Introduction to Gradient Descent

Abu Talha

Department of Data Science and Engineering, IISER Bhopal

What is Gradient?

This is probably the closest thing to a regular **derivative**[1] that you will notice. A gradient essentially tells how much a surface or some quantity changes from one point in space / time to another. The Gradient (also called slope) of a line shows how steep it is.

Physical significance of Gradient :

The gradient[2] is **a measurement of how much something shifts from one point to another point in a given feature space.**

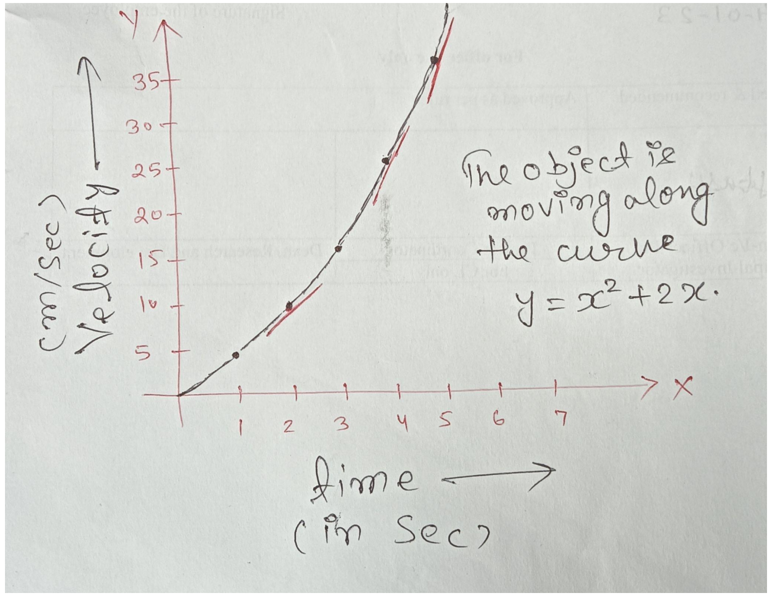
In other words, Gradient of a function at any point tells about the direction of change of that function.

Source: (1) Calculus by Gilbert Strang & Edwin Herman, MIT & University of Wisconsin-Stevens Point.
(2) https://qsstudy.com/physical-significance-gradient

What is derivative?

In mathematics, the derivative of a function of a real variable **measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value)**. Derivative is a fundamental tool of calculus.

For example, the ***derivative of the position of a moving object with respect to time is the object's velocity***: this measures how quickly the position of the object changes when time advances.



Source: https://en.wikipedia.org/wiki/Derivative

What is derivative?

The derivative of a function of a *single variable* at a chosen input value, when it exists, is the *slope* of the *tangent* line to the graph of the function at that point.

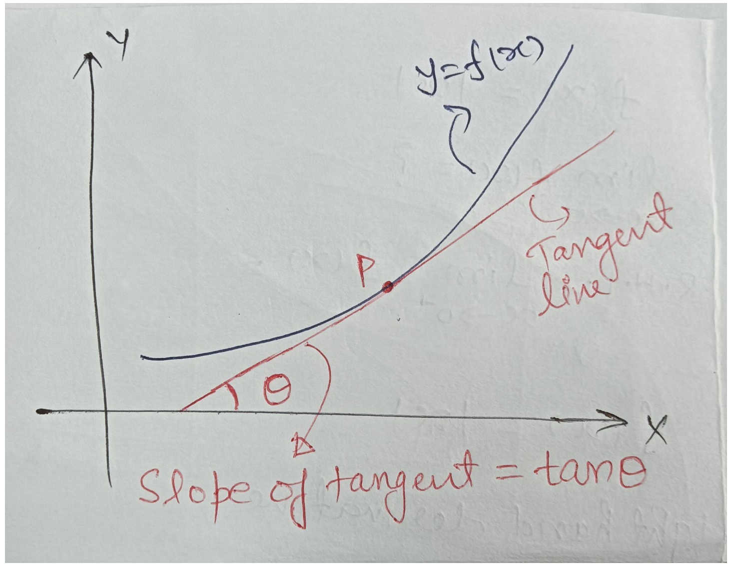
What is tangent ?

The tangent line to a curve at a given point is the straight line that "just touches" the curve at that point.
In other words, the **tangent represents the instantaneous rate of change** of the given function at a point.

