Dérivée

Vidéo ■ partie 1.1. Définition de la dérivée
Vidéo ■ partie 1.2. Dérivée d'une composition
Vidéo ■ partie 1.3. Dérivation automatique
Vidéo ■ partie 1.4. Fonctions d'activation
Vidéo ■ partie 1.5. Minimums et maximums

La notion de dérivée joue un rôle clé dans l'étude des fonctions. Elle permet de déterminer les variations d'une fonction et de trouver ses extremums. Une formule fondamentale pour la suite sera la formule de la dérivée d'une fonction composée.

Ceux qui sont à l'aise en mathématiques peuvent se rendre directement à la deuxième section de ce chapitre consacrée à la « dérivation automatique ».

1. Dérivée

1.1. Définition

Soit $f : I \to \mathbb{R}$ une fonction, où I est un intervalle ouvert de $\mathbb{R}$ (par exemple du type $]a, b[$). Soit $x_0 \in I$.

Définition.
La dérivée de f en x_0, si elle existe, est le nombre

$$\boxed{f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.}$$

C'est donc la limite du taux d'accroissement $\frac{f(x)-f(x_0)}{x-x_0}$ lorsque x tend vers x_0. Nous noterons la dérivée de f en x_0 indifféremment sous la forme :

$$\boxed{f'(x_0) \qquad \text{ou} \qquad \frac{\mathrm{d}f}{\mathrm{d}x}(x_0).}$$

Remarque.
Une dérivée n'existe pas toujours. Dans ce cours nous supposerons par défaut que la dérivée est bien définie, c'est-à-dire que f est ***dérivable*** en x_0. Nous préciserons explicitement les situations pour lesquelles ce n'est pas le cas. Comme la limite est unique, la dérivée de f en x_0 ne peut prendre qu'une seule valeur.

> **Exemple.**
> Calculons la dérivée en $x_0 = 1$ de la fonction f définie par $f(x) = x^2$. On commence par réécrire le taux

d'accroissement :

$$\frac{f(x)-f(1)}{x-1} = \frac{x^2-1}{x-1} = \frac{(x-1)(x+1)}{x-1} = x+1.$$

Ce taux d'accroissement tend vers 2 lorsque x tend vers 1, donc $f'(1)=2$.
Plus généralement, on montre que $f'(x_0)=2x_0$:

$$\frac{f(x)-f(x_0)}{x-x_0} = \frac{x^2-x_0^2}{x-x_0} = \frac{(x-x_0)(x+x_0)}{x-x_0} = x+x_0 \xrightarrow[x\to x_0]{} 2x_0.$$

Exemple.
On connaît la limite suivante :

$$\frac{\exp(x)-1}{x} \xrightarrow[x\to 0]{} 1.$$

Interprétons ceci en termes de dérivée. Soit $f(x)=\exp(x)$. Alors la limite ci-dessus s'écrit :

$$\frac{f(x)-f(0)}{x-0} \xrightarrow[x\to 0]{} 1,$$

c'est-à-dire $f'(0)=1$. Autrement dit, la dérivée de l'exponentielle en 0 vaut 1.

Pour chaque x_0 en lequel la fonction f est dérivable, on associe un nombre $f'(x_0)$, ce qui nous permet de définir une nouvelle fonction.

Définition.
La fonction qui à x associe $f'(x)$ est la ***fonction dérivée*** de f. On la notera de l'une des façons suivantes :

$$x \mapsto f'(x) \qquad \text{ou} \qquad f' \qquad \text{ou} \qquad \frac{\mathrm{d}f}{\mathrm{d}x}.$$

Pour le premier exemple avec $f(x)=x^2$, nous avons montré que $f'(x)=2x$ (et donc $f'(1)=2$). L'exponentielle possède la propriété fondamentale que sa dérivée est aussi l'exponentielle : si $f(x)=\exp(x)$ alors $f'(x)=\exp(x)$. On retrouve bien que $f'(0)=\exp(0)=1$.

1.2. Calcul approché de valeurs

Connaître la dérivée d'une fonction en un point permet d'approcher les valeurs de la fonction autour de ce point. Commençons par un petit changement de variable. Posons $h=x-x_0$, ce qui revient à écrire $x=x_0+h$ (et considérer un intervalle centré en x_0). Comme on s'intéresse aux valeurs de x qui tendent vers x_0, cela revient à dire que h tend vers 0. Le taux d'accroissement devient $\frac{f(x_0+h)-f(x_0)}{h}$ et on a :

$$\frac{f(x_0+h)-f(x_0)}{h} \xrightarrow[h\to 0]{} f'(x_0).$$

Cela fournit une valeur approchée de f en x_0+h, pourvu que h soit proche de 0 :

$$\boxed{f(x_0+h) \simeq f(x_0)+hf'(x_0).}$$

Démonstration. Comme

$$\frac{f(x_0+h)-f(x_0)}{h} \xrightarrow[h\to 0]{} f'(x_0),$$

alors pour h suffisamment petit :

$$\frac{f(x_0+h)-f(x_0)}{h} \simeq f'(x_0).$$

En multipliant de part et d'autre par h, on obtient l'estimation voulue. □

Exemple.
On souhaite trouver une valeur approchée de $\sin(0.01)$ sans calculatrice. Posons $f(x) = \sin(x)$. On sait que $f'(x) = \cos(x)$. Avec $x_0 = 0$, on a $f(x_0) = \sin(0) = 0$. On se doute bien que $\sin(0.01)$ sera proche de 0, mais on veut faire mieux. Posons $h = 0.01$ et calculons $f'(x_0) = \sin'(0) = \cos(0) = 1$. Donc notre formule s'écrit, pour h proche de 0 :

$$\sin(h) = f(x_0 + h) \simeq f(x_0) + hf'(x_0) = 0 + h \cdot 1 = h.$$

Et donc pour $h = 0.01$ on a $\sin(0.01) \simeq 0.01$ On vérifie à la calculatrice que $\sin(0.01) = 0.00999983\ldots$, donc notre approximation est très bonne. (Attention, il faut d'abord sélectionner les radians comme unité d'angle sur la calculatrice.)

Exemple.
Justifions la formule

$$\sqrt{1+h} \simeq 1 + \frac{1}{2}h,$$

valable pour des valeurs de h proches de 0.
Soit $f(x) = \sqrt{x}$ et $x_0 = 1$. On sait que $f'(x) = \frac{1}{2}\frac{1}{\sqrt{x}}$ donc $f'(x_0) = \frac{1}{2}$. Pour h proche de 0 :

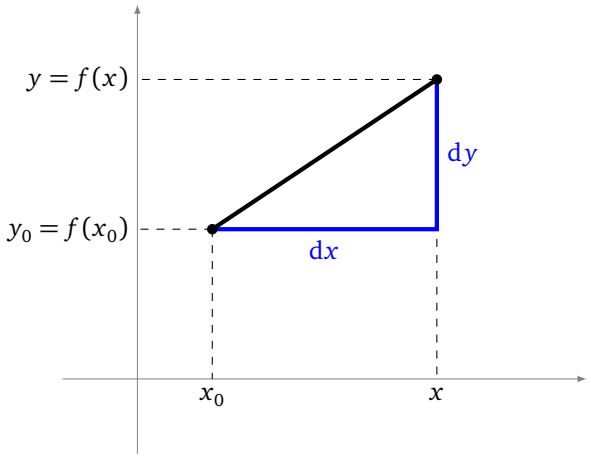
$$\sqrt{1+h} = f(x_0 + h) \simeq f(x_0) + hf'(x_0) = 1 + \frac{1}{2}h.$$

Par exemple avec $h = 0.1$ on obtient :

$$\sqrt{1.1} \simeq 1 + 0.5 \times 0.1 = 1.05$$

et la calculatrice donne $\sqrt{1.1} = 1.0488\ldots$

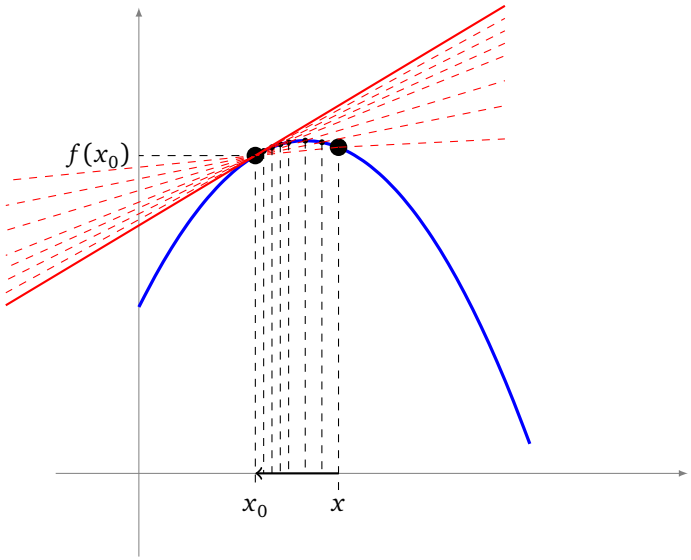
Remarque.
Voici quelques explications sur la notation $\frac{\mathrm{d}f}{\mathrm{d}x}$, préférée par les physiciens, et qu'il faut bien comprendre car nous allons la généraliser plus tard.
La notation « $\mathrm{d}x$ » représente un élément infinitésimal de la variable x, c'est-à-dire la valeur $x - x_0$, avec x très proche de x_0 (c'est-à-dire $x \to x_0$). $\mathrm{d}y$ ou encore $\mathrm{d}f$ représente la variation correspondante de la fonction, c'est-à-dire la valeur $f(x) - f(x_0)$, pour les mêmes valeurs de x et x_0. Ainsi $\frac{\mathrm{d}f}{\mathrm{d}x}(x_0)$ représente le quotient de ces deux valeurs, autrement dit le taux d'accroissement pris à la limite.



1.3. Tangente

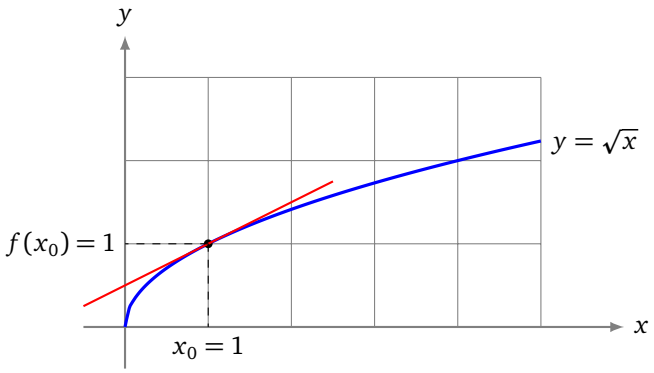
L'interprétation géométrique de la dérivée est essentielle ! Le coefficient directeur de la tangente au graphe de f en x_0 est $f'(x_0)$.
Une équation de la ***tangente*** au point $(x_0, f(x_0))$ est donc :

$$\boxed{y = (x - x_0)f'(x_0) + f(x_0)}$$

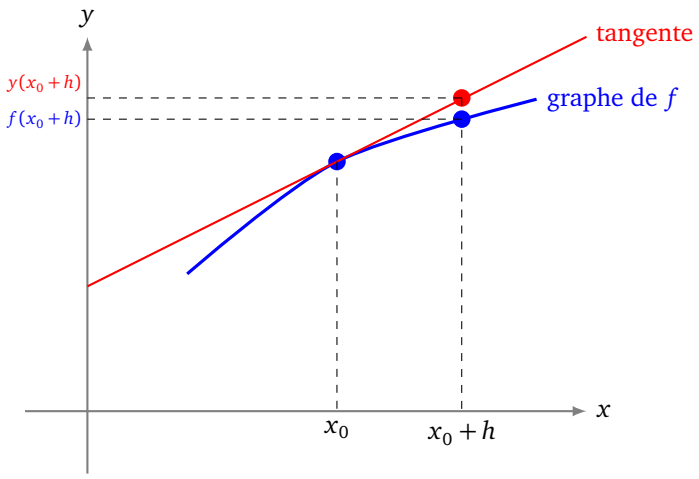


Justification : la droite qui passe par les points $(x_0, f(x_0))$ et $(x, f(x))$ a pour coefficient directeur $\frac{f(x)-f(x_0)}{x-x_0}$. À la limite, lorsque $x \to x_0$, on trouve que le coefficient directeur de la tangente est $f'(x_0)$. Voir l'illustration ci-dessus pour des valeurs de x tendant vers x_0.

