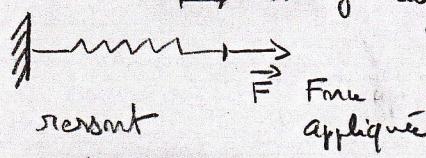


## VI

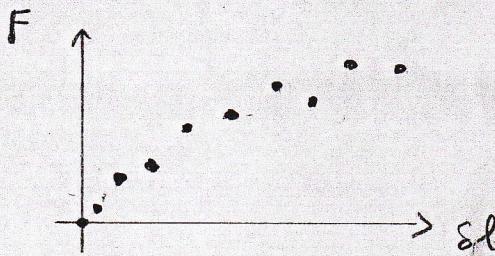
## Optimisation sans contrainte

Etant donné  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  suffisamment régulière (typiquement  $C^2$ ), nous allons étudier d'un point de vue analytique et numérique l'existence d'extrema de  $f$  (un extremum = un minimum ou un maximum). On parle d'optimisation sans contrainte (un problème d'optimisation avec contrainte étant posé sur un sous-ensemble de  $\mathbb{R}^n$ ). Il s'agira pour nous de trouver un minimum d'une fonction  $f$ , sans perte de généralité car maximiser une fonction  $f$  revient à chercher un minimum de  $f = -\tilde{f}$ .

Exemples :Méthode des mailles canes :

  $\delta l$  allongement du ressort obtenu expérimentalement  
ressort  $F$  force appliquée non linéaire

$$\text{On suppose : } F = V'(\delta l) = k \delta l + \alpha \delta l^2 + \beta \delta l^3$$



Pb : déterminer les paramètres  $k, \alpha, \beta$  qui correspondent le mieux aux mesures expérimentales ( $\mathbf{p}$  mesures  $(F_i, \delta l_i)$ )

Le système  $\begin{cases} F_i = k \delta l_i + \alpha \delta l_i^2 + \beta \delta l_i^3 \\ 1 \leq i \leq p \end{cases}$  ( $\mathbf{p}$  éq., 3 inconnues) ne

possède en général pas de solution pour  $p \geq 3$ . Soit  $\mathbf{x} = (k \ \alpha \ \beta)^T$

Ce système s'écrit  $A\mathbf{x} = \mathbf{F}$  où  $A \in M_{p \times 3}(\mathbb{R})$  et  $\mathbf{F} = (F_1, \dots, F_p)^T \in \mathbb{R}^p$ .

Méthode des mailles canes : on cherche  $\mathbf{x} \in \mathbb{R}^3$  qui minimise  $\|A\mathbf{x} - \mathbf{F}\|_2$ , ou de manière équivalente minimise  $\|A\mathbf{x} - \mathbf{F}\|_2^2 = f(\mathbf{x})$

La fonction  $\|Ax - F\|_2^2$  est quadratique, et nous verrons plus loin que la solution de ce problème est donnée par un système linéaire ayant une solution unique.

- Résolution d'un système linéaire  $Ax = b$ , avec

$A$  symétrique définie positive. Exemple d'application :

Résolution de l'équation de Poisson par différences finies.

Nous verrons que la solution du système minimise

$f(x) = \frac{1}{2} t_x A x - t_b x$ . Cette propriété permet d'introduire de nouvelles méthodes de résolution du système.

- Exemple de minimisation d'une fonction non quadratique :

$$f(x) = \frac{1}{2} \sum_{\substack{1 \leq i, j \leq N \\ i \neq j}} V(\|x_i - x_j\|)$$

Cristal constitué de  $N$  atomes, de positions  $x_i \in \mathbb{R}^3$ , interagissant par paires via un potentiel  $V$ .

$f$  représente l'énergie potentielle totale du cristal. La forme d'équilibre du cristal à  $T=0$  Kelvin est donnée par un minimum de l'énergie potentielle  $f$ .

Sous certaines hypothèses sur  $V$  et en deux dimensions d'espace, il a été démontré très récemment (F. Theil, 2005) que ce minimum est atteint pour un arrangement périodique des atomes.

# 1) Quelques résultats de base en calcul différentiel et optimisation:

## a) Etude locale des fonctions à n variables

Par la suite  $\Omega$  désigne un ouvert de  $\mathbb{R}^m$  et  $f$  une fonction de  $\Omega$  dans  $\mathbb{R}$ .

Def  $f$  est différentiable en  $x \in \Omega$  si il existe une application linéaire  $T: \mathbb{R}^m \rightarrow \mathbb{R}$  telle que pour  $h \approx 0$ :

$$f(x+h) = f(x) + Th + o(\|h\|)$$

L'application  $T$  est alors unique et on note  $T = Df(x)$ .  
 $T$  est appelée différentielle de  $f$  au point  $x$ .

Remarque: Si  $f$  est différentiable en  $x$  alors elle est continue en  $x$ .

Lemme Si  $f$  est différentiable en  $x$ , alors  $\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_m}(x)$  existent et  $Df(x)h = \left( \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_m}(x) \right) h$ .

