

Optimisation et réseaux de neurones

23 février 2024

Table des matières

1 Fonctions de plusieurs variables et optimisation	5
I Définition et exemples	5
1 Définition	5
2 Deux variables	6
3 Trois variables	6
4 n variables	7
II Graphe	7
1 Définition	7
2 Tranches	9
3 Minimum, maximum	11
4 Recherche élémentaire d'un minimum	14
III Lignes de niveau	16
1 Définition	16
2 Exemples	18
3 Surfaces quadratiques	19
4 Régression linéaire	24
IV Dérivées partielles	26
1 Définition	26
2 Calculs	26
3 Interprétation géométrique	27
V Gradient	28
1 Définition	28
2 Tangentes aux lignes de niveau	29
3 Lignes de plus forte pente	31
4 Dérivée directionnelle	31
5 Surface de niveau	33
6 Calcul approché	34
7 Minimum et maximum	35
VI Descente de gradient classique	40
1 Où est le minimum ?	40
2 Exemple en deux variables	42
3 Exemples en une variable	43
4 Algorithme du gradient	47
VII Optimisation	48
1 Faire varier le pas	48
2 Régression linéaire $y = ax + b$	49
VIII Descente de gradient stochastique	52
1 Petits pas à petits pas	53
2 Différentes fonctions d'erreurs	57
3 Descente par lots	57
IX Accélérations	61

1	Moment	61
2	Nesterov	62
3	Vocabulaire	62
2	Réseaux de neurones	63
I	Perceptron	63
1	Perceptron linéaire	63
A	Compétences attendues à l'issue de ce cours	69

Fonctions de plusieurs variables et optimisation

De nombreux phénomènes dépendent de plusieurs paramètres, par exemple le volume d'un gaz dépend de la température et de la pression ; l'altitude z d'un terrain dépend des coordonnées (x, y) du lieu.

Dans les réseaux de neurones, les fonctions de plusieurs variables interviennent de deux manières :

- lors de l'utilisation d'un réseau. C'est la partie la plus facile et la plus fréquente. On utilise un réseau déjà bien paramétré pour répondre à une question (Est-ce une photo de chat ? Tourner à droite ou à gauche ? Quel pion déplacer au prochain coup ?). La réponse est un calcul direct obtenu en évaluant une fonction de plusieurs variables.
- lors de la paramétrisation du réseau. C'est la partie difficile et le but de ce cours. Quels paramètres choisir pour définir ce réseau afin qu'il réponde au problème ? Ces paramètres seront choisis comme minimum d'une fonction de plusieurs variables. Une des difficultés est qu'il peut y avoir des milliers de paramètres à gérer.

I- Définition et exemples

1. Définition

Nous allons étudier les fonctions de deux variables, mais aussi de trois variables et plus généralement de n variables. Ces fonctions sont donc de la forme

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

où n est un entier naturel supérieur ou égal à 1.

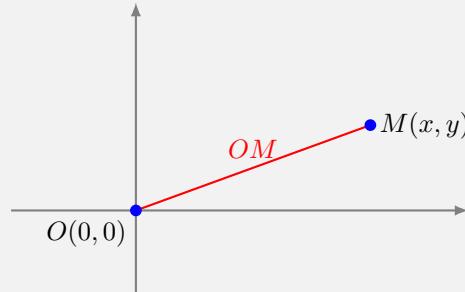
Un élément de l'ensemble de départ est un vecteur de type $x = (x_1, \dots, x_n)$. À chacun de ces vecteurs, f associe un nombre réel $f(x_1, \dots, x_n)$. On pourrait aussi limiter l'ensemble de départ à une partie E de \mathbb{R}^n .

2. Deux variables

Lorsque $n = 2$, on préfère noter les variables (x, y) plutôt que (x_1, x_2) . Voici quelques exemples.

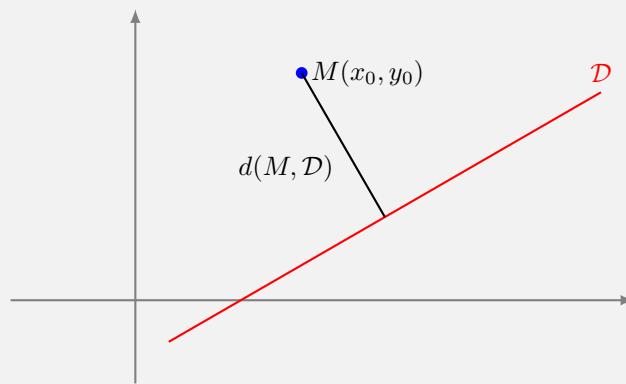
Exemple 2.1

- $f(x, y) = 2x + 3y^2 + 1$.
- $f(x, y) = \cos(xy)$.
- $f(x, y) = \sqrt{x^2 + y^2}$. f renvoie la distance entre un point $M(x, y)$ et l'origine $O(0, 0)$.



- L'équation physique $PV = nRT$ implique $T = \frac{1}{nR}PV$: la température d'un gaz s'exprime en fonction de son volume et de la pression (n et R sont des constantes).
- La distance entre une droite fixée \mathcal{D} d'équation $ax + by + c = 0$ et un point $M(x_0, y_0)$ est donnée par une fonction de deux variables :

$$d(x_0, y_0) = d(M, \mathcal{D}) = \frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}.$$

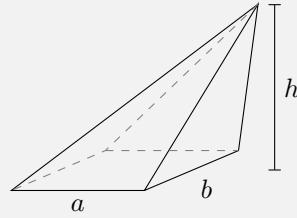


3. Trois variables

Exemple 3.1

1. $f(x, y, z) = ax + by + cz + d$. Cette fonction f est une fonction affine (a, b, c, d sont des constantes).
2. $f(x, y, z) = x^2 + y^2 + z^2$ qui donne la distance au carré entre un point $M(x, y, z)$ et l'origine $O(0, 0, 0)$.
3. Le volume d'un cône à base rectangulaire dépend des longueurs des côtés a et b de la base et de la hauteur h :

$$V = f(a, b, h) = \frac{1}{3}abh.$$



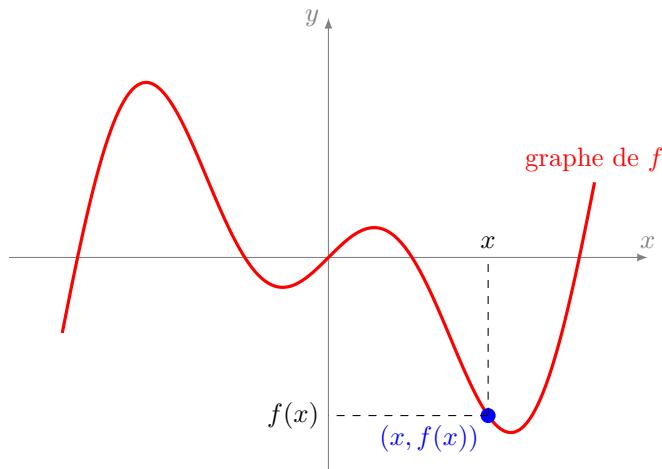
4. n variables

Exemple 4.1

1. $f(x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n + a_0$ une fonction affine (les a_i sont des constantes).
2. $f(x_1, \dots, x_n) = \sqrt{(x_1 - a_1)^2 + \dots + (x_n - a_n)^2}$ exprime la distance entre les points $M(x_1, \dots, x_n)$ et $A(a_1, \dots, a_n)$ dans \mathbb{R}^n .

II- Graphe

Le cas le plus simple, et déjà connu, est celui des fonctions d'une seule variable $f : \mathbb{R} \rightarrow \mathbb{R}$. C'est l'ensemble de tous les points du plan de la forme $(x, f(x))$. Voici le graphe de la fonction $x \mapsto x \cos x$.



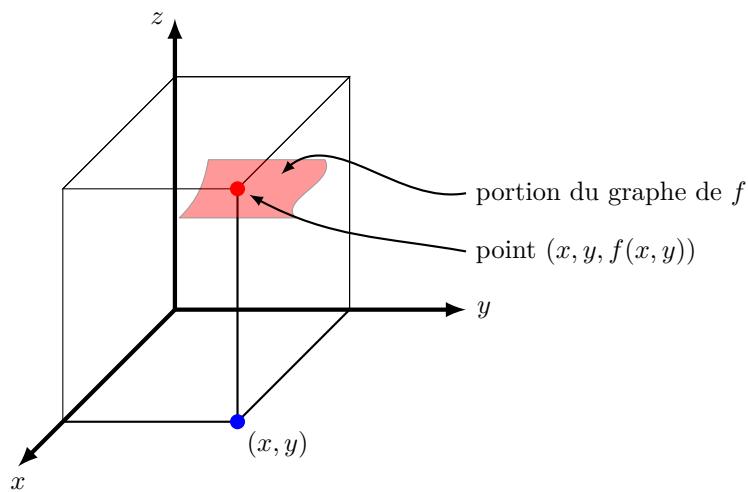
1. Définition

Définition II.1

Le **graphe** \mathcal{G}_f d'une fonction de deux variables $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ est l'ensemble des points de \mathbb{R}^3 ayant pour coordonnées $(x, y, f(x, y))$, pour (x, y) parcourant \mathbb{R}^2 . Le graphe est donc :

$$\mathcal{G}_f = \{(x, y, z) \in \mathbb{R}^3 \mid (x, y) \in \mathbb{R}^2 \text{ et } z = f(x, y)\}.$$

Dans le cas de deux variables, le graphe d'une fonction est une surface tracée dans l'espace.



Exemple 1.1

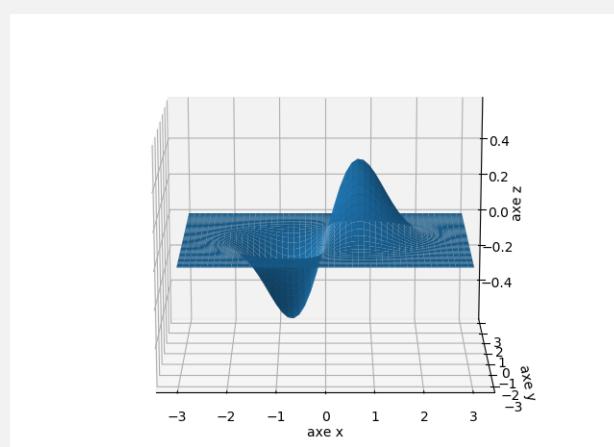
On souhaite tracer le graphe de la fonction définie par :

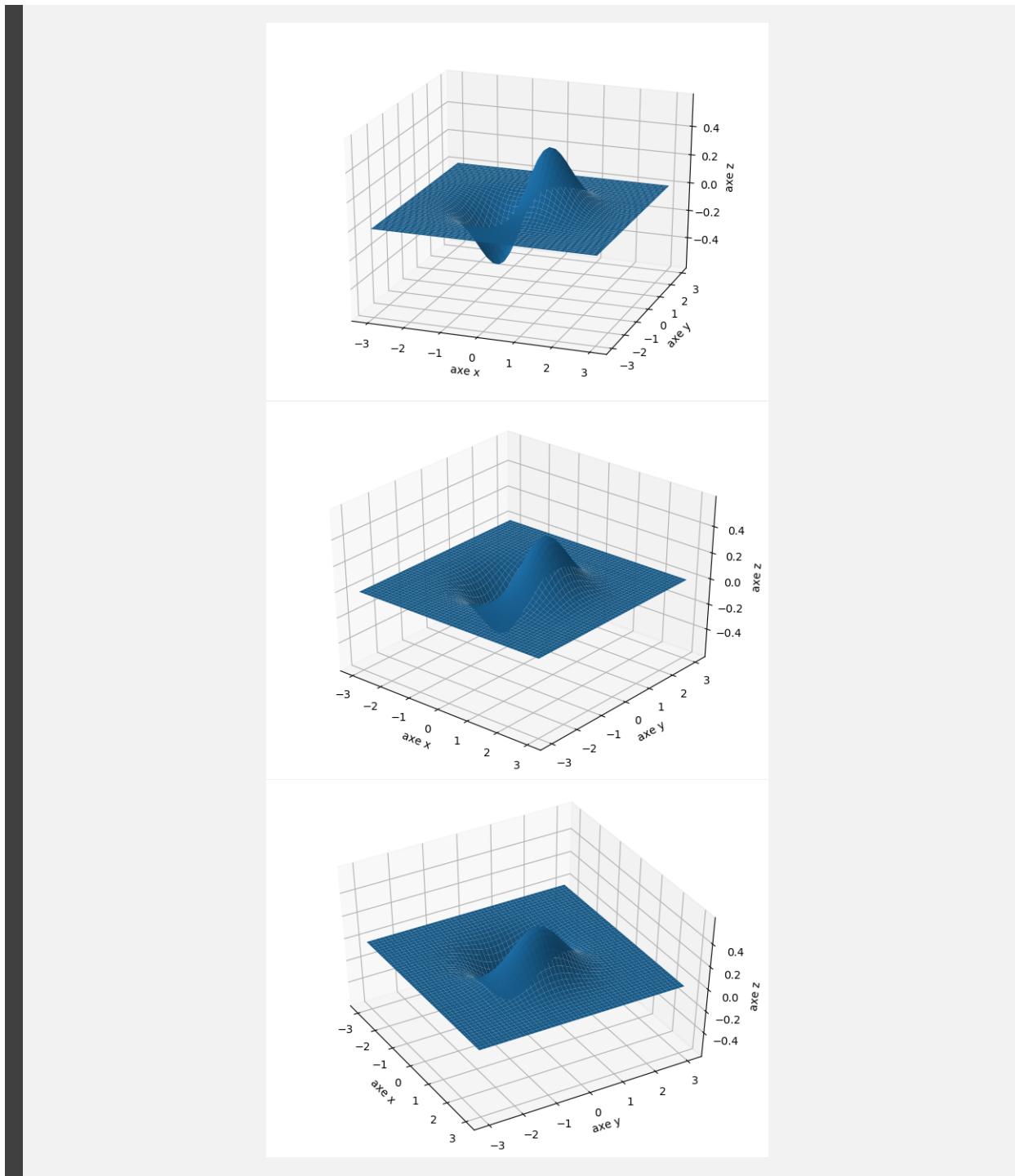
$$f(x, y) = xe^{-x^2-y^2}.$$

On commence par tracer quelques points à la main :

- si $(x, y) = (0, 0)$ alors $f(x, y) = f(0, 0) = 0$ donc le point de coordonnées $(0, 0, 0)$ appartient au graphe.
- Comme $f(1, 0) = 1/e$ alors le point de coordonnées $(1, 0, 1/e)$ appartient au graphe.
- Pour n'importe quel y , on a $f(0, y) = 0$ donc la droite de l'espace d'équation $(x = 0$ et $z = 0)$ est incluse dans le graphe.
- Notons $r = \sqrt{x^2 + y^2}$ la distance entre le point de coordonnées (x, y) et l'origine $(0, 0)$ alors on a la formule $f(x, y) = xe^{-r^2}$. Pour un point éloigné de l'origine, r est grand, donc e^{-r^2} est très petit, et $f(x, y)$ est très proche de 0.

Voici différentes vues de ce graphe.





2. Tranches

Afin de tracer le graphe d'une fonction de deux variables, on peut découper la surface en « tranches ». On fixe par exemple une valeur y_0 et on trace dans le plan (xOz) le graphe de la fonction d'une variable

$$f|_{y_0} : x \mapsto f(x, y_0).$$

Géométriquement, cela revient à tracer l'intersection du graphe de f et du plan d'équation ($y = y_0$). On recommence pour plusieurs valeurs de y_0 , ce qui nous donne des tranches du graphe de f et nous donne une bonne idée du graphe complet de f .

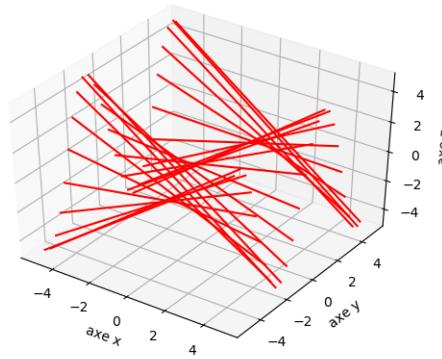
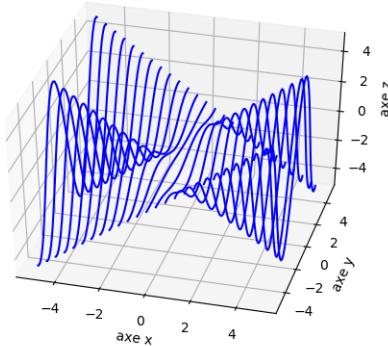
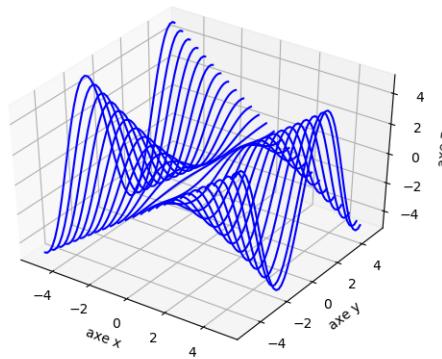
On peut faire le même travail en fixant des valeurs x_0 avec les fonctions :

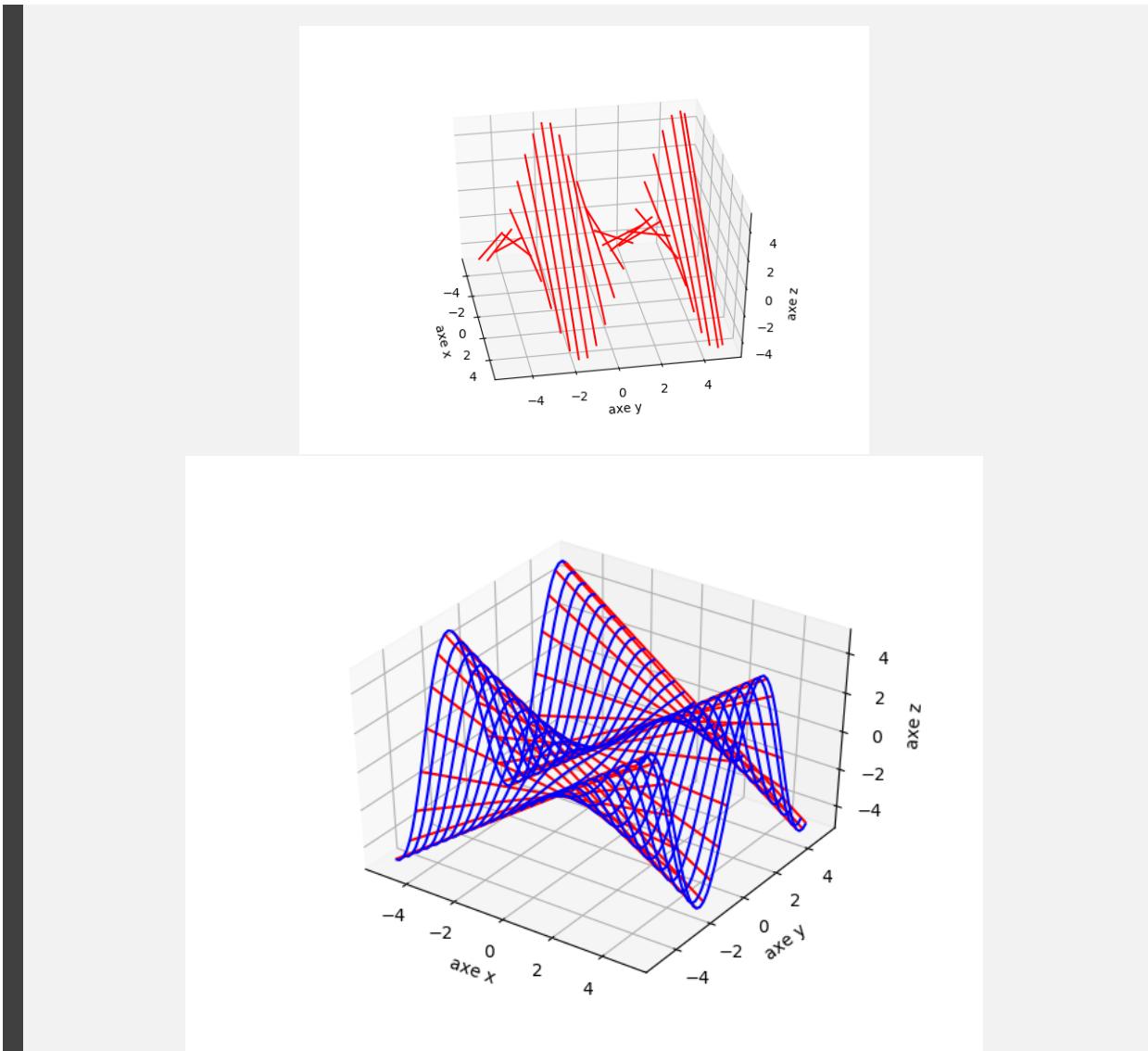
$$f|_{x_0} : y \mapsto f(x_0, y).$$

Exemple 2.1

On souhaite tracer le graphe de la fonction définie par :

$$f(x, y) = x \sin(y).$$





En haut les tranches pour lesquelles x est constant (deux points de vue), au milieu les tranches pour lesquelles y est constant (deux points de vue). Lorsque l'on rassemble les tranches (à x constant et à y constant), on reconstitue la surface (dernière figure).

3. Minimum, maximum

Pour des fonctions de deux variables (ou plus) il existe une notion de minimum et de maximum.

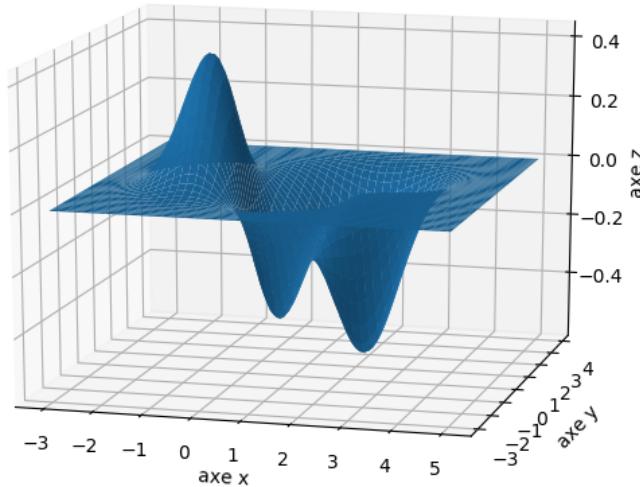
Définition II.2

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ une fonction.

- f atteint un **minimum global** en $(x_0, y_0) \in \mathbb{R}^2$ si pour tout $(x, y) \in \mathbb{R}^2$, on a $f(x, y) \geq f(x_0, y_0)$.
- f atteint un **minimum local** en $(x_0, y_0) \in \mathbb{R}^2$ si il existe un intervalle ouvert I contenant x_0 et un intervalle ouvert J contenant y_0 tels que pour tout $(x, y) \in I \times J$, on a $f(x, y) \geq f(x_0, y_0)$.

Exemple 3.1

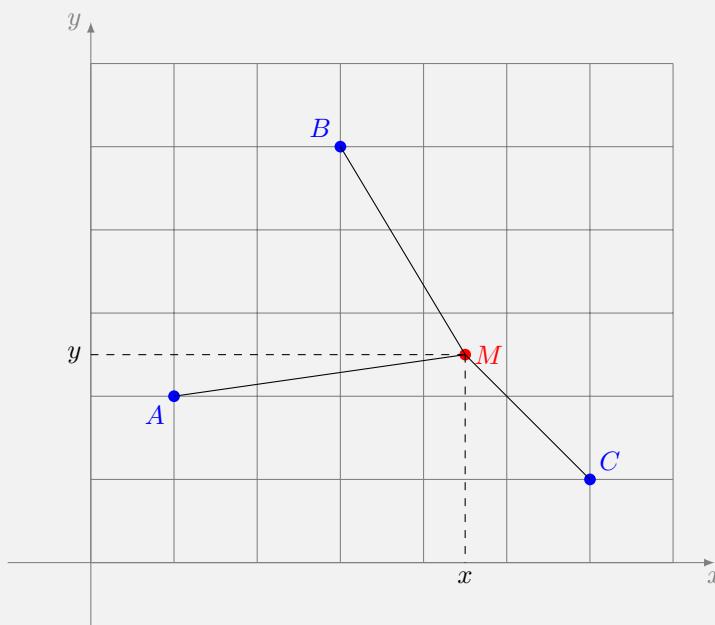
Voici l'exemple d'une fonction qui admet deux minimums locaux. L'un est aussi un minimum global. Elle admet un maximum local qui est aussi global.



Trouver les bons paramètres d'un réseau de neurones nous amènera à trouver le minimum d'une fonction de plusieurs (et même de centaines de) variables. Voyons un exemple en deux variables.

Exemple 3.2

Étant donnés trois points $A(1, 2)$, $B(3, 5)$ et $C(6, 1)$, il s'agit de trouver un point $M(x, y)$ qui « approche au mieux » ces trois points. Il faut expliciter une fonction à minimiser pour définir correctement le problème. Nous décidons de prendre la somme des carrés des distances.