Exemple.
Voici le graphe de la fonction définie par $f(x) = \sqrt{x}$. La dérivée est $f'(x) = \frac{1}{2}\frac{1}{\sqrt{x}}$. La tangente en $x_0 = 1$ a donc pour équation $y = (x-1)\frac{1}{2} + 1$, autrement dit c'est $y = \frac{1}{2}x + \frac{1}{2}$.



La tangente en x_0 est la droite qui « approche » au mieux le graphe de f autour x_0. Voici l'interprétation géométrique de l'approximation étudiée dans la section précédente : pour x proche de x_0, au lieu de lire les valeurs $f(x)$ sur le graphe de f, on lit les valeurs approchées $y(x) = (x - x_0)f'(x_0) + f(x_0)$ sur la tangente en x_0.



1.4. Formules usuelles

Voici les expressions des dérivées de fonctions classiques. Elles sont à connaître sur le bout des doigts.

Fonction	Dérivée
x^n	$nx^{n-1} \quad (n \in \mathbb{Z})$
$\frac{1}{x}$	$-\frac{1}{x^2}$
$\sqrt{x}$	$\frac{1}{2}\frac{1}{\sqrt{x}}$
x^α	$\alpha x^{\alpha-1} \quad (\alpha \in \mathbb{R})$
e^x	e^x
$\ln x$	$\frac{1}{x}$
$\cos x$	$-\sin x$
$\sin x$	$\cos x$
$\tan x$	$1+\tan^2 x = \frac{1}{\cos^2 x}$

Ces formules, conjuguées aux opérations décrites dans la proposition ci-dessous, permettent de calculer un très grand nombre de dérivées.

Proposition 1.
À partir de deux fonctions dérivables $f : I \to \mathbb{R}$ et $g : I \to \mathbb{R}$, on calcule la dérivée des opérations élémentaires suivantes :

- ***Somme***

$$\boxed{(f+g)' = f' + g'}$$

  Autrement dit $(f+g)'(x) = f'(x) + g'(x)$ pour tout $x \in I$.
- ***Produit par un scalaire***

$$\boxed{(\lambda f)' = \lambda f'}$$

  où λ est un réel fixé. Autrement dit $(\lambda f)'(x) = \lambda f'(x)$.
- ***Produit***

$$\boxed{(f \times g)' = f'g + fg'}$$

Autrement dit $(f \times g)'(x) = f'(x)g(x) + f(x)g'(x)$.

- ***Quotient***

$$\boxed{\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}}$$

Autrement dit $\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$ *(si* $g(x) \neq 0$*).*

> **Exemple.**
> - Soit $f_1(x) = x^2 + \cos(x)$ alors $f_1'(x) = 2x - \sin(x)$. La dérivée d'une somme est la somme des dérivées.
> - Soit $f_2(x) = x^3 \times \sin(x)$ alors $f_2'(x) = 3x^2 \sin(x) + x^3 \cos(x)$. Nous avons appliqué la formule de la dérivée d'un produit.
> - Soit $f_3(x) = \frac{\ln(x)}{e^x}$ alors $f_3'(x) = \frac{\frac{1}{x}e^x - \ln(x)e^x}{e^{2x}}$ en appliquant la formule de la dérivée d'un quotient et la relation $(e^x)^2 = e^{2x}$.

Démonstration. Voyons comment prouver une des formules, par exemple $(f \times g)' = f'g + fg'$.
Fixons $x_0 \in I$. Nous allons réécrire le taux d'accroissement de $f(x) \times g(x)$:

$$\begin{aligned}\frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0} &= \frac{f(x) - f(x_0)}{x - x_0}g(x) + \frac{g(x) - g(x_0)}{x - x_0}f(x_0) \\ &\xrightarrow[x \to x_0]{} f'(x_0)g(x_0) + g'(x_0)f(x_0).\end{aligned}$$

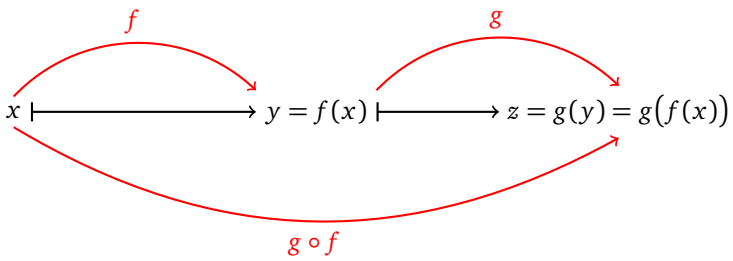
Ceci étant vrai pour tout $x_0 \in I$, la fonction $f \times g$ est dérivable sur I et a pour dérivée $f'g + fg'$. □

1.5. Dérivée d'une composition

Soient $f : \mathbb{R} \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$ deux fonctions. La ***composition*** $g \circ f$ est définie par :

$$g \circ f(x) = g\big(f(x)\big).$$

Il faut bien faire attention à l'ordre : on calcule d'abord $f(x)$ puis on évalue la quantité $g\big(f(x)\big)$ à l'aide de la fonction g.



La dérivée d'une composition est la formule fondamentale de tout ce cours ! En effet, c'est cette formule qui permet de calculer les bons paramètres pour un réseau de neurones.
Voici la formule :

Proposition 2.
La dérivée de $g \circ f$ *est donnée par :*

$$\boxed{\big(g \circ f\big)'(x) = g'\big(f(x)\big) \cdot f'(x)}$$

Autrement dit, si $F(x) = g\big(f(x)\big)$ alors $F'(x) = g'\big(f(x)\big) \cdot f'(x)$.
Voici une façon pratique et mnémotechnique de retenir la formule : on note $y = f(x)$ et $z = g(y) = g \circ f(x)$.
Alors

$$\boxed{\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x}}$$

qui est exactement notre formule :

$$\frac{\mathrm{d}g \circ f}{\mathrm{d}x}(x) = \frac{\mathrm{d}g}{\mathrm{d}y}(y) \times \frac{\mathrm{d}f}{\mathrm{d}x}(x).$$

Pourquoi cette formule est-elle si facile à retenir ? C'est comme si les dérivées se comportaient comme des fractions que l'on pouvait simplifier. Bien sûr c'est juste la notation qui permet cela.

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\cancel{\mathrm{d}y}} \times \frac{\cancel{\mathrm{d}y}}{\mathrm{d}x}$$

Démonstration. La preuve est similaire à celle du produit de deux fonctions, si on suppose que $f(x) \neq f(x_0)$ pour $x \neq x_0$ (autour de x_0), en écrivant cette fois :

$$\begin{aligned}\frac{g \circ f(x) - g \circ f(x_0)}{x - x_0} &= \frac{g\big(f(x)\big) - g\big(f(x_0)\big)}{f(x) - f(x_0)} \times \frac{f(x) - f(x_0)}{x - x_0} \\ &\xrightarrow[x \to x_0]{} g'\big(f(x_0)\big) \times f'(x_0).\end{aligned}$$

□

Exemple.
Commençons par un exemple simple avec $F(x) = \sin(3x)$. C'est la composition de $f(x) = 3x$ avec $g(x) = \sin(x)$: $F(x) = g \circ f(x)$. On sait que $f'(x) = 3$ et $g'(x) = \cos(x)$ donc $g'(f(x)) = \cos(3x)$. Ainsi

$$F'(x) = 3\cos(3x).$$

Une autre façon de le voir est de poser $y = 3x$, $z = \sin(y)$ $(= \sin(3x) = F(x))$. On a $\frac{\mathrm{d}z}{\mathrm{d}y} = \cos(y)$ et $\frac{\mathrm{d}y}{\mathrm{d}x} = 3$, ainsi :

$$F'(x) = \frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x} = \cos(y) \times 3.$$

Et comme $y = 3x$, on obtient :

$$F'(x) = 3\cos(3x).$$

Exemple.
Soit $F(x) = \ln(\cos(x))$. C'est la composition de $f(x) = \cos(x)$ avec $g(x) = \ln(x)$: $F(x) = g \circ f(x)$. On sait que $f'(x) = -\sin(x)$. D'autre part $g'(x) = \frac{1}{x}$, donc $g'(f(x)) = \frac{1}{f(x)} = \frac{1}{\cos(x)}$. Ainsi :

$$F'(x) = \frac{-\sin(x)}{\cos(x)}.$$

Exemple.
Soit $F(x) = \sqrt{\tan(x)}$. Cette fois utilisons la notation des physiciens avec $y = \tan(x)$ et $z = \sqrt{y}$ $(= F(x))$. Alors :

$$F'(x) = \frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{1}{2}\frac{1}{\sqrt{y}} \times (1 + \tan^2(x)).$$

Et comme $y = \tan(x)$, on obtient :

$$F'(x) = \frac{1}{2}\frac{1 + \tan^2(x)}{\sqrt{\tan(x)}}.$$

Voici ce que donnent les dérivées des compositions des fonctions usuelles. Les formules s'obtiennent en appliquant directement la formule donnée plus haut, mais sont tout de même à connaître par cœur. Ici u désigne une fonction $x \mapsto u(x)$.

Fonction	Dérivée
u^n	$nu'u^{n-1} \quad (n \in \mathbb{Z})$
$\frac{1}{u}$	$-\frac{u'}{u^2}$
$\sqrt{u}$	$\frac{1}{2}\frac{u'}{\sqrt{u}}$
u^α	$\alpha u'u^{\alpha-1} \quad (\alpha \in \mathbb{R})$
e^u	$u'e^u$
$\ln u$	$\frac{u'}{u}$
$\cos u$	$-u' \sin u$
$\sin u$	$u' \cos u$
$\tan u$	$u'(1+\tan^2 u) = \frac{u'}{\cos^2 u}$

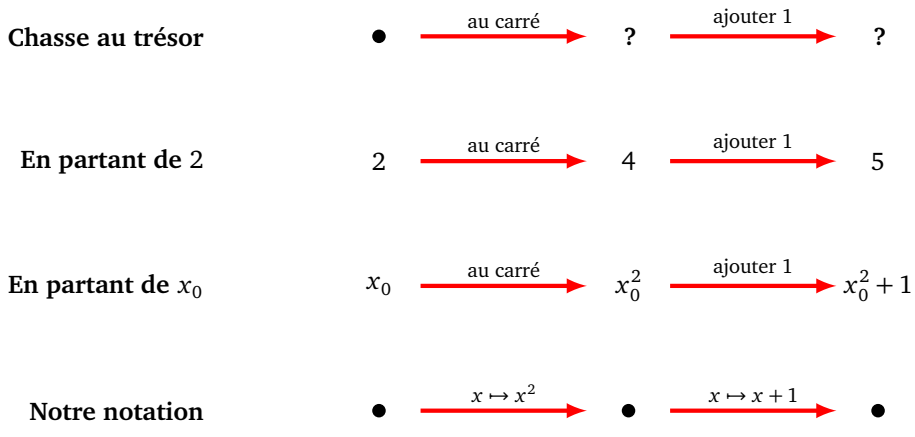
La formule de la dérivée de composition se généralise : si x est une fonction de t, y est une fonction de x et z est une fonction de y alors :

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x} \times \frac{\mathrm{d}x}{\mathrm{d}t}.$$

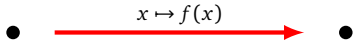
2. Dérivation automatique

2.1. Graphe de calcul

Un graphe de calcul est comme une chasse au trésor : on effectue des calculs étape par étape jusqu'à obtenir le résultat final. Voici un exemple ci-dessous.

