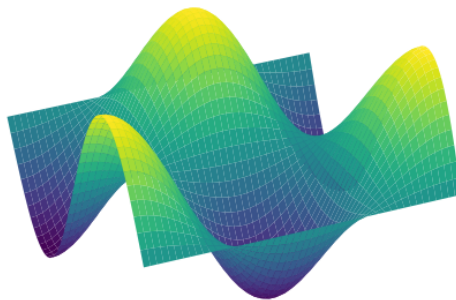


Descente de gradient

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.

1. Descente de gradient classique

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.



1.1. 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.

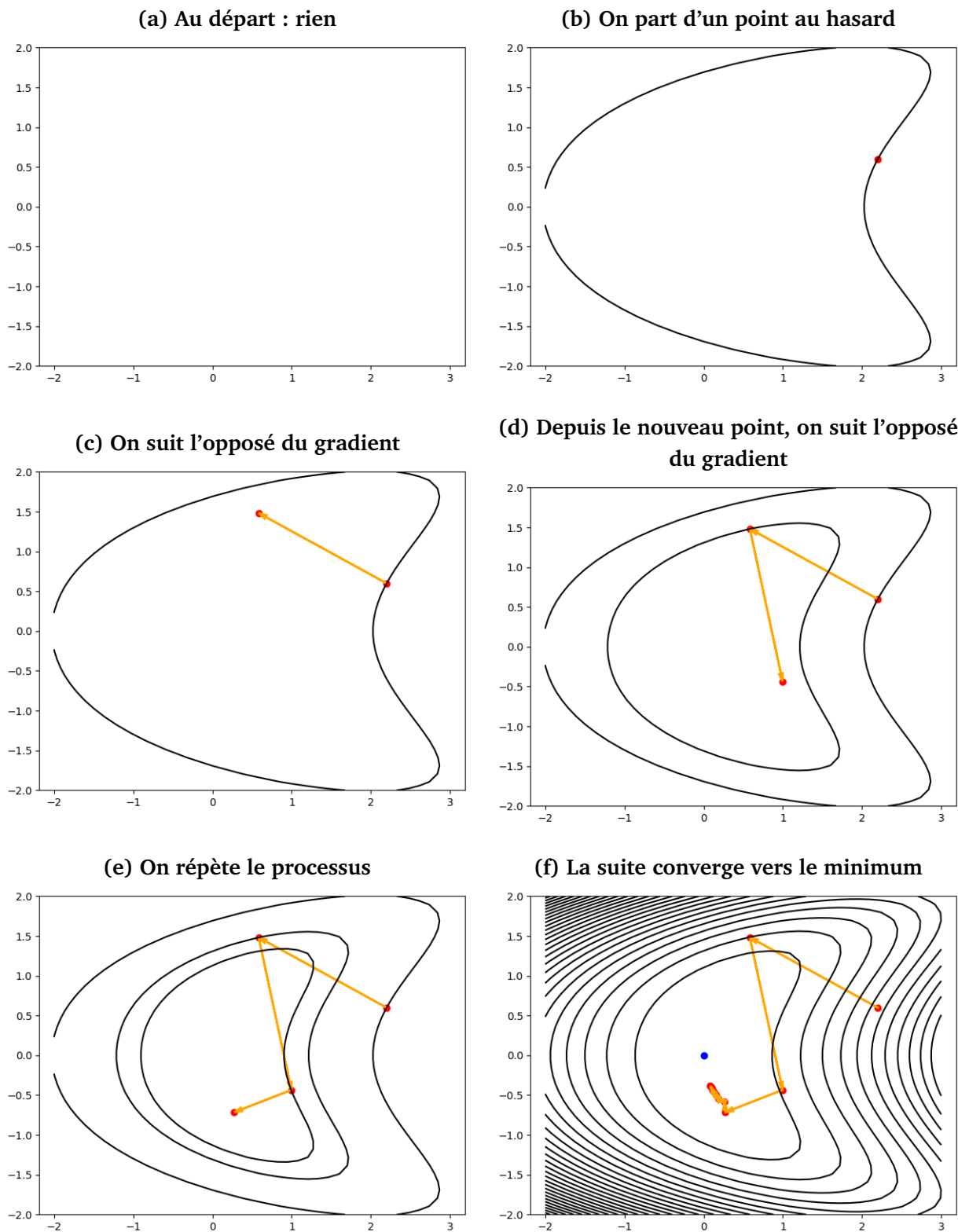
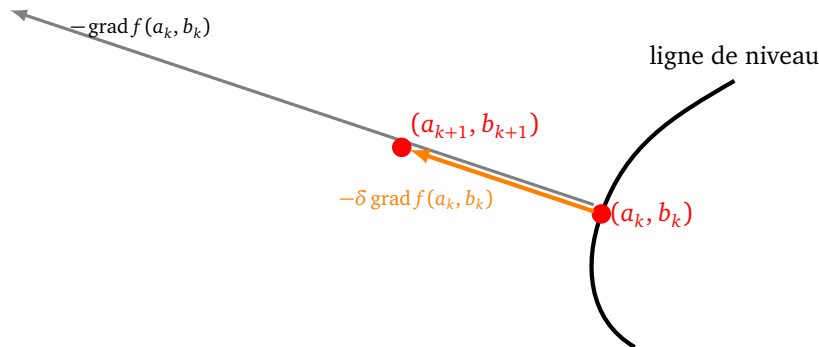


