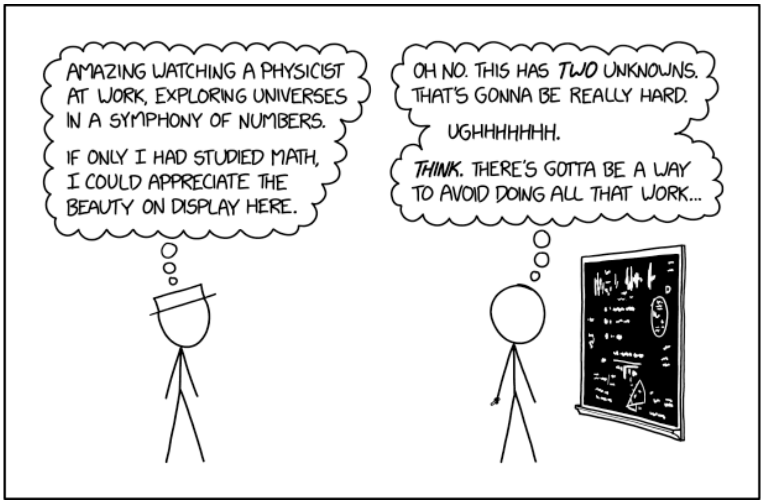
Mini-syllabus for algebra and calculus

Camille Gontier (camille.gontier@unibe.ch)

Introduction to biological and computational learning


AMAZING WATCHING A PHYSICIST AT WORK, EXPLORING UNIVERSES IN A SYMPHONY OF NUMBERS.
IF ONLY I HAD STUDIED MATH, I COULD APPRECIATE THE BEAUTY ON DISPLAY HERE.
OH NO. THIS HAS TWO UNKNOWNS. THAT'S GONNA BE REALLY HARD.
UGHHHHHHH.
THINK. THERE'S GOTTA BE A WAY TO AVOID DOING ALL THAT WORK...

https://xkcd.com/2207

Contents

1 Introduction

Students taking the class *”Introduction to Biological and Computational Learning”* are expected to have basic knowledge of calculus and algebra, and especially to know how to differentiate common functions and perform basic operations on vectors. This short syllabus is intended to make sure all students can follow and benefit from the class.

Questions and suggestions are very much welcome: camille.gontier@unibe.ch

Since a short animation is often worth a long explanation, interested readers are also encouraged to watch the following videos:

- Essence of linear algebra: https://youtu.be/fNk_zzaMoSs
- Essence of linear calculus: https://youtu.be/9vKqVkMQHKk

2 Algebra and calculus cookbook

2.1 Algebra

You are used to manipulate "classical" real numbers: 3, 0, -2, $\frac{1}{3}$, e^2, π, etc. Such numbers are called **scalars**, and you know how to add, subtract, multiply, and divide them.

We are now going to introduce another class of numbers, which are called **vectors**. A vector $\mathbf{x}$ of dimension n can be understood as n scalars put together:

$$\mathbf{x} = [x_1,\ x_2,\ ...\ x_n]$$

For instance, $\mathbf{x} = [7, -3, 2]$ is a 3-dimension vector. Scalars and vectors are simply numbers of different dimensions. A scalar is a 1-dimension vector, while vectors are an extension of scalars.

A scalar is the position of a point in a 1-dimension space (i.e. on a line):

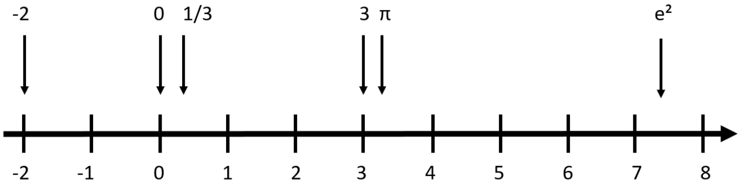


Figure 1: A scalar can be understood as a position on the line of real numbers.