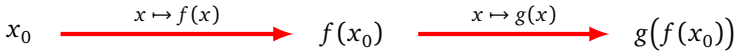


On part d'un nombre et on doit calculer des résultats en fonction des opérations précisées sur les flèches (ici mettre le nombre au carré, puis ajouter 1). Si on part par exemple de 2 : on l'élève au carré, puis on ajoute 1 pour obtenir le résultat 5. Si on partait d'un réel quelconque x_0, cela donne $x_0^2 + 1$. Le dernier diagramme est le graphe de calcul tel que nous allons le représenter dans la suite.
Plus généralement, on représente l'évaluation par une fonction f comme ceci :

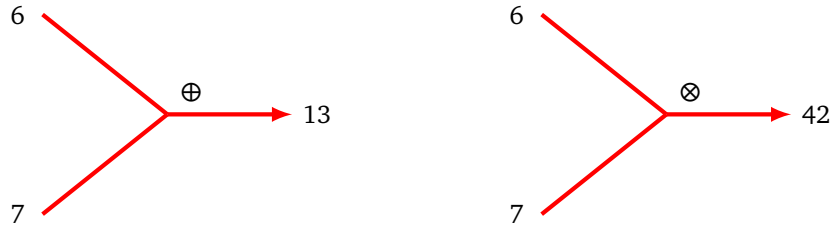


ce qui permet de représenter facilement la composition de deux fonctions $g \circ f$:

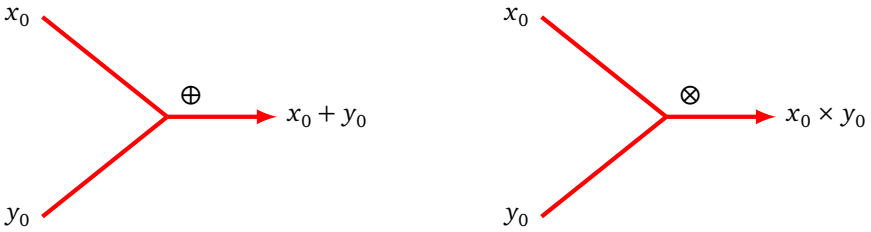
qui correspond au calcul :



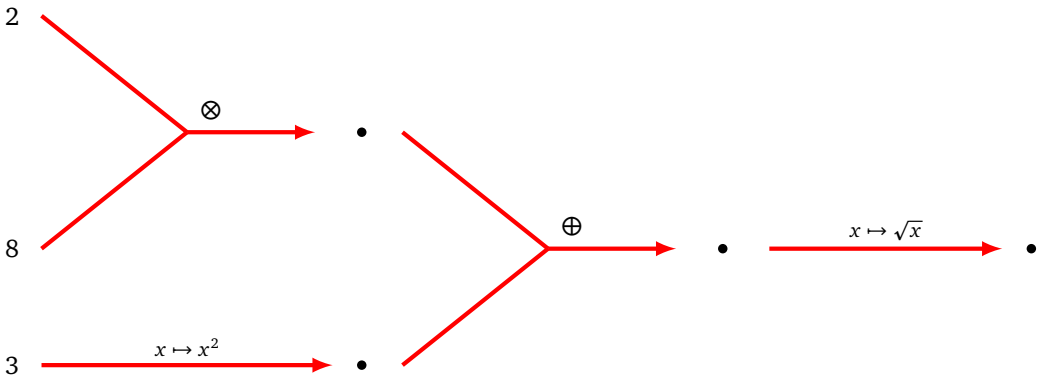
Pour les additions et les multiplications, chaque graphe possède deux entrées et une sortie.
Par exemple :



Et plus généralement :

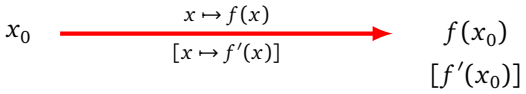


Voici un graphe de calcul obtenu en combinant plusieurs opérations. Vérifier que le résultat final (tout à droite) vaut 5. Écrire le calcul algébrique effectué en remplaçant les nombres 2, 8 et 3 par trois variables x_0, y_0 et z_0.



2.2. Dérivation automatique (cas simple)

Nous allons enrichir nos graphes de calcul à l'aide de la dérivée : en plus d'écrire le résultat du calcul $f(x_0)$, on ajoute entre crochets la valeur de la dérivée $f'(x_0)$. On appelle cette valeur la ***dérivée locale***. Pour aider à effectuer les calculs, on peut rappeler sous la flèche la formule de la fonction dérivée.



Voici trois exemples :

$$2 \xrightarrow[[x \mapsto \frac{1}{x}]]{x \mapsto \ln(x)} \underset{[\frac{1}{2}]}{\ln(2)} \qquad -1 \xrightarrow[[x \mapsto -2e^{-2x}]]{x \mapsto e^{-2x}} \underset{[-2e^2]}{e^2} \qquad 3 \xrightarrow{x \mapsto x^2} \underset{[6]}{9}$$

Bien sûr, il n'est pas obligatoire d'écrire la formule $[x \mapsto f'(x)]$ sous la flèche quand on connaît bien ses dérivées !

Voici comment évaluer la dérivée d'une composition. Prenons la fonction $F(x) = \cos(2x^2)$ pour laquelle on souhaite par exemple calculer $F'(3)$.

Une première méthode serait de calculer $F'(x)$ (quel que soit x), puis évaluer $F'(3)$. Voici une autre méthode : on pose $f(x) = 2x^2$ et $g(y) = \cos(y)$ alors $F(x) = g \circ f(x)$ (on aurait pu noter $g(x) = \cos(x)$ mais utiliser le nom y est plus pratique pour la suite).

$$3 \xrightarrow{x \mapsto 2x^2} \bullet \xrightarrow{y \mapsto \cos(y)} \bullet$$

Première étape : graphe de calcul. On écrit le graphe de calcul correspondant à cette composition :

$$3 \xrightarrow{x \mapsto 2x^2} 18 \xrightarrow{y \mapsto \cos(y)} \cos(18)$$

Seconde étape : ajout des dérivées locales. On calcule les valeurs de dérivées locales :

$$3 \xrightarrow[[x \mapsto 4x]]{x \mapsto 2x^2} \underset{[12]}{18} \xrightarrow[[y \mapsto -\sin(y)]]{y \mapsto \cos(y)} \underset{[-\sin(18)]}{\cos(18)}$$

Troisième étape : calcul de la dérivée de la composition. On calcule la dérivée de la composition comme le produit des dérivées locales.

On note $\{\{F'(3)\}\}$ cette dérivée, entre accolades doubles.

$$x = 3 \xrightarrow[[x \mapsto 4x]]{x \mapsto 2x^2} \underset{[12]}{y = 18} \xrightarrow[[y \mapsto -\sin(y)]]{y \mapsto \cos(y)} \underset{[-\sin(18)]}{z = \cos(18)}$$

$$\times$$

$$\{\{-12\sin(18)\}\}$$

Pour des raisons que l'on expliquera plus tard, il faut prendre l'habitude de calculer le produit de la droite vers la gauche : on effectue $(-\sin(18)) \times 12$ et non $12 \times (-\sin(18))$.

Reprenons la même fonction, mais cette fois pour calculer $F'(4)$:

Exemple avec $x = 4$ $\quad 4 \xrightarrow{x \mapsto 2x^2} \bullet \xrightarrow{y \mapsto \cos(y)} \bullet$

Graphe de calcul $\quad 4 \xrightarrow{x \mapsto 2x^2} 32 \xrightarrow{y \mapsto \cos(y)} \cos(32)$

Dérivées locales $\quad 4 \xrightarrow[[x \mapsto 4x]]{x \mapsto 2x^2} \underset{[16]}{32} \xrightarrow[[y \mapsto -\sin(y)]]{y \mapsto \cos(y)} \underset{[-\sin(32)]}{\cos(32)}$

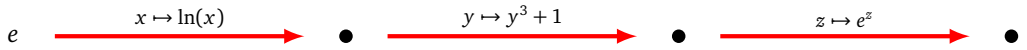
Dérivée en $x = 4$ $\quad x = 4 \xrightarrow[[x \mapsto 4x]]{x \mapsto 2x^2} \underset{[16]}{y = 32} \xrightarrow[[y \mapsto -\sin(y)]]{y \mapsto \cos(y)} \underset{[-\sin(32)]}{z = \cos(32)}$

$$\times$$

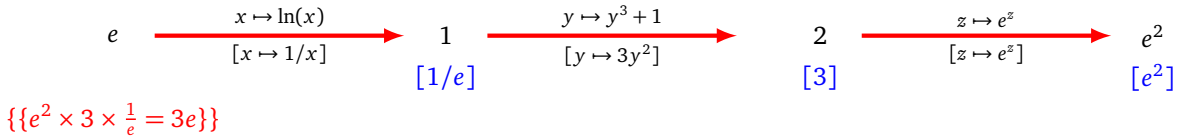
$$\{\{-16\sin(32)\}\}$$

On trouve $F'(4) = -16\sin(32)$.

Plus de fonctions. Le principe est le même sur des exemples plus compliqués, ici avec trois fonctions : $F = h \circ g \circ f$.



Qui donne :

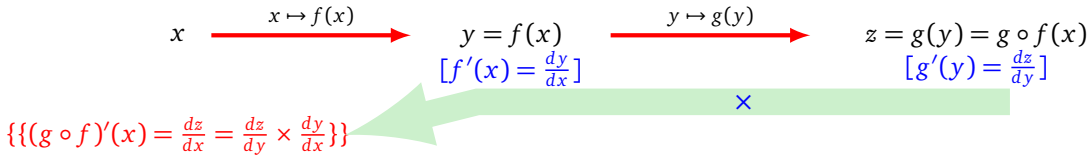


Répondre aux questions suivantes :

- Quelle est la fonction composée ?
- Quelle est la dérivée calculée ? Combien vaut-elle ? Vérifier les calculs !
- Combien vaudrait la dérivée en $x = 1$? En $x = 2$?

Justification. Pourquoi ces opérations donnent-elles le bon résultat ? Il s'agit juste de la réécriture de la formule de la dérivée d'une composition !
Pour $F = g \circ f$, les dérivées locales sont $f'(x)$ et $g'(y)$ (en notant $y = f(x)$), le produit des dérivées locales est donc $f'(x) \times g'(y) = f'(x) \times g'(f(x)) = F'(x)$. C'est encore plus facile à comprendre en notant $y = f(x)$, $z = g(y) = g \circ f(x)$: il s'agit juste de la formule :

$$\frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}x}.$$

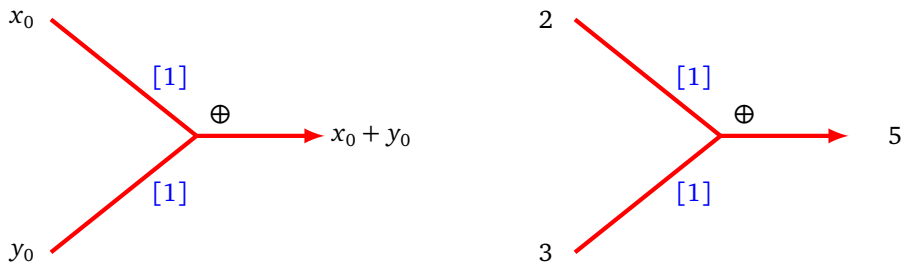


Pour des compositions plus compliquées, par exemple avec trois compositions, c'est la réécriture de la formule :

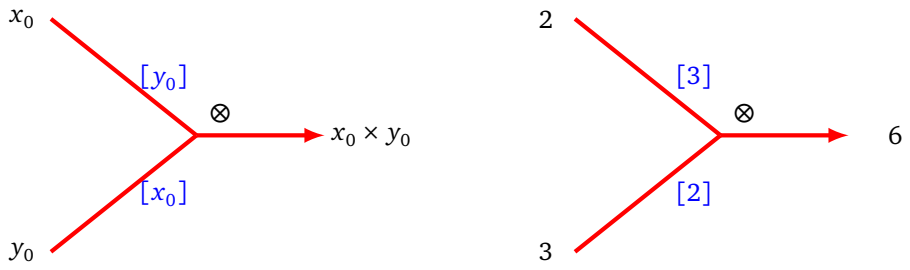
$$\frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\mathrm{d}F}{\mathrm{d}z} \times \frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x}.$$

2.3. Dérivation automatique (cas général)

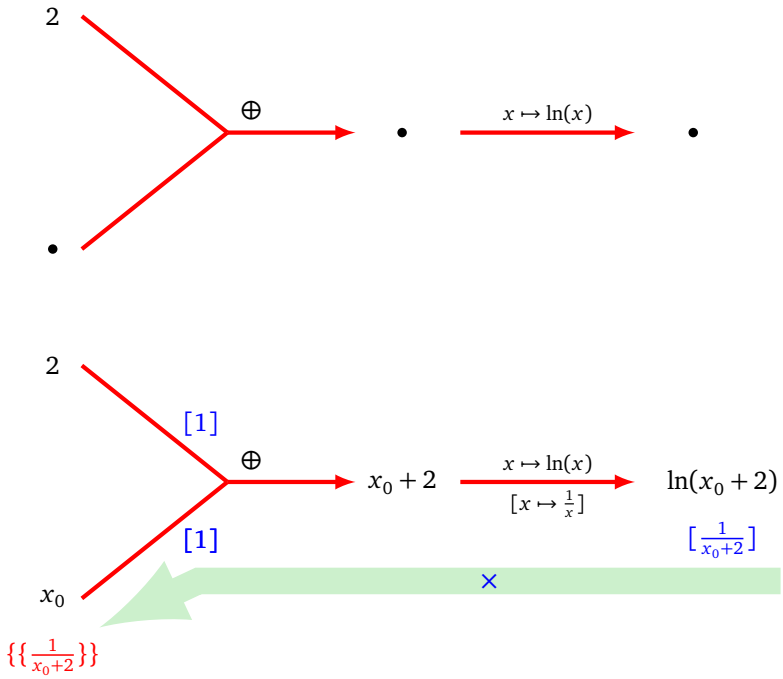
Pour les additions et multiplications, on ajoute des dérivées locales sur les arêtes. Les formules seront justifiées dans le chapitre « Gradient ». Pour l'addition, les deux dérivées locales valent 1 (le principe à gauche ci-dessous, un exemple à droite).



