Introduction aux outils de l'intelligence artificielle
Optimisation et descente de gradient

V. Monbet

[1] Université de Rennes, UFR Mathématiques

Outline

Un problème classique de machine learning : classification

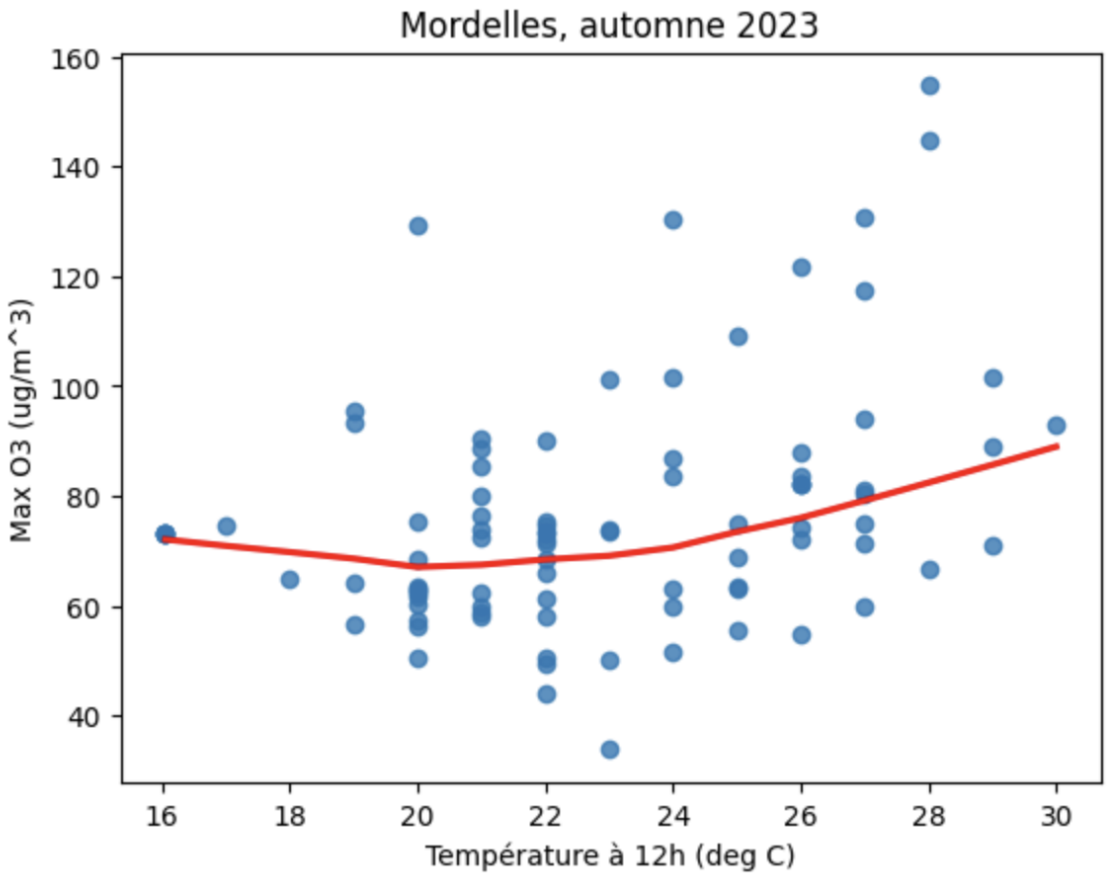
Données-exemples étiquetées :

Identifier :

- **Fabriquer la fonction** : $f($  $) =$ Tomate

$$f : \mathbb{R}^{128\times128} \rightarrow \{\text{Pomme, Poire, Tomate, Vache Chien, Cheval}\}$$

Un autre problème classique de machine learning : régression



Fabriquer la fonction f : T12 $\mapsto$ max03

Apprendre à partir d'exemples

Exemples : données $(x_1, y_1), \cdots, (x_n, y_n)$

- x : image, courbe, variable météo, etc
- y : étiquette (tomate/cheval/...), concentration d'ozone, hauteur de vague, performance sportive, etc

Apprendre : trouver une fonction $\hat{f}$ telle que

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(y_i, f(x_i))$$

où L est une fonction de perte et $\mathcal{F}$ une famille de fonctions.

Les problèmes d'optimisation
sont une des étapes clé du machine learning.

Exemple : Régression linéaire simple

En régression linéaire simple,

- $\mathcal{F}$ est l'ensemble des droites
- $L(y_i, f(x_i)) = \sum_{i=1}^{n} (y_i - f(x_i))^2$ est l'erreur aux moindres carrés.
- Le problème d'optimisation s'écrit encore

$$(\widehat{\beta}_0, \widehat{\beta}_1) \in \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

On remarque que la minimisation en f est remplacée par une minimisation en (β_0, β_1). La famille de fonctions est explicite dans la fonction de perte.

Outline

Problème d'optimisation

On note $L : w \in \Omega \mapsto L(w) \in \mathbb{R}$ la fonction de perte [1] (ou fonction objectif ou fonction de coût).

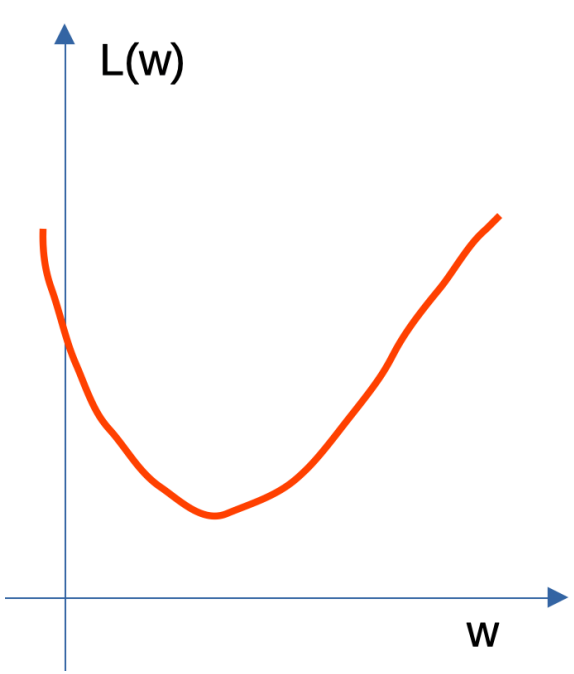
On cherche le point $\hat{w}$ tel que

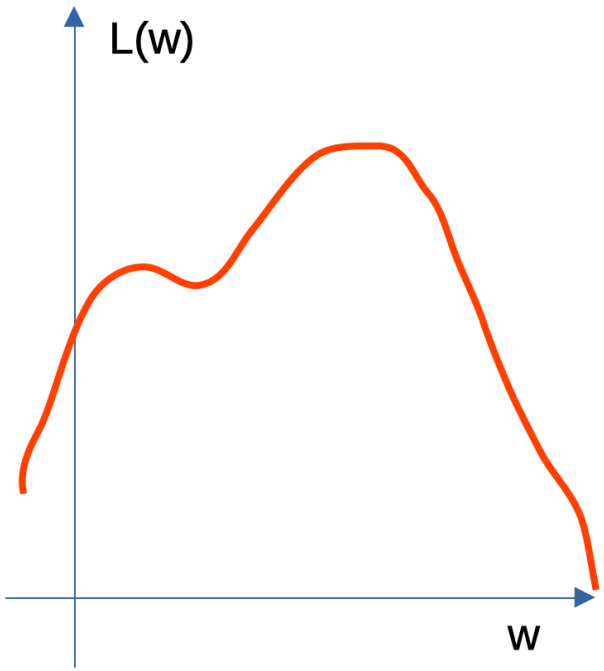
$$L(\hat{w}) = \min_{w \in \Omega} L(w)$$