Remarques - Par abus de langage, on confond souvent l'application linéaire  $Df(x)$  et sa matrice  $\left( \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_m}(x) \right)$  appelée matrice Jacobienne de  $f$  au point  $x$ .

- Le fait que  $\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_m}(x)$  existent n'implique pas que  $f$  est différentiable en  $x$ .

- On appelle  $\nabla f(x) = \left( \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_m}(x) \right)^T$  le gradient de  $f$  en  $x$ . Alors  $Df(x)h = \nabla f(x) \cdot h$ , où  $\cdot$  désigne le produit scalaire nouvel sur  $\mathbb{R}^m$ .

Def  $f$  est différentiable sur un ouvert  $\Omega$  si elle est différentiable en tout point de  $\Omega$ .

Def  $f: \Omega \rightarrow \mathbb{R}$  est  $C^1$  si  $f$  est différentiable sur  $\Omega$  et si l'application  $\Omega \rightarrow \mathbb{R}^m : x \mapsto \nabla f(x)$  est continue.

On peut montrer le résultat suivant:

Lemma:  $f: \Omega \rightarrow \mathbb{R}$  est  $C^1$  si et seulement si ses dérivées partielles  $\frac{\partial f}{\partial x_i}$  ( $i=1, \dots, n$ ) existent et sont continues sur  $\Omega$ .

Def:  $f: \Omega \rightarrow \mathbb{R}$  est  $C^2$  si  $f$  est  $C^1$  sur  $\Omega$  et ses dérivées partielles  $\frac{\partial^2 f}{\partial x_i \partial x_j}$  ( $i, j = 1, \dots, n$ ) sont  $C^1$  sur  $\Omega$ .

Lemme de Schwarz: (version simplifiée)

Soit  $f: \Omega \rightarrow \mathbb{R}$  de classe  $C^2$ . Alors pour tout  $x \in \Omega$ ,  $\forall i, j = 1, \dots, n$ :

$$\frac{\partial}{\partial x_i} \left( \frac{\partial f}{\partial x_j} \right) (x) = \frac{\partial}{\partial x_j} \left( \frac{\partial f}{\partial x_i} \right) (x).$$

Remarque: on notera ces dérivées  $\frac{\partial^2 f}{\partial x_i \partial x_j} (x)$ .

Théorème: (formule de Taylor à l'ordre 2)

Soit  $f: \Omega \rightarrow \mathbb{R}$  de classe  $C^2$ . Pour tout  $x \in \Omega$  et  $h \approx 0$ :

$$f(x+h) = f(x) + \nabla f(x) \cdot h + \frac{1}{2} {}^t h H_f(x) h + o(\|h\|^2)$$

avec  $H_f(x) \in M_n(\mathbb{R})$  définie par

$$(H_f(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

$H_f(x)$  est appelée matrice hessienne de  $f$  en  $x$ . (autre notation:  $H_f(x)$ )

Remarques:

- $H_f(x)$  est symétrique d'après le lemme de Schwarz.
- on appelle  $D^2 f(x)$  (différentielle seconde de  $f$  en  $x$ ) la forme bilinéaire symétrique  $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  définie par  $D^2 f(x)(h, y) = {}^t h H_f(x) y$

Def:  $f: \Omega \rightarrow \mathbb{R}$  admet un minimum local en  $x \in \Omega$  si il existe un voisinage ouvert  $U$  de  $x$  tel que  $f(x) \leq f(y) \forall y \in U$ .  $f$  admet un maximum local en  $x \in \Omega$  si  $f(x) \geq f(y) \forall y \in U$ .

Remarques: Supposons  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ . On parle de minimum ou maximum global lorsque  $U = \mathbb{R}^n$ .

Lemme: Lorsque les inégalités sont strictes (pour  $y \neq x$ ) on parle de minimum ou maximum strict.

Lemme: Soit  $f: \Omega \rightarrow \mathbb{R}$  de classe  $C^1$ . ( $\Omega$  ouvert de  $\mathbb{R}^n$ ). Si  $f$  admet un extrémum local en  $x \in \Omega$  alors  $\nabla f(x) = 0$ .

Remarques

- faux en général si  $\Omega$  n'est pas un ouvert (un extrémum peut être atteint sur le bord de  $\Omega$  dans que  $\nabla f$  s'y annule)
- $\nabla f(x) = 0$  peut être résolu p. ex. par la méthode de Newton.
- on peut avoir  $\nabla f(x) = 0$  dans que  $f$  admette un extrémum en  $x$ . Ex:  $f(x,y) = x^2 - y^2$  en  $(x,y) = (0,0)$

Lemme: Soit  $f: \Omega \rightarrow \mathbb{R}$  de classe  $C^2$ . On suppose qu'il existe  $x \in \Omega$  tel que  $\nabla f(x) = 0$ . Alors :

- Si les valeurs propres de  $Hf(x)$  sont  $> 0$ ,  $f$  admet un minimum local strict en  $x$ .
- Si les valeurs propres de  $Hf(x)$  sont  $< 0$ ,  $f$  admet un maximum local strict en  $x$ .
- Si les valeurs propres de  $Hf(x)$  sont  $\neq 0$  et pas toutes de même signe,  $f$  n'admet pas d'extrémum au point  $x$  ( $x$  est appelé un "point selle").