A 2-dimension vector $\mathbf{x} = [x_1,\ x_2]$ is the position of a point in a 2-dimension space (i.e. on a plane); you can understand x_1 and x_2 as latitude and longitude of a point on a map, for instance:

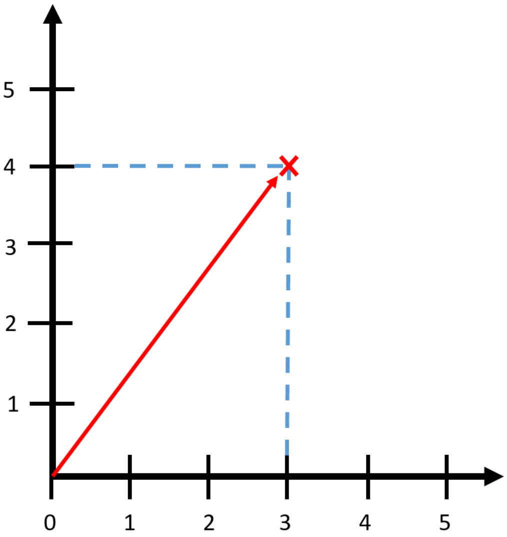
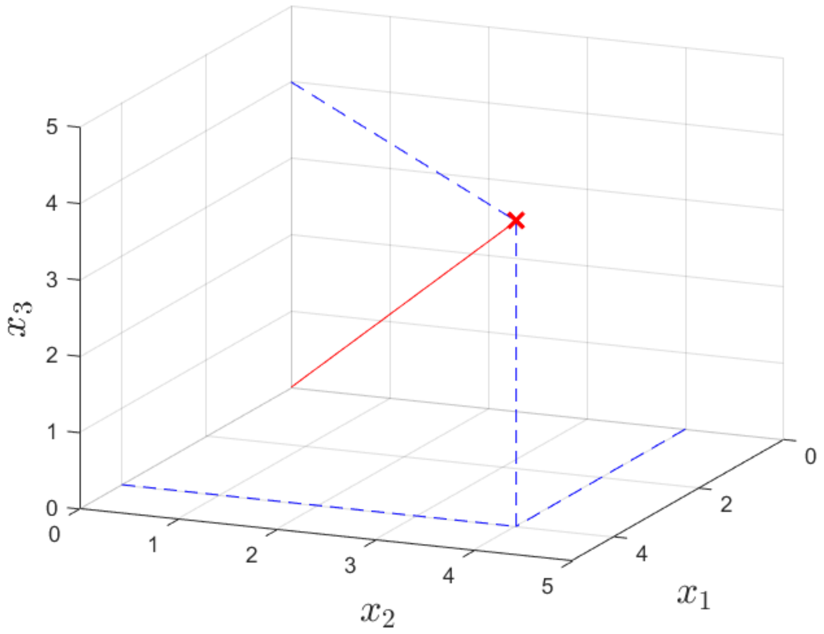


Figure 2: Here, the vector $[3, 4]$ indicates a position on the plane.

A 3-dimension vector $\mathbf{x} = [x_1,\ x_2,\ x_3]$ is the position of a point in a 3-dimension space (i.e. in a volume):



And so on.

Light text is usually used for scalars (x) and bold for vectors ($\mathbf{x}$). Just like the four classical operations on scalars (addition, subtraction, multiplication, division), it is possible to define different operations on vectors. The most useful are the following ones:

- **Multiplication by a scalar**: a vector $\mathbf{x} = [x_1,\ x_2,\ ...\ x_n]$ can be multiplied by a scalar α:

$$\alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix} \tag{1}$$

  Examples:

$$-3 \begin{bmatrix} 5 \\ -1 \end{bmatrix} = \begin{bmatrix} -15 \\ 3 \end{bmatrix}$$

$$2 \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \\ -2 \end{bmatrix}$$

- **Addition**: 2 vectors **of the same dimension** can be added:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \tag{2}$$

Examples:

$$\begin{bmatrix}5\\-3\end{bmatrix}+\begin{bmatrix}2\\2\end{bmatrix}=\begin{bmatrix}7\\-1\end{bmatrix}$$

$$\begin{bmatrix}5\\-3\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}=\begin{bmatrix}3\\-5\end{bmatrix}$$

$$2\begin{bmatrix}4\\5\\-3\end{bmatrix}-\begin{bmatrix}1\\2\\2\end{bmatrix}=\begin{bmatrix}7\\8\\-8\end{bmatrix}$$