Il s'agit donc de minimiser la fonction f suivante, qui correspond à une fonction distance

(aussi appelée fonction erreur ou bien fonction coût) :

$$f(x, y) = MA^2 + MB^2 + MC^2 = (x - 1)^2 + (y - 2)^2 + (x - 3)^2 + (y - 5)^2 + (x - 6)^2 + (y - 1)^2.$$

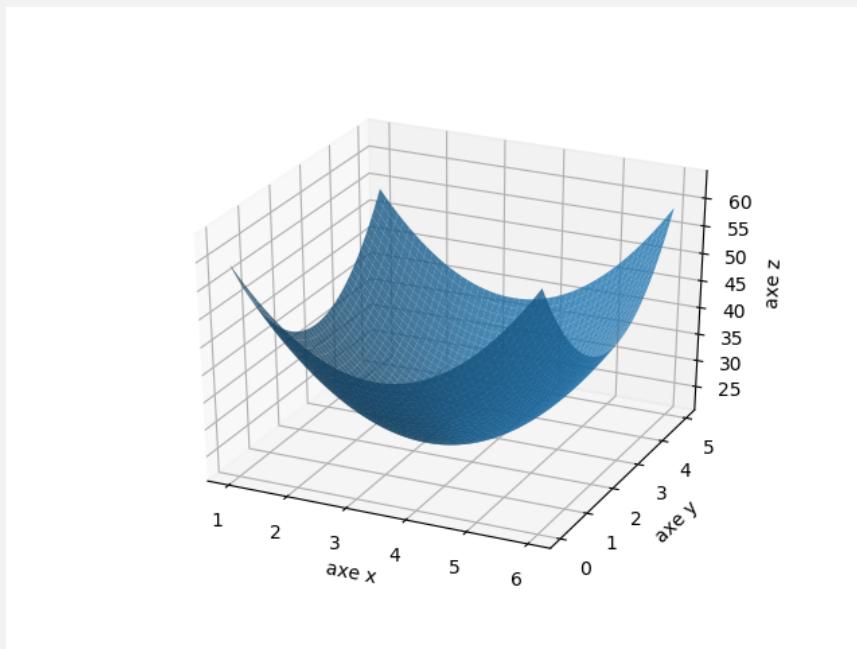
En développant on trouve :

$$f(x, y) = 3x^2 + 3y^2 - 20x - 16y + 76.$$

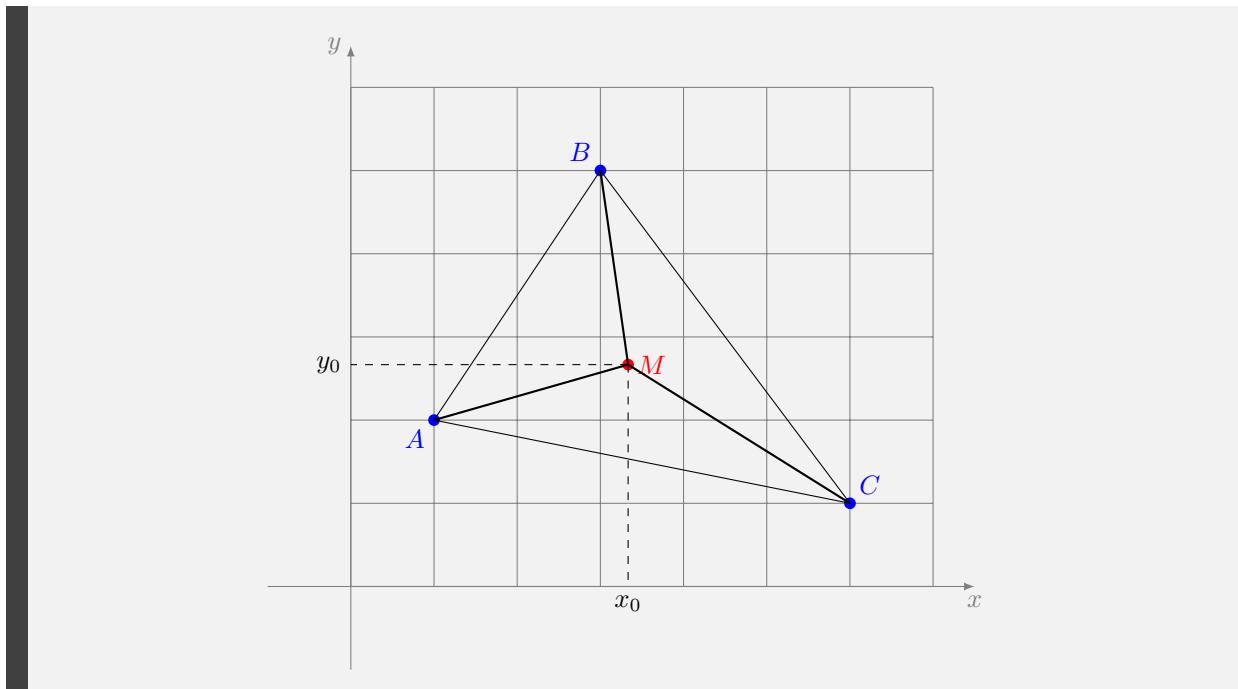
Le graphe de f nous suggère qu'il existe un unique minimum qui est le minimum global de f . Par recherche graphique ou par les méthodes décrites dans la section suivante, on trouverait une solution approchée. En fait la solution géométrique exacte est l'isobarycentre des points (autrement dit le centre de gravité du triangle ABC), ainsi :

$$(x_0, y_0) = \left(\frac{10}{3}, \frac{8}{3} \right) \simeq (3.33, 2.66)$$

pour lequel f atteint son minimum $z_0 = f(x_0, y_0) = \frac{64}{3} \simeq 21.33$.



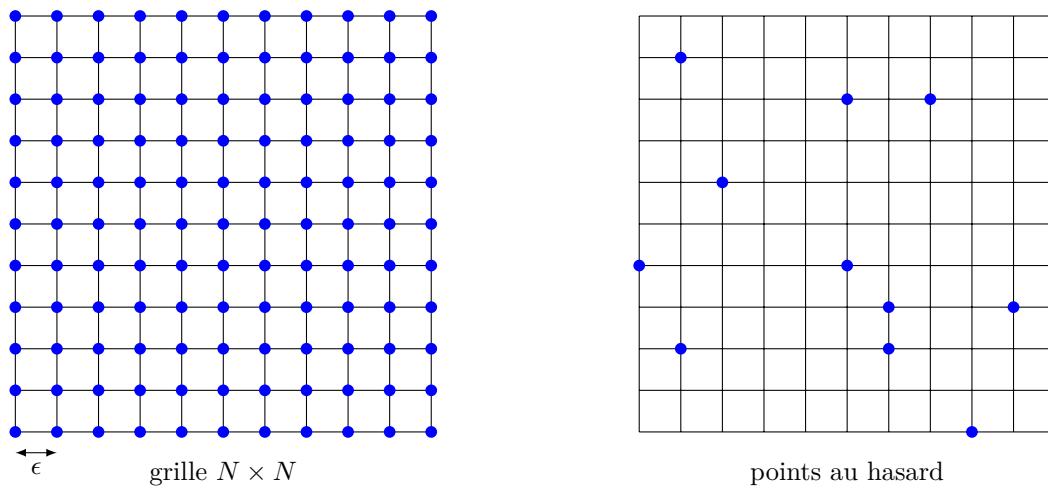
Le point qui convient le mieux à notre problème et en lequel notre fonction distance est minimale est donc le point de coordonnées $(\frac{10}{3}, \frac{8}{3})$. Attention, un autre choix de la fonction distance f pourrait conduire à une autre solution (voir l'exemple de la prochaine section).



4. Recherche élémentaire d'un minimum

Voici trois techniques pour trouver les valeurs approchées des coordonnées du point en lequel une fonction de plusieurs variables atteint son minimum. Ces techniques sont valables quelque soit le nombre de variables même si ici elles ne sont décrites que dans le cas de deux variables.

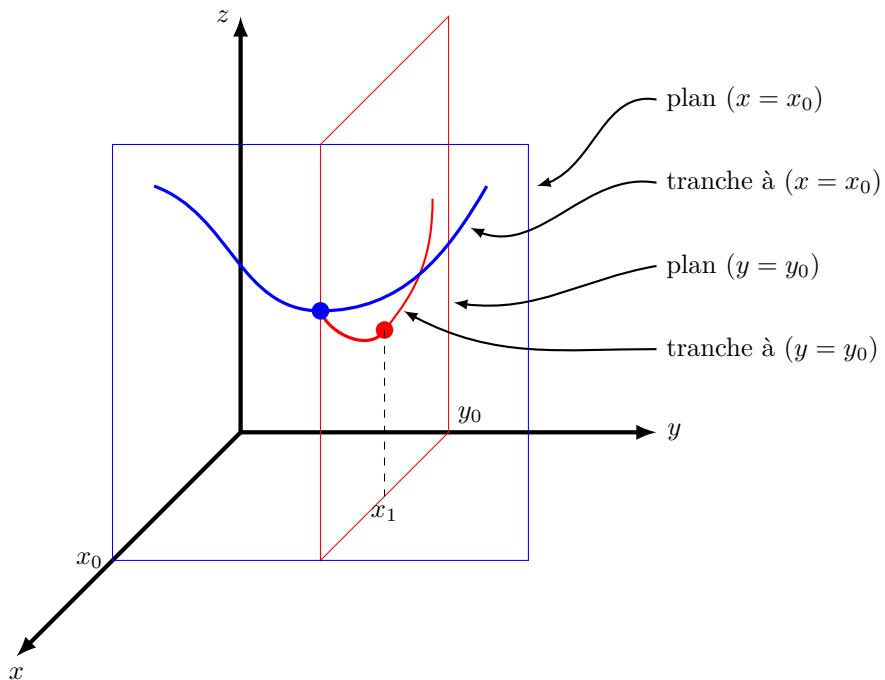
Recherche sur une grille. On calcule $f(x, y)$ pour (x, y) parcourant une grille. On retient le point (x_0, y_0) en lequel $z_0 = f(x_0, y_0)$ est le plus petit. Si on le souhaite, on peut affiner la grille autour de ce point, en diminuant le pas ϵ pour améliorer l'approximation. C'est une technique qui demande N^2 calculs pour une grille de largeur N (et même N^n pour une fonction de n variables) ce qui peut être énorme.



Recherche au hasard. Cela peut sembler incongru mais choisir quelques coordonnées (x, y) au hasard, calculer chaque valeur $z = f(x, y)$ et comme auparavant retenir le point (x_0, y_0) correspondant au z_0 minimal n'est pas ridicule! Un ordinateur peut tester plusieurs millions de points en quelques secondes. Bien sûr il y a de fortes chances de ne trouver qu'une solution approchée. C'est aussi une technique que l'on retrouvera plus tard : partir d'un point au hasard pour ensuite construire une suite de points convergeant vers un minimum. Et si cela n'est pas concluant, il faudra repartir d'un autre point tiré au hasard.

Recherche par tranche. L'idée est de se ramener à des fonctions d'une seule variable. En effet, pour les fonctions d'une variable, on sait qu'il faut chercher les minimums là où la dérivée s'annule.

On part d'une valeur x_0 (au hasard!). On cherche le minimum sur la tranche $x = x_0$, c'est-à-dire que l'on cherche le minimum de la fonction d'une variable $y \mapsto f(x_0, y)$. On trouve une valeur y_0 qui réalise un minimum. On change alors de direction en étudiant maintenant la tranche $y = y_0$ (pour le y_0 que l'on vient d'obtenir), on obtient l'abscisse x_1 du minimum de la fonction d'une variable $x \mapsto f(x, y_0)$. On recommence depuis le début à partir de ce x_1 . On obtient ainsi une suite de points (x_i, y_i) avec des valeurs $z_i = f(x_i, y_i)$ de plus en plus petites. On peut espérer tendre vers un minimum.



Sur la figure ci-dessus, on part d'une tranche choisie au hasard, donnée par $x = x_0$. Cette tranche définit le graphe d'une fonction d'une variable. On se déplace sur cette courbe jusqu'à atteindre le minimum de cette tranche, en une valeur y_0 . On considère la tranche perpendiculaire donnée par $y = y_0$. On se déplace sur la courbe jusqu'à atteindre le minimum de cette tranche, en une valeur x_1 . On pourrait continuer avec une nouvelle tranche $x = x_1$, etc.

Les descriptions données ici sont assez informelles, d'une part parce qu'il est difficile d'énoncer des théorèmes qui garantissent d'atteindre un minimum local et d'autre part parce qu'aucune technique ne garantit d'atteindre un minimum global. Lorsque l'on étudiera le gradient, nous obtiendrons une méthode plus efficace.

Exemple 4.1

On reprend le problème précédent, à savoir trouver un point M qui approche au mieux les trois points A , B et C , mais cette fois on choisit la « vraie » distance comme fonction d'erreur :

$$g(x, y) = MA + MB + MC = \sqrt{(x - 1)^2 + (y - 2)^2} + \sqrt{(x - 3)^2 + (y - 5)^2} + \sqrt{(x - 6)^2 + (y - 1)^2}.$$

On applique ces trois techniques pour chercher le minimum de g en se limitant au carré $[0, 6] \times [0, 6]$.

1. Avec une grille $N \times N$. Par exemple pour $N = 100$, on évalue g en 10 000 points. On trouve $(x_{\min}, y_{\min}) \simeq (2.91, 2.97)$ pour une valeur $z_{\min} \simeq 7.84$.

2. Avec un tirage aléatoire de 1000 points, on trouve par exemple : $(x_{\min}, y_{\min}) \simeq (2.88, 2.99)$ et $z_{\min} \simeq 7.84$. Le résultat est similaire à la méthode précédente, bien qu'on ait effectué 10 fois moins de calculs.

3. Par les tranches. On part de la tranche ($x = 0$). On pose $x_0 = 0$, la fonction à étudier est donc

$$g|_{x_0}(y) = \sqrt{1 + (y - 2)^2} + \sqrt{9 + (y - 5)^2} + \sqrt{36 + (y - 1)^2}.$$

C'est une fonction de la seule variable y pour laquelle on possède des techniques efficaces de recherche de minimum. On trouve que le minimum de $g|_{x_0}$ est atteint en $y_0 \simeq 2.45$. On recommence avec cette fois la tranche ($y = y_0$) et on cherche le minimum de la fonction $g|_{y_0}(x) = g(x, y_0)$. On trouve que cette fonction atteint son minimum en $x_1 \simeq 2.84$. On recommence ce processus jusqu'à atteindre la précision souhaitée. Ainsi en 5 étapes on obtient une valeur approchée assez précise du minimum $(x_{\min}, y_{\min}) \simeq (2.90579, 2.98464)$ et $z_{\min} \simeq 7.83867$.

La première conclusion à tirer de ce qui précède est que pour résoudre un problème il faut définir correctement une fonction d'erreur, c'est-à-dire celle que l'on cherche à minimiser. La solution trouvée dépend de cette fonction d'erreur choisie. Enfin, la méthode des tranches est une méthode efficace pour trouver un minimum d'une fonction, mais nous en découvrirons une encore meilleure en utilisant le gradient.

III- Lignes de niveau

1. Définition

Définition III.1

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ une fonction de deux variables. La **ligne de niveau** $z = c \in \mathbb{R}$ est l'ensemble de tous les points (x, y) vérifiant $f(x, y) = c$:

$$L_c = \{(x, y) \in \mathbb{R}^2 \mid f(x, y) = c\}.$$

La ligne de niveau c est une courbe du plan \mathbb{R}^2 .

On peut aussi définir une **courbe de niveau**, c'est l'ensemble des points de l'espace obtenus comme intersection du graphe \mathcal{G}_f et du plan $z = c$ qui est horizontal et « d'altitude » c . Ce sont donc tous les points $(x, y, f(x, y))$ avec $f(x, y) = c$. On obtient la courbe de niveau en translatant la ligne de niveau d'une altitude c .

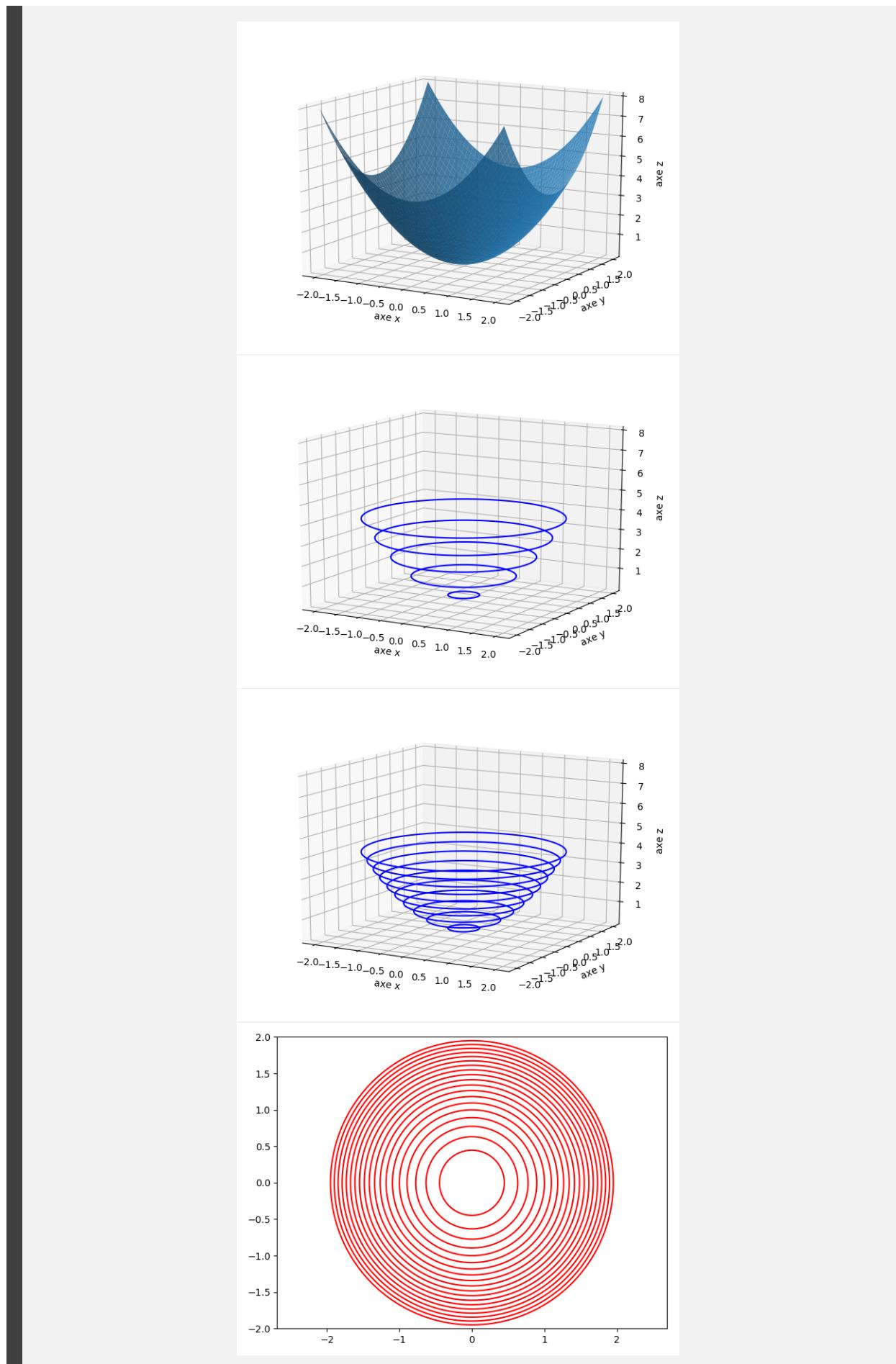
Exemple 1.1

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x, y) = x^2 + y^2$.

- Si $c < 0$, la ligne de niveau L_c est vide (aucun point n'est d'altitude négative).
- Si $c = 0$, la ligne de niveau L_0 se réduit à $\{(0, 0)\}$.
- Si $c > 0$, la ligne de niveau L_c est le cercle du plan de centre $(0, 0)$ et de rayon \sqrt{c} . On « remonte » L_c à l'altitude $z = c$: la courbe de niveau est alors le cercle horizontal de l'espace de centre $(0, 0, c)$ et de rayon \sqrt{c} .

Le graphe est alors une superposition de cercles horizontaux de l'espace de centre $(0, 0, c)$ et de rayon \sqrt{c} , $c > 0$.

Ci-dessous : (a) la surface, (b) 5 courbes de niveau, (c) 10 courbes de niveau, (d) les lignes de niveau dans le plan.



2. Exemples

Exemple 2.1

Sur une carte topographique, les lignes de niveau représentent les courbes ayant la même altitude.

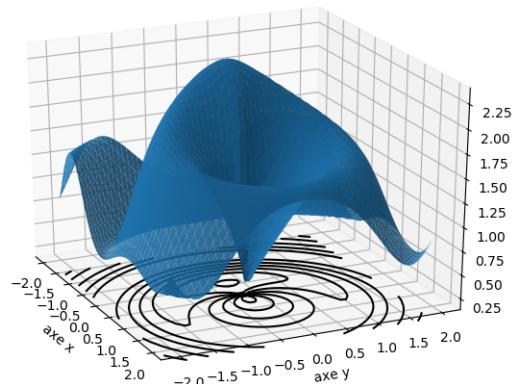


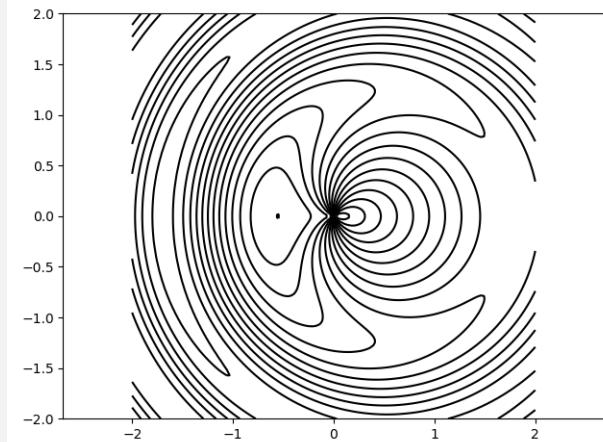
- Ici, une carte *Open Street Map* avec au centre le mont Gerbier de Jonc (source de la Loire, 1551 m).
- Les lignes de niveau correspondent à des altitudes équidistantes de 10 m (par exemple, pour $c = 1400$, $c = 1410$, $c = 1420\ldots$).
- Lorsque les lignes de niveau sont très espacées, le terrain est plutôt plat ; lorsque les lignes sont rapprochées le terrain est pentu.
- Par définition, si on se promène en suivant une ligne de niveau, on reste toujours à la même altitude !

Exemple 2.2

Voici le graphe et les lignes de niveau de la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par

$$f(x, y) = \frac{\sin(r^2 - x)}{r} + 1 \quad \text{où } r = \sqrt{x^2 + y^2}.$$





3. Surfaces quadratiques

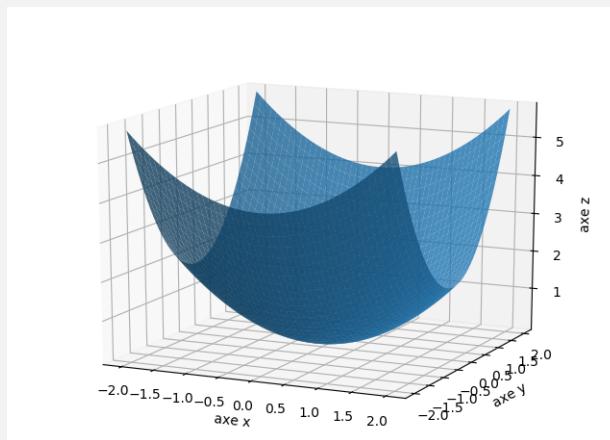
Ce sont des exemples à bien comprendre car ils seront importants pour la suite du cours.

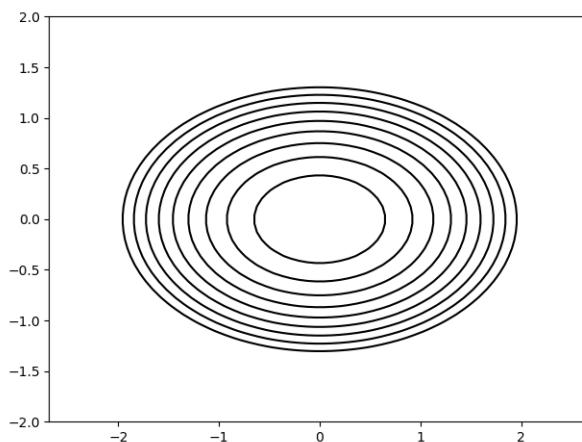
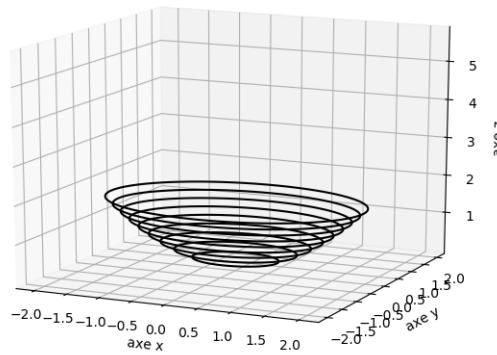
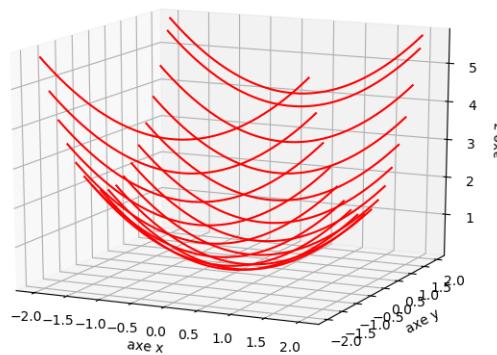
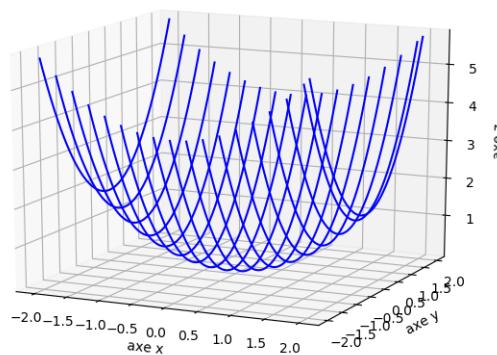
Exemple 3.1

$$f(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1$$

- Les tranches sont des paraboles.
- Les lignes de niveau sont des ellipses.
- Le graphe est donc un **paraboloïde elliptique**.

Ci-dessous : (a) la surface, (b) les tranches avec x constant, (c) les tranches avec y constant, (d) les courbes de niveau, (e) les lignes de niveau dans le plan.



