Gradient

Vidéo ■ partie 7.1. Dérivées partielles
Vidéo ■ partie 7.2. Gradient et géométrie
Vidéo ■ partie 7.3. Gradient et minimum/maximum
Vidéo ■ partie 7.4. Différentiation automatique
Vidéo ■ partie 7.5. Gradient pour un réseau de neurones

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. Dérivées partielles

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

1.1. Définition

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

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

Plus généralement, pour une fonction $f : \mathbb{R}^n \to \mathbb{R}$ de plusieurs variables, $\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n)$ est la dérivée partielle de f par rapport à la variable x_i au point $(x_1, \ldots, x_n) \in \mathbb{R}^n$. C'est la dérivée en x_i de la fonction d'une variable $x_i \mapsto f(x_1, \ldots, x_n)$ 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 $\frac{\mathrm{d}f}{\mathrm{d}x}(x)$ 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 $f'_x(x, y)$, $f'_y(x, y)$.

Remarque.
Pour une fonction d'une variable $f : \mathbb{R} \to \mathbb{R}$, on distingue le nombre dérivé $f'(x_0)$ et la fonction dérivée f' définie par $x \mapsto f'(x)$. Il en est de même avec les dérivées partielles. Pour $f : \mathbb{R}^2 \to \mathbb{R}$:

- $\dfrac{\partial f}{\partial x}(x_0, y_0)$ et $\dfrac{\partial f}{\partial y}(x_0, y_0)$ sont des nombres réels.
- $\dfrac{\partial f}{\partial x}$ et $\dfrac{\partial f}{\partial y}$ sont des fonctions de deux variables, par exemple :

$$\begin{array}{rccc} \dfrac{\partial f}{\partial x} : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x, y) & \longmapsto & \dfrac{\partial f}{\partial x}(x, y) \end{array}$$

1.2. Calculs

La 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.
Calculer les dérivées partielles de la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ 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.
Pour $f : \mathbb{R}^3 \to \mathbb{R}$ 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.
Soit $f : \mathbb{R}^n \to \mathbb{R}$ définie par $f(x_1, \ldots, x_n) = x_1^2 + x_2^2 + \cdots + x_n^2$, alors pour $i = 1, \ldots, n$:

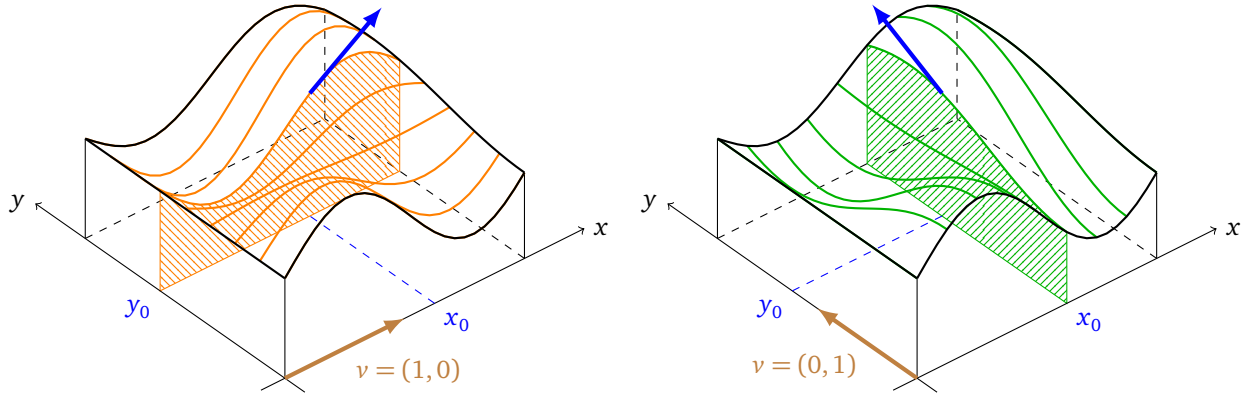
$$\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) = 2x_i.$$

1.3. Interprétation géométrique

Pour une fonction d'une variable, la dérivée est la pente de la tangente au graphe de la fonction (le graphe est 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 est 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) (en orange). Sur la figure de droite, la dérivée partielle $\frac{\partial f}{\partial y}$ indique la pente de la tranche parallèle à l'axe (Oy) (en vert).



2. 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 croît le plus vite.

2.1. Définition

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

$$\operatorname{grad} f(x_0,y_0) = \begin{pmatrix} \dfrac{\partial f}{\partial x}(x_0,y_0) \\ \dfrac{\partial f}{\partial y}(x_0,y_0) \end{pmatrix}.$$

Les physiciens et les anglo-saxons notent souvent $\nabla f(x,y)$ pour $\operatorname{grad} f(x,y)$. Le symbole ∇ se lit « nabla ».
Plus généralement, pour $f : \mathbb{R}^n \to \mathbb{R}$, le gradient de f en $(x_1,\dots,x_n) \in \mathbb{R}^n$ est le vecteur :

$$\operatorname{grad} f(x_1,\dots,x_n) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(x_1,\dots,x_n) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(x_1,\dots,x_n) \end{pmatrix}.$$

Exemple.

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

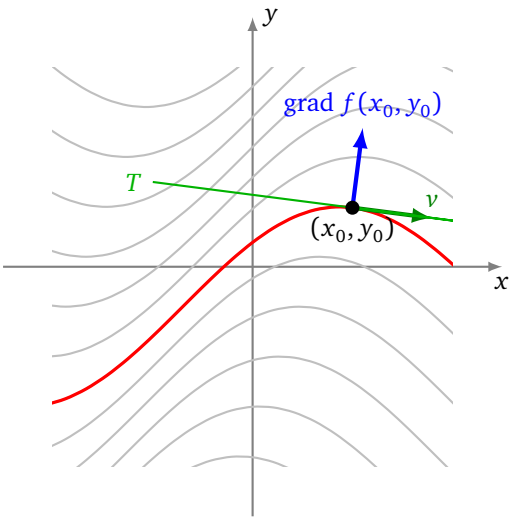
Remarque.
Le gradient est un élément de $\mathbb{R}^n$ écrit comme un vecteur colonne. Parfois, pour alléger l'écriture, on peut aussi l'écrire sous la forme d'un vecteur ligne.

2.2. Tangentes aux lignes de niveau

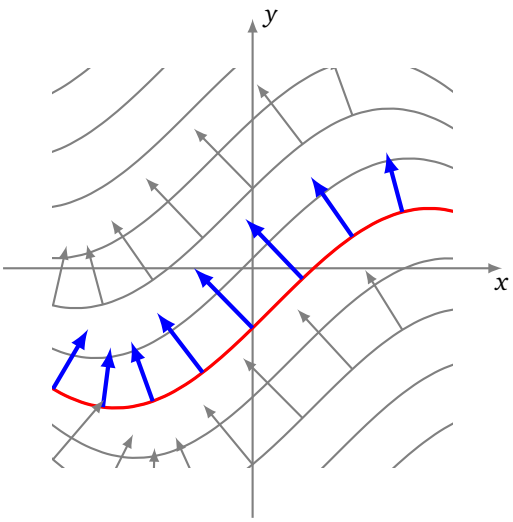
Soit $f : \mathbb{R}^2 \to \mathbb{R}$ une fonction différentiable. On considère les lignes de niveau $f(x,y) = k$.

Proposition 1.
Le vecteur gradient $\operatorname{grad} f(x_0,y_0)$ *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) (en rouge), un vecteur tangent v en ce point et la tangente à la ligne de niveau (en vert). Le vecteur gradient est un vecteur du plan et est orthogonal à la ligne de niveau en ce point (en bleu).



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 ».



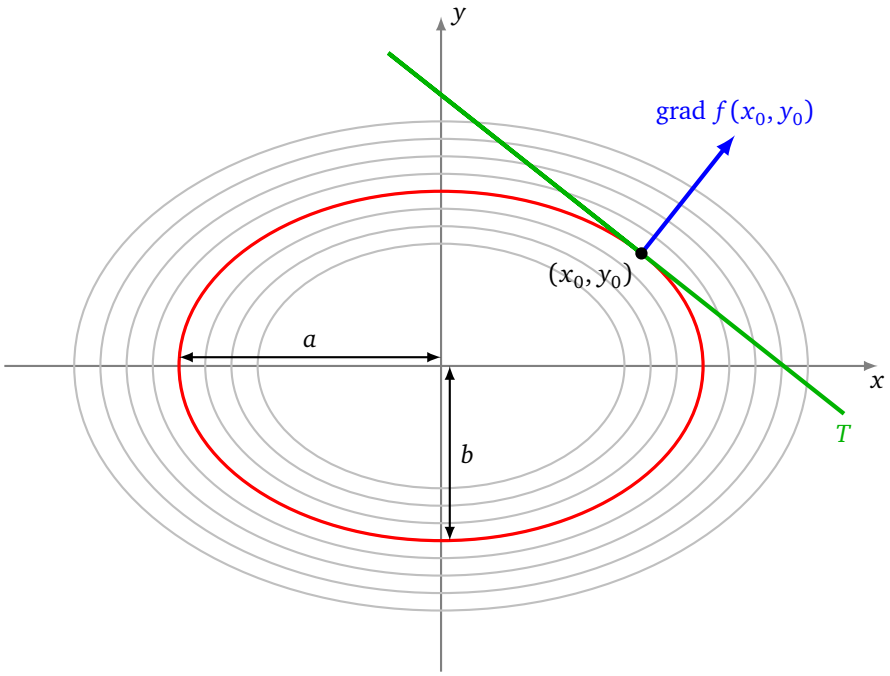
Dans le cadre de notre étude, nous nous intéressons à l'équation de la tangente.

