Calculus as such pervades all domains and has a lot to be studied if we had that goal in mind. In our context we will largely focus on derivatives and optimization as that is what we need for ML and AI models. Goal is to understand things intuitively and to know fundamentals of what goes on at the back end of model development in terms of maths and stats being applied on the data and how it relates to the outcomes that you get and practices that you follow. We will start our discussion with derivatives .

Introduction to Derivatives

derivatives at a very basic level are rate of change of one quantity with respect to another . They play a pivotal role in optimization . When we know how a quantity changes with respect to another , it allows to find when does it reach its maximum or minimum as desired . We will not get into optimization right away, first lets familiarize ourselves with finding derivatives.

Consider the relationship

$$f(x) = 3x$$

Lets try to find out rate of change of f with respect to x . If I change x to $x + h$, thus changing it by an amount h, f changes from $f(x)$ to $f(x + h)$.

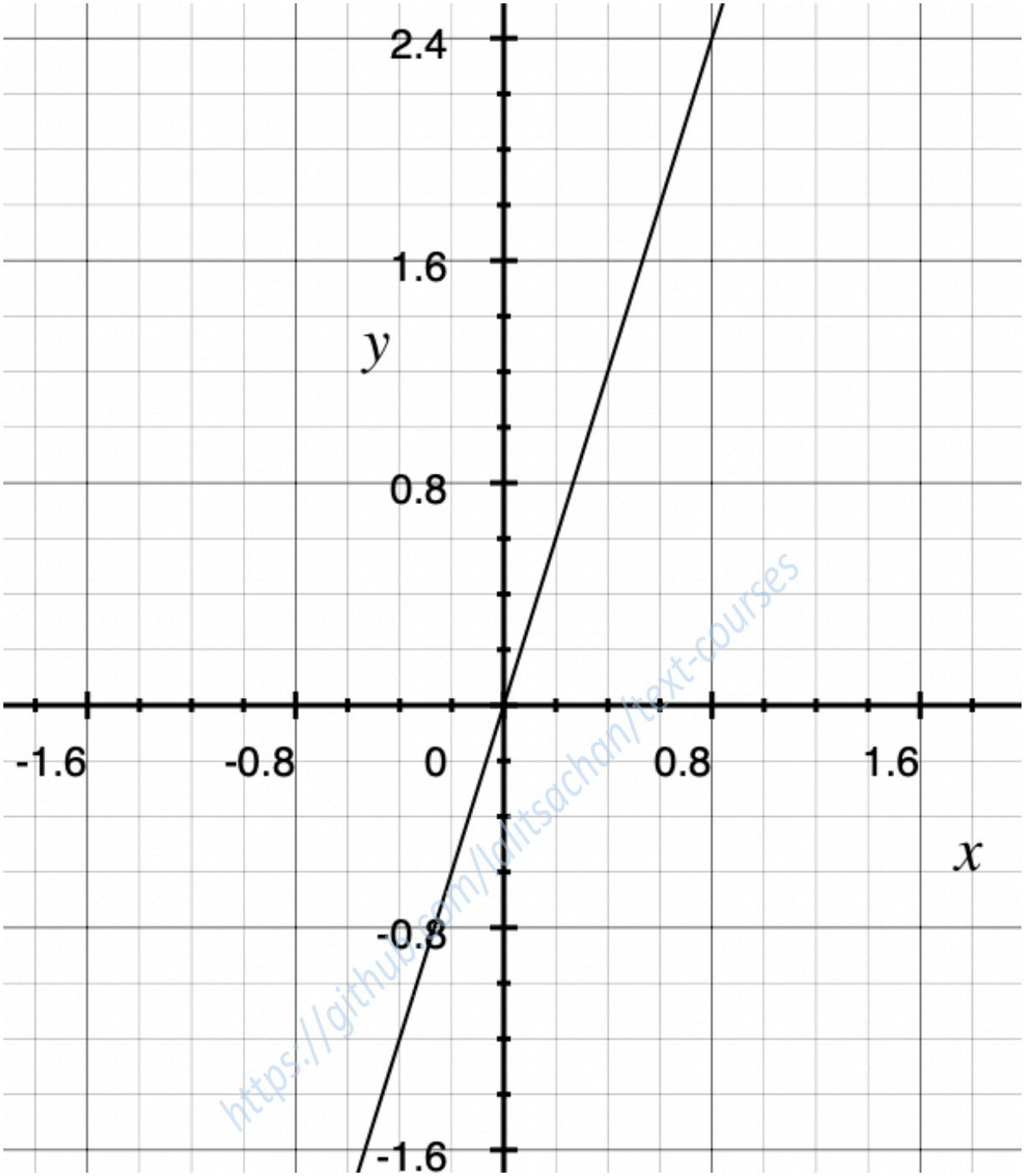
we can define the rate of change of f with respect to x then as

$$= \frac{\partial f}{\partial x} = \frac{f(x+h) - f(x)}{x + h - x} = \frac{f(x+h) - f(x)}{h}$$

You might be familiar with derivative of f with respect to x being written as $\frac{df}{df}$ or $f'(x)$ or sometimes simply f' . Lets complete this exercise for $f(x) = 3x$.

$$\begin{aligned} f'(x) &= \frac{f(x+h) - f(x)}{h} \\ f'(x) &= \frac{3(x+h) - 3x}{h} \\ f'(x) &= \frac{3h}{h} \\ f'(x) &= 3 \end{aligned}$$