- **Element-wise multiplication** (also called **Hadamard product**): 2 vectors **of the same dimension** can be element-wise multiplied. The result is a vector of the same dimension as the operands, where the i^{th} element is the product of each of the i^{th} elements of the operands:

$$\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}\odot\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}=\begin{bmatrix}x_1y_1\\x_2y_2\\\vdots\\x_ny_n\end{bmatrix} \tag{3}$$

Examples:

$$\begin{bmatrix}5\\-3\end{bmatrix}\odot\begin{bmatrix}2\\2\end{bmatrix}=\begin{bmatrix}10\\-6\end{bmatrix}$$

$$\begin{bmatrix}5\\4\end{bmatrix}\odot\begin{bmatrix}2\\2\end{bmatrix}=\begin{bmatrix}10\\8\end{bmatrix}$$

- **Dot product**:

$$\mathbf{x}\cdot\mathbf{y}=\sum_{i=1}^{n}x_iy_i=x_1y_1+x_2y_2+...+x_ny_n$$

Examples:

$$\begin{bmatrix}5\\-3\end{bmatrix}\cdot\begin{bmatrix}2\\2\end{bmatrix}=4$$

$$\begin{bmatrix}5\\5\\-3\end{bmatrix}\cdot\begin{bmatrix}2\\1\\2\end{bmatrix}=9$$

Be careful that the operands of a dot product are **vectors**, but the result is a **scalar** (which is why it is sometimes also called a scalar product). The dot product between $\mathbf{x}$ and $\mathbf{y}$ is also sometimes denoted as $\mathbf{x}^T\mathbf{y}$ or $\langle \mathbf{x}, \mathbf{y} \rangle$.

At this point, one question you may start to ask yourself is "*what is the point of using vectors, which are more complicated objects, rather than scalars ?*"

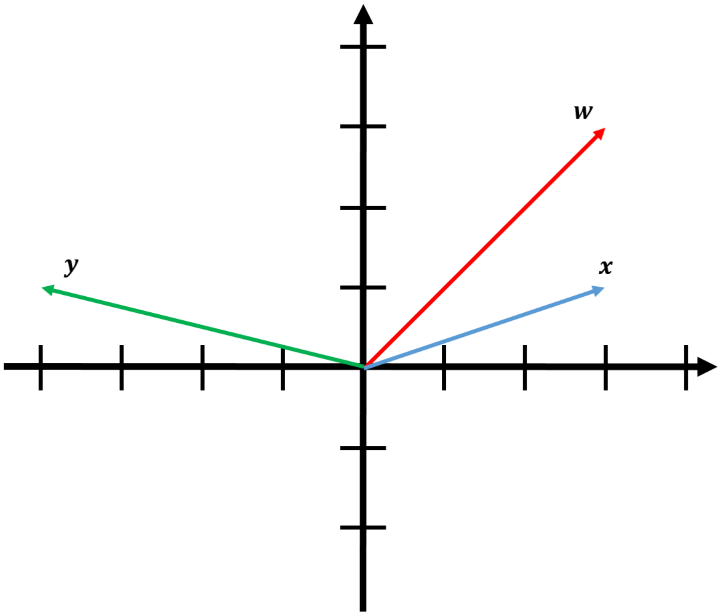
- **They allow to simplify the notations of large data sets**: the weights of each stimulus in overshadowing, or the coordinates of the separation plane in a n-dimensional classification task, are w_1, w_2, ... w_n. Instead, we can simply write $\mathbf{w} = [w_1, w_2, ...w_n]$.
- **They allow to handle operations between data sets**: in overshadowing, assuming that we have n stimuli x_1, x_2, ... x_n, each with a weight w_1, w_2, ... w_n, then the expected reward is

  $$\bar{r} = \sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + ... + w_n x_n$$

  Similarly, in linear classification, if we want to verify that a point $\mathbf{x}$ is correctly classified using $\mathbf{w}$, we have to compute a n-term sum.