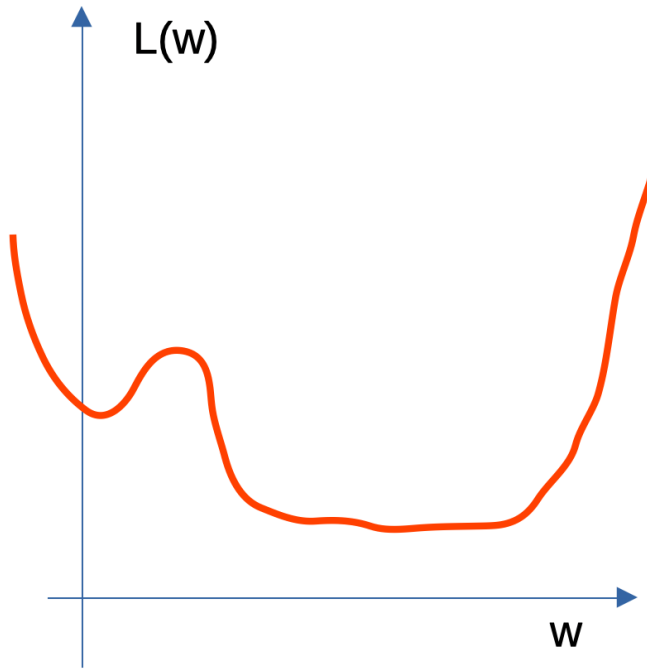
- Sous quelles conditions sur L et Ω ce problème admet-il une solution ?
- La solution est-elle unique ?
- La solution est-elle calculable ?
- Si on ne peut pas calculer la solution peut-on l'approcher ? Avec quel(s) algorithme(s) ?

1. pour simplifier les notations, on oublie pour le moment la dépendance en x_i et y_i

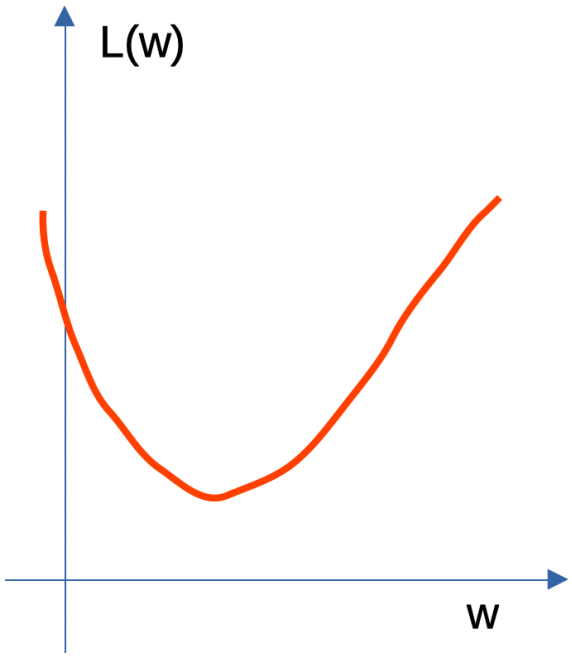
Commenter ces 3 exemples en ayant en tête les questions précédentes.



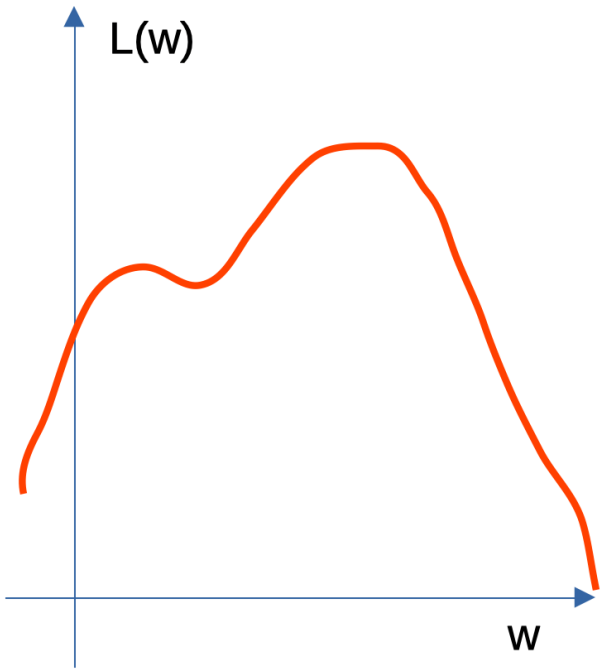




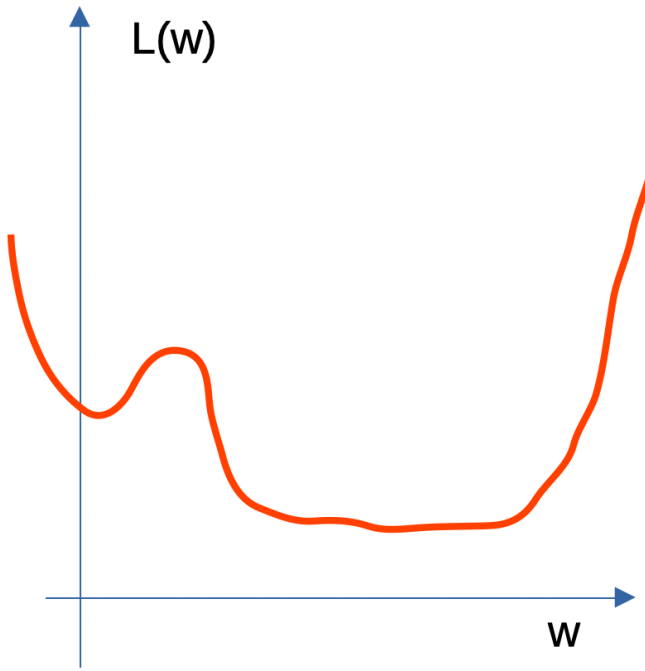
Commenter ces 3 exemples.