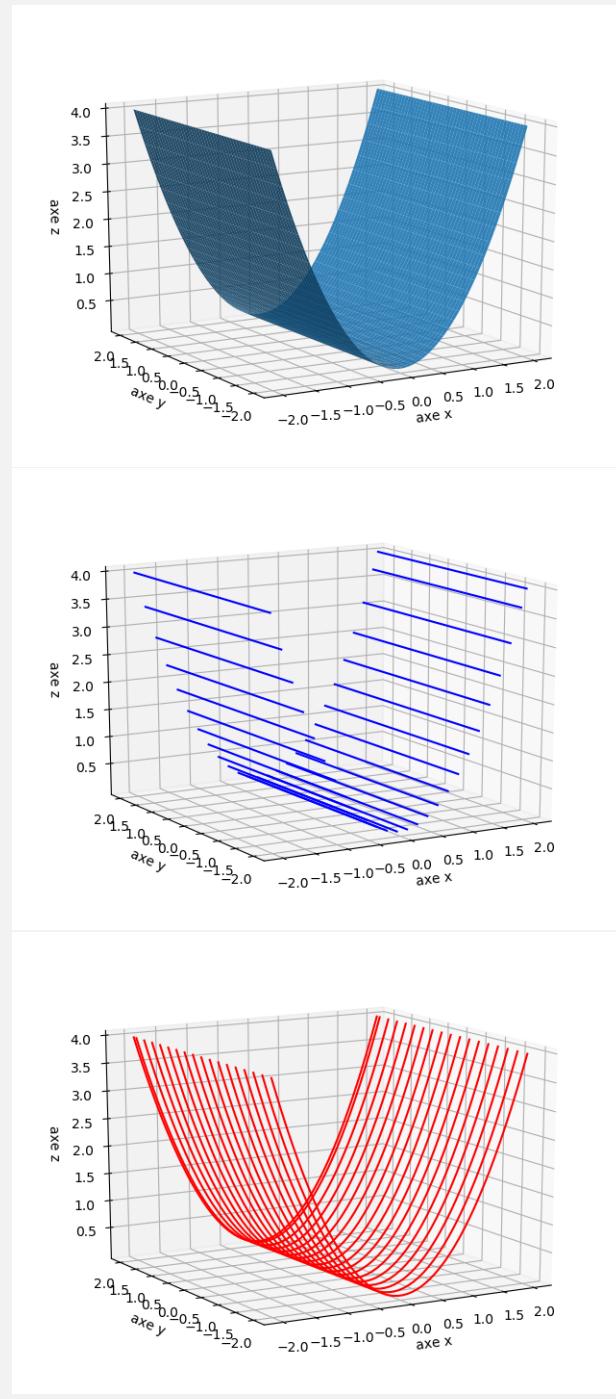


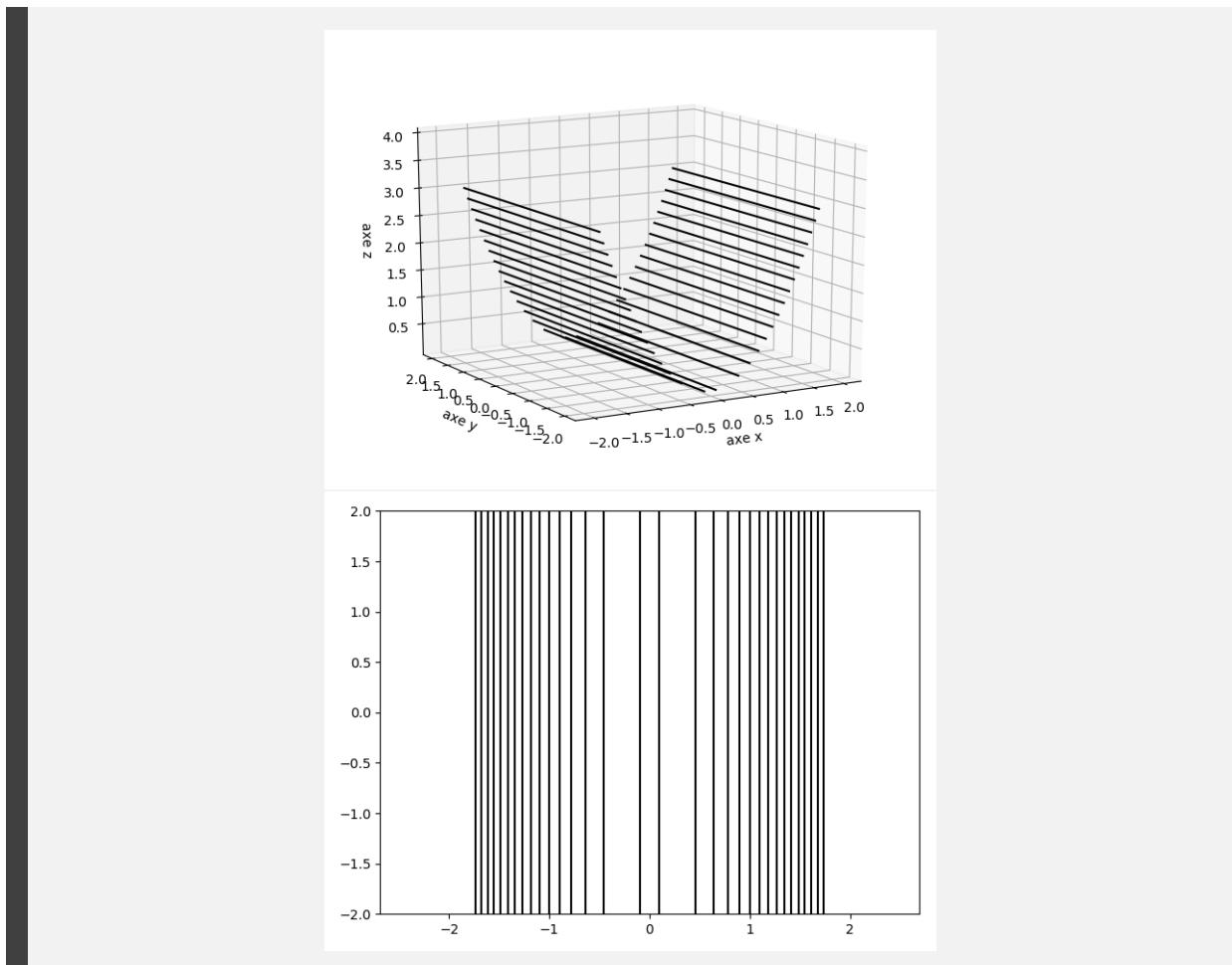
Exemple 3.2

$$f(x, y) = x^2$$

- Les tranches obtenues en coupant selon des plans $y = y_0$ sont des paraboles. Dans l'autre direction, ce sont des droites.
- Les lignes de niveau sont des droites.
- Le graphe est donc un **cylindre parabolique**.

Ci-dessous : (a) la surface, (b) les tranches avec x constant, (c) les tranches avec y constant, (d) les courbes de niveau, (e) les lignes de niveau dans le plan.



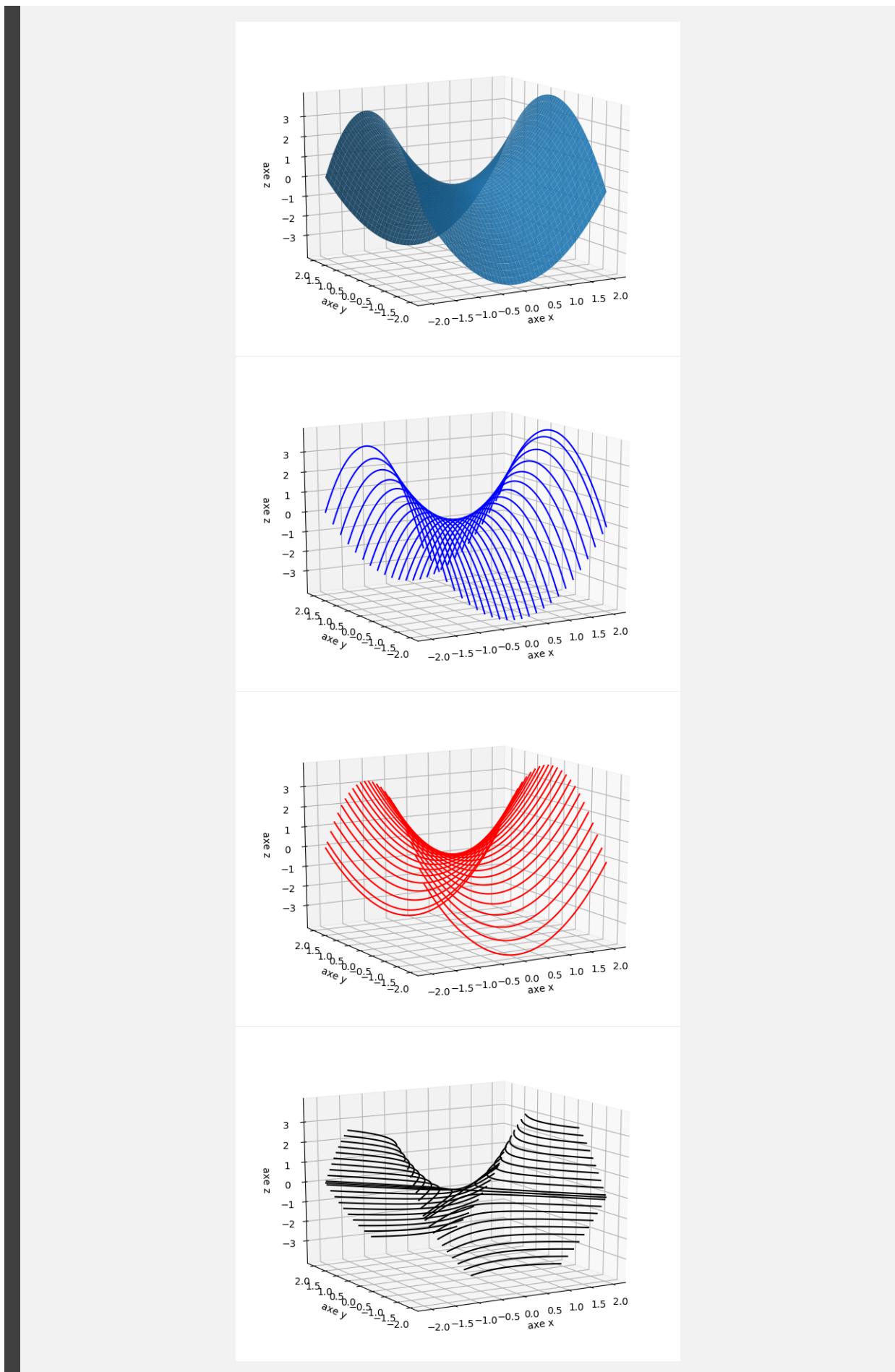


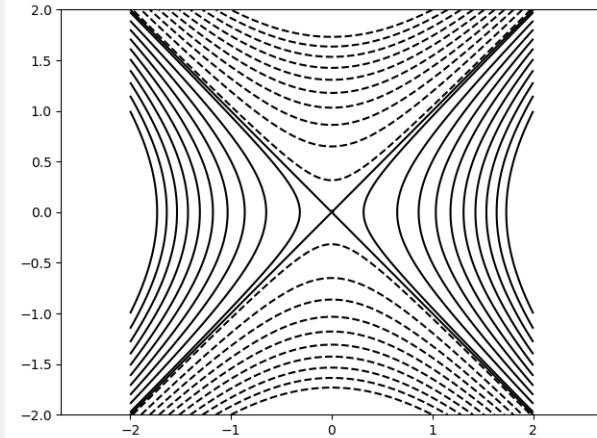
Exemple 3.3

$$f(x, y) = x^2 - y^2$$

- Les tranches sont des paraboles, tournées vers le haut ou vers le bas selon la direction de la tranche.
- Les lignes de niveau sont des hyperboles.
- Le graphe est donc un **paraboloïde hyperbolique** que l'on appelle aussi la **selle de cheval**.
- Un autre nom pour cette surface est un **col** (en référence à un col en montagne). En effet le point $(0, 0, 0)$, est le point de passage le moins haut pour passer d'un versant à l'autre de la montagne.

Ci-dessous : (a) la surface, (b) les tranches avec x constant, (c) les tranches avec y constant, (d) les courbes de niveau, (e) les lignes de niveau dans le plan (en pointillé les lignes de niveau négatif).





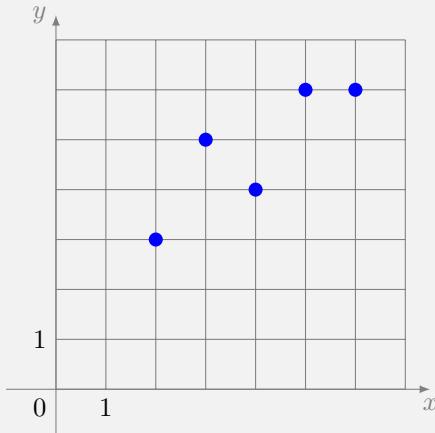
4. Régression linéaire

Exemple 4.1

On se donne des points du plan, comme ci-dessous. On s'aperçoit qu'ils sont à peu près alignés et on souhaite trouver l'équation d'une droite :

$$y = ax + b$$

qui les approche au mieux.



Formalisons un peu le problème : on se donne N points $A_i(x_i, y_i)$, $i = 1, \dots, N$. Pour une droite \mathcal{D} d'équation $y = ax + b$, la distance entre A_i et la droite \mathcal{D} est donnée par la formule :

$$d(A_i, \mathcal{D}) = \frac{|ax_i - y_i + b|}{\sqrt{1 + a^2}}.$$

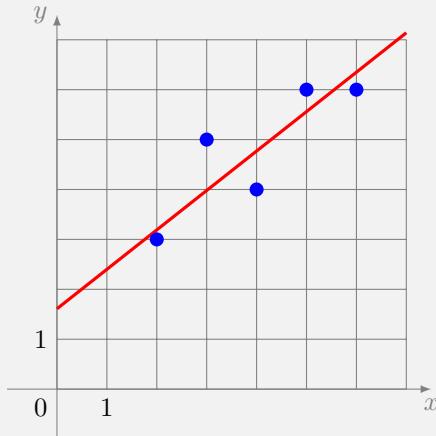
Pour se débarrasser des valeurs absolues et des racines carrées, on élève au carré et on décide que la droite qui approche au mieux tous les points A_i est la droite qui minimise la fonction

$$f(a, b) = \sum_{i=1}^N d(A_i, \mathcal{D})^2 = \frac{1}{1 + a^2} \sum_{i=1}^N (ax_i - y_i + b)^2.$$

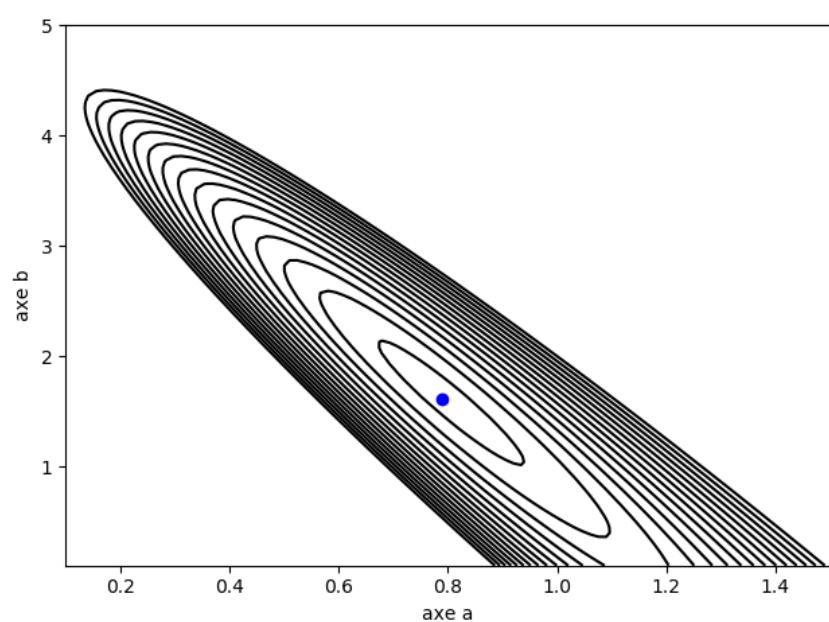
Pour les 5 points du dessin initial : $A_1(2, 3)$, $A_2(3, 5)$, $A_3(4, 4)$, $A_4(5, 6)$ et $A_5(6, 6)$, il s'agit donc de trouver (a, b) qui minimise la fonction

$$f(a, b) = \frac{1}{1 + a^2} ((2a - 3 + b)^2 + (3a - 5 + b)^2 + (4a - 4 + b)^2 + (5a - 6 + b)^2 + (6a - 6 + b)^2).$$

On trace le graphe de f , les lignes de niveau de f , et on utilise les techniques à notre disposition pour trouver qu'un minimum global est réalisé en $(a, b) \simeq (0.8, 1.6)$, ce qui permet de tracer une droite $y = ax + b$ solution.



Lorsque l'on dessine les lignes de niveau, on s'aperçoit que le minimum se trouve dans une région plate et allongée. Cela signifie que, bien que le minimum (a_{\min}, b_{\min}) soit unique, il existe beaucoup de (a, b) tels que $f(a, b)$ soit proche de la valeur minimale $f(a_{\min}, b_{\min})$. De plus, ces points (a, b) peuvent être assez éloignés de la solution (a_{\min}, b_{\min}) (par exemple tous les points de la zone ovale autour du minimum). Ce qui signifie que beaucoup de droites très différentes approchent la solution optimale.



IV- Dérivées partielles

Pour une fonction de plusieurs variables, il existe une dérivée pour chacune des variables, qu'on appelle dérivée partielle.

1. Définition

Définition IV.1

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. La **dérivée partielle** $\frac{\partial f}{\partial x}(x_0, y_0)$ de f par rapport à la variable x au point $(x_0, y_0) \in \mathbb{R}^2$ est la dérivée en x_0 de la fonction d'une variable $x \mapsto f(x, y_0)$.

De même $\frac{\partial f}{\partial y}(x_0, y_0)$ est la dérivée partielle de f par rapport à la variable y au point (x_0, y_0) .

Comme d'habitude et sauf mention contraire, nous supposerons que toutes les dérivées partielles existent.

Autrement dit, en revenant à la définition de la dérivée comme une limite :

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} \quad \text{et} \quad \frac{\partial f}{\partial y}(x_0, y_0) = \lim_{h \rightarrow 0} \frac{f(x_0, y_0 + h) - f(x_0, y_0)}{h}.$$

Plus généralement, pour une fonction $f : \mathbb{R}^n \rightarrow \mathbb{R}$ de plusieurs variables, $\frac{\partial f}{\partial x_i}(x_1, \dots, x_n)$ est la dérivée partielle de f par rapport à la variable x_i au point $(x_1, \dots, x_n) \in \mathbb{R}^n$. C'est la dérivée en x_i de la fonction d'une variable $x_i \mapsto f(x_1, \dots, x_n)$ où l'on considère fixes les variables x_j pour $j \neq i$.

Notations.

$$\frac{\partial f}{\partial x}(x, y) \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y)$$

sont les analogues de l'écriture $\frac{df}{dx}(x)$ pour l'écriture de la dérivée lorsqu'il n'y a qu'une seule variable. Le symbole « ∂ » se lit « d rond ». Une autre notation est $\partial_x f(x, y)$, $\partial_y f(x, y)$ ou bien encore $f'_x(x, y)$, $f'_y(x, y)$.

2. Calculs

Le calcul d'une dérivée partielle n'est pas plus compliqué que le calcul d'une dérivée.

Méthode. Pour calculer une dérivée partielle par rapport à une variable, il suffit de dériver par rapport à cette variable en considérant les autres variables comme des constantes.

Exemple 2.1

Calculer les dérivées partielles de la fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x, y) = x^2 e^{3y}$.

Solution.

Pour calculer la dérivée partielle $\frac{\partial f}{\partial x}$, par rapport à x , on considère que y est une constante et on dérive $x^2 e^{3y}$ comme si c'était une fonction de la variable x uniquement :

$$\frac{\partial f}{\partial x}(x, y) = 2x e^{3y}.$$

Pour l'autre dérivée $\frac{\partial f}{\partial y}$, on considère que x est une constante et on dérive $x^2 e^{3y}$ comme si c'était une fonction de y :

$$\frac{\partial f}{\partial y}(x, y) = 3x^2 e^{3y}.$$

Exemple 2.2

Pour $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ définie par $f(x, y, z) = \cos(x + y^2)e^{-z}$ on a :

$$\begin{aligned}\frac{\partial f}{\partial x}(x, y, z) &= -\sin(x + y^2)e^{-z}, \\ \frac{\partial f}{\partial y}(x, y, z) &= -2y \sin(x + y^2)e^{-z}, \\ \frac{\partial f}{\partial z}(x, y, z) &= -\cos(x + y^2)e^{-z}.\end{aligned}$$

Exemple 2.3

Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ définie par $f(x_1, \dots, x_n) = x_1^2 + x_2^2 + \dots + x_n^2$, alors pour $i = 1, \dots, n$:

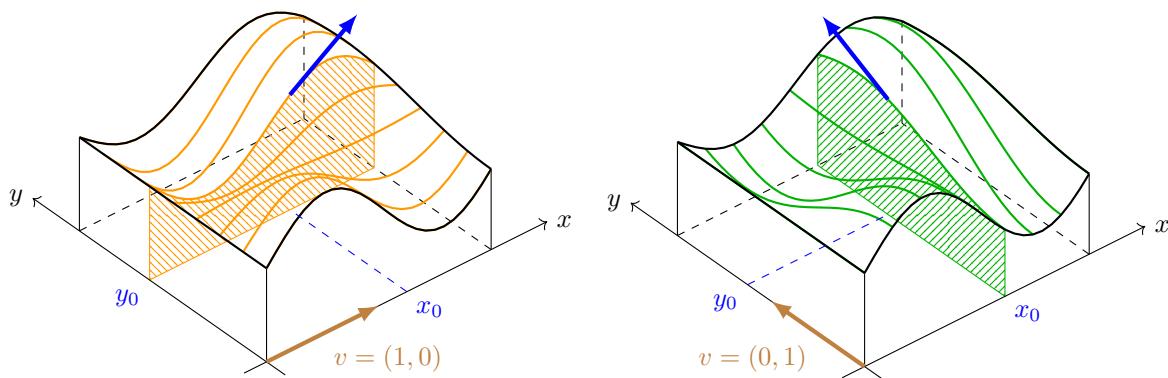
$$\frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = 2x_i.$$

3. Interprétation géométrique

Pour une fonction d'une variable, la dérivée est la pente de la tangente au graphe de la fonction (le graphe étant alors une courbe). Pour une fonction de deux variables $(x, y) \mapsto f(x, y)$, les dérivées partielles indiquent les pentes au graphe de f selon certaines directions (le graphe étant ici une surface). Plus précisément :

- $\frac{\partial f}{\partial x}(x_0, y_0)$ est la pente du graphe de f en (x_0, y_0) suivant la direction de l'axe (Ox). En effet cette pente est celle de la tangente à la courbe $z = f(x, y_0)$ et est donnée par la dérivée de $x \mapsto f(x, y_0)$ en x_0 , c'est donc bien $\frac{\partial f}{\partial x}(x_0, y_0)$.
- $\frac{\partial f}{\partial y}(x_0, y_0)$ est la pente du graphe de f en (x_0, y_0) suivant la direction de l'axe (Oy).

Sur la figure de gauche, la dérivée partielle $\frac{\partial f}{\partial x}$ indique la pente de la tranche parallèle à l'axe (Ox). Sur la figure de droite, la dérivée partielle $\frac{\partial f}{\partial y}$ indique la pente de la tranche parallèle à l'axe (Oy).



V- Gradient

Le gradient est un vecteur dont les coordonnées sont les dérivées partielles. Il a de nombreuses applications géométriques car il donne l'équation des tangences aux courbes et surfaces de niveau. Surtout, il indique la direction dans laquelle la fonction varie le plus vite.

Le gradient est un vecteur qui remplace la notion de dérivée pour les fonctions de plusieurs variables. On sait que la dérivée permet de décider si une fonction est croissante ou décroissante. De même, le vecteur gradient indique la direction dans laquelle la fonction croît ou décroît le plus vite. Nous allons voir comment calculer de façon algorithmique le gradient grâce à la « différentiation automatique ».

1. Définition

Définition V.1

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ une fonction admettant des dérivées partielles. Le **gradient** de f en $(x_0, y_0) \in \mathbb{R}^2$, noté $\overrightarrow{\text{grad}} f(x_0, y_0)$, est le vecteur :

$$\overrightarrow{\text{grad}} f(x_0, y_0) = \begin{pmatrix} \frac{\partial f}{\partial x}(x_0, y_0) \\ \frac{\partial f}{\partial y}(x_0, y_0) \end{pmatrix}.$$

Les physiciens et les anglo-saxons notent souvent $\nabla f(x, y)$ pour $\overrightarrow{\text{grad}} f(x, y)$. Le symbole ∇ se lit « nabla ».

Plus généralement, pour $f : \mathbb{R}^n \rightarrow \mathbb{R}$, le gradient de f en $(x_1, \dots, x_n) \in \mathbb{R}^n$ est le vecteur :

$$\overrightarrow{\text{grad}} f(x_1, \dots, x_n) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_1, \dots, x_n) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x_1, \dots, x_n) \end{pmatrix}.$$

Exemple 1.1

- $f(x, y) = x^2y^3$, $\overrightarrow{\text{grad}} f(x, y) = \begin{pmatrix} 2xy^3 \\ 3x^2y^2 \end{pmatrix}$. Au point $(x_0, y_0) = (2, 1)$, $\overrightarrow{\text{grad}} f(2, 1) = \begin{pmatrix} 4 \\ 12 \end{pmatrix}$.
- $f(x, y, z) = x^2 \sin(yz)$, $\overrightarrow{\text{grad}} f(x, y, z) = \begin{pmatrix} 2x \sin(yz) \\ x^2 z \cos(yz) \\ x^2 y \cos(yz) \end{pmatrix}$.
- $f(x_1, \dots, x_n) = x_1^2 + x_2^2 + \dots + x_n^2$, $\overrightarrow{\text{grad}} f(x_1, \dots, x_n) = \begin{pmatrix} 2x_1 \\ \vdots \\ 2x_n \end{pmatrix}$.

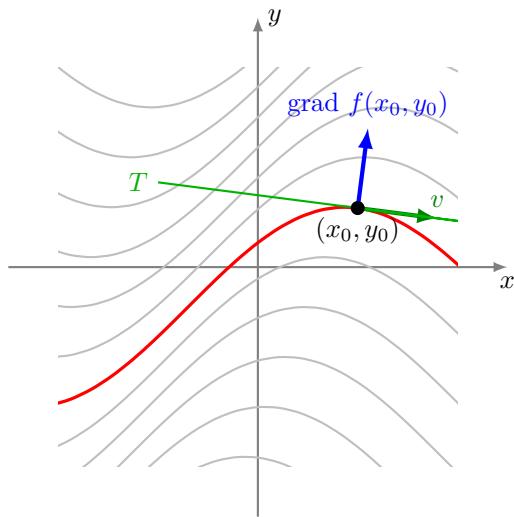
2. Tangentes aux lignes de niveau

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ une fonction différentiable. On considère les lignes de niveau $f(x, y) = k$.