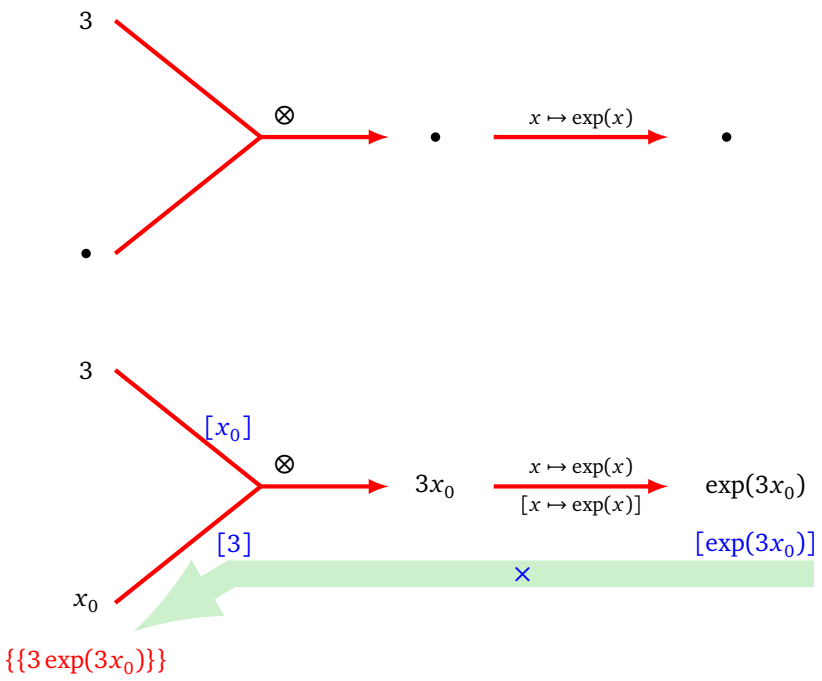
Pour la multiplication, les dérivées locales correspondent aux valeurs d'entrée x_0 et y_0, mais échangées : $[y_0]$ et $[x_0]$ (le principe à gauche ci-dessous, un exemple à droite).



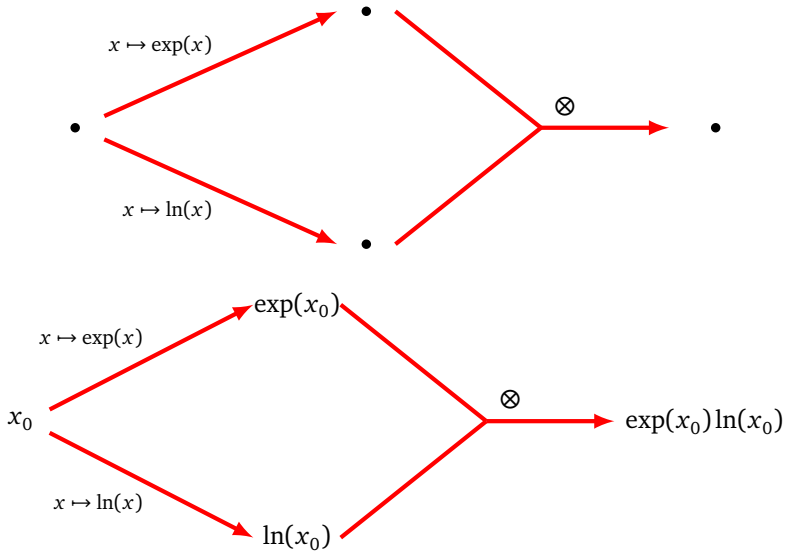
Pour calculer la dérivée d'une composition, le principe est le même qu'auparavant : la dérivée est le produit des dérivées locales le long des arêtes joignant la valeur x_0 à la valeur $F(x_0)$.
Par exemple, voici le graphe de calcul correspondant à $F(x) = \ln(x+2)$, on trouve $F'(x) = \frac{1}{x+2}$.



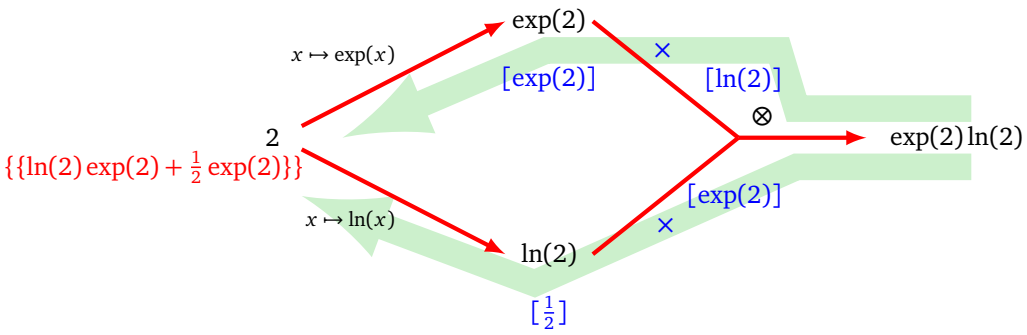
Voici le graphe de calcul correspondant à $F(x) = \exp(3 \times x)$, on trouve $F'(x) = 3\exp(3x)$.



Il existe une nouvelle situation, lorsque l'on a besoin de dupliquer la variable d'entrée. Considérons par exemple la fonction $F(x) = \exp(x) \times \ln(x)$. L'entrée « • » tout à gauche (sur la figure ci-dessous) représente une même valeur x_0 utilisée deux fois.



Voici le graphe de calcul correspondant à $F(2)$ pour l'entrée $x_0 = 2$, complété par les dérivées locales (entre crochets) et la dérivée globale par rapport à la variable d'entrée (entre accolades doubles) :



La dérivée $F'(2)$ est la somme de termes, calculée entre accolades :

$$F'(2) = \{\{\ln(2)\exp(2) + \frac{1}{2}\exp(2)\}\} = \left(\ln(2) + \frac{1}{2}\right)e^2.$$

Bilan.

> Pour calculer la dérivée d'une fonction composée F en x_0 :
> - pour chaque entrée, calculer le produit des dérivées locales (qui sont entre crochets) le long des chemins allant de la valeur finale $F(x_0)$ en revenant vers l'entrée,
> - puis, calculer la somme de ces produits (écrite entre accolades doubles).

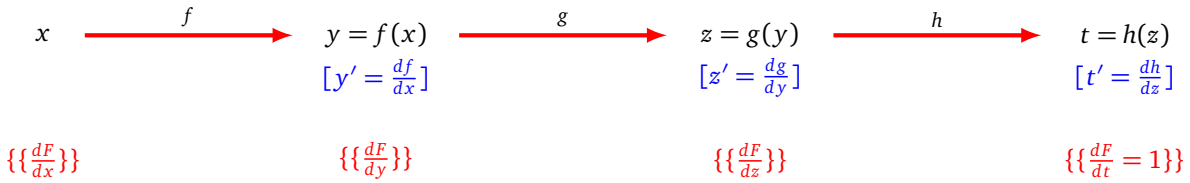
2.4. Dérivées séquentielles

Cette section est plus technique et peut être omise en première lecture.

Soit $F(x) = h \circ g \circ f(x)$ une composition de trois fonctions. Notons

$$y = f(x), \qquad z = g(y), \qquad t = h(z).$$

On appelle ***dérivées séquentielles***, les dérivées de F par rapport à chacune des variables (ici x, y, z, t).



Dans ce graphe de calcul, les dérivées locales sont entre crochets. Les dérivées séquentielles sont entre accolades doubles (lues en partant de la droite à partir de la figure ci-dessus) :

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \frac{\mathrm{d}t}{\mathrm{d}t} = 1, \qquad \frac{\mathrm{d}F}{\mathrm{d}z} = \frac{\mathrm{d}t}{\mathrm{d}z} = [t'],$$

$$\frac{\mathrm{d}F}{\mathrm{d}y} = \frac{\mathrm{d}t}{\mathrm{d}y} = \frac{\mathrm{d}t}{\mathrm{d}z} \times \frac{\mathrm{d}z}{\mathrm{d}y} = [t'] \times [z'], \qquad \frac{\mathrm{d}F}{\mathrm{d}x} = \frac{\mathrm{d}t}{\mathrm{d}x} = \frac{\mathrm{d}t}{\mathrm{d}z} \times \frac{\mathrm{d}z}{\mathrm{d}y} \times \frac{\mathrm{d}y}{\mathrm{d}x} = [t'] \times [z'] \times [y'].$$

On calcule donc facilement la dérivée $\frac{\mathrm{d}F}{\mathrm{d}x}$ mais on a aussi comme étapes intermédiaires $\frac{\mathrm{d}F}{\mathrm{d}y}$ et $\frac{\mathrm{d}F}{\mathrm{d}z}$.
Pourquoi est-il pertinent de calculer les dérivées séquentielles de la droite vers la gauche ? Sur l'exemple ci-dessus les dérivées séquentielles sont :

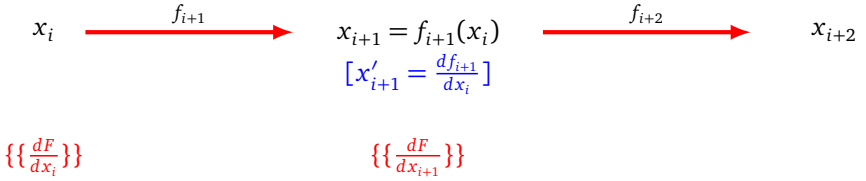
$$\{\{1\}\}, \qquad \{\{t'\}\}, \qquad \{\{t' \times z'\}\}, \qquad \{\{t' \times z' \times y'\}\}.$$

Pour éviter de calculer beaucoup de produits, on remarque que chaque nouveau produit est la multiplication de la dérivée séquentielle précédente avec la dérivée locale :

$$\{\{1\}\}, \qquad \{\{t'\}\} = \{\{1\}\} \times [t'], \qquad \{\{t' \times z'\}\} = \{\{t'\}\} \times [z'], \qquad \{\{t' \times z' \times y'\}\} = \{\{t' \times z'\}\} \times [y'].$$

Plus généralement, si $F = f_n \circ \cdots \circ f_2 \circ f_1$ et $x_1 = f_1(x_0)$, $x_2 = f_2(x_1)$,..., $x_n = f_n(x_{n-1})$ alors :

$$\left\{\left\{\frac{\mathrm{d}F}{\mathrm{d}x_i}\right\}\right\} = \left\{\left\{\frac{\mathrm{d}F}{\mathrm{d}x_{i+1}}\right\}\right\} \times \left[\frac{\mathrm{d}f_{i+1}}{\mathrm{d}x_i}\right].$$



Pour obtenir toutes les dérivées séquentielles $\{\{\frac{\mathrm{d}F}{\mathrm{d}x_i}\}\}$, il y a donc n produits à calculer en partant de $\{\{\frac{\mathrm{d}F}{\mathrm{d}x_n}\}\}$ et en terminant par $\{\{\frac{\mathrm{d}F}{\mathrm{d}x_0}\}\}$. Si on n'utilise pas cette formule de récurrence alors on a $\frac{n(n+1)}{2}$ produits à calculer, ce qui est beaucoup plus si n est grand.