L admet un minimum unique
L est convexe



L admet un maximum
Les minimums de *L* sont infinis



L admet une infinité de minima

Exemple de la régression linéaire

Supposons que l'ordonnée à l'origine est fixée et qu'on cherche uniquement la pente β_1 de la droite de régression. On se ramène ainsi à un problème d'optimisation en dimension 1.

$$\min_{\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

On a vu que β_1 est solution du problème si

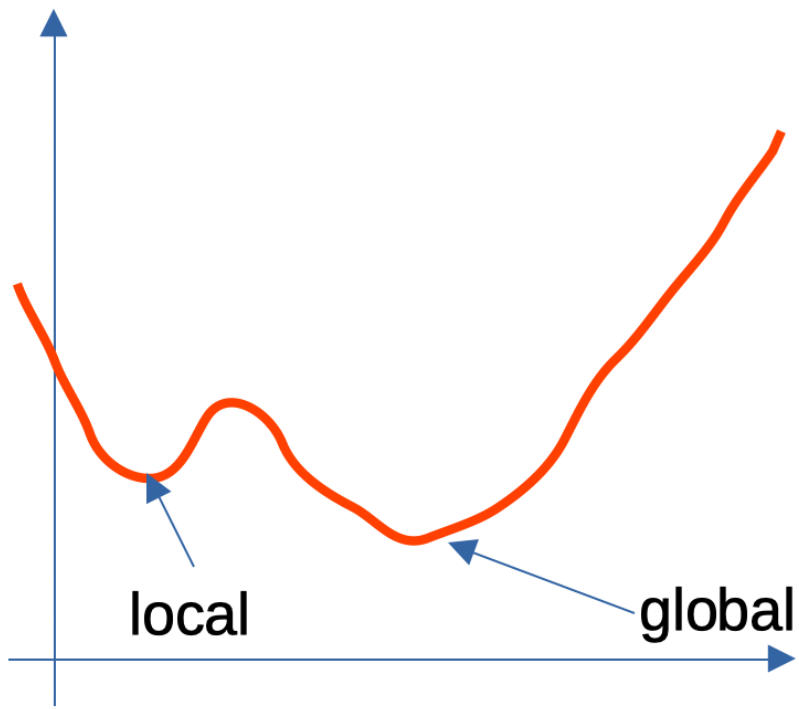
- la dérivée $L'(\beta_1) = 0$;
- la dérivée seconde $L''(\beta_1) > 0$.

Ce résultat se généralise.

Condition nécessaire d'extremum local

Theorem

Condition nécessaire d'extremum local
Soit $L : \mathbb{R} \to \mathbb{R}$ une fonction à valeurs réelles. Si la fonction L admet un extremum local en un point w et si elle est dérivable en ce point, alors $L'(w) = 0$.
Si $L''(w) \geq 0$ alors ce point est un minimum.

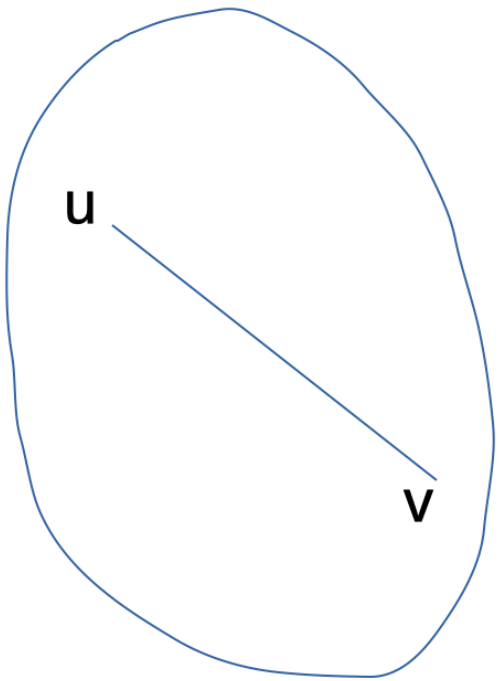


Ensemble convexe

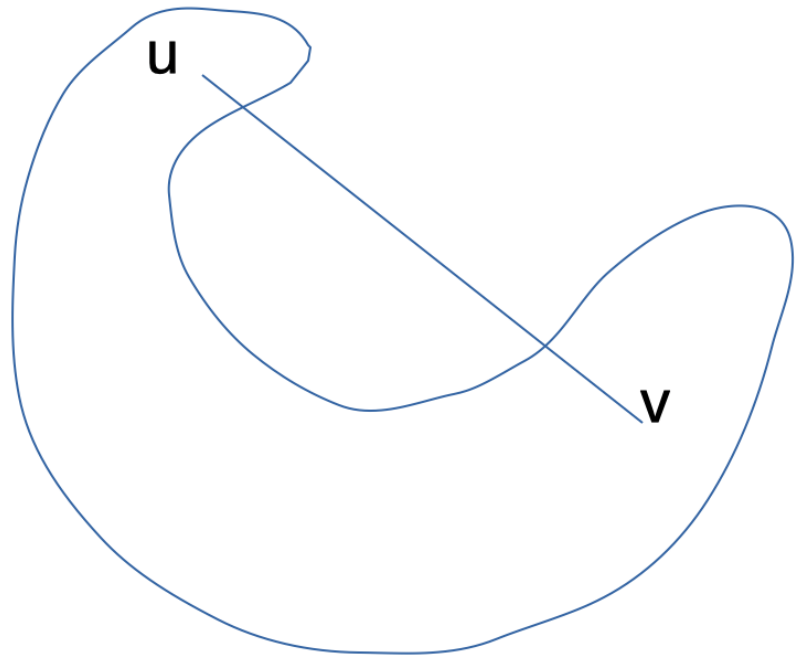
Definition

Ensemble convexe
On dit qu'un ensemble $\mathcal{W} \subset \mathbb{R}^n$ est convexe, si

$$\forall u, v \in \mathcal{W}, \quad \forall \lambda \in [0,1], \quad \lambda u + (1-\lambda)v \in \mathcal{W}$$



$\mathcal{W}$ convexe



$\mathcal{W}$ non convexe

Exemples d'ensembles convexes

- Un sous espace vectoriel est convexe
- Un hyperplan est convexe
- Toute intersection d'ensembles convexes est convexe.
- Un hyper-rectangle est convexe

