INU Champollion

Licence Info 3ème Année

# Maths pour l'I.A.

Thierry Montaut

# Plan du cours

# Formule de Taylor à l'ordre 2

Soit $f : \mathbb{R}^n \to \mathbb{R}$ de classe $\mathcal{C}^2$ et $a \in U$
On cherche à évaluer plus finement l'approximation fournie par la formule de Taylor à l'ordre 2 :

$$f(x) = f(a) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a).(x_i - a_i) + \|(x_i - a_i)\|\varepsilon(x_i - a_i)$$

## Théorème

*(formule de Taylor à l'ordre 2)*
*Soit* $f : \mathbb{R}^n \to \mathbb{R}$ *de classe* $\mathcal{C}^2$ *et* $a \in U$. *On note* $x = a + h$.
*Il existe une fonction* $\varepsilon : \mathbb{R} \to \mathbb{R}$ *telle que* $\varepsilon(h) \underset{h \to 0_{\mathbb{R}^n}}{\longrightarrow} 0$, *vérifiant :*

$$f(x) - f(a) = df_a(h) + \frac{1}{2} Q_a(h) + \|h\|^2 \varepsilon(h)$$

*où*

$$df_a(h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a).h_i$$

$$Q_a(h) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) h_i h_j$$

## Définition

*La matrice* $H_a \in \mathcal{M}_n(\mathbb{R})$ *définie par*

$$h_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

*est appelée matrice hessienne de* $f$ *en* $a$*.*
*Elle est carrée, symétrique et réelle donc...diagonalisable dans une BON de vecteurs propres.*

# Extrema d'une fonction numérique

## Définition

*Soit* $a \in U$ *et*

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

• *On dit que* $f$ *admet un minimum local (respectivement un minimum local strict) en* $a$ *s'il existe un réel* $\alpha > 0$ *tel que :*

$$\forall x \in U \cap B(a,\alpha), \ \ f(x) \geq f(a)$$

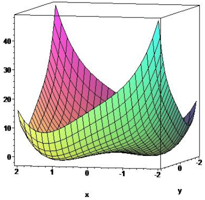
*(respectivement tel que :*

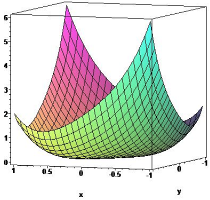
$$\forall x \in U \cap B(a,\alpha), \ \ f(x) > f(a))$$

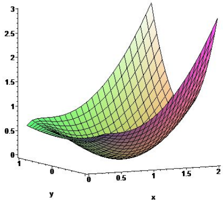
• *On dit que* $f$ *admet un minimum global (respectivement un minimum global strict) en* $a$ *s'il existe un réel* $\alpha > 0$ *tel que :*

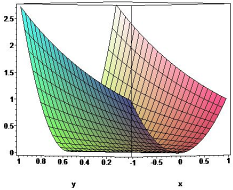
$$\forall x \in U, \ \ f(x) \geq f(a)$$

# Quelques minima



## Définition

• *On retrouve des définitions analogues avec un maximum, local ou global, strict ou pas...*

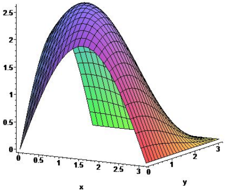
• *On appelle plus généralement extremum un maximum ou un minimum.*

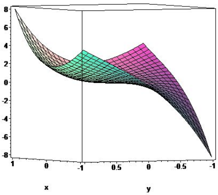
*Rien que pour faire mon intéressant, j'utiliserai les pluriels latins en "a".*

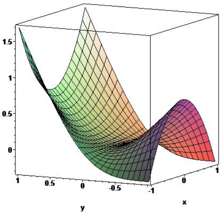
**Remarque** : $f$ présente un extremum local en $a$ si que $f(a+h)-f(a)$ est de signe constant sur un voisinage de 0.

Au contraire, $f$ ne présente pas d'extremum local en $a$ si $f(a+h)-f(a)$ change de signe sur TOUT voisinage de 0.

# Un maximum et deux points selle



# Condition nécessaire

## Théorème

*existence d'extrema : condition nécessaire*
*Soit* $f : \mathbb{R}^n \to \mathbb{R}$ *une fonction différentiable sur* $U$ *et* $a \in \mathbb{R}^n$.
*Si* $f$ *présente un extremum local en* $a$ *alors*

$$\forall i \in [\![1, n]\!], \quad \frac{\partial f}{\partial x_i}(a) = 0.$$

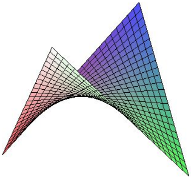
*Un tel point est dit point singulier (ou point critique...ou stationnaire...).*

**Remarque** : Cette condition n'est pas suffisante :

**Exercice 1** : Soit la fonction

$$f \ : \left( \begin{array}{c} \mathbb{R}^2 \rightarrow \mathbb{R} \\ (x,y) \mapsto xy \end{array} \right. .$$

Montrer que $(0,0)$ est un point stationnaire de $f$ mais qu'aussi près qu'on le veut de $(0,0)$, il existe des points $x$ et $y$ tels que $f(x) > f((0,0))$ et $f(y) < f((0,0))$.