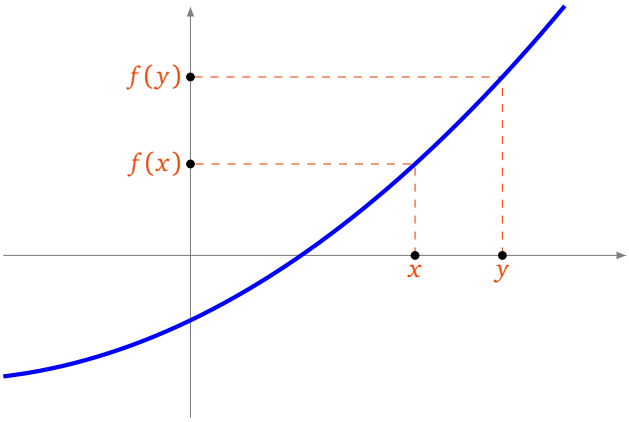
3. Étude de fonctions

3.1. Variations

Soit une fonction $f : I \to \mathbb{R}$.

Définition.
Pour tout x et y de I :

- f est ***croissante*** si $x \leqslant y$ implique $f(x) \leqslant f(y)$.
- f est ***strictement croissante*** si $x < y$ implique $f(x) < f(y)$.
- f est ***décroissante*** si $x \leqslant y$ implique $f(x) \geqslant f(y)$.
- f est ***strictement décroissante*** si $x < y$ implique $f(x) > f(y)$.



La croissance et la dérivée sont liées par la proposition suivante.

Proposition 3.
Soit $f : I \to \mathbb{R}$ une fonction dérivable, définie sur un intervalle I. Alors :

- *f est croissante sur I $\iff$ $f'(x) \geqslant 0$ pour tout $x \in I$,*
- *f est décroissante sur I $\iff$ $f'(x) \leqslant 0$ pour tout $x \in I$,*
- *f est constante sur I $\iff$ $f'(x) = 0$ pour tout $x \in I$.*

On peut améliorer un peu certains résultats : par exemple si $f'(x) > 0$ pour tout $x \in I$ alors f est strictement croissante (la réciproque est fausse comme le montre l'exemple défini par $f(x) = x^3$).

Démonstration. Nous allons seulement prouver l'implication : f est croissante $\implies$ $f'(x) \geqslant 0$. Nous admettrons les autres propriétés.
Soit f une fonction croissante. Fixons x_0 et soit $x \geqslant x_0$. Alors comme f est croissante on a $f(x) \geqslant f(x_0)$. Ainsi le taux d'accroissement $\frac{f(x)-f(x_0)}{x-x_0}$ est positif, pour tout $x > x_0$. Ainsi à la limite (lorsque $x \to x_0$ en ayant $x > x_0$) on a : $f'(x_0) = \lim_{x \to x_0^+} \frac{f(x)-f(x_0)}{x-x_0} \geqslant 0$.

□

Définition.
Une fonction $f : I \to J$ est une ***bijection*** si chaque $y \in J$ admet un unique antécédent $x \in I$, c'est-à-dire tel que $f(x) = y$.

Proposition 4.
Soit $f : I \to \mathbb{R}$ une fonction. Soit $J = f(I)$ l'ensemble des valeurs prises par f sur I. Si f est strictement croissante sur I (ou bien strictement décroissante sur I) alors la fonction $f : I \to J$ est bijective.

Démonstration. Par définition de J, tout $y \in J$ admet un antécédent $x \in I$ par f. Comme f est strictement croissante alors cet antécédent est unique. □

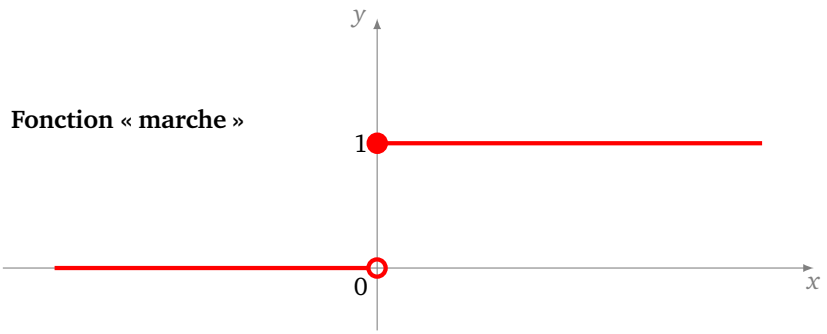
3.2. Fonctions d'activation

Une ***fonction d'activation*** n'est rien d'autre qu'une fonction $f : \mathbb{R} \to \mathbb{R}$ qui sera dans la suite associée à un neurone. Nous allons en étudier plusieurs exemples. À une ***entrée*** $x \in \mathbb{R}$, on associe une ***sortie*** $y = f(x)$.

Fonction marche de Heaviside.

C'est la fonction ***marche d'escalier*** définie par la formule suivante :

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



Cela justifie le nom de fonction d'activation. Imaginons un composant électronique qui allume une diode lorsque l'intensité est positive. On modélise la situation ainsi :

- $x \in \mathbb{R}$ est l'intensité donnée en entrée,
- $y = 0$ (diode éteinte) ou $y = 1$ (diode allumée) est la valeur de sortie,
- la relation entre les deux est donnée par la fonction de Heaviside $y = H(x)$: la diode est activée lorsque l'intensité x est positive.

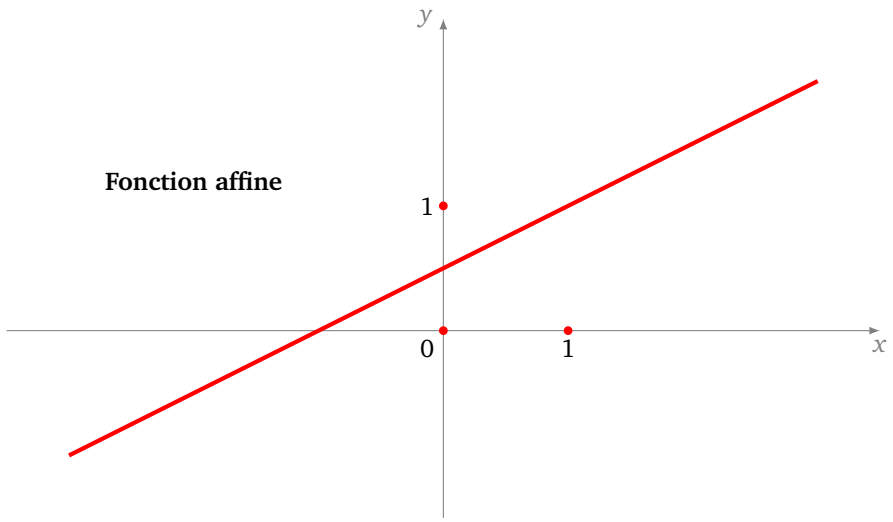
Remarque.
Quelques caractéristiques de H :

- la fonction a l'avantage d'être très simple et de séparer clairement les résultats en deux catégories (0 et 1),
- c'est aussi un inconvénient car il n'y a pas de nuance possible (du genre « plutôt oui/plutôt non »),
- la fonction H n'est pas dérivable en 0 (ni même continue), ce qui sera gênant pour la suite,
- avoir posé $H(0) = 1$ est un choix arbitraire, on aurait pu choisir une autre valeur comme 0 (certains préfèrent la valeur $\frac{1}{2}$).

Fonction affine.

Une ***fonction affine*** est définie par :

$$f(x) = ax + b$$



Remarque.

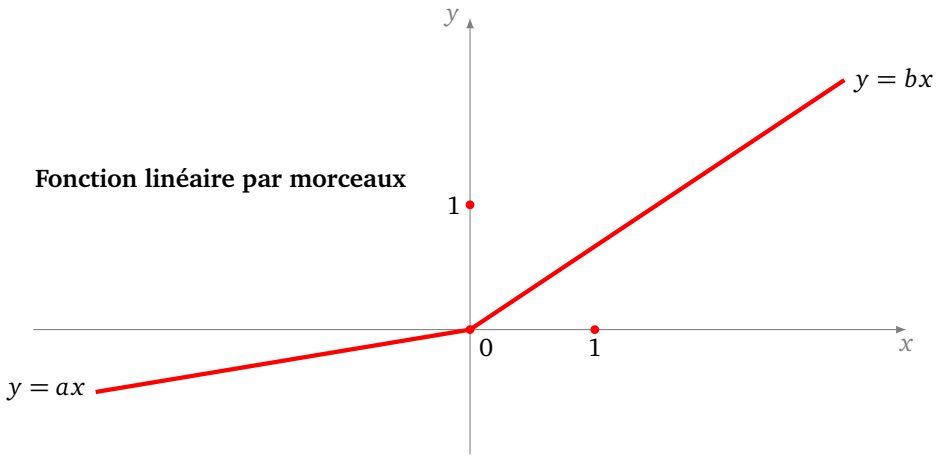
- Les fonctions affines seront à la base des réseaux de neurones. Elles sont simples.
- Si $a = 1$ et $b = 0$ alors $f(x) = x$, c'est la fonction identité : la sortie est égale à l'entrée. Nous l'utiliserons plus tard dans certains cas.
- Si $b = 0$ la sortie est proportionnelle à l'entrée.
- Le principal inconvénient est le suivant : si $f(x) = ax + b$ et $g(x) = cx + d$ sont deux fonctions affines alors la composition $h = g \circ f$ est encore une fonction affine. Ainsi, avec des fonctions d'activation qui sont affines, par composition on n'obtiendra que des fonctions affines, ce qui n'est pas assez riche pour les problèmes de classification.

ReLU et fonctions linéaires par morceaux.

Une fonction définie par

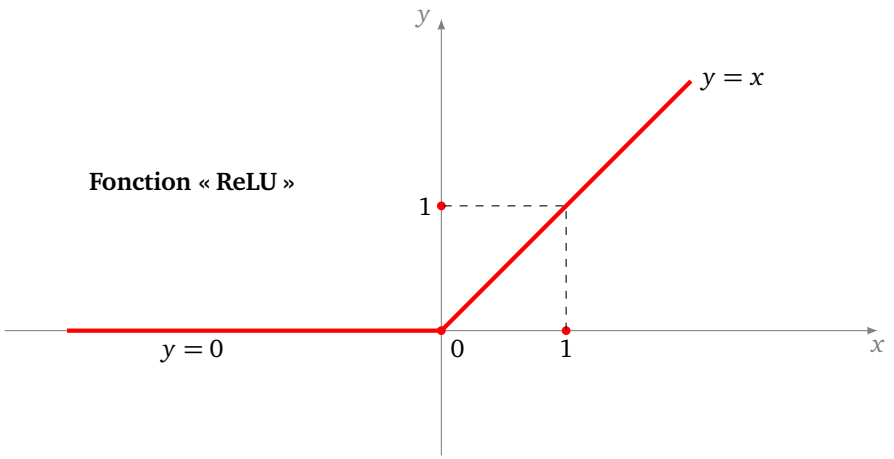
$$\begin{cases} f(x) = ax & \text{si } x < 0 \\ f(x) = bx & \text{si } x \geqslant 0 \end{cases}$$

est un exemple de ***fonction linéaire par morceaux*** (elle est linéaire à gauche et linéaire à droite).



La plus utilisée est la fonction ReLU (pour *Rectified Linear Unit*), définie par

$$\begin{cases} f(x) = 0 & \text{si } x < 0 \\ f(x) = x & \text{si } x \geqslant 0 \end{cases}$$



Remarque.

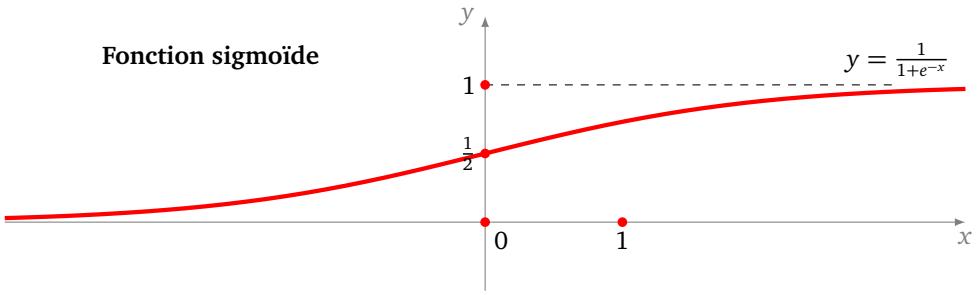
- Attention ces fonctions ne sont pas dérivables en 0 (sauf dans le cas $a = b$).
- La fonction ReLU conjugue les bénéfices d'une fonction continue, avec une activation (ici la sortie est non nulle pour une entrée strictement positive) et d'une sortie proportionnelle à l'entrée (pour les entrées positives).