https://github.com/lalitsachan/text-courses



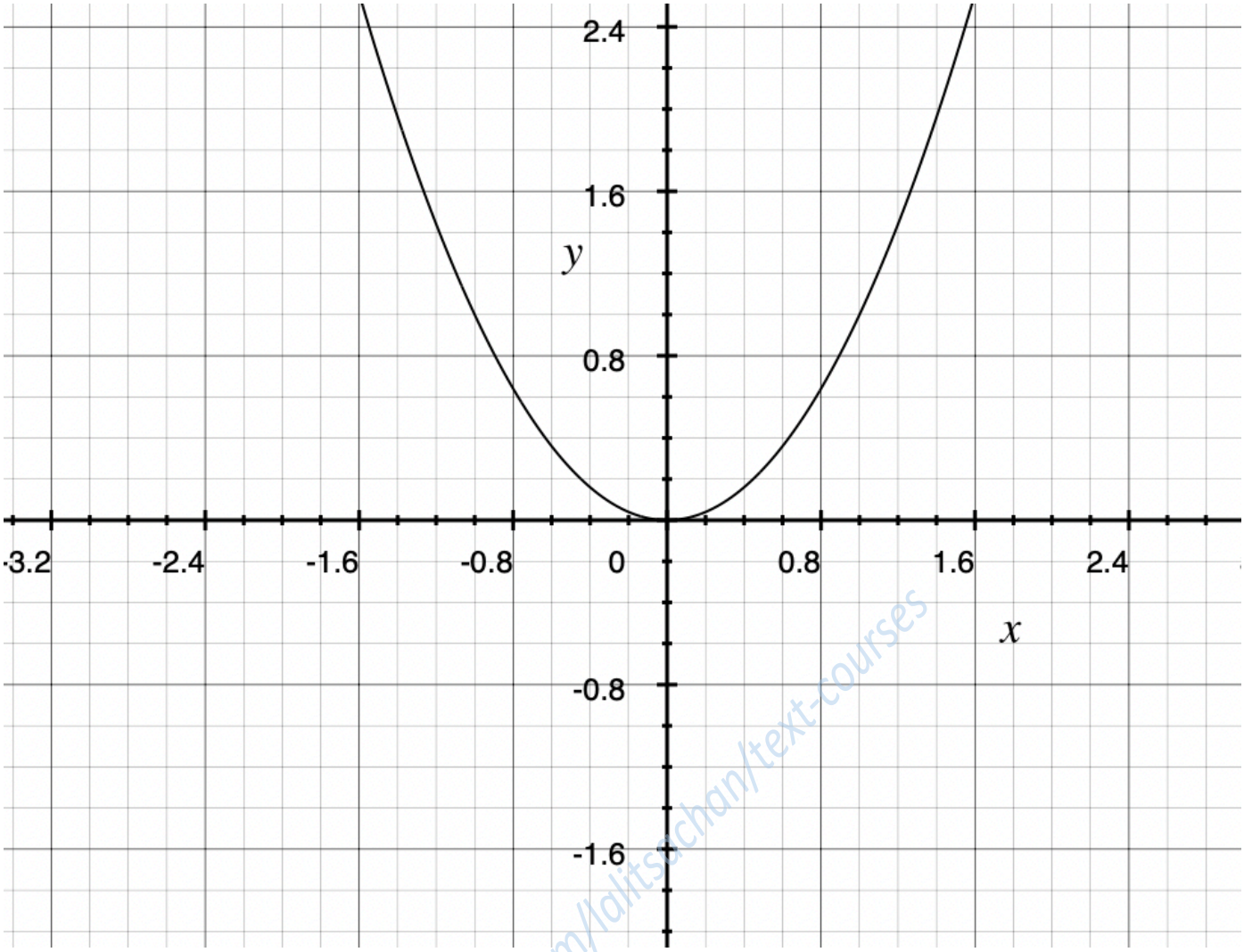
https://github.com/lalitsachan/text-courses

So rate of change of f with respect to x is 3, which makes sense as the relationship $f(x) = 3x$ easily suggests that, if we change x by some amount , f will simply change by 3 times that amount .

However as evident in the figure above and then in our results , this rate is constant and does not change irrespective whatever amount of change h we consider. Which is not always the case .

Consider another relationship :

$$f(x) = x^2$$



Look at the figure , clearly rate of change of f with respect to x is not constant as was the case was earlier . We can see that for small value of x , f changes at much slower pace in comparison to higher values of x . As x increases; f starts to rise much faster, which implies that, rate of change of f with respect to x is also increasing continuously.

The kind of functional relationships we will be working with are going to be of the same kind where rate of change of f changes with values of x . This is called instantaneous rate of change. If we apply the usual method with large values of h , we will get a bad approximation of this instantaneous rate of change which assumes that rate of change remains the same for that interval which isn't true. In order to get true value for this instantaneous rate of change around any given value of x, we will have to consider the change amount h to be infinitesimal tending to 0 . Formally a more generalized method for finding a derivative or rate of change of f with respect to x will be :

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

Considering this modified definition lets find derivative of $f(x) = x^2$.

https://github.com/lalitsachan/text-courses

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$
$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h}$$
$$f'(x) = \lim_{h \to 0} \frac{x^2 + h^2 + 2hx - x^2}{h}$$
$$f'(x) = \lim_{h \to 0} \frac{h^2 + 2hx}{h}$$
$$f'(x) = \lim_{h \to 0} h + 2x$$
$$f'(x) = 2x$$

Using this method we can find out derivative of any function; but thankfully we don't really need to do this. People who came before us have painfully gone through this complex process and derived derivatives for many standard functions , here is a not so exhaustive list .

$f(x)$	$f'(x)$
$Constant$	0
x^n	nx^{n-1}
e^x	e^x
$ln(x)$	$\frac{1}{x}$
a^x	$a^x\ ln(x)$
$sin(x)$	$cos(x)$
$cos(x)$	$-sin(x)$

https://github.com/lalitsachan/text-courses

Problem with lists like these is that they don't help. Reason being that functions that we get to work with are not simply the ones shown in the table above [or in even a more exhaustive version of the same somewhere]

Usually the functions that we get to work with are some complex combinations of the simple functions that we see in the tables like these . Lets see how to find derivatives of these :

Derivatives of compound functions

$$F(x) = af(x) + bg(x)$$

$$F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}$$
$$F'(x) = \lim_{h \to 0} \left(\frac{af(x+h) + bg(x+h) - af(x) + bg(x)}{h} \right)$$
$$F'(x) = \lim_{h \to 0} \left(a \left(\frac{f(x+h) - f(x)}{h} \right) + b \left(\frac{g(x+h) - g(x)}{h} \right) \right)$$
$$F'(x) = a \lim_{h \to 0} \left(\frac{f(x+h) - f(x)}{h} \right) + b \lim_{h \to 0} \left(\frac{g(x+h) - g(x)}{h} \right)$$
$$=> \mathbf{F'(x) = af'(x) + bg'(x)}$$

an example for the same will be :

$$F(x) = 2x^4 + sin(x) + ln(x)$$
$$F'(x) = 8x^3 + cos(x) + \frac{1}{x}$$

$F(x) = f(x) * g(x)$

$$F'(x) = \lim_{h \to 0} \frac{f(x+h) * g(x+h) - f(x) * g(x)}{h}$$
$$F'(x) = \lim_{h \to 0} \frac{f(x+h) * g(x+h) - f(x+h) * g(x) + f(x+h) * g(x) - f(x) * g(x)}{h}$$
$$F'(x) = \lim_{h \to 0} \frac{f(x+h) * g(x+h) - f(x+h) * g(x) + f(x+h) * g(x) - f(x) * g(x)}{h}$$
$$F'(x) = \lim_{h \to 0} f(x+h) \frac{g(x+h) - g(x)}{h} + \lim_{h \to 0} g(x) \frac{f(x+h) - f(x)}{h}$$
$$F'(x) = \lim_{h \to 0} f(x+h) * \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} + g(x) * \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$
$$=> \mathbf{F'(x) = f(x) * g'(x) + f'(x) * g(x)}$$

an example of this kind :

$$F(x) = x^2 * ln(x)$$
$$=> \text{consider } f(x) = x^2 \text{ and } g(x) = ln(x)$$
$$=> \text{which gives us } f'(x) = 2x \text{ and } g'(x) = \frac{1}{x}$$
$$F'(x) = x^2 * \frac{1}{x} + 2x * ln(x)$$
$$= x + 2x * ln(x)$$