Figure (a). Au départ nous n'avons aucune information globale sur f . La seule opération que l'on s'autorise c'est calculer $\text{grad } f(a, b)$ en certains points.

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 : $-\text{grad } f(a_0, b_0)$. On sait d'une part que la ligne de niveau est orthogonale à ce gradient et surtout que dans la direction de $-\text{grad } f(a_0, b_0)$, 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.

1.2. 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 :

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

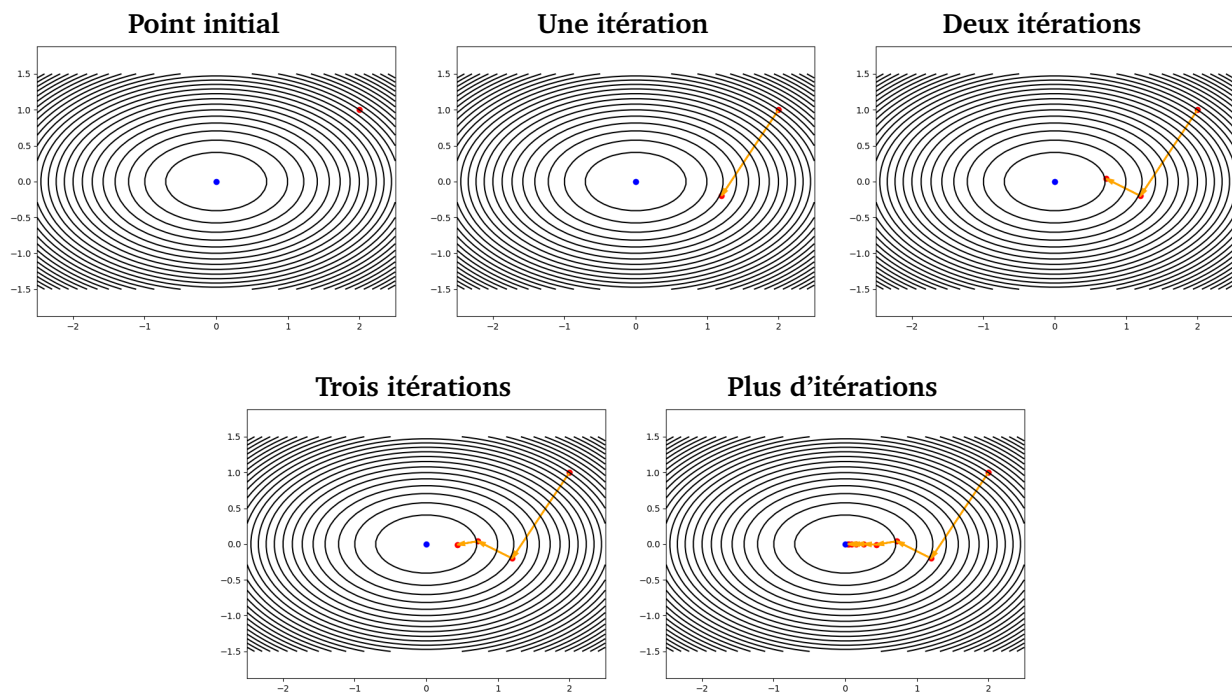
Tout d'abord, on part d'un point $(a_0, b_0) = (2, 1)$ par exemple. Même si nous n'en avons pas besoin pour notre construction, on a $f(a_0, b_0) = 7$. On calcule $\text{grad } f(a_0, b_0) = (4, 6)$. On fixe le facteur $\delta = 0.2$. On se déplace dans la direction opposée à ce gradient :

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

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).