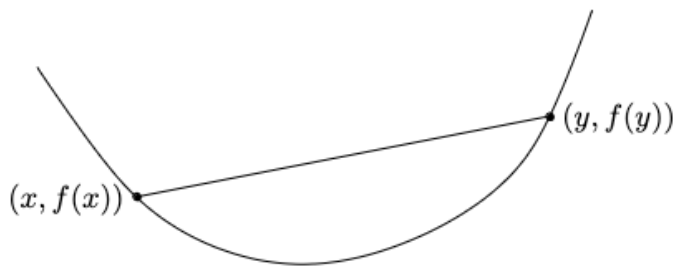
Fonction convexe

Definition

Fonction convexe
Soit $\mathcal{W}$ un ensemble convexe, soit $f : w \in \mathcal{W} \mapsto f(w) \in \mathbb{R}$ une fonction à valeurs réelles,
- f est convexe si et seulement si

$$\forall x, y \in \mathcal{W}, \forall \lambda \in [0,1], \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

- f est convexe si et seulement si pour tout x_0,
l'application qui a $x \in \mathcal{W} \mapsto \frac{f(x)-f(x_0)}{x-x_0}$ est croissante

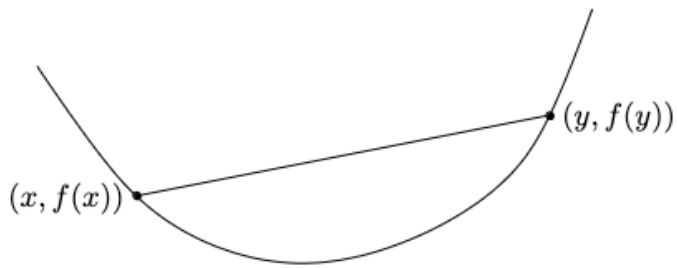


Si f dérivable alors f est convexe ssi sa dérivée est croissante et réciproquement.

A retenir : si f est convexe la corde qui relie $(x, f(x))$ et $(y, f(y))$ est au-dessus de la courbe représentant f pour tout points x et y.

Exemples de fonctions convexes

- Sur $\mathbb{R}$, la fonction $x \mapsto x^2$ est convexe
- Sur $\mathbb{R}$, la fonction $x \mapsto |x|$ est convexe mais pas strictement
- Sur $\mathbb{R}^2$, la fonction $x \mapsto ||x||^2$ est convexe
- Sur $\mathbb{R}^2$, la fonction de perte des moindres carrées de la régression linéaire (simple) est une fonction quadratique, elle est donc convexe.

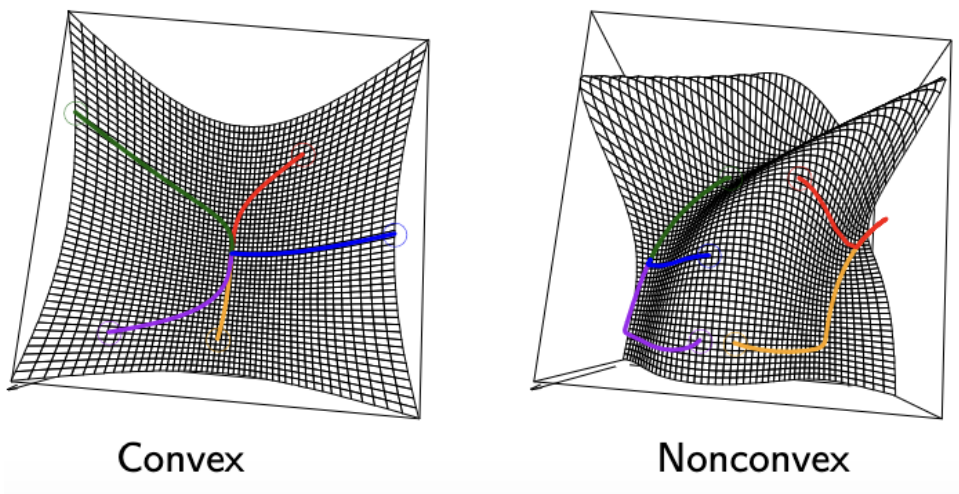


Extremum global

Theorem

Minimum de fonctions convexes
Soit $\mathcal{W}$ un ensemble convexe,

1. *Si une fonction convexe $L : \mathcal{W} \to \mathbb{R}$ admet un minimum local en un point w, elle y admet en fait un minimum global sur $\mathcal{W}$.*
2. *Une fonction $L : \mathcal{W} \to \mathbb{R}$ strictement convexe admet au plus un minimum local qui est en fait un minimum global strict.*
3. *Soit L une fonction convexe définie sur un ouvert convexe $\Omega \subset \mathbb{R}^n$. Alors un point $w \in \Omega$ est un minimum global de L si et seulement si $\nabla L(w) = 0$*



A ce stade du cours, il faut lire $\nabla L(w)$ comme la dérivée de L en w.

Outline

Quand la solution n'a pas de forme analytique

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L(y_i, f(x_i))$$

Si la fonction f appartient à une famille paramétrique, le problème s'écrit

$$\hat{w} = \arg\min_{w \in \Omega} \sum_{i=1}^{n} L(y_i, f(w; x_i))$$

Si la forme de f est complexe [2], l'équation

$$\sum_{i=1}^{n} L'(y_i, f(w, x_i) = 0$$

peut ne pas avoir de solution analytique ou une solution difficile à calculer.

On utilise un algorithme d'optimisation numérique
pour approcher la solution du problème de minimisation.

2. au sens de compliqué

Algorithme de descente de gradient

Soit L une fonction strictement convexe sur un ouvert Ω.

On considère le problème d'optimisation

$$\min_{w \in \Omega} L(w)$$

et on cherche le point $\hat{w} \in \Omega$ qui réalise le minimum.

Idées :

- On part d'un point quelconque w_0 (par exemple $w_0 = 0$)
- On choisit une direction de descente s ie un vecteur s tel que

$$L(w_0 + s) \leq L(w_0)$$

- On fait $w_1 = w_0 + s$ et on recommence.

Choix de s

D'après la formule de Taylor,

$$L(y) \approx L(w_0) + \underbrace{L'(w_0)(y - w_0)}_{\text{approx. lin.}} + \underbrace{\frac{1}{2\alpha}(y - w_0)^2}_{\text{approx. quad.}}$$