# Condition suffisante cas d'une fonction de $\mathbb{R}^n \to \mathbb{R}$.

Si $a$ est un point singulier de $f$ d'après la formule de Taylor à l'ordre 2,

$$\Delta(h) = f(a+h) - f(a) = \frac{1}{2} Q_a(h) + \|h\|^2 \varepsilon(h)$$

Donc $f$ présente un extremum en $a$ ssi $\Delta$ est localement de signe constant.

## Théorème

*Si a est un point singulier de* $f : U \subset \mathbb{R}^n \to \mathbb{R}$ *de classe* $\mathcal{C}^2$ *sur U alors :*

1. *Si toutes les valeurs propres de* $H_a$ *sont strictement positives, alors* $f$ *présente un minimum local strict en a.*
2. *Si toutes les valeurs propres de* $H_a$ *sont strictement négatives, alors* $f$ *présente un maximum local strict en a.*
3. *Si* $H_a$ *admet deux valeurs propres non nulles de signe opposé alors* $f$ *n'a pas d'extremum en a (on dit alors que a est un point col ou un point selle suivant qu'on préfère la montagne ou les indiens)*
4. *Dans le cas ou les valeurs propres sont de même signe mais que 0 est valeur propre on ne peut conclure sans une étude plus fine...*

# Condition suffisante cas d'une fonction de $\mathbb{R}^2 \to \mathbb{R}$.

Dans le cas d'une fonction de deux variables, la matrice hessienne est de taille 2.
On pose (en utilisant les notations dites de Monge :

$$p = \frac{\partial f}{\partial x}(a); \quad q = \frac{\partial f}{\partial y}(a); \quad r = \frac{\partial^2 f}{\partial x^2}(a); \quad s = \frac{\partial^2 f}{\partial x \partial y}(a); \quad t = \frac{\partial^2 f}{\partial y^2}(a)$$

Alors

$$H_a = \begin{pmatrix} r & s \\ s & t \end{pmatrix}$$

On note enfin

$$\Delta = rt - s^2.$$

Alors

## Théorème

*existence d'extrema : condition suffisante*

*Soit $f$ une fonction de classe $\mathcal{C}^2$ d'un ouvert $U$ de $\mathbb{R}^2$ sur $\mathbb{R}$, et $a \in U$ un point stationnaire de $f$.*

1. *Si $\Delta > 0$, alors $f$ admet un extremum local strict en $a$. (dans ce cas $rt > 0$ donc $r$ et $t$ sont de même signe)*
   - *Si $r$ (ou $t$) $> 0$ : C'est un minimum.*
   - *Si $r$ (ou $t$) $< 0$ : C'est un maximum.*
2. *Si $\Delta < 0$, alors $f$ n'admet pas d'extremum en $a$. mais un point selle.*
3. *Si $\Delta = 0$, il faudra effectuer une étude plus fine "à la main"...*

**Exercice 2** : Montrer que la fonction :

$$f \ : \left( \begin{array}{c} \mathbb{R}^2 \to \mathbb{R} \\ (x,y) \mapsto x^3 + y^3 - 3xy \end{array} \right.,$$

admet deux points stationnaires. En étudiant le signe de $\Delta$ pour chacun d'eux, montrer que l'un est un minimum local strict et que l'autre est un point col.

**Exercice 3** : Étudier de même les extrema des fonctions suivantes de $\mathbb{R}^2$ dans $\mathbb{R}$.

**1°)** $x^3 + 3xy^2 - 15x - 12y$,
**2°)** $x^4 + y^4 - 4xy$,
**3°)** $(x - y)^2 + (x + y)^4$,
**4°)** $(x - y)^2 + (x + y)^3$,
(Étudier la restriction à la droite d'équation $y = x$.)
**5°)** $x^2y + \ln(1 + y^2)$
(Étudier la restriction à la courbe d'équation $y = x^3$.)

# Minimisation numérique : Descente de gradient

- La méthode de descente de gradient est une méthode numérique permettant de déterminer une valeur approchée des extrema d'une fonction d'une ou plusieurs variables réelles.
- On l'utilise dans les cas où on ne sait pas résoudre exactement le problème de minimisation et où des valeurs approchées de ce minimum suffise.
- Cette méthode est très répandue en Machine Learning notamment pour l'optimisation des problèmes de régression et la phse d'apprentissage des réseaux de neurones.

