Math - Differential Calculus

Calculus is the study of continuous change. It has two major subfields: *differential calculus*, which studies the rate of change of functions, and *integral calculus*, which studies the area under the curve. In this notebook, we will discuss the former.

Differential calculus is at the core of Deep Learning, so it is important to understand what derivatives and gradients are, how they are used in Deep Learning, and understand what their limitations are.

Note: the code in this notebook is only used to create figures and animations. You do not need to understand how it works (although I did my best to make it clear, in case you are interested).

Slope of a straight line

+ Code   + Text

Show code

What is the slope of the following line?

Show code
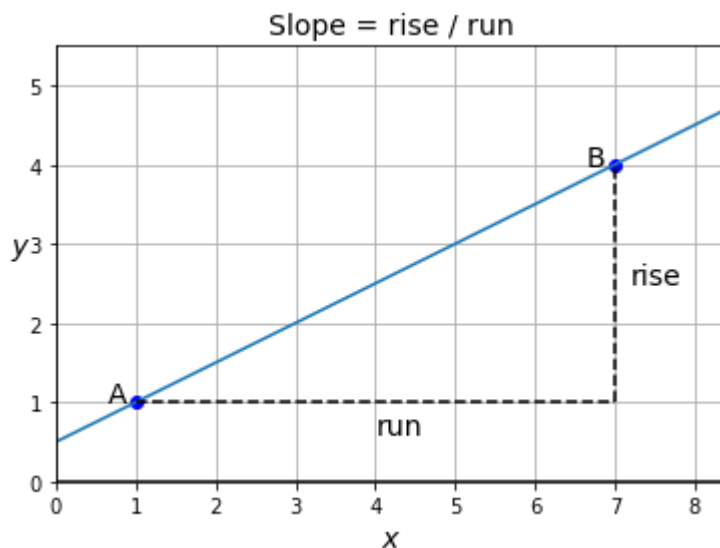
As you probably know, the slope of a (non-vertical) straight line can be calculated by taking any two points A and B on the line, and computing the "rise over run":
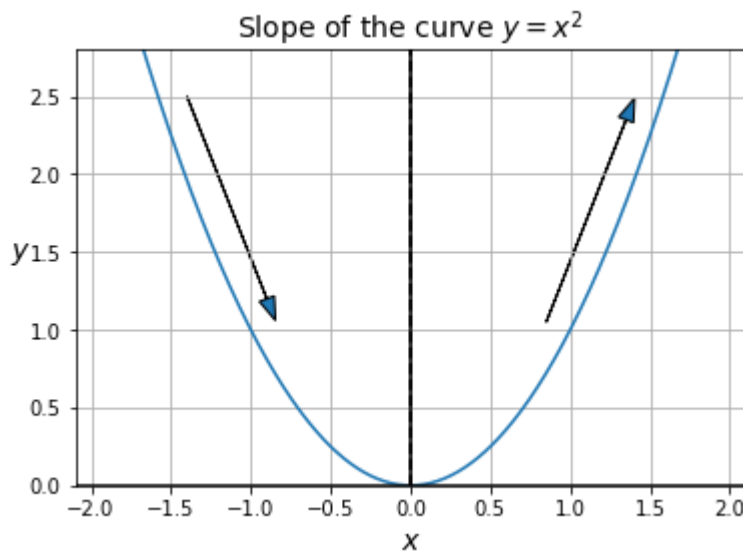
$$slope = \frac{rise}{run} = \frac{\Delta y}{\Delta x} = \frac{y_B - y_A}{x_B - x_A}$$

In this example, the rise is 3, and the run is 6, so the slope is 3/6 = 0.5.

▾ Defining the slope of a curve

But what if you want to know the slope of something other than a straight line? For example, let's consider the curve defined by $y = f(x) = x^2$:

Show code

Obviously, the slope varies: on the left (i.e., when $x < 0$), the slope is negative (i.e., when we move from left to right, the curve goes down), while on the right (i.e., when $x > 0$) the slope is positive (i.e., when we move from left to right, the curve goes up). At the point $x = 0$, the slope is equal to 0 (i.e., the curve is locally flat). The fact that the slope is 0 when we reach a minimum (or indeed a maximum) is crucially important, and we will come back to it later.

How can we put numbers on these intuitions? Well, say we want to estimate the slope of the curve at a point A. We can do this by taking another point B on the curve, not too far away, and then computing the slope between these two points:

Show code

Slope of the curve $y = x^2$ at $x_A = -1$
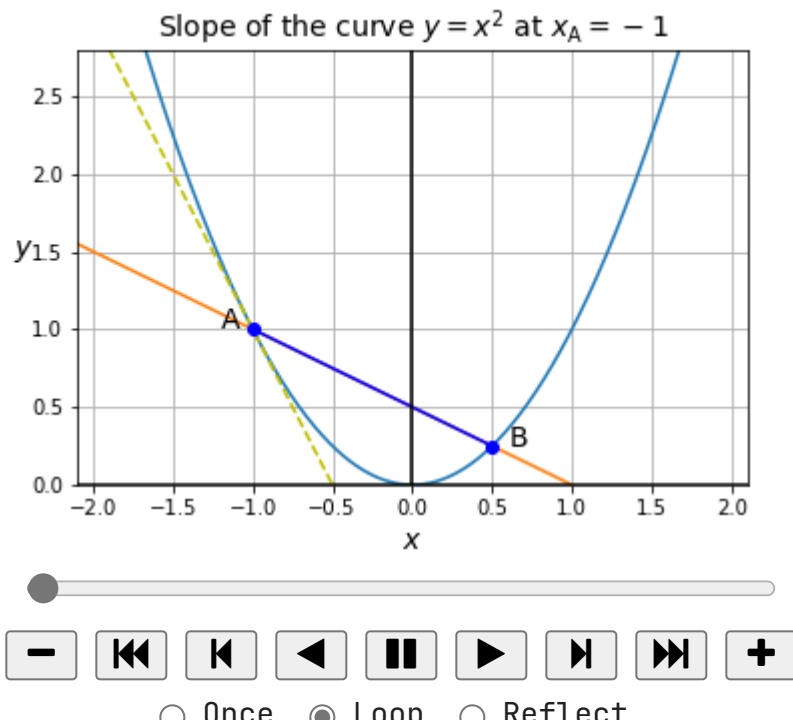
As you can see, when point B is very close to point A, the (AB) line becomes almost indistinguishable from the curve itself (at least locally around point A). The (AB) line gets closer and closer to the **tangent** line to the curve at point A: this is the best linear approximation of the curve at point A.

So it makes sense to define the slope of the curve at point A as the slope that the (AB) line approaches when B gets infinitely close to A. This slope is called the **derivative** of the function f at $x = x_A$. For example, the derivative of the function $f(x) = x^2$ at $x = x_A$ is equal to $2x_A$ (we will see how to get this result shortly), so on the graph above, since the point A is located at $x_A = -1$, the tangent line to the curve at that point has a slope of -2.

▾ Differentiability

Note that some functions are not quite as well-behaved as x^2. For example, consider the function $f(x) = |x|$, the absolute value of x:

Show code

Slope of the curve $y = |x|$ at $x_A = 0$
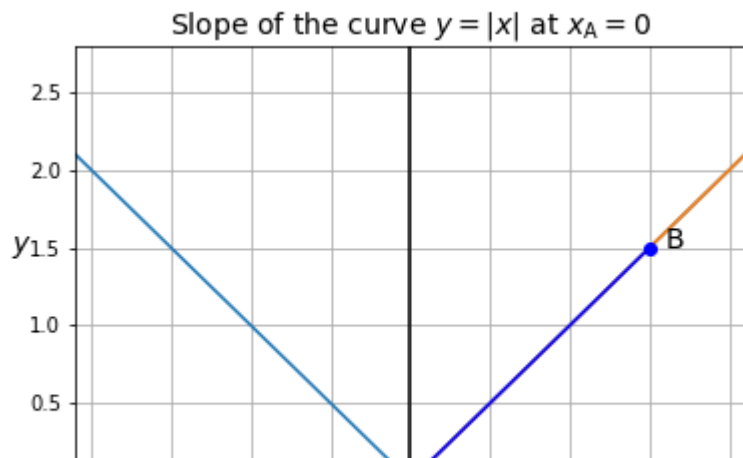
No matter how much you zoom in on the origin (the point at $x = 0, y = 0$), the curve will always look like a V. The slope is -1 for any $x < 0$, and it is +1 for any $x > 0$, but **at $x = 0$, the slope is undefined**, since it is not possible to approximate the curve $y = |x|$ locally around the origin using a straight line, no matter how much you zoom in on that point.