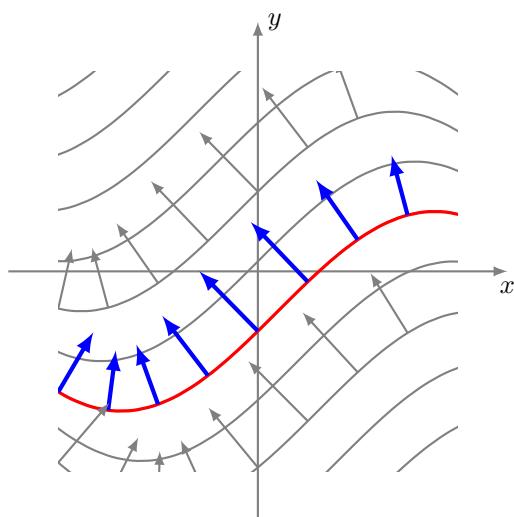
Proposition 1.1

Le vecteur gradient $\overrightarrow{\text{grad}} f(x_0, y_0)$ est orthogonal à la ligne de niveau de f passant au point (x_0, y_0) .

Sur ce premier dessin, sont dessinés la ligne de niveau passant par le point (x_0, y_0) , un vecteur tangent v en ce point et la tangente à la ligne de niveau. Le vecteur gradient est un vecteur du plan qui est orthogonal à la ligne de niveau en ce point.



À chaque point du plan, on peut associer un vecteur gradient. Ce vecteur gradient est orthogonal à la ligne de niveau passant par ce point. Nous verrons juste après comment savoir s'il est orienté « vers le haut » ou « vers le bas ».



Dans le cadre de notre étude, nous nous intéressons à l'équation de la tangente.

Proposition 1.2

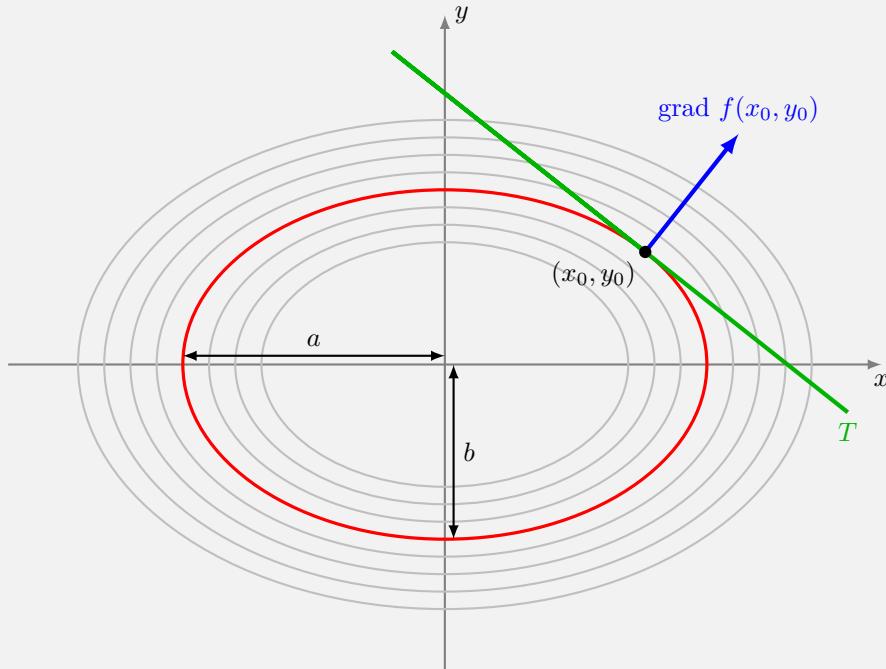
Au point (x_0, y_0) , l'équation de la tangente à la ligne de niveau de f est :

$$\frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) = 0$$

pourvu que le gradient de f en ce point ne soit pas le vecteur nul.

Exemple 2.1 : Tangentes à une ellipse

Trouver les tangentes à l'ellipse \mathcal{E} d'équation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.



Cette ellipse \mathcal{E} est la ligne de niveau $f(x, y) = 1$ de la fonction $f(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}$. Les dérivées partielles en (x_0, y_0) sont :

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{2x_0}{a^2} \quad \text{et} \quad \frac{\partial f}{\partial y}(x_0, y_0) = \frac{2y_0}{b^2}.$$

L'équation de la tangente à l'ellipse \mathcal{E} en ce point est donc :

$$\frac{2x_0}{a^2}(x - x_0) + \frac{2y_0}{b^2}(y - y_0) = 0.$$

Mais comme $\frac{x_0^2}{a^2} + \frac{y_0^2}{b^2} = 1$, l'équation de la tangente se simplifie en $\frac{x_0}{a^2}x + \frac{y_0}{b^2}y = 1$.

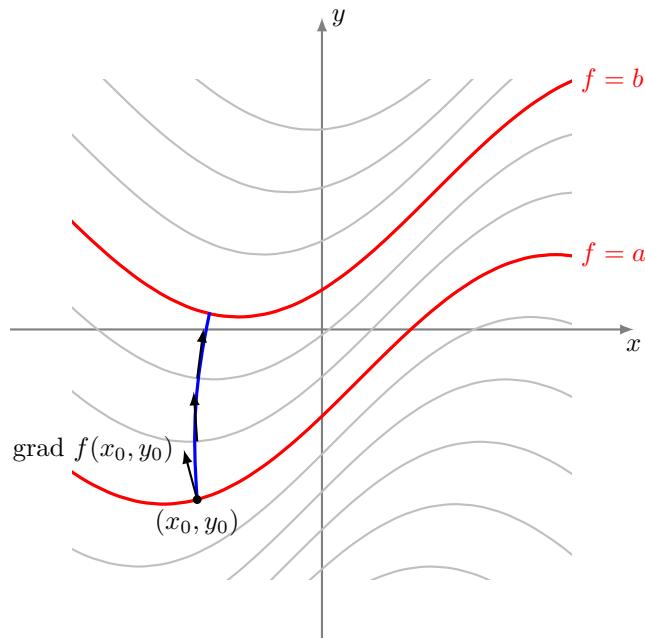
3. Lignes de plus forte pente

Considérons les lignes de niveau $f(x, y) = k$ d'une fonction $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. On se place en un point (x_0, y_0) . On cherche dans quelle direction se déplacer pour augmenter au plus vite la valeur de f .

Proposition 1.3

Le vecteur gradient $\overrightarrow{\text{grad}} f(x_0, y_0)$ indique la direction de plus grande pente à partir du point (x_0, y_0) .

Autrement dit, si l'on veut, à partir d'un point donné (x_0, y_0) de niveau a , passer au niveau $b > a$ le plus vite possible alors il faut démarrer en suivant la direction du gradient $\overrightarrow{\text{grad}} f(x_0, y_0)$.



Comme illustration, un skieur de descente, voulant optimiser sa course, choisira en permanence de s'orienter suivant la plus forte pente, c'est-à-dire dans le sens opposé au gradient.

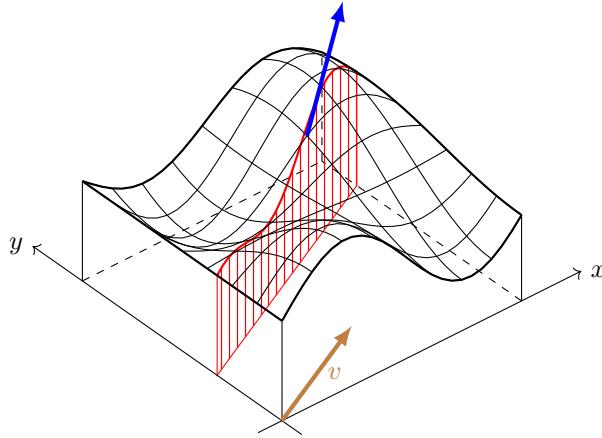
4. Dérivée directionnelle

Pour prouver que le gradient indique la ligne de la plus grande pente, nous avons besoin de généraliser la notion de dérivée partielle. Ce passage est plus technique et peut être ignoré en première lecture.

Soit $v = \begin{pmatrix} h \\ k \end{pmatrix}$ un vecteur du plan. La **dérivée directionnelle** de f suivant le vecteur v en (x_0, y_0) est le nombre :

$$D_v f(x_0, y_0) = h \frac{\partial f}{\partial x}(x_0, y_0) + k \frac{\partial f}{\partial y}(x_0, y_0).$$

La dérivée directionnelle correspond à la pente de la fonction pour la tranche dirigée par le vecteur v .



Remarque : pour $v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ alors $D_v f(x_0, y_0) = \frac{\partial f}{\partial x}(x_0, y_0)$ et pour $v = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ alors $D_v f(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0)$.

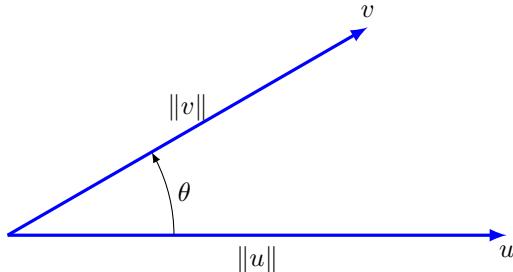
On rappelle que le **produit scalaire** de deux vecteurs $u = \begin{pmatrix} x \\ y \end{pmatrix}$ et $v = \begin{pmatrix} x' \\ y' \end{pmatrix}$ est donné par

$$\langle u | v \rangle = xx' + yy'.$$

On sait que le produit scalaire se calcule aussi géométriquement par :

$$\langle u | v \rangle = \|u\| \cdot \|v\| \cdot \cos(\theta)$$

où θ est l'angle entre u et v .



Ainsi, on peut réécrire la dérivée directionnelle sous la forme :

$$D_v f(x_0, y_0) = \langle \overrightarrow{\text{grad}} f(x_0, y_0) | v \rangle.$$

On peut maintenant prouver que le gradient indique la ligne de plus grande pente.

Démonstration. La dérivée suivant le vecteur non nul v au point (x_0, y_0) décrit la variation de f autour de ce point lorsqu'on se déplace dans la direction v . La direction selon laquelle la croissance est la plus grande est celle du gradient de f . En effet,

$$D_v f(x_0, y_0) = \langle \overrightarrow{\text{grad}} f(x_0, y_0) | v \rangle = \|\overrightarrow{\text{grad}} f(x_0, y_0)\| \cdot \|v\| \cdot \cos \theta$$

où θ est l'angle entre le vecteur $\overrightarrow{\text{grad}} f(x_0, y_0)$ et le vecteur v . Le maximum est atteint lorsque l'angle $\theta = 0$, c'est-à-dire lorsque v pointe dans la même direction que $\overrightarrow{\text{grad}} f(x_0, y_0)$. \square

5. Surface de niveau

Les résultats présentés ci-dessus pour les fonctions de deux variables se généralisent aux fonctions de trois variables ou plus. Commençons avec trois variables et une fonction $f : \mathbb{R}^3 \rightarrow \mathbb{R}$. Rappelons qu'un plan de \mathbb{R}^3 passant par (x_0, y_0, z_0) et de vecteur normal $n = (a, b, c)$ a pour équation cartésienne :

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0.$$

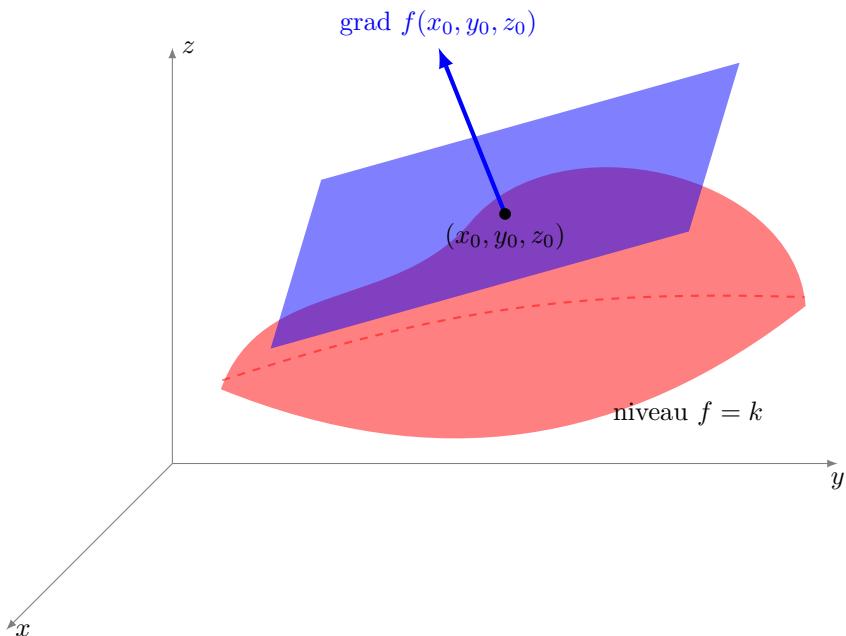
De même qu'il existe une droite tangente pour les lignes de niveau, il existe un **plan tangent** à une surface de niveau.

Proposition 1.4

Le vecteur gradient $\overrightarrow{\text{grad}} f(x_0, y_0, z_0)$ est orthogonal à la surface de niveau de f passant au point (x_0, y_0, z_0) . Autrement dit, l'équation du plan tangent à la surface de niveau de f en (x_0, y_0, z_0) est

$$\frac{\partial f}{\partial x}(x_0, y_0, z_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0, z_0)(y - y_0) + \frac{\partial f}{\partial z}(x_0, y_0, z_0)(z - z_0) = 0$$

pourvu que le gradient de f en ce point ne soit pas le vecteur nul.



Plus généralement pour $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $\overrightarrow{\text{grad}} f(x_1, \dots, x_n)$ est orthogonal à l'espace tangent à l'hypersurface de niveau $f = k$ passant par le point $(x_1, \dots, x_n) \in \mathbb{R}^n$ et ce vecteur gradient $\overrightarrow{\text{grad}} f(x_1, \dots, x_n)$ indique la direction de plus grande pente à partir du point (x_1, \dots, x_n) .

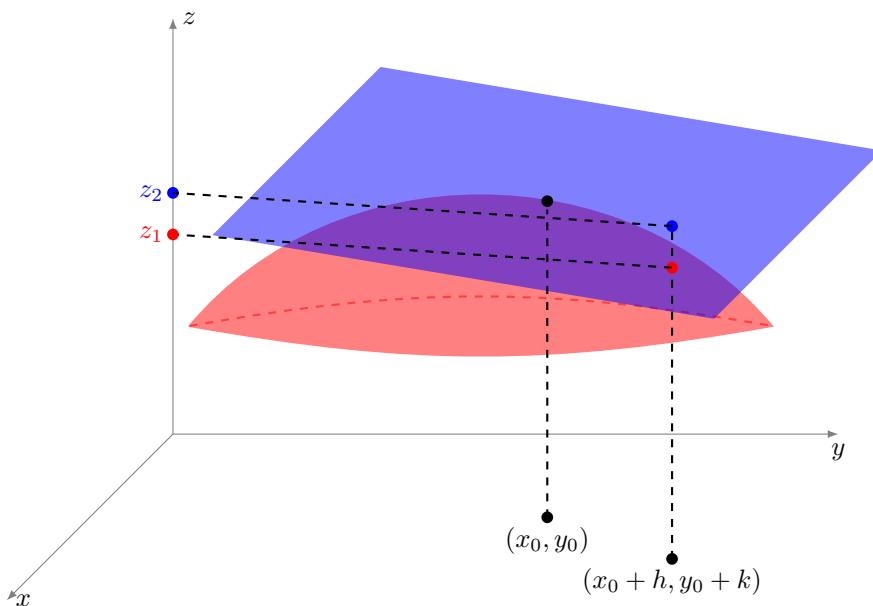
6. Calcul approché

Rappelez-vous que la dérivée nous a permis de faire des calculs approchés, par exemple pour estimer $\sqrt{1.01}$ sans calculatrice (voir le chapitre « Dérivée »). Voici, en deux variables, l'analogue de la formule pour une variable :

$$f(x_0 + h, y_0 + k) \simeq f(x_0, y_0) + h \frac{\partial f}{\partial x}(x_0, y_0) + k \frac{\partial f}{\partial y}(x_0, y_0).$$

Cette approximation est valable pour h et k petits.

L'interprétation géométrique est la suivante : on approche le graphe de f en (x_0, y_0) par le plan tangent au graphe en ce point. Sur la figure ci-dessous sont représentés : le graphe de f , le plan tangent au-dessus du point (x_0, y_0) . La valeur $z_1 = f(x_0 + h, y_0 + k)$ est la valeur exacte donnée par le point de la surface au dessus de $(x_0 + h, y_0 + k)$. On approche cette valeur par $z_2 = f(x_0, y_0) + h \frac{\partial f}{\partial x}(x_0, y_0) + k \frac{\partial f}{\partial y}(x_0, y_0)$ donnée par le point du plan tangent au dessus de $(x_0 + h, y_0 + k)$.



Exemple 6.1

Valeur approchée de $f(1.002, 0.997)$ si $f(x, y) = x^2y$.

Solution. Ici $(x_0, y_0) = (1, 1)$, $h = 2 \times 10^{-3}$, $k = -3 \times 10^{-3}$, $\frac{\partial f}{\partial x}(x, y) = 2xy$, $\frac{\partial f}{\partial y}(x, y) = x^2$, donc $\frac{\partial f}{\partial x}(x_0, y_0) = 2$, $\frac{\partial f}{\partial y}(x_0, y_0) = 1$. Ainsi

$$f(1 + h, 1 + k) \simeq f(1, 1) + 2h + k$$

donc

$$f(1.002, 0.997) \simeq 1 + 2 \times 2 \times 10^{-3} - 3 \times 10^{-3} \simeq 1.001.$$

Avec une calculatrice, on trouve $f(1.002, 0.997) = 1.000992$: l'approximation est bonne.

7. Minimum et maximum

Définition V.2

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$.

- La fonction f admet un **minimum local** en (x_0, y_0) s'il existe un disque D centré en ce point tel que

$$f(x, y) \geq f(x_0, y_0) \quad \text{pour tout } (x, y) \in D.$$

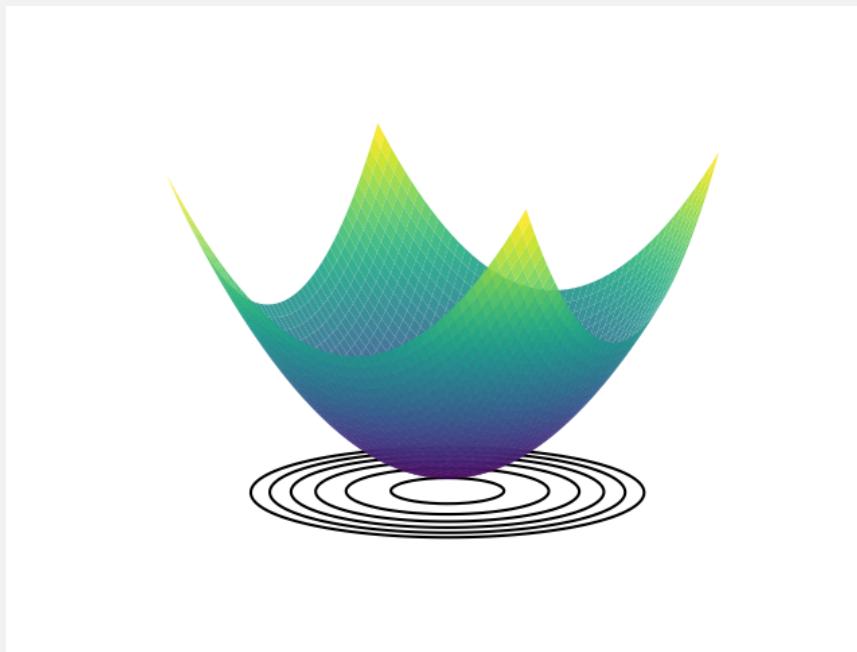
- La fonction f admet un **maximum local** en (x_0, y_0) pour lequel

$$f(x, y) \leq f(x_0, y_0) \quad \text{pour tout } (x, y) \in D.$$

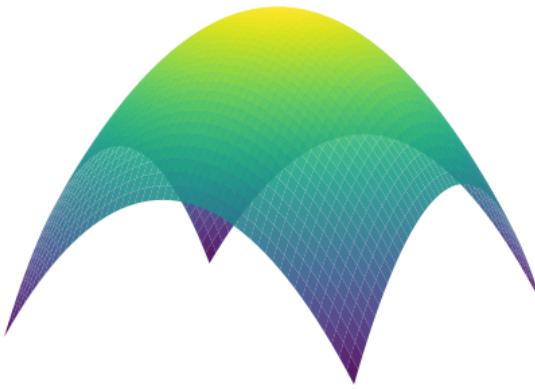
- On parle d'un **extremum local** pour un minimum ou un maximum local.

Exemple 7.1

L'exemple type de minimum est celui de la fonction $f(x, y) = x^2 + y^2$ en $(0, 0)$. Voici son graphe et ses lignes de niveau.



La fonction $f(x, y) = -x^2 - y^2$ admet, elle, un maximum en $(0, 0)$.



Proposition 1.5

Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. Si f admet un minimum ou un maximum local en (x_0, y_0) alors le gradient est le vecteur nul en ce point, autrement dit :

$$\frac{\partial f}{\partial x}(x_0, y_0) = 0 \quad \text{et} \quad \frac{\partial f}{\partial y}(x_0, y_0) = 0.$$

Démonstration. Prenons le cas d'un minimum local. La fonction d'une variable $x \mapsto f(x, y_0)$ admet aussi un minimum en x_0 donc sa dérivée est nulle en x_0 , c'est-à-dire $\frac{\partial f}{\partial x}(x_0, y_0) = 0$. De même $y \mapsto f(x_0, y)$ admet un minimum en y_0 donc $\frac{\partial f}{\partial y}(x_0, y_0) = 0$. \square

Dans la suite du cours nous chercherons les points pour lesquels une fonction donnée présente un minimum local. D'après la proposition précédente, ces points sont à chercher parmi les points en lesquels le gradient s'annule. On dira que (x_0, y_0) est un **point critique** de f si les deux dérivées partielles $\frac{\partial f}{\partial x}(x_0, y_0)$ et $\frac{\partial f}{\partial y}(x_0, y_0)$ s'annulent simultanément.

Exemple 7.2

Chercher les points en lesquels $f(x, y) = x^2 - y^3 + xy$ peut atteindre son minimum.

Recherche des points critiques. On calcule

$$\frac{\partial f}{\partial x}(x, y) = 2x + y \quad \text{et} \quad \frac{\partial f}{\partial y}(x, y) = -3y^2 + x.$$

On cherche les points (x, y) en lesquels les deux dérivées partielles s'annulent. Par l'annulation de la première dérivée, on a $2x + y = 0$ donc $y = -2x$. Par l'annulation de la seconde dérivée, on a $-3y^2 + x = 0$ ce qui donne par substitution $-12x^2 + x = 0$, ainsi $x(-12x + 1) = 0$. Donc soit $x = 0$ et alors on a $y = 0$, soit $x = \frac{1}{12}$ et alors $y = -\frac{1}{6}$. Bilan : il y a deux points critiques :

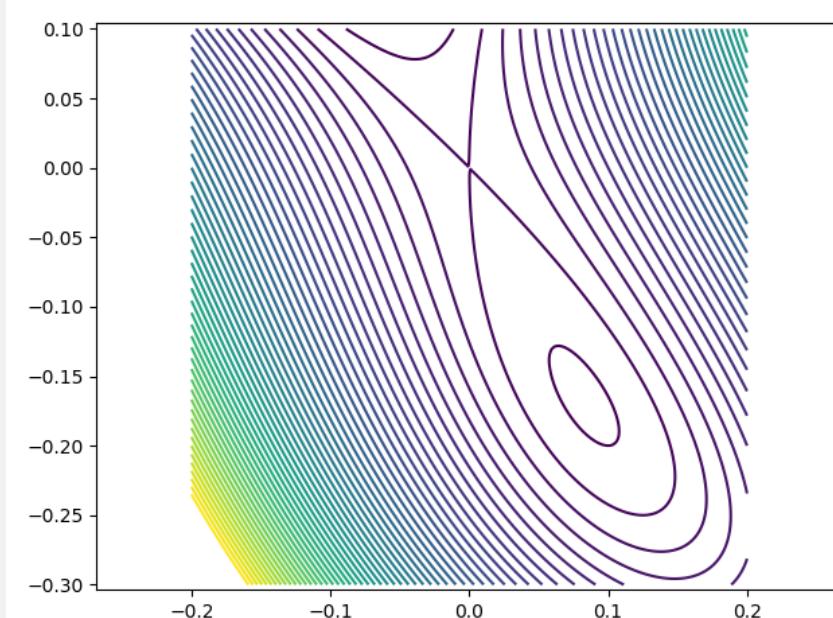
$$(0, 0) \quad \text{et} \quad \left(\frac{1}{12}, -\frac{1}{6}\right).$$

Étude du point critique $(0, 0)$. On a $f(0, 0) = 0$ mais on remarque que $f(0, y) = -y^3$

qui peut être négatif ou positif (selon le signe de y proche de 0), donc en $(0, 0)$ il n'y a ni minimum ni maximum.

Étude du point critique $(\frac{1}{12}, -\frac{1}{6})$. Il existe un critère (que l'on ne décrira pas ici) qui permet de dire qu'en ce point f admet un minimum local.

Sur le dessin ci-dessous, le minimum est situé à l'intérieur du petit ovale, l'autre point critique en $(0, 0)$ correspond à l'intersection de la ligne de niveau $f = 0$ avec elle-même.

