n} f(x)$ dels mètodes basats en exploracions lineals és:

Algorithm *mètodes basats en exploracions lineals*
    Punt inicial x^0, $k = 0$
    while x^k no és solució **do**
        Calcular direcció de moviment (de descens) d
        Calcular longitud de pas α
        $x^{k+1} = x^k + \alpha d$
        $k := k + 1$
    end_while
    Return: $x^* = x^k$
End_algorithm

- Els tres passos clau estan marcats en vermell.

Direccions de moviment: direccions de descens I

- La direcció d usada per $x^{k+1} = x^k + \alpha d$ ha de ser de descens: ha de reduir “localment” el valor de f per a α petita: $f(x^{k+1}) < f(x^k)$.

Definició

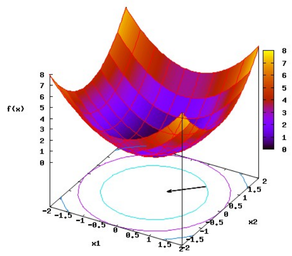
d és direcció de descens de f a x^k si $\exists \bar{\alpha} : \forall \alpha \in (0, \bar{\alpha})\ f(x^k + \alpha d) < f(x^k)$.

Exemple (ja vist):

$$f(x_1, x_2) = x_1^2 + x_2^2 \qquad x^k = [1\ 1]^T \qquad d = [-1\ \ -1]^T$$

$$f(x^k + \alpha d) = (1-\alpha)^2 + (1-\alpha)^2 = 2 - 4\alpha + 2\alpha^2 = f(x^k) + \alpha(2\alpha - 4)$$

Per a $\alpha \in (0, 2)$, $f(x^k + \alpha d) < f(x^k)$ i d és direcció de descens.



Optimització no lineal sense restriccions | Mètodes basats en exploracions lineals: gradient i Newton

Direccions de moviment: direccions de descens II

- Caracterització de direccions de descens:

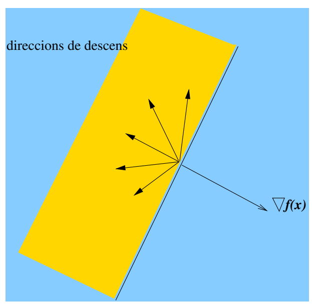
Proposició

Si $\nabla f(x^k)^T d < 0$ *llavors d és direcció de descens de f a x^k.*

- Quines d són de descens?
  Com que $\|\nabla f(x^k)\| \neq 0$ i $\|d\| \neq 0$ tenim que

$$\nabla f(x^k)^T d = \|\nabla f(x^k)\| \|d\| \cos(\nabla f(x^k), d) < 0 \quad \text{si } \cos(\nabla f(x^k), d) < 0$$

Si l’angle entre $\nabla f(x^k)$ i d és superior a $\pi/2$ radians la direcció és de descens.



jordi.castro@upc.edu

La direcció de descens més ràpid *(steepest descent)*

- La direcció de descens més ràpid de f al punt x^k és la solució de

$$\begin{array}{ll} \min\limits_{d \in \mathbb{R}^n} & \nabla f(x^k)^T d \\ \text{s. a} & \|d\|_2 = 1. \end{array}$$

  (S'imposa $\|d\|_2 = 1$ perquè sino el problema lineal seria il·limitat.)
- Com que $\|d\|_2 = 1$ i $\cos(t) \geq -1$ calculem el mínim directament:

$$\nabla f(x^k)^T d = \|\nabla f(x^k)\| \|d\| \cos(\nabla f(x^k), d) \geq -\|\nabla f(x^k)\| \Rightarrow d = \frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|}$$

- Usarem $d = -\nabla f(x^k)$, que fa un angle de π radians amb $\nabla f(x^k)$.
- Iteracions del mètode del gradient o direcció de descens més ràpid:

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

- Mètode originalment proposat per Cauchy el 1847.

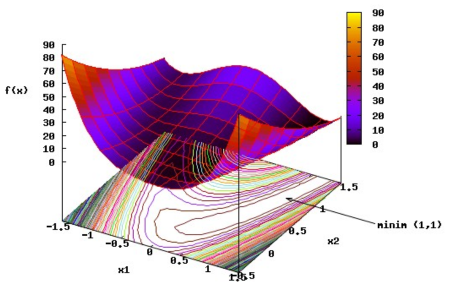
Exemple direcció de descens més ràpid I

- Funció de Rosenbrock en $\mathbb{R}^2$ ("banana function"):

$$f(x) = 10(x_2 - x_1^2)^2 + (1 - x_1)^2$$

$$\nabla f(x) = \begin{bmatrix} -40(x_2 - x_1^2)x_1 - 2(1 - x_1) \\ 20(x_2 - x_1^2) \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 120x_1^2 - 40x_2 + 2 & -40x_1 \\ -40x_1 & 20 \end{bmatrix}$$



- Mínim local estricte: $x^* = [1 \;\; 1]^T$, $\nabla f(x^*) = [0 \;\; 0]^T$ i $\nabla^2 f(x^*) = \begin{bmatrix} 82 & -40 \\ -40 & 20 \end{bmatrix}$ és definida positiva. De fet és el mínim global (encara que f no és convexa).

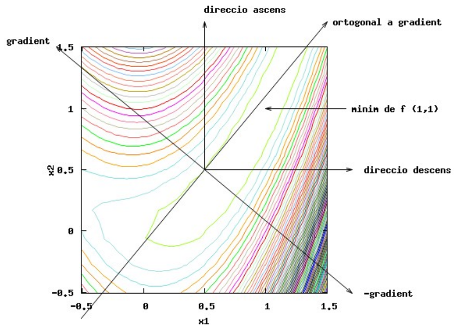
Exemple direcció de descens més ràpid II

- En el punt $x^k = [1/2 \quad 1/2]^T$

$$f(x^k) = \frac{7}{8} \qquad \nabla f(x^k) = \begin{bmatrix} -6 \\ 5 \end{bmatrix} \qquad \nabla^2 f(x^k) = \begin{bmatrix} 12 & -20 \\ -20 & 20 \end{bmatrix}$$

- Comprovem condició de descens $\nabla f(x^k)^T d$ en vàries direccions:

  - $d^1 = -\nabla f(x^k) = [6 \quad -5]^T$ (descens).
  - $d^2 = [1 \ \ 0]^T$ (descens).
  - $d^3 = [0 \ \ 1]^T$ (ascens).
  - $d^4 = [5 \ \ 6]^T$ (ortogonal a $\nabla f(x^k)$).