The function $f(x) = |x|$ is said to be **non-differentiable** at $x = 0$: its derivative is undefined at $x = 0$. This means that the curve $y = |x|$ has an undefined slope at that point. However, the function $f(x) = |x|$ is **differentiable** at all other points.

In order for a function $f(x)$ to be differentiable at some point x_A, the slope of the (AB) line must approach a single finite value as B gets infinitely close to A.

This implies several constraints:

- First, the function must of course be **defined** at x_A. As a counterexample, the function $f(x) = \dfrac{1}{x}$ is undefined at $x_A = 0$, so it is not differentiable at that point.
- The function must also be **continuous** at x_A, meaning that as x_B gets infinitely close to x_A, $f(x_B)$ must also get infinitely close to $f(x_A)$. As a counterexample, $f(x) = \begin{cases} -1 \text{ if } x < 0 \\ +1 \text{ if } x \geq 0 \end{cases}$ is not continuous at $x_A = 0$, even though it is defined at that point: indeed, when you approach it from the negative side, it does not approach infinitely close to $f(0) = +1$. Therefore, it is not continuous at that point, and thus not differentiable either.
- The function must not have a **breaking point** at x_A, meaning that the slope that the (AB) line approaches as B approaches A must be the same whether B approaches from the left side or from the right side. We already saw a counterexample with $f(x) = |x|$, which is both defined and continuous at $x_A = 0$, but which has a breaking point at $x_A = 0$: the slope of the curve $y = |x|$ is -1 on the left, and +1 on the right.
- The curve $y = f(x)$ must not be **vertical** at point A. One counterexample is $f(x) = \sqrt[3]{x}$, the cubic root of x: the curve is vertical at the origin, so the function is not differentiable at $x_A = 0$, as you can see in the following animation:

Slope of the curve $y = \sqrt[3]{x}$ at $x_A = 0$

○ Once   ◉ Loop   ○ Reflect
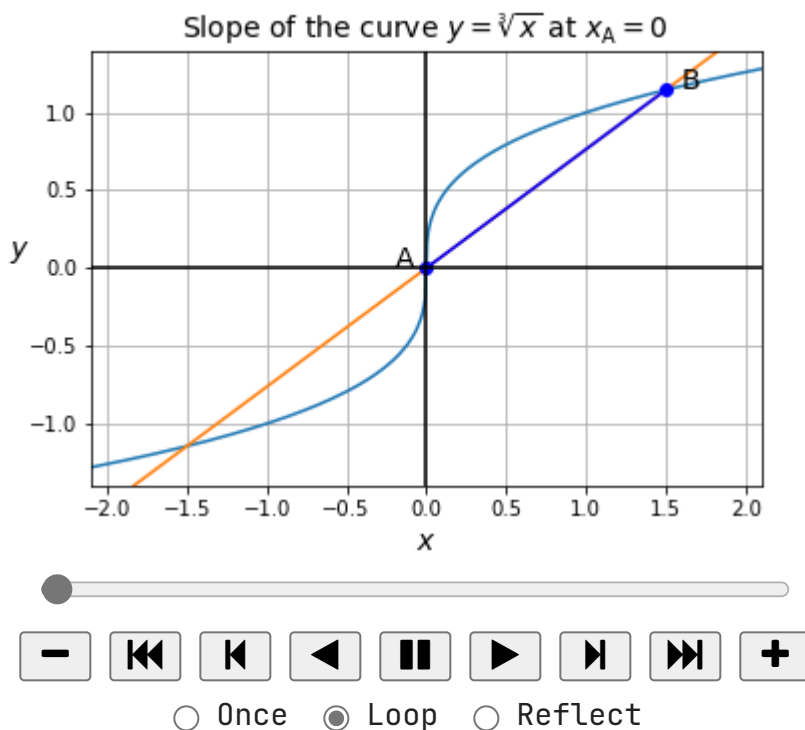
Now let's see how to actually differentiate a function (i.e., find its derivative).

▾ Differentiating a function

The previous discussion leads to the following definition:

The **derivative** of a function $f(x)$ at $x = x_A$ is noted $f'(x_A)$, and it is defined as:

$$f'(x_A) = \lim_{x_B \to x_A} \frac{f(x_B) - f(x_A)}{x_B - x_A}$$

Don't be scared, this is simpler than it looks! You may recognize the *rise over run* equation $\frac{y_B - y_A}{x_B - x_A}$ that we discussed earlier. That's just the slope of the (AB) line. And the notation $\lim_{x_B \to x_A}$ means that we are making x_B approach infinitely close to x_A. So in plain English, $f'(x_A)$ is the value that the slope of the (AB) line approaches when B gets infinitely close to A. This is just a formal way of saying exactly the same thing as earlier.

Example: finding the derivative of x^2

Let's look at a concrete example. Let's see if we can determine what the slope of the $y = x^2$ curve is, at any point A (try to understand each line, I promise it's not that hard):

$$
\begin{aligned}
f'(x_A) &= \lim_{x_B \to x_A} \frac{f(x_B) - f(x_A)}{x_B - x_A} \\
&= \lim_{x_B \to x_A} \frac{x_B^2 - x_A^2}{x_B - x_A} && \text{since } f(x) = x^2 \\
&= \lim_{x_B \to x_A} \frac{(x_B - x_A)(x_B + x_A)}{x_B - x_A} && \text{since } x_A^2 - x_B^2 = (x_A - x_B)(x_A + x_B) \\
&= \lim_{x_B \to x_A} (x_B + x_A) && \text{since the two } (x_B - x_A) \text{ cancel out} \\
&= \lim_{x_B \to x_A} x_B + \lim_{x_B \to x_A} x_A && \text{since the limit of a sum is the sum of the l} \\
&= x_A + \lim_{x_B \to x_A} x_A && \text{since } x_B \text{ approaches } x_A \\
&= x_A + x_A && \text{since } x_A \text{ remains constant when } x_B \text{ appro} \\
&= 2x_A
\end{aligned}
$$

That's it! We just proved that the slope of $y = x^2$ at any point A is $f'(x_A) = 2x_A$. What we have done is called **differentiation**: finding the derivative of a function.

Note that we used a couple of important properties of limits. Here are the main properties you need to know to work with derivatives:

- $\lim\limits_{x \to k} c = c$ if c is some constant value that does not depend on x, then the limit is just c.
- $\lim\limits_{x \to k} x = k$ if x approaches some value k, then the limit is k.
- $\lim\limits_{x \to k} [f(x) + g(x)] = \lim\limits_{x \to k} f(x) + \lim\limits_{x \to k} g(x)$ the limit of a sum is the sum of the limits
- $\lim\limits_{x \to k} [f(x) \times g(x)] = \lim\limits_{x \to k} f(x) \times \lim\limits_{x \to k} g(x)$ the limit of a product is the product of the limits

Important note: in Deep Learning, differentiation is almost always performed automatically by the framework you are using (such as TensorFlow or PyTorch). This is called auto-diff, and I did another notebook on that topic. However, you should still make sure you have a good understanding of derivatives, or else they will come and bite you one day, for example when you use a square root in your cost function without realizing that its derivative approaches infinity when x approaches 0 (tip: you should use $\sqrt{x + \epsilon}$ instead, where ϵ is some small constant, such as 10^{-4}).

You will often find a slightly different (but equivalent) definition of the derivative. Let's derive it from the previous definition. First, let's define $\epsilon = x_B - x_A$. Next, note that ϵ will approach 0 as x_B approaches x_A. Lastly, note that $x_B = x_A + \epsilon$. With that, we can reformulate the definition above like so:

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

While we're at it, let's just rename x_A to x, to get rid of the annoying subscript A and make the equation simpler to read:

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

Okay! Now let's use this new definition to find the derivative of $f(x) = x^2$ at any point x, and (hopefully) we should find the same result as above (except using x instead of x_A):

$$
\begin{aligned}
f'(x) &= \lim_{\epsilon \to 0} \frac{f(x + \epsilon) - f(x)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{(x + \epsilon)^2 - x^2}{\epsilon} && \text{since } f(x) = x^2 \\
&= \lim_{\epsilon \to 0} \frac{x^2 + 2x\epsilon + \epsilon^2 - x^2}{\epsilon} && \text{since } (x + \epsilon)^2 = x^2 + 2x\epsilon + \epsilon^2 \\
&= \lim_{\epsilon \to 0} \frac{2x\epsilon + \epsilon^2}{\epsilon} && \text{since the two } x^2 \text{ cancel out} \\
&= \lim_{\epsilon \to 0} (2x + \epsilon) && \text{since } 2x\epsilon \text{ and } \epsilon^2 \text{ can both be divided by } \epsilon \\
&= 2x
\end{aligned}
$$

Yep! It works out.