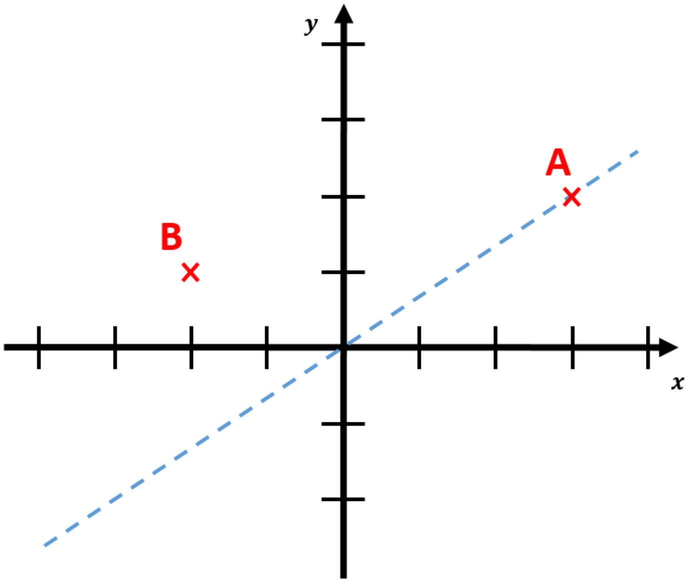
  It is much simpler to write it as $\mathbf{w} \cdot \mathbf{x}$.
- **They provide a graphical interpretation of data**: an interesting property of the dot product in a 2-dimensional plane is that if the angle between two vectors is smaller than 90°, their dot product is positive; reciprocally, if their angle is larger than 90°, their dot product will be negative.

  Consider for instance the three vectors $\mathbf{w} = [3, 3]$, $\mathbf{x} = [3, 1]$, and $\mathbf{y} = [-4, 1]$:



You can verify that $\mathbf{w} \cdot \mathbf{x} = 3 \times 3 + 3 \times 1 = 12$, while $\mathbf{w} \cdot \mathbf{y} = 3 \times -4 + 3 \times 1 = -9$. The idea is the same in higher dimensions: if a dot product is positive, it means that both vectors are pointing towards the same part of the plane; reciprocally, if a dot product is negative, the vectors are heading towards opposite sides. A particular case arises when $\mathbf{w} \cdot \mathbf{x} = 0$: in this case, the angle between vectors $\mathbf{w}$ and $\mathbf{x}$ is exactly equal to 90° (they are said to be **orthogonal**).

Firstly, recall that, in 2 dimensions, a straight line is classically described by the equation $y = ax + b$, where a is called the **slope** of the line, and b its **intercept**. For a given a and b, a line is made of all the points which coordinates $[x, y]$ are such that $y = ax + b$. On the picture below, the equation of the blue dashed line is $y = ax + b$ with $a = \frac{2}{3}$ and $b = 0$. Point $A = [3, 2]$ is part of the line, but $B = [-2, 1]$ is not (check yourself).



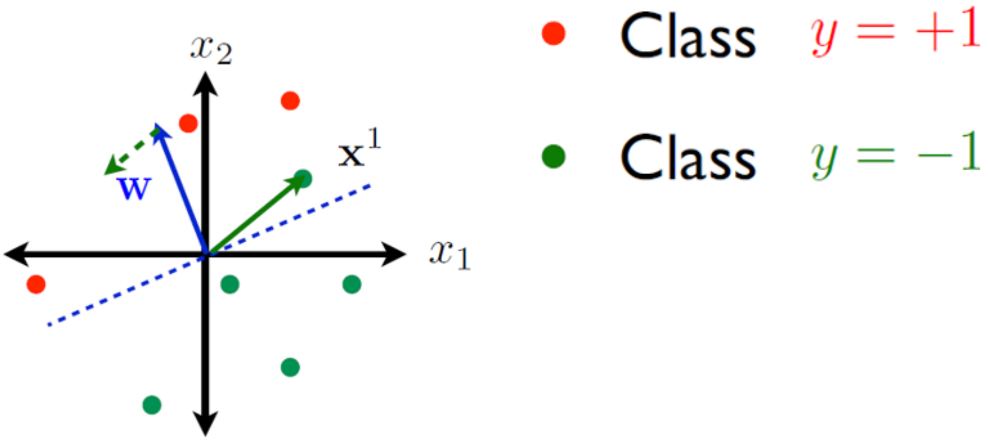
Interestingly, this definition can also be written as a dot product. Indeed, the equality $y = ax + b$ can be transformed into $-ax + y = b$, which is equivalent to $b = \mathbf{w} \cdot \mathbf{x}$ where we define the vectors $\mathbf{w} = [-a, 1]$ and $\mathbf{x} = [x, y]$. **$b = \mathbf{w} \cdot \mathbf{x}$ is the equation of a line in a 2-dimensional plane** (or of an hyperplane in more dimensions). As you will see, this is particularly practical to describe a linear classifier.