Proposition 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 (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 :

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

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

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

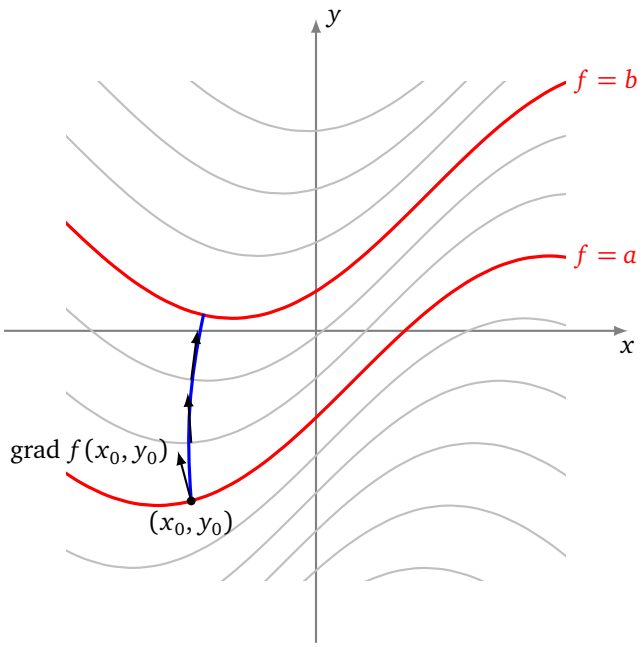
> Mais comme $\frac{x_0^2}{a^2} + \frac{y_0^2}{b^2} = 1$, l'équation de la tangente se simplifie en $\frac{x_0}{a^2}x + \frac{y_0}{b^2}y = 1$.

2.3. Lignes de plus forte pente

Considérons les lignes de niveau $f(x,y) = k$ d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}$. 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 3.
Le vecteur gradient $\operatorname{grad} f(x_0, y_0)$ *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), passer du niveau a au niveau $b > a$ le plus vite possible alors il faut démarrer en suivant la direction du gradient $\operatorname{grad} f(x_0, y_0)$.



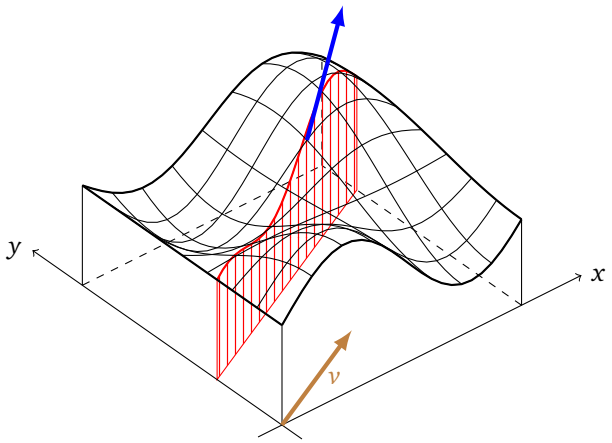
Comme illustration, un skieur de descente, voulant optimiser sa course, choisira en permanence de s'orienter suivant la plus forte pente, c'est-à-dire dans le sens opposé au gradient.

2.4. 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 $v = \binom{h}{k}$ 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.



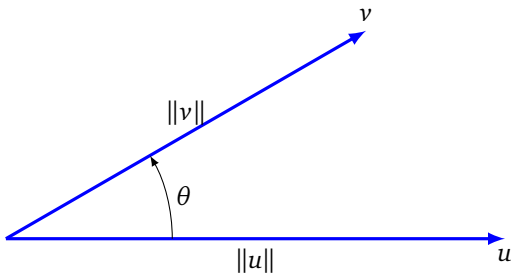
Remarque : pour $v = \binom{1}{0}$ alors $D_v f(x_0, y_0) = \frac{\partial f}{\partial x}(x_0, y_0)$ et pour $v = \binom{0}{1}$ alors $D_v f(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0)$.
On rappelle que le ***produit scalaire*** de deux vecteurs $u = \binom{x}{y}$ et $v = \binom{x'}{y'}$ est donné par

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

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

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

où θ est l'angle entre u et v.



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

$$D_v f(x_0, y_0) = \langle \operatorname{grad} f(x_0, y_0) \mid v \rangle.$$

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,

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

où θ est l'angle entre le vecteur $\operatorname{grad} f(x_0, y_0)$ 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 $\operatorname{grad} f(x_0, y_0)$. □

2.5. 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 $f : \mathbb{R}^3 \to \mathbb{R}$. 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 :

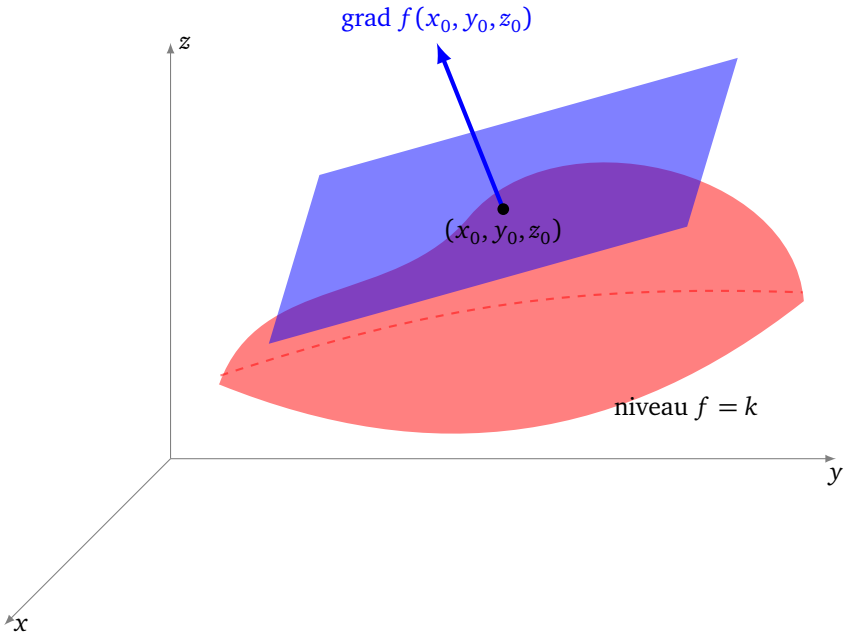
$$a(x - x_0) + b(y - y_0) + c(z - z_0) + d = 0.$$

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

Proposition 4.
Le vecteur gradient $\operatorname{grad} f(x_0, y_0, z_0)$ *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.*



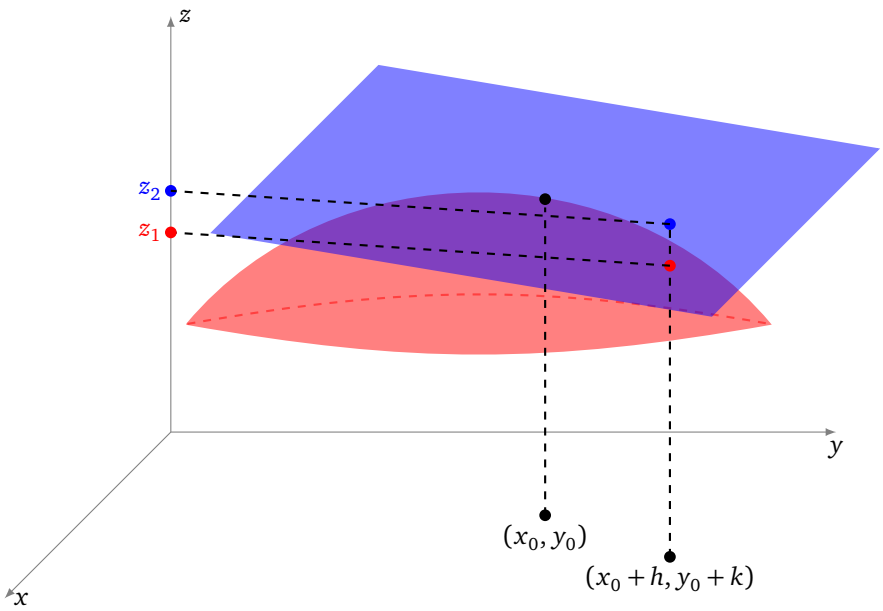
Plus généralement pour $f : \mathbb{R}^n \to \mathbb{R}$, $\operatorname{grad} f(x_1, \ldots, x_n)$ est orthogonal à l'espace tangent à l'hypersurface de niveau $f = k$ passant par le point $(x_1, \ldots, x_n) \in \mathbb{R}^n$ et le vecteur gradient $\operatorname{grad} f(x_1, \ldots, x_n)$ indique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

2.6. 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 (en rouge), le plan tangent au-dessus du point (x_0, y_0) (en bleu). 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.
Valeur approchée de $f(1.002, 0.997)$ si $f(x,y) = x^2 y$.

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

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

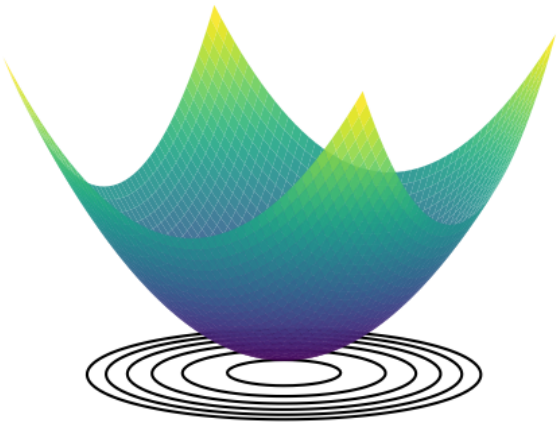
Avec une calculatrice, on trouve $f(1.002, 0.997) = 1.000992$: l'approximation est bonne.