Remarque: Si  $Hf(x)$  n'est pas inversible, la nature du point  $x$  (extrémum de  $f$  ou non) dépend des termes d'ordre supérieur dans le développement de Taylor de  $f$  en  $x$ . Ex:  $f(x,y) = x^2 \pm$

Lemme 6: Soit  $f: \mathbb{R} \rightarrow \mathbb{R}$  de classe  $C^1$ .

Soit  $x_0 \in \mathbb{R}$  tel que  $\nabla f(x_0) \neq 0$ . L'équation  $f(x) = f(x_0)$  définit localement (pour  $x \approx x_0$ ) une hypersurface  $S$  (de dimension  $n-1$ ), qui admet un plan tangent en tout point  $x \approx x_0$ . Le plan tangent à  $S$  en  $x_0$  est orthogonal à  $\nabla f(x_0)$ .

Lemme: Sous les hypothèses précédentes,  $\nabla f(x_0)$  est orientée dans le sens des valeurs de  $f$  croissantes. Plus précisément:

$$\frac{d}{d\epsilon} f(x_0 + \epsilon \nabla f(x_0)) \Big|_{\epsilon=0} = \|\nabla f(x_0)\|^2 > 0.$$

b) Conditions suffisantes pour l'existence et l'unicité d'un minimum.

Voyons d'abord une condition suffisante pour l'existence d'un minimum.

Théorème Soit  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  continue et telle que  $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ .

Alors il existe  $x \in \mathbb{R}^n$  tel que  $f(x) \leq f(y) \quad \forall y \in \mathbb{R}^n$ .

R On dit que  $f$  admet un minimum global en  $x$ .

demo Si  $\|y\| \geq R$  avec  $R$  assez grand,  $f(x) \geq f(y)$ .

Donc  $\inf_{y \in \mathbb{R}^n} f(y) = \inf_{\|y\| \leq R} f(y) = f(x)$  avec  $\|x\| \leq R$ ,

puisque  $f$  est continue et que la boule  $\|y\| \leq R$  est compacte.

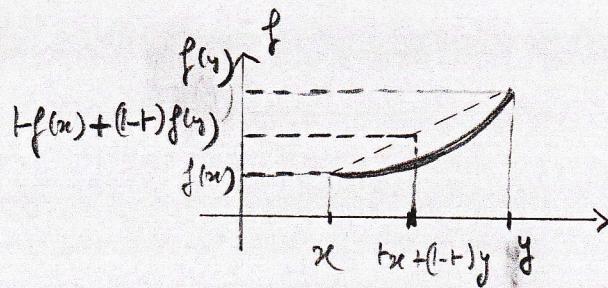
Le minimum de  $f$  peut ne pas être unique. Nous allons donner maintenant une condition suffisante d'unicité.

Def  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  est convexe si  $\forall x, y \in \mathbb{R}^n, \forall t \in [0, 1]$

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$$

Remarque Si  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  est convexe alors elle est continue sur  $\mathbb{R}^n$ .

Interprétation en dim 1



← fonction convexe  
(graphique au dessus de toute corde)

Def.  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  est strictement convexe si  $\forall x, y \in \mathbb{R}^n$  tels que  $x \neq y$ ,  $\forall t \in ]0, 1[$ ,

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y)$$

Théorème 2 Si  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  est strictement convexe, il existe au plus un  $x \in \mathbb{R}^n$  tel que  $f(x) = \min_{y \in \mathbb{R}^n} f(y)$ .

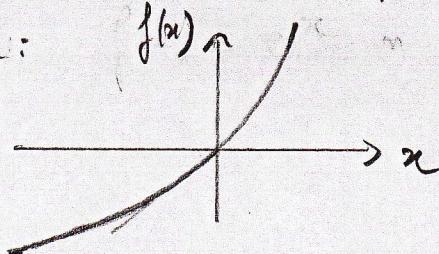
dém Supposons l'existence de deux minima en  $x_1$  et  $x_2$ . Alors

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2) = f(x_1) = \min_{x \in \mathbb{R}^n} f(x).$$

On arrive alors à une contradiction. □

Remarque Ce théorème ne donne pas l'existence d'un min.

Exemple:



← fonction strictement convexe qui tend vers  $+\infty$  lorsque  $x \rightarrow -\infty$

Avec les résultats de théorèmes 1 et 2 on obtient :

Théorème 3 Soit  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  strictement convexe et telle que  $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ . Alors il existe un unique  $x \in \mathbb{R}^n$  tel que  $f(x) = \min_{y \in \mathbb{R}^n} f(y)$ .

Nous allons maintenant relier les notions de point critique ( $\nabla f(x) = 0$ ) et minimum pour les fonctions convexes.