In the example below, the separation line is the dashed blue line, and the vector $\mathbf{w}$ is perpendicular to it. The slope and the intercept of the separation line are respectively $a = \frac{2}{3}$ and $b = 0$, then $\mathbf{w} = [-\frac{2}{3}, 1]$, and the separation line is characterized by the equality $\mathbf{w} \cdot \mathbf{x} = 0$ (i.e. you can verify yourself that, for any point $\mathbf{x}$, $\mathbf{x}$ belongs to the line if and only if $\mathbf{w} \cdot \mathbf{x} = 0$).

If we wish to check to which class does a given point $\mathbf{x}$ belong (i.e. to formulate on which side of the separation line it is), we simply have to compute $\mathbf{w} \cdot \mathbf{x}$.

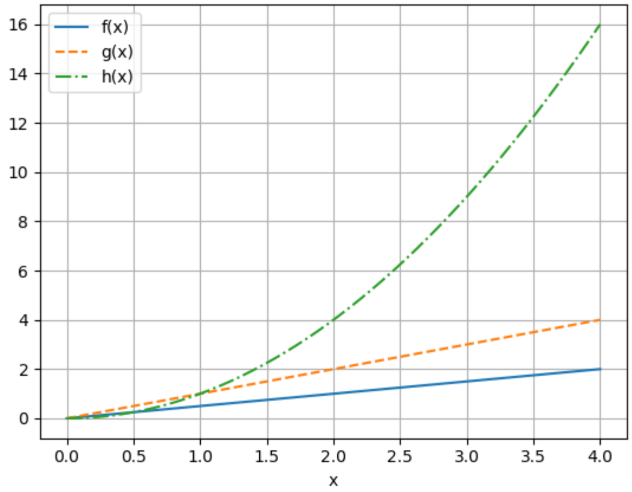
- If $\mathbf{x}$ is in the upper-left part of the plane, then $\mathbf{w} \cdot \mathbf{x} > 0$. Indeed, $\mathbf{w}$ and $\mathbf{x}$ are pointing towards the same area.
- If $\mathbf{x}$ is in the lower-right part of the plane, then $\mathbf{w} \cdot \mathbf{x} < 0$. Which means that $\mathbf{w}$ and $\mathbf{x}$ are pointing towards opposite parts of the plane.
- If $\mathbf{x}$ is on the separation line, then $\mathbf{w} \cdot \mathbf{x} = 0$ ($\mathbf{w}$ and $\mathbf{x}$ are orthogonal).

In this example, the point $\mathbf{x}_1$ is misclassified.