What is the slope of a tangent?

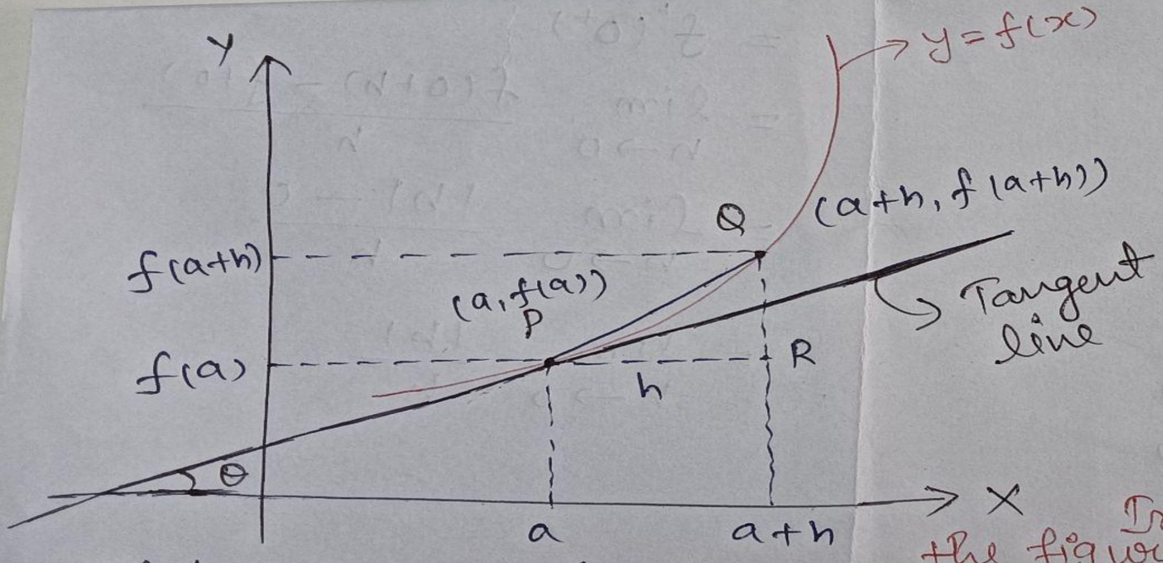
If a tangent line makes an angle ϴ with the positive X- axis, then slope of tangent is equal to tan ϴ .

The slope of the tangent at a point is equal to the derivative of the function at the same point .



Source : https://www.cuemath.com/geometry/tangent

How we define derivative at any point ?



Slope of tangent = $\tan\theta$

The derivative of $y=f(x)$ w.r.t. x at point 'a' is given by

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

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

$$= \lim_{h\to 0} \frac{\text{change in } y}{\text{change in } x}$$

In the limiting process, as $h \to 0$, from the figure, the point Q tendes to point P, then

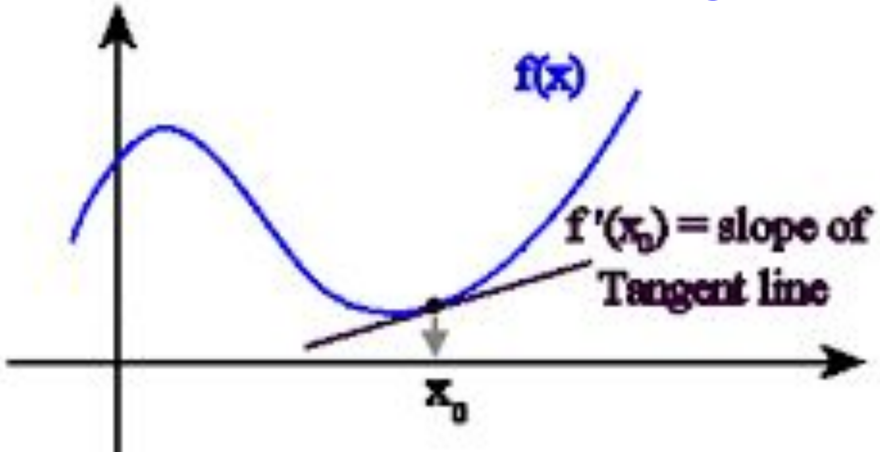
$$f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = \lim_{Q\to P} \frac{QR}{PR} = \text{slope of Tangent}$$

