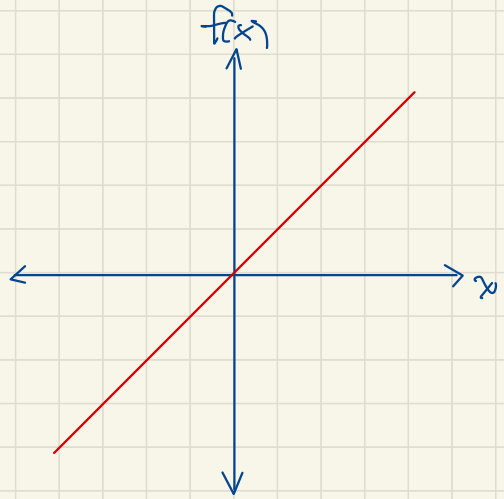
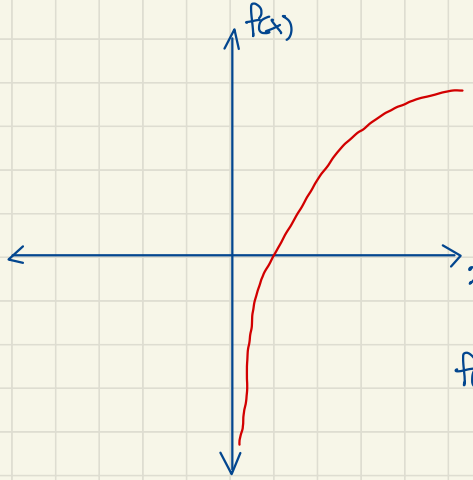
Calculus

Calculus is a subject in mathematics that studies how values of a function grow or shrink.

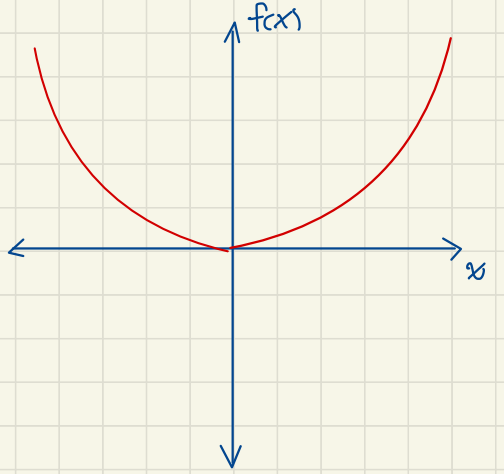
A function maps each value of x) to another value (output) f(x).
We can plot the f(x) to understand its shape. Notice where f(x) intersects with x axis and y axis and where maximum and minimum are.



$f(x) = x$



$f(x) = \log(x)$



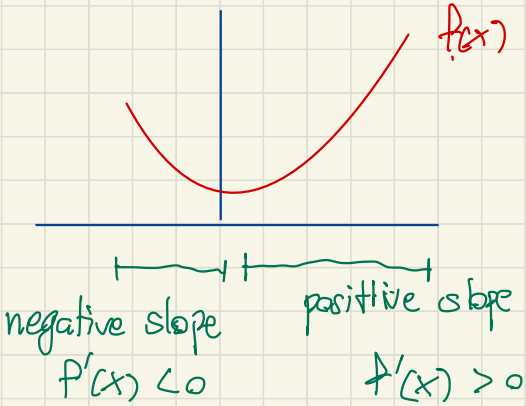
$f(x) = x^2$

Derivative

The derivative of $f(x)$ is a function that measures the rate of change or slope at a point.

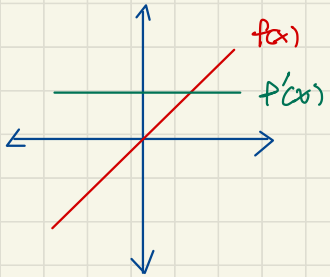
Positive slope means if we increase x at that point for a little bit then $f(x)$ will increase.

We use calculus to find $f'(x)$ or $\frac{d}{dx} f(x)$ derivative of $f(x)$. For example,



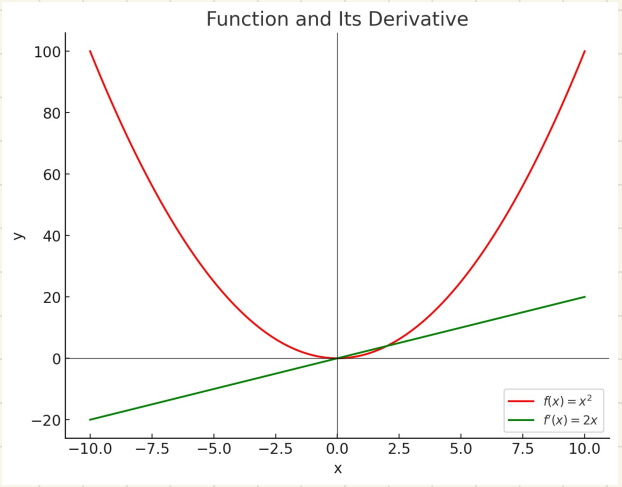
$f(x) = x$

$\frac{d}{dx} f(x) = 1$



$f(x) = x^2$

$\frac{d}{dx} f(x) = 2x$



Mathematicians come up with a few formulas for finding $\frac{d}{dx} f(x)$ of many functions such as:

$$\frac{d}{dx} x^n = n \cdot x^{(n-1)} \qquad \frac{d}{dx} c = 0$$

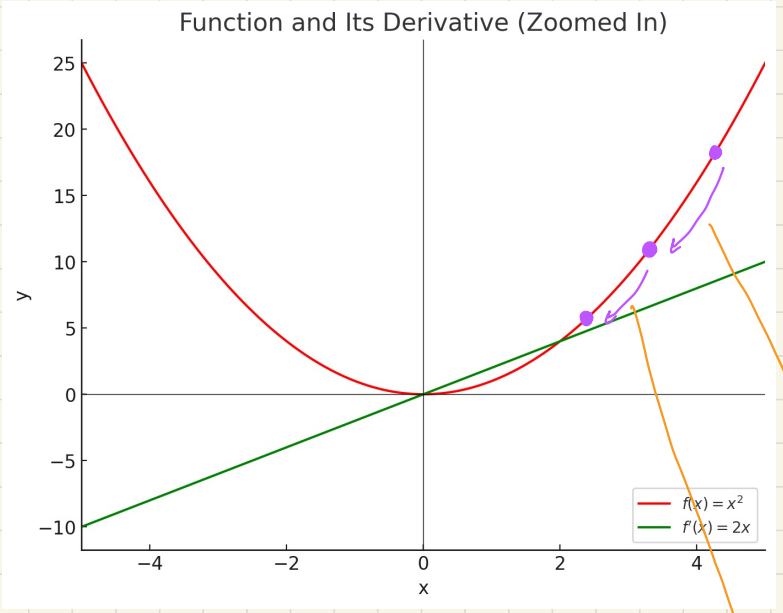
$$\frac{d}{dx} ax + b = \frac{d}{dx} ax + \frac{d}{dx} b$$

$$= a$$

$$\frac{d}{dx} \log(x) = \frac{1}{x} \quad \text{and many more}$$

Derivatives are useful in finding the minimum value of $f(x)$. because $f'(x)$ tell us whether increasing or decreasing x will lead to the decrease in $f(x)$

In NLP, x is the parameter and
$f(x)$ is a loss function that we want to minimize



Start at a random point

$f(x) = x^2$

$f'(x) = 2x$

$x = 2$

$f'(4) = 2 \cdot 4 = 8$

Update $x \leftarrow x - \alpha \cdot f'(x)$

$x \leftarrow 4 - 0.1 \cdot 8$

$x = 3.2$

$f(4) = 4^2 = 16$

decreasing

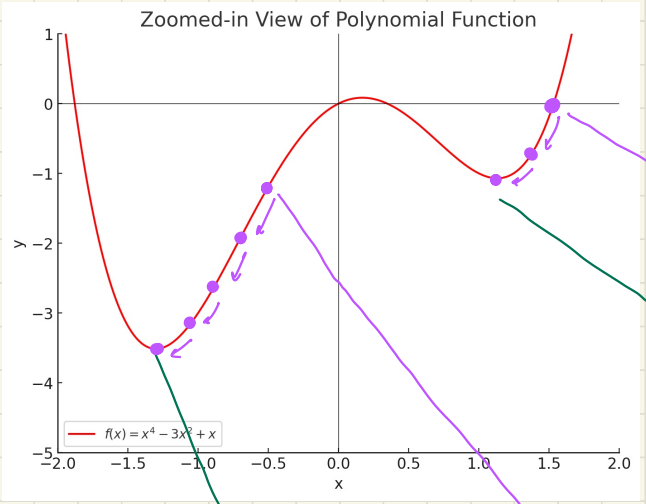
$f(3.2) = 3.2^2 = 10.24$

$f'(3.2) = 2 \cdot 3.2 = 6.4$

Update $x \leftarrow 3.2 - 0.1 \cdot 6.4$ decreasing

$x = 2.56$

$f(2.56) = 6.55$



This is a case where gradient descent gives us 'local minimum' instead of global minimum.

we start at the wrong place.

local minimum

In NLP, the loss function is quite complex, and we can never know whether we obtain the global minimum or not. We are just approximating the 'best' model.

global minimum

we start at the right place

Partial Derivative

Some functions have many variables such as

$$f(x_1, x_2, x_3) = x_1 + x_2^2 + x_3^3 + 10$$

We can find $f'(x_1, x_2, x_3)$ to find the rate of change (slope) with respect to each variable. This is called partial derivative.

$$\frac{\partial}{\partial x_1} f(x_1, x_2, x_3) = 1$$

$$\frac{\partial}{\partial x_2} f(x_1, x_2, x_3) = 2x_2$$

$$\frac{\partial}{\partial x_3} f(x_1, x_2, x_3) = 3x_3^2$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

If we think of the variables as a vector then the gradient of $f(x_1, x_2, x_3)$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 1 \\ 2x_2 \\ 3x_3^2 \end{bmatrix}$$

In NLP, loss function has many variables (parameters) and is much more complicated.

$L(W, b)$ — b: bias vector; W: weight matrix

In multivariable cases, gradient descent algorithm remains unchanged

start with random values
compute gradient over all parameters
update the parameters

$$W := W_{old} - \alpha \cdot \nabla L(W)$$