2.7. Minimum et maximum

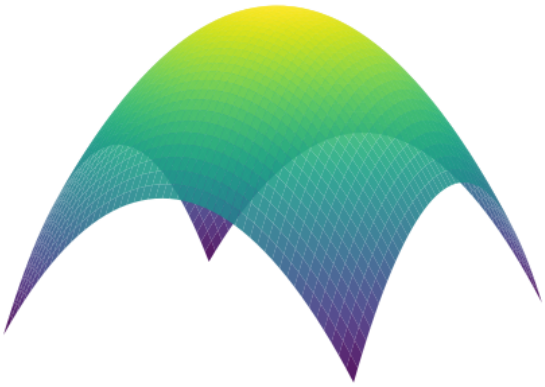
Définition.
Soit $f : \mathbb{R}^2 \to \mathbb{R}$.

- La fonction f admet un ***minimum local*** en (x_0, y_0) s'il existe un disque D centré en ce point tel que
$$f(x,y) \geqslant f(x_0, y_0) \quad \text{pour tout } (x,y) \in D.$$
- De même pour un ***maximum local*** en (x_0, y_0) pour lequel
$$f(x,y) \leqslant f(x_0, y_0) \quad \text{pour tout } (x,y) \in D.$$
- On parle d'un ***extremum local*** pour un minimum ou un maximum local.

Exemple.
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 5.
Soit $f : \mathbb{R}^2 \to \mathbb{R}$. 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 :

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

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 $\frac{\partial f}{\partial y}(x_0, y_0) = 0$. □

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.

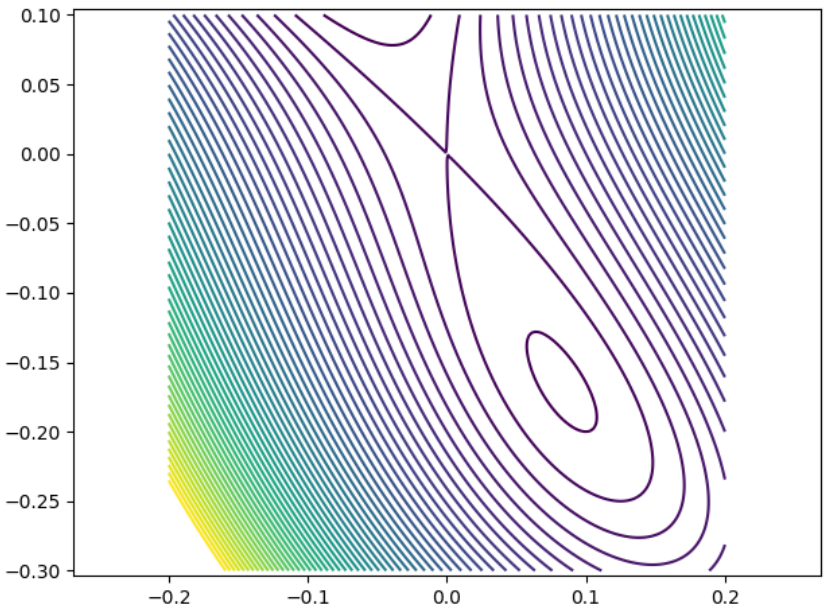
Exemple.
Chercher les points en lesquels $f(x, y) = x^2 - y^3 + xy$ peut atteindre son minimum.
Recherche des points critiques. On calcule

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

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 :

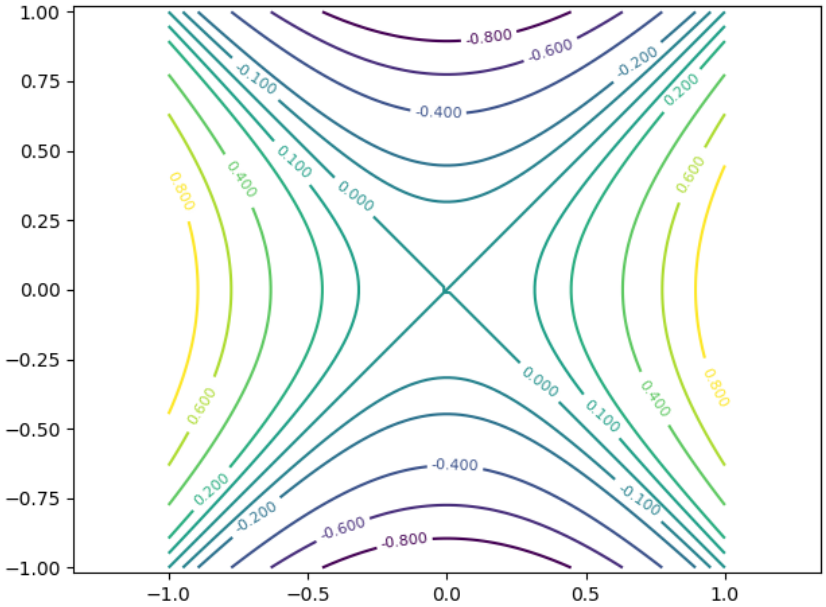
$$(0, 0) \quad \text{et} \quad \left(\frac{1}{12}, -\frac{1}{6}\right).$$

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.



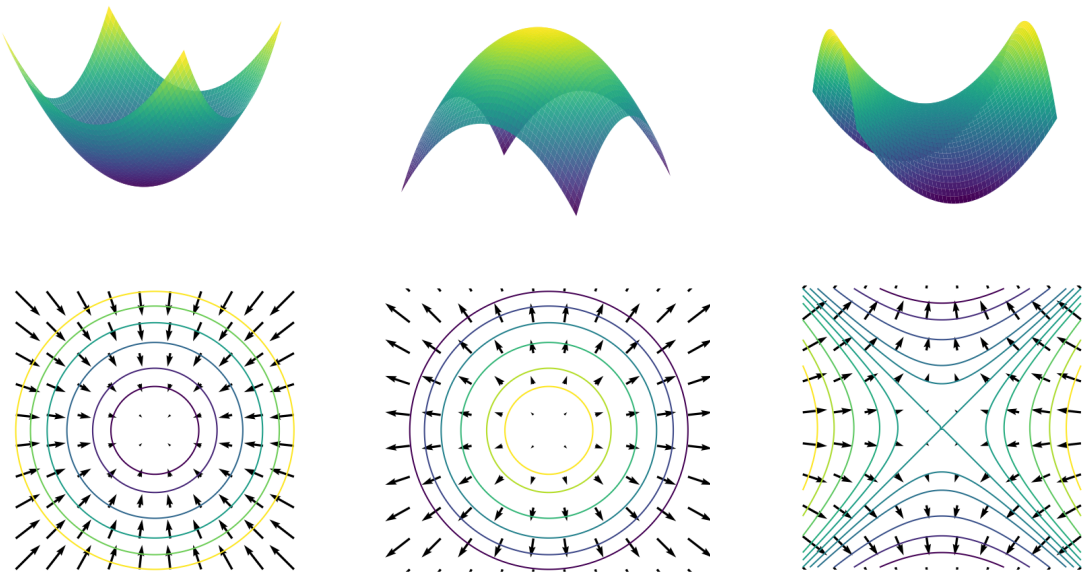
Sur l'exemple précédent, nous avons assez facilement calculé les points critiques à partir des deux équations à deux inconnues. Il faut prendre garde que ce n'est pas un système linéaire et que dans le cas d'une fonction plus compliquée il aurait été impossible de déterminer exactement les points critiques.
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.
Soit $f(x, y) = x^2 - y^2$. Voici son graphe vu sous trois angles différents et ses lignes de niveau.



Comme il est 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 $-\operatorname{grad} f(x, y)$. 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.



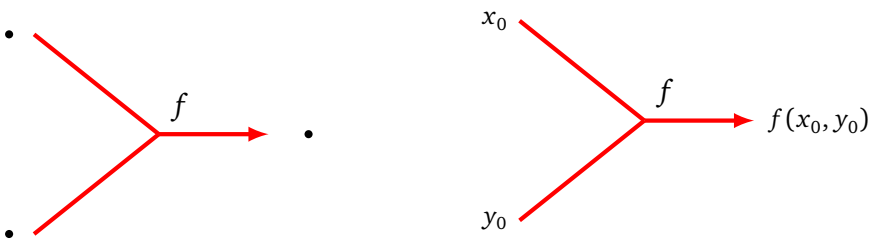
Cas d'un minimum **Cas d'un maximum** **Cas d'un point-selle**

3. Différentiation automatique

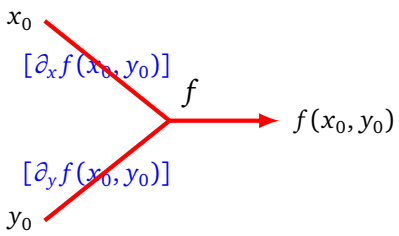
Dans le chapitre « Dérivée », nous avons vu comment calculer la dérivée d'une fonction composée à l'aide de son graphe de calcul. Nous allons faire de même pour les dérivées partielles des fonctions de plusieurs variables afin de calculer le gradient d'une fonction définie par un réseau de neurones.

3.1. Différentiation automatique

Graphe de calcul. Voici le graphe de calcul d'une fonction f de deux variables (schéma de principe à gauche, évaluation en (x_0, y_0) à droite).