La fonction sigmoïde.

La *fonction sigmoïde* est définie par :

$$\sigma(x) = \frac{1}{1+e^{-x}}.$$

Le graphe de la fonction σ possède une forme de « sigma » comme le symbole intégrale $\int$.

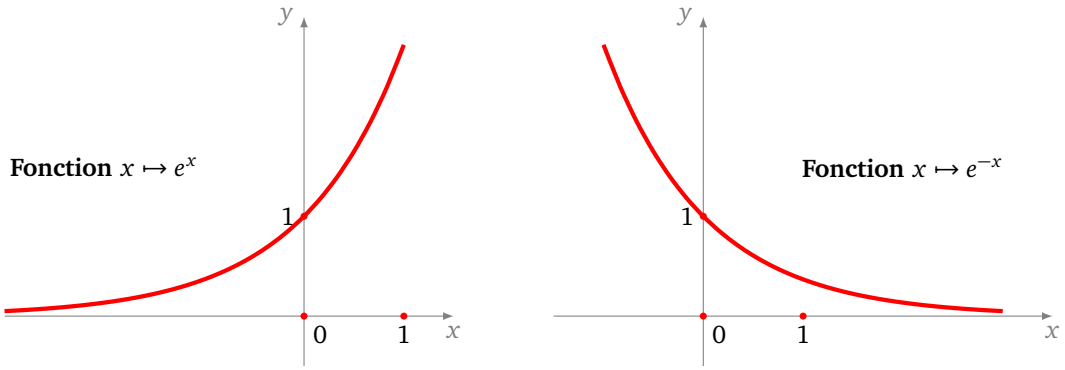


La fonction ressemble un peu à la fonction de Heaviside mais elle à l'avantage d'être continue et dérivable. Elle propose une transition douce de 0 à 1, la transition étant assez linéaire autour de 0.
Étudions en détails cette fonction.

Proposition 5.

1. *La fonction σ est strictement croissante.*
2. *La limite en $-\infty$ est 0, la limite en $+\infty$ en 1.*
3. *La fonction σ est continue et dérivable et*
$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$
4. *La tangente en $x = 0$ a pour équation $y = \frac{1}{4}x + \frac{1}{2}$.*
5. *La fonction dérivée vérifie aussi la relation :*
$$\boxed{\sigma'(x) = \sigma(x)(1-\sigma(x)).}$$

La dernière relation montre un autre avantage de la fonction σ : si on a calculé $\sigma(x_0)$ alors, sans effort supplémentaire, on sait calculer $\sigma'(x_0)$.
Pour mémoire, on rappelle l'allure des graphes des fonctions $x \mapsto e^x$ (à gauche) et $x \mapsto e^{-x}$ (à droite) en se souvenant que $e^{-x} = 1/e^x$.



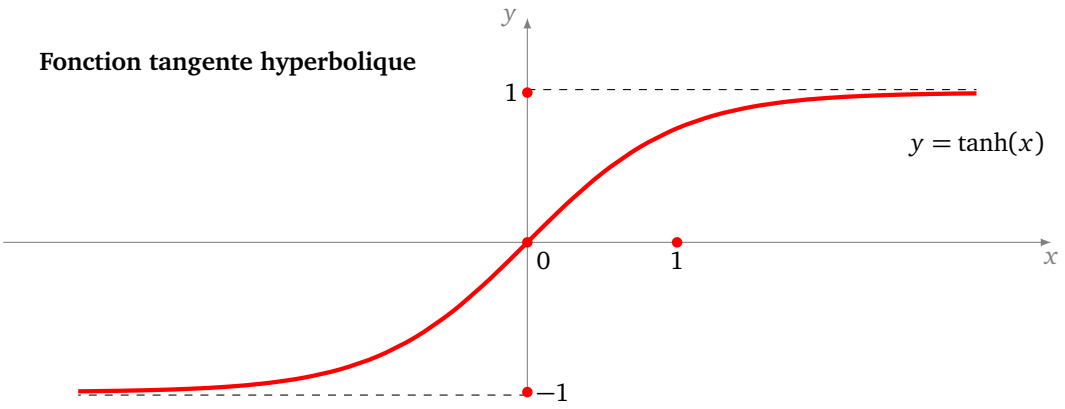
Démonstration.

1. La fonction $x \mapsto e^{-x}$ est strictement décroissante, donc la fonction $x \mapsto 1+e^{-x}$ l'est aussi. Ainsi son inverse $x \mapsto \sigma(x) = 1/(1+e^{-x})$ est une fonction strictement croissante.
2. e^{-x} tend vers 0 en $+\infty$ donc $\sigma(x) = 1/(1+e^{-x})$ tend vers 1 lorsque x tend vers $+\infty$. De même e^{-x} tend vers $+\infty$ en $-\infty$ donc $\sigma(x)$ tend vers 0 lorsque x tend vers $-\infty$.
3. On calcule la dérivée de σ par la formule de la dérivée d'un quotient. Comme la dérivée est strictement positive cela prouve à nouveau que σ est strictement croissante.
4. Au passage, on a $\sigma'(0) = \frac{1}{4}$. La tangente en $x = 0$ a donc pour équation $y = \frac{1}{4}x + \frac{1}{2}$. Et, anticipant sur la suite, on calcule que la dérivée seconde $\sigma''(0) = 0$. La courbe possède donc un point d'inflexion en $x = 0$, ce qui justifie le caractère très linéaire du graphe autour de $x = 0$.
5. Enfin :
$$\sigma(x)\big(1-\sigma(x)\big) = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma'(x).$$

□

La fonction tangente hyperbolique.
La ***tangente hyberbolique*** est définie par :

$$\text{th}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$



Elle a des propriétés similaires à la fonction sigmoïde mais varie de -1 à $+1$. C'est une fonction impaire : son graphe est symétrique par rapport à l'origine. On pourrait refaire une étude complète de cette fonction mais celle-ci est liée à celle de la fonction σ.

Proposition 6.

$$\boxed{\text{th}(x) = 2\sigma(2x) - 1.}$$

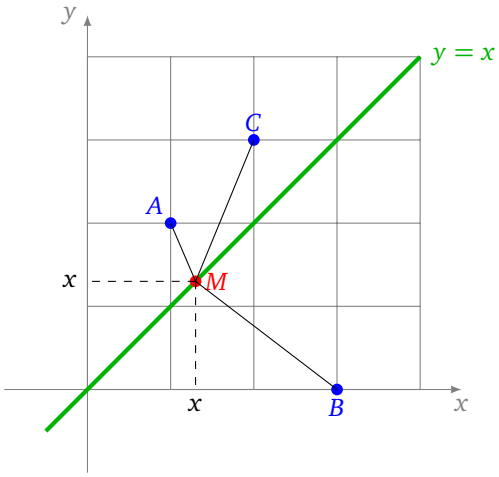
Démonstration.

$$2\sigma(2x) - 1 = \frac{2}{1+e^{-2x}} - 1 = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \text{th}(x).$$

L'avant-dernière égalité s'obtient en multipliant numérateur et dénominateur par e^x. □

3.3. Minimum et maximum

Pour configurer un réseau de neurones, il faut fixer des paramètres réels (appelés *poids*). Les bons paramètres sont ceux qui permettent au réseau de répondre correctement à la problématique (du genre « Est-ce une photo de chat ? »). Il n'y a pas de méthode directe pour trouver immédiatement quels sont les meilleurs paramètres. Ceux-ci sont obtenus par essais/erreurs, nous y reviendrons.
Pour le moment, la question est de savoir : que sont de « bons » paramètres ?
Voyons le cas d'un seul paramètre avec le problème suivant : on se donne trois points du plan $A(1,2)$, $B(3,0)$ et $C(2,3)$. On cherche le point $M(x,x)$ qui a la contrainte d'être sur la droite d'équation $(y = x)$ et qui approche au mieux les trois points A, B et C.



Il faut préciser ce que l'on entend par « approche au mieux ». Nous souhaitons que la somme des carrés des distances entre M et chacun des points soit la plus petite possible. Nous allons définir une fonction d'erreur et chercher le point M qui minimise cette erreur. Définissons la fonction d'erreur :

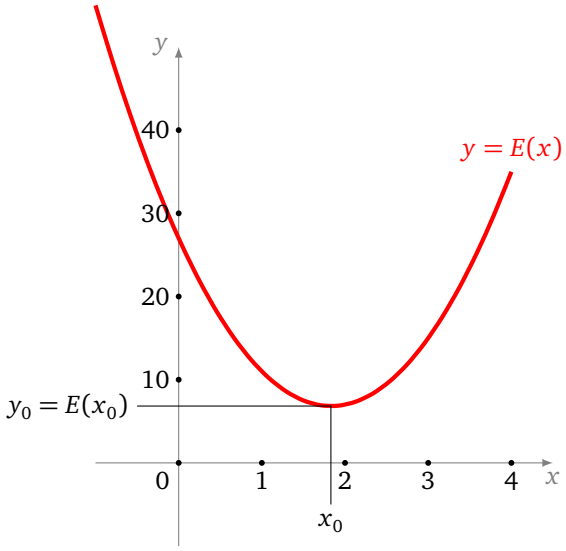
$$E(x) = AM^2 + BM^2 + CM^2.$$

La formule algébrique est la suivante :

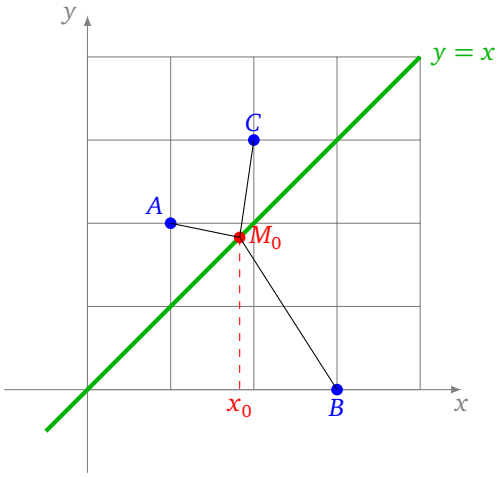
$$E(x) = \big((x-1)^2 + (x-2)^2\big) + \big((x-3)^2 + (x-0)^2\big) + \big((x-2)^2 + (x-3)^2\big).$$

Il s'agit de trouver la valeur x de sorte que $E(x)$ soit le plus petit possible. Pour cela, nous pouvons développer les termes de $E(x)$ et réécrire :

$$E(x) = ax^2 + bx + c \quad \text{avec } a = 6,\ b = -22,\ c = 27.$$



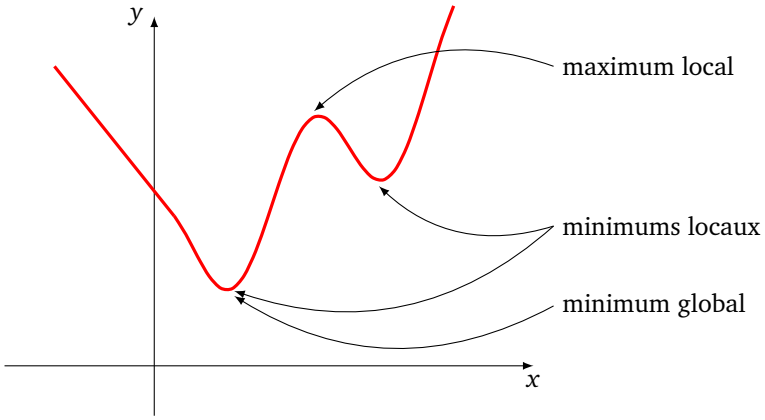
La courbe d'erreur est donc une parabole, la valeur la plus petite possible est atteinte en $x_0 = -\frac{b}{2a} = \frac{22}{12} = 1.83\ldots$ qui correspond au sommet de la parabole. Pour cette valeur x_0, on trouve $E(x_0) = \frac{41}{6} = 6.83\ldots$ et tout autre paramètre $x \in \mathbb{R}$ vérifie $E(x) \geqslant E(x_0)$.
Voici la configuration minimale avec $M_0 = (x_0, x_0)$.



De façon plus générale nous avons les définitions suivantes.

Définition.
Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction.