$F(x) = \frac{f(x)}{g(x)}$

goes without saying that the in the interval in consideration will not have $g(x) = 0$

https://github.com/lalitsachan/text-courses

$$F'(x) = \lim_{h \to 0} \frac{\frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)}}{h}$$
$$F'(x) = \lim_{h \to 0} \frac{g(x) * f(x+h) - f(x) * g(x+h)}{h * g(x) * g(h+x)}$$
$$F'(x) = \lim_{h \to 0} \frac{g(x) * f(x+h) - g(x) * f(x) + g(x) * f(x) - f(x) * g(x+h)}{h * g(x) * g(h+x)}$$
$$F'(x) = \left(\lim_{h \to 0} \frac{1}{g(h+x)} * \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \right) - \left(\frac{f(x)}{g(x)} * \lim_{h \to 0} \frac{1}{g(x+h)} * \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \right)$$
$$F'(x) = \frac{f'(x)}{g(x)} - \frac{f(x) * g'(x)}{(g(x))^2}$$
$$=> \mathbf{F'(x) = \frac{f'(x) * g(x) - f(x) * g'(x)}{(g(x))^2}}$$

an example of this kind ：

$$F(x) = \frac{sin(x)}{2^x}$$
$$=> \text{consider } f(x) = sin(x) \text{ and } g(x) = 2^x$$
$$=> \text{which gives us } f'(x) = cos(x) \text{ and } g'(x) = 2^x * ln(2)$$
$$F'(x) = \frac{cos(x) * 2^x - sin(x) * 2^x * ln(2)}{2^{2x}}$$

Nested Function $F(x) = f(g(x))$

This is the final kind that we need to get used to and that will complete the discussion on compound function derivatives .

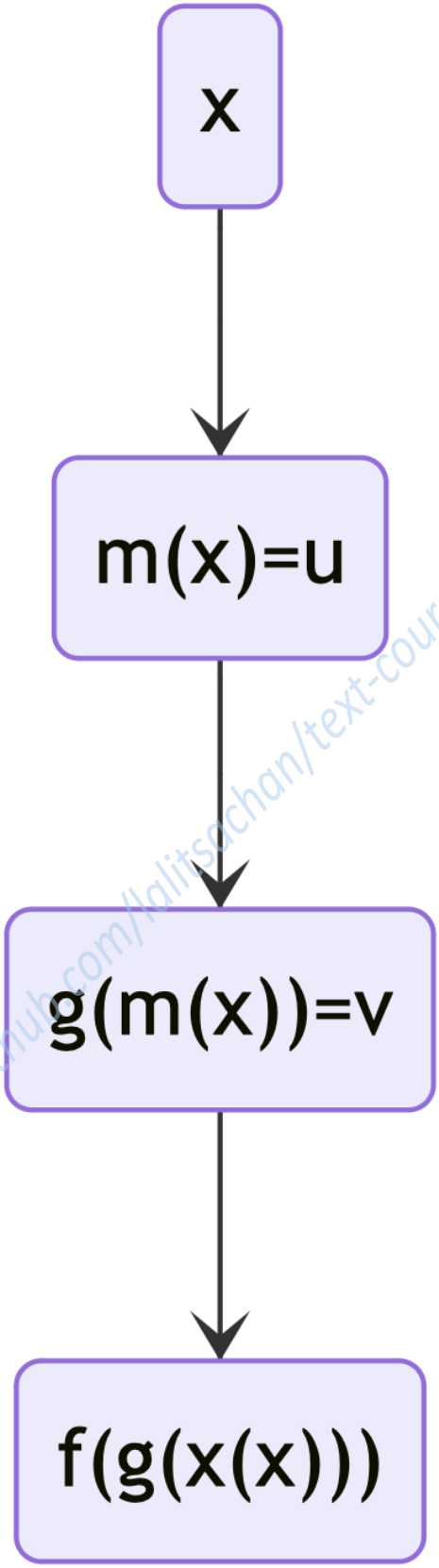
there are many ways to understand this, but we will focus on a method which uses chains of derivatives and is fairly intuitive to understand .

consider this [although, generalizing this requires some mathematical nuance that we are going to ignore to avoid complexity in our discussion and it will not affect our final result]

$$m(x) = u$$
$$g(m(x)) = g(u) = v$$
$$f(g(m(x))) = f(g(u)) = f(v)$$

consider this diagram of dependence

https://github.com/lalitsachan/text-courses



https://github.com/lalitsachan/text-courses

now if we introduce a change in x as Δx, it will cause change in u as Δu, which will further cause flow of change forward as Δv and then Δf. We can use this as :

$$
\begin{aligned}
F'(x) &= \frac{\Delta f}{\Delta x} \\
&= \frac{\Delta f}{\Delta v} * \frac{\Delta v}{\Delta u} * \frac{\Delta u}{\Delta x} \\
&= \frac{\partial f}{\partial v} * \frac{\partial v}{\partial u} * \frac{\partial u}{\partial x}
\end{aligned}
$$

This might look complex , but it simply implies that when you encounter nested functions, simply start at the outermost function , ignoring what it is a function of and use the standard formula for derivative and then go one step inside and repeat until you reach to x .Lets understand this with couple of examples

consider $F(x) = ln(sin(x^{e^x}))$

Now the outermost function is $ln(\cdots)$, we know its derivative is $\frac{1}{\cdots}$, inside the next function is $sin(\cdots)$, and we know its derivative is $cos(\cdots)$, then we have $x^{\cdots}$ and its derivative is $(\cdots)x^{\cdots-1}$, finally we get to e^x which has e^x as its derivative. So derivative of $F(x)$ can be written as :

$$F'(x) = \frac{1}{sin(x^{e^x})} * cos(x^{e^x}) * \left(e^x x^{e^x-1}\right) * e^x$$

I want to emphasize here that we will never get such complex derivatives in ML, but the intuition that we develop here will surely come in handy. Lets look at another example which is much close to what we might encounter.