Le résultat suivant fournit une caractérisation utile de la convexité pour les fonctions  $C^1$ .

Lemme 1 Soit  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  de classe  $C^1$ .

- $f$  est convexe si et seulement si

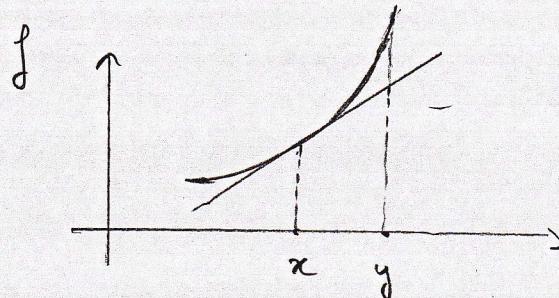
$$\forall x, y \in \mathbb{R}^n, \quad f(y) \geq f(x) + Df(x)(y-x)$$

- $f$  est strictement convexe si et seulement si

$$\forall x, y \in \mathbb{R}^n \text{ avec } x \neq y, \quad f(y) > f(x) + Df(x)(y-x)$$

(Résultat admis)

Interprétation en dim 1:



← fonction  
convexe  
(graphique dessus  
de toute tangente)

Cette propriété donne notamment le résultat suivant:

Théorème 4 Soit  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  de classe  $C^1$  et convexe.

$$\text{Alors } f(x) = \min_{y \in \mathbb{R}^n} f(y) \Leftrightarrow \nabla f(x) = 0$$

demo

$$\bullet \Rightarrow \text{ voir § 2), } \quad \text{ et } \quad \text{ et } \quad \text{ et }$$

$$\bullet \Leftarrow f(y) \geq f(x) + \underbrace{Df(x)(y-x)}_{=0} \quad \forall y \in \mathbb{R}^n.$$

Remarque: On peut donc calculer numériquement les minima de fonctions convexes en recherchant les zéros de  $x \mapsto \nabla f(x)$ . (p.ex. par la méth. de Newton)

On admettra la caractérisation suivante de la convexité pour des fonctions  $C^2$ :

Lemme 2 Soit  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  de classe  $C^2$ . Alors

$$\bullet \quad f \text{ est convexe} \Leftrightarrow \nabla^2 f(x) \text{ est définie positive } \forall x \in \mathbb{R}^n, \quad \forall y \in \mathbb{R}^n.$$

• si  $\nabla^2 f(x)$  est symétrique définie positive  $\forall x \in \mathbb{R}^n$  alors  $f$  est strictement

Remarque: pour  $f(x) = \frac{x^4}{12}$  (strictement convexe),  $Hf(x) = x^2$   
 $\Rightarrow Hf(0) = 0$  n'est pas sym. def. positive.

Application Soit  $A \in M_n(\mathbb{R})$  avec  $A$  symétrique définie positive

Soit  $A \in M_n(\mathbb{R})$  et  $f(x) = \frac{1}{2} b^T x - \frac{1}{2} x^T A x$ .  $A = (a_{ij})_{1 \leq i,j \leq n}$

$$f(x) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_j x_i - \sum_{i=1}^n b_i x_i$$

$$(Hf)_{ij} = \frac{1}{2} (a_{ij} + a_{ji}) \Rightarrow Hf = \frac{1}{2} (A + {}^t A) = A \text{ sym def} >$$

$\Rightarrow f$  est strictement convexe.

De plus,  $f(x) \rightarrow +\infty$  quand  $\|x\| \rightarrow +\infty$  car

( $\lambda$  désigne la plus petite valeur propre de  $A$ , qui est  $> 0$ )

$$f(x) \geq \frac{\lambda}{2} \|x\|_2^2 - \|b\|_2 \|x\|_2 \rightarrow +\infty \text{ quand } \|x\|_2 \rightarrow +\infty.$$

Donc (Thm 3) il existe un unique  $x \in \mathbb{R}^n$  /  $\underset{\mathbb{R}^n}{\operatorname{Min}} f = f(x)$ .

Cette propriété est équivalente à  $\nabla f(x) = 0$  (Thm 4).

$$\frac{\partial f}{\partial x_i} = \frac{1}{2} \sum_{j=1}^n (a_{ij} + a_{ji}) x_j - b_i \Rightarrow \nabla f(x) = \frac{1}{2} (A + {}^t A)x - b.$$

Puisque  $A$  est symétrique,  $\nabla f(x) = Ax - b$ . Donc

$$Ax = b \Leftrightarrow f(x) = \underset{\mathbb{R}^n}{\operatorname{Min}} f, \text{ avec } f(x) = \frac{1}{2} b^T x - \frac{1}{2} x^T A x$$

Cela permet de reformuler la résolution du système  $Ax = b$  comme un problème de minimisation.

## 2) Quelques méthodes numériques pour l'optimisation dans contraintes:

Nous abandonnons maintenant le calcul numérique d'un minimum de  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  de classe  $C^1$ . On suppose que  $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ , de sorte que ce minimum existe.