▾ Notations

A word about notations: there are several other notations for the derivative that you will find in the literature:

$$f'(x) = \frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$$

This notation is also handy when a function is not named. For example $\dfrac{\mathrm{d}}{\mathrm{d}x}\left[x^2\right]$ refers to the derivative of the function $x \mapsto x^2$.

Moreover, when people talk about the function $f(x)$, they sometimes leave out "(x)", and they just talk about the function f. When this is the case, the notation of the derivative is also simpler:

$$f' = \frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}f$$

The f' notation is Lagrange's notation, while $\dfrac{\mathrm{d}f}{\mathrm{d}x}$ is Leibniz's notation.

There are other less common notations, such as Newton's notation $\dot{u}$ (assuming $u = f(x)$) or

▾ Plotting the tangent to a curve

Let's use the equation $f'(x) = 2x$ to plot the tangent to the $y = x^2$ curve at various values of x (you can click on the play button under the graphs to play the animation):

    Show code

$y = f(x) = x^2$ and the tangent at $x = x_A$
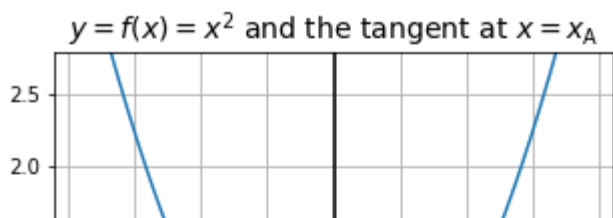
Note: consider the tangent line to the curve $y = f(x)$ at some point A. What is its equation? Well, since the tangent is a straight line, its equation must look like:

$$y = \alpha x + \beta$$

where α is the slope of the line, and β is the offset (i.e., the y coordinate of the point at which the line crosses the vertical axis). We already know that the slope of the tangent line at point A is the derivative of $f(x)$ at that point, so:

$$\alpha = f'(x_A)$$

But what about the offset β? Well we also know that the tangent line touches the curve at point A, so we know that $\alpha x_A + \beta = f(x_A)$. So:

$$\beta = f(x_A) - f'(x_A)x_A$$

So we get the following equation for the tangent:

$$y = f(x_A) + f'(x_A)(x - x_A)$$

For example, the tangent to the $y = x^2$ curve is given by:

$$y = x_A{}^2 + 2x_A(x - x_A) = 2x_A x - x_A^2$$

| — | | |◀◀| | |◀| | |◀| | |⫿| | |▶| | |▶| | |▶▶| | |+| |

▾ Differentiation rules

One very important rule is that **the derivative of a sum is the sum of the derivatives**. More precisely, if we define $f(x) = g(x) + h(x)$, then $f'(x) = g'(x) + h'(x)$. This is quite easy to prove:

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

$$= \lim_{\epsilon \to 0} \frac{g(x+\epsilon) + h(x+\epsilon) - g(x) - h(x)}{\epsilon} \qquad \text{using } f(x) = g(x) + h(x)$$

$$= \lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x) + h(x+\epsilon) - h(x)}{\epsilon} \qquad \text{just moving terms around}$$

$$= \lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x)}{\epsilon} + \lim_{\epsilon \to 0} \frac{h(x+\epsilon) - h(x)}{\epsilon} \qquad \text{since the limit of a sum is th}$$

$$= g'(x) + h'(x) \qquad \text{using the definitions of } g'(x$$

Similarly, it is possible to show the following important rules (I've included the proofs at the end of this notebook, in case you're curious):

	Function f	Derivative f'
Constant	$f(x) = c$	$f'(x) = 0$
Sum	$f(x) = g(x) + h(x)$	$f'(x) = g'(x) + h'(x)$
Product	$f(x) = g(x)h(x)$	$f'(x) = g(x)h'(x) + g'(x)h(x)$
Quotient	$f(x) = \dfrac{g(x)}{h(x)}$	$f'(x) = \dfrac{g'(x)h(x) - g(x)h'(x)}{h^2(x)}$
Power	$f(x) = x^r$ with $r \neq 0$	$f'(x) = rx^{r-1}$
Exponential	$f(x) = \exp(x)$	$f'(x) = \exp(x)$
Logarithm	$f(x) = \ln(x)$	$f'(x) = \dfrac{1}{x}$
Sin	$f(x) = \sin(x)$	$f'(x) = \cos(x)$
Cos	$f(x) = \cos(x)$	$f'(x) = -\sin(x)$
Tan	$f(x) = \tan(x)$	$f'(x) = \dfrac{1}{\cos^2(x)}$
Chain Rule	$f(x) = g(h(x))$	$f'(x) = g'(h(x))\,h'(x)$

Let's try differentiating a simple function using the above rules: we will find the derivative of $f(x) = x^3 + \cos(x)$. Using the rule for the derivative of sums, we find that $f'(x) = \dfrac{\mathrm{d}}{\mathrm{d}x}[x^3] + \dfrac{\mathrm{d}}{\mathrm{d}x}[\cos(x)]$. Using the rule for the derivative of powers and for the cos function, we find that $f'(x) = 3x^2 - \sin(x)$.

Let's try a harder example: let's find the derivative of $f(x) = \sin(2x^2) + 1$. First, let's define $u(x) = \sin(x) + 1$ and $v(x) = 2x^2$. Using the rule for sums, we find that $u'(x) = \dfrac{d}{dx}[sin(x)] + \dfrac{d}{dx}[1]$. Since the derivative of the $\sin$ function is $\cos$, and the derivative of constants is 0, we find that $u'(x) = \cos(x)$. Next, using the product rule, we find that $v'(x) = 2\dfrac{d}{dx}[x^2] + \dfrac{d}{dx}[2]\,x^2$. Since the derivative of a constant is 0, the second term cancels out. And since the power rule tells us that the derivative of x^2 is $2x$, we find that $v'(x) = 4x$. Lastly, using the chain rule, since $f(x) = u(v(x))$, we find that $f'(x) = u'(v(x))\,v'(x) = \cos(2x^2)\,4x$.

Let's plot f followed by f', and let's use $f'(x_\mathbf{A})$ to find the slope of the tangent at some point $\mathbf{A}$:

Show code

The chain rule

The chain rule is easier to remember using Leibniz's notation:

If $f(x) = g(h(x))$ and $y = h(x)$, then: $\dfrac{\mathrm{d}f}{\mathrm{d}x} = \dfrac{\mathrm{d}f}{\mathrm{d}y}\dfrac{\mathrm{d}y}{\mathrm{d}x}$

Indeed, $\dfrac{\mathrm{d}f}{\mathrm{d}y} = f'(y) = f'(h(x))$ and $\dfrac{\mathrm{d}y}{\mathrm{d}x} = h'(x)$.

It is possible to chain many functions. For example, if $f(x) = g(h(i(x)))$, and we define $y = i(x)$ and $z = h(y)$, then $\dfrac{\mathrm{d}f}{\mathrm{d}x} = \dfrac{\mathrm{d}f}{\mathrm{d}z}\dfrac{\mathrm{d}z}{\mathrm{d}y}\dfrac{\mathrm{d}y}{\mathrm{d}x}$. Using Lagrange's notation, we get

$$f'(x) = g'(z)\,h'(y)\,i'(x) = g'(h(i(x)))\,h'(i(x))\,i'(x)$$

The chain rule is crucial in Deep Learning, as a neural network is basically a long composition of functions. For example, a 3-layer dense neural network corresponds to the following function: $f(\mathbf{x}) = \mathrm{Dense}_3(\mathrm{Dense}_2(\mathrm{Dense}_1(\mathbf{x})))$ (in this example, Dense_3 is the output layer).