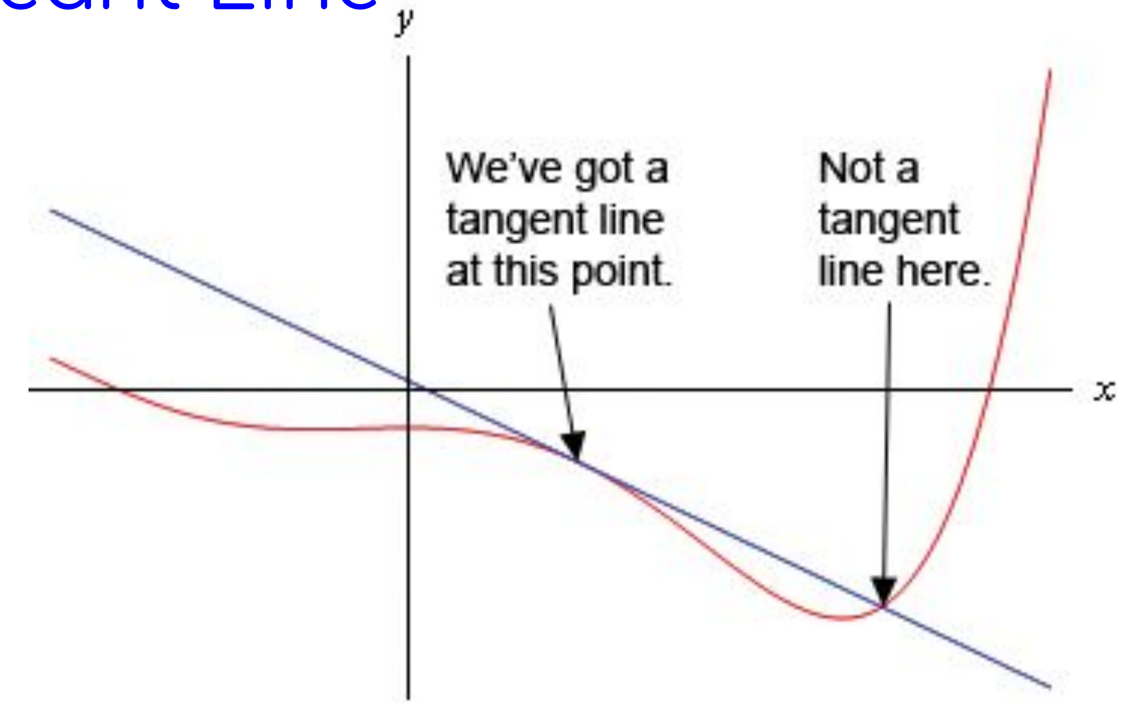
What is the physical significance of derivative?

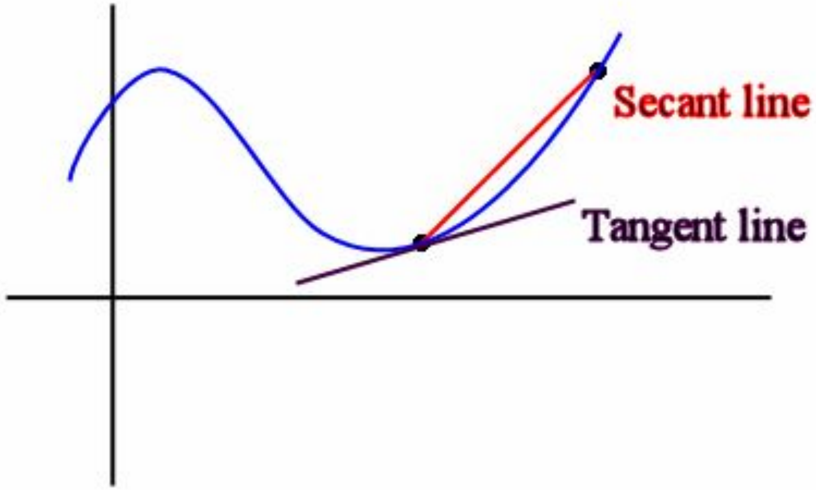
The derivative is defined as an instantaneous rate of change at a given point.