Nous allons d'abord voir des méthodes de gradient, qui sont des algorithmes itératifs utilisant uniquement  $f$  et  $\nabla f$ .

L'exemple le plus simple d'une telle méthode est l'algorithme du gradient à pas constant :

$$\begin{cases} x_{k+1} = x_k - \rho \nabla f(x_k) \\ x_0 \in \mathbb{R}^n \text{ donné.} \end{cases} \quad \underline{\rho > 0 \text{ fixé}}$$

Cette méthode est motivée par la propriété que  $-\nabla f$  est orientée dans le sens des valeurs de  $f$  décroissantes (cf p. 6)

On dit alors que  $-\nabla f(x_k)$  est une direction de descente en  $x_k$ .

La méthode du gradient à pas constant est assez peu utilisée en pratique car elle conduit facilement à des instabilités numériques. Par exemple, pour  $f(x) = x^4$  (fonction strictement convexe) on obtient  $x_{k+1} = x_k(1 - 4\rho x_k^2)$ . Si  $x_0^2 \geq \frac{1}{\rho}$ , on montre par récurrence que  $|x_{k+1}| \geq 3|x_k|$  (car  $1 - 4\rho x_k^2 \leq -3$ )

et donc  $|x_k| \xrightarrow{k \rightarrow \infty} +\infty$ . L'efficacité de la méthode consiste dans le fait

Pour éviter ce type de phénomène on peut considérer la méthode de plus grande pente (ou steepest-descent method) dans laquelle  $\rho$  est adapté à chaque itération de manière optimale :

$$\begin{cases} x_0 \in \mathbb{R}^n \text{ donné} \\ x_{k+1} = x_k - \rho_k \nabla f(x_k) \end{cases}$$

$$f(x_k - \rho_k \nabla f(x_k)) = \min_{\rho \geq 0} f(x_k - \rho \nabla f(x_k))$$

A chaque étape de l'itération il faut donc résoudre un problème de minimisation en une dimension; plus précisément minimiser la fonction  $\varphi: \mathbb{R}^+ \rightarrow \mathbb{R}$

$$p \mapsto f(x_k - p Df(x_k)) := \varphi(p),$$

un minimum étant atteint en  $p = p_k$ . (Ce minimum existe (sans être nécessairement unique) puisque  $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ ).

Calcul de  $p_k$ : il y a plusieurs possibilités

- méthode de Newton ou méthode de la seconde pour résoudre  $\varphi'(p) = 0$ . Noter que c'est une condition nécessaire mais en général non pas suffisante pour obtenir un minimum.

Cependant, si  $f$  est convexe alors  $\varphi$  est aussi convexe (c'est la restriction de  $f$  à une droite passant par  $x_k$ );

Dans ce cas  $\varphi'(p) = 0 \Leftrightarrow \varphi(p) = \min_{y \in \mathbb{R}} \varphi(y)$

pour tout  $y \in \mathbb{R}$ , c'est à dire lorsque  $p = p_k$ .

- dichotomie: Posons  $a=0$ .

On suppose  $\varphi: [a, b] \rightarrow \mathbb{R}$  unimodale, c.a.d.  $\exists p^* \in ]a, b[$  tel que  $\varphi' < 0$  sur  $]a, p^*[$  et  $\varphi' > 0$  sur  $]p^*, b[$ . On pose  $\delta = \frac{b-a}{4}$ ,  $x_i = a + i\delta$ . Selon la position relative des  $f(x_i)$  ( $i=1, 2, 3$ ) on peut choisir  $a' < b'$  tels que  $\varphi$  soit unimodale sur  $[a', b'] \subset [a, b]$  et  $b' - a' = \frac{1}{2}(b-a)$ . On recommence l'opération sur  $[a', b']$  jusqu'à atteindre la précision souhaitée.

- Cas particulier d'une fonction quadratique:

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

$A \in M_n(\mathbb{R})$  symétrique  
définie positive,  $b \in \mathbb{R}^n$

Néanmoins  $\nabla \mathbf{x}_k = \nabla f(\mathbf{x}_k) = A\mathbf{x}_k - \mathbf{b} \neq 0$  (sinon le min est déjà atteint!) (1)

$$\varphi'(p_k) = 0 \Leftrightarrow \mathbf{r}_{k+1} \cdot \mathbf{r}_k = 0 \Leftrightarrow (\underbrace{A\mathbf{x}_k - p_k A\mathbf{r}_k - \mathbf{b}}_{A\mathbf{x}_{k+1}}) \cdot \mathbf{r}_k = 0$$

On obtient donc explicitement :

$$p_k = \frac{\|\mathbf{r}_k\|_2^2}{\mathbf{r}_k^T A \mathbf{r}_k}$$

avec  $\mathbf{r}_k^T A \mathbf{r}_k \neq 0$  puisque  $A$  est symétrique définie positive.

Remarque:

En pratique le calcul de  $p_k$  n'a pas besoin d'être réalisé avec une très grande précision.