2.2 Calculus

Let's consider the three functions f, g, and h plotted here below:



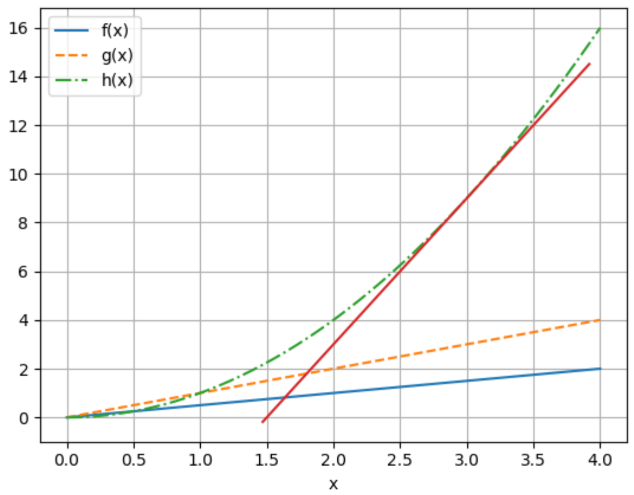
As you can see, all three functions are increasing, but h looks like it is increasing faster than g, which is itself increasing faster than f. How can we quantify this observation ? The solution is to use their derivatives f', g', and h'. *The derivative f' of f tells you how and by how much f is increasing.*

Let's illustrate this with simple linear functions. If you go on the mapping platform of the Swiss Confederation, and look for "Slope over 30°", you can display all the slopes over 57% [1]. What does a 57% = 0.57 slope means ? It means that, if you walked 100 meters in the horizontal direction, you moved 57 meters in the vertical direction (i.e. you gained 57 meters in altitude). More generally, a path with a slope α means that if you walked x units of length in the horizontal plane, you have climbed αx units in the vertical plane. α quantifies how steep your path is.

We can apply it to the functions f and g above. For f, we start at $x = 0$ and $f(x) = 0$, and move up to $x = 4$ and $f(x) = 2$. By moving 4 units on the horizontal axis, we gain 2 units on the vertical axis. This gives us a slope of $\frac{2}{4} = 0.5$. Similarly, g has a slope of $\frac{4}{4} = 1$. These quantities are the derivatives of f and g, and tell us how they are increasing: g is increasing twice as fast as f.

It is very simple for f and g, which are *linear functions* (they are represented by a straight line with a constant slope). But how can we compute the slope of h, which plot is curved ? The solution is to use the line which is tangent to the curve of h at a given point x. For instance, on the figure below, we plotted the tangent of h at $x = 3$:

[1] A slope above 57% (i.e. above 30°) indicates a high risk of avalanche. Stay safe.



This line has a slope of 6 (check yourself). This means that *the derivative $h'(x)$ of h evaluated at $x = 3$ is 6*. This is higher than the derivatives of f and g, which confirms our intuition that h is increasing faster.

The derivative $f'(x)$ of a function f evaluated at x is the slope of the line tangent to f at x. More formally, the slope of a function f indicates what will be the effect on the y axis of a change on the x axis: $\frac{\text{change in } y}{\text{change in } x}$, which is also denoted as $\frac{\Delta y}{\Delta x}$[2]. If we increase x by ϵ, then $\Delta x = x + \epsilon - x = \epsilon$. Consequently, $\Delta y = f(x + \epsilon) - f(x)$, so the slope will be $\frac{f(x+\epsilon)-f(x)}{\epsilon}$. The derivative at x, or instantaneous rate of change, corresponds to this ratio when ϵ gets very small:

$$f'(x) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon) - f(x)}{\epsilon}$$

Finally, note also that we are not only interested in the value of the derivative, but also in its sign. A negative derivative indicates a decreasing function; a positive derivative indicates an increasing function.

Computing a slope is quite easy for linear functions (like f and g), but becomes more complicated for non-constant functions like h. Luckily, you only have to remember the derivatives of the most common functions, as well as some differentiation rules, to be able to differentiate any function:

$y = f(x)$	$\frac{dy}{dx} = f'(x)$
α	0
x	1
x^2	$2x$
x^3	$3x^2$
$\frac{1}{x}$	$-\frac{1}{x^2}$
x^n	nx^{n-1}

(if you remember that $\frac{1}{x} = x^{-1}$, you will see that all these rules are just a consequence of the last one.)

Here are the most common differentiation rules:

- **Multiplication by a constant**: if α is a constant (i.e. a scalar number that does not depend on x) then $(\alpha f)' = \alpha f'$. Be careful that this relation only holds if α is a constant with respect to x: if α depends on x, then you have to apply the rule for differentiating a product below.