Consider $F(x) = (2x^2 + 3)^2$

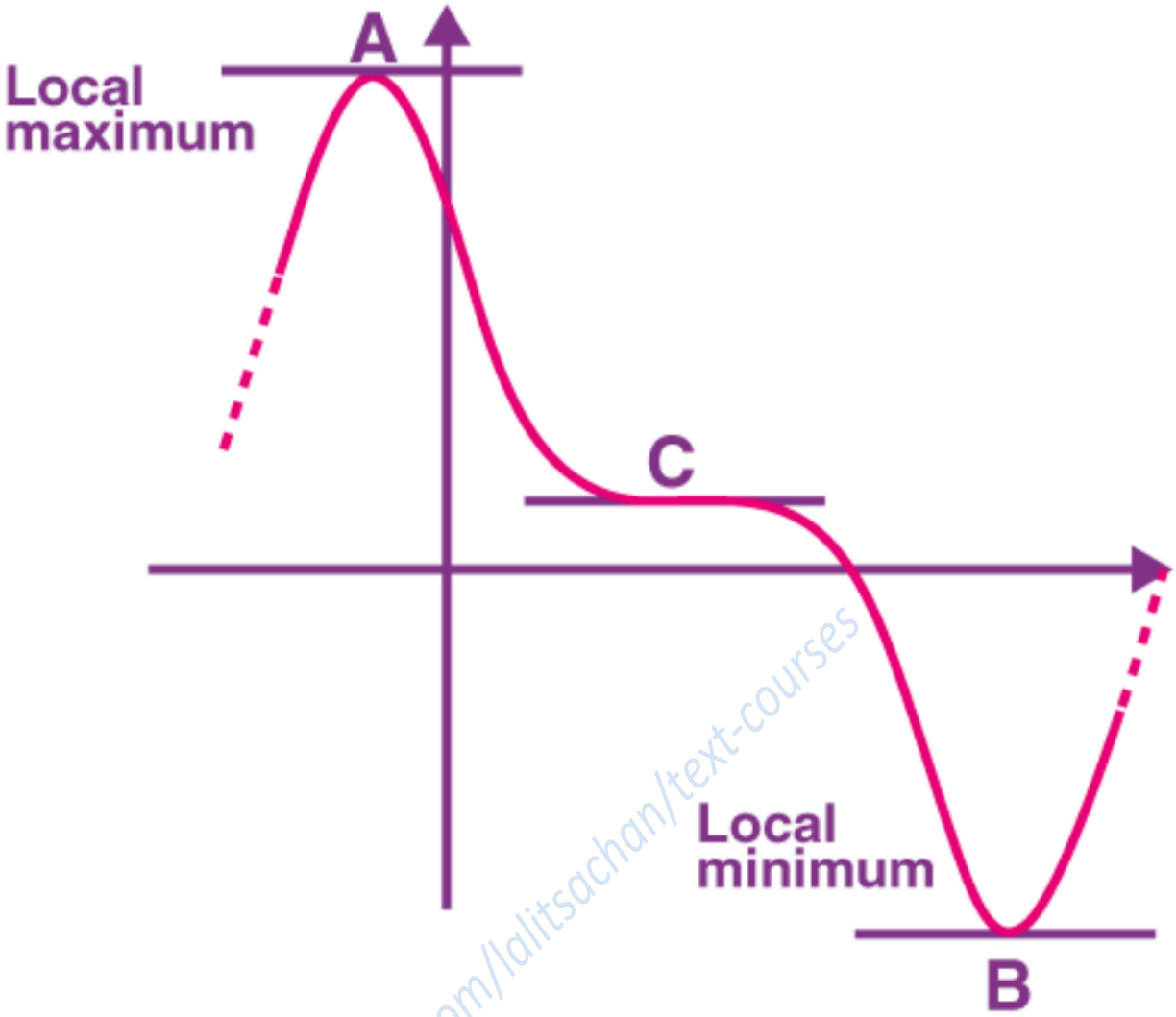
the outermost function here is $(\cdots)^2$, we know its derivative will be $2 * (\cdots)$, and then we go inside find $2x^2 + 3$, which has derivative $4x$. So we can find derivative of $F(x)$ as :

$$F'(x) = 2(2x^2 + 3) * 4x$$

Derivatives and Maxima/Minima of Functions

Now we come to how derivatives relate to optimal points [maxima/minima] of a function and finally understand what all the hoopla was about the derivatives.

https://github.com/lalitsachan/text-courses



https://github.com/lalitsachan/text-courses

The points marked as local minima/maxima are of interest to us when we are looking for optimal points of a function. Majority of the discussion in machine learning comes down to finding parameter values for which a cost function is minimum.

Note that at these points of the function, rate of change is 0 or derivative of the function is 0. For example for a local minima; function is decreasing hence the derivative is negative , function reaches to its minima and then starts increasing with a positive rate of change or positive derivative, so right at the minima; derivative is zero. Same stands true for local maxima as well .
[we are going to leave out the discussion on saddle points for now]

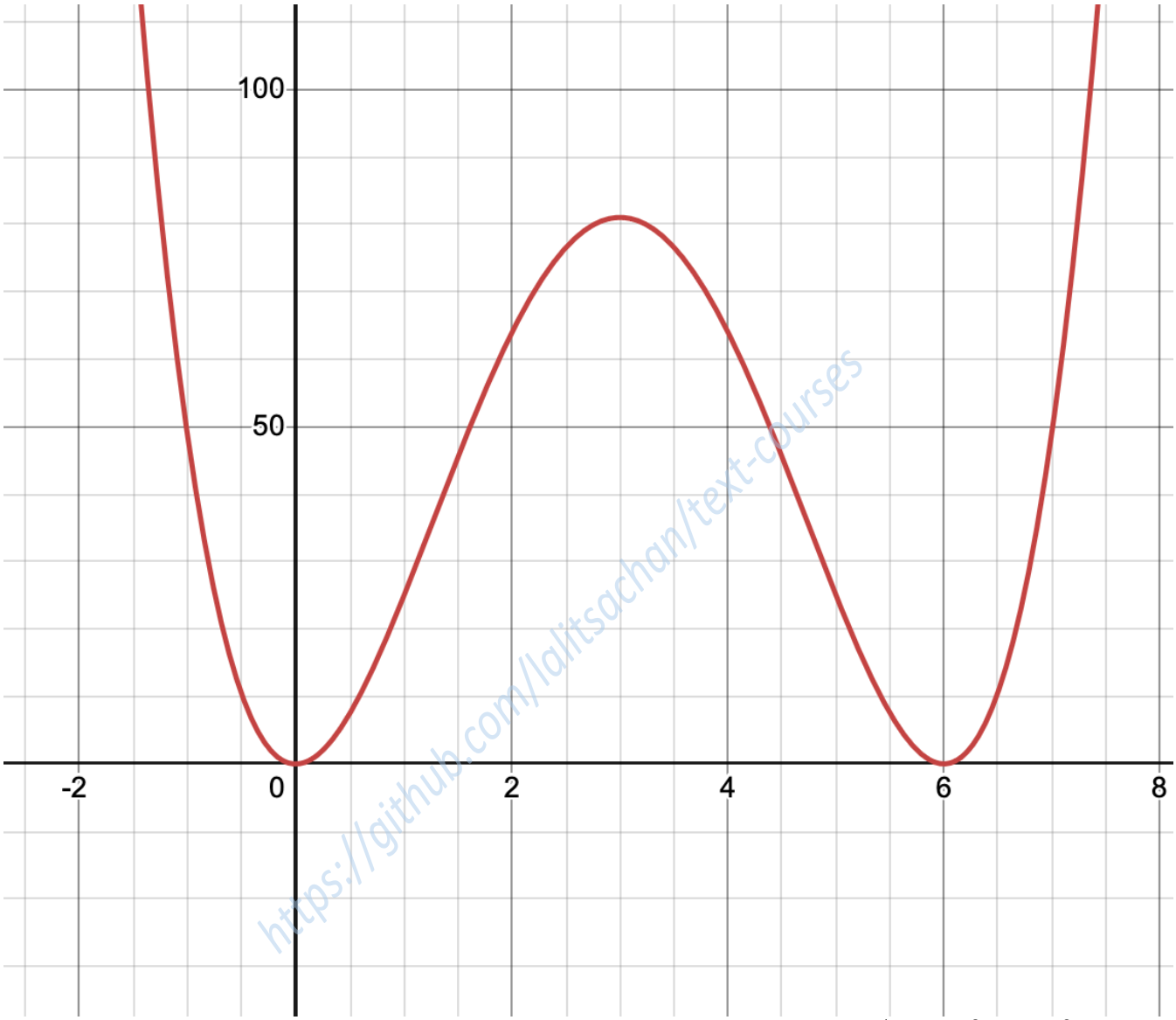
`Note that there is no general method available to find a global maxima or minima for a generic function and thats why our discussions are going to be around local minima and maxima`

Since derivatives are 0 at both minima and maxima, we need some more information to differentiate between minima and maxima.

lets say that the function achieves its minima at $x = x_1$. if we consider a small interval around $x = x_1$, say $(x_1 - \epsilon, x_1 + \epsilon)$, we know that f' , the rate of chagne of f is negative in the interval $(x_1 - \epsilon, x_1)$, because f is decreasing. Then at $x = x_1$, f' is 0 . After its minima, function starts to increase; implying that in the interval $(x_1, x_1 + \epsilon)$, since f is increasing , f' will be positive.