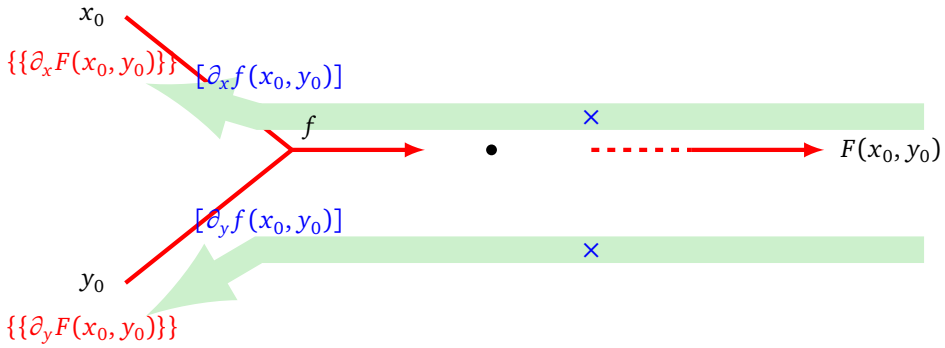


Dérivées locales. Voici la règle pour les dérivées locales à rajouter à chaque branche (entre crochets).

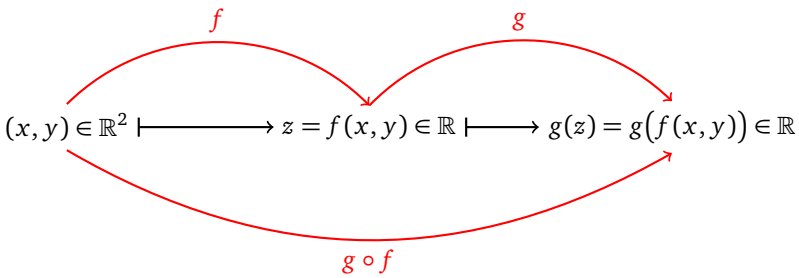


On note $\partial_x f$ comme raccourci de la fonction $\frac{\partial f}{\partial x}$.

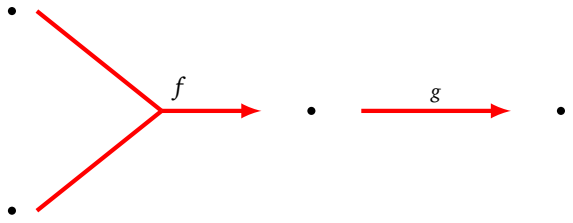
Dérivées partielles. On obtient chacune des dérivées partielles d'une composition F comme le produit des dérivées locales le long des branches allant de la sortie $F(x, y)$ vers l'entrée x (ou y).



Formule mathématique. Soit $f : \mathbb{R}^2 \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$.



La composition $F = g \circ f : \mathbb{R}^2 \to \mathbb{R}$ correspond au graphe de calcul dessiné ci-dessous :



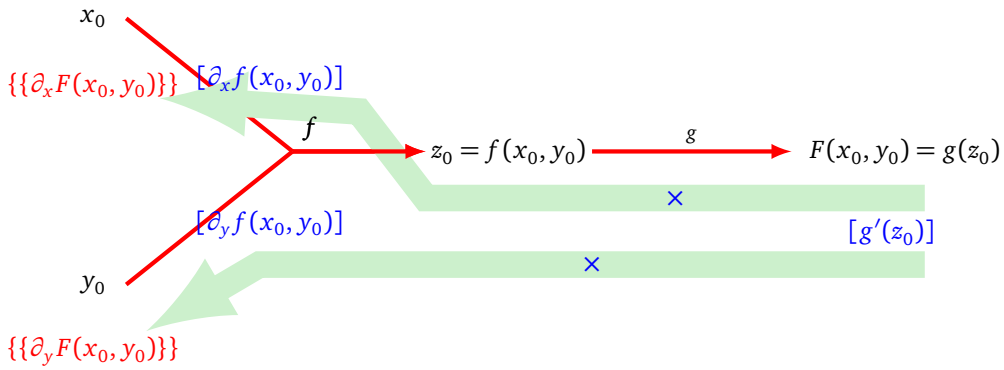
On a donc :

$$F(x, y) = g \circ f(x, y) = g\big(f(x, y)\big).$$

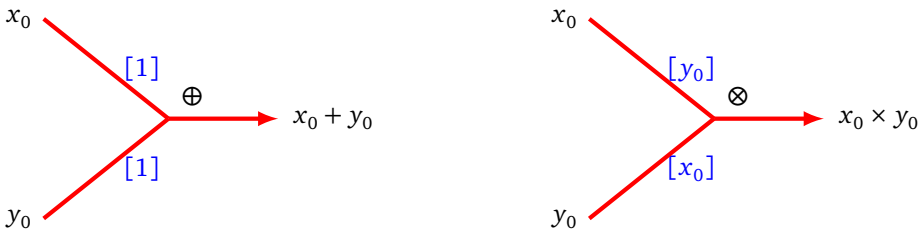
Les dérivées partielles de F sont données par les formules :

$$\boxed{\begin{aligned} \frac{\partial F}{\partial x}(x_0, y_0) &= \frac{\partial f}{\partial x}(x_0, y_0) \cdot g'\big(f(x_0, y_0)\big) \\ \frac{\partial F}{\partial y}(x_0, y_0) &= \frac{\partial f}{\partial y}(x_0, y_0) \cdot g'\big(f(x_0, y_0)\big) \end{aligned}}$$

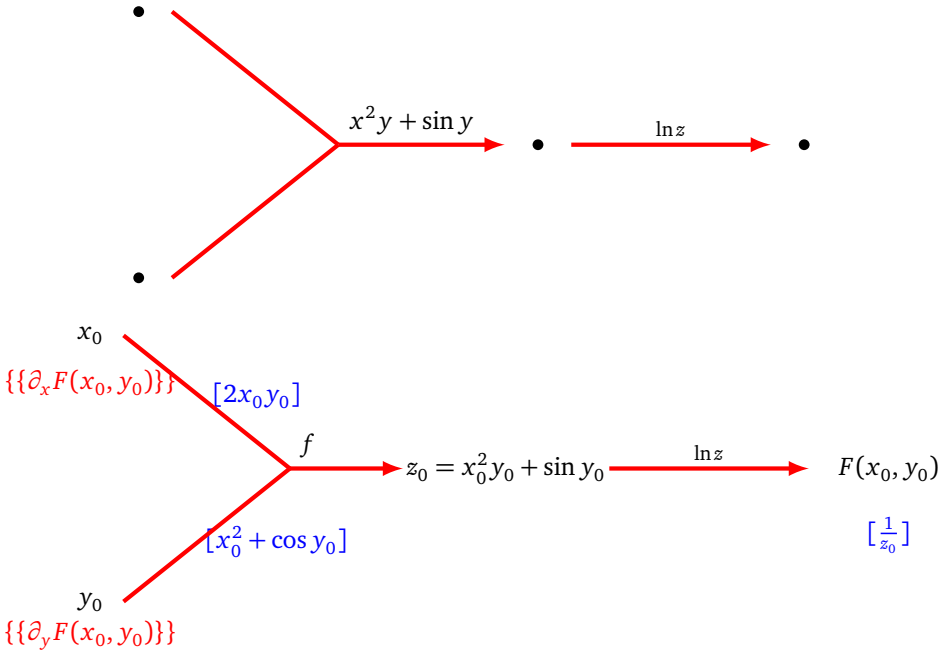
La preuve de la formule pour $\frac{\partial F}{\partial x}(x_0, y_0)$ découle directement de la formule de la dérivée d'une composition pour la fonction d'une seule variable $x \mapsto F(x, y_0)$. Il en est de même pour l'autre dérivée partielle. Ces formules justifient notre règle de calcul : la dérivée partielle est le produit des dérivées locales le long de chacune des branches.



Addition et multiplication.
Dans le cas $f(x, y) = x + y$ et $f(x, y) = x \times y$, on retrouve les dérivées locales déjà utilisées dans le cas d'une seule variable.



Exemple. Soit $F(x, y) = \ln(x^2 y + \sin y)$. On souhaite calculer $\operatorname{grad} F(3, 2)$. Nous allons montrer comment calculer $\operatorname{grad} F(x_0, y_0)$ pour x_0 et y_0 quelconques, puis nous reprendrons les calculs depuis le début dans le cas $(x_0, y_0) = (3, 2)$.