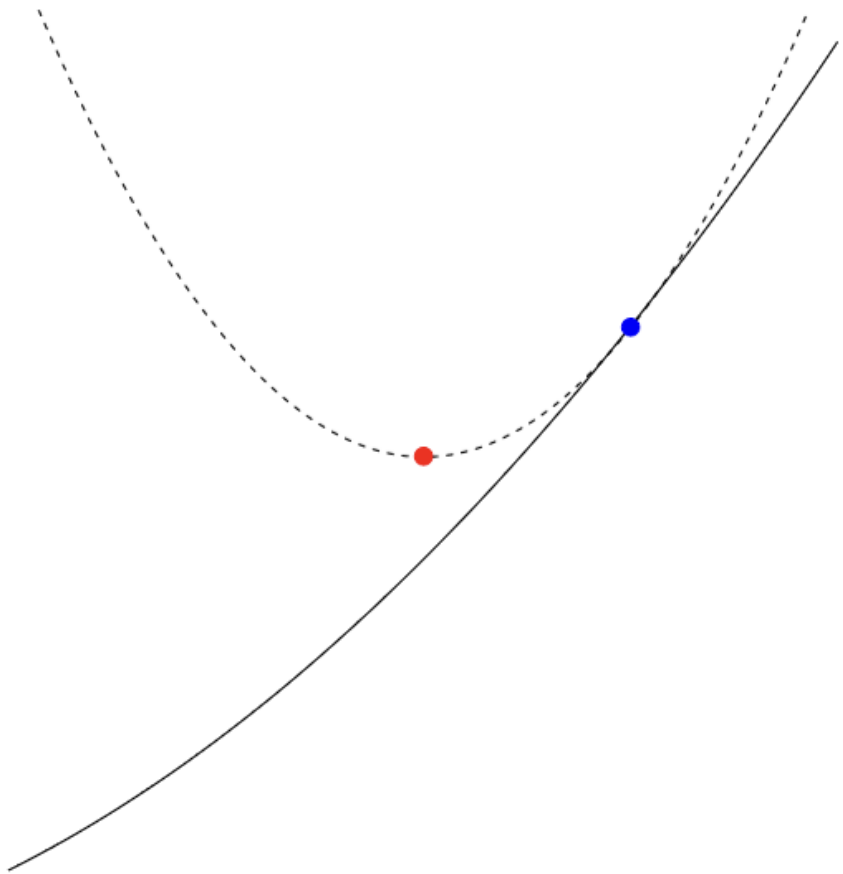
On choisit le $s = (y - w_0)$ tel que y minimise l'approximation. On obtient $s = -\alpha L'(w_0)$

On vérifie

$$L(w_0 - \alpha L'(w_0)) \approx L(w_0) - \alpha \underbrace{L'(w_0)L'(w_0)}_{\geq 0}$$

on a donc bien

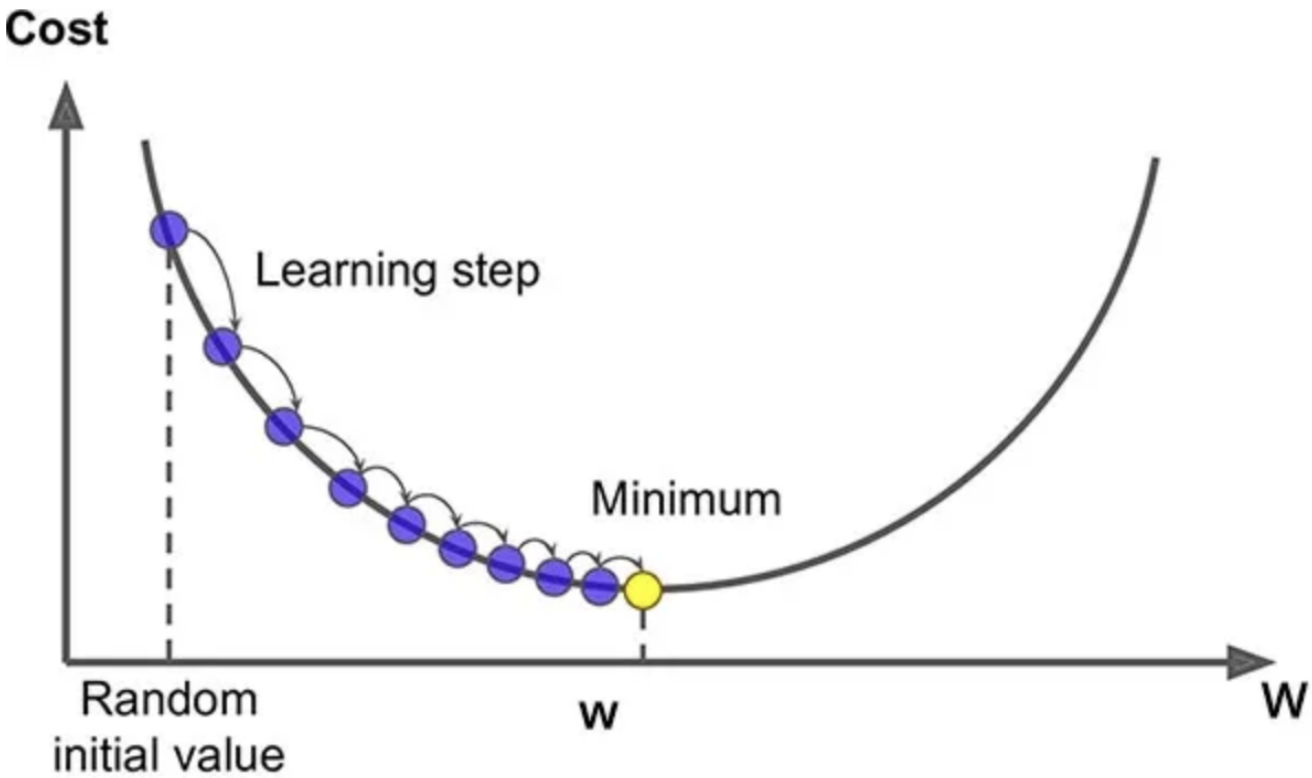
$$L(w_0 - \alpha L'(w_0)) \leq L(w_0)$$



Le point bleu est w_0 et le point rouge est
$w_1 = \arg\min_y L(w_0) + L'(w_0)(y - w_0) + \frac{1}{2\alpha}(y - w_0)^2$

Algorithme de descente de gradient

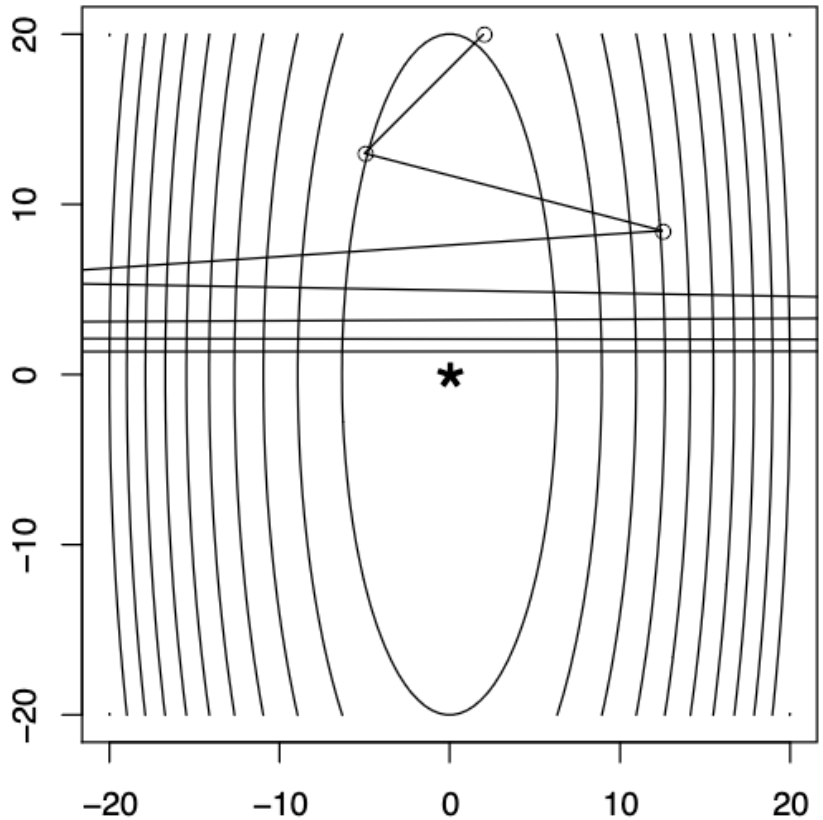
Initialisation : choisir un point $w_0 \in \Omega$
Tant que $L'(w) > \epsilon$ faire
$w \leftarrow w_{k+1} = w_k - \alpha_k L'(w_k)$