Now if you consider the entire interval $(x_1 - \epsilon, x_1 + \epsilon)$; f' goes from being negative to zero and then positive. In the other words we can say that f' is monotonically increasing in this interval. Which leads to the conclusion that rate of change of f' , that is f'' will be positive at $x = x_1$. we can summarise this as follows

- At minima x_i of f, $f'(x_i) = 0$ and $f''(x_i) > 0$
- With similar logic we can say that at maxima x_j of f, $f'(x_j) = 0$ and $f''(x_j) < 0$



https://github.com/lalitsachan/text-courses

Lets work on an example to understand this better . Consider the function $f(x) = x^4 - 12x^3 + 36x^2$, as you can see in its plot, it has some maximas and minimas. Lets try to find out those optimal points with what we have learnt about the derivations.

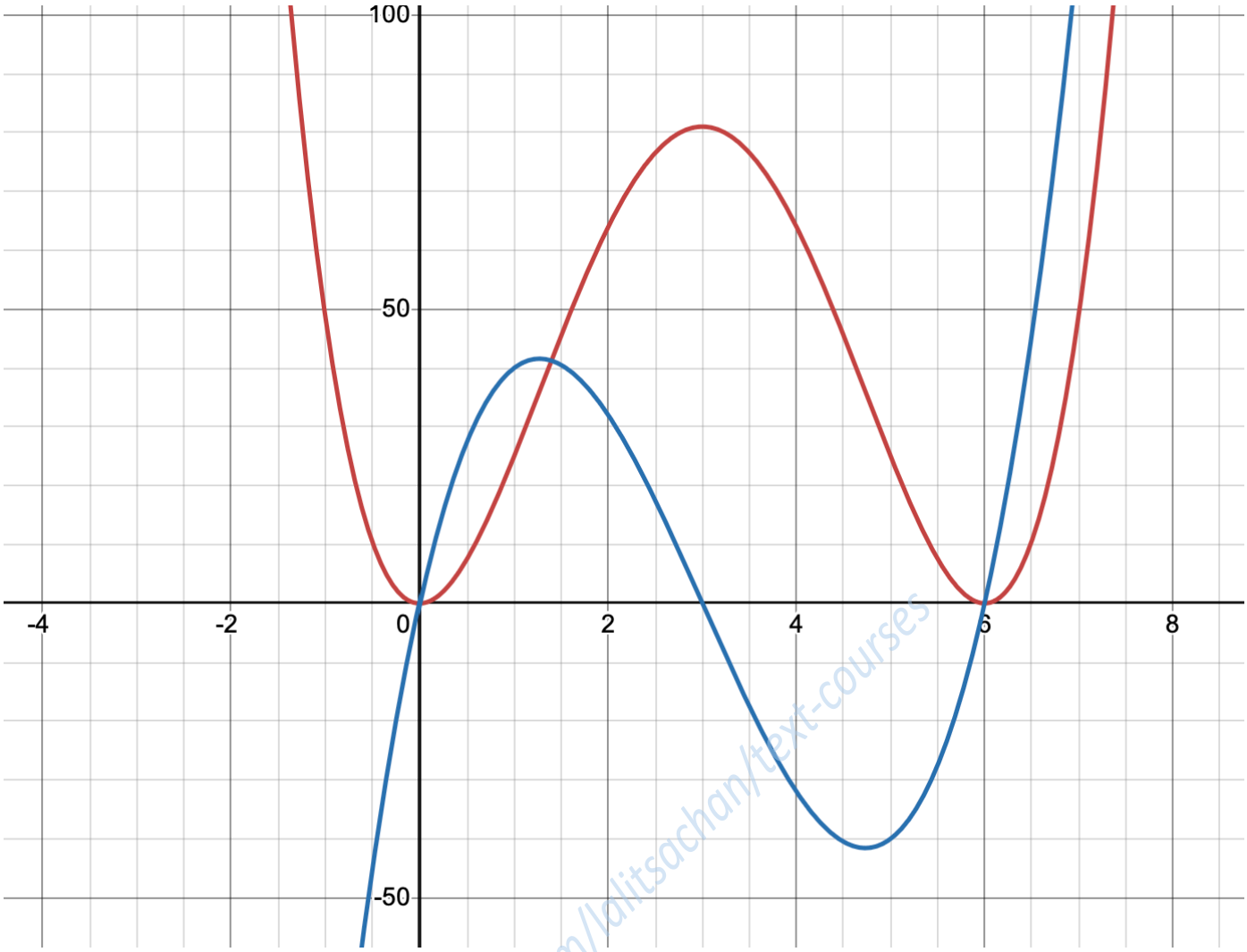
For optimal points :