On peut montrer que la méthode de la plus grande pente converge pour toute condition initiale  $\mathbf{x}_0$  si  $f$  est strictement convexe. La CV est linéaire et peut donc être assez lente.

Pour avoir une convergence plus rapide, on peut utiliser la méthode de Newton pour résoudre  $\nabla f(\mathbf{x}) = 0$ . En particulier, si  $f$  est convexe on obtient ainsi forcément un minimum de  $f$  (voir thm 4 p. 8). Il existe par ailleurs des variantes moins coûteuses que Newton et efficaces, comme la méthode de Broyden.

Une autre méthode beaucoup utilisée est la méthode du gradient conjugué. Soit  $f: \mathbb{R}^m \rightarrow \mathbb{R}$  de classe  $C^2$ , avec  $\nabla f(\mathbf{x}) \xrightarrow[\|\mathbf{x}\| \rightarrow +\infty]{} +\infty$  et  $H_f(\mathbf{x})$  symétrique définie positive  $\forall \mathbf{x} \in \mathbb{R}^m$ .  $f$  possède alors un minimum global strict  $\bar{\mathbf{x}} \in \mathbb{R}^m$ . La méthode du gradient conjugué utilise un direction de descente plus efficace que  $\nabla f(\mathbf{x}_k)$ , qui fait également appel à  $\nabla f(\mathbf{x}_{k-1})$ . Nous allons étudier cette méthode lorsque  $f$  est une fonction quadratique, mais elle s'applique dans un cadre plus général.

### 3) Méthode du gradient conjugué pour une fonction quadratique

On considère  $f(x) = \frac{1}{2} \|Ax - b\|^2$  avec  $A \in M_n(\mathbb{R})$  symétrique

définie positive et  $b \in \mathbb{R}^n$ . Nous avons vu que  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  admet un minimum global strict en  $x = \bar{x}$  avec  $A\bar{x} = b$ .

La méthode du gradient conjugué définit une suite  $(x_k)$  qui converge vers  $\bar{x}$ . Nous allons voir que la convergence se fait en un nombre fini d'itérations  $\leq n$ ; de ce point de vue la méth. du gradient conjugué est donc à classer parmi les méthodes directes. Cependant,

à cause des erreurs d'arrondis, cette propriété n'est pas vérifiée en pratique (plus particulièrement pour de grands systèmes) et la méthode est plutôt considérée comme itérative. On contrôlera donc cet algorithme par un nombre maximal d'itérations et par un test d'arrêt.

#### a) Description de la méthode:

On notera par la suite  $r_k = \nabla f(x_k) = Ax_k - b$ . Si  $r_k = 0$  alors l'algorithme s'arrête ( $x_k = \bar{x}$ ).

#### i) Initialisation:

- on fixe  $x_0 \in \mathbb{R}^n$ .

- Si  $r_0 = 0$  alors l'algorithme s'arrête car  $x_0 = \bar{x}$ .

- Si  $r_0 \neq 0$ , on calcule  $x_1$  par la méthode de plus grande pente.

On pose  $w_0 = \nabla f(x_0)$ .  $-w_0$  est direction de descente pour calculer  $x_1$ .

$$x_1 = x_0 - \rho_0 w_0, \quad f(x_0 - \rho_0 w_0) = \min_{\rho \geq 0} f(x_0 - \rho w_0)$$

Remarque minimum explicite car on minimise un polynôme de degré 2 en  $\rho$ .

#### ii) Itération (échec pour calculer $x_2$ )

On suppose connus  $x_k$  et  $w_{k-1}$  ( $-w_{k-1}$  est la direction de descente utilisée pour calculer  $x_k$ ).

- Si  $r_k = 0$  alors l'algorithme s'arrête car  $x_k = \bar{x}$ .

- Si  $r_k \neq 0$ : on pose

$$w_k = r_k + \theta_k w_{k-1},$$

( $-w_k$  = direction de descente pour calculer  $x_{k+1}$ )

$$\theta_k = \frac{t_{r_k} (r_k - r_{k-1})}{\|r_{k-1}\|_2^2} \quad (1)$$

$$x_{k+1} = x_k - p_k w_k, \quad f(x_k - p_k w_k) = \min_{p \geq 0} f(x_k - p w_k)$$

Dans le cas présent où  $f$  est quadratique, la valeur de  $p_k$  est connue explicitement (voir le lemme 1 qui suit).

Nous allons montrer les résultats suivants : (en particulier,  $r_k \neq 0$  implique  $w_k \neq 0$  puisque  $r_k \perp w_{k-1}$ )

### Lemme 1

i)  $f(x_{k+1}) = \min_{\theta \in \mathbb{R}} \min_{p \geq 0} f[x_k - p(r_k + \theta w_{k-1})] \quad (\text{cela motive le choix (1)})$

ii)  $t_{r_k} w_{k-1} = 0, \quad p_k = \frac{\|r_k\|_2^2}{t_{w_k} A w_k} \quad (2)$

iii)  $t_{w_k} A w_{k-1} = 0 \quad (w_k \text{ et } w_{k-1} \text{ sont dits "A-conjugués"})$

Lemme 2 :  $t_{r_k} r_{k-1} = 0$  et (1) se simplifie en :

$$\theta_k = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} \quad (3)$$

Remarque: les formules (1) et (3) sont équivalentes pour une fonction  $f$  quadratique. Pour  $f$  plus générale, (1) correspond à la méthode de Polak-Ribiére et (3) à celle de Fletcher-Reeves. La méthode du gradient conjugué dans le cas quadratique est due à Hestenes et Steifel (1952).