Derivatives and optimization

When trying to optimize a function $f(x)$, we look for the values of x that minimize (or maximize) the function.

It is important to note that when a function reaches a minimum or maximum, assuming it is differentiable at that point, the derivative will necessarily be equal to 0. For example, you can check the above animation, and notice that whenever the function f (in the upper graph) reaches a maximum or minimum, then the derivative f' (in the lower graph) is equal to 0.

So one way to optimize a function is to differentiate it and analytically find all the values for which the derivative is 0, then determine which of these values optimize the function (if any). For example, consider the function $f(x) = \dfrac{1}{4}x^4 - x^2 + \dfrac{1}{2}$. Using the derivative rules (specifically, the sum rule, the product rule, the power rule and the constant rule), we find that $f'(x) = x^3 - 2x$. We look for the values of x for which $f'(x) = 0$, so $x^3 - 2x = 0$, and therefore $x(x^2 - 2) = 0$. So $x = 0$, or $x = \sqrt{2}$ or $x = -\sqrt{2}$. As you can see on the following graph of $f(x)$, these 3 values correspond to local extrema. Two global minima $f(\sqrt{2}) = f(-\sqrt{2}) = -\dfrac{1}{2}$ and one local maximum $f(0) = \dfrac{1}{2}$.
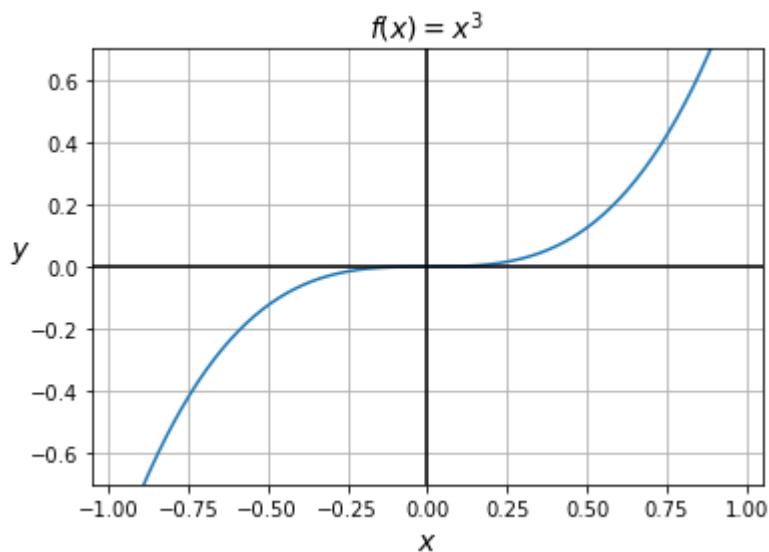
$$y = f(x) = \tfrac{1}{4}x^4 - x^2 + 5$$
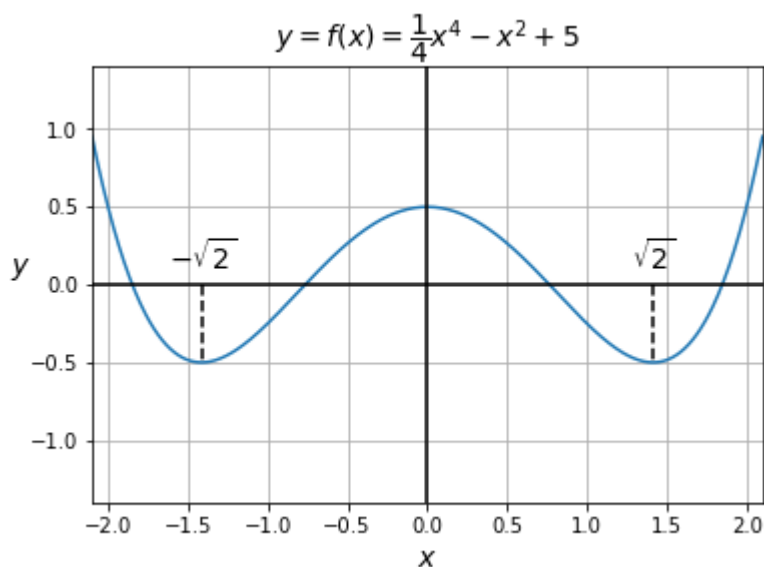
If a function has a local extremum at a point x_A and is differentiable at that point, then $f'(x_A) = 0$. However, the reverse is not always true. For example, consider $f(x) = x^3$. Its derivative is $f'(x) = x^2$, which is equal to 0 at $x_A = 0$. Yet, this point is *not* an extremum, as you can see on the following diagram. It's just a single point where the slope is 0.

$$f(x) = x^3$$

So in short, you can optimize a function by analytically working out the points at which the derivative is 0, and then investigating only these points. It's a beautifully elegant solution, but it

requires a lot of work, and it's not always easy, or even possible. For neural networks, it's practically impossible.

Another option to optimize a function is to perform **Gradient Descent** (we will consider minimizing the function, but the process would be almost identical if we tried to maximize a function instead): start at a random point x_0, then use the function's derivative to determine the slope at that point, and move a little bit in the downwards direction, then repeat the process until you reach a local minimum, and cross your fingers in the hope that this happens to be the global minimum.

At each iteration, the step size is proportional to the slope, so the process naturally slows down as it approaches a local minimum. Each step is also proportional to the learning rate: a parameter of the Gradient Descent algorithm itself (since it is not a parameter of the function we are optimizing, it is called a **hyperparameter**).

Here is an animation of this process for the function $f(x) = \dfrac{1}{4}x^4 - x^2 + \dfrac{1}{2}$:

Show code

$y = f(x) = \frac{1}{4}x^4 - x^2 + \frac{1}{2}$ and the tangent at $x = x_A$
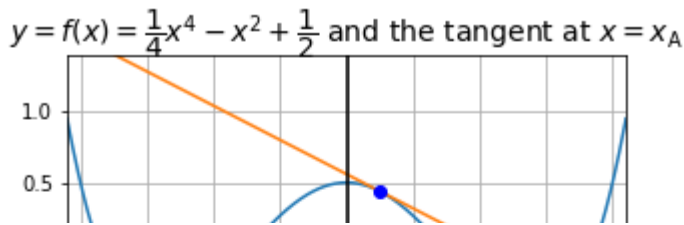
In this example, we started with $x_0 = \frac{1}{4}$, so Gradient Descent "rolled down" towards the minimum value at $x = \sqrt{2}$. But if we had started at $x_0 = -\frac{1}{4}$, it would have gone towards $-\sqrt{2}$. This illustrates the fact that the initial value is important: depending on x_0, the algorithm may converge to a global minimum (hurray!) or to a poor local minimum (boo!) or stay stuck on a plateau, such as a horizontal inflection point (boo!).

There are many variants of the Gradient Descent algorithm, discussed in Chapter 11 of the book. These are the ones we care about in Deep Learning. They all rely on the derivative of the cost function with regards to the model parameters (we will discuss functions with multiple parameters later in this notebook).

Higher order derivatives

What happens if we try to differentiate the function $f'(x)$? Well, we get the so-called second order derivative, noted $f''(x)$, or $\dfrac{d^2 f}{dx^2}$. If we repeat the process by differentiating $f''(x)$, we get the third-order derivative $f'''(x)$, or $\dfrac{d^3 f}{dx^3}$. And we could go on to get higher order derivatives.

What's the intuition behind second order derivatives? Well, since the (first order) derivative represents the instantaneous rate of change of f at each point, the second order derivative represents the instantaneous rate of change of the rate of change itself, in other words, you can think of it as the **acceleration** of the curve: if $f''(x) < 0$, then the curve is accelerating "downwards", if $f''(x) > 0$ then the curve is accelerating "upwards", and if $f''(x) = 0$, then the curve is locally a straight line. Note that a curve could be going upwards (i.e., $f'(x) > 0$) but also be accelerating downwards (i.e., $f''(x) < 0$): for example, imagine the path of a stone thrown upwards, as it is being slowed down by gravity (which constantly accelerates the stone downwards).

Deep Learning generally only uses first order derivatives, but you will sometimes run into some optimization algorithms or cost functions based on second order derivatives.

Partial derivatives

Up to now, we have only considered functions with a single variable x. What happens when there are multiple variables? For example, let's start with a simple function with 2 variables: $f(x, y) = \sin(xy)$. If we plot this function, using $z = f(x, y)$, we get the following 3D graph. I also plotted some point A on the surface, along with two lines I will describe shortly.