$$f'(x) = 4x^3 - 36x^2 + 72x = 0$$

solution of this equation will give us the optimal points

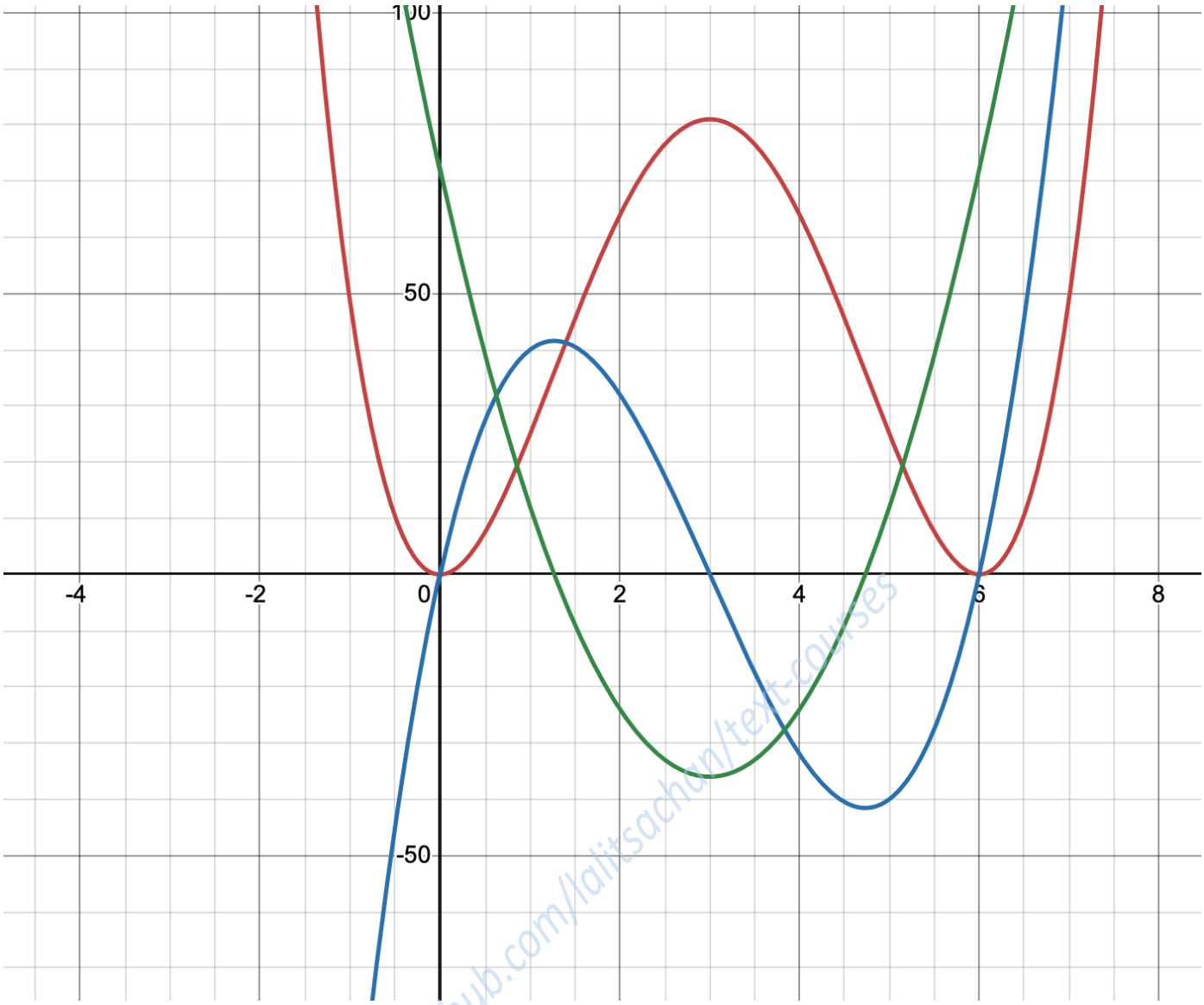
$$\begin{aligned} 4x^3 - 36x^2 + 72x &= 0 \\ x(x^2 - 9x + 18) &= 0 \\ x(x-3)(x-6) &= 0 \end{aligned}$$

we get $x = 0, x = 3, x = 6$ as optimal points. Now to determine which one of these is minima, which is maxima, we will need to look at the second derivative.



blue curve here represents f', see how it crosses x-axis or is zero at optimal points. look at this with f'' and you will see how at the optimal points f'' becomes positive or negative. [postive at minima, negative at maxima]

https://github.com/lalitsachan/text-courses



https://github.com/lalitsachan/text-courses

$$
\begin{aligned}
f''(x) &= 12x^2 - 72x + 72 \\
f''(x)\Big|_{x=0} &= 72 && > 0 \ \textbf{minima} \\
f''(x)\Big|_{x=3} &= -36 && < 0 \ \textbf{maxima} \\
f''(x)\Big|_{x=6} &= 72 && > 0 \ \textbf{minima}
\end{aligned}
$$

so now we know that function has two minimas at $x = 0$ and $x = 6$ and has one maxima at $x = 3$.

Taylor's Expansion

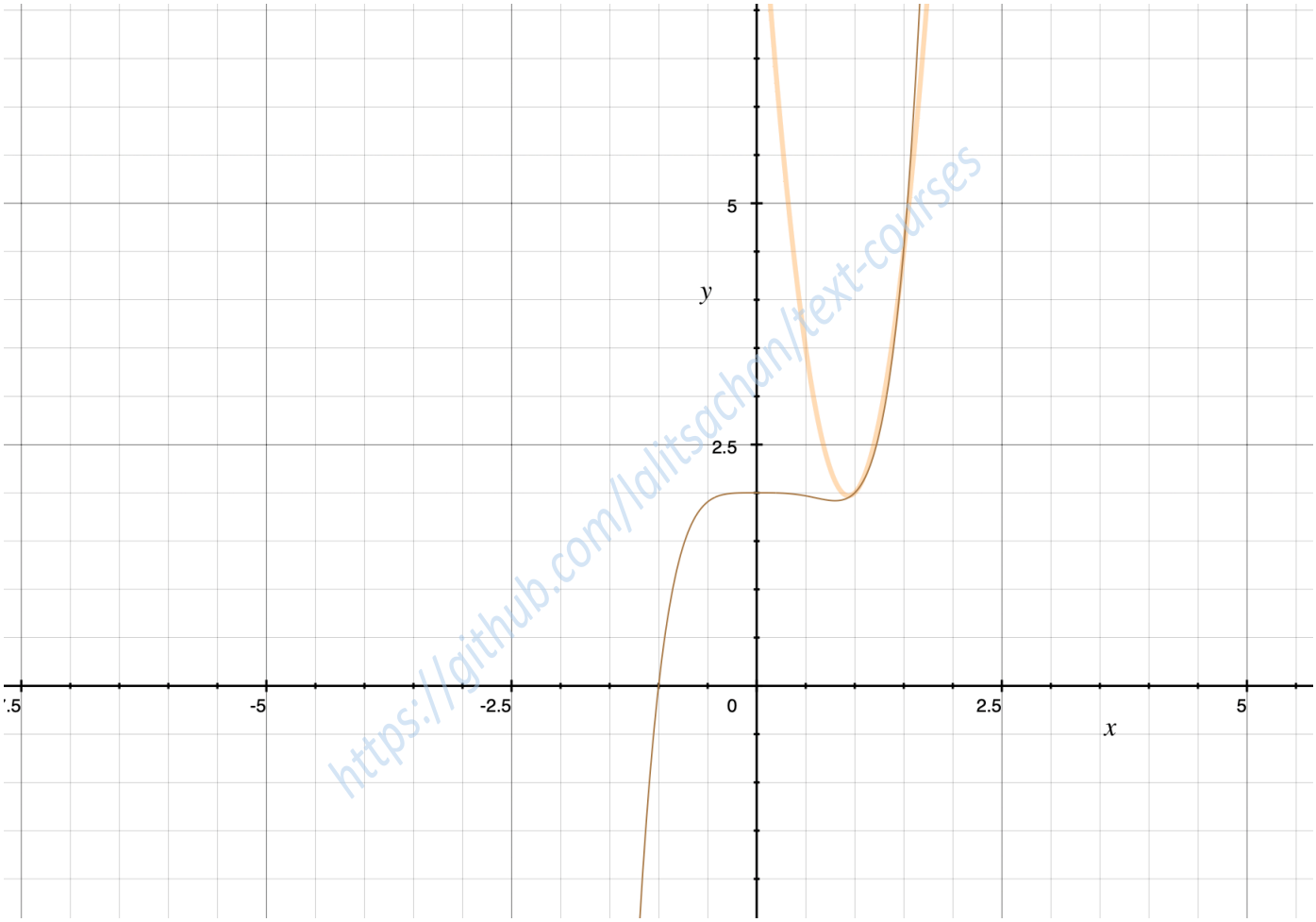
Another use case that you will come across in ML using derivatives is approximating functions with some amenable alternatives using taylor's expansion formula .