dy/dx = tan ϴ = change in y/ change in x

Tangent vs Secant Line





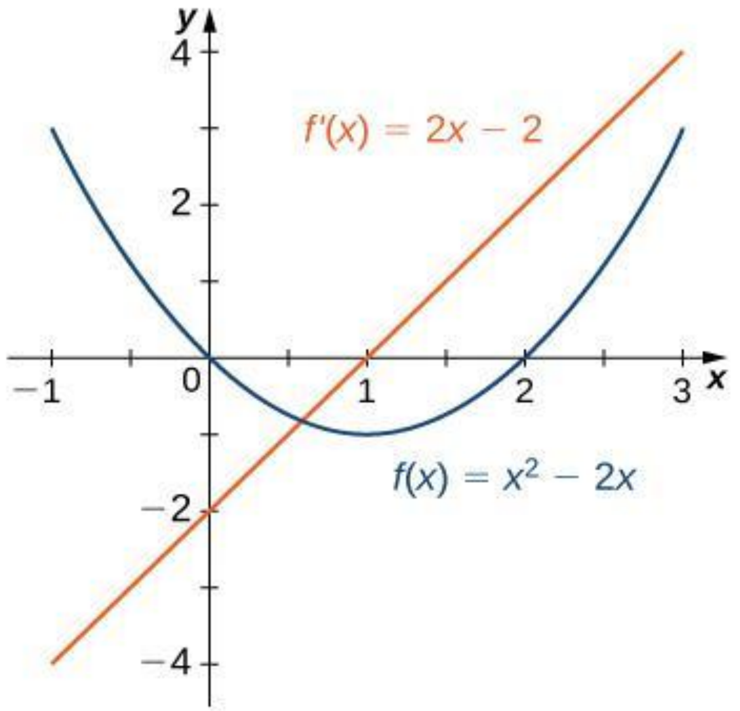


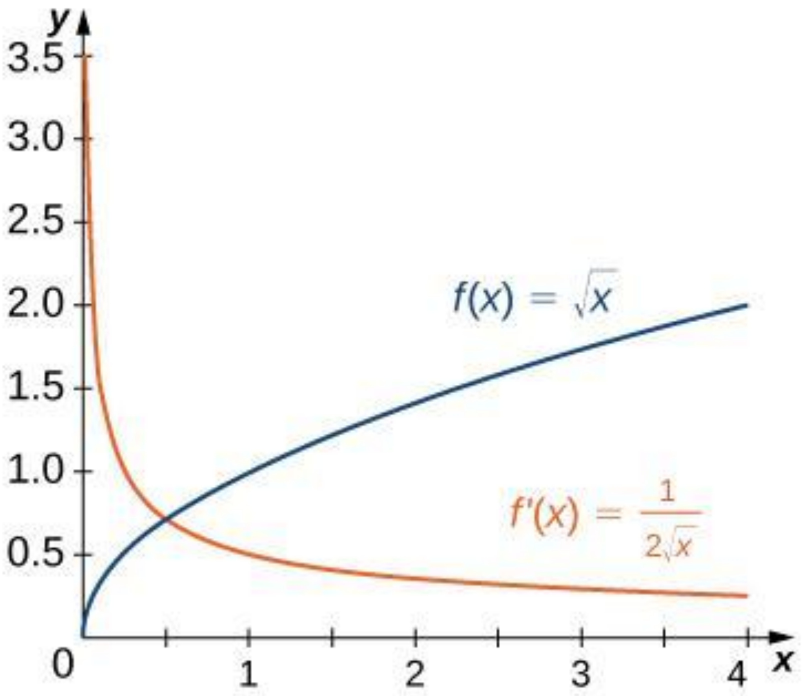
A **secant** is a line that intersects a curve at a minimum of two distinct points.

Source:https://www.intmath.com/differentiation

Illustrations: Derivative of a function:

Let us consider few examples on derivatives graphically.