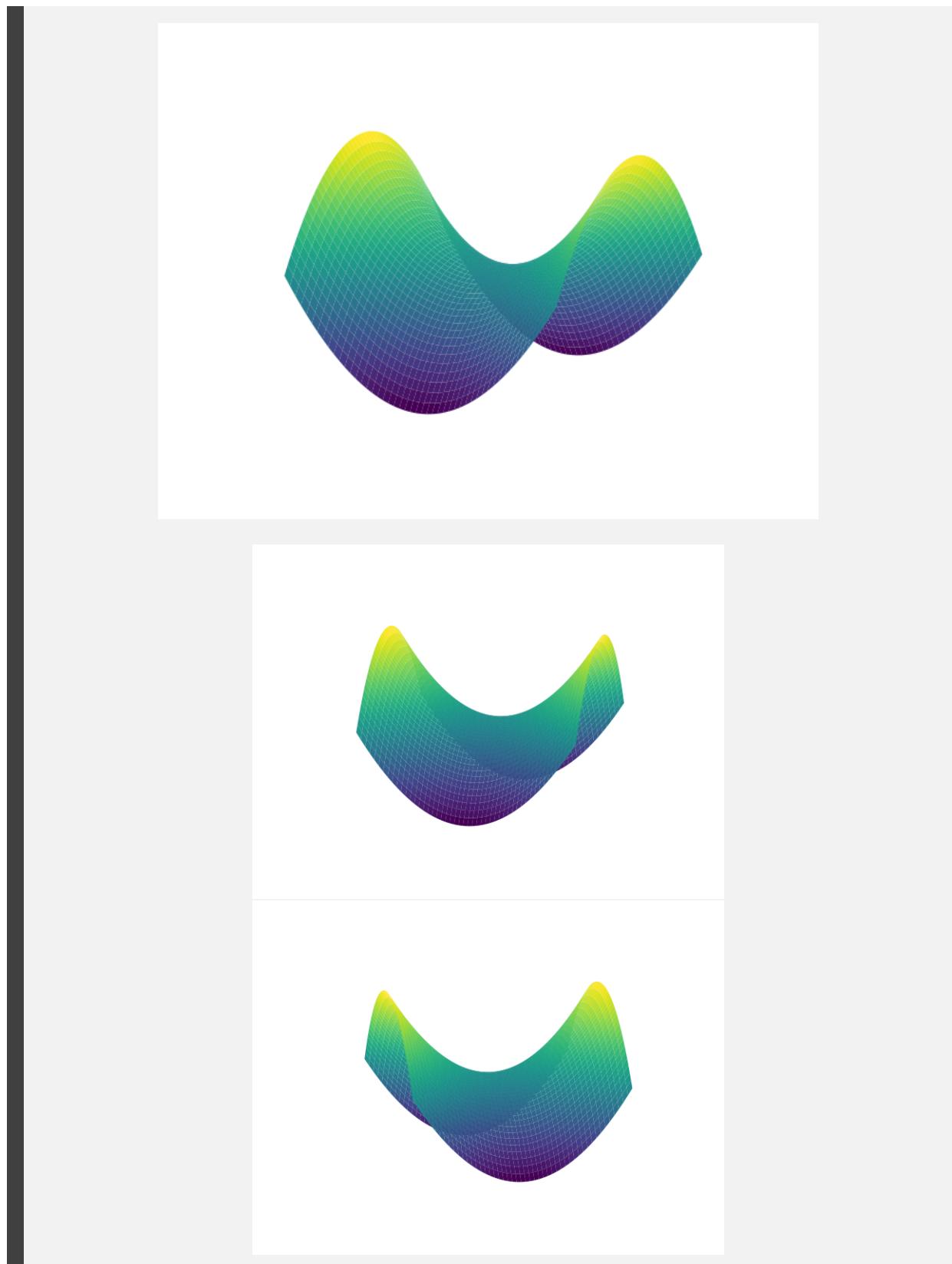


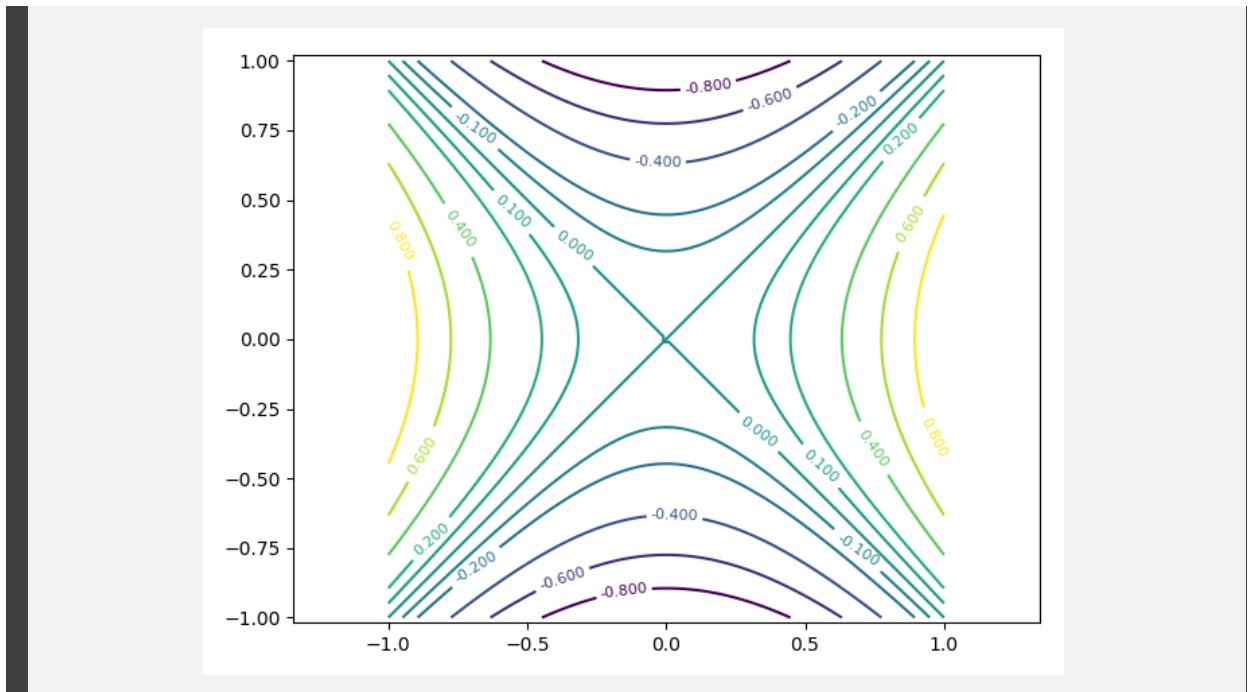
Sur l'exemple précédent, nous avons assez facilement calculé les points critiques à partir des deux équations à deux inconnues. Il faut prendre garde que ce n'est pas un système linéaire et que dans le cas d'une fonction plus compliquée il aurait été impossible de déterminer exactement les points critiques.

On note aussi dans l'exemple précédent que certains points critiques ne sont ni des maximums ni des minimums. L'exemple type, illustré ci-dessous, est celui d'un **col** appelé aussi **point-selle** en référence à sa forme de selle de cheval.

Exemple 7.3

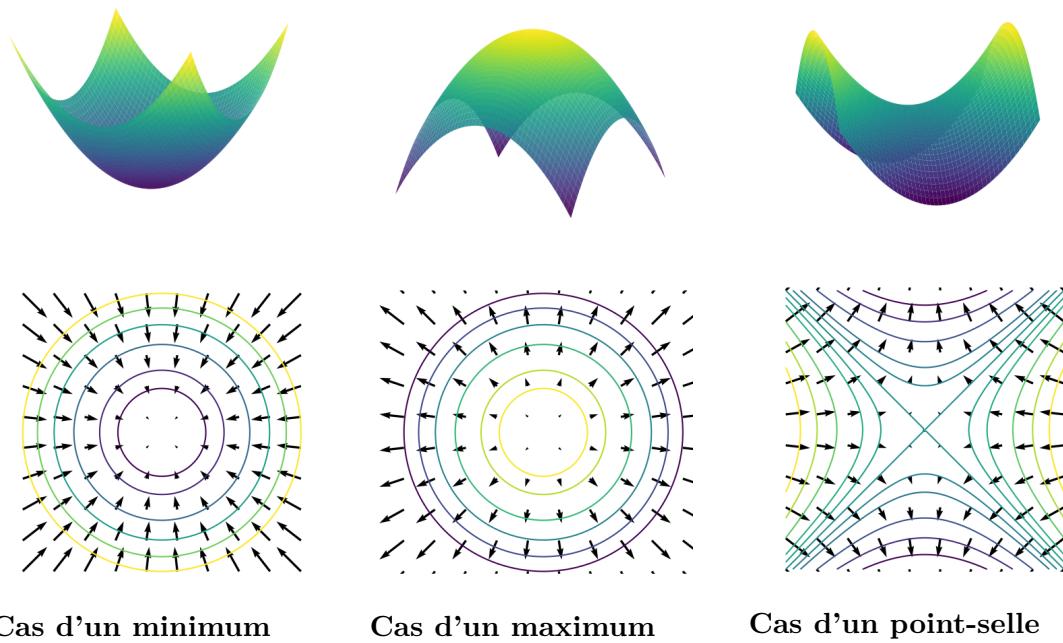
Soit $f(x, y) = x^2 - y^2$. Voici son graphe vu sous trois angles différents et ses lignes de niveau.





Comme il peut être difficile de calculer les points critiques de façon exacte, nous allons utiliser des méthodes numériques. L'idée qui sera détaillée dans le prochain chapitre est la suivante : comme le gradient indique la direction dans laquelle la fonction f croît le plus rapidement, nous allons suivre la direction opposée au gradient, pour laquelle f décroît le plus rapidement. Ainsi, partant d'un point (x_0, y_0) au hasard, on sait dans quelle direction se déplacer pour obtenir un nouveau point (x_1, y_1) en lequel f est plus petite. Et on recommence.

Sur les trois dessins ci-dessous, on a dessiné les lignes de niveau d'une fonction f ainsi que les vecteurs $-\text{grad } f(x, y)$. On voit que ces vecteurs pointent bien vers le minimum (figure de gauche), s'éloignent d'un maximum (figure centrale), le cas d'un point-selle est spécial (figure de droite). Dans tous les cas, la longueur des vecteurs gradients diminue à l'approche du point critique.

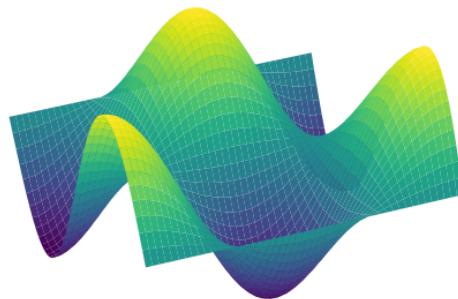


VI-

Descente de gradient classique

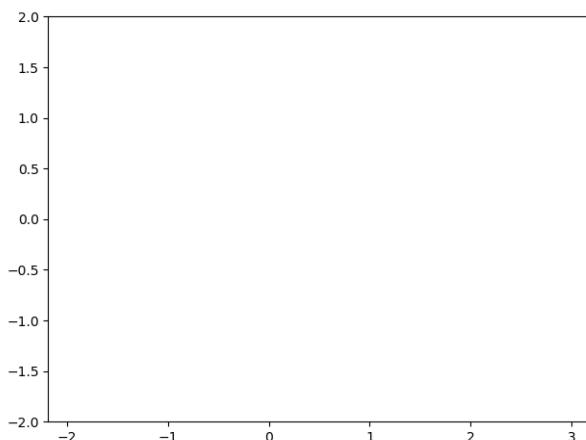
L'objectif de la méthode de descente de gradient est de trouver un minimum d'une fonction de plusieurs variables le plus rapidement possible. L'idée est très simple, on sait que le vecteur opposé au gradient indique une direction vers des plus petites valeurs de la fonction, il suffit donc de suivre d'un pas cette direction et de recommencer. Cependant, afin d'être encore plus rapide, il est possible d'ajouter plusieurs paramètres qui demandent pas mal d'ingénierie pour être bien choisis.

Imaginons une goutte d'eau en haut d'une colline. La goutte d'eau descend en suivant la ligne de plus grande pente et elle s'arrête lorsqu'elle atteint un point bas. C'est exactement ce que fait la descente de gradient : partant d'un point sur une surface, on cherche la pente la plus grande en calculant le gradient et on descend d'un petit pas, on recommence à partir du nouveau point jusqu'à atteindre un minimum local.

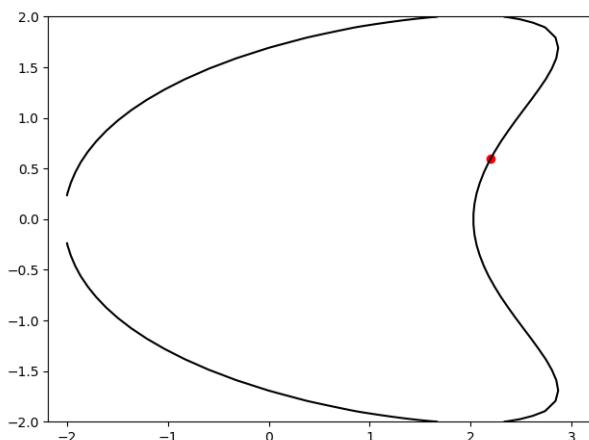
**1. Où est le minimum ?**

On nous donne une fonction f de deux variables (a, b) et nous cherchons un point (a_{\min}, b_{\min}) en lequel f atteint un minimum. Voici la méthode expliquée par des dessins sur lesquels ont été tracées des lignes de niveau.

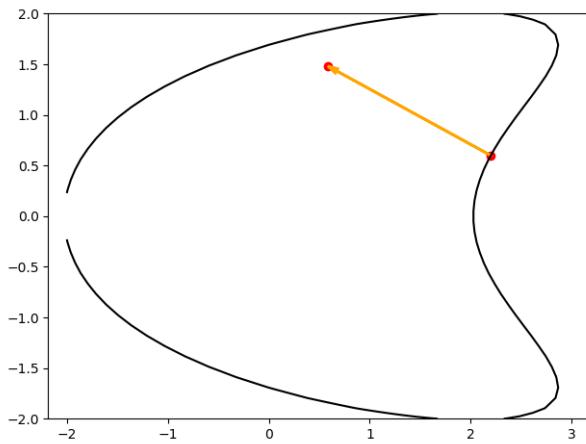
(a) Au départ : rien



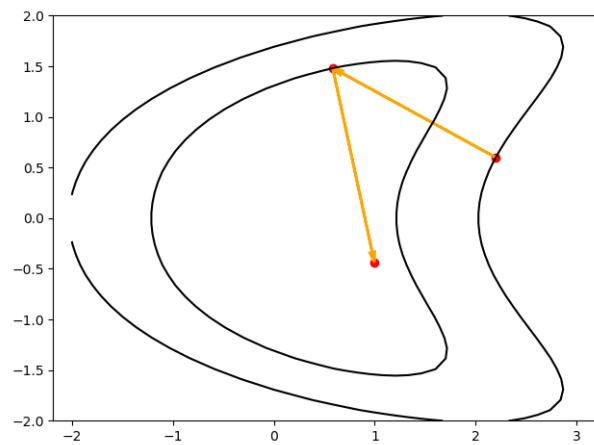
(b) On part d'un point au hasard



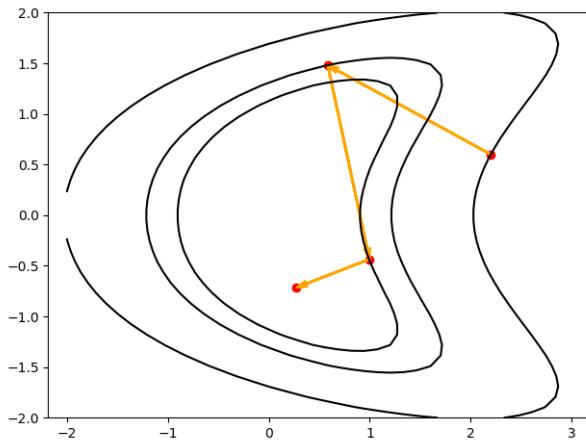
(c) On suit l'opposé du gradient



(d) Depuis le nouveau point, on suit l'opposé du gradient



(e) On répète le processus



(f) La suite converge vers le minimum

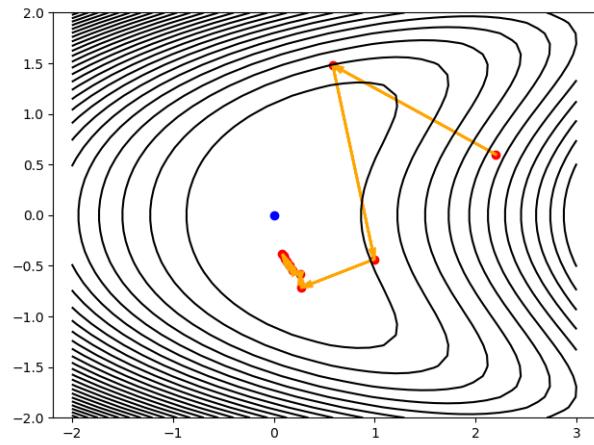
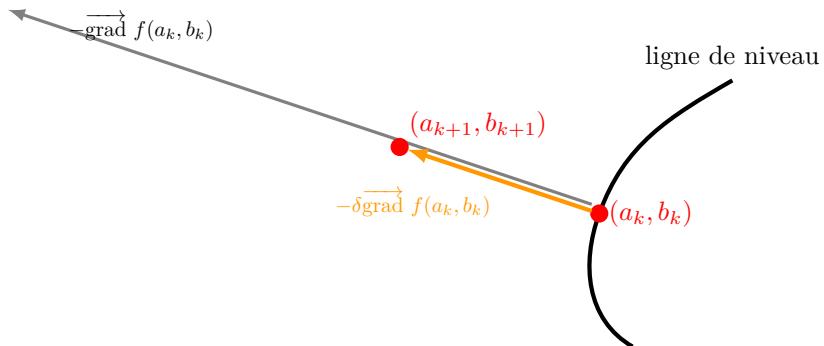


Figure (a). Au départ nous n'avons aucune information globale sur f . La seule opération que l'on s'autorise c'est calculer $\overrightarrow{\text{grad}} f(a, b)$ en certains points.

Figure (b). On choisit un point (a_0, b_0) au hasard. Si on note $c_0 = f(a_0, b_0)$ la valeur de f en ce point, on sait que la ligne de niveau ($f = c_0$) passe par (a_0, b_0) .

Figure (c). On calcule en ce point le gradient de f . On trace l'opposé du gradient : $-\overrightarrow{\text{grad}} f(a_0, b_0)$. On sait d'une part que la ligne de niveau est orthogonale à ce gradient et surtout que dans la direction de $-\overrightarrow{\text{grad}} f(a_0, b_0)$, les valeurs de f vont diminuer.



On se dirige alors dans la direction opposée au gradient d'un facteur δ (par exemple $\delta = 0.1$). On arrive à un point noté (a_1, b_1) . Par construction, si δ est assez petit, la valeur $c_1 = f(a_1, b_1)$ est plus petite que c_0 .

Figure (d). On recommence depuis (a_1, b_1) . On calcule l'opposé du gradient en (a_1, b_1) , on se dirige dans cette nouvelle direction pour obtenir un point (a_2, b_2) où $c_2 = f(a_2, b_2) < c_1$.

Figure (e). On itère le processus pour obtenir une suite de points (a_k, b_k) pour lesquels f prend des valeurs de plus en plus petites.

Figure (f). On choisit de s'arrêter (selon une condition préalablement établie) et on obtient une valeur approchée (a_N, b_N) du point (a_{\min}, b_{\min}) en lequel f atteint son minimum.

Évidemment avec la vision globale de la fonction, on se dit qu'on aurait pu choisir un point de départ plus près et que certaines directions choisies ne sont pas les meilleures. Mais souvenez-vous que l'algorithme est « aveugle », il ne calcule pas les valeurs de f en les (a_k, b_k) et n'a pas connaissance du comportement de f au voisinage de ces points.

2. Exemple en deux variables

Prenons l'exemple de $f(a, b) = a^2 + 3b^2$ dont le minimum est bien évidemment atteint en $(0, 0)$ et appliquons la méthode du gradient.

Nous aurons besoin de calculer la valeur du gradient en certains points par la formule :

$$\overrightarrow{\text{grad}} f(a, b) = \left(\frac{\partial f}{\partial a}(a, b), \frac{\partial f}{\partial b}(a, b) \right) = (2a, 6b).$$

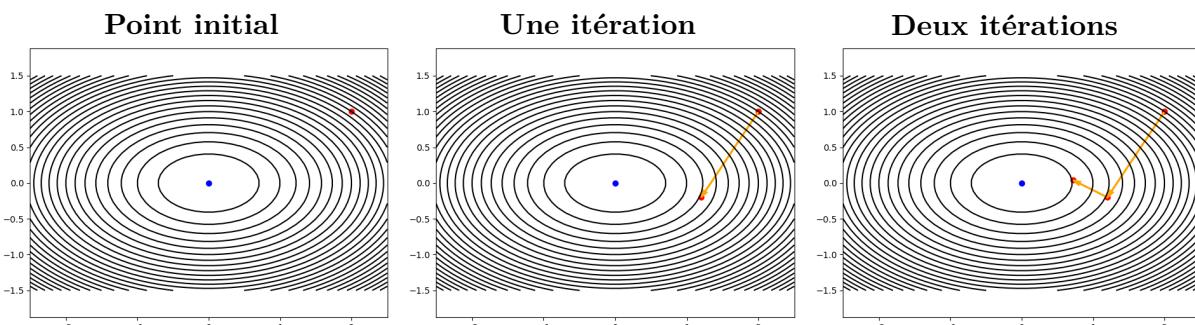
Tout d'abord, on part d'un point $(a_0, b_0) = (2, 1)$ par exemple. Même si nous n'en avons pas besoin pour notre construction, on a $f(a_0, b_0) = 7$. On calcule $\overrightarrow{\text{grad}} f(a_0, b_0) = (4, 6)$. On fixe le facteur $\delta = 0.2$. On se déplace dans la direction opposée à ce gradient :

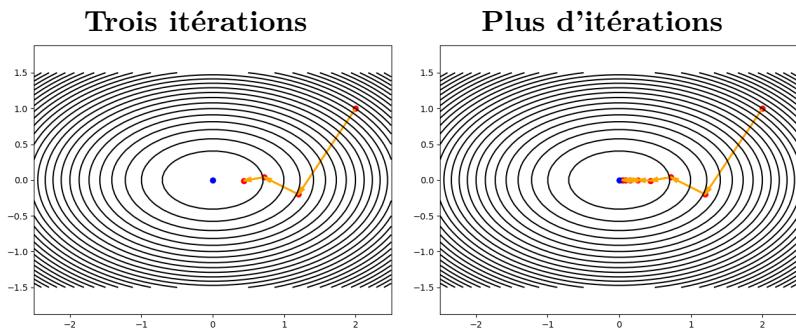
$$(a_1, b_1) = (a_0, b_0) - \delta \overrightarrow{\text{grad}} f(a_0, b_0) = (2, 1) - 0.2(4, 6) = (2, 1) - (0.8, 1.2) = (1.2, -0.2).$$

On note que $f(a_1, b_1) = 1.56$ est bien plus petit que $f(a_0, b_0)$. On recommence ensuite depuis (a_1, b_1) . En quelques étapes les valeurs de f tendent vers la valeur minimale et, dans notre cas, la suite converge vers $(0, 0)$ (les valeurs sont approchées).

k	(a_k, b_k)	$\overrightarrow{\text{grad}} f(a_k, b_k)$	$f(a_k, b_k)$
0	$(2, 1)$	$(4, 6)$	7
1	$(1.2, -0.2)$	$(2.4, -1.20)$	1.56
2	$(0.72, 0.04)$	$(1.44, 0.24)$	0.523
3	$(0.432, -0.008)$	$(0.864, -0.048)$	0.186
4	$(0.2592, 0.0016)$	$(0.5184, 0.0096)$	0.067
5	$(0.15552, -0.00032)$	$(0.31104, -0.00192)$	0.024
...			
10	$(0.012, 1.02 \cdot 10^{-7})$	$(0.024, 6.14 \cdot 10^{-7})$	0.00014
...			
20	$(7.31 \cdot 10^{-5}, 1.04 \cdot 10^{-14})$	$(1.46 \cdot 10^{-4}, 6.29 \cdot 10^{-14})$	$5.34 \cdot 10^{-9}$

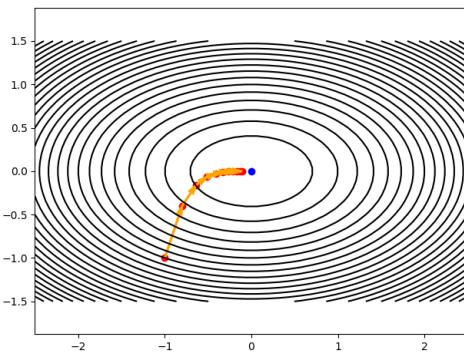
Voici les graphiques des premières itérations :





Que se passe-t-il si l'on part d'un autre point ? Partons cette fois de $(a_0, b_0) = (-1, -1)$ et fixons le pas à $\delta = 0.1$. Alors $(a_1, b_1) = (-0.8, -0.4)$, $(a_2, b_2) = (-0.64, -0.16)\dots$ La suite converge également vers $(0, 0)$.

Partant de $(-1, -1)$ avec $\delta = 0.1$



3. Exemples en une variable

La descente de gradient fonctionne aussi très bien pour les fonctions d'une seule variable et sa visualisation est instructive.

Exemple 3.1 : C

Considérons la fonction $f : \mathbb{R} \rightarrow \mathbb{R}$ définie par

$$f(a) = a^2 + 1.$$

Il s'agit de trouver la valeur en laquelle f atteint son minimum, c'est clairement $a_{\min} = 0$ pour lequel $f(a_{\min}) = 1$. Retrouvons ceci par la descente de gradient.

Partant d'une valeur a_0 quelconque, la formule de récurrence est :

$$a_{k+1} = a_k - \delta \overrightarrow{\text{grad}} f(a)$$

où δ est le pas, choisi assez petit, et $\overrightarrow{\text{grad}} f(a) = f'(a) = 2a$. Autrement dit :

$$a_{k+1} = a_k - 2\delta a_k.$$

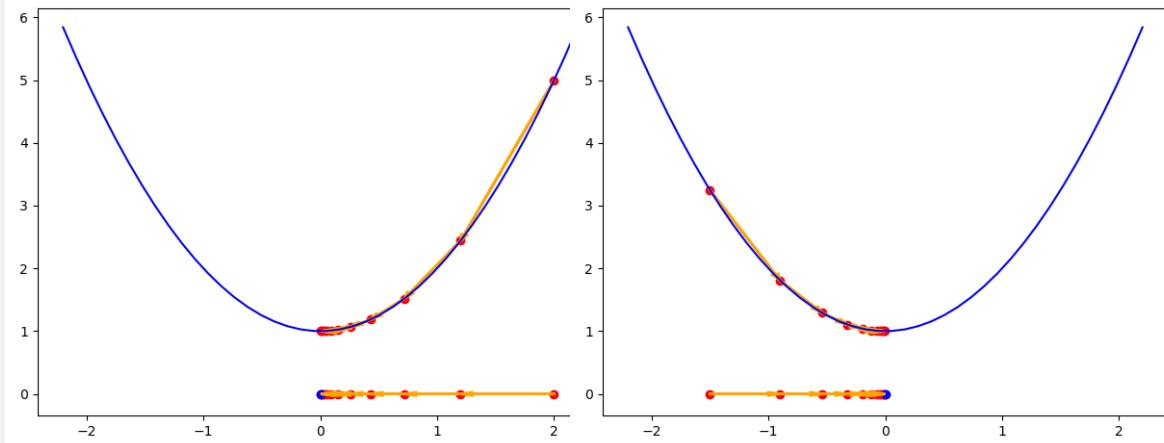
Voici le tableau des valeurs pour un pas $\delta = 0.2$ et une valeur initiale $a_0 = 2$.

k	a_k	$f'(a_k) = \overrightarrow{\text{grad}} f(a_k)$	$f(a_k)$
0	2	4	5
1	1.2	2.4	2.44
2	0.72	1.44	1.5184
3	0.43	0.86	1.1866
4	0.25	0.5184	1.0671
5	0.15	0.31	1.0241
6	0.093	0.186	1.0087
7	0.055	0.111	1.0031
8	0.033	0.067	1.0011
9	0.020	0.040	1.0004
10	0.012	0.024	1.0001

Voici la version graphique de ces 10 premières itérations (figure de gauche). Si l'on change le point initial, ($a_0 = -1.5$ sur la figure de droite) alors la suite (a_k) converge vers la même valeur $a_{\min} = 0$.

$$\delta = 0.2 \quad a_0 = 2$$

$$\delta = 0.2 \quad a_0 = -1.5$$



Il faut bien comprendre ce graphique : la suite des points (a_k) se lit sur l'axe des abscisses. Les vecteurs montrent les itérations. Il est plus facile de comprendre l'algorithme sur le graphe de f . Sur ce graphe, on reporte les points $(a_k, f(a_k))$, ce qui permet de bien comprendre que les valeurs $f(a_k)$ décroissent rapidement. On note aussi que le gradient (ici $f'(a_k)$) diminue à l'approche du minimum, ce qui se traduit par des vecteurs (c'est-à-dire l'écart entre deux points successifs) de plus en plus petits.