Taylor's theorem — Let $k \geq 1$ be an integer and let the function f be k times differentiable at the point $x = a$. Then there exists a function h_k such that:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k,$$
$$\text{and}$$
$$\lim_{x \to a} h_k(x) = 0.$$

usually the way its used is that we assume as $x \to a$, higher powers of $(x-a)$ will be too close to zero and can be safely ignored . for example lets say we have a high degree polynomial $f(x) = x^5 - x^4 + 2$, then we can approximate this around $x = 1$ as this

$$\begin{aligned} f(x) &\approx f(1) + f'(1)(x-1) + \frac{f''(1)}{2}(x-1)^2 \\ &= 2 + (x-1) + 8(x-1)^2 \end{aligned}$$



slim brown curve is the original function and light orange is the second degree polynomial shown above.

you can see how close this approximation is to the original function around $x = 1$, you can make the approximation better by including further higher order terms. Utility of these approximation lies in the idea that this gives much simplified functions in the interval of interest which are easier to math around than the original function.

This is the standard format , another variation of the same which is used when we need to approximate $f(x+a)$ as $a \to 0$

consider the original form

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k$$

replace x with $x+a$

$$f(x+a) = f(a) + f'(a)x + \frac{f''(a)}{2!}x^2 + \cdots + \frac{f^{(k)}(a)}{k!}x^k + h_k(x+a)x^k$$

now interchange x and a and treat this as function of a and use the approximation as $a \to 0$

$$f(x+a) = f(x) + f'(x)a + \frac{f''(x)}{2!}a^2 + \cdots + \frac{f^{(k)}(x)}{k!}a^k + h_k(x+a)a^k$$

we will come across usage of above expression a couple tims in our discussions [remember that in the approximation we will use only few initial terms in the expression and ignore the higher order terms]

Going beyond single variables

What we just went through is called single variable calculus or scalar calculus. These ideas are going to be helpful in essence but when it comes to working with data and models in ML, we need a version where we can use the same ideas but in context of how our data and data operations are represented, that is as vectors and matrices .

I have found using the analogies from the world of physics makes these discussions much more intuitive, so i will use them from time to time, but eventually the usage of these ideas is going to be in context of ML cost function optimization and other such things in the course.

https://github.com/lalitsachan/text-courses

$\frac{\partial \vec{a}}{\partial b}$: derivative of a vector w.r.t. a scalar

An example of this kind of derivative can be a vector quantity such as velocity in 3-D plane, where we take its derivative with respect to a scalar such as time. This will get us acceleration of the object in all three dimensions separately through the space . In general for a vector $\vec{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$, the derivative with respect to a scalar b is defined as :

$$\frac{\partial \vec{a}}{\partial b} = \begin{bmatrix} \frac{\partial a_1}{\partial b} & \frac{\partial a_2}{\partial b} & \cdots & \frac{\partial a_n}{\partial b} \end{bmatrix}$$

an example for this kind would be :

$$\vec{k} = \begin{bmatrix} m^2 & 2m & 12 \end{bmatrix}$$
$$\frac{\partial \vec{k}}{\partial m} = \begin{bmatrix} 2m & 2 & 0 \end{bmatrix}$$

$\frac{\partial k}{\partial \vec{v}}$: derivative of a scalar w.r.t. a vector

An example of this kind can be when we take the derivative of work done with respect to displacement, we get Force, which is a vector .

in general if we have a scalar which is a function of multiple components like $k(v_1, v_2, \cdots, v_n)$, with respect to the vector $\vec{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$, the derivative $\frac{\partial k}{\partial \vec{v}}$ is defined as :

$$\frac{\partial k}{\partial \vec{v}} = \begin{bmatrix} \frac{\partial k}{\partial v_1} & \frac{\partial k}{\partial v_2} & \cdots & \frac{\partial k}{\partial v_n} \end{bmatrix}$$

an example of the kind :

$$
\begin{aligned}
k(m_1, m_2, m_3) &= m_1^2 + 2m_2m_3 \\
\vec{m} &= [\, m_1 \quad m_2 \quad m_3] \\
\frac{\partial k}{\partial \vec{m}} &= [\, 2m_1 \quad 2m_3 \quad 2m_2]
\end{aligned}
$$

this is also called gradient of the scalar with respect to the given vector and written as ∇k

$\frac{\partial \vec{a}}{\partial \vec{b}}$: derivative of a vector w.r.t. a vector

here a is vector valued function and b is vector of the components which make a.

consider that

$$
\begin{aligned}
b &= [\, b_1 \quad b_2 \quad \cdots \quad b_n] \\
\text{and } a &= [\, g_1(b) \quad g_2(b) \quad \cdots \quad g_n(b)]
\end{aligned}
$$

then derivative of a with respect to b is known as Jacobian, a matrix defined as

$$
J = \frac{\partial a}{\partial b} = \begin{bmatrix} \frac{\partial g_1}{\partial b_1} & \frac{\partial g_1}{\partial b_2} & \cdots & \frac{\partial g_1}{\partial b_n} \\ \frac{\partial g_2}{\partial b_1} & \frac{\partial g_2}{\partial b_2} & \cdots & \frac{\partial g_2}{\partial b_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial b_1} & \frac{\partial g_n}{\partial b_2} & \cdots & \frac{\partial g_n}{\partial b_n} \end{bmatrix}
$$

https://github.com/lalitsachan/text-courses

An example of this can be force applied on a moving charged particle in a magnetic field. Both the force and the velocity of the particle are vectors. The result jacobian, is just a collection of stacked gradients , where each row of jacobian gives you gradient of the scalar valued component of the vector valued function. For example lets say the force vector is $\vec{F} = [\, f_x \quad f_y \quad f_z]$ and velocity vector is $\vec{V} = [\, v_x \quad v_y \quad v_z]$

$$
J = \frac{\partial \mathbf{F}}{\partial \mathbf{V}} = \begin{bmatrix} \frac{\partial f_x}{\partial v_x} & \frac{\partial f_x}{\partial v_y} & \frac{\partial f_x}{\partial v_z} \\ \frac{\partial f_y}{\partial v_x} & \frac{\partial f_y}{\partial v_y} & \frac{\partial f_y}{\partial v_z} \\ \frac{\partial f_z}{\partial v_x} & \frac{\partial f_z}{\partial v_y} & \frac{\partial f_z}{\partial v_z} \end{bmatrix}
$$

Here is another way you will find people talking about Jacobian : It is first order derivative of a vector valued function

Hessian

In context of optimisation , we have seen that we need to take second order derivative also in order to determine if the optimal point in question is a minima or maxima.

consider a scaler valued function $f(w_1, w_2)$ and vector $\vec{w} = [w_1 \; w_2]$. the first order derivative of f with respect to $\vec{w}$ is called gradient as we have seen before .

$$
\nabla f = \begin{bmatrix} \frac{\partial f}{w_1} & \frac{\partial f}{w_2} \end{bmatrix}
$$

now if we take derivative of ∇f with respect to $\vec{w}$ again , it is called second order derivative , since now ∇f and w , both are vectors , we run into scenario similar to Jacobian . The process is same but the resulting matrix is called a Hessian .

$$H = \frac{\partial \nabla f}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} \end{bmatrix}$$

Although we arrive at Jacobian and Hessian following similar path but do note that Jacobian is first order derivative of a vector valued function where as Hessian is second order derivative of a scalar valued function, which is pretty evident if you look at individual values of the matrices.