  Example: if $f(x) = 3x$, then $f'(x) = 3 \times 1 = 3$

[2] Δ is the classical symbol to indicate the evolution of a quantity. Intuitively, $\Delta x =$ ”new value of x” - ”old value of x”.

- **Derivative of a sum**: $(f+g)' = f' + g'$

  Example 1: if $f(x) = x^2 + x$, then $f'(x) = 2x + 1$

  Example 2: if $f(x) = 3x^3 - 2x^2$, then $f'(x) = 3 \times 3x^2 - 2 \times 2x = 9x^2 - 4x$

- **Derivative of a product**: $(fg)' = fg' + f'g$

  Example 1: if $f(x) = x(3x+5)$, then $f'(x) = x \times 3 + 1 \times (3x+5) = 6x + 5$

  Example 2: if $f(x) = -3x^3(2x^2+5x)$, then $f'(x) = -9x^2 \times (2x^2+5x) - 3x^3 \times (4x+5) = -30x^4 - 60x^3$

- **Derivative of a quotient**: $\left(\frac{f}{g}\right)' = \frac{f'g - g'f}{g^2}$

  Example 1: if $f(x) = \frac{-2x+7}{2x}$, then $f'(x) = \frac{-2 \times 2x - 2 \times (-2x+7)}{4x^2}$

  Example 2: if $f(x) = \frac{x^2 - 5x}{2x^2}$, then $f'(x) = \frac{(2x-5) \times 2x^2 - 4x \times (x^2 - 5x)}{4x^4}$

- **Derivative of a composition of functions (chain rule)**: Consider two functions f and g. Assuming that the conditions for derivating f, g, and $f \circ g$ are met, the chain rule yields

  $$(f \circ g)' = (f' \circ g) \cdot g'$$

  or, using a different notation,

  $$(f(g(x)))' = f'(g(x)) \cdot g'(x)$$

  The chain rule thus allows to compute the derivative of the composition of functions. An intuitive way to understand it is to rewrite the derivatives as fractions (using Leibniz's notation):

  $$(f(g(x)))' = \frac{\Delta f(g(x))}{\Delta x} = \frac{\Delta f(g(x))}{\Delta g(x)} \frac{\Delta g(x)}{\Delta x} = f'(g(x)) \cdot g'(x)$$

  Example 1: if $f(x) = (2x-3)^2$, then $f'(x) = 2(2x-3) \times 2$

  Example 2: if $f(x) = (-2x-3)^2$, then $f'(x) = 2(-2x-3) \times -2$

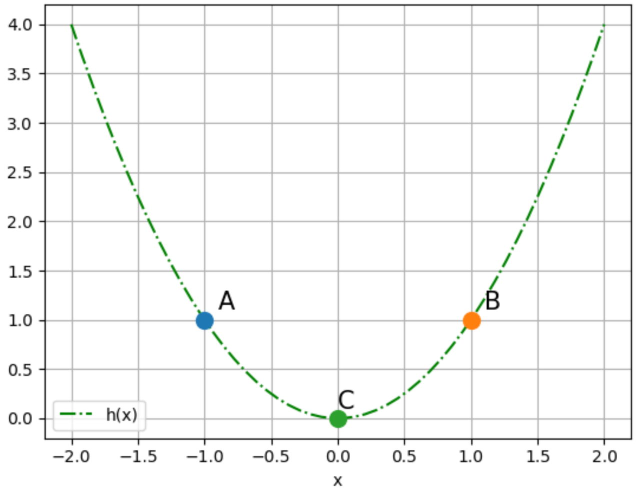
  Example 3: if $f(x) = (-5x^2+2x-3)^3$, then $f'(x) = 3(-5x^2+2x-3)^2 \times (-10x+2)$

This is all very nice, but why are derivatives so useful ? They are a powerful tool to perform **optimization**, i.e. to find the maximum (or equivalently the minimum) of a function.