On appelle α le taux d'apprentissage (*learning rate*).

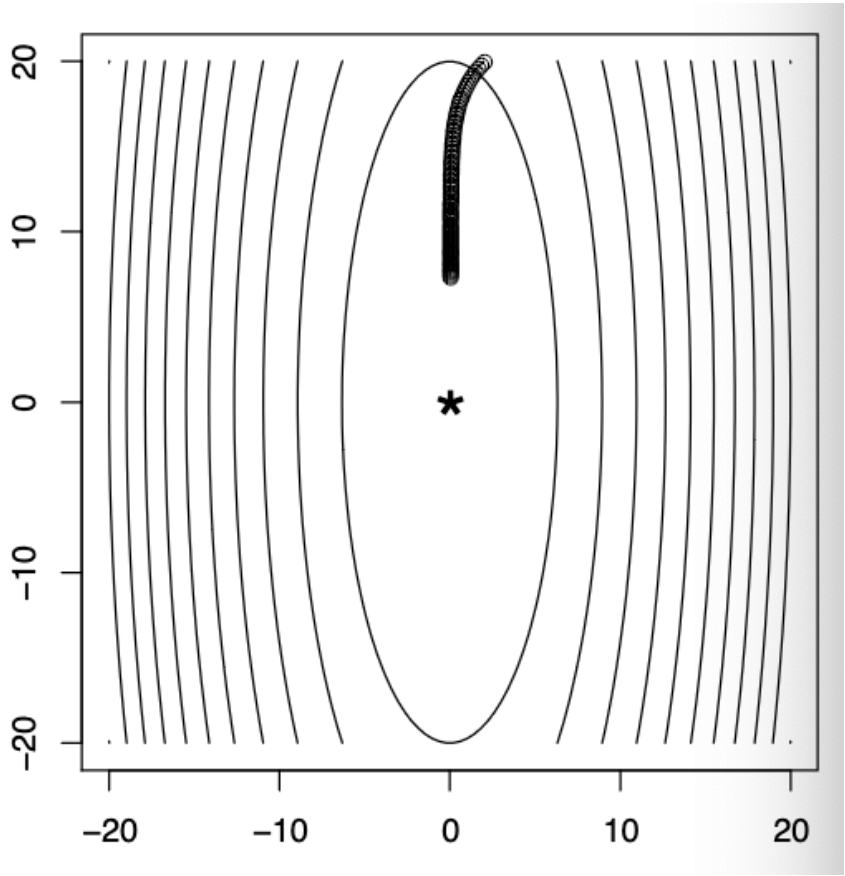
Choix du taux d'apprentissage alpha

On peut choisir $\alpha_k = \alpha$ pour $k = 1, 2, \cdots$, mais l'algorithme de descente du gradient va diverger si α est trop grand.

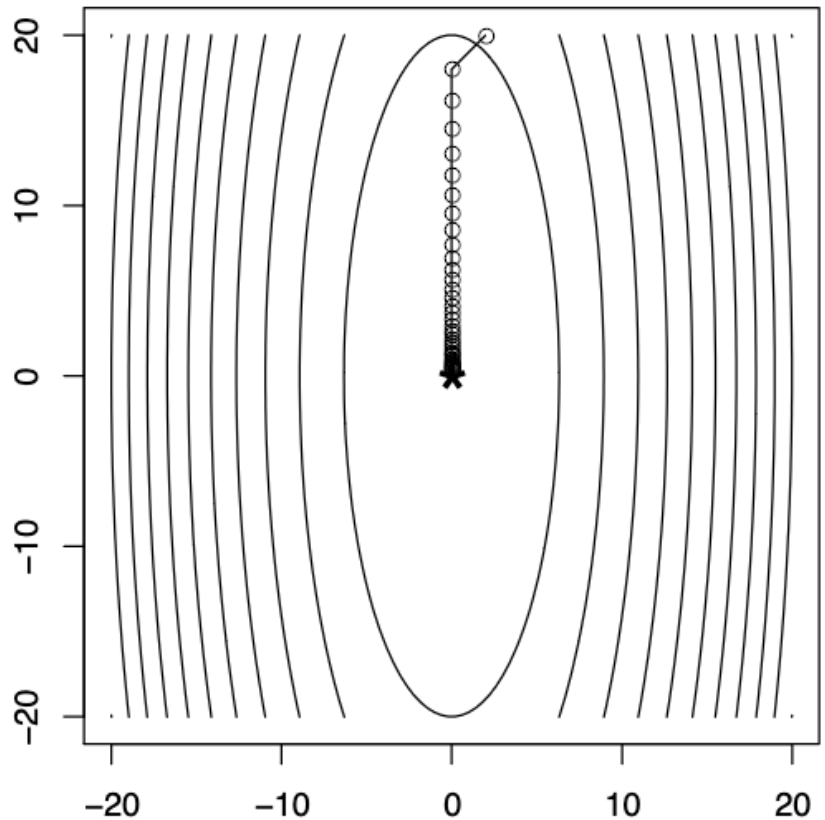


$$f(x) = (10x_1^2 + x_2)/2$$

L'algorithme de descente du gradient va être trop lent si α est trop petit.



L'algorithme de descente du gradient converge bien si α est juste bien.