This process of derivative of a scalar resulting in a vector , derivative of vector resulting in a matrix can be further continued . But we will start to venture beyond simple 2D matrices [tensors], and that for now is not a needed direction for our discussion here.

While these ideas will help you in reading/understanding ML papers and implementing them in your work, here are some quick results which will come in handy very often

Useful Results

1

if we have two vectors [with matching dimensions] $\vec{a}$ and $\vec{c}$, then derivative of their dot product $\vec{a} \cdot \vec{c}$, with respect to one of the vectors will result in the other vectors.

$$\frac{\partial \vec{a} \cdot \vec{c}}{\partial \vec{c}} = \vec{a}$$
$$\text{and}$$
$$\frac{\partial \vec{a} \cdot \vec{c}}{\partial \vec{a}} = \vec{c}$$

to derive this consider $\vec{a}$ to be a row vector $\vec{a} = [a_1 \quad a_2 \quad a_3]$ and $\vec{b}$ to be a column vector $\vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

.Recall our linear algebra discussion, Dot product $\vec{a} \cdot \vec{c}$ is a scalar

$$\vec{a} \cdot \vec{c} = a_1 c_1 + a_2 c_2 + a_3 c_3$$

when you take the derivative of this scalar with respect to the vector say $\vec{a} = [a_1 \quad a_2 \quad a_3]$, you will get a vector

$$\frac{\partial \vec{a} \cdot \vec{c}}{\partial \vec{a}} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \vec{c}$$

2

Consider vector x or x^T , its derivative with respect to itself x will be a Jacobian ,in fact an Identity matrix eventually. Consider $\vec{x} = [x_1 \quad x_2 \quad x_3]$

$$\frac{\partial \vec{x}}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \frac{\partial x_2}{\partial x_1} & \frac{\partial x_3}{\partial x_1} \\ \frac{\partial x_1}{\partial x_2} & \frac{\partial x_2}{\partial x_2} & \frac{\partial x_3}{\partial x_2} \\ \frac{\partial x_1}{\partial x_3} & \frac{\partial x_2}{\partial x_3} & \frac{\partial x_3}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

3

https://github.com/lalitsachan/text-courses

Consider a column vector b with dimension $n \times 1$. A square matrix A with dimensions $n \times n$.

given these dimensions , we know that the quantity $b^T Ab$ is defined , the dimensions align, more importantly; its a scalar . Now if we take the derivative of this scalar quantity with respect to the vector b, lets see what we get .

$$\frac{\partial b^T Ab}{\partial b} = b^T \frac{\partial Ab}{\partial b} + \frac{\partial b^t}{\partial b} Ab$$
$$\frac{\partial b^T Ab}{\partial b} = b^T A + IAb$$
$$\frac{\partial b^T Ab}{\partial b} = b^T A + Ab$$

We will need some adjustment here since given that b is a column vector with dimensions $n \times 1$, that makes $b^T A$ a row vector of dimension $1 \times n$ and Ab a column vector of dimension $n \times 1$, we can write $b^t A$ as its transpose bA^T to make them both column vectors so that the sum makes sense . continuing from where we left :

$$\frac{\partial b^T Ab}{\partial b} = b^T A + Ab$$
$$\frac{\partial b^T Ab}{\partial b} = bA^T + Ab$$
$$\frac{\partial b^T Ab}{\partial b} = (A + A^T)b$$

further if A is a symmetric matrix that is $A^T = A$, then this further simplifies :

$$\frac{\partial b^T Ab}{\partial b} = (A + A^T)b$$
$$\frac{\partial b^T Ab}{\partial b} = 2Ab = 2A^T b$$

https://github.com/lalitsachan/text-courses

An Example to bring everything together

Consider these matrices and vectors :

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}^{n \times 1}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}^{n \times (p+1)}, W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}^{(p+1) \times 1}$$

Consider that X and Y are observed data points and W contains unknown parameters. Now consider a scalar valued function $F = (Y - XW)^T (Y - XW)$. Can you find out for what values of W [the elements in it], F will take its minimum value.

From our derivative discussion we know that if F is minimum at some particular W , we will have $\frac{\partial F}{\partial W} = 0$ for that. So we can solve that equation to find out that particular W . Before we go through that process , lets open up F and simplify .

$$F = (Y - XW)^T (Y - XW)$$
$$F = (Y^T - W^T X^T)(Y - XW)$$
$$F = Y^T Y - W^T X^T Y - Y^T XW + W^T X^T XW$$

if you go about evaluating the dimensions of the terms in the expression above, you will realize that all of them are scalars. If you recall scalars are unaffected by transposing. So we can replace $W^T X^T Y$ with $(W^T X^T Y)^T = Y^T XW$ in the expression without an issue. Continuing with the process above.

$$
\begin{aligned}
F &= Y^T Y - W^T X^T Y - Y^T XW + W^T X^T XW \\
F &= Y^T Y - Y^T XW - Y^T XW + W^T X^T XW \\
F &= Y^T Y - 2Y^T XW + W^T X^T XW
\end{aligned}
$$

Now lets calculate the derivative :

$$
\begin{aligned}
\frac{\partial F}{\partial W} &= \frac{\partial Y^T Y}{\partial W} - 2\frac{\partial Y^T XW}{\partial W} + \frac{\partial W^T X^T XW}{\partial W} \\
\frac{\partial F}{\partial W} &= 0 - 2Y^T X + 2X^T XW
\end{aligned}
$$

we used the result 3 from earlier , considering $X^T X$ being a symmetric matrix. to find the value of W for which F is minimum we can equate $\frac{\partial F}{\partial W}$ to 0 and solve for W.

$$
\begin{aligned}
\frac{\partial F}{\partial W} &= -2Y^T X + 2X^T XW = 0 \\
X^T XW &= Y^T X \\
W &= (X^T X)^{-1} Y^T X
\end{aligned}
$$

In order to ensure that its a minima , we need a little more mathematical details which i have left out to keep things simple. However i will mention them below so that you have a thread to pull on if you want more details .

F is a quadratic function. Quadratic functions are either convex or concave depending on the curvature, and from there it follows that the optimal point is either global minima or global maxima. $\frac{\partial^2 F}{\partial^2 W} = X^T X$. Since $X^T X$ is positive semi-definite matrix , it implies that F is convex and hence value obtained above for W represents a minima.

Lets see how it all come to pass in our next chapter where we start with linear models

https://github.com/lalitsachan/text-courses