Show code

$$z = f(x, y) = \sin(xy)$$
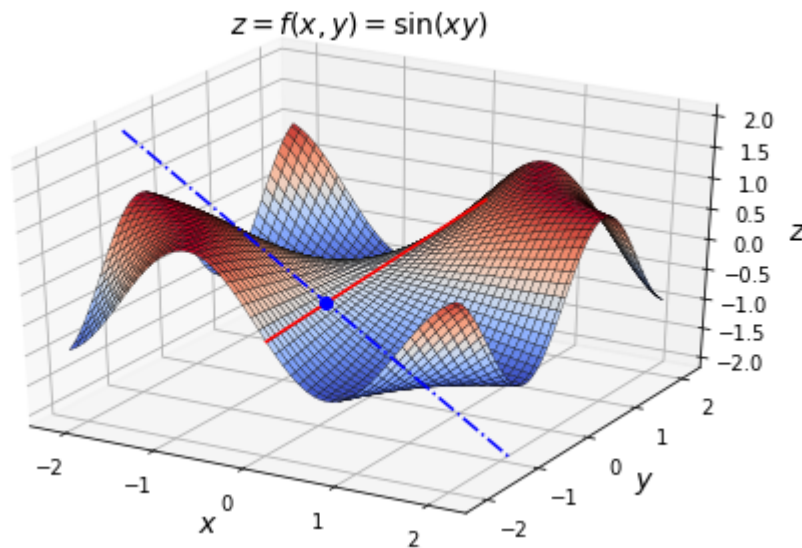
If you were to stand on this surface at point A and walk along the x axis towards the right (increasing x), your path would go down quite steeply (along the dashed blue line). The slope along this axis would be negative. However, if you were to walk along the y axis, towards the back (increasing y), then your path would almost be flat (along the solid red line), at least locally: the slope along that axis, at point A, would be very slightly positive.

As you can see, a single number is no longer sufficient to describe the slope of the function at a given point. We need one slope for the x axis, and one slope for the y axis. One slope for each variable. To find the slope along the x axis, called the **partial derivative of f with regards to x**, and noted $\dfrac{\partial f}{\partial x}$ (with curly ∂), we can differentiate $f(x, y)$ with regards to x while treating all other variables (in this case just y) as constants:

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

If you use the derivative rules listed earlier (in this example you would just need the product rule and the chain rule), making sure to treat y as a constant, then you will find:

$$\frac{\partial f}{\partial x} = y\cos(xy)$$

Similarly, the partial derivative of f with regards to y is defined as:

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

All variables except for y are treated like constants (just x in this example). Using the derivative rules, we get:

$$\frac{\partial f}{\partial y} = x\cos(xy)$$

We now have equations to compute the slope along the x axis and along the y axis. But what about the other directions? If you were standing on the surface at point A, you could decide to walk in any direction you choose, not just along the x or y axes. What would the slope be then? Shouldn't we compute the slope along every possible direction?

Well, it can be shown that if all the partial derivatives are defined and continuous in a neighborhood around point A, then the function f is **totally differentiable** at that point, meaning that it can be locally approximated by a plane P_A (the tangent plane to the surface at point A). In this case, having just the partial derivatives along each axis (x and y in our case) is sufficient to perfectly characterize that plane. Its equation is:

$$z = f(x_A, y_A) + (x - x_A)\frac{\partial f}{\partial x}(x_A, y_A) + (y - y_A)\frac{\partial f}{\partial y}(x_A, y_A)$$

In Deep Learning, we will generally be dealing with well-behaved functions that are totally differentiable at any point where all the partial derivatives are defined, but you should know that some functions are not that nice. For example, consider the function:

$$h(x, y) = \begin{cases} 0 \text{ if } x = 0 \text{ or } y = 0 \\ 1 \text{ otherwise} \end{cases}$$

At the origin (i.e., at $(x, y) = (0, 0)$), the partial derivatives of the function h with respect to x and y are both perfectly defined: they are equal to 0. Yet the function can clearly not be approximated by a plane at that point. It is not totally differentiable at that point (but it is totally differentiable at any point off the axes).

Gradients

So far we have considered only functions with a single variable x, or with 2 variables, x and y, but the previous paragraph also applies to functions with more variables. So let's consider a function f

with n variables: $f(x_1, x_2, \ldots, x_n)$. For convenience, we will define a vector $\mathbf{x}$ whose components are these variables:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Now $f(\mathbf{x})$ is easier to write than $f(x_1, x_2, \ldots, x_n)$.

The gradient of the function $f(\mathbf{x})$ at some point $\mathbf{x}_A$ is the vector whose components are all the partial derivatives of the function at that point. It is noted $\nabla f(\mathbf{x}_A)$, or sometimes $\nabla_{\mathbf{x}_A} f$:

$$\nabla f(\mathbf{x}_A) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(\mathbf{x}_A) \\ \dfrac{\partial f}{\partial x_2}(\mathbf{x}_A) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(\mathbf{x}_A) \end{pmatrix}$$

Assuming the function is totally differentiable at the point $\mathbf{x}_A$, then the surface it describes can be approximated by a plane at that point (as discussed in the previous section), and the gradient vector is the one that points towards the steepest slope on that plane.

Gradient Descent, revisited

In Deep Learning, the Gradient Descent algorithm we discussed earlier is based on gradients instead of derivatives (hence its name). It works in much the same way, but using vectors instead of scalars: simply start with a random vector $\mathbf{x}_0$, then compute the gradient of f at that point, and perform a small step in the opposite direction, then repeat until convergence. More precisely, at each step t, compute $\mathbf{x}_t = \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1})$. The constant η is the learning rate, typically a small value such as 10^{-3}. In practice, we generally use more efficient variants of this algorithm, but the general idea remains the same.

In Deep Learning, the letter $\mathbf{x}$ is generally used to represent the input data. When you *use* a neural network to make predictions, you feed the neural network the inputs $\mathbf{x}$, and you get back a prediction $\hat{y} = f(\mathbf{x})$. The function f treats the model parameters as constants. We can use more explicit notation by writing $\hat{y} = f_{\mathbf{w}}(\mathbf{x})$, where $\mathbf{w}$ represents the model parameters and indicates that the function relies on them, but treats them as constants.

However, when *training* a neural network, we do quite the opposite: all the training examples are grouped in a matrix $\mathbf{X}$, all the labels are grouped in a vector $\mathbf{y}$, and both $\mathbf{X}$ and $\mathbf{y}$ are treated as constants, while $\mathbf{w}$ is treated as variable: specifically, we try to minimize the cost function $\mathcal{L}_{\mathbf{X},\mathbf{y}}(\mathbf{w}) = g(f_{\mathbf{X}}(\mathbf{w}), \mathbf{y})$, where g is a function that measures the "discrepancy" between the predictions $f_{\mathbf{X}}(\mathbf{w})$ and the labels $\mathbf{y}$, where $f_{\mathbf{X}}(\mathbf{w})$ represents the vector containing the predictions for each training example. Minimizing the loss function is usually performed using Gradient Descent (or a variant of GD): we start with random model parameters $\mathbf{w}_0$, then we compute $\nabla \mathcal{L}(\mathbf{w}_0)$ and we use this gradient vector to perform a Gradient Descent step, then we repeat the process until convergence. It is crucial to understand that the gradient of the loss function is with regards to the model parameters $\mathbf{w}$ (*not* the inputs $\mathbf{x}$).

Jacobians

Until now, we have only considered functions that output a scalar, but it is possible to output vectors instead. For example, a classification neural network typically outputs one probability for each class, so if there are m classes, the neural network will output a d-dimensional vector for each input.