Justifions l'algorithme et l'intervention du gradient dans le cas d'une variable. Si la fonction est croissante sur un intervalle, $f'(a) > 0$ pour tout a dans cet intervalle et la formule

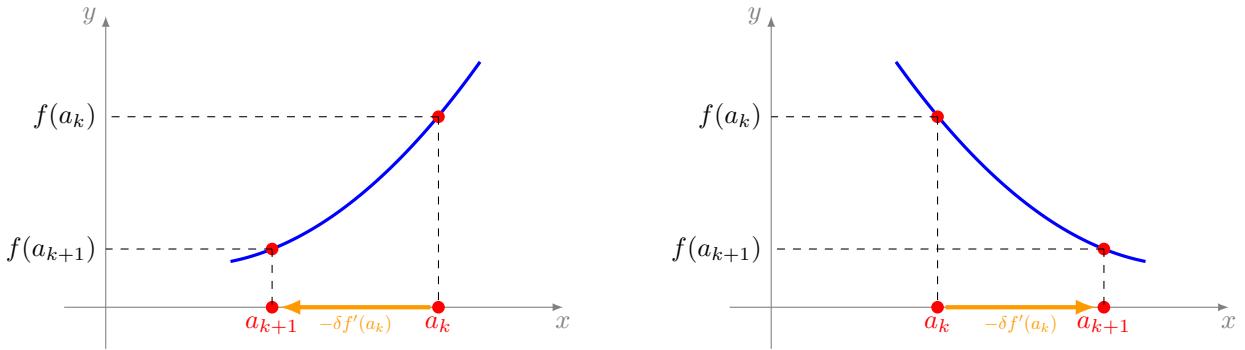
$$a_{k+1} = a_k - \delta f'(a_k) \quad \text{donne} \quad a_{k+1} < a_k.$$

Ainsi $f(a_{k+1}) < f(a_k)$ et l'ordonnée du point $(a_{k+1}, f(a_{k+1}))$ est donc inférieure à celle du point $(a_k, f(a_k))$. Par contre, si f est décroissante alors $f'(a) < 0$ et

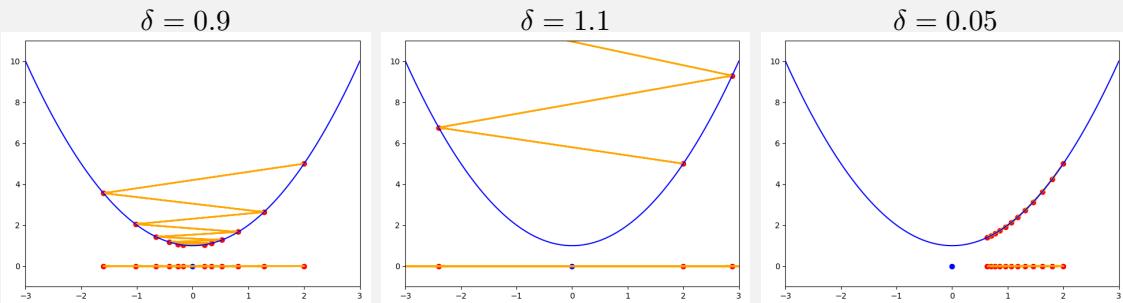
$$a_{k+1} = a_k - \delta f'(a_k) \quad \text{donne} \quad a_{k+1} > a_k,$$

ce qui implique de nouveau $f(a_{k+1}) < f(a_k)$ (car f est décroissante).

Dans tous les cas, l'ordonnée du point $(a_{k+1}, f(a_{k+1}))$ est inférieure à celle du point $(a_k, f(a_k))$.

**Exemple 3.2**

Le choix du paramètre δ est important. Reprenons la fonction f définie par $f(x) = x^2 + 1$ et testons différentes « mauvaises » valeurs du pas δ (avec toujours $a_0 = 2$).



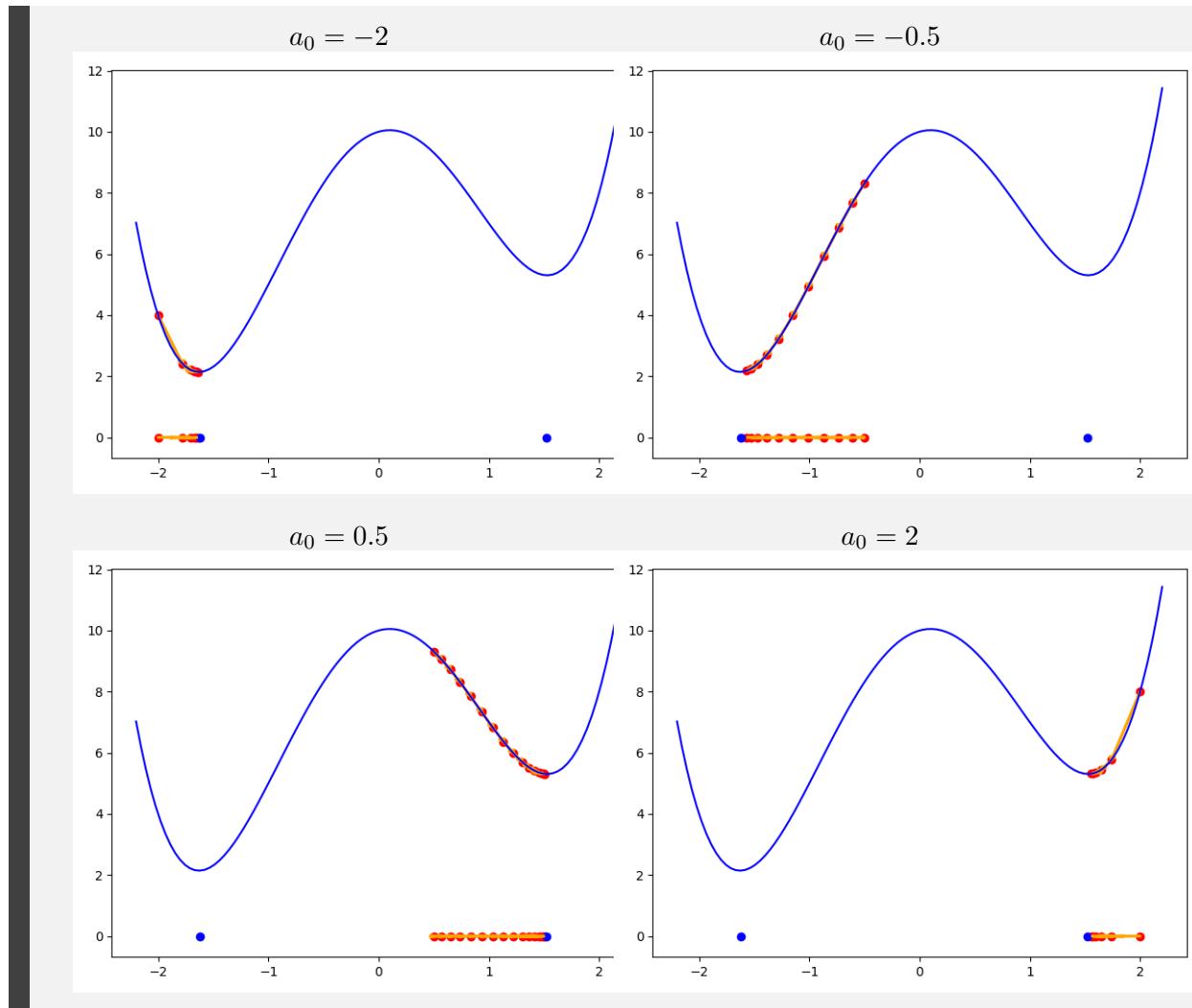
- Pour $\delta = 0.9$, la suite (a_k) tend bien vers $a_{\min} = 0$. Les ordonnées sont bien décroissantes mais comme δ est trop grand, la suite des points oscille de part et d'autre du minimum.
- Pour $\delta = 1.1$, la suite (a_k) diverge. Les ordonnées augmentent, la suite des points oscille et s'échappe. Cette valeur de δ ne donne pas de convergence vers un minimum.
- Pour $\delta = 0.05$, la suite (a_k) tend bien vers a_{\min} mais, comme δ est trop petit, il faudrait beaucoup d'itérations pour arriver à une approximation raisonnable.

Exemple 3.3

Le choix du point de départ est également important surtout lorsqu'il existe plusieurs minima locaux. Soit la fonction f définie par :

$$f(a) = a^4 - 5a^2 + a + 10.$$

Cette fonction admet deux minima locaux. La suite (a_k) de la descente de gradient converge vers l'un de ces deux minima selon le choix du point initial a_0 (ici $\delta = 0.02$).



Exemple 3.4

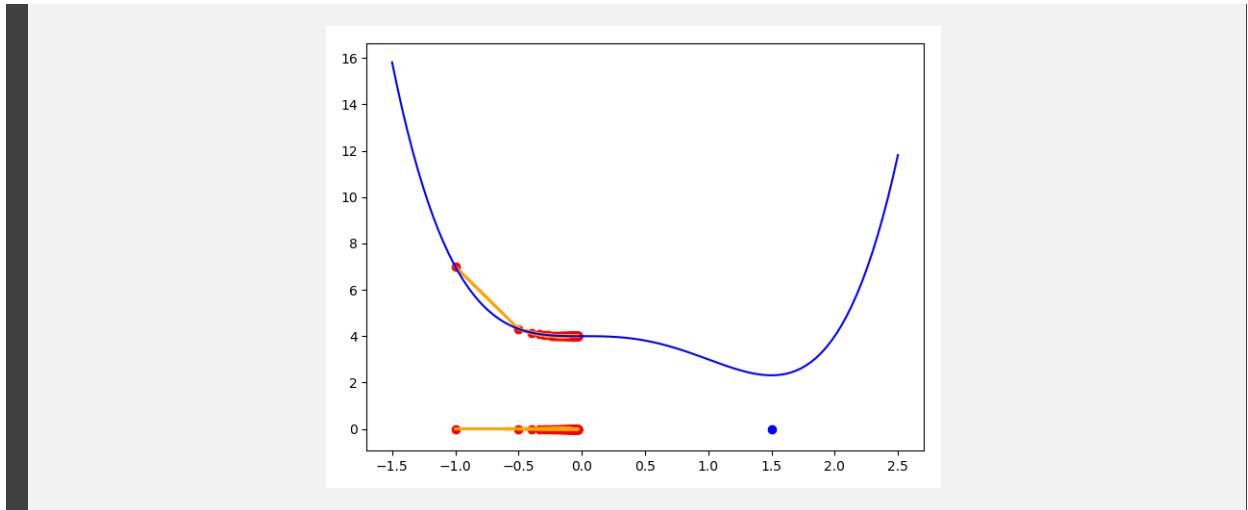
Les points-selles posent également problème.

La fonction f définie par

$$f(a) = a^4 - 2a^3 + 4$$

a pour dérivée $f'(a) = 4a^3 - 6a^2$ qui s'annule en $a = 0$ qui est l'abscisse d'un point-selle (ni un minimum ni un maximum, en fait la fonction est strictement décroissante autour de $a = 0$). La dérivée s'annule aussi en $a = \frac{3}{2}$ où est atteint le minimum global.

Voici les 100 premières itérations pour la descente de gradient en partant de $a_0 = -1$ (avec $\delta = 0.05$) : la suite a_k converge vers 0 qui n'est pas le minimum recherché.



4. Algorithme du gradient

Formalisons un peu les choses pour mettre en évidence l'idée générale et les problèmes techniques qui surviennent.

Algorithme de la descente de gradient.

Soit une fonction $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $P \mapsto f(P)$ de plusieurs variables, avec $P = (a_1, \dots, a_n)$, dont on sait calculer le gradient $\overrightarrow{\text{grad}} f(P)$.

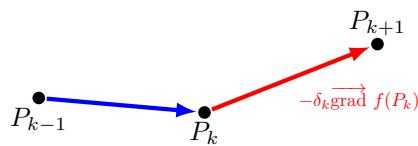
Données.

- Un point initial $P_0 \in \mathbb{R}^n$.
- Un niveau d'erreur $\epsilon > 0$.

Itération. On calcule une suite de points $P_1, P_2, \dots \in \mathbb{R}^n$ par récurrence de la façon suivante. Supposons que l'on ait déjà obtenu le point P_k :

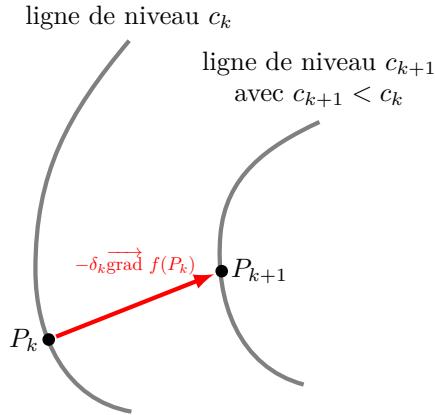
- on calcule $\overrightarrow{\text{grad}} f(P_k)$,
- on choisit un pas δ_k et on calcule

$$P_{k+1} = P_k - \delta_k \overrightarrow{\text{grad}} f(P_k).$$



Arrêt. On s'arrête lorsque $\|\overrightarrow{\text{grad}} f(P_k)\| \leq \epsilon$.

- Évidemment, plus on choisit le point initial P_0 proche d'un minimum local, plus l'algorithme va aboutir rapidement. Mais comme on ne sait pas où est ce minimum local (c'est ce que l'on cherche), le plus simple est de choisir un P_0 au hasard.
- Le choix du pas δ_k est crucial. On sait que l'on peut choisir δ_k assez petit de façon à avoir $f(P_{k+1}) \leq f(P_k)$ car dans la direction de $-\overrightarrow{\text{grad}} f(P_k)$ la fonction f décroît.



On peut fixer à l'avance un pas δ commun à toutes les itérations, par exemple $\delta = 0.01$. On pourrait également tester à chaque itération plusieurs valeurs de δ par balayage ($\delta = 0.001$, puis $\delta = 0.002\dots$) et choisir pour δ_k celui en lequel f prend la plus petite valeur.

- Le critère d'arrêt assure qu'en P_k le gradient est très petit. Cela ne garantit pas que ce point soit proche d'un minimum local (et encore moins d'un minimum global). Souvenez-vous : en un minimum local le gradient est nul, mais ce n'est pas parce que le gradient est nul que l'on a atteint un minimum local, cela pourrait être un point-selle voire un maximum local.
- Dans la pratique, on ne définira pas de seuil d'erreur ϵ , mais un nombre d'itérations fixé à l'avance.
- Il est important de calculer $\vec{\text{grad}} f(a_1, \dots, a_n)$ rapidement. On pourrait bien sûr calculer une approximation de chacune des dérivées partielles $\frac{\partial f}{\partial a_i}(a_1, \dots, a_n)$ comme un limite. Mais pour gagner en temps et en précision, on préfère que ce calcul soit fait à l'aide de son expression exacte.

VII- Optimisation

Nous allons d'une part résoudre des problèmes d'optimisation : quelle droite approche au mieux un nuage de points, et d'autre part étudier comment améliorer le choix du pas δ .

1. Faire varier le pas

On se concentre d'abord sur le choix du **pas** δ (*learning rate*).

Rappelons tout d'abord que lorsque l'on se rapproche d'un point minimum, le gradient tend vers 0. Le vecteur $\vec{\text{grad}} f(P_k)$ tend donc vers 0 à l'approche du minimum, même si δ reste constant.

Cependant, il faut choisir δ ni trop grand, ni trop petit : δ ne doit pas être trop grand car sinon les points P_k vont osciller autour du minimum, mais si δ est trop petit alors les points P_k ne s'approcheront du minimum qu'au bout d'un temps très long. Une solution est de faire varier δ . Pour les premières itérations, on choisit un δ_k assez grand, puis de plus en plus petit au fil des itérations.

Voici différentes formules possibles, à chaque fois δ_0 est le pas initial (par exemple $\delta_0 = 0.1$ ou $\delta_0 = 0.01$).

Décroissance linéaire.

$$\delta_k = \frac{\delta_0}{k + 1}.$$

Décroissance quadratique.

$$\delta_k = \frac{\delta_0}{(k + 1)^2}.$$

Décroissance exponentielle.

$$\delta_k = \delta_0 e^{-\beta k}$$

où β est une constante positive.

Décroissance linéaire utilisée par keras

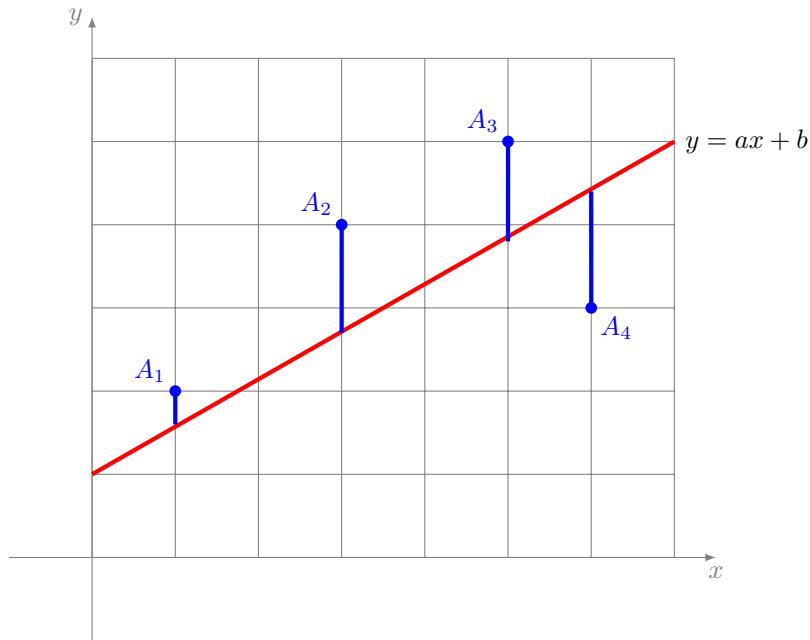
$$\delta_k = \frac{\delta_0}{\alpha k + 1}$$

où $\alpha \geq 0$ est une constante (appelée *decay*). Si $\alpha = 0$ alors δ_k est constant (et vaut δ_0). L'usage courant est d'utiliser des valeurs de α entre 10^{-4} et 10^{-6} .

Terminons par rappeler que le bon choix d'un δ ou des δ_k n'a rien d'évident, il s'obtient soit par test à la main, soit par des expérimentations automatiques, mais à chaque fois il doit être adapté à la situation.

2. Régression linéaire $y = ax + b$

On considère un ensemble de N points $A_i = (x_i, y_i)$, $i = 1, \dots, N$. L'objectif est de trouver l'équation $y = ax + b$ de la droite qui approche au mieux tous ces points. Précisons ce que veut dire « approcher au mieux » : il s'agit de minimiser la somme des carrés des distances verticales entre les points et la droite.



La formule qui donne l'erreur est :

$$E(a, b) = \sum_{i=1}^N (y_i - (ax_i + b))^2,$$

autrement dit

$$E(a, b) = (y_1 - (ax_1 + b))^2 + \dots + (y_N - (ax_N + b))^2.$$

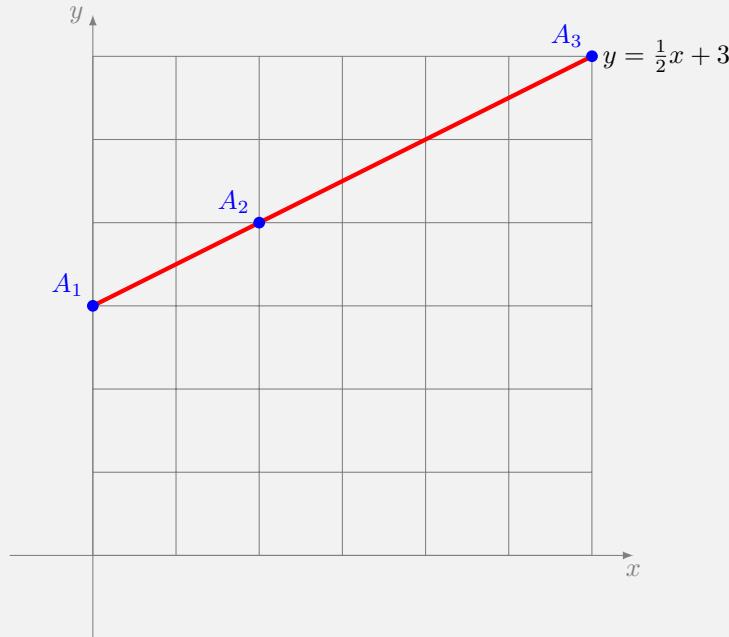
Remarquons que l'on a toujours $E(a, b) \geq 0$. Si par exemple tous les points sont alignés, alors on peut trouver a et b tels que $E(a, b) = 0$. Quand ce n'est pas le cas, on cherche a et b qui rendent $E(a, b)$ le plus petit possible. Il s'agit donc bien ici de minimiser une fonction de deux variables (les variables sont a et b).

Nous allons appliquer la méthode de la descente de gradient à la fonction $E(a, b)$. Pour cela nous aurons besoin de calculer son gradient :

$$\overrightarrow{\text{grad}} E(a, b) = \left(\frac{\partial E}{\partial a}(a, b), \frac{\partial E}{\partial b}(a, b) \right) = \left(\sum_{i=1}^N -2x_i(y_i - (ax_i + b)), \sum_{i=1}^N -2(y_i - (ax_i + b)) \right).$$

Exemple 2.1

Prenons d'abord l'exemple de trois points $A_1 = (0, 3)$, $A_2 = (2, 4)$ et $A_3 = (6, 6)$ qui sont alignés.



La fonction $E(a, b)$ s'écrit :

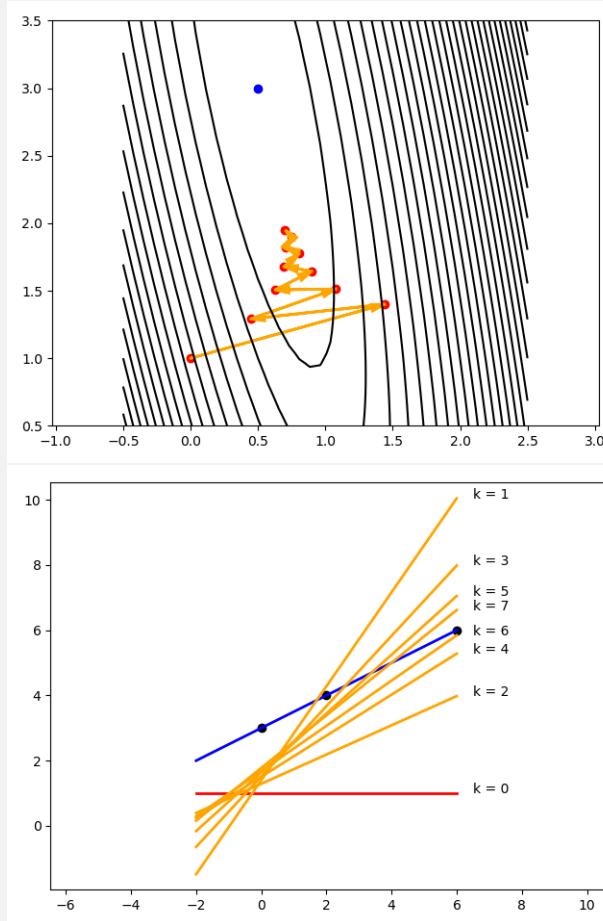
$$E(a, b) = (3 - b)^2 + (4 - (2a + b))^2 + (6 - (6a + b))^2.$$

Partons arbitrairement de $(a_0, b_0) = (0, 1)$ (qui correspond à la droite horizontale d'équation $y = 1$). Voici les valeurs successives (a_k, b_k) obtenues par la méthode de descente de gradient pour un pas $\delta = 0.02$.

k	(a_k, b_k)	$\overrightarrow{\text{grad}} E(a_k, b_k)$	$E(a_k, b_k)$
0	(0, 1)	(-72, -20)	38
1	(1.44, 1.4)	(49.60, 5.44)	18.96
2	(0.44, 1.29)	(-31.50, -11.08)	10.28
3	(1.07, 1.51)	(22.44, 0.32)	6.24
4	(0.62, 1.50)	(-13.57, -6.89)	4.27
5	(0.90, 1.64)	(10.35, -1.72)	3.24

Au bout de 100 itérations, on obtient $a_{100} \simeq 0.501$ et $b_{100} \simeq 2.99$ (avec un gradient et une erreur presque nuls). C'est bien la droite $y = \frac{1}{2}x + 3$ qui passe par les trois points.

Sur la figure de gauche ci-dessous, sont dessinés, dans le plan de coordonnées (a, b) , les premiers points (a_k, b_k) qui convergent (lentement et en oscillant) vers $(\frac{1}{2}, 3)$. Sur la figure de droite sont tracées, dans le plan de coordonnées (x, y) , les droites d'équation $y = a_k x + b_k$ pour les premières valeurs de k . Il est beaucoup plus difficile d'appréhender la convergence des droites (vers la droite d'équation $y = \frac{1}{2}x + 3$) que celle des points de la figure de gauche.



Exemple 2.2

À partir des données des 5 points suivants, quelle ordonnée peut-on extrapoler pour le point d'abscisse $x = 6$?