Let's imagine you are on a walking trail leading to the top of a hill. The path has a constant slope, and you wish to reach the top of the hill (that is to say, to maximize your altitude). The slope is positive, which means that, if you walk forwards, your altitude is going to increase. Reciprocally, if you walk backwards, your altitude is going to decrease. You just have to know the sign of the slope (whether it is positive or negative) to know in which direction you have to walk to increase (or decrease) your altitude. Same goes for the function f described in the figures above. f has a constant positive slope: so if you want to ”maximize your altitude” (i.e. increase the value of $f(x)$) you have to ”move forwards” (i.e. increase the value of x). Reciprocally, if x decreases, so does $f(x)$.

Direction you have to follow in order to reach the top of the linear Niesen (left) or of the parabolic Gantrisch (right). On the left, the slope is constant, and does not depend on where you are. On the right, the slope depends on your position: you have to compute the tangent line (dashed line) to the path at your position.

We can make the same exercise for the function h, for which a broader plot is below:



So far, we used derivatives to find the maximum of a function. Now, assume you wish to find the minimum of this function.[3] From the plot, it is obvious that the minimum is reached for $x = 0$, but we can formalize this intuition using derivatives. When you are sitting at point A:

- h is decreasing;
- Its derivative is negative (if you draw the line that is tangent to h at point A, you obtain a negative slope);
- If you want to minimize $h(x)$, you have to ”move forwards” (i.e. increase x).

All these propositions are equivalent. Just by knowing the sign of the derivative, you know in which direction you have to move from A to advance towards the minimum of the function. Similarly, if you are sitting at B, then h is increasing, so its derivative is positive, so you have to move ”backwards” to minimize $h(x)$.

And what happens at C ? There, the derivative is zero, meaning that you should not move at all: you have reached the minimum of the function. Congratulation !

At first sight, using the derivatives seems to be an overcomplicated method to find the extremum of a function. After all, it is much easier to simply plot the function and to look for its extremum. On the graph above, it is obvious that the minimum of h is reached at $x = 0$. But the main point of using its derivative is that it is a **local** method. Let's go back to the example of a hiker wishing to reach the top of a mountain. If you know where you are and what the mountain looks like, you can indeed directly aim for the summit. But the hiker might also be lost in the fog, without a map or GPS: in that case, you can only rely on the slope of the path right beneath your feet, that is to say on local information. Same goes for analytical functions: some of them are too complicated or high-dimensional to be simply plotted, and we might only have access to local derivatives to look for the direction in which the extremum lies.

These simple examples illustrate how derivation can be used to find the minimum of complicated functions. In machine learning, we wish to find the value of the parameters $\mathbf{w}$ that will minimize the error function $E(\mathbf{w})$. But E might be a complicated function, and $\mathbf{w}$ a multi-dimensional vector, so we cannot simply plot $E(\mathbf{w})$ to find its minimum. However, we can compute its derivative, which tells us how we should modify $\mathbf{w}$ in order to move closer to the minimum. This method is called **gradient descent**, and is the basis of many optimization techniques.

[3] Finding the minimum or the maximum of a function are actually equivalent problems, since maximizing f is equivalent to minimizing $-f$. In practice, optimization problems are usually formulated as minimization problems.

3 Additional resources

- On STDP: Jesper Sjöström and Wulfram Gerstner (2010) Spike-timing dependent plasticity. Scholarpedia, 5(2):1362.
- On vector and matrix computation: The Matrix Cookbook (Kaare Brandt Petersen and Michael Syskind Pedersen)
- On biological learning: Dayan, P., Abbott, L. (2001). Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Cambridge, MA, USA: MIT Press.
- On neural networks and backpropagation: Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015
- On the need to have a quantitative (i.e. math-oriented) approach to biology: Bialek, W. (2017). Perspectives on theory at the interface of physics and biology. Reports on Progress in Physics, 81(1), 012601.
- On neuron networks and deep learning: Le Cun, Y. (2019). Quand la machine apprend: la révolution des neurones artificiels et de l'apprentissage profond. Odile Jacob.
- On the biological plausibility of backpropagation: Sun, W., Zhao, X. and Spruston, N. Bursting potentiates the neuro–AI connection. Nat Neurosci (2021).