Exemple direcció de descens més ràpid III

- $\nabla f(x^k)^T d^4 = 0$: no sabem si és d'ascens o descens. Usant

$$f(x^k + \alpha d^4) = f(x^k) + \alpha \nabla f(x^k)^T d^4 + \frac{1}{2}\alpha^2 (d^4)^T \nabla^2 f(x^k) d^4 + o(\alpha^2)$$

caldria mirar el signe de $(d^4)^T \nabla^2 f(x^k) d^4$

$$[5 \ \ 6] \begin{bmatrix} 12 & -20 \\ -20 & 20 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} = [-60 \ \ 20] \begin{bmatrix} 5 \\ 6 \end{bmatrix} = -300 + 120 = -180 < 0.$$

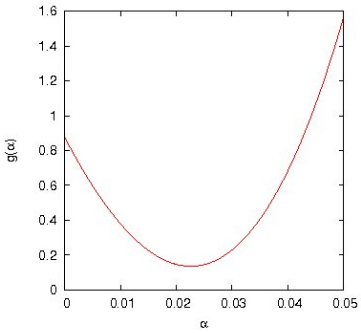
d^4 és direcció de descens.

Exemple direcció de descens més ràpid IV

- La direcció $-\nabla f(x^k)$ és de descens només localment. Si $\alpha \gg 0$ llavors $f(x^k - \alpha \nabla f(x^k)) > f(x^k)$:

$$x^k - \alpha \nabla f(x^k) = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} - \alpha \begin{bmatrix} -6 \\ 5 \end{bmatrix} = \begin{bmatrix} 1/2 + 6\alpha \\ 1/2 - 5\alpha \end{bmatrix}$$

$$g(\alpha) = f(x^k - \alpha \nabla f(x^k)) = 10((1/2 - 5\alpha) - (1/2 + 6\alpha)^2)^2 + (1 - (1/2 + 6\alpha))^2$$



- Cal fer una exploració lineal per trobar la millor α^* (una bona α serà suficient a la pràctica).

Algorisme del mètode del gradient

Algorithm *Gradient o "Steepest descent"*
 Punt inicial x^0, $k = 0$
 while x^k no és solució **do**
  $d^k = -\nabla f(x^k)$
  Calcular α^k que satisfà condicions d'Armijo-Wolfe
  $x^{k+1} = x^k + \alpha^k d^k$
  $k := k + 1$
 end_while
 Return: $x^* = x^k$
End_algorithm

(Intenteu programar-ho en Matlab/octave. Useu la funció matlab `fminbnd` per calcular α^k per exploració lineal exacta. Proveu-lo amb la funció de Rosenbrock començant, per exemple, des del punt $x^0 = [-1.2 \;\; 1]^T$.)

La direcció de Newton

- S'obté a partir d'aproximació quadràtica de f al punt x^k:

$$f(x^k+d) = f(x^k) + \nabla f(x^k)^T d + \tfrac{1}{2} d^T \nabla^2 f(x^k) d + R_2(\|d\|)$$
$$\text{on} \quad R_2(\|d\|) = o(\|d\|^2) \qquad R_2(\|d\|) = O(\|d\|^3)$$

$$f(x^k+d) \approx m_k(d) = f(x^k) + \nabla f(x^k)^T d + \tfrac{1}{2} d^T \nabla^2 f(x^k) d$$

- Notació: $f(t) = O(g(t)$ si $\exists M : |f(t)| \leq M|g(t)|$ per a tot $t \geq t_0$ o $0 \leq t \leq t_0$ ($f(t)$ i $g(t)$ són del mateix ordre).
- Escrivim aproximació quadràtica en funció de x: $d = x - x^k$

$$m_k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k)$$

- Suposant $\nabla^2 f(x^k)$ és semidef. pos. ($m_k(d)$ convexa) calculem

$$\min_{d \in \mathbb{R}^n} \quad m_k(d) = f(x^k) + \nabla f(x^k)^T d + \tfrac{1}{2} d^T \nabla^2 f(x^k) d$$
$$0 = \nabla m_k(d) = \nabla f(x^k) + \nabla^2 f(x^k) d$$
$$d = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

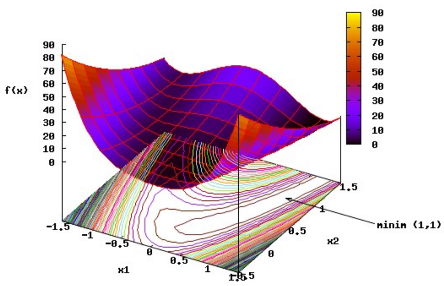
Exemple mètode de Newton I

- Funció de Rosenbrock en $\mathbb{R}^2$ ("banana function"):

$$f(x) = 10(x_2 - x_1^2)^2 + (1 - x_1)^2$$

$$\nabla f(x) = \begin{bmatrix} -40(x_2 - x_1^2)x_1 - 2(1 - x_1) \\ 20(x_2 - x_1^2) \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 120x_1^2 - 40x_2 + 2 & -40x_1 \\ -40x_1 & 20 \end{bmatrix}$$



- Mínim local estricte: $x^* = [1 \ \ 1]^T$, $\nabla f(x^*) = [0 \ \ 0]^T$ i $\nabla^2 f(x^*) = \begin{bmatrix} 82 & -40 \\ -40 & 20 \end{bmatrix}$ és definida positiva. De fet és el mínim global (encara que f no és convexa).