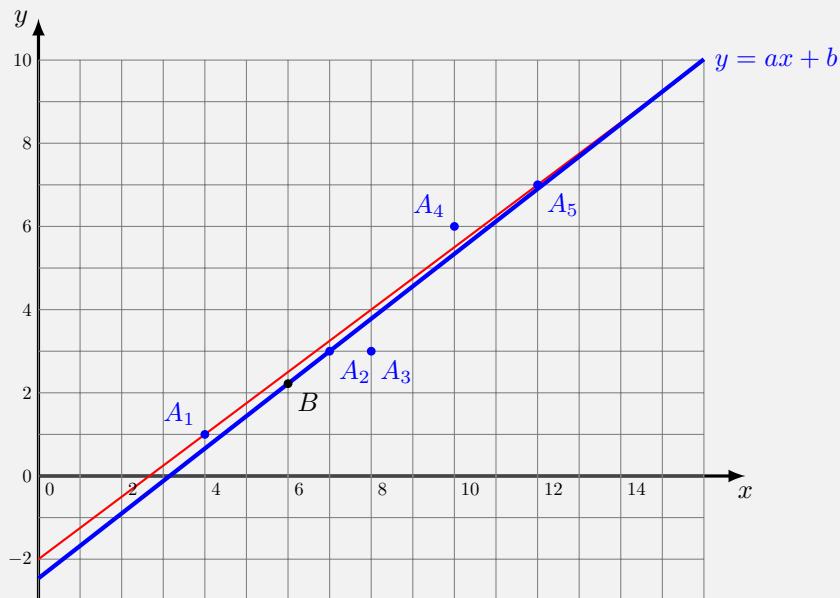
$$A_1 = (4, 1), \quad A_2 = (7, 3), \quad A_3 = (8, 3), \quad A_4 = (10, 6), \quad A_5 = (12, 7).$$



Ces 5 points sont à peu près alignés. On calcule la meilleure droite de régression linéaire par

la descente de gradient. Cela revient par exemple à minimiser la fonction $E(a, b)$ présentée ci-dessus, mais actualisée avec nos données. On fixe un pas $\delta = 0.001$. On ne choisit pas le point initial (a_0, b_0) au hasard. Plus on part d'un point proche de la solution, plus la suite convergera rapidement. On trace la droite qui passe par le premier point A_1 et le dernier point A_5 . Cette droite a pour équation $y = \frac{3}{4}x - 2$ et est déjà une droite qui approche assez bien les 5 points. Prenons cette droite comme point de départ, c'est-à-dire posons $(a_0, b_0) = (\frac{3}{4}, -2)$. La descente de gradient conduit au bout de 1000 itérations à $a \simeq 0.78$ et $b \simeq -2.46$, pour l'équation de la droite de régression linéaire.

Sur le dessin ci-dessous sont tracées la droite initiale qui passe par les points A_1 et A_5 et la droite de régression linéaire.



N'oublions pas de répondre à la question initiale. Selon notre modèle linéaire, pour $x = 6$, on doit avoir $y = ax + b \simeq 2.22$ (le point B de la figure ci-dessus).

Remarque : il existe une formule directe pour calculer exactement les coefficients a et b de la droite de régression linéaire, mais ce n'est pas l'esprit de ce cours.

VIII-

Descente de gradient stochastique

La descente de gradient stochastique (abrégée en *sgd*) est une façon d'optimiser les calculs de la descente de gradient pour une fonction d'erreur associée à une grande série de données. Au lieu de calculer un gradient (compliqué) et un nouveau point pour l'ensemble des données, on calcule un gradient (simple) et un nouveau point par donnée, il faut répéter ce processus pour chaque donnée.

1. Petits pas à petits pas

Revenons à l'objectif visé par la régression linéaire.

On considère des données (X_i, y_i) , $i = 1, \dots, N$ où $X_i \in \mathbb{R}^\ell$ et $y_i \in \mathbb{R}$. Ces données proviennent d'observations ou d'expérimentations.

Il s'agit de trouver une fonction $F : \mathbb{R}^\ell \rightarrow \mathbb{R}$ qui modélise au mieux ces données, c'est-à-dire telle que

$$F(X_i) \simeq y_i.$$

Pour l'**entrée** X_i , la valeur y_i est la **sortie attendue**, alors que $F(X_i)$ est la **sortie produite** par notre modèle.

Pour mesurer la pertinence de la fonction F , on introduit la fonction d'**erreur totale** qui mesure l'écart entre la sortie attendue et la sortie produite :

$$E = \sum_{i=1}^N E_i = \sum_{i=1}^N (y_i - F(X_i))^2.$$

Cette erreur totale est une somme d'**erreurs locales** :

$$E_i = (y_i - F(X_i))^2.$$

Le but du problème est de déterminer la fonction F qui minimise l'erreur E . Par exemple, dans le cas de la régression linéaire, il fallait trouver les paramètres a et b pour définir $F(x) = ax + b$, ou bien, pour deux variables, les paramètres a , b , c pour définir $F(x, y) = ax + by + c$.

Considérons une fonction erreur $E : \mathbb{R}^n \rightarrow \mathbb{R}$ qui dépend de n paramètres a_1, \dots, a_n (qui définissent l'expression de la fonction F).

Descente de gradient classique. Pour minimiser l'erreur et déterminer les meilleurs paramètres, on peut appliquer la méthode du gradient classique.

On part d'un point $P_0 = (a_1, \dots, a_n) \in \mathbb{R}^n$, puis on applique la formule de récurrence :

$$\overrightarrow{P_{k+1}} = \overrightarrow{P_k} - \delta \overrightarrow{\text{grad}} E(P_k).$$

Pour appliquer cette formule, il faut calculer des gradients $\overrightarrow{\text{grad}} E(P_k)$, or

$$\overrightarrow{\text{grad}} E(P_k) = \sum_{i=1}^N \overrightarrow{\text{grad}} E_i(P_k).$$

Il faut donc calculer une somme de N termes à chaque itération, ce qui pose des problèmes d'efficacité pour de grandes valeurs de N .

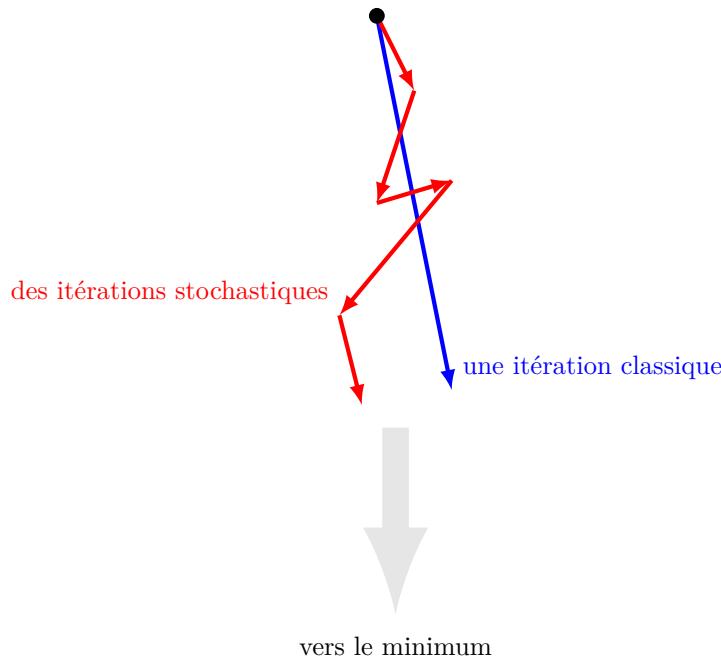
Descente de gradient stochastique.

Pour diminuer la quantité de calculs, l'idée est de considérer à chaque itération un seul gradient E_i à la place de E . C'est-à-dire :

$$\overrightarrow{P_{k+1}} = \overrightarrow{P_k} - \delta \overrightarrow{\text{grad}} E_i(P_k)$$

pour une seule erreur E_i (correspondant à la donnée numéro i). L'itération suivante se basera sur l'erreur E_{i+1} .

Quel est l'intérêt de cette méthode ? Dans la méthode de gradient classique, on calcule à chaque itération un « gros » gradient (associé à la totalité des N données) qui nous rapproche d'un grand pas vers le minimum. Ici on calcule N « petits » gradients qui nous rapprochent du minimum.



Voici les premières itérations de cet algorithme.

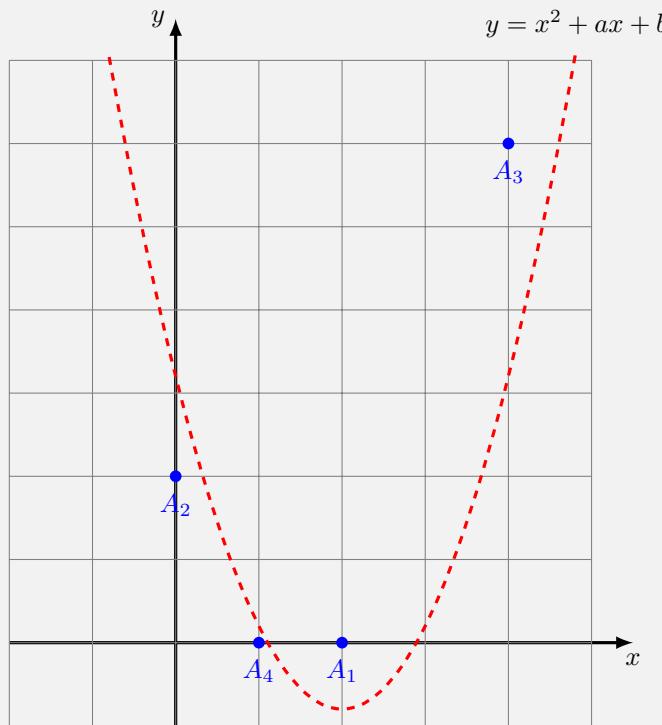
- On part d'un point P_0 .
- On calcule $P_1 = P_0 - \delta \overrightarrow{\text{grad}} E_1(P_0)$. C'est la formule du gradient, mais seulement pour l'erreur locale E_1 (juste à partir de la première donnée (X_1, y_1)).
- On calcule $P_2 = P_1 - \delta \overrightarrow{\text{grad}} E_2(P_1)$. C'est la formule du gradient, mais seulement pour l'erreur locale E_2 .
- On itère encore et encore.
- On calcule $P_N = P_{N-1} - \delta \overrightarrow{\text{grad}} E_N(P_{N-1})$. C'est la formule du gradient, mais seulement pour l'erreur locale E_N . À ce stade de l'algorithme, nous avons tenu compte de toutes les données.
- On calcule $P_{N+1} = P_N - \delta \overrightarrow{\text{grad}} E_1(P_N)$. On recommence pour P_N et l'erreur locale E_1 .
- Etc. On s'arrête au bout d'un nombre d'étapes fixé à l'avance ou lorsque l'on est suffisamment proche du minimum.

Exemple 1.1

On considère les quatre points :

$$A_1 = (2, 0), \quad A_2 = (0, 2), \quad A_3 = (4, 6), \quad A_4 = (1, 0).$$

Comme ces points ne sont clairement pas alignés, on cherche un modèle pour les placer au mieux sur une parabole.



On va ici chercher des coefficients a et b tels que les points soient proches de la parabole d'équation $y = x^2 + ax + b$. On note donc $F(x) = x^2 + ax + b$ et on souhaite $F(x_i) \simeq y_i$ pour les points $A_i = (x_i, y_i)$, $i = 1, \dots, 4$.

Les fonctions d'erreurs locales sont :

$$E_i(a, b) = (y_i - (x_i^2 + ax_i + b))^2.$$

La fonction d'erreur globale est :

$$E(a, b) = E_1(a, b) + E_2(a, b) + E_3(a, b) + E_4(a, b).$$

Voici les premières itérations pour chacune des méthodes, la descente de gradient classique (à gauche), la descente de gradient stochastique (à droite) toutes les deux en partant du point $(a_0, b_0) = (1, 1)$ et avec $\delta = 0.01$.

Descente de gradient

Descente de gradient stochastique

k	(a_k, b_k)	$E(a_k, b_k)$	k	(a'_k, b'_k)	$E(a'_k, b'_k)$
0	(1, 1)	284	0	(1, 1)	284
			1	(0.72, 0.86)	236.43
			2	(0.72, 0.88)	237.41
1	(-0.54, 0.52)	84.87	3	(-0.38, 0.60)	100.74
			4	(-0.40, 0.58)	97.917
			5	(-0.55, 0.50)	83.245
			6	(-0.55, 0.53)	83.913
			7	(-1.22, 0.37)	36.48
			8	(-1.22, 0.36)	36.29

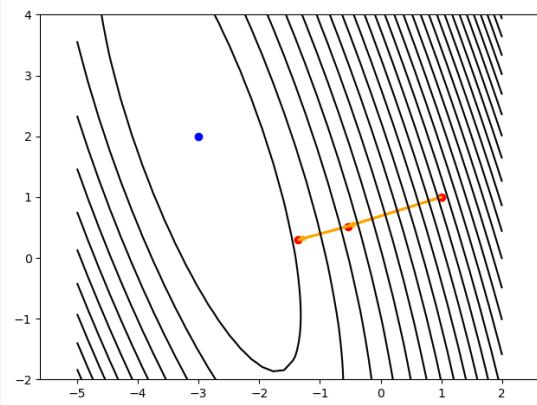
Au bout de 200 itérations la descente de gradient classique conduit à $(a_{200}, b_{200}) \simeq (-2.9981, 1.9948)$ (chaque donnée a été utilisée 200 fois).

Cela correspond à 800 itérations de la descente de gradient stochastique (chacune des 4 données a été utilisée 200 fois) cela conduit à $(a'_{800}, b'_{800}) \simeq (-2.9984, 1.9954)$. La limite

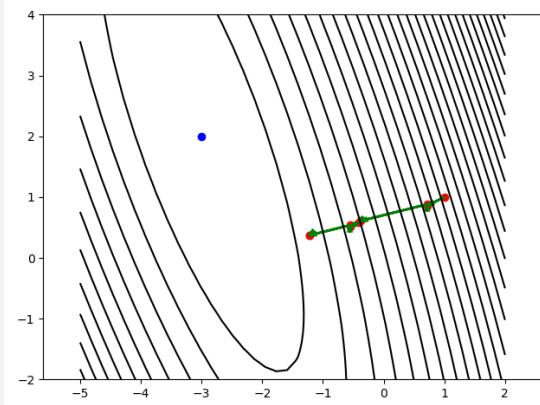
cherchée étant $(a, b) = (-3, 2)$, avec $E(a, b) = 0$, les deux méthodes convergent à la même vitesse. Chaque calcul de gradient de la méthode stochastique est très simple, mais il faut plus d'itérations.

Voici les points des premières itérations correspondant au tableau ci-dessus.

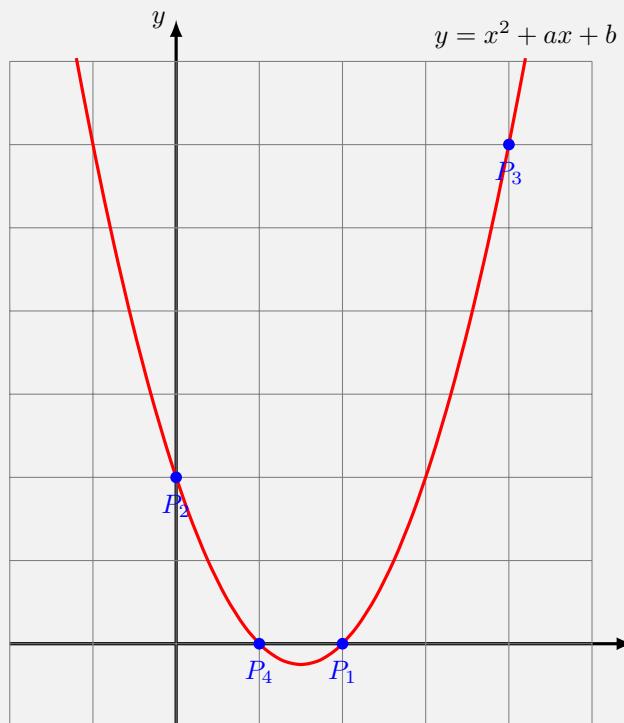
Descente de gradient classique



Descente de gradient stochastique



Conclusion : sur cet exemple les points sont exactement sur la parabole d'équation $y = x^2 + ax + b$ avec $a = -3$ et $b = 2$. Bien sûr cette méthode est encore plus intéressante lorsqu'il s'agit de trouver une parabole qui ne contient pas l'ensemble des points donnés.



Terminons par des remarques plus techniques. Tout d'abord, la formule précise de la descente de gradient stochastique est :

$$P_{k+1} = P_k - \delta \overrightarrow{\text{grad}} E_{(k\%N)+1}(P_k)$$

où $k\%N$ est « k modulo N ».

La descente de gradient stochastique est une méthode qui peut être plus efficace :

- Tout d'abord elle n'utilise qu'une donnée à la fois et évite ainsi les problèmes de mémoire de la descente classique pour laquelle il faut manipuler toutes les données à chaque itération.

- Toujours dans le cas où l'on a beaucoup de données, la descente de gradient stochastique peut converger en deux ou trois passages sur l'ensemble des données, alors que la descente classique nécessite toujours plusieurs itérations (voir la section 3 plus loin).
- Avec la méthode stochastique, on calcule des gradients en des points qui sont plus proches du minimum. Attention cependant, certains petits pas peuvent aller dans la mauvaise direction.
- Le caractère aléatoire de ces petits pas est parfois un avantage, par exemple pour s'échapper d'un point-selle.

2. Différentes fonctions d'erreurs

Il existe différentes formules pour calculer l'erreur entre la sortie attendue y_i et la sortie produite $F(x_i)$.

On considère une série de valeurs y_i , $i = 1, \dots, N$ (fournie par observations ou expérimentations) qui sont approchées par des valeurs $F(x_i)$ produites par une formule issue d'un réseau de neurones par exemple. Le but est d'obtenir $F(x_i)$ le plus proche possible de y_i , pour tout $i = 1, \dots, N$. Pour savoir si l'objectif est atteint, on mesure l'écart entre ces valeurs.

Erreur quadratique moyenne.

$$E = \frac{1}{N} \sum_{i=1}^N (y_i - F(x_i))^2.$$

C'est la formule la plus classique (en anglais *minimal squared error* ou *mse*). Bien entendu $E \geq 0$ quels que soient les $F(x_i)$ et $E = 0$ si et seulement si $y_i = F(x_i)$ pour tous les $i = 1, \dots, N$. C'est presque la formule que l'on a utilisée pour la régression linéaire (il n'y avait pas le facteur $\frac{1}{N}$).

Erreur absolue moyenne.

$$E = \frac{1}{N} \sum_{i=1}^N |y_i - F(x_i)|.$$

C'est une formule plus naturelle, mais moins agréable à manipuler à cause de la valeur absolue.

Noter que pour ces deux formules, l'erreur globale E est la moyenne d'erreurs locales $E_i = (y_i - F(x_i))^2$ (ou bien $E_i = |y_i - F(x_i)|$). Les erreurs locales sont indépendantes les unes des autres, ce qui est la base de la descente de gradient stochastique. (Une formule d'erreur du type $E = y_1 y_2 - F(x_1)F(x_2)$ ne permettrait pas la descente stochastique.)

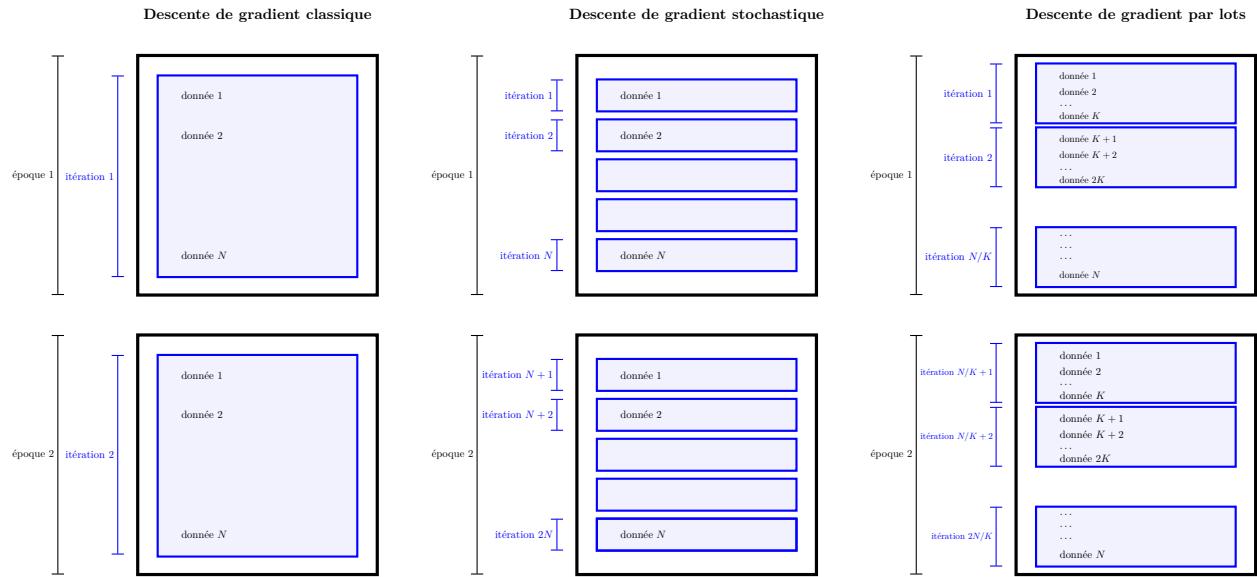
Il existe d'autres formules d'erreur, en particulier si la sortie attendue est du type 0 ou 1 ou bien si la sortie produite est une probabilité $0 \leq p \leq 1$.

3. Descente par lots

Il existe une méthode intermédiaire entre la descente de gradient classique (qui tient compte de toutes les données à chaque itération) et la descente de gradient stochastique (qui n'utilise qu'une seule donnée à chaque itération).

La descente de gradient par **lots** (ou **mini-lots**, *mini-batch*) est une méthode intermédiaire : on divise les données par paquets de taille K . Pour chaque paquet (appelé « lot »), on calcule un gradient et on effectue une itération.

Au bout de N/K itérations, on a parcouru tout le jeu de données : cela s'appelle une **époque**.



La formule est donc

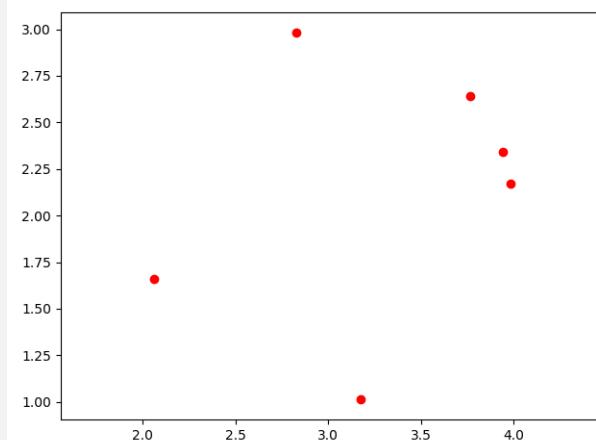
$$P_{k+1} = P_k - \delta \overrightarrow{\text{grad}} (E_{j_0+1} + E_{j_0+2} + \dots + E_{j_0+K})(P_k).$$

Pour P_{k+2} , on repart de P_{k+1} et on utilise le gradient de la fonction $E_{j_0+K+1} + E_{j_0+K+2} + \dots + E_{j_0+2K}$.

- Pour $K = 1$, c'est exactement la descente de gradient stochastique. Pour $K = N$, c'est la descente de gradient classique.
- Cette méthode combine le meilleur des deux mondes : la taille des données utilisées à chaque itération peut être adaptée à la mémoire et le fait de travailler par lots évite les pas erratiques de la descente stochastique pure.
- On peut par exemple choisir $2 \leq K \leq 32$ et profiter du calcul parallèle en calculant $\overrightarrow{\text{grad}} (E_1 + \dots + E_K)$, par le calcul de chacun des $\overrightarrow{\text{grad}} E_i$ sur K processeurs, puis en additionnant les résultats.
- Il est d'usage de mélanger au hasard les données (X_i, y_i) avant chaque époque.

Exemple 3.1

Voyons un exemple d'interpolation circulaire. Les 6 points ci-dessous sont sur un cercle. Comment déterminer son centre et son rayon ?



Pour des points (x_i, y_i) , $i = 1, \dots, N$, on mesure la distance globale par rapport au cercle \mathcal{C}

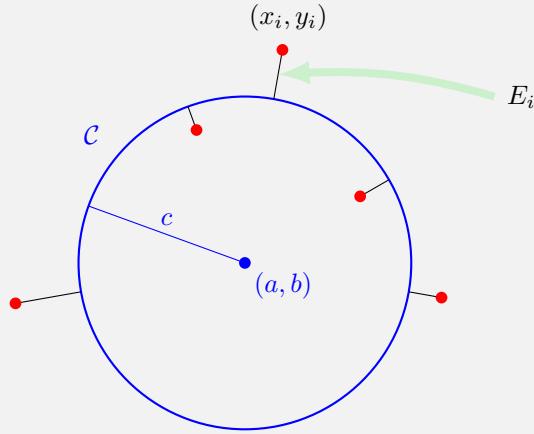
de centre (a, b) et de rayon c par la formule d'erreur :

$$E(a, b, c) = \sum_{i=1}^N E_i(a, b, c)$$

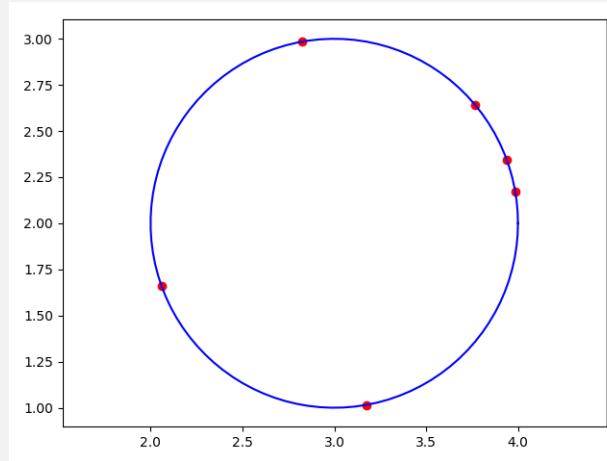
où

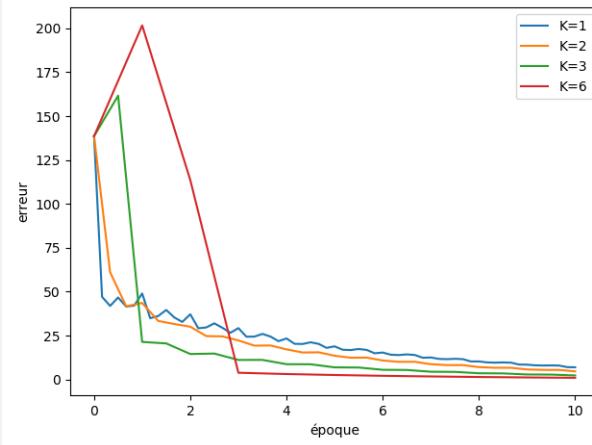
$$E_i(a, b, c) = ((x_i - a)^2 + (y_i - b)^2 - c^2)^2.$$

En effet, E_i mesure en quelque sorte la distance entre le point (x_i, y_i) et le cercle \mathcal{C} de centre (a, b) et de rayon c . Donc $E_i = 0$ si et seulement si $(x_i, y_i) \in \mathcal{C}$, sinon $E_i > 0$.