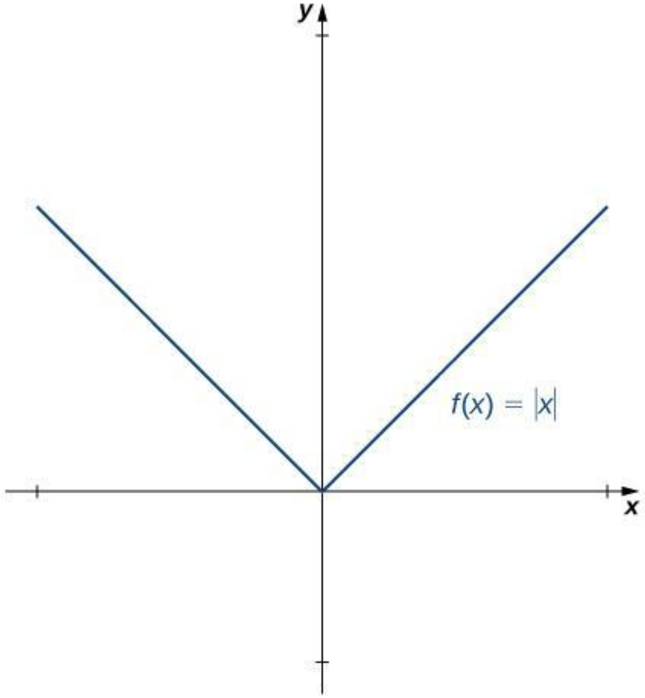
The derivative $f'(x)<0$ where the function f(x) is decreasing and $f'(x)>0$, where f(x) is increasing. The derivative is zero where the function has a horizontal tangent.

The derivative $f'(x)$ is positive everywhere because the function f(x)is increasing

Source:https://math.libretexts.org/Bookshelves/Calculus. Calculus by Gilbert Strang & Edwin Herman, MIT & University of Wisconsin-Stevens Point.

Case Study : When Derivative does not exist?

The function $f(x)=|x|$ is continuous at 0,but is not differentiable at 0.



$f(x) = |x|$

$$\boxed{|x| = \begin{cases} x \,;\, x \geq 0 \\ -x \,;\, x < 0 \end{cases}}$$

$f'(0) = ?$

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

$$= \lim_{h \to 0} \frac{|h| - 0}{h}$$

$$f'(0) = \lim_{h \to 0} \frac{|h|}{h} = \begin{cases} 1 \,;\, h \geq 0 \\ -1 \,;\, h < 0 \end{cases}$$

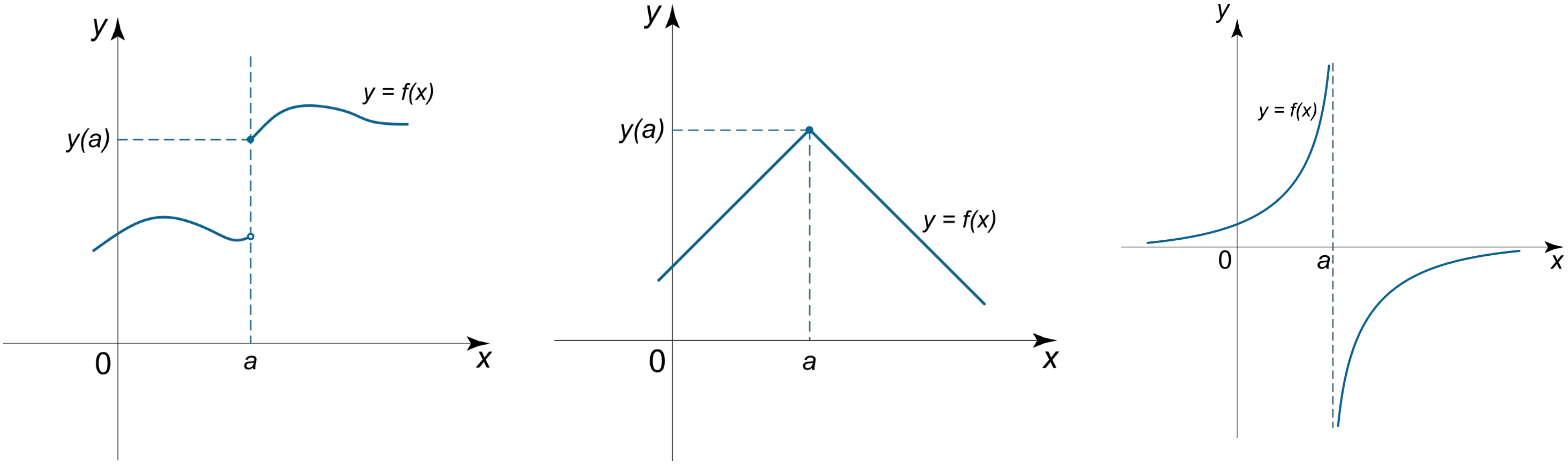
$\Rightarrow$ limit is not unique.