Exemple mètode de Newton II

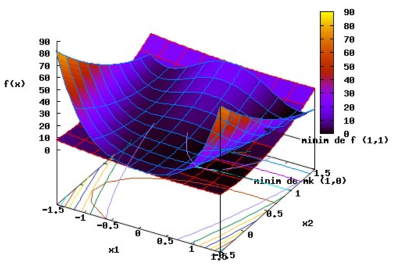
- Calculem model quadràtic $m_k(x)$ en $x^k = [0 \;\; 0]^T$:

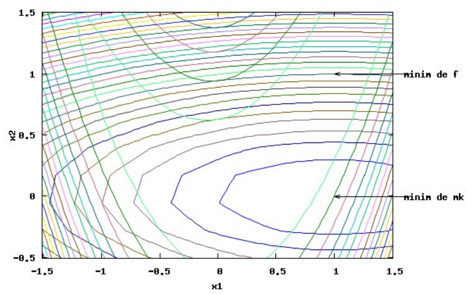
$$x^k = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad f(x^k) = 1 \qquad \nabla f(x^k) = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \qquad \nabla^2 f(x^k) = \begin{bmatrix} 2 & \\ & 20 \end{bmatrix}$$

$$\begin{aligned} m_k(x) &= f(x^k) + \nabla f(x^k)^T (x - x^k) + \tfrac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k) \\ &= 1 + [-2 \;\; 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \tfrac{1}{2}[x_1 \;\; x_2] \begin{bmatrix} 2 & \\ & 20 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_1^2 + 10x_2^2 - 2x_1 + 1 \end{aligned}$$

- El mínim de $m_k(x)$ és $[1 \;\; 0]^T$.

Exemple mètode de Newton III



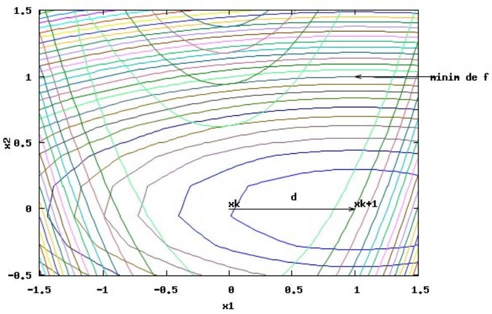


- La direcció de Newton al punt $x^k = [0 \;\; 0]^T$ és $d = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$:

$$\nabla^2 f(x^k) d = -\nabla f(x^k) \quad \Leftrightarrow \quad \begin{bmatrix} 2 & \\ & 20 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = - \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad \Rightarrow \quad d = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

- El nou punt és: $x^{k+1} = x^k + d = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, que és el mínim de $m_k(x)$.

Exemple mètode de Newton IV



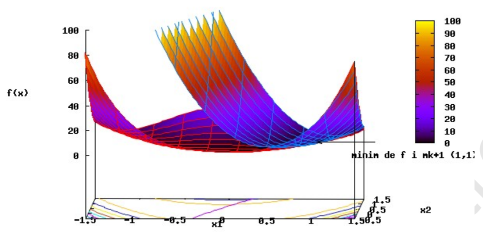
- Calculem model quadràtic $m_{k+1}(x)$ en $x^{k+1} = [1 \;\; 0]^T$:

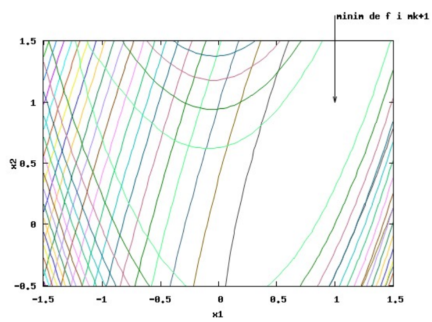
$$x^{k+1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad f(x^{k+1}) = 10 \quad \nabla f(x^{k+1}) = \begin{bmatrix} 40 \\ -20 \end{bmatrix} \quad \nabla^2 f(x^{k+1}) = \begin{bmatrix} 122 & -40 \\ -40 & 20 \end{bmatrix}$$

$$\begin{aligned} m_{k+1}(x) &= f(x^{k+1}) + \nabla f(x^{k+1})^T(x - x^{k+1}) + \tfrac{1}{2}(x - x^{k+1})^T \nabla^2 f(x^{k+1})(x - x^{k+1}) \\ &= 10 + [40 \;\; -20] \begin{bmatrix} x_1 - 1 \\ x_2 \end{bmatrix} + \tfrac{1}{2}[x_1 - 1 \;\; x_2] \begin{bmatrix} 122 & -40 \\ -40 & 20 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 \end{bmatrix} \\ &= 61x_1^2 + 10x_2^2 - 40x_1x_2 - 82x_1 + 20x_2 + 31 \end{aligned}$$

Exemple mètode de Newton V

- El mínim de $m_{k+1}(x)$ és $[1 \;\; 1]^T$.



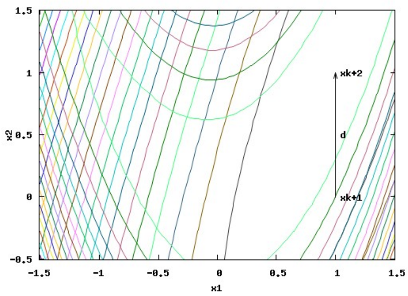


- La direcció de Newton al punt $x^{k+1} = [1 \;\; 0]^T$ és $d = -(\nabla^2 f(x^{k+1}))^{-1} \nabla f(x^{k+1})$:

$$\nabla^2 f(x^{k+1}) d = -\nabla f(x^{k+1}) \Leftrightarrow \begin{bmatrix} 122 & -40 \\ -40 & 20 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = -\begin{bmatrix} 40 \\ -20 \end{bmatrix} \quad \Rightarrow \quad d = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Exemple mètode de Newton VI

- El nou punt és: $x^{k+2} = x^{k+1} + d = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, que és el mínim de $m_{k+1}(x)$, i també la solució de $\min f(x)$.



Algorisme del mètode de Newton

Algorithm *Mètode de Newton*
  Punt inicial x^0, $k = 0$
  while x^k no és solució **do**
    $d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$
    Calcular α^k que satisfà condicions d'Armijo-Wolfe
    $x^{k+1} = x^k + \alpha^k d^k$
    $k := k + 1$
  end_while
  Return: $x^* = x^k$
End_algorithm

(Intenteu programar-ho en Matlab/octave. Useu la funció matlab `fminbnd` per calcular α^k per exploració lineal exacta. Proveu-lo amb la funció de Rosenbrock començant, per exemple, des del punt $x^0 = [-1.2 \;\; 1]^T$.)