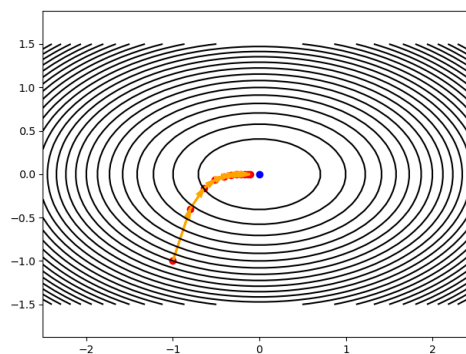
k	(a_k, b_k)	$\text{grad } f(a_k, b_k)$	$f(a_k, b_k)$
0	(2, 1)	(4, 6)	7
1	(1.2, -0.2)	(2.4, -1.20)	1.56
2	(0.72, 0.04)	(1.44, 0.24)	0.523
3	(0.432, -0.008)	(0.864, -0.048)	0.186
4	(0.2592, 0.0016)	(0.5184, 0.0096)	0.067
5	(0.15552, -0.00032)	(0.31104, -0.00192)	0.024
...			
10	(0.012, $1.02 \cdot 10^{-7}$)	(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 $(a_1, b_1) = (-0.8, -0.4)$, $(a_2, b_2) = (-0.64, -0.16)$... La suite converge également vers $(0, 0)$.

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



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.

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 :

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

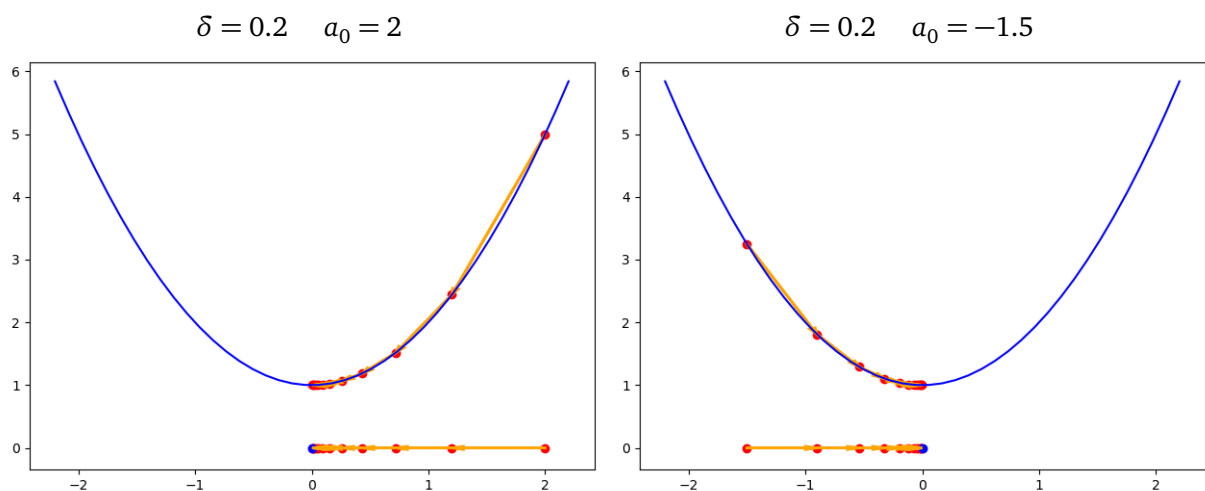
où δ est le pas, choisi assez petit, et $\operatorname{grad} f(a) = f'(a) = 2a$. Autrement dit :

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

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

k	a_k	$f'(a_k) = \operatorname{grad} f(a_k)$	$f(a_k)$
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 $a_{\min} = 0$.