- f atteint un ***minimum global*** en $x_0 \in \mathbb{R}$ si pour tout $x \in \mathbb{R}$, on a $f(x) \geqslant f(x_0)$.
- f atteint un ***minimum local*** en $x_0 \in \mathbb{R}$ si il existe un intervalle ouvert I, contenant x_0 et tel que pour tout $x \in I$, on a $f(x) \geqslant f(x_0)$.

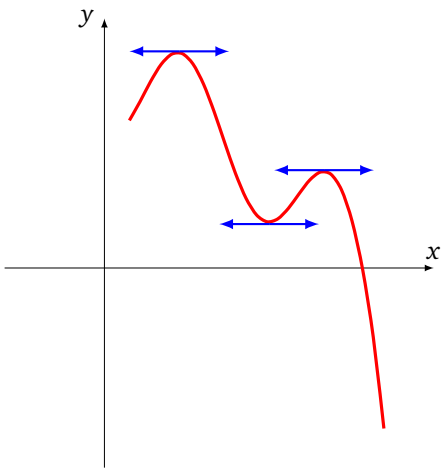


- On définirait de façon analogue un ***maximum global*** et un ***maximum local***.
- Comme son nom l'indique, pour savoir si on a un minimum local, on ne regarde que ce qui se passe autour de x_0 (l'intervalle I peut être très petit autour de x_0).
- Un minimum global est aussi un minimum local, mais l'inverse n'est pas vrai en général.

Comme on l'a vu, trouver la meilleure solution d'un problème, correspond à trouver le minimum global d'une fonction. C'est un problème difficile en général. Par contre, il est facile de trouver des valeurs qui sont les candidats à être des minimums locaux.

Proposition 7.
Si f atteint un minimum local ou un maximum local en x_0 alors $f'(x_0) = 0$.

On appellera ***point critique*** une valeur x_0 vérifiant $f'(x_0) = 0$. Voici l'interprétation géométrique : si f atteint un minimum local ou un maximum local en x_0, alors la tangente au graphe en x_0 est horizontale.



Attention ! La réciproque de cette proposition n'est pas vraie. Il se peut que $f'(x_0) = 0$ mais que f n'atteigne ni un minimum local, ni un maximum local en x_0. Nous y reviendrons.

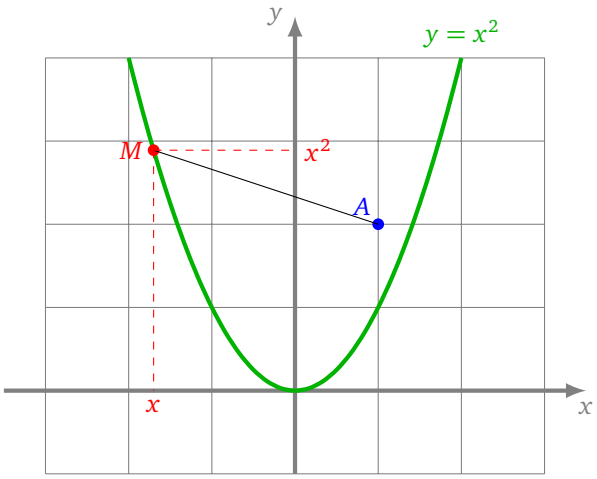
Démonstration. Prenons le cas d'un minimum local en x_0.

- Pour x proche de x_0 avec $x > x_0$, on a $f(x_0) \leqslant f(x)$, car en x_0 on atteint un minimum local. Donc le taux d'accroissement $\frac{f(x)-f(x_0)}{x-x_0}$ est positif. En passant à la limite, lorsque $x \to x_0$ avec $x > x_0$, on obtient que $f'(x_0) \geqslant 0$.
- Pour x proche de x_0 avec $x < x_0$, le taux d'accroissement $\frac{f(x)-f(x_0)}{x-x_0}$ est négatif (car cette fois $x - x_0 < 0$), à la limite, lorsque $x \to x_0$ avec $x < x_0$, on obtient que $f'(x_0) \leqslant 0$.
- Conclusion : $f'(x_0) \geqslant 0$ et $f'(x_0) \leqslant 0$, donc $f'(x_0) = 0$.

□

Reprenons l'exemple de notre fonction $E(x) = 6x^2 - 22x + 27$, et plus généralement celui d'une fonction $E(x) = ax^2 + bx + c$, alors $E'(x) = 2ax + b$. Le seul point en lequel la dérivée s'annule est $x_0 = -\frac{b}{2a}$ qui correspond bien au sommet de la parabole et est la valeur en laquelle la fonction E atteint son minimum (ou bien son maximum si on avait $a < 0$). Sur cet exemple le minimum local est aussi un minimum global. On termine avec un autre exemple un peu plus sophistiqué.

> **Exemple.**
> Quel point de la parabole d'équation $(y = x^2)$ est le plus près du point A de coordonnées $(1, 2)$?



Les coordonnées d'un point M appartenant à cette parabole sont de la forme (x, x^2). La distance entre A et M est donc :

$$AM = \sqrt{(x-1)^2 + (x^2-2)^2}.$$

Mais trouver le plus petit AM équivaut à trouver le plus petit AM^2. On définit donc comme fonction à minimiser

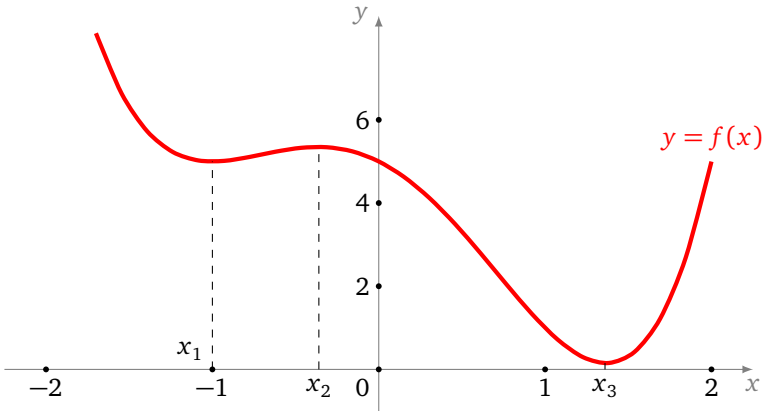
$$f(x) = AM^2 = (x-1)^2 + (x^2-2)^2 = x^4 - 3x^2 - 2x + 5.$$

Étudions cette fonction f.

- Tout d'abord $f'(x) = 4x^3 - 6x - 2$.
- Il faut chercher les valeurs en lesquelles f s'annule. Ici il se trouve que $f'(-1) = 0$, on peut donc factoriser $f'(x) = 2(x+1)(2x^2 - 2x - 1)$. Ainsi $f'(x)$ s'annule en deux autres valeurs, les solutions de $2x^2 - 2x - 1 = 0$, c'est-à-dire en

$$x_1 = -1, \quad x_2 = \frac{1-\sqrt{3}}{2}, \quad x_3 = \frac{1+\sqrt{3}}{2}.$$

- On peut dresser le tableau de variations de f puis tracer son graphe.



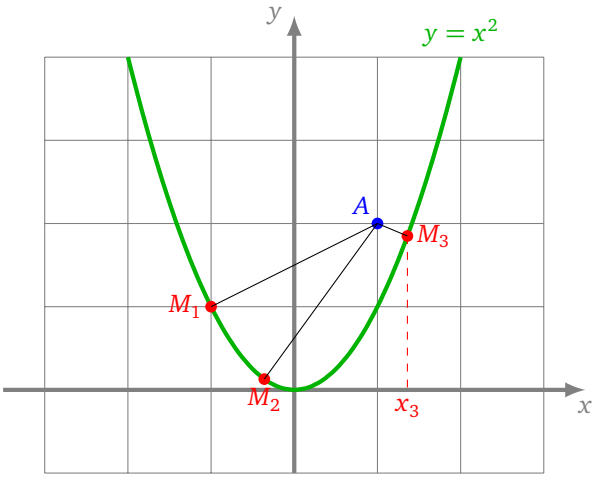
- Maximum : f admet un maximum local en x_2, mais f n'a pas de maximum global.
- Minimums : f admet un minimum local en x_1 et x_3. L'un des deux est un minimum global, mais lequel ? Il suffit de comparer $f(x_1)$ et $f(x_3)$:

$$f(x_1) = 5, \qquad f(x_3) = \frac{11 - 6\sqrt{3}}{4}.$$

Ainsi le minimum global est atteint en x_3 avec

$$x_3 = \frac{1+\sqrt{3}}{2} = 1.36\ldots \quad \text{et} \quad f(x_3) = 0.15\ldots$$

- Conclusion : le point de la parabole le plus proche du point A est le point M_3 d'abscisse x_3 et d'ordonnée x_3^2.



- Remarque importante. Noter que le point M_1 (correspondant à x_1) est moins éloigné de A que ses voisins proches, mais que ce n'est pas lui la solution du problème : M_1 correspond à un minimum local, mais pas un minimum global.

4. Dérivée seconde

Nous avons vu que le nombre dérivé $f'(x_0)$ permet d'approcher les valeurs de $f(x)$ pour x proche de x_0. La dérivée seconde permet de faire un peu mieux.

4.1. Définition

Soit $f : I \to \mathbb{R}$ une fonction dérivable, c'est-à-dire telle que $f'(x)$ existe pour tout $x \in I$. Nous notons f' ou $x \mapsto f'(x)$ ou $\frac{\mathrm{d}f}{\mathrm{d}x}$ la fonction dérivée.

Définition.
La ***dérivée seconde*** en x_0 (si elle existe) est la dérivée de la fonction f' en x_0, c'est-à-dire :

$$(f')'(x_0).$$

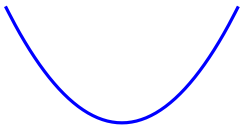
Nous noterons ce nombre $f''(x_0)$ ou bien $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x_0)$.

Si la dérivée seconde existe pour tout $x \in I$, on obtient une fonction dérivée seconde notée f'' ou $x \mapsto f''(x)$ ou $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$.

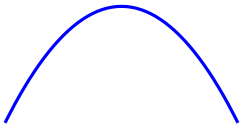
Exemple.
- Soit $f(x) = x^3$. Calculons $f''(-2)$. On sait que $f'(x) = 3x^2$. La dérivée de la fonction f' est la fonction f'' définie par $f''(x) = 6x$. Donc $f''(-2) = -12$.
- Soit $f(x) = \sin(x)$ alors $f'(x) = \cos(x)$ et la dérivée seconde est $f''(x) = -\sin(x)$.
- Soit $f(x) = \exp(x^2)$, alors $f'(x) = 2x\exp(x^2)$ et $f''(x) = (2 + 4x^2)\exp(x^2)$.

4.2. Concavité et minimum local

Le signe de la dérivée seconde informe sur l'allure du graphe : si la dérivée seconde est positive sur un intervalle, alors la fonction y est convexe. La courbe est en forme de ∪. Le modèle est $f(x) = x^2$ (dont la dérivée seconde est $f''(x) = 2$, qui est bien positive). Si la dérivée seconde est négative sur un intervalle, alors la fonction y est concave. La courbe est en forme de ∩. Le modèle est $f(x) = -x^2$ (avec $f''(x) = -2$).



fonction convexe, $f''(x) > 0$



fonction concave, $f''(x) < 0$

Cela va nous aider à déterminer si un point en lequel la dérivée s'annule, c'est-à-dire un point critique, est un minimum local ou un maximum local.

Proposition 8.

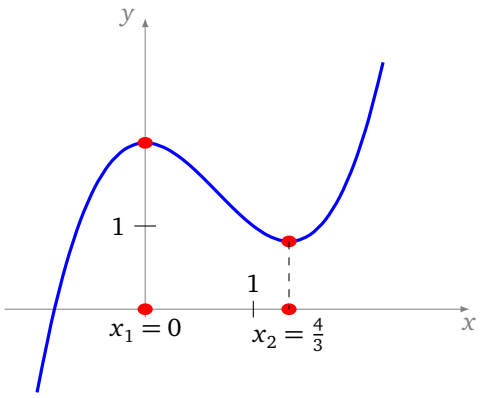
- *Si* $f'(x_0) = 0$ *et* $f''(x_0) > 0$ *alors* f *atteint un minimum local en* x_0.
- *Si* $f'(x_0) = 0$ *et* $f''(x_0) < 0$ *alors* f *atteint un maximum local en* x_0.