Propietats direcció de Newton

- Si $\nabla^2 f(x^k)$ és def. pos. llavors $(\nabla^2 f(x^k))^{-1}$ és def. pos. (es deixa com exercici) i $d = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$ és de descens

$$\nabla f(x^k)^T d = -\nabla f(x^k)^T(\nabla^2 f(x^k))^{-1}\nabla f(x^k) < 0.$$

  Fixem-nos que qualsevol B_k definida positiva garanteix que $d = -B_k^{-1}\nabla f(x^k)$ és de descens:
  - $B_k = I$: direcció de màxim descens.
  - $B_k = \nabla^2 f(x^k)$: direcció de Newton.
  - B_k aproximació definida positiva de $\nabla^2 f(x^k)$: direcció quasi-Newton.
- A diferència del mètode del gradient $\alpha = 1$ és el pas "natural" a la direcció de Newton. Si es pot, s'usa $\alpha = 1$.
- La direcció de Newton no garanteix descens si $\nabla^2 f(x^k)$ no és definida positiva. Fins i tot pot no existir si $\nabla^2 f(x^k)$ és singular.
  - Cal usar variants que modifiquen la Hessiana.

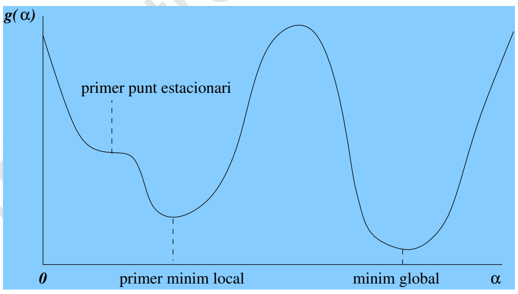
Exploració lineal exacta i inexacta

- Exploració lineal: busquem α per a $x^{k+1} = x^k + \alpha d^k$.
- Idealment exploració lineal exacta, però és computacionalment costosa.

$$\alpha^* = \arg\min_{\alpha > 0} g(\alpha) = f(x^k + \alpha d^k)$$

on $$g'(\alpha) = \frac{d}{d\alpha} f(x^k + \alpha d^k) = \nabla f(x^k + \alpha d^k)^T d^k.$$

És un problema d'optimització d'una variable i una restricció.



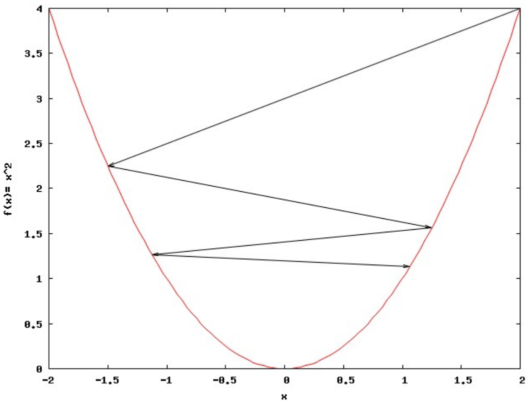
- A la pràctica n'hi ha prou amb una α subòptima que garanteixi que $\{x^k\}_{k\geq 0}$ convergeix a un punt estacionari: Exploració lineal inexacta.

Condicions a imposar a α insuficients I

- Quina condició imposem a α per garantir convergència a $\nabla f(x^*) = 0$?
  No és suficient imposar $f(x^k + \alpha d^k) < f(x^k)$.
- Exemple 1 ($\|\alpha d^k\|$ llarg, reducció de f petita):
  - $f(x) = x^2 \qquad x^0 = 2$
  - $d^k = (-1)^{k+1} \qquad \alpha^k = 2 + \dfrac{3}{2^{k+1}}$
  - Seqüència de punts $x^{k+1} = x^k + \alpha^k d^k$:

k	x^k	d^k	α^k
0	2	-1	$2+3/2$
1	$-3/2$	1	$2+3/4$
2	$5/4$	-1	$2+3/8$
3	$-9/8$	1	$2+3/16$
4	$17/16$	-1	$2+3/32$

  - $x^k = (-1)^k(1 + \frac{1}{2^k})$
  - $\lim_{k\to\infty} f(x^k) = 1 \neq 0$. No convergeix al punt estacionari.

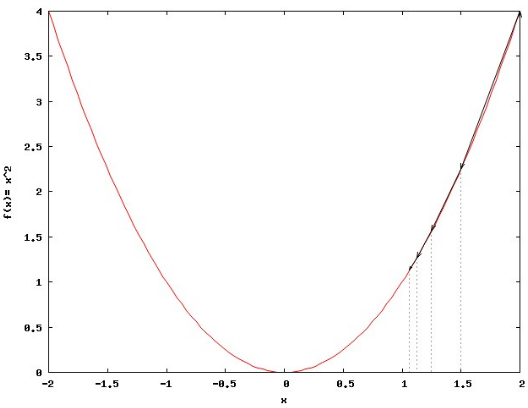


Condicions a imposar a α insuficients II

- Exemple 2 ($\|\alpha d^k\|$ massa petit):
  - $f(x) = x^2 \qquad x^0 = 2$
  - $d^k = -1 \qquad \alpha^k = \dfrac{1}{2^{k+1}}$
  - Seqüència de punts $x^{k+1} = x^k + \alpha^k d^k$:

k	x^k	d^k	α^k
0	2	-1	$1/2$
1	$3/2$	-1	$1/4$
2	$5/4$	-1	$1/8$
3	$9/8$	-1	$1/16$
4	$17/16$	-1	$1/32$

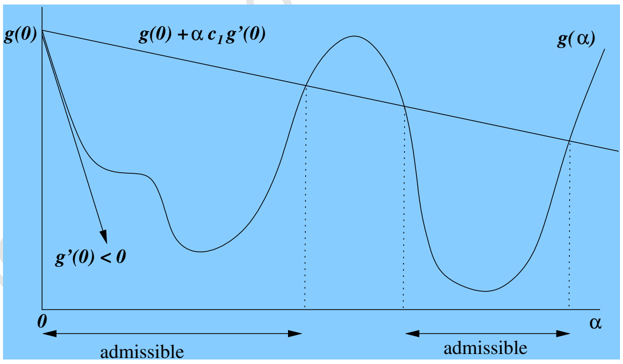
  - $x^k = (1 + \frac{1}{2^k})$
  - $\lim_{k\to\infty} f(x^k) = 1 \neq 0$. No convergeix al punt estacionari.