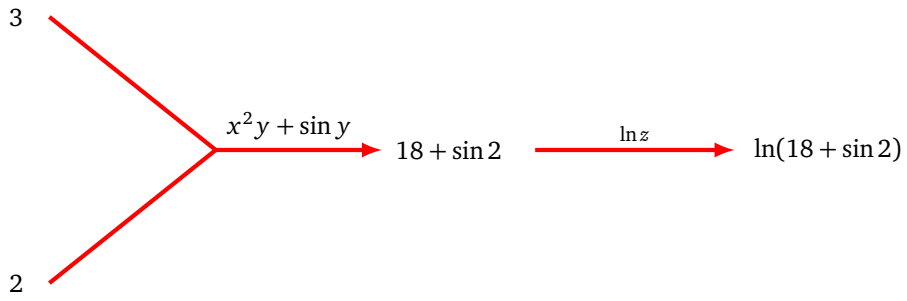


On obtient les dérivées partielles comme produit des dérivées locales :

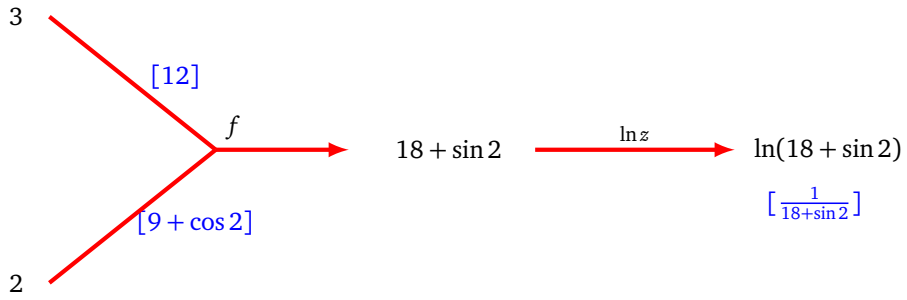
$$\frac{\partial F}{\partial x}(x_0,y_0)=\left[\frac{1}{z_0}\right]\times[2x_0y_0]=\frac{2x_0y_0}{x_0^2y_0+\sin y_0},$$

$$\frac{\partial F}{\partial y}(x_0,y_0)=\left[\frac{1}{z_0}\right]\times[x_0^2+\cos y_0]=\frac{x_0^2+\cos y_0}{x_0^2y_0+\sin y_0}.$$

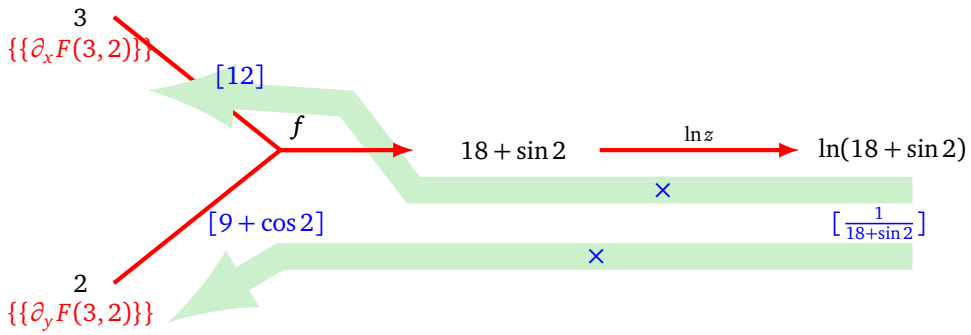
Dans la pratique, pour les réseaux de neurones, on ne calcule jamais l'expression formelle de $\operatorname{grad} F(x_0,y_0)$ mais seulement des gradients en des valeurs (x_0,y_0) données. On reprend donc à chaque fois les étapes ci-dessus mais uniquement pour des valeurs numériques.
La première étape est de calculer les valeurs des fonctions (de la gauche vers la droite).



La seconde étape est de calculer toutes les dérivées locales. On utilise les valeurs de l'étape précédente et la connaissance des formules de chacune des dérivées des fonctions élémentaires (ici $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ et $\frac{\mathrm{d}\ln}{\mathrm{d}z}$).



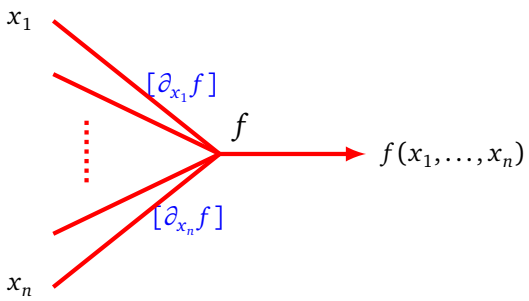
On calcule le produit des dérivées locales le long des arêtes.



On obtient les dérivées partielles :

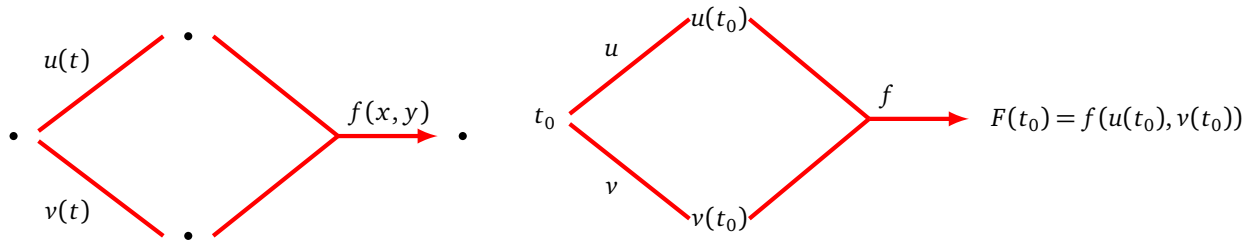
$$\frac{\partial F}{\partial x}(3,2)=\left[\frac{1}{18+\sin 2}\right]\times[12]=\frac{12}{18+\sin 2}\qquad\text{et}\qquad\frac{\partial F}{\partial y}(3,2)=\left[\frac{1}{18+\sin 2}\right]\times[9+\cos 2]=\frac{9+\cos 2}{18+\sin 2}.$$

Règle générale. Dans le cas de n entrées $(x_1,\dots,x_n)$, la règle des dérivées locales se généralise naturellement : on associe à la branche numéro i la dérivée locale $\frac{\partial f}{\partial x_i}$.

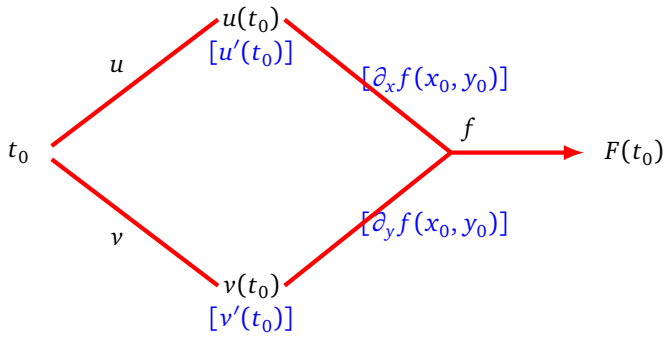


3.2. Différentiation automatique (suite)

Graphe de calcul. Voici le graphe de calcul d'une situation que l'on a déjà rencontrée dans le cas d'une seule variable, mais dont la formule se justifie par les fonctions de deux variables. Il s'agit du graphe de calcul de $F(t) = f\big(u(t), v(t)\big)$ où $u : \mathbb{R} \to \mathbb{R}$, $v : \mathbb{R} \to \mathbb{R}$ et $f : \mathbb{R}^2 \to \mathbb{R}$. L'objectif est de calculer $F'(t)$.

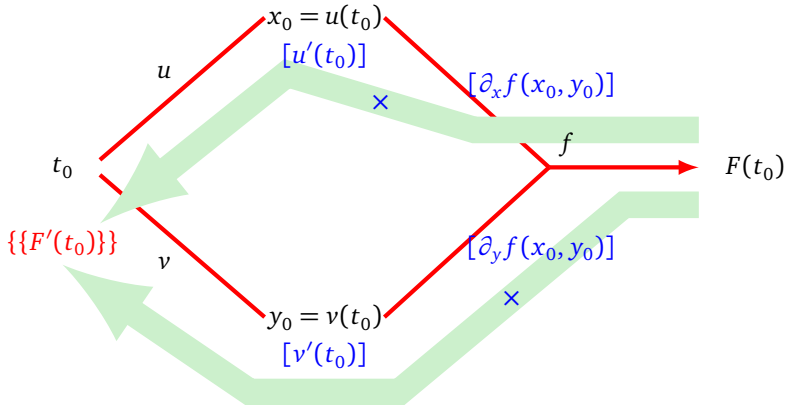


Dérivées locales. On calcule les dérivées locales comme d'habitude.

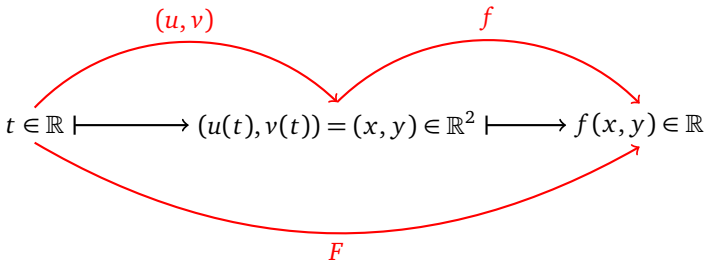


Dérivée. La dérivée s'obtient en deux étapes :

- on calcule le produit des dérivées locales le long des chemins partant de chaque arête sortante jusqu'à la sortie,
- puis on calcule la somme de ces produits.