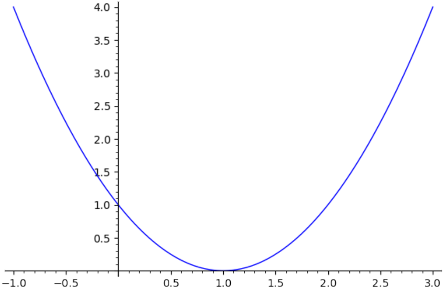
# Minimisation numérique : Descente de gradient

- L'idée centrale de cette méthode est que la dérivée (et plus généralement le vecteur gradient) donne la direction et le sens de plus grande augmentation de la fonction $f$. Symétriquement, l'opposé du vecteur gradient donne la direction et le sens de plus grande diminution.

# Cas d'une fonction réelle

Exemple : On cherche à déterminer le minimum de la fonction réelle définie par $f(x) = x^2 - 2x + 1$.

Cette fonction est de classe $\mathcal{C}^1$ sur $\mathbb{R}$, on sait parfaitement étudier sa dérivée, ses variations et donc établir qu'elle possède un unique minimum en 1 valant 0.



# Descente de gradient

Lorsque la résolution analytique est trop complexe, on cherche à utiliser une méthode numérique et itérative d'approximation de ce minimum : Si $f$ est de classe $\mathcal{C}^1$ :

Initialiser $x_0$ à une valeur quelconque
Tant qu'il n'y a pas convergence :

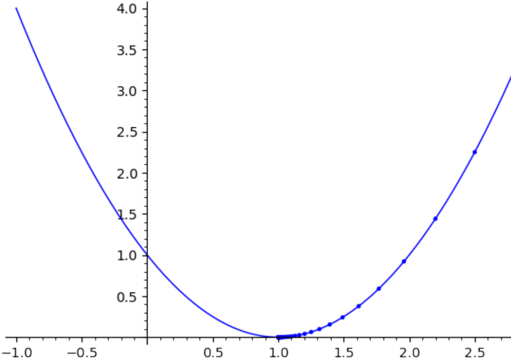
$$x_{k+1} = x_k - \alpha . f'(x_k)$$

# Descente de gradient

- $f'(x_k)$ donne le sens de plus grande variation. On utilise - pour une minimisation et + pour une recherche de maximum
- $\alpha$ est le pas de la méthode. C'est un paramètre important qui permet d'assurer la convergence et la vitesse de convergence de la méthode.
- La condition de sortie de boucle peut dépendre du nombre d'itérations, de la différence $x_{k+1} - x_k$ ou de la valeur de la dérivée $f'(x_k)$.

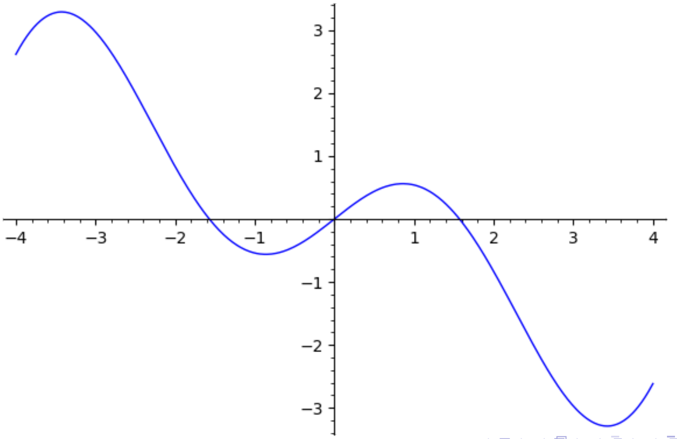
```
Entrée [7]: X=grad1(f,2.5,0.1,0.001,100)

1  :  2.20000000000000
2  :  1.96000000000000
3  :  1.76800000000000
4  :  1.61440000000000
5  :  1.49152000000000
6  :  1.39321600000000
7  :  1.31457280000000
8  :  1.25165824000000
9  :  1.20132659200000
10  :  1.16106127360000
11  :  1.12884901888000
12  :  1.10307921510400
13  :  1.08246337208320
14  :  1.06597069766656
15  :  1.05277655813325
```



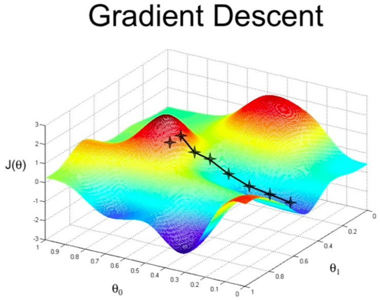
# Les défauts de la méthode

- Peut se faire piéger dans un minimum local si la fonction n'est pas convexe.
- La convergence peut être très lente dans les zones où la dérivée est très faible.



# Cas d'une fonction de plusieurs variables

L'intérêt de cette méthode est qu'elle se généralise très simplement à une fonction de plusieurs variables en remplaçant la dérivée par le vecteur gradient



# Descente de gradient

Si $f$ est de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$ :

```
Initialiser X_0 à une valeur quelconque
Tant qu'il n'y a pas convergence :
```

$$X_{k+1} = X_k - \alpha . \bigtriangledown f(X_k)$$