Condicions d'Armijo-Wolfe I

1. Condició de decrement suficient (per evitar el problema de l'exemple 1). Anomenada condició d'Armijo (1966)

$$
\begin{array}{rcll}
g(\alpha) & \leq & g(0) + \alpha c_1 g'(0) & \qquad c_1 \in (0,1) \\
 & & \Updownarrow & \\
f(x^k + \alpha d^k) & \leq & f(x^k) + \alpha c_1 \nabla f(x^k) d^k & \qquad c_1 \in (0,1)
\end{array}
$$

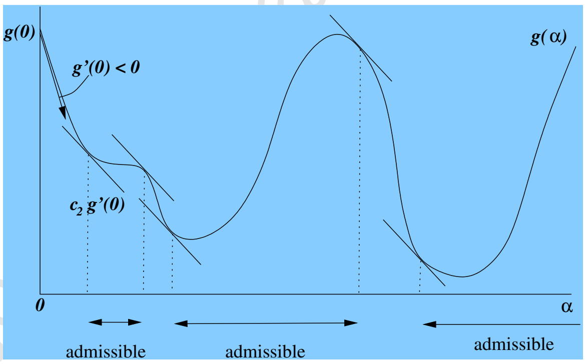


Valor habitual $c_1 = 10^{-4}$, que facilita trobar α.

Condicions d'Armijo-Wolfe II

2. Condició de corbatura (per evitar passos petits com a l'exemple 2).

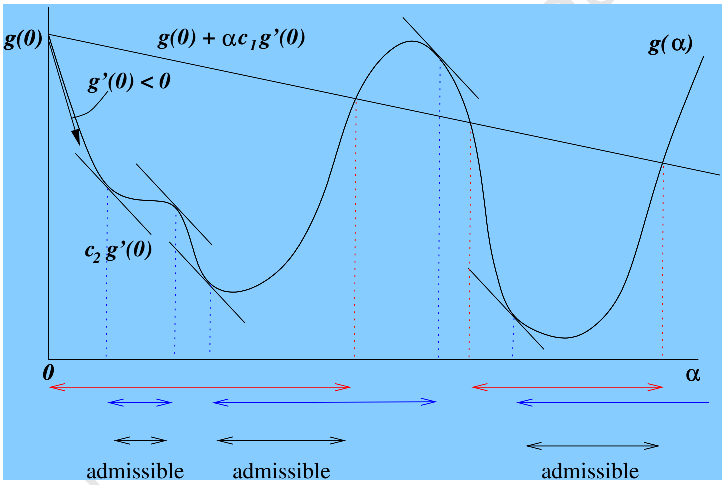
$$
\begin{array}{rcll}
g'(\alpha) & \geq & c_2 g'(0) & \qquad 0 < c_1 < c_2 < 1 \\
 & & \Updownarrow & \\
\nabla f(x^k + \alpha d^k)^T d^k & \geq & c_2 \nabla f(x^k) d^k & \qquad 0 < c_1 < c_2 < 1
\end{array}
$$



Valor habitual $c_2 = 0.9$, que evita passos petits i facilita trobar α.

Condicions d'Armijo-Wolfe III

- Busquem α que satisfaci les dues condicions d'Armijo-Wolfe:



- Demostrarem que si les α satisfan aquestes condicions, i les d^k també satisfan unes altres condicions, llavors la seqüència de punts $x^{k+1} = x^k + \alpha^k d^k$ convergirà globalment a un punt estacionari.

Exploració lineal per retrocès ("backtracking")

- Busquem un α^k que verifiqui la primera condició de decrement suficient.
- El mètode de backtracking és molt senzill d'implementar:

> **Algorithm** *exploració lineal per "backtracking"*
>   Escollir $\bar{\alpha} > 0$, $\rho \in (0,1)$, $c_1 \in (0,1)$;
> $\alpha := \bar{\alpha}$;
>   **while** NOT $(f(x^k + \alpha d^k) \leq f(x^k) + \alpha c_1 \nabla f(x^k) d^k)$ **do**
> $\alpha := \rho\alpha$;
>   **end_while**
>   **Return:** $\alpha^k = \alpha$;
> **End_algorithm**

- Al mètode de Newton és natural usar $\bar{\alpha} = 1$.
- El punt anterior a α^k és α^k/ρ, pel que el propi procediment de backtracking evita punts molt propers a 0, garantint-se també la segona condició de Wolfe per a algun $c_2 < 1$.

Convergència global i local

- Els mètodes d'optimització generen seqüències de punts $x^0, x^1, \ldots, x^k, \ldots$
- Un algorisme té convergència global si finalitza en una solució independentment del punt inicial: $\lim_{k\to\infty} x^k = x^*$ per a tot x^0.
- La convergència local estudia la velocitat a la que ens atansem al punt x^* quan estem a prop d'ell.
- El teorema de Zoutendijk que veurem a continuació tracta la convergència global dels mètodes basats en exploracions lineals.

Teorema de Zoutendijk

Denotem per θ_k l'angle entre la direcció d^k i la direcció de màxim descens $-\nabla f(x^k)$, tal que

$$\cos\theta_k = \frac{-\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\| \|d^k\|}.$$

Teorema (de Zoutendijk)

Considerem el procediment iteratiu $x^{k+1} = x^k + \alpha^k d^k$ tal que (1) d^k és direcció de descens, i (2) α^k satisfà les condicions d'Armijo-Wolfe. Suposem que (3) f està afitada inferiorment en $\mathbb{R}^n$ i que (4) $f \in \mathcal{C}^1$ en un conjunt obert $\mathcal{N}$ que conté el conjunt de nivell $\mathcal{L} = \{x : f(x) \le f(x^0)\}$, on x^0 és el punt inicial d'iteració. També suposem (5) que el gradient ∇f és Lipschitz continu en $\mathcal{N}$, és a dir, existeix una constant $L > 0$ tal que

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall\ x, y \in \mathcal{N}.$$

Llavors

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^k)\|^2 < \infty.$$

Convergència global mètodes d'exploración lineal

- La condició de Zoutendijk

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^k)\|^2 < \infty$$

implica

$$\lim_{k\to\infty} \cos^2 \theta_k \|\nabla f(x^k)\|^2 = 0.$$

- Si d^k garanteix que θ_k (angle entre d^k i $-\nabla f(x^k)$) és inferior a $\pi/2$

$$\cos \theta_k \geq \delta > 0 \quad \forall k$$

i per tant

$$\lim_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

- El mètode és globalment convergent a un punt estacionari.

Convergència global del mètode de màxim descens (o mètode del gradient)

- $d^k = -\nabla f(x^k)$ i per tant $\theta_k = 0$ i $\cos \theta_k = 1$.
- Per tant, si les longituds de pas α^k verifiquen les condicions d'Armijo-Wolfe, el mètode de màxim descens (o del gradient) té convergència global a un punt estacionari.