Formule mathématique. La situation est cette fois la suivante :



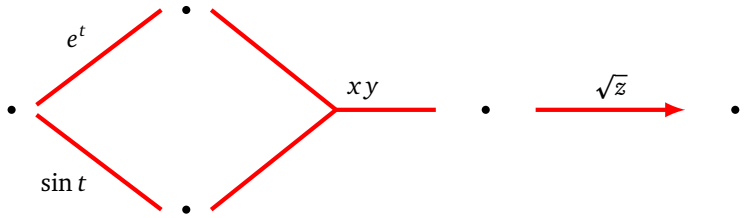
La formule de dérivation de la composition de

$$F(t) = f\big(u(t), v(t)\big)$$

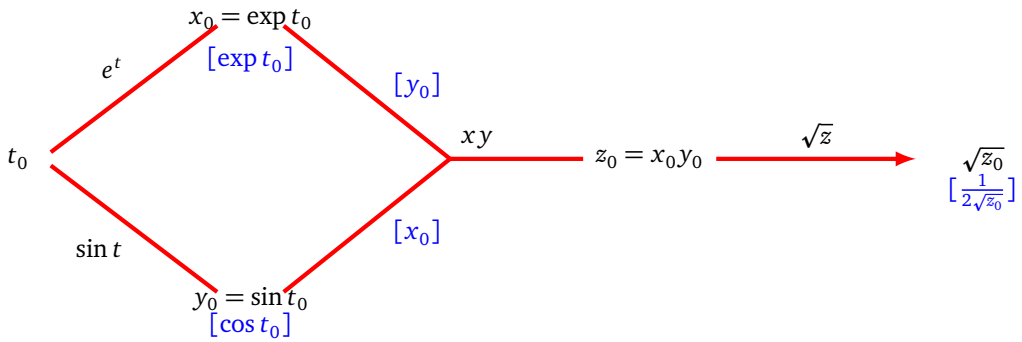
est :

$$\boxed{F'(t_0) = u'(t_0)\frac{\partial f}{\partial x}(u(t_0), v(t_0)) \ + \ v'(t_0)\frac{\partial f}{\partial y}(u(t_0), v(t_0))}$$

Exemple. Soit $F(t) = \sqrt{\exp(t)\sin(t)}$. On souhaite calculer $F'(1)$. On commence par calculer $F'(t_0)$ en général avant de tout reprendre dans la cas $t_0 = 1$. Voici le graphe de calcul :



Une fois complété avec les dérivées locales cela donne :



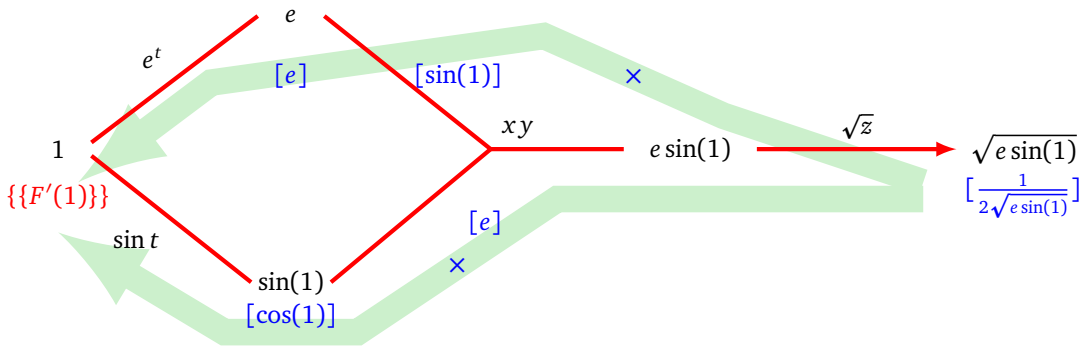
On trouve ainsi :

$$F'(t_0) = \left[\frac{1}{2\sqrt{z_0}}\right] \cdot [y_0] \cdot [\exp t_0] \ + \ \left[\frac{1}{2\sqrt{z_0}}\right] \cdot [x_0] \cdot [\cos t_0]$$

et donc

$$F'(t_0) = \frac{\exp t_0(\sin y_0 + \cos y_0)}{2\sqrt{\exp t_0 \sin y_0}}.$$

Reprenons tout depuis le début pour calculer $F'(1)$ en oubliant que l'on a déjà trouvé la formule générale :



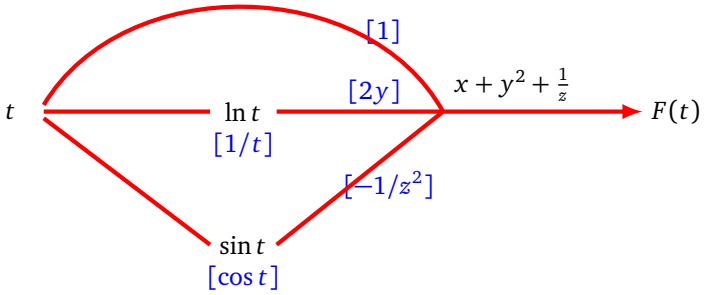
On trouve ainsi :

$$F'(1) = \left[\frac{1}{2\sqrt{e\sin(1)}}\right] \cdot [\sin(1)] \cdot [e] + \left[\frac{1}{2\sqrt{e\sin(1)}}\right] \cdot [e] \cdot [\cos(1)]$$

et donc

$$F'(1) = \frac{e(\sin(1)+\cos(1))}{2\sqrt{e\sin(1)}}.$$

Règle générale. Dans le cas de n sorties, on somme sur toutes les arêtes sortantes comme dans la situation ci-dessous.

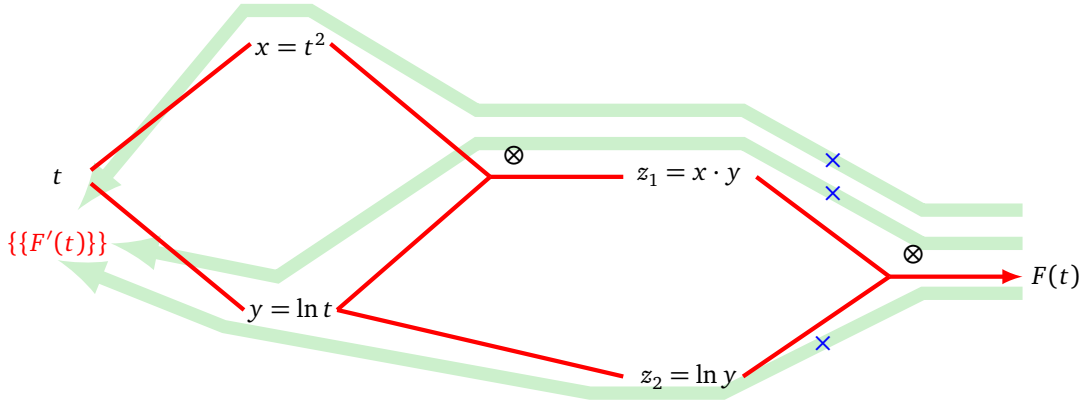


La fonction est $F(t) = t + (\ln t)^2 + \frac{1}{\sin t}$ et en sommant on trouve bien $F'(t) = 1 + \frac{2\ln t}{t} - \frac{\cos t}{\sin^2 t}$.

Un autre exemple. On termine par un exemple plus compliqué : on souhaite calculer la dérivée de

$$F(t) = t \cdot \ln t \cdot \ln(\ln t).$$

Voici le graphe de calcul que l'on utilise (noter qu'avec ce graphe, on ne calcule qu'une seule fois $\ln t$ dont le résultat est réutilisé pour calculer $\ln(\ln t)$).

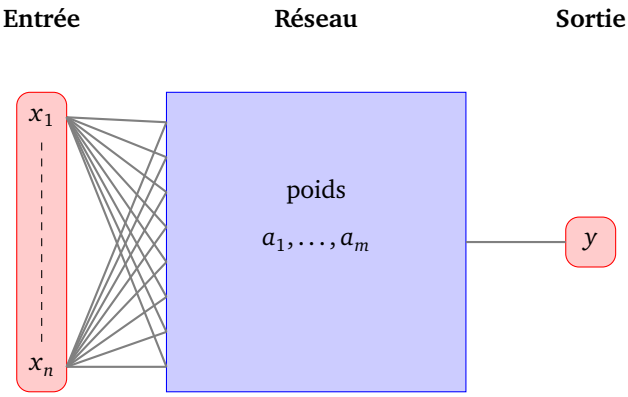


La dérivée s'obtient comme somme sur tous les chemins de la sortie à l'entrée. C'est donc un peu plus compliqué que ce l'on pense au premier abord : il faut ici faire la somme selon trois chemins différents, car l'arête sortant vers le bas de l'entrée t se sépare en deux au niveau de y. Nous allons voir comment gérer cette difficulté dans la section suivante.

4. Gradient pour un réseau de neurones

4.1. Fonction associée à un réseau

Pour un réseau de neurones $\mathcal{R}$ ayant n entrées $(x_1,\dots,x_n)$ et une seule sortie, nous associons une fonction : $F:\mathbb{R}^n\to\mathbb{R}$, $(x_1,\dots,x_n)\mapsto F(x_1,\dots,x_n)$. La situation est en fait plus compliquée. Jusqu'ici les paramètres du réseau étaient donnés. À partir de maintenant les variables du problème ne seront plus les entrées mais les poids du réseau. Notons $a_1,\dots,a_m$ ces poids (l'ensemble des coefficients et des biais). Si l'entrée est fixée et que les poids sont les variables du réseau alors on pourrait considérer que ce même réseau $\mathcal{R}$ définit la fonction $\widetilde{F}:\mathbb{R}^m\to\mathbb{R}$, $(a_1,\dots,a_m)\mapsto\widetilde{F}(a_1,\dots,a_m)$.