Si $f'(x_0) = 0$ et $f''(x_0) = 0$ alors on ne peut rien dire en général. Il faut étudier chaque fonction au cas par cas car tout peut arriver : un minimum local, un maximum local ou bien ni l'un ni l'autre.

Exemple.
Cherchons les minimums et maximums locaux de la fonction f définie par :

$$f(x) = x^3 - 2x^2 + 2.$$

- **Points critiques.** On calcule $f'(x) = 3x^2 - 4x = x(3x-4)$ et on cherche les valeurs en lesquelles f' s'annule. Ici f' s'annule en $x_1 = 0$ et en $x_2 = 4/3$.
- **Dérivée seconde en x_1.** On calcule $f''(x) = 6x - 4$. En x_1 on a $f''(x_1) = f''(0) = -4$, donc par la proposition **??**, f admet un maximum local en x_1.
- **Dérivée seconde en x_2.** En x_2 on a $f''(x_2) = f''(\frac{4}{3}) = +4$, donc par la proposition **??**, f admet un minimum local en x_2.



Noter que f n'admet aucun maximum global ni aucun minimum global sur $\mathbb{R}$.

Démonstration. Prenons l'exemple d'un x_0 tel que $f'(x_0) = 0$ et $f''(x_0) > 0$. Nous supposons dans ce cours que les fonctions sont suffisamment régulières pour nos besoins. Ici nous supposons donc que f'' est continue, cela implique que $f''(x) > 0$ pour x dans un petit intervalle ouvert $]a, b[$ contenant x_0.
On peut dresser le tableau de variation de f sur $[a, b]$:

- $f''(x)$ est positive, donc f' est strictement croissante sur $[a,b]$.
- Comme f' s'annule en x_0 alors $f'(x)$ est négatif pour x dans $[a,x_0[$ et positif pour x dans $]x_0,b]$.
- Ainsi f est décroissante sur $[a,x_0]$ puis croissante sur $[x_0,b]$.
- Donc en x_0, f atteint bien un minimum local.

□

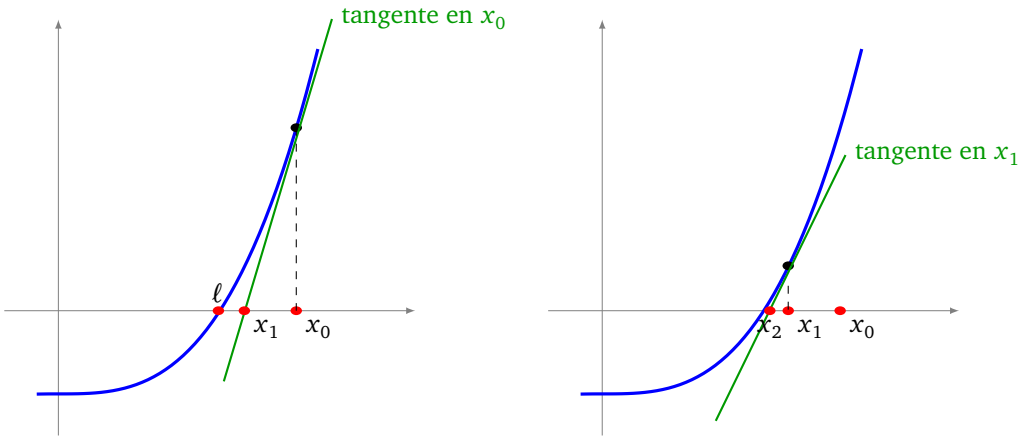
5. Méthode de Newton

La méthode de Newton est une façon efficace de trouver une solution approchée d'une équation $f(x)=0$. Dans le paramétrage d'un réseau de neurones, il ne s'agit pas de trouver une valeur en laquelle la fonction s'annule mais une valeur en laquelle la fonction d'erreur atteint son minimum. Alors à quoi bon étudier la méthode de Newton ?
Tout d'abord, pour trouver le minimum de f, il faut identifier les points en lesquels f' s'annule. Il s'agit donc de résoudre l'équation $f'(x)=0$ ce qui peut se faire par la méthode de Newton (appliquée à f'). Ensuite, la méthode de Newton consiste à suivre le graphe selon la direction de la tangente. C'est une idée fondamentale que l'on retrouvera plus tard lors de l'étude de la « descente de gradient » pour minimiser des fonctions de plusieurs variables.

5.1. Principe

Pour chercher une valeur approchée de ℓ de $[0,1]$ telle que $f(\ell)=0$, on pourrait par exemple calculer $f(0.00)$, $f(0.01)$, $f(0.02)$,...,$f(1.00)$ et choisir pour ℓ celui qui donne la valeur la plus proche de 0. C'est une méthode très lente (il faut ici évaluer 100 fois la fonction f) et peu précise (la précision sera de l'ordre 10^{-2}, soit deux chiffres exacts après la virgule).
On va voir une autre méthode plus efficace pour obtenir une valeur approchée d'une solution ℓ de $f(\ell)=0$. L'idée de la méthode de Newton est d'utiliser la tangente :

- on part d'une valeur x_0 quelconque,
- on trace la tangente au graphe de f au point d'abscisse x_0,
- cette tangente recoupe l'axe des abscisses en un point d'abscisse x_1 (figure de gauche),
- cette valeur x_1 est plus proche de ℓ que ne l'est x_0 (sous réserve que f satisfasse des hypothèses raisonnables),
- on recommence à partir de x_1 : on trace la tangente, elle recoupe l'axe des abscisses, on obtient une valeur x_2... (figure de droite).



Partons d'un x_0 quelconque et voyons comment calculer x_1, la valeur suivante de la suite. L'équation de la tangente en x_0 est :

$$y=(x-x_0)f'(x_0)+f(x_0).$$

On cherche le point (x_1, y_1) en lequel la tangente recoupe l'axe des abscisses. Un tel point vérifie donc l'équation de la tangente avec en plus $y_1 = 0$, ainsi :

$$0 = (x_1 - x_0)f'(x_0) + f(x_0).$$

Donc

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

5.2. Suite

On va ainsi définir par récurrence une suite (x_n). L'équation de la tangente en une valeur x_n est donnée par $y = f'(x_n)(x - x_n) + f(x_n)$.
En partant d'une valeur x_0, on obtient une formule de récurrence, pour $n \geqslant 0$:

$$\boxed{x_0 \quad \text{est donné} \qquad \text{et} \qquad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{pour } n \geqslant 0.}$$

Pour que cette méthode fonctionne, il faut tout de même partir d'une valeur x_0 pas trop éloignée de la solution ℓ cherchée.

Exemple.
On cherche une valeur approchée de $\sqrt[3]{100}$.

- Soit f définie par $f(x) = x^3 - 100$. L'unique valeur en laquelle f s'annule est bien le réel ℓ qui vérifie $\ell^3 = 100$, c'est-à-dire $\ell = \sqrt[3]{100}$.
- On calcule $f'(x) = 3x^2$.
- On part de $x_0 = 10$ par exemple.
- On calcule $f(x_0) = 900$ et $f'(x_0) = 300$. Par la formule de récurrence :
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 10 - \frac{900}{300} = 7$$
- On calcule $f(x_1) = f(7) = 243$, $f'(x_1) = f'(7) = 147$ et donc
$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = \frac{262}{49} = 5.3469\ldots$$
- Puis $x_3 = 4.7305\ldots$
- Puis $x_4 = 4.6432\ldots$
- Puis $x_5 = 4.6415894\ldots$ qui a déjà 5 chiffres après la virgule de corrects.
- Et $x_6 = 4.641588833612\ldots$ qui a 12 chiffres après la virgule de corrects.

Pour le plaisir et pour appréhender la vitesse extraordinaire de convergence, voici le terme x_{13} (après $n = 13$ itérations) qui donne 1000 chiffres exacts après la virgule pour $\sqrt[3]{100}$.

4.
641588833612778892410076350919446576551349125011243637650692858684777869692844826189959070897571379841543308228265404820510270287495774377362322395030214650941774267196509162954521460897633669381041162860653359655138485386961949615722782627731576754883017169207448098556934156362916689287996611195246166796700772935481246868717652590657254713373415318486044646217897776989721242854691446396378952687182014556020545505584013855637728160923375212916394808607478395557739842572739466861092267994060507040244202985417730012040741023241387966317327003410606737496780919282092017340424063011069619562088429

> 61382086140688243128977537380423154514270094830453
> 92228536191076596869380984898548880361175285974621
> 54055836006657079466872481552449410810797956289034
> 29915634432197012553305943982255082950700356942961
> 35930373637320850012819533899529732642969005258003
> 48542372295980181985171325034488774422768662481827
> 78772197705070286933786368810026273023928761425717
> 31457550215784934201074501143121549887585570334866

5.3. Algorithme

Algorithme de la méthode de Newton.

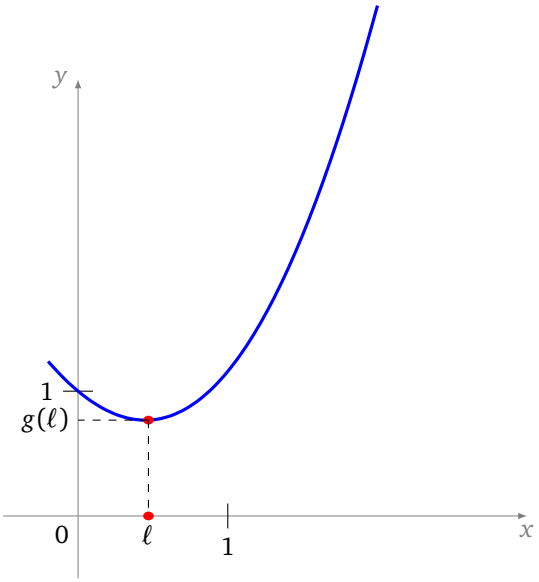
- — Entrée : une fonction f, une valeur de départ x_0, un nombre d'itérations n.
  — Sortie : une valeur approchée de ℓ tel que $f(\ell) = 0$.
- Poser $x = x_0$.
- Répéter n fois :

$$x \leftarrow x - \frac{f(x)}{f'(x)}$$

- À la fin renvoyer x (qui approche une solution ℓ).

5.4. Exemple

Objectif. Trouver la valeur de ℓ en laquelle la fonction $g(x) = x^2 - \sin(x) + 1$ atteint son minimum.



Méthode. Il n'existe pas de méthode exacte pour trouver cette solution, nous allons donner une solution approchée de ℓ.
Calculons la dérivée de f ainsi que sa dérivée seconde :

$$g'(x) = 2x - \cos(x), \qquad g''(x) = 2 + \sin(x).$$

Une étude de fonction très simple montre que : (a) $g''(x) > 0$, la dérivée est donc strictement croissante ; (b) la dérivée ne s'annule qu'en une seule valeur ℓ qui est la solution de $g'(x) = 0$ autrement dit $\cos(x) = 2x$. On pose donc $f(x) = 2x - \cos(x)$ (c'est-à-dire $f(x) = g'(x)$) et on applique la méthode de Newton à f. Partons de $x_0 = 0$. Voici la suite obtenue et l'erreur commise à chaque étape.

$$\begin{array}{l|l} x_0 = 0 & \epsilon_0 \leqslant 1 \\ x_1 = 0.5 & \epsilon_1 \leqslant 10^{-1} \\ x_2 = 0.4506266930\ldots & \epsilon_2 \leqslant 10^{-3} \\ x_3 = 0.4501836475\ldots & \epsilon_3 \leqslant 10^{-7} \\ x_4 = 0.4501836112\ldots & \epsilon_4 \leqslant 10^{-15} \end{array}$$

Conclusion. Une valeur approchée de ℓ est donc donnée par

$$\ell \simeq 0.4501836112$$

et en ce point $f(\ell) = 0$ (donc $g'(\ell) = 0$) et g atteint sa valeur minimale :

$$g(\ell) \simeq 0.7675344248.$$

Partir d'une autre valeur, par exemple $x_0 = 10$, conduit rapidement à la même solution.

$$\begin{aligned} x_0 &= 10 \\ x_1 &= -4.3127566511\ldots \\ x_2 &= -1.4932228819\ldots \\ x_3 &= 1.5615319504\ldots \\ x_4 &= 0.5235838839\ldots \\ x_5 &= 0.4511295428\ldots \\ x_6 &= 0.4501837766\ldots \\ x_7 &= 0.4501836112\ldots \end{aligned}$$