Convergència global del mètode de Newton

- $d^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$.
- Si $\nabla^2 f(x^k)$ definida positiva d^k és de descens.
- Si $\nabla^2 f(x^k)$ té un nombre de condició uniformement afitat, és a dir,

$$\exists M : \quad \|\nabla^2 f(x^k)\| \, \|(\nabla^2 f(x^k))^{-1}\| \leq M \quad \forall k$$

  (o el que és el mateix, $\nabla^2 f(x^k)$ no és singular ni propera a ser-ho) llavors es demostra que

$$\cos\theta_k \geq \frac{1}{M}$$

  (es deixa com exercici).
- Llavors per la condició de Zoutendijk tenim convergència global (si α^k verifiquen Armijo-Wolfe).

Convergència local i velocitat de convergència

- Els mètodes d'optimització generen seqüències de punts $x^0, x^1, \ldots, x^k, \ldots$
- La convergència local estudia la velocitat a la que ens atansem al punt x^* quan estem a prop d'ell.
  - Convergència lineal (Exemple: $x^k = 2^{-k}$)

$$\exists r \in (0,1) : \frac{||x^{k+1} - x^*||}{||x^k - x^*||} \leq r \quad \forall k \text{ suficientment gran}$$

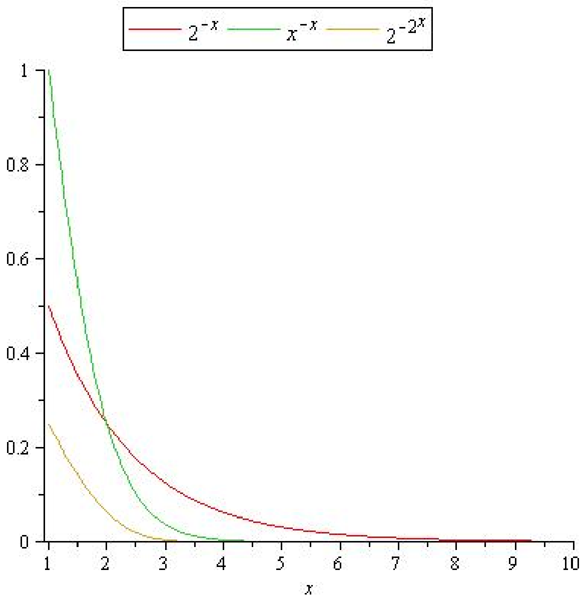
  - Convergència superlineal (Exemple: $x^k = k^{-k}$)

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^*||}{||x^k - x^*||} = 0$$

  - Convergència quadràtica (Exemple: $x^k = 2^{-2^k}$)

$$\exists M \in \mathbb{R} : \frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} \leq M \quad \forall k \text{ suficientment gran}$$

Exemple de velocitat de convergència



k	2^{-k}	k^{-k}	2^{-2^k}
1	0.500000	1.000000	0.250000
2	0.250000	0.250000	0.062500
3	0.125000	0.037037	0.003906
4	0.062500	0.003906	0.000015
5	0.031250	0.000320	0.000000
6	0.015625	0.000021	
7	0.007812	0.000001	
8	0.003906	0.000000	
9	0.001953		
10	0.000976		
11	0.000488		
12	0.000244		
13	0.000122		
14	0.000061		

Mètode del gradient en funcions quadràtiques

$$f(x) = \frac{1}{2}x^TQx - b^Tx \quad \nabla f(x) = Qx - b \quad \nabla^2 f(x) = Q \text{ simètrica i def. pos.}$$

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

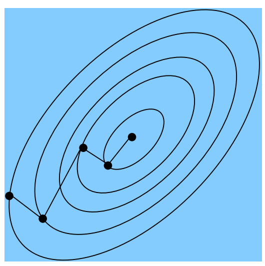
- Es pot calcular analíticament α^k per exploració lineal exacta

$$\alpha^k = \arg\min_{\alpha>0} f(x - \alpha\nabla f(x^k)) \quad \Rightarrow \quad \alpha^k = \frac{\nabla f(x^k)^T \nabla f(x^k)}{\nabla f(x^k)^T Q \nabla f(x^k)} > 0.$$

(Es deixa com exercici.)

- Si es fa exploració lineal exacta, dues direccions consecutives $d^k = -(Qx^k - b)$ i $d^{k+1} = -(Qx^{k+1} - b)$ són ortogonals.

(Es deixa com exercici.)



Convergència en funcions quadràtiques

- Definim $\|x\|_Q^2 = x^T Qx$ (norma ponderada per Q simètrica i definida positiva)
- Q definida positiva $\Rightarrow$ valors propis $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

Teorema

El mètode del màxim descens, (1) amb exploració lineal exacta $x^k = \arg\min_{\alpha>0} f(x^k + \alpha d^k)$, (2) aplicat a una funció quadràtica $f(x) = \frac{1}{2}x^T Qx - b^T x$, Q simètrica i definida positiva, satisfà

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} = \frac{\|x^{k+1} - x^*\|_Q^2}{\|x^k - x^*\|_Q^2} \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2$$

on λ_1 i λ_n són el menor i major valor propi de Q.

(Demostració a [Luenberger, Ye (2008)])

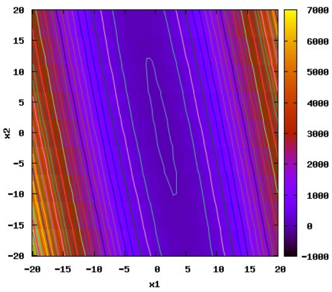
El mètode del gradient té convergència lineal.

Exemple convergència en funcions quadràtiques

$$f(x) = \frac{1}{2}x^T Qx - b^T x$$

$Q = \begin{bmatrix} 20 & 4 \\ 4 & 1 \end{bmatrix}$ $b = \begin{bmatrix} 24 \\ 5 \end{bmatrix}$ $x^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

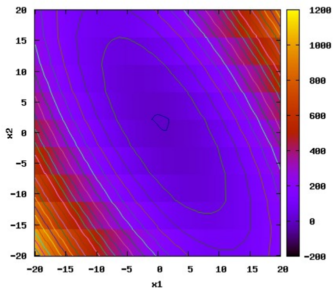
$\lambda_1 = 0.19224$ $\lambda_2 = 20.80776$ $\left(\frac{\lambda_2 - \lambda_1}{\lambda_2 + \lambda_1}\right)^2 = 0.96372$



Convergència en 118 iteracions, $\|\nabla f(x^k)\| = 7 \cdot 10^{-5}$

$Q = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ $b = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ $x^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\lambda_1 = 0.38197$ $\lambda_2 = 2.61803$ $\left(\frac{\lambda_2 - \lambda_1}{\lambda_2 + \lambda_1}\right)^2 = 0.55556$



Convergència en 32 iteracions, $\|\nabla f(x^k)\| = 8 \cdot 10^{-6}$

Convergència en funcions no lineals

És pràcticament la mateixa que per al cas quadràtic.