Ce dont nous aurons besoin pour la suite et que nous allons calculer dans ce chapitre c'est le gradient de $\widetilde{F}$ par rapport aux poids $(a_1,\dots,a_m)$, autrement dit, il s'agit de calculer :

$$\frac{\partial\widetilde{F}}{\partial a_j}.$$

Pour concilier les deux points de vue (entrées et poids), on dira qu'un réseau de neurones ayant des entrées $(x_1,\dots,x_n)$ et des poids $(a_1,\dots,a_m)$ définit la fonction :

$$\begin{array}{rccc}\widehat{F}: & \mathbb{R}^n\times\mathbb{R}^m & \longrightarrow & \mathbb{R}\\ & (x_1,\dots,x_n),(a_1,\dots,a_m) & \longmapsto & \widehat{F}(x_1,\dots,x_n,a_1,\dots,a_m)\end{array}$$

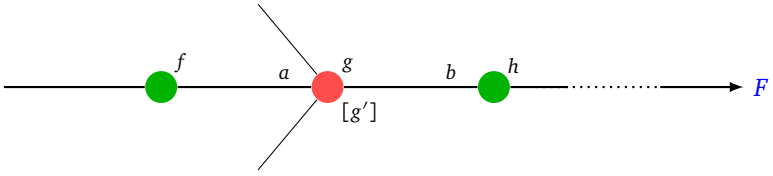
Remarque.
Ce n'est pas tout à fait la fonction $\widetilde{F}$ dont on voudra calculer le gradient mais notre attention se portera sur une fonction d'erreur E de la forme $E=(\widetilde{F}-y_0)^2$, où $\widetilde{F}$ est la fonction définie ci-dessus correspondant à une certaine entrée et à la sortie y_0. On calcule facilement les dérivées partielles de E à partir de celles de $\widetilde{F}$ par la formule :

$$\frac{\partial E}{\partial a_j}=2\frac{\partial\widetilde{F}}{\partial a_j}(\widetilde{F}-y_0).$$

Voir le chapitre « Rétropropagation » pour plus de détails.

4.2. Formule du gradient

On considère la portion suivante d'un réseau de neurones :



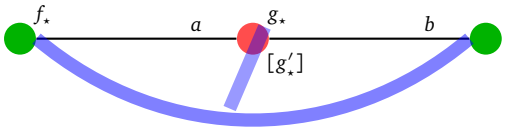
On s'intéresse à une seule arête entrante du neurone central rouge, celle qui porte le poids a.

- a et b sont des poids,
- f, g, h sont des fonctions d'activation,
- f', g', h' sont leur dérivées,
- F est la fonction associée au réseau complet.

Pour distinguer la fonction de sa valeur en un point, on notera f la fonction et $f_\star$ la valeur de la fonction à la sortie du neurone correspondant.
Voici la formule pour calculer la dérivée partielle de F par rapport au coefficient a, connaissant la dérivée partielle par rapport au coefficient b.

$$\boxed{\frac{\partial F}{\partial a} = f_\star \cdot \frac{g'_\star}{g_\star} \cdot b \cdot \frac{\partial F}{\partial b}}$$

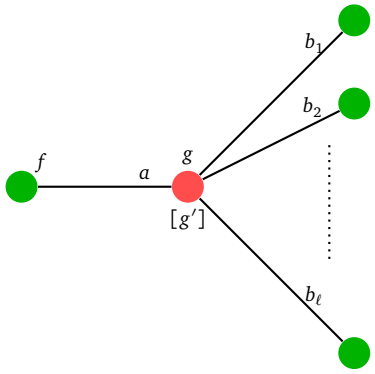
Voici un schéma pour retenir cette « formule du sourire » : on multiplie les coefficients $f_\star$, $g'_\star$, b et la dérivée partielle par rapport à b le long de l'arc et on divise par le coefficient $g_\star$ au bout du segment.



Il est à noter que dans la formule, seule l'arête portant le coefficient a intervient, les autres arêtes entrantes n'interviennent pas (et ne sont pas représentées). Par contre, dans le cas de plusieurs arêtes sortantes il faut calculer la somme des formules précédentes sur chaque arête :

$$\boxed{\frac{\partial F}{\partial a} = \sum_{i=1}^{\ell} f_\star \cdot \frac{g'_\star}{g_\star} \cdot b_i \cdot \frac{\partial F}{\partial b_i}}$$

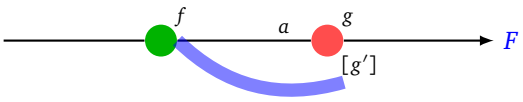
Cette somme comporte autant de termes que d'arêtes sortantes (il n'y a pas à énumérer tous les chemins entre le sommet et la sortie comme auparavant dans la différentiation automatique).



Pour pouvoir calculer toutes les dérivées partielles, on procède par récurrence, en partant de la fin puis en revenant en arrière de proche en proche. Voici la formule d'initialisation associée aux coefficients en sortie de réseau,

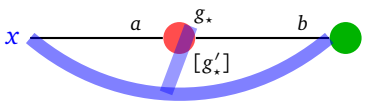
$$\boxed{\frac{\partial F}{\partial a} = f_\star \cdot g'_\star}$$

C'est la « formule du demi-sourire » :



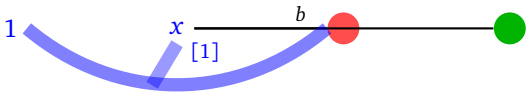
Voici quelques situations particulières, mais qui sont simplement des applications de la formule du sourire.
Formule à l'entrée. On applique la formule du sourire avec x à la place de f.

$$\frac{\partial F}{\partial a} = x \cdot \frac{g_\star'}{g_\star} \cdot b \cdot \frac{\partial F}{\partial b}$$



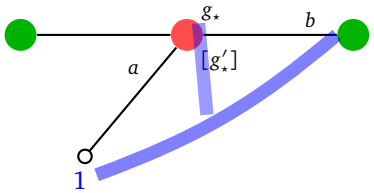
Dérivée partielle par rapport aux variables d'entrées. On applique la formule du sourire en ajoutant des coefficients virtuels égaux à 1 (la dérivée de x par rapport à x est 1) :

$$\frac{\partial F}{\partial x} = \frac{b}{x} \cdot \frac{\partial F}{\partial b}$$



Cas d'un biais. On applique la formule du sourire en ajoutant un coefficient virtuel égal à 1 :

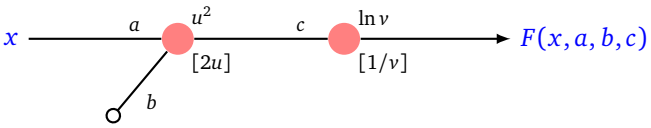
$$\frac{\partial F}{\partial a} = \frac{g_\star'}{g_\star} \cdot b \cdot \frac{\partial F}{\partial b}$$



Remarque.
Ces formules ne sont pas valables lorsque $g_\star = 0$. Nous verrons lors de la preuve de ces formules comment régler ce problème.

4.3. Premier exemple

Voici un réseau très simple.



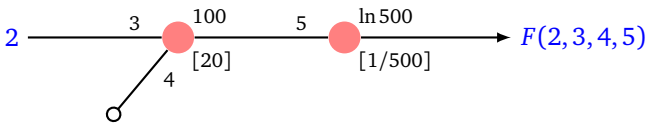
Avec :

- une entrée x, une sortie $F(x)$,
- trois poids a, b, c,
- des fonctions d'activation (plutôt fantaisistes) $u \mapsto u^2$ et $v \mapsto \ln v$.

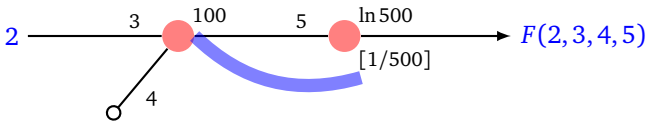
Nous avions l'habitude de considérer la fonction $x \mapsto F(x)$ mais dorénavant les poids sont de nouvelles variables, nous devons donc étudier la fonction $\widehat{F}$ introduite ci-dessus que nous noterons encore F dans la suite :

$$F(x, a, b, c) = \ln\left(c(ax + b)^2\right).$$

Nous souhaitons calculer les dérivées partielles de F par rapport aux poids a, b, c. Nous ne souhaitons pas obtenir une formule générale mais juste la valeur exacte de ces dérivées partielles en un point précis. Nous choisissons l'exemple de $(x, a, b, c) = (2, 3, 4, 5)$. On récrit le réseau avec les valeurs des poids, les valeurs des fonctions et les valeurs des dérivées locales.



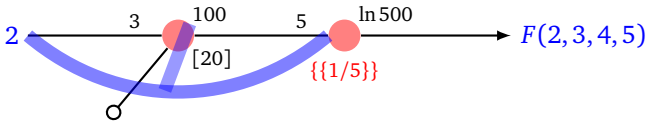
Calcul de la dérivée partielle par rapport à c. On part de la sortie pour l'initialisation. On applique la formule du demi-sourire.



$$\frac{\partial F}{\partial c} = 100 \times \frac{1}{500} = \frac{1}{5}.$$

Calcul de la dérivée partielle par rapport à a. On applique la formule du sourire :

$$\frac{\partial F}{\partial a} = 2 \times \frac{20}{100} \times 5 \times \frac{\partial F}{\partial c}.$$

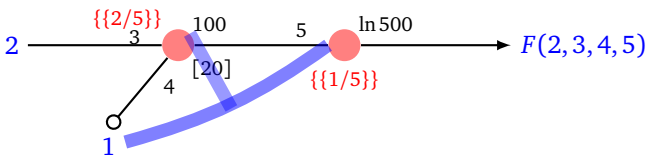


Mais on a déjà calculé $\frac{\partial F}{\partial c} = \frac{1}{5}$ (entre accolades doubles), donc :

$$\frac{\partial F}{\partial a} = \frac{2}{5}.$$

Calcul de la dérivée partielle par rapport à b. On applique la formule du sourire (en posant 1 pour le coefficient manquant) :

$$\frac{\partial F}{\partial b} = 1 \times \frac{20}{100} \times 5 \times \frac{\partial F}{\partial c}.$$

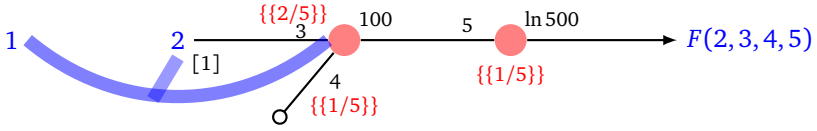


Donc

$$\frac{\partial F}{\partial b} = \frac{1}{5}.$$

Calcul de la dérivée partielle par rapport à x. On peut aussi calculer cette dérivée partielle, même si nous n'en aurons pas besoin dans les autres chapitres.