Recherche linéaire

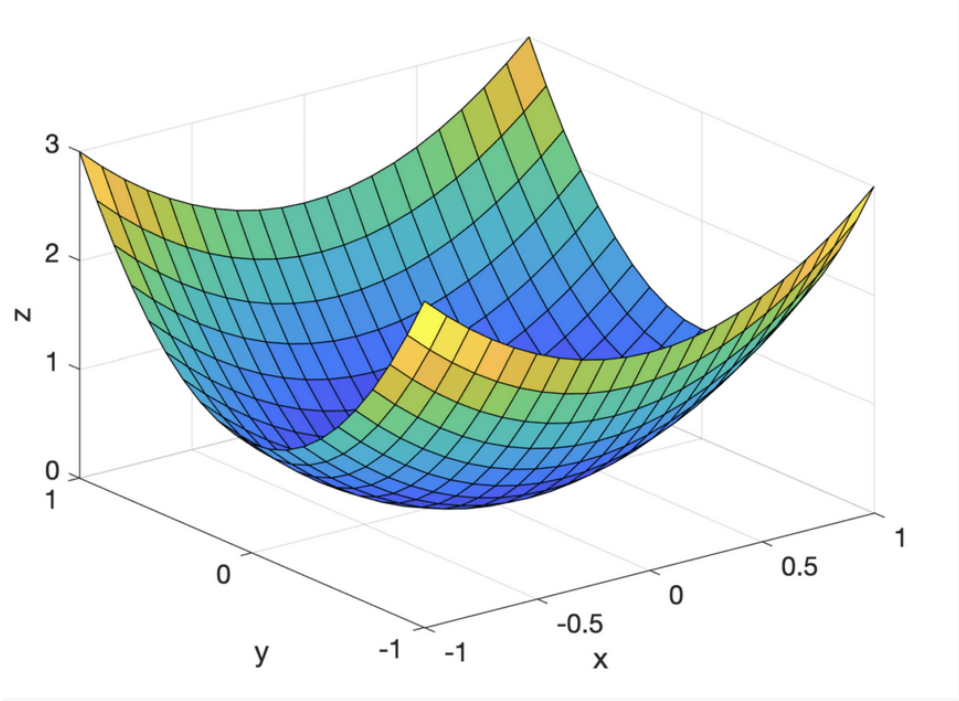
Les algorithmes de descente de gradient incluent parfois une étape de recherche linéaire pour optimiser, à chaque itération, le taux d'apprentissage α.

$$\alpha_k^* = \arg\min_{a \in \mathbb{R}} L(w_k - aL'(w_k))$$

Outline

Fonctions à plusieurs variables

Une fonction a plusieurs variables est une fonction f qui à un vecteur $(x_1, \ldots, x_p)$ associe $f(x_1, \ldots, x_p)$.



Par exemple, la fonction de perte des moindres carrées de la régression linéaire simple :

$$(\beta_0, \beta_1) \mapsto L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Comment calculer des "dérivées" ?

La notion de dérivée se généralise aux fonctions à plusieurs variables. On parle alors de dérivée partielle.

La dérivée partielle de f en fonction de x_j est définie par

$$\frac{\partial f(x_1, \ldots, x_p)}{\partial x_j} = \lim_{h \to 0} \frac{f(x_1, \cdots, x_{j-1}, x_j + h, x_{j+1}, \cdots, x_p) - f(x_1, \ldots, x_p)}{h}$$

Les règles de dérivation habituelles s'appliquent.

Exemple : calculer les dérivées partielles de

$$f(x, y) = \sqrt{x^2 + y^2}$$

Gradient

Definition

Le gradient de la fonction $f : (x_1, \ldots, x_p) \mapsto f(x_1, \ldots, x_p)$ est le vecteur des dérivées partielles.

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_p} \end{pmatrix}$$

où $\mathbf{x} = (x_1, \ldots, x_p)$.

On remarque que le gradient à la même dimension que le vecteur $\mathbf{x}$.

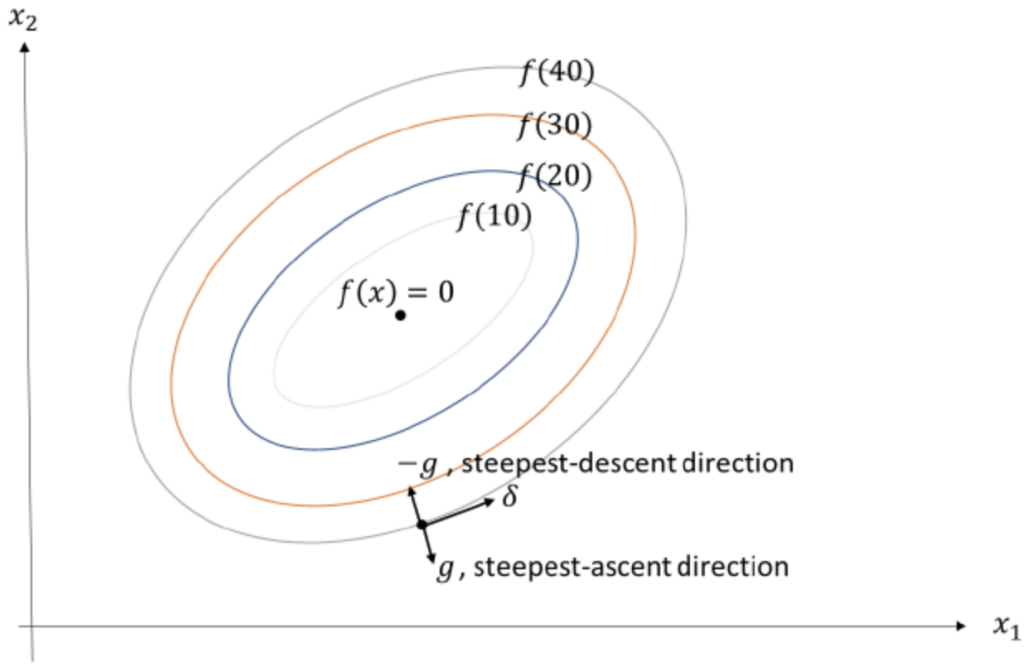
Exemples
- calculer le gradient de la fonction

$$f(x_1, x_2) = x_1^2(x_1 + x_2)$$

- calculer le gradient de la fonction

$$f(x_1, x_2, x_3) = (x_1 - x_2 - ax_3)^2$$

Le gradient donne la direction de plus forte pente



steepest-descent and steepest-ascent method

Le gradient est toujours orthogonal aux lignes de niveau : il donne la direction la plus efficace pour quitter la ligne de niveau en un point donné.

Points critiques d'une fonction à plusieurs variables

Les points critiques d'une fonction f à plusieurs variables sont les points $\mathbf{w}^*$ tels que

$$\nabla f(\mathbf{w}^*) = 0$$

Ce sont des extremums ou des points selle [3].

3. généralisation des points d'inflexion

Extremum d'une fonction à plusieurs variables

Soit $L : \Omega \in \mathbb{R}^p \rightarrow \mathbb{R}$ une fonction convexe [4] qui à $\mathbf{w}$ associe $L(\mathbf{w})$.