Il faut bien comprendre ce graphique : la suite des points (a_k) (points rouges) se lit sur l'axe des abscisses. Les vecteurs (orange) montrent les itérations. Il est plus facile de comprendre l'algorithme sur le graphe de f (en bleu). Sur ce graphe, on reporte les points $(a_k, f(a_k))$ (en rouge), 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

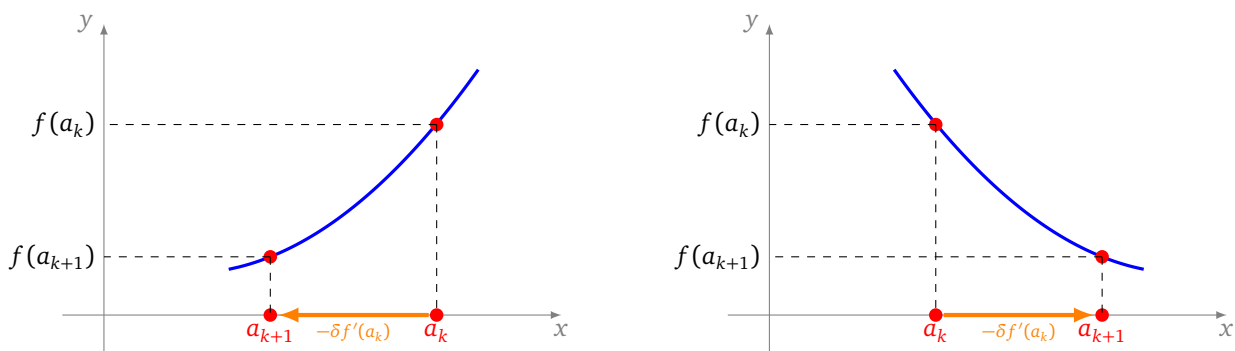
$$a_{k+1} = a_k - \delta f'(a_k) \quad \text{donne} \quad a_{k+1} < a_k.$$

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

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

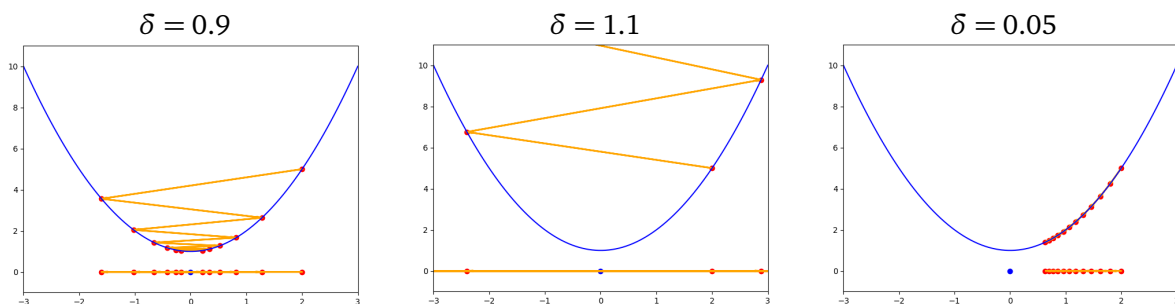
ce qui implique de nouveau $f(a_{k+1}) < f(a_k)$ (car f est décroissante).

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.

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.