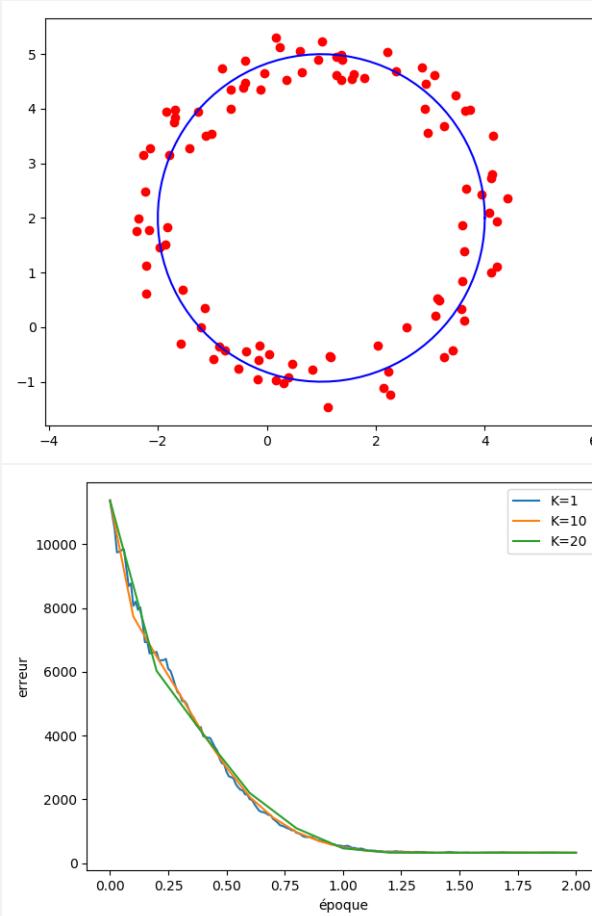
Dans notre exemple, les $N = 6$ points sont exactement situés sur le cercle de centre (a, b) et de rayon c . En appliquant la descente de gradient par lots avec $\delta = 0.01$ et $(a_0, b_0, c_0) = (1, 1, 2)$, on trouve $(a, b) = (3, 2)$ et $c = 1$. Ci-dessous, à droite, nous avons représenté les valeurs de l'erreur totale E (qui tend vers 0) en fonction du nombre d'époques et ceci pour différentes tailles du lot : $K = 1$ (descente stochastique), $K = 2$, $K = 3$ et $K = 6$ (descente classique).





On remarque qu'au bout de 10 époques la valeur de l'erreur est à peu près la même quelle que soit la taille K de l'échantillon. Par contre, l'évolution au départ est différente. Par exemple pour $K = 1$, l'erreur fluctue à la hausse ou à la baisse à chaque itération.

Cette méthode présente bien sûr davantage d'intérêt quand les points ne sont pas exactement sur un cercle. Il s'agit alors de trouver le meilleur cercle qui convient, c'est-à-dire de trouver le minimum (cette fois non nul) de E . Voici un exemple de $N = 100$ points tirés au hasard autour du cercle de centre $(a, b) = (1, 2)$ et de rayon $c = 3$. La descente de gradient est appliquée avec $\delta = 0.001$ et $(a_0, b_0, c_0) = (1, 1, 1)$ pour des lots de différentes tailles $K = 1$, $K = 10$ et $K = 20$. On remarque que deux époques suffisent pour avoir convergence et que plus la taille K de l'échantillon est grande plus la convergence est régulière vers le minimum.

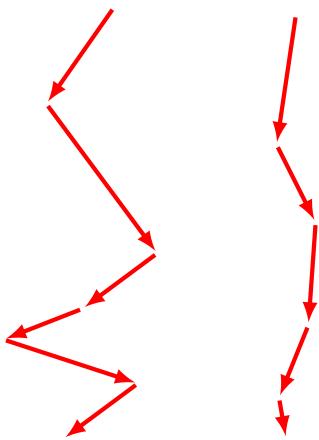


IX- Accélérations

Le choix du pas δ n'est pas la seule amélioration possible de la méthode du gradient, nous allons voir comment la modifier à l'aide du « moment ». Commençons par revenir à l'analogie de la descente du gradient classique qui correspond à une goutte d'eau qui descend une montagne : la goutte emprunte le chemin qui suit la courbe de plus grande pente, quitte à serpenter et osciller lors de la descente. Imaginons que l'on lance maintenant une balle assez lourde du haut de la même montagne. Cette balle va suivre, comme la goutte d'eau, le chemin de la plus forte pente, mais une fois lancée elle va acquérir de l'inertie, appelée **moment**, qui va atténuer ses changements de direction. Ainsi la balle ne s'embarrasse pas des petits aléas du terrain et dévale la pente plus rapidement que la goutte d'eau.

Nos petits aléas de terrain à nous viennent du fait que l'on ne calcule pas exactement le gradient de la fonction d'erreur en utilisant tout le jeu de données à chaque fois, mais seulement un échantillon. Cela peut conduire à certains gradients mal orientés. L'inertie de la balle est en quelque sorte la mémoire de la trajectoire passée qui corrige les mouvements erratiques.

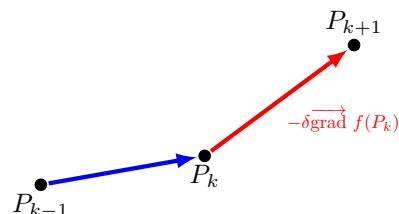
Sur la figure de gauche, la descente classique (la goutte d'eau), sur la figure de droite, la descente de gradient avec le moment (la balle).



1. Moment

Rappelons la formule de la descente de gradient classique :

$$P_{k+1} = P_k - \delta \overrightarrow{\text{grad}} f(P_k).$$



Considérons nos points comme une particule qui voyage au cours du temps. Alors le vecteur $\overrightarrow{P_{k-1}P_k}$ correspond à la vitesse de cette particule et est appelé le **moment** au point P_k .

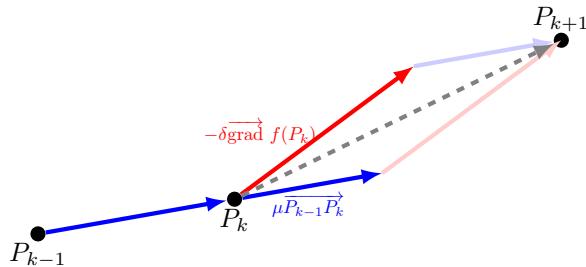
La formule de la descente de gradient avec moment est :

$$P_{k+1} = P_k + \mu \overrightarrow{P_{k-1}P_k} - \delta \overrightarrow{\text{grad}} f(P_k).$$

où $\mu, \delta \in \mathbb{R}$. Cette formule peut être définie pour $k = 0$ si on suppose que la particule est immobile au départ, c'est-à-dire en posant $\overrightarrow{P_{-1}P_0} = \vec{0}$.

On peut prendre par exemple $\mu \in [0.5, 0.9]$ et $\delta = 0.01$.

Schématiquement, au point P_k nous avons deux vecteurs : un qui provient du moment $\mu \overrightarrow{P_{k-1}P_k}$ (la mémoire du passé) et un qui provient du gradient $-\delta \overrightarrow{\text{grad}} f(P_k)$ (qui projette vers l'avenir). La somme permet de calculer le point suivant P_{k+1} .



2. Nesterov

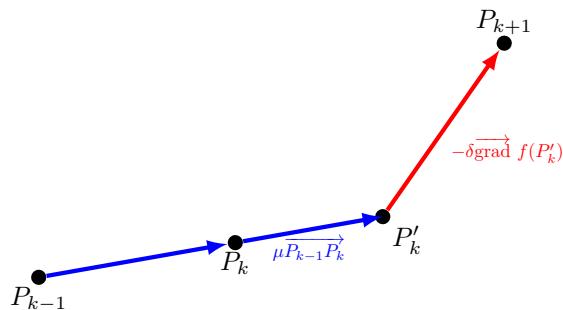
Dans la méthode précédente, le moment et le gradient sont calculés au même point P_k . La méthode de Nesterov est une variante de cette méthode. Elle consiste à appliquer d'abord le moment, pour obtenir un point P'_k , puis de calculer le gradient en ce point (et non en P_k).

La formule est donc

$$P_{k+1} = P_k + \mu \overrightarrow{P_{k-1}P_k} - \delta \overrightarrow{\text{grad}} f(P_k + \mu \overrightarrow{P_{k-1}P_k}).$$

Autrement dit, si on note P'_k le point $P_k + \mu \overrightarrow{P_{k-1}P_k}$ alors

$$P_{k+1} = P'_k - \delta \overrightarrow{\text{grad}} f(P'_k).$$



C'est un petit avantage par rapport à la méthode du moment puisqu'on calcule le gradient au point P'_k qui est censé être plus près de la solution P_{\min} que P_k .

3. Vocabulaire

Terminons par un petit résumé du vocabulaire avec sa traduction en anglais :

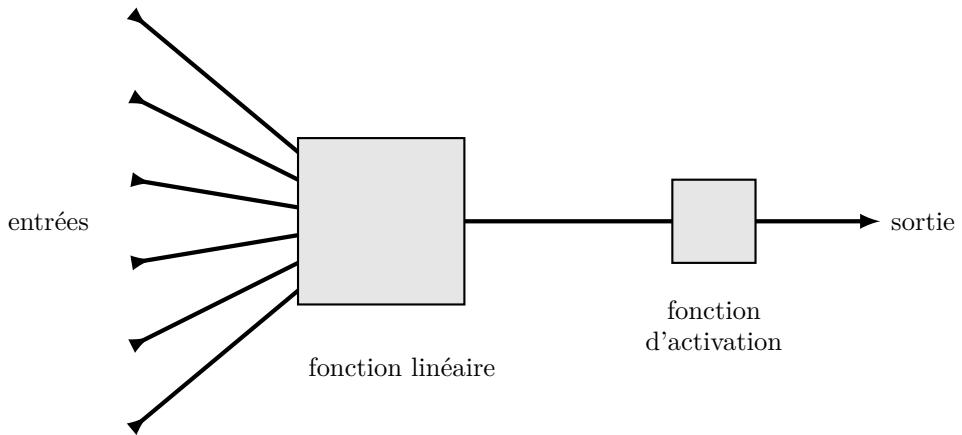
- descente de gradient (classique), *(batch) gradient descent*,
- descente de gradient stochastique, *sgd* pour *stochastic gradient descent*,
- descente de gradient par lots, *mini-batch gradient descent*,
- pas δ , *learning rate*,
- erreur quadratique moyenne, *mse* pour *minimal squared error*,
- moment, *momentum*,
- époque, *epoch*.

Réseaux de neurones

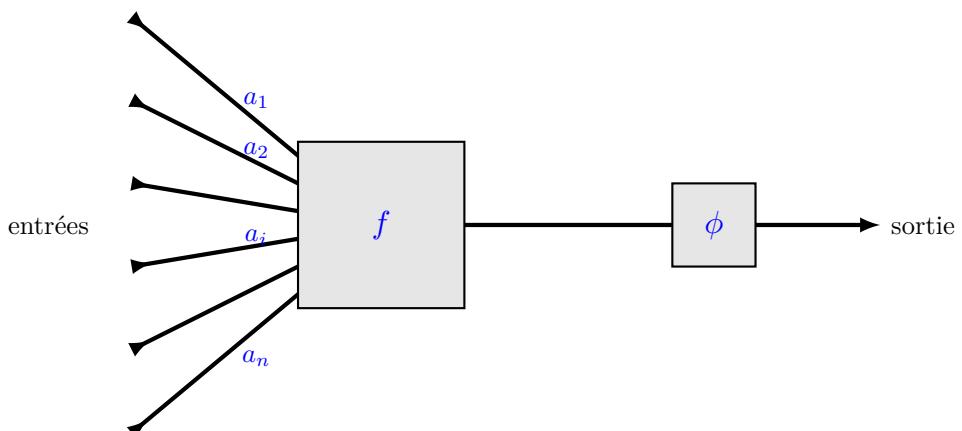
I- Perceptron

1. Perceptron linéaire

Le principe du perceptron linéaire est de prendre des valeurs en entrées, de faire un calcul simple et de renvoyer une valeur en sortie. Les calculs dépendent de paramètres propres à chaque perceptron.



Le calcul effectué par un perceptron se décompose en deux phases : un calcul par une fonction linéaire f , suivi d'une fonction d'activation ϕ .



Détaillons chaque phase.

- **Partie linéaire.** Le perceptron est d'abord muni de **poids** a_1, \dots, a_n qui déterminent une fonction linéaire

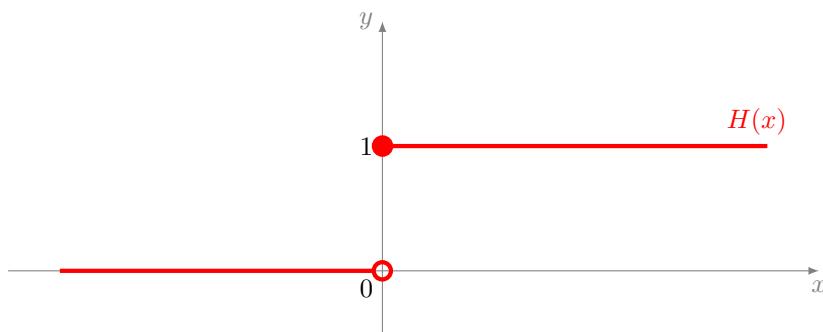
$$f(x_1, \dots, x_n) = a_1x_1 + a_2x_2 + \dots + a_nx_n.$$

- **Fonction d'activation.** La valeur renvoyée par la fonction linéaire f est ensuite composée par une fonction d'activation ϕ .

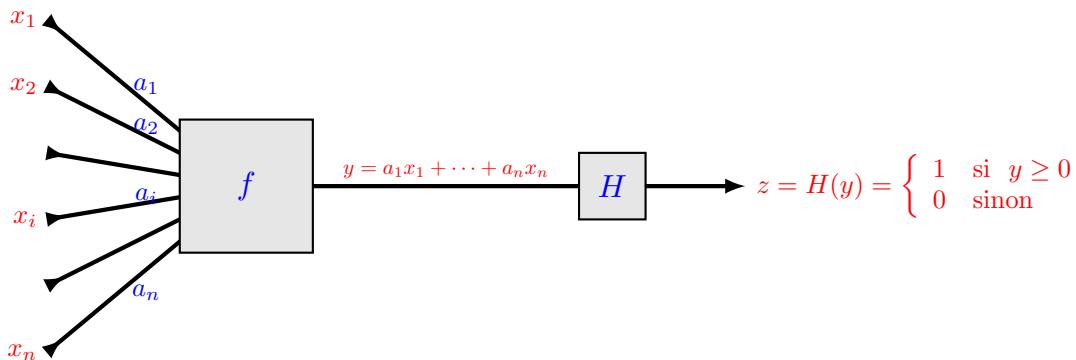
- **Sortie.** La valeur de sortie est donc $\phi(a_1x_1 + a_2x_2 + \dots + a_nx_n)$.

Dans ce chapitre, la fonction d'activation sera (presque) toujours la fonction marche de Heaviside :

$$\begin{cases} H(x) = 1 & \text{si } x \geq 0, \\ H(x) = 0 & \text{si } x < 0. \end{cases}$$



Voici ce que fait un perceptron linéaire de poids a_1, \dots, a_n et de fonction d'activation la fonction marche de Heaviside :

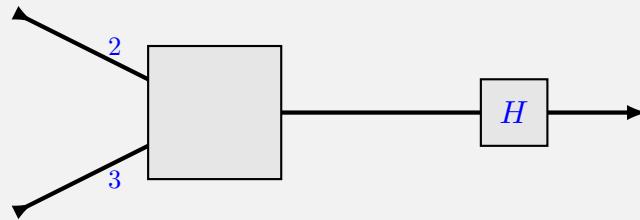


On peut donc définir ce qu'est un perceptron. Un **perceptron linéaire** à n variables et de fonction d'activation la fonction marche de Heaviside est la donnée de n coefficients réels a_1, \dots, a_n auxquels est associée la fonction $F : \mathbb{R} \rightarrow \mathbb{R}$ définie par $F = H \circ f$, c'est-à-dire :

$$\begin{cases} F(x_1, \dots, x_n) = 1 & \text{si } a_1x_1 + a_2x_2 + \dots + a_nx_n \geq 0, \\ F(x_1, \dots, x_n) = 0 & \text{sinon.} \end{cases}$$

Exemple 1.1

Voici un perceptron à deux entrées. Il est défini par les poids $a = 2$ et $b = 3$.



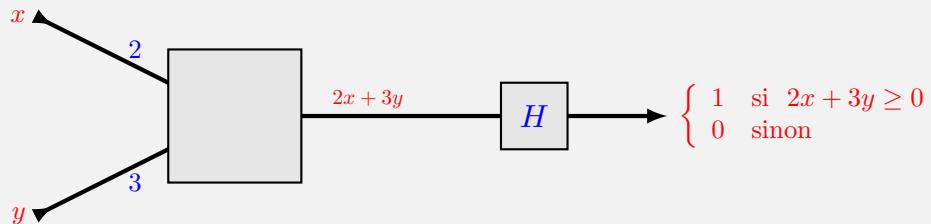
— **Formule.**

Notons x et y les deux réels en entrée. La fonction linéaire f est donc

$$f(x, y) = 2x + 3y.$$

La valeur en sortie est donc :

$$\begin{cases} F(x, y) = 1 & \text{si } 2x + 3y \geq 0 \\ F(x, y) = 0 & \text{sinon.} \end{cases}$$



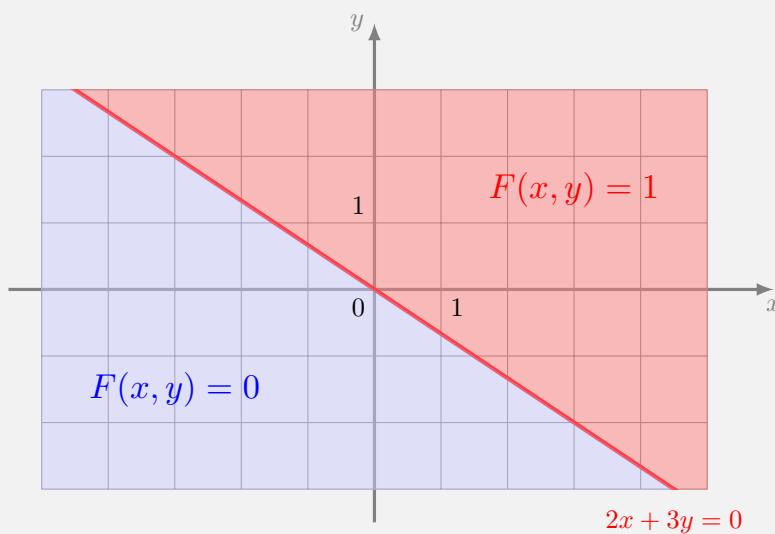
— **Évaluation.** Utilisons ce perceptron comme une fonction. Que renvoie le perceptron pour la valeur d'entrée $(x, y) = (4, -1)$? On calcule $f(x, y) = 2x + 3y = 5$. Comme $f(x, y) \geq 0$, alors la valeur de sortie est donc $F(x, y) = 1$.

Recommençons avec $(x, y) = (-3, 1)$. Cette fois $f(x, y) = -3 < 0$ donc $F(x, y) = 0$.

L'entrée $(x, y) = (6, -4)$ est « à la limite » car $f(x, y) = 0$ (0 est l'abscisse critique pour la fonction marche de Heaviside). On a $F(x, y) = 1$.

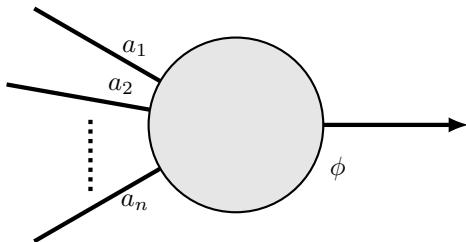
— **Valeurs de la fonction.**

La fonction F prend seulement deux valeurs : 0 ou 1 . La frontière correspond aux points (x, y) tels que $f(x, y) = 0$, c'est-à-dire à la droite $2x + 3y = 0$. Pour les points au-dessus de la droite (ou sur la droite) la fonction F prend la valeur 1 ; pour les points en-dessous de la droite, la fonction F vaut 0 .

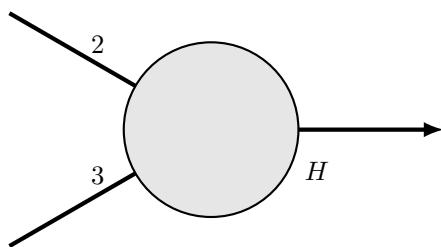


Notation.

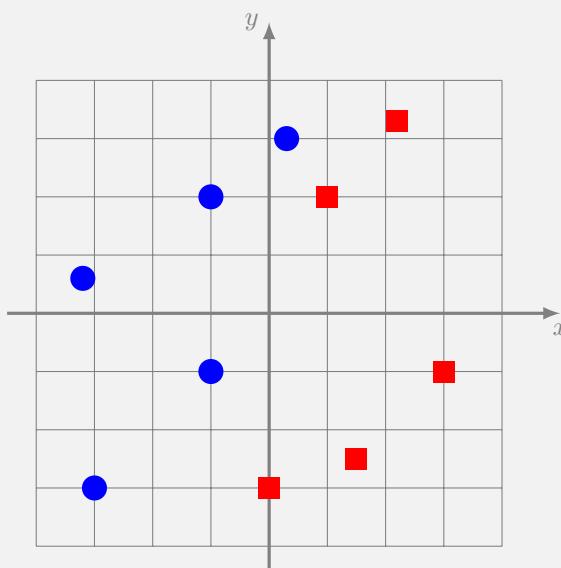
Nous représentons un perceptron par une forme plus condensée : sous la forme d'un **neurone**, avec des poids sur les arêtes d'entrées. Nous précisons en indice la fonction d'activation utilisée ϕ . Si le contexte est clair cette mention est omise.



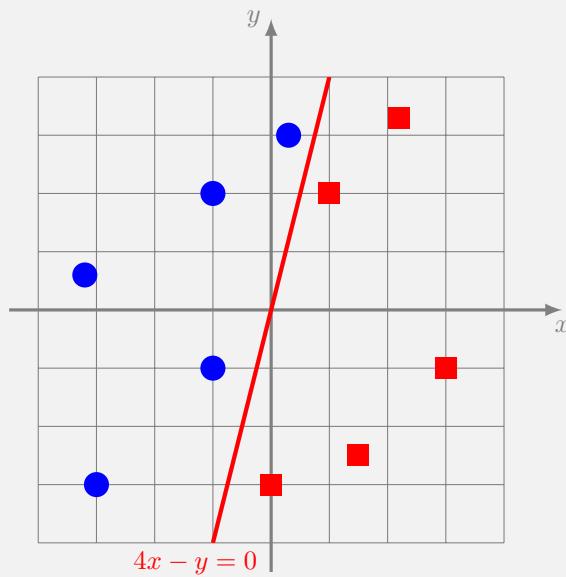
Voici le neurone à deux variables de l'exemple précédent.

**Exemple 1.2**

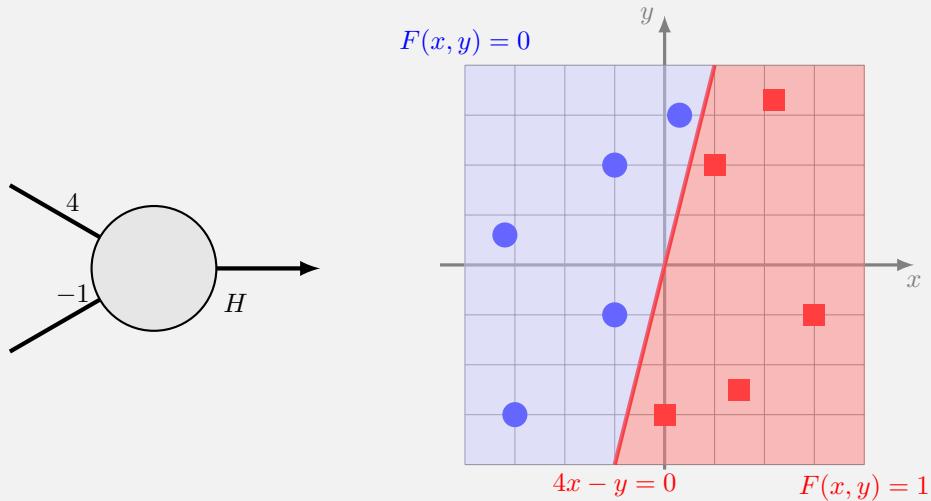
Voici deux catégories de points : des ronds bleus et des carrés rouges. Comment trouver un perceptron qui les sépare ?

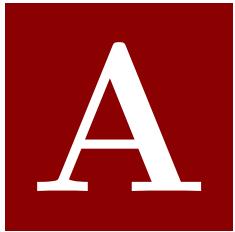


Il s'agit donc de trouver les deux poids a et b d'un perceptron, dont la fonction associée F vérifie $F(x, y) = 1$ pour les coordonnées des carrés et $F(x, y) = 0$ pour les ronds.



Trouvons une droite qui les sépare. Par exemple, la droite d'équation $4x - y = 0$ sépare les ronds des carrés. On définit donc le neurone avec les poids $a = 4$ et $b = -1$. Si (x, y) sont les coordonnées d'un carré alors on a bien $F(x, y) = 1$ et pour un rond $F(x, y) = 0$.





Compétences attendues à l'issue de ce cours