On cherche la solution w^* du problème

$$\min_{\mathbf{w} \in \Omega} L(\mathbf{w})$$

Une condition nécessaire est $\nabla f(\mathbf{w}^*) = 0$

4. fonction convexe : $f(\mathbf{x} + \lambda \mathbf{y}) \leq f(\mathbf{x}) + \lambda f(\mathbf{y}), \forall \mathbf{x}, \mathbf{y} \in \Omega, \forall \lambda \in [0, 1]$

Algorithme de descente du gradient

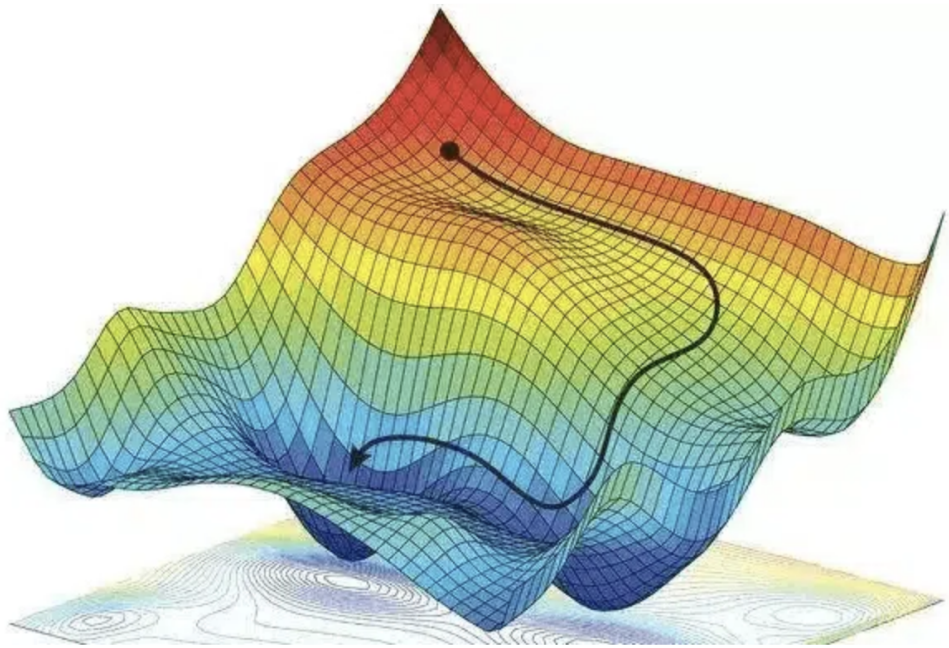
Soit $L : \Omega \in \mathbb{R}^p \rightarrow \mathbb{R}$ une fonction convexe qui à $\mathbf{w}$ associe $L(\mathbf{w})$.

On cherche la solution w^* du problème

$$\min_{\mathbf{w} \in \Omega} L(\mathbf{w})$$

L'algorithme de descente du gradient permet d'approcher la solution

Initialisation : choisir un point $\mathbf{w}_0 \in \Omega$
Tant que $\nabla L(\mathbf{w}) > \epsilon$ faire
$\mathbf{w} \leftarrow \mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \nabla L(\mathbf{w}_k)$



Algorithme de descente du gradient

$$L(\mathbf{y}) \approx L(\mathbf{w}) + \nabla L(\mathbf{w})^T(\mathbf{y} - \mathbf{w})$$

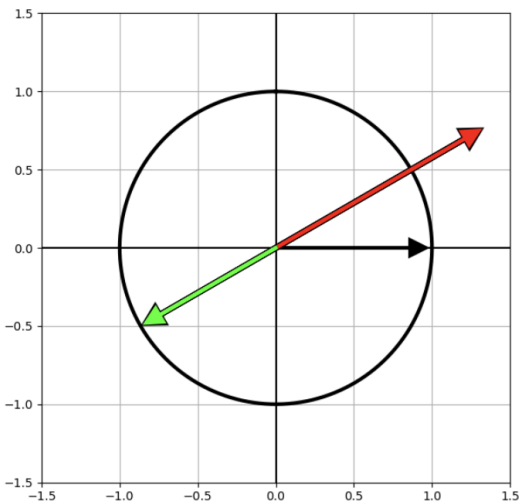
Le vecteur $\mathbf{d} = (\mathbf{y} - \mathbf{w})$ donne une direction générale, centrée au point $\mathbf{w}$.
On cherche $\mathbf{d}$ qui correspond au minimum de

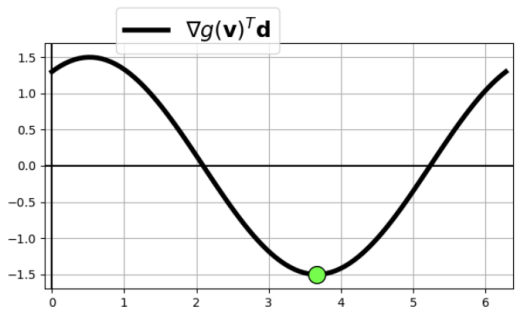
$$L(\mathbf{w}) + \nabla L(\mathbf{w})^T\mathbf{d}$$

Or le produit scalaire s'écrit par définition

$$\nabla L(\mathbf{w})^T\mathbf{d} = \|\nabla L(\mathbf{w})\|^2\|\mathbf{d}\|^2\cos(\theta)$$

où θ est l'angle entre $\nabla L(\mathbf{w})$ et $\mathbf{d}$
$\nabla L(\mathbf{w})^T\mathbf{d}$ atteint son minimum quand $cos(\theta)$ est minimum car les longueurs $\|\nabla L(\mathbf{w})\|$ et $\|\mathbf{d}\|$ sont fixes. Ainsi $\mathbf{d}$ doit pointer dans la direction de $-\nabla L(\mathbf{w})$.





Outline

Fonction de perte localement convexe

On verra dans la suite que les fonctions de perte du machine learning sont rarement convexes.

Dans ce cas, l'algorithme de la descente du gradient converge vers un minimum local qui est dans le "bassin d'attraction" du point de départ $\mathbf{w}_0$.

Une façon de trouver le minimum global est d'essayer plusieurs points de départ puis de comparer les solutions obtenues et conserver la meilleure.

Outline

Take home messages

- Les fonctions strictement convexes admettent un minimum et il est unique.
- Une fonction non convexe peut admettre plusieurs minima locaux.
- Pour une fonction à plusieurs variables, le gradient est le vecteur des dérivées partielles.
- Si une fonction L est différentiable et que son gradient est nul en un point w^* alors $L(w^*)$ est un optimum ou un point selle.
- En un point quelconque w, l'opposé du gradient donne la direction de plus forte pente de la surface au voisinage de w.
- L'algorithme de descente du gradient permet d'approcher un minimum local d'une fonction.
- Dans l'algorithme de descente du gradient, le taux d'apprentissage a un effet sur le nombre d'itérations nécessaire pour attendre le minimum.