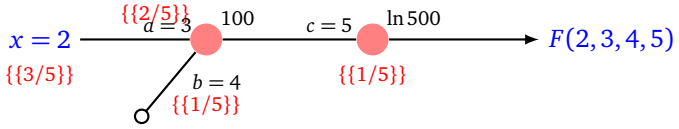
$$\frac{\partial F}{\partial x} = 1 \times \frac{1}{2} \times 3 \times \frac{\partial F}{\partial a}.$$



donc

$$\frac{\partial F}{\partial x} = \frac{3}{5}.$$

Bilan. Ainsi $F(2,3,4,5) = \ln 500$ et on a calculé les dérivées partielles (entre doubles accolades) pour chacune des variables x, a, b, c.



Vérification. On peut vérifier nos formules en calculant directement les dérivées partielles à partir de l'expression :

$$F(x,a,b,c) = \ln\big(c(ax+b)^2\big).$$

Par exemple :

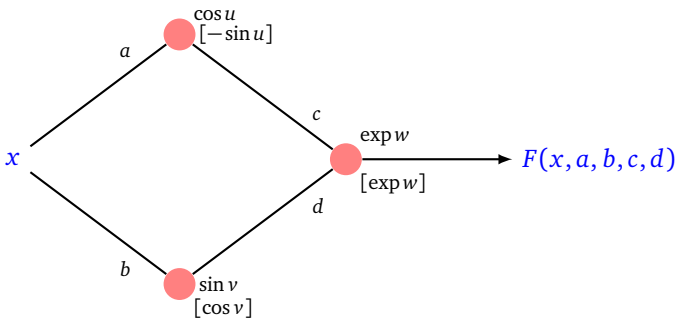
$$\frac{\partial F}{\partial a}(x,a,b,c) = \frac{2cx(ax+b)}{c(ax+b)^2} = \frac{2x}{ax+b}$$

et on a bien

$$\frac{\partial F}{\partial a}(2,3,4,5) = \frac{4}{10} = \frac{2}{5}.$$

4.4. Second exemple

Pour le réseau suivant, on associe comme précédemment une fonction F.



On considère les poids a, b, c, d comme des variables, la fonction $F : \mathbb{R}^5 \to \mathbb{R}$ s'écrit donc :

$$F(x,a,b,c,d) = \exp\big(c\cos(ax) + d\sin(bx)\big).$$

On va noter $u = ax$ et $v = bx$, ainsi $\cos u$ et $\sin v$ sont les sorties des deux premiers neurones. On note $w = c\cos u + d\sin v$. La sortie du troisième neurone (qui est aussi la valeur de F) est alors $\exp w$.
On part de la sortie pour l'initialisation. Il y a deux dérivées partielles à calculer.
Calcul de la dérivée partielle par rapport à c. On applique la formule du demi-sourire :

$$\frac{\partial F}{\partial c} = \cos u \cdot \exp w.$$

Calcul de la dérivée partielle par rapport à d. On applique de nouveau la formule du demi-sourire :

$$\frac{\partial F}{\partial d} = \sin v \cdot \exp w.$$

Calcul de la dérivée partielle par rapport à a. On applique la formule du sourire :

$$\frac{\partial F}{\partial a} = x \cdot \frac{-\sin u}{\cos u} \cdot c \cdot \frac{\partial F}{\partial c}$$

donc

$$\frac{\partial F}{\partial a} = -xc\sin u \exp w.$$

Calcul de la dérivée partielle par rapport à b. On applique la formule du sourire :

$$\frac{\partial F}{\partial b} = x \cdot \frac{\cos v}{\sin v} \cdot d \cdot \frac{\partial F}{\partial d}$$

donc

$$\frac{\partial F}{\partial b} = xd\cos v \exp w.$$

Calcul de la dérivée partielle par rapport à x. Cette dérivée partielle s'obtient comme la somme de deux termes correspondant aux deux arêtes sortantes :

$$\frac{\partial F}{\partial x} = a \cdot \frac{1}{x} \cdot \frac{\partial F}{\partial a} + b \cdot \frac{1}{x} \cdot \frac{\partial F}{\partial b}$$

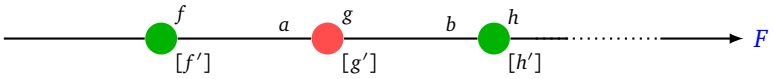
donc

$$\frac{\partial F}{\partial x} = (-ac \sin u + bd \cos v) \exp w.$$

Vérification. En effectuant les substitutions $u = ax$, $v = bx$ et $w = c\cos(ax) + d\sin(bx)$, on retrouve les dérivées partielles attendues, par exemple

$$\frac{\partial F}{\partial x} = \big(-ac\sin(ax) + bd\cos(bx)\big) \exp\big(c\cos(ax) + d\sin(bx)\big).$$

4.5. Preuve et formule générale



Préliminaires.

$$\frac{\partial h}{\partial g} = b \cdot h'_\star \tag{1}$$

Preuve : on a $h_\star = h(bg_\star)$, la formule s'obtient en dérivant $g \mapsto h(bg)$ par rapport à la variable g.

$$\frac{\partial g}{\partial a} = f_\star \cdot g'_\star \tag{2}$$

Preuve : on a $g_\star = g(\cdots + af_\star + \cdots)$, la formule s'obtient en dérivant $a \mapsto g(\cdots + af + \cdots)$ par rapport à la variable a.

$$\frac{\partial h}{\partial b} = g_\star \cdot h'_\star \tag{3}$$

Preuve : c'est la même formule que l'équation (2) mais cette fois pour $b \mapsto h(\cdots + bg + \cdots)$.

Formule générale.

$$\frac{\partial F}{\partial a} = \frac{\partial F}{\partial g} \cdot f_\star \cdot g'_\star \tag{4}$$

Preuve : c'est la formule

$$\frac{\partial F}{\partial a} = \frac{\partial F}{\partial g} \cdot \frac{\partial g}{\partial a}$$

(valable car g est une fonction d'une seule variable) suivie de l'application de l'équation (2).

$$\frac{\partial F}{\partial g} = \frac{\partial F}{\partial h} \cdot b \cdot h'_\star \tag{5}$$

Preuve : c'est la formule

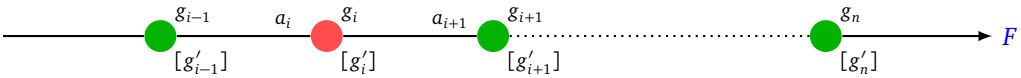
$$\frac{\partial F}{\partial g} = \frac{\partial F}{\partial h} \cdot \frac{\partial h}{\partial g}$$

suivie de l'application de l'équation (1).
Dans le cas d'un neurone de sortie on a :

$$\frac{\partial F}{\partial g} = 1 \tag{6}$$

car comme g est la dernière fonction, on a $F_\star = g_\star$.

Algorithme.



Voici comment calculer toutes les dérivées partielles voulues (y compris dans le cas $g_\star = 0$ qui avait été exclu dans la formule du sourire).

- On part du neurone de sortie pour lequel on initialise le processus par la formule (6) ce qui donne $\frac{\partial F}{\partial g_n} = 1$.
- On procède par récurrence à rebours. On suppose que l'on a déjà calculé $\frac{\partial F}{\partial g_{i+1}}$ On en déduit :

  $$\frac{\partial F}{\partial g_i} = \frac{\partial F}{\partial g_{i+1}} \cdot a_{i+1} \cdot g'_{i+1,\star}$$

  par la formule (5).
- Cela permet de calculer les dérivées partielles par rapport aux poids à l'aide de la formule (4) :

  $$\frac{\partial F}{\partial a_i} = \frac{\partial F}{\partial g_i} \cdot g_{i-1,\star} \cdot g'_{i,\star}.$$

Preuve de la formule du sourire.
On va exprimer $\frac{\partial F}{\partial a}$ directement en fonction de $\frac{\partial F}{\partial b}$.
Par les équations (4) et (5), on a d'une part :

$$\frac{\partial F}{\partial a} = \frac{\partial F}{\partial h} \cdot b \cdot h'_\star \cdot f_\star \cdot g'_\star \tag{7}$$

et d'autre part :

$$\frac{\partial F}{\partial b} = \frac{\partial F}{\partial h} \cdot \frac{\partial h}{\partial b}.$$

Donc, en utilisant l'équation (3), on obtient :

$$\frac{\partial F}{\partial b} = \frac{\partial F}{\partial h} \cdot g_\star \cdot h'_\star. \tag{8}$$

Cette dernière équation permet de calculer $\frac{\partial F}{\partial h}$ en fonction $\frac{\partial F}{\partial b}$. Ainsi des équations (7) et (8), on obtient la formule du sourire :

$$\frac{\partial F}{\partial a} = f_\star \cdot \frac{g'_\star}{g_\star} \cdot b \cdot \frac{\partial F}{\partial b}.$$

Dans le cas de plusieurs arêtes sortantes, il s'agit de faire une somme comme on l'a déjà vue lors de la différentiation automatique.