### b) Preuve du lemme 1.

Nous allons montrer successivement ii), i) et iii).

Tout d'abord, puisque  $f(x_{k-1} - \rho w_{k-1}) = \min_{\rho \geq 0} f(x_{k-1} - \rho w_{k-1})$  (1)

on a  $\nabla f(x_{k-1} - \rho w_{k-1}) \cdot w_{k-1} = 0$  soit  $\nabla f(x_{k-1} - \rho w_{k-1}) \cdot w_{k-1} = 0 \Rightarrow$  on a montré la 1<sup>e</sup> égalité.

Pour  $w = x_k + \theta w_{k-1}$  on a (polynôme du second degré en  $\rho$ )

$$\begin{aligned} f(x_k - \rho w) &= f(x_k) - \rho \nabla f(x_k) \cdot w + \frac{1}{2} \rho^2 t_w A w \\ &= f(x_k) - \rho \nabla f(x_k) \cdot w + \frac{1}{2} \rho^2 t_w A w \end{aligned}$$

Puisque  $\nabla f(x_k) \cdot w_{k-1} = 0$ ,  $\nabla f(x_k) \cdot w$  est indépendant de  $\theta$  et on obtient :

$$f(x_k - \rho w) = f(x_k) - \rho \|\nabla f(x_k)\|^2 + \frac{1}{2} \rho^2 t_w A w \quad (5)$$

Le minimum de ce polynôme de degré 2 est atteint en :

$$\rho_\theta = \frac{\|\nabla f(x_k)\|^2}{t_w A w} \quad (2^e \text{ égalité de (5)})$$

et vaut

$$f(x_k - \rho_\theta w) = f(x_k) - \frac{1}{2} \frac{\|\nabla f(x_k)\|^2}{t_w A w}$$

Pour minimiser  $f(x_k - \rho_\theta w)$  suivant  $\theta$  il faut minimiser  $t_w A w$ .  
c'est à dire  $|w|$ . Il faut choisir pour cela  $w = w_k$  tel que  
 $t_w A w_k w_{k-1} = 0$ , ce qu'on notera  $w_k \perp w_{k-1}$ :

$$\langle \nabla f(x_k) + \theta w_{k-1}, \nabla f(x_k) + \theta w_{k-1} \rangle = \|\nabla f(x_k)\|^2 + 2\theta \langle \nabla f(x_k), w_{k-1} \rangle + \theta^2 \|w_{k-1}\|^2$$

$$\text{Minimum pour } \theta = \theta_k = -\frac{\langle \nabla f(x_k), w_{k-1} \rangle}{\|w_{k-1}\|^2}. \quad \text{Donc (5)}$$

$$\text{Donc : } w_k = \nabla f(x_k) - w_{k-1} \frac{\langle \nabla f(x_k), w_{k-1} \rangle}{\|w_{k-1}\|^2} \quad (6)$$

d'où  $w_k \perp w_{k-1}$ . Afin de montrer le lemme (1), il reste à montrer que (5) correspond bien à (1).

a) (1). D'une part:

$$r_k - r_{k-1} = A(r_k - r_{k-1}) = -\rho_{k-1} Aw_{k-1} \text{ donc}$$

$$^t r_k (r_k - r_{k-1}) = -\rho_{k-1} \langle r_k, w_{k-1} \rangle \quad (7)$$

D'autre part

$$\|w_{k-1}\|^2 = (Aw_{k-1}, w_{k-1}) = -\frac{1}{\rho_{k-1}} (A(r_k - r_{k-1}), w_{k-1})$$

$$= -\frac{1}{\rho_{k-1}} (r_k - r_{k-1}, w_{k-1})$$

$$= \frac{1}{\rho_{k-1}} (r_{k-1}, w_{k-1}) \quad (\text{car } \langle r_k, w_{k-1} \rangle = 0)$$

$$= \frac{1}{\rho_{k-1}} (r_{k-1}, r_{k-1} + \theta_{k-1} w_{k-2})$$

$$= \frac{1}{\rho_{k-1}} \|r_{k-1}\|^2 \quad (\text{car } \langle r_{k-1}, w_{k-2} \rangle = 0)$$

$$\text{donc } \|r_k\|^2 = \rho_{k-1} \|w_{k-1}\|^2. \quad (8)$$

Avec (5), (7) et (8) on obtient donc:

$$\frac{^t r_k (r_k - r_{k-1})}{\|r_{k-1}\|^2} = -\frac{\langle r_k, w_{k-1} \rangle}{\|w_{k-1}\|^2} = \theta_k.$$