In Deep Learning we generally only need to differentiate the loss function, which almost always outputs a single scalar number. But suppose for a second that you want to differentiate a function $\mathbf{f}(\mathbf{x})$ which outputs d-dimensional vectors. The good news is that you can treat each *output* dimension independently of the others. This will give you a partial derivative for each input dimension and each output dimension. If you put them all in a single matrix, with one column per input dimension and one row per output dimension, you get the so-called **Jacobian matrix**.

$$
\mathbf{J_f(x_A)} =
\begin{pmatrix}
\dfrac{\partial f_1}{\partial x_1}(\mathbf{x_A}) & \dfrac{\partial f_1}{\partial x_2}(\mathbf{x_A}) & \cdots & \dfrac{\partial f_1}{\partial x_n}(\mathbf{x_A}) \\[2ex]
\dfrac{\partial f_2}{\partial x_1}(\mathbf{x_A}) & \dfrac{\partial f_2}{\partial x_2}(\mathbf{x_A}) & \cdots & \dfrac{\partial f_2}{\partial x_n}(\mathbf{x_A}) \\[2ex]
\vdots & \vdots & \ddots & \vdots \\[2ex]
\dfrac{\partial f_m}{\partial x_1}(\mathbf{x_A}) & \dfrac{\partial f_m}{\partial x_2}(\mathbf{x_A}) & \cdots & \dfrac{\partial f_m}{\partial x_n}(\mathbf{x_A})
\end{pmatrix}
$$

The partial derivatives themselves are often called the **Jacobians**. It's just the first order partial derivatives of the function $\mathbf{f}$.

▾ Hessians

Let's come back to a function $f(\mathbf{x})$ which takes an n-dimensional vector as input and outputs a scalar. If you determine the equation of the partial derivative of f with regards to x_i (the ith component of $\mathbf{x}$), you will get a new function of $\mathbf{x}$: $\dfrac{\partial f}{\partial x_i}$. You can then compute the partial derivative of this function with regards to x_j (the jth component of $\mathbf{x}$). The result is a partial derivative of a partial derivative: in other words, it is a **second order partial derivatives**, also called a **Hessian**. It is noted $\mathbf{x}$: $\dfrac{\partial^2 f}{\partial x_j x_i}$. If $i \neq j$ then it is called a **mixed second order partial derivative**. Or else, if $j = i$, it is noted $\dfrac{\partial^2 f}{\partial x_i{}^2}$

Let's look at an example: $f(x, y) = \sin(xy)$. As we showed earlier, the first order partial derivatives of f are: $\dfrac{\partial f}{\partial x} = y\cos(xy)$ and $\dfrac{\partial f}{\partial y} = x\cos(xy)$. So we can now compute all the Hessians (using the derivative rules we discussed earlier):

- $\dfrac{\partial^2 f}{\partial x^2} = \dfrac{\partial f}{\partial x}[y\cos(xy)] = -y^2\sin(xy)$

- $\dfrac{\partial^2 f}{\partial y\, \partial x} = \dfrac{\partial f}{\partial y}[y\cos(xy)] = \cos(xy) - xy\sin(xy)$

- $\dfrac{\partial^2 f}{\partial x\, \partial y} = \dfrac{\partial f}{\partial x}[x\cos(xy)] = \cos(xy) - xy\sin(xy)$

- $\dfrac{\partial^2 f}{\partial y^2} = \dfrac{\partial f}{\partial y}[x\cos(xy)] = -x^2\sin(xy)$

Note that $\dfrac{\partial^2 f}{\partial x\, \partial y} = \dfrac{\partial^2 f}{\partial y\, \partial x}$. This is the case whenever all the partial derivatives are defined and continuous in a neighborhood around the point at which we differentiate.

The matrix containing all the Hessians is called the **Hessian matrix**:

$$\mathbf{H}_f(\mathbf{x_A}) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1{}^2}(\mathbf{x_A}) & \dfrac{\partial^2 f}{\partial x_1\, \partial x_2}(\mathbf{x_A}) & \cdots & \dfrac{\partial^2 f}{\partial x_1\, \partial x_n}(\mathbf{x_A}) \\ \dfrac{\partial^2 f}{\partial x_2\, \partial x_1}(\mathbf{x_A}) & \dfrac{\partial^2 f}{\partial x_2{}^2}(\mathbf{x_A}) & \cdots & \dfrac{\partial^2 f}{\partial x_2\, \partial x_n}(\mathbf{x_A}) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\, \partial x_1}(\mathbf{x_A}) & \dfrac{\partial^2 f}{\partial x_n\, \partial x_2}(\mathbf{x_A}) & \cdots & \dfrac{\partial^2 f}{\partial x_n{}^2}(\mathbf{x_A}) \end{pmatrix}$$

There are great optimization algorithms which take advantage of the Hessians, but in practice Deep Learning almost never uses them. Indeed, if a function has n variables, there are n^2 Hessians:

since neural networks typically have several millions of parameters, the number of Hessians would exceed thousands of billions. Even if we had the necessary amount of RAM, the computations would be prohibitively slow.

A few proofs

Let's finish by proving all the derivative rules we listed earlier. You don't have to go through all these proofs to be a good Deep Learning practitioner, but it may help you get a deeper understanding of derivatives.

Constant: $f(x) = c$

$$
\begin{aligned}
f'(x) &= \lim_{\epsilon \to 0} \frac{f(x + \epsilon) - f(x)}{\epsilon} && \text{by definition} \\
&= \lim_{\epsilon \to 0} \frac{c - c}{\epsilon} && \text{using } f(x) = c \\
&= \lim_{\epsilon \to 0} 0 && \text{since } c - c = 0 \\
&= 0 && \text{since the limit of a constant is that constant}
\end{aligned}
$$

Product rule: $f(x) = g(x)h(x)$

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

$$= \lim_{\epsilon \to 0} \frac{g(x+\epsilon)h(x+\epsilon) - g(x)h(x)}{\epsilon} \qquad\qquad \text{usi}$$

$$= \lim_{\epsilon \to 0} \frac{g(x+\epsilon)h(x+\epsilon) - g(x)h(x+\epsilon) + g(x)h(x+\epsilon) - g(x)h(x)}{\epsilon} \qquad\qquad \text{sul}$$

$$= \lim_{\epsilon \to 0} \frac{g(x+\epsilon)h(x+\epsilon) - g(x)h(x+\epsilon)}{\epsilon} + \lim_{\epsilon \to 0} \frac{g(x)h(x+\epsilon) - g(x)h(x)}{\epsilon} \qquad\qquad \text{sin}$$

$$= \lim_{\epsilon \to 0} \left[\frac{g(x+\epsilon) - g(x)}{\epsilon} h(x+\epsilon) \right] + \lim_{\epsilon \to 0} \left[g(x) \frac{h(x+\epsilon) - h(x)}{\epsilon} \right] \qquad\qquad \text{fac}$$

$$= \lim_{\epsilon \to 0} \left[\frac{g(x+\epsilon) - g(x)}{\epsilon} h(x+\epsilon) \right] + g(x) \lim_{\epsilon \to 0} \frac{h(x+\epsilon) - h(x)}{\epsilon} \qquad\qquad \text{tal}$$

$$= \lim_{\epsilon \to 0} \left[\frac{g(x+\epsilon) - g(x)}{\epsilon} h(x+\epsilon) \right] + g(x)h'(x) \qquad\qquad \text{usi}$$

$$= \lim_{\epsilon \to 0} \left[\frac{g(x+\epsilon) - g(x)}{\epsilon} \right] \lim_{\epsilon \to 0} h(x+\epsilon) + g(x)h'(x) \qquad\qquad \text{sin}$$

$$= \lim_{\epsilon \to 0} \left[\frac{g(x+\epsilon) - g(x)}{\epsilon} \right] h(x) + h(x)g'(x) \qquad\qquad \text{sin}$$

$$= g'(x)h(x) + g(x)h'(x) \qquad\qquad \text{usi}$$

Note that if $g(x) = c$ (a constant), then $g'(x) = 0$, so the equation simplifies to:

$$f'(x) = c\,h'(x)$$

Chain rule: $f(x) = g(h(x))$

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

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

$$= \lim_{\epsilon \to 0} \left[\frac{h(x + \epsilon) - h(x)}{h(x + \epsilon) - h(x)} \frac{g(h(x + \epsilon)) - g(h(x))}{\epsilon} \right]$$ multiplying and di

$$= \lim_{\epsilon \to 0} \left[\frac{h(x + \epsilon) - h(x)}{\epsilon} \frac{g(h(x + \epsilon)) - g(h(x))}{h(x + \epsilon) - h(x)} \right]$$ swapping the deno

$$= \lim_{\epsilon \to 0} \left[\frac{h(x + \epsilon) - h(x)}{\epsilon} \right] \lim_{\epsilon \to 0} \left[\frac{g(h(x + \epsilon)) - g(h(x))}{h(x + \epsilon) - h(x)} \right]$$ the limit of a produ

$$= h'(x) \lim_{\epsilon \to 0} \left[\frac{g(h(x + \epsilon)) - g(h(x))}{h(x + \epsilon) - h(x)} \right]$$ using the definitior

$$= h'(x) \lim_{\epsilon \to 0} \left[\frac{g(u) - g(v)}{u - v} \right]$$ using $u = h(x + \epsilon)$

$$= h'(x) \lim_{u \to v} \left[\frac{g(u) - g(v)}{u - v} \right]$$ since h is continuc

$$= h'(x) g'(v)$$ using the definitio

$$= h'(x) g'(h(x))$$ since $v = h(x)$

Exponential: $f(x) = \exp(x) = e^x$

There are several equivalent definitions of the number e. One of them states that e is the unique

positive number for which $\lim_{\epsilon \to 0} \dfrac{e^\epsilon - 1}{\epsilon} = 1$. We will use this in this proof:

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

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

$$= \lim_{\epsilon \to 0} \frac{e^x e^\epsilon - e^x}{\epsilon}$$ using the fact that $x^{a+b} = x^a x^b$

$$= \lim_{\epsilon \to 0} \left[e^x \frac{e^\epsilon - 1}{\epsilon} \right]$$ factoring out e^x

$$= \lim_{\epsilon \to 0} e^x \lim_{\epsilon \to 0} \frac{e^\epsilon - 1}{\epsilon}$$ the limit of a product is the product of the limits

$$= \lim_{\epsilon \to 0} e^x$$ since $\lim_{\epsilon \to 0} \dfrac{e^\epsilon - 1}{\epsilon} = 1$

$$= e^x$$ since e^x does not depend on ϵ

Logarithm: $f(x) = \ln(x)$

Another definition of the number e is:

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

By defining $\epsilon = \frac{1}{n}$, we can rewrite the previous definition as:

$$e = \lim_{\epsilon \to 0} (1 + \epsilon)^{1/\epsilon}$$

This will come in handy in a second:

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

$$= \lim_{\epsilon \to 0} \frac{\ln(x + \epsilon) - \ln(x)}{\epsilon} \qquad \text{using } f(x) = \ln(x)$$

$$= \lim_{\epsilon \to 0} \frac{\ln\left(\dfrac{x + \epsilon}{x}\right)}{\epsilon} \qquad \text{since } \ln(a) - \ln(b) = \ln\left(\frac{a}{b}\right)$$

$$= \lim_{\epsilon \to 0} \left[\frac{1}{\epsilon} \ln\left(1 + \frac{\epsilon}{x}\right)\right] \qquad \text{just moving things around a bit}$$

$$= \lim_{\epsilon \to 0} \left[\frac{1}{xu} \ln(1 + u)\right] \qquad \text{defining } u = \frac{\epsilon}{x} \text{ and thus } \epsilon = xu$$

$$= \lim_{u \to 0} \left[\frac{1}{xu} \ln(1 + u)\right] \qquad \text{replacing } \lim_{\epsilon \to 0} \text{ with } \lim_{u \to 0} \text{ since } \lim_{\epsilon \to 0} u = 0$$

$$= \lim_{u \to 0} \left[\frac{1}{x} \ln\left((1 + u)^{1/u}\right)\right] \qquad \text{since } a\ln(b) = \ln(b^a)$$

$$= \frac{1}{x} \lim_{u \to 0} \left[\ln\left((1 + u)^{1/u}\right)\right] \qquad \text{taking } \frac{1}{x} \text{ out since it does not depend on } \epsilon$$

$$= \frac{1}{x} \ln\left(\lim_{u \to 0} (1 + u)^{1/u}\right) \qquad \text{taking } \ln \text{ out since it is a continuous function}$$

$$= \frac{1}{x} \ln(e) \qquad \text{since } e = \lim_{u \to 0} (1 + u)^{1/u}$$

$$= \frac{1}{x} \qquad \text{since } \ln(e) = 1$$

Power rule: $f(x) = x^r$, with $r \neq 0$

Let's define $g(x) = e^x$ and $h(x) = \ln(x^r)$. Since $a = e^{\ln(a)}$, we can rewrite f as $f(x) = g(h(x))$, which allows us to use the chain rule:

$$f'(x) = h'(x)g'(h(x))$$

We know the derivative of the exponential: $g'(x) = e^x$. We also know the derivative of the natural logarithm: $\ln'(x) = \dfrac{1}{x}$ so $h'(x) = \dfrac{r}{x}$. Therefore:

$$f'(x) = \frac{r}{x}e^{\ln(x^r)}$$

Since $e^{\ln(a)} = a$, this equation simplifies to:

$$f'(x) = \frac{r}{x}x^r$$

And finally:

$$f'(x) = rx^{r-1}$$

Note that the power rule works for any $r \neq 0$, including negative numbers and real numbers. For example:

- if $f(x) = \dfrac{1}{x} = x^{-1}$, then $f'(x) = -x^{-2} = -\dfrac{1}{x^2}$.
- if $f(x) = \sqrt{x} = x^{1/2}$, then $f'(x) = \dfrac{1}{2}x^{-1/2} = \dfrac{1}{}$

Multiplicative inverse: $f(x) = \dfrac{1}{h(x)}$

First, let's define $g(x) = \dfrac{1}{x}$. This leads to $f(x) = g(h(x))$. Now we can use the chain rule:

$$f'(x) = h'(x)g'(h(x))$$

Since $g(x) = x^{-1}$, we can use the power rule to find $g'(x) = -\dfrac{1}{x^2}$

Finally, we get:

$$f'(x) = -\frac{h'(x)}{h^2(x)}$$

Quotient rule: $f(x) = \dfrac{g(x)}{h(x)}$

Let's rewrite $f(x)$ as a product: $f(x) = g(x)u(x)$ with $u(x) = \dfrac{1}{h(x)}$

We can now use the product rule to get:

$$f(x) = g'(x)u(x) + g(x)u'(x)$$

Replacing $u(x)$ with $\dfrac{1}{h(x)}$ and using the result of the previous section to replace $u'(x)$ with $\dfrac{-h'(x)}{h^2(x)}$, we get:

$$f(x) = g'(x)\frac{1}{h(x)} + g(x)\frac{-h'(x)}{h^2(x)}$$

Now we multiply and divide the first term by $h(x)$:

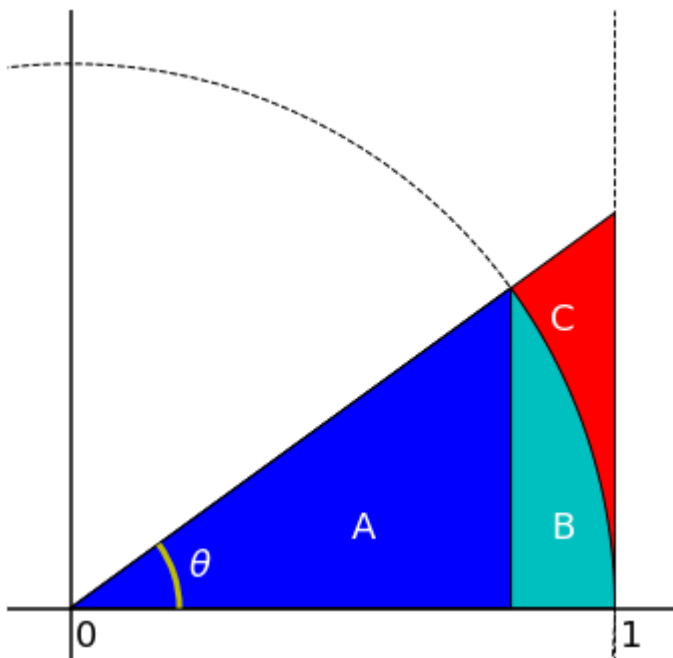$$f(x) = \frac{g'(x)h(x)}{h^2(x)} - \frac{g(x)h'(x)}{h^2(x)}$$

And finally:

$$f(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h^2(x)}$$

▾ Sin: $f(x) = \sin(x)$

For this proof we will first need to prove that $\lim\limits_{\theta \to 0} \dfrac{\sin(\theta)}{\theta} = 1$. One way to do that is to consider the following diagram:

Show code

The circle is the unit circle (radius=1).