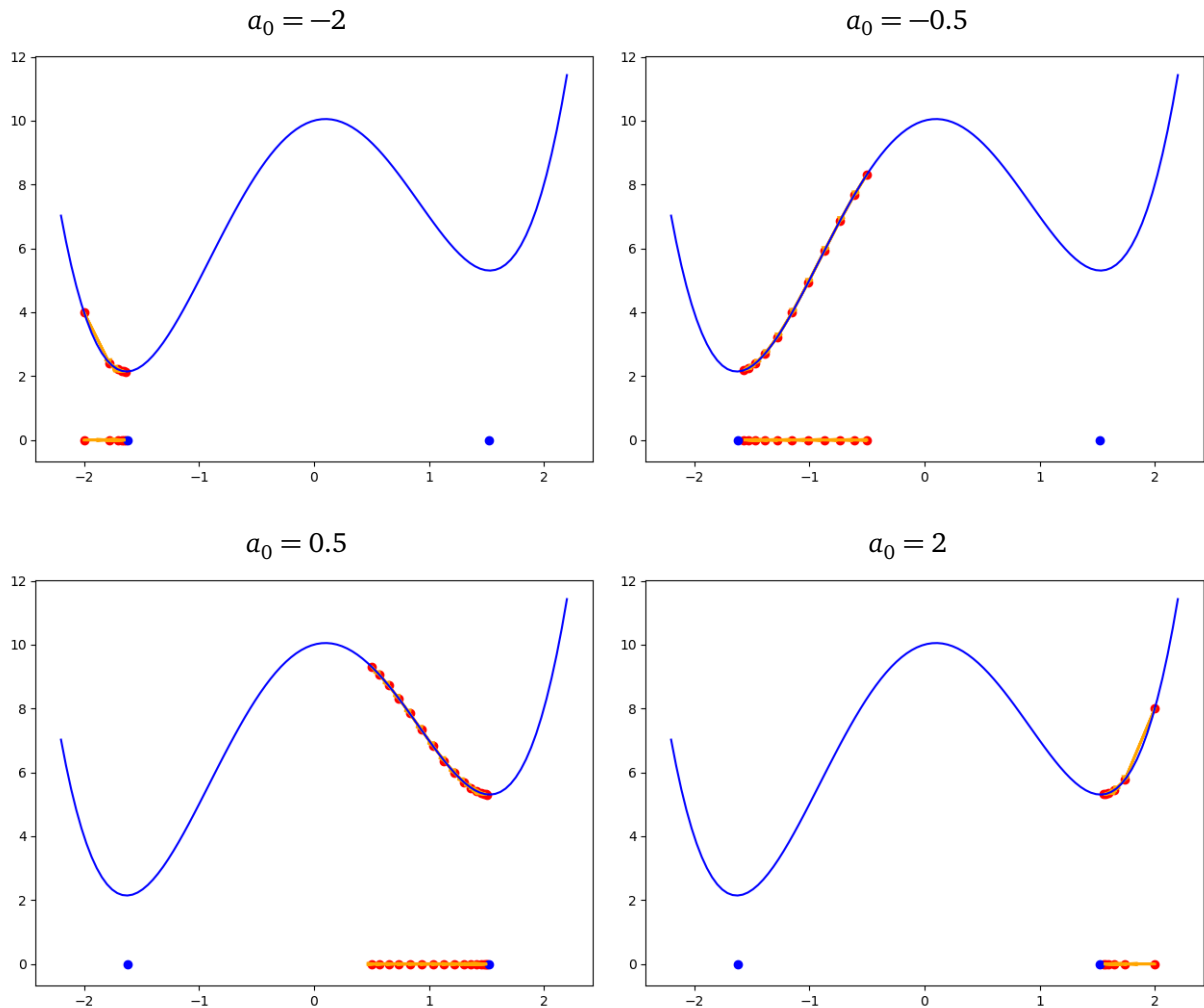
Exemple.

Le choix du point de départ est également important surtout lorsqu'il existe plusieurs minimums locaux.

Soit la fonction f définie par :

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

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



Exemple.

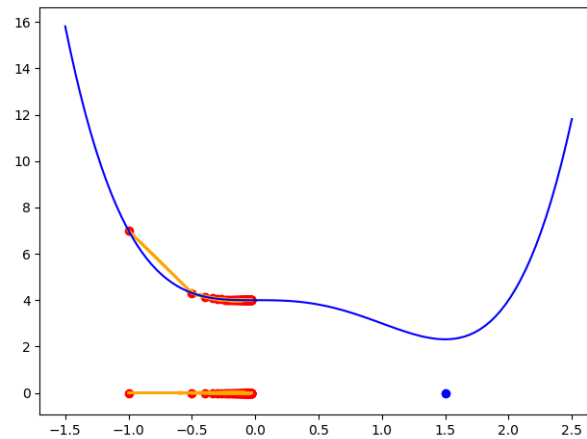
Les points-selles posent également problème.

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.



1.4. 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 $\text{grad } f(P)$.

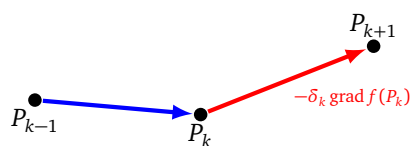
Données.

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

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



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

Remarque.

- 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 $-\text{grad } f(P_k)$ la fonction f décroît.