On obtient donc bien la formule (1) plus simple pour le calcul de  $\theta_k$ .

c) Convergence de la méthode du gradient conjugué et preuve du lemme 2<sup>(1)</sup>

Supposons  $\pi_k \neq 0$  pour  $k=0, \dots, n-1$  (si  $\pi_k = 0$  l'algorithme converge) si cela implique  $P_k \neq 0$  pour  $k=0, \dots, n-1$ , et donc

Lemme 3 Pour tout  $k=1, \dots, n$  on a :

$$(P_k) \quad \left\{ \begin{array}{l} \pi_k \cdot w_q = 0 \quad \text{pour } q=0, \dots, k-1 \\ \langle w_{k+1}, Aw_q \rangle = 0 \quad \text{pour } q=0, \dots, k-1, \quad \langle x, y \rangle := \langle Ax, y \rangle \\ \pi_k \cdot \pi_q = 0 \quad \text{pour } q=0, \dots, k-1. \end{array} \right.$$

demo : par récurrence. On considère les produits scalaires  $\langle x, y \rangle = \langle Ax, y \rangle = \langle x, Ay \rangle$ .

•)  $P_0$  est vraie :  $\pi_0 \cdot \pi_0 = x_0 \cdot w_0 = 0$  (condition d'optimalité de  $P_0$ )

$$\langle w_1, w_0 \rangle = 0 \quad (\text{d'après le lemme 1}).$$

• Supposons  $P_k$  vraie et montrons  $P_{k+1}$  : ( $k \leq n-1$ ) et on a :

$$\pi_{k+1} \cdot w_k = 0 \quad (\text{condition d'optimalité de } P_k)$$

$$\pi_{k+1} \cdot w_q = (Ax_{k+1} - b, w_q) = (A(x_{k+1} - x_k) + Ax_k - b, w_q)$$

$$= -P_k \langle w_k, w_q \rangle + \pi_k \cdot w_q$$

$$= 0 \quad \text{pour } q=0, \dots, k-1 \quad (\text{hyp. de récurrence } P_k).$$

$$\text{Donc } \pi_{k+1} \cdot w_q = 0 \quad \text{pour } q=0, \dots, k.$$

$$\text{Par ailleurs, } \pi_{k+1} \cdot \pi_q = \pi_{k+1} \cdot (w_q - \theta_q w_{q-1}) \quad (\text{avec } \theta_0 := 0 \text{ car } \pi_0 = w_0)$$

$$\text{donc } \pi_{k+1} \cdot \pi_q = 0 \quad \text{pour } q=0, \dots, k.$$

Ensuite  $\langle w_{k+1}, w_k \rangle = 0$  + d'après le lemme 1, et pour

$$q=0, \dots, k-1:$$

$$\langle w_{k+1}, w_q \rangle = \langle \pi_{k+1}, w_q \rangle + \theta_{k+1} \langle w_k, w_q \rangle = \langle \pi_{k+1}, w_q \rangle \quad \text{par l'hyp de récurrence } (P_k)$$

$$r_{q+1} - r_q = A(x_{q+1} - x_q) = -\rho_q A w_q \text{ alors}$$

$$\langle w_{q+1}, w_q \rangle = \langle r_{q+1}, w_q \rangle = \frac{1}{\rho_q} (r_{q+1} - r_{q+1} + r_q) \quad (\rho_q \neq 0)$$

$$= 0 \quad \begin{array}{l} \text{(car } \rho_q \neq 0 \text{ et } \\ \text{car } 0 \leq q \leq k-1 \text{ et } \\ \text{car } \langle x_{q+1}, r_p \rangle = 0 \text{ pour } p = 0, \dots, k \text{.)} \end{array}$$

Cela prouve P<sub>k</sub> par récurrence.

□

En conclusion, la famille  $(w_0, \dots, w_{n-1})$  est libre car les  $w_i$  sont deux à deux orthogonaux pour le produit scalaire  $\langle x, y \rangle = {}^t x A y$ . C'est donc une base de  $\mathbb{R}^n$ . Puisque  $r_m$  est orthogonal à  $w_0, \dots, w_{m-1}$ , on a donc  $r_m = 0$ . Nous avons donc montré que  $Ax_m = b$ , i.e.  $x_m =$

Théorème :

Soit  $A \in M_n(\mathbb{R})$  symétrique définie positive,  $b \in \mathbb{R}^n$  et  $f(x) = \frac{1}{2} {}^t x A x - {}^t b x$ . Alors l'algorithme du gradient conjugué donné au § a) définit une suite  $(x_k)_{k=0, \dots, p}$  avec  $p \leq n$  et  $Ax_p = b$ . On a  $f(x_p) = \min_{x \in \mathbb{R}^n} f(x)$ .

Remarque Le cas  $p < n$  est exceptionnel.

Enfin, nous avons montré dans le lemme 3 que  $r_k \cdot r_{k-1} = 0$ , ce qui prouve le lemme 2.