Assuming $0 < \theta < \dfrac{\pi}{2}$, the area of the blue triangle (area A) is equal to its height ($\sin(\theta)$) times its base ($\cos(\theta)$) divided by 2. So $A = \dfrac{1}{2}\sin(\theta)\cos(\theta)$.

The unit circle has an area of π, so the circular sector (in the shape of a pizza slice) has an area of $A + B = \pi\dfrac{\theta}{2\pi} = \dfrac{\theta}{2}$.

Next, the large triangle ($A + B + C$) has an area equal to its height ($\tan(\theta)$) multiplied by its base (of length 1) divided by 2, so $A + B + C = \dfrac{\tan(\theta)}{2}$.

When $0 < \theta < \dfrac{\pi}{2}$, we have $A < A + B < A + B + C$, therefore:

$$\frac{1}{2}\sin(\theta)\cos(\theta) < \frac{\theta}{2} < \frac{\tan(\theta)}{2}$$

We can multiply all the terms by 2 to get rid of the $\dfrac{1}{2}$ factors. We can also divide by $\sin(\theta)$, which is strictly positive (assuming $0 < \theta < \dfrac{\pi}{2}$), so the inequalities still hold:

$$cos(\theta) < \frac{\theta}{\sin(\theta)} < \frac{\tan(\theta)}{\sin(\theta)}$$

Recall that $\tan(\theta) = \dfrac{\sin(\theta)}{\cos(\theta)}$, so the last term simplifies like so:

$$cos(\theta) < \frac{\theta}{\sin(\theta)} < \frac{1}{\cos(\theta)}$$

Since all these terms are strictly positive when $0 < \theta < \dfrac{\pi}{2}$, we can take their inverse and change the direction of the inequalities:

Inequalities (1)

$$\frac{1}{cos(\theta)} > \frac{\sin(\theta)}{\theta} > \cos(\theta)$$

assuming $0 < \theta < \dfrac{\pi}{2}$

Now since $\sin(-\theta) = -\sin(\theta)$, we see that $\dfrac{\sin(-\theta)}{-\theta} = \dfrac{\sin(\theta)}{\theta}$. Moreover, $\cos(-\theta) = \cos(\theta)$, and therefore $\dfrac{1}{\cos(-\theta)} = \dfrac{1}{\cos(\theta)}$. Replacing the terms in the inequalities (1), we get:

$$\frac{1}{\cos(-\theta)} > \frac{\sin(-\theta)}{-\theta} > \cos(-\theta)$$

assuming $0 < \theta < \dfrac{\pi}{2}$

In other words, the inequalities (1) also hold for $-\dfrac{\pi}{2} < \theta < 0$:

Inequalities (2)

$$\frac{1}{\cos(\theta)} > \frac{\sin(\theta)}{\theta} > \cos(\theta)$$

assuming $-\dfrac{\pi}{2} < \theta < \dfrac{\pi}{2}$ and $\theta \neq 0$

Since cos is a continuous function, $\lim\limits_{\theta \to 0} \cos(\theta) = \cos(0) = 1$. Similarly,

$$\lim_{\theta \to 0} \frac{1}{\cos(\theta)} = \frac{1}{\cos(0)} = 1.$$

Since the inequalities (2) tell us that $\dfrac{\sin(\theta)}{\theta}$ is squeezed between $\dfrac{1}{\cos(\theta)}$ and $\cos(\theta)$ when θ is

close to 0, and since both of these approach 1 when θ approaches 0, we can use the **squeeze theorem** (also called the **sandwich theorem**) to conclude that $\dfrac{\sin(\theta)}{\theta}$ must also approach 1 when θ

approaches 0.

We have proven that:

$$\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} = 1$$

Now the second thing we need to prove before we can tackle the derivative of the sin function is

the fact that $\lim\limits_{\theta \to 0} \dfrac{\cos(\theta) - 1}{\theta} = 0$. Here we go:

$$\lim_{\theta \to 0} \frac{\cos(\theta) - 1}{\theta} = \lim_{\theta \to 0} \frac{\cos(\theta) - 1}{\theta} \frac{\cos(\theta) + 1}{\cos(\theta) + 1}$$ multiplying and dividing by $\cos$

$$= \lim_{\theta \to 0} \frac{\cos^2(\theta) - 1}{\theta(\cos(\theta) + 1)}$$ since $(a - 1)(a + 1) = a^2 - 1$

$$= \lim_{\theta \to 0} \frac{-\sin^2(\theta)}{\theta(\cos(\theta) + 1)}$$ since $\cos^2(\theta) - 1 = -\sin^2(\theta)$

$$= -\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} \frac{\sin(\theta)}{\cos(\theta) + 1}$$ just rearranging the terms

$$= -\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} \lim_{\theta \to 0} \frac{\sin(\theta)}{\cos(\theta) + 1}$$ since the limit of a product is the

$$= -\lim_{\theta \to 0} \frac{\sin(\theta)}{\cos(\theta) + 1}$$ since $\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} = 1$

$$= -\frac{0}{1 + 1}$$ since $\lim_{\theta \to 0} \sin(\theta) = 0$ and $\lim_{\theta \to 0} \cos($

$$= 0$$

We have proven that:

$$\lim_{\theta \to 0} \frac{\cos(\theta) - 1}{\theta} = 0$$

With all that, we are finally ready to prove the fact that the derivative of the sin function is the cos function:

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

$$= \lim_{\theta \to 0} \frac{\sin(x + \theta) - \sin(x)}{\theta}$$ using $f(x) = \sin(x)$

$$= \lim_{\theta \to 0} \frac{\cos(x)\sin(\theta) + \sin(x)\cos(\theta) - \sin(x)}{\theta}$$ since $\sin(a + b) = \cos(a)$

$$= \lim_{\theta \to 0} \frac{\cos(x)\sin(\theta)}{\theta} + \lim_{\theta \to 0} \frac{\sin(x)\cos(\theta) - \sin(x)}{\theta}$$ since the limit of a sum i

$$= \cos(x) \lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} + \sin(x) \lim_{\theta \to 0} \frac{\cos(\theta) - 1}{\theta}$$ bringing out $\cos(x)$ and

$$= \cos(x) \lim_{\theta \to 0} \frac{\sin(\theta)}{\theta}$$ since $\lim_{\theta \to 0} \frac{\cos(\theta) - 1}{\theta} = 0$

$$= \cos(x)$$ since $\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} = 1$

Cosine: $f(x) = \cos(x)$

Since we have proven that $\sin'(x) = \cos(x)$, proving that $\cos'(x) = -\sin(x)$ will be much easier.

$$
\begin{aligned}
f'(x) &= \lim_{\theta \to 0} \frac{f(x + \theta) - f(x)}{\theta} & & \text{by definition} \\[2mm]
&= \lim_{\theta \to 0} \frac{\cos(x + \theta) - \cos(x)}{\theta} & & \text{using } f(x) = \cos(x) \\[2mm]
&= \lim_{\theta \to 0} \frac{\sin\left(x + \dfrac{\pi}{2} + \theta\right) - \sin\left(x + \dfrac{\pi}{2}\right)}{\theta} & & \text{since } \cos(x) = \sin\left(x + \dfrac{\pi}{2}\right) \\[2mm]
&= \lim_{\theta \to 0} \frac{\sin(z + \theta) - \sin(z)}{\theta} & & \text{using } z = x + \dfrac{\pi}{2} \\[2mm]
&= \sin'(z) & & \text{using the definition of } \sin'(z) \\[1mm]
&= \cos(z) & & \text{since we proved that } \sin'(z) = \cos(\\[1mm]
&= \cos\left(x + \dfrac{\pi}{2}\right) & & \text{using the definition of } z \\[1mm]
&= -\sin(x) & & \text{using this well-known rule of trigo}
\end{aligned}
$$

▾ Tangent: $f(x) = \tan(x)$

Since $\tan(x) = \dfrac{\sin(x)}{\cos(x)}$, we can use the quotient rule to get:

$$
f'(x) = \frac{\sin'(x)\cos(x) - \sin(x)\cos'(x)}{\cos^2(x)}
$$

Since $\sin'(x) = \cos(x)$ and $\cos'(x) = -\sin(x)$, we get:

$$
f'(x) = \frac{\cos^2(x) + \sin^2(x)}{\cos^2(x)}
$$

And finally, since $\cos^2(x) + \sin^2(x) = 1$, we get:

$$
f'(x) = \frac{1}{\cos^2(x)}
$$