Teorema

Suposem que el mètode del màxim descens (1) amb exploració lineal exacta $x^k = \arg\min_{\alpha>0} f(x^k + \alpha d^k)$, (2) aplicat a $f : \mathbb{R}^n \to \mathbb{R}$, (3) convergeix a un punt x^* on la Hessiana $\nabla^2 f(x^*)$ és definida positiva.

Llavors, per a k prou gran

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \le r^2 \qquad \text{on} \qquad r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1\right).$$

on λ_1 i λ_n són el menor i major valor propi de $\nabla^2 f(x^*)$.

(Demostració a [Luenberger, Ye (2008)])

Convergència local del mètode de Newton i α^k

- A mida que ens apropem a x^* pot demostrar-se (i s'observa a la pràctica) que la longitud de pas $\alpha^k = 1$ verificarà les condicions d'Armijo-Wolfe.
- Per tant podem suposar que a prop de l'òptim (o per k suficientment gran) les iteracions del mètode de Newton seran

$$x^{k+1} = x^k + d^k \qquad\qquad d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k).$$

Convergència local del mètode de Newton

El mètode de Newton té convergència quadràtica.

Teorema

Suposem que

1. $f \in \mathcal{C}^2$.
2. $\nabla^2 f(x)$ *és Lipschitz contínua en un entorn* $\mathcal{N}$ *d'una solució* x^*, *és a dir,*
$$\exists\, L > 0 \quad : \quad \|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathcal{N}.$$
3. *A la solució* x^* *es verifiquen les condicions suficients d'optimalitat* ($\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ *és definida positiva).*
4. *Considerem la iteració* $x^{k+1} = x^k + d^k$, *on* $d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$ ($\alpha^k = 1$).

Llavors

- **(i)** *Si* x^0 *és prou proper a* x^*, *la seqüència* $\{x^k\}$ *convergeix a* x^*.
- **(ii)** *La seqüència* $\{x^k\}$ *convergeix quadràticament.*
- **(iii)** *La seqüència* $\{\|\nabla f(x^k)\|\}$ *convergeix quadràticament a 0.*

Gradient i Newton amb la funció de Rosenbrock I

- Punt inicial $x^0 = [-2 \quad 1]^T$ per als dos mètodes.
- Resultats amb un codi "acadèmic" (no-comercial) escrit en Matlab.

Mètode del màxim descens.

k	$f(x^k)$	α^k	$\|\nabla f(x^k)\|$
0	9.9000000e+01	0.0000000e+00	2.5321137e+02
1	3.5384414e+00	1.2500000e-02	2.8051479e+01
2	2.8487298e-01	9.9714185e-03	8.1723192e+00
3	7.9346618e-02	6.2748965e-03	2.0643391e-01
4	7.4166874e-02	2.3623952e-01	1.1643898e+00
⋮	⋮	⋮	⋮
496	2.0061607e-04	1.8711002e-02	1.7328277e-02
497	1.9766057e-04	1.9689489e-02	1.7644906e-02
498	1.9474726e-04	1.8713998e-02	1.7074314e-02
499	1.9187636e-04	1.9699026e-02	1.7389026e-02
500	1.8904649e-04	1.8716956e-02	1.6823893e-02

S'atura per màxim d'iteracions "lluny" de l'òptim ($x^{500} = [1.0137 \quad 1.0279]^T$).

Gradient i Newton amb la funció de Rosenbrock II

Mètode de Newton.

k	$f(x^k)$	α^k	$\|\nabla f(x^k)\|$	definida
0	9.9000000e+01	0.0000000e+00	2.5321137e+02	+
1	8.7073952e+00	1.0000000e+00	6.0905697e+00	+
2	8.5567773e+00	2.8000000e-01	3.5547527e+01	+
3	4.0146814e+00	1.0000000e+00	5.0656188e+00	+
4	2.9055843e+00	2.8000000e-01	7.8324868e+00	+
5	1.7745925e+00	1.0000000e+00	5.0638727e+00	+
6	9.5739882e-01	1.0000000e+00	2.4174271e+00	+
7	4.9140658e-01	1.0000000e+00	2.4059962e+00	+
8	1.8495509e-01	1.0000000e+00	6.5620535e-01	+
9	6.7969770e-02	1.0000000e+00	2.2116522e+00	+
10	8.8169985e-03	1.0000000e+00	1.0475959e-01	+
11	5.6685478e-04	1.0000000e+00	2.9539216e-01	+
12	1.1913120e-06	1.0000000e+00	1.1974893e-03	+
13	1.4918114e-11	1.0000000e+00	4.9345722e-05	+
14	8.8071795e-22	1.0000000e+00	3.2619050e-11	+

S'atura a l'òptim ($x^{14} = [1.0 \quad 1.0]^T$) on ($\|\nabla f(x^k)\| \approx 0$).

Modificació de la Hessiana

- Si $\nabla^2 f(x^k)$ no és definida positiva $d^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$ no té per què ser de descens (ni existir) i no garantiríem les condicions del Teorema de Zoutendijk: no podem garantir convergència global.
- Al mètode de Newton modificat usen B_k si $\nabla^2 f(x^k)$ no és prou def. pos.

Algorithm *Newton modificat*
  Punt inicial x^0, $k = 0$
  while x^k no és solució **do**
    Factoritzar $B_k = \nabla^2 f(x^k) + E_k$ on
      • $E_k = 0$ si $\nabla^2 f(x^k)$ és prou definida positiva
      • sino E_k s'escull de forma que B_k sigui prou definida positiva
    Calcular d^k: $B_k d^k = -\nabla f(x^k)$
    Calcular α^k que verifica Armijo-Wolfe
    $x^{k+1} = x^k + \alpha^k d^k$
    $k := k + 1$
  end_while
  Return: $x^* = x^k$
End_algorithm

Convergència global del mètode de Newton modificat

- Igual que al mètode de Newton, si B_k té un nombre de condició uniformement afitat, és a dir,

$$\exists C : \quad \|B_k\| \, \|B_k^{-1}\| \leq C \quad \forall k$$

(o el que és el mateix, B_k no és singular ni propera a ser-ho) llavors es demostra que

$$\cos\theta_k \geq \frac{1}{C}$$

i pel Teorema de Zoutendijk

$$\lim_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

- Llavors el mètode de Newton modificat té convergència global.

Convergència local del mètode de Newton modificat

- Si $\{x^k\} \to x^*$ i $\nabla^2 f(x^*)$ és "suficientment definida positiva", de forma que $E_k = 0$ a partir de un k' determinat. Llavors Newton modificat es comporta com Newton i

  el mètode de Newton modificat té convergència local quadràtica.

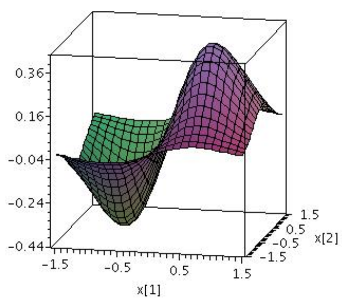
- Si $\{x^k\} \to x^*$ i $\nabla^2 f(x^*)$ no és "suficientment definida positiva" i és gairebé singular, $E_k \neq 0$ per a tot k, i la velocitat de convergència pot ser només lineal.

Alguns tipus de modificacions

1. Modificació de la descomposició espectral.
   - $\nabla^2 f(x^k) = V\Lambda V^T$, $V = [v_1|\ldots|v_n]$ matriu de vectors propis, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ matriu de valors propis.
   - Si $\nabla^2 f(x^k)$ no és "prou" def. pos., algun $\lambda_i < \epsilon$ i el modifiquem $\lambda_i + \delta_i$:
   $$B_k = V(\Lambda + \Delta)V^T = V\Lambda V^T + V\Delta V^T = \nabla^2 f(x^k) + E_k$$
   - Inconvenient: cal fer la descomposició espectral de $\nabla^2 f(x^k)$ (costós).
2. Modificació per addició d'una matriu diagonal
   - $B_k = \nabla^2 f(x^k) + \tau I$. Si τ "prou" gran llavors B_k "prou" def. pos.
   - Inconvenient: no sabem el valor de τ; es pot anar modificant iterativament fins que es prova (fent la factorització de Cholesky) que B_k que és "prou" def. pos. (costós).
3. Modificació de la factorització de Cholesky
   - Factorització de Cholesky: $\nabla^2 f(x^k) = LDL^T$, on $D = \text{diag}(d_1, \ldots, d_n)$. Si $\nabla^2 f(x^k)$ no és "prou" def. pos., algun $d_i < \epsilon$.
   - Es modifiquen (s'incrementen) els $d_i < \epsilon$ durant la factorització de Cholesky.
   - Finalment tenim una factorització: $LDL^T = P\nabla^2 f(x^k)P^T + E_k$.
   - Eficient.

Newton vs. Newton modificat: exemple numèric I

- Considerem el problema



$$f(x_1, x_2) = x_1 \cdot e^{-x_1^2 - x_2^2}$$

$$\nabla f(x) = e^{-x_1^2 - x_2^2} \begin{bmatrix} 1 - 2x_1^2 \\ -2x_1 x_2 \end{bmatrix}$$

$$\nabla^2 f(x) = e^{-x_1^2 - x_2^2} \begin{bmatrix} -6x_1 + 4x_1^3 & -2x_2 + 4x_1^2 x_2 \\ -2x_2 + 4x_1^2 x_2 & -2x_1 + 4x_1 x_2^2 \end{bmatrix}$$

- La solució usant les condicions d'optimalitat:

$$\nabla f(x) = 0 \quad \Rightarrow \quad x^+ = \begin{bmatrix} +\sqrt{\frac{1}{2}} \\ 0 \end{bmatrix} \quad x^- = \begin{bmatrix} -\sqrt{\frac{1}{2}} \\ 0 \end{bmatrix} \quad \nabla^2 f(x^-) = e^{-\frac{1}{2}} \begin{bmatrix} 4\sqrt{\frac{1}{2}} & \\ & 2\sqrt{\frac{1}{2}} \end{bmatrix}$$

$\nabla^2 f(x^-)$ és definida positiva i x^- és mínim. L'altre punt x^+ és un màxim.

Newton vs. Newton modificat: exemple numèric II

- Considerem el mètode de Newton a partir de $x^0 = [0 \;\; 0]^T$.

$$\nabla f(x^0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \nabla^2 f(x^0) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \Rightarrow \quad \nexists\, d^0 : \nabla^2 f(x^0) d^0 = -\nabla f(x^0)$$

La direcció de Newton no està definida. El mètode de Newton aborta, no pot anar a x^1.

- Considerem el mètode de Newton modificat a partir del mateix x^0

Mètode de Newton modificat (modificació LDL^T)

k	$f(x^k)$	α^k	$\|\nabla f(x^k)\|$	definida
0	0.0000000e+00	0.0000000e+00	1.0000000e+00	S+
1	-2.4292358e-02	2.1132487e-01	9.1176318e-02	S+
2	-4.1752913e-01	4.8000000e+00	2.0561645e-01	I
3	-4.2885970e-01	1.0000000e+00	8.7562332e-03	+
4	-4.2888194e-01	1.0000000e+00	3.0705391e-05	+
5	-4.2888194e-01	1.0000000e+00	3.8857291e-10	+

S'atura a l'òptim ($x^5 = [-0.70710678 \;\;\; 0]^T \approx [-\sqrt{\frac{1}{2}} \;\; 0]^T$) on ($\|\nabla f(x^k)\| \approx 0$).

Bibliografia

- D.P. Bertsekas, *Nonlinear Programming, 2nd Ed.*, 1999, Athena Scientific, Belmont, USA.
- D.G. Luenberger, Y. Ye., *Linear and Nonlinear Programming, 3rd Ed.*, 2008, Springer, New York, USA.
- J. Nocedal, S.J. Wright, *Numerical Optimization, 2nd Ed.*, 2006, Springer, New York, USA.