$\Rightarrow$ $f'(0)$ does not exist.

What is the limit of a function?

The limit of a function is a value of the function as the input of the function gets closer or approaches some number. The limit of a function is always concerned with the behavior of the function at a particular point.

Functions whose derivative does not exist



In the above graph the function is **not** continuous at point "a" , therefore the derivative of the function at point "a" does **not** exist.

Remark : The derivative of a function does not exist always.

Source: https://www.vedantu.com/question-answer

What will happen when we have a function of several variables (i.e. two or more than two variables) ?

In this case, we need **partial derivatives** of the function.

What is Partial Derivative?

A **Partial Derivative** of a function of several variables is, its derivative with respect to one of those variables, with the others held constant.

Example :

The body mass index (BMI) is a measure that uses your height and weight to work out your health status.

Illustration: Partial Derivative

BMI- is a person's weight in kilograms (or pounds) divided by the square of height in meters (or feet)

$$BMI = \text{weight (lb)} / [\text{height (in)}]^2 \times 703$$

Illustration from the table : In particular, If we keep the height 67 inches fixed, then we can observe the BMI with respective to corresponding weights.

Similarly, If we keep the weight 143 pounds fixed, we can see the change in BMI with the different heights value viz 58,59,60,61,71 inches.

Source:

2) https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/
3) https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/

Illustration : Partial Derivative (Continued)

Body Mass Index Table

	Normal						Overweight					Obese										Extreme Obesity														
BMI	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**
Height (inches)																**Body Weight (pounds)**																				
58	91	96	100	105	110	115	119	124	129	134	138	143	148	153	158	162	167	172	177	181	186	191	196	201	205	210	215	220	224	229	234	239	244	248	253	258
59	94	99	104	109	114	119	124	128	133	138	143	148	153	158	163	168	173	178	183	188	193	198	203	208	212	217	222	227	232	237	242	247	252	257	262	267
60	97	102	107	112	118	123	128	133	138	143	148	153	158	163	168	174	179	184	189	194	199	204	209	215	220	225	230	235	240	245	250	255	261	266	271	276
61	100	106	111	116	122	127	132	137	143	148	153	158	164	169	174	180	185	190	195	201	206	211	217	222	227	232	238	243	248	254	259	264	269	275	280	285
62	104	109	115	120	126	131	136	142	147	153	158	164	169	175	180	186	191	196	202	207	213	218	224	229	235	240	246	251	256	262	267	273	278	284	289	295
63	107	113	118	124	130	135	141	146	152	158	163	169	175	180	186	191	197	203	208	214	220	225	231	237	242	248	254	259	265	270	278	282	287	293	299	304
64	110	116	122	128	134	140	145	151	157	163	169	174	180	186	192	197	204	209	215	221	227	232	238	244	250	256	262	267	273	279	285	291	296	302	308	314
65	114	120	126	132	138	144	150	156	162	168	174	180	186	192	198	204	210	216	222	228	234	240	246	252	258	264	270	276	282	288	294	300	306	312	318	324
66	118	124	130	136	142	148	155	161	167	173	179	186	192	198	204	210	216	223	229	235	241	247	253	260	266	272	278	284	291	297	303	309	315	322	328	334
67	121	127	134	140	146	153	159	166	172	178	185	191	198	204	211	217	223	230	236	242	249	255	261	268	274	280	287	293	299	306	312	319	325	331	338	344
68	125	131	138	144	151	158	164	171	177	184	190	197	203	210	216	223	230	236	243	249	256	262	269	276	282	289	295	302	308	315	322	328	335	341	348	354
69	128	135	142	149	155	162	169	176	182	189	196	203	209	216	223	230	236	243	250	257	263	270	277	284	291	297	304	311	318	324	331	338	345	351	358	365
70	132	139	146	153	160	167	174	181	188	195	202	209	216	222	229	236	243	250	257	264	271	278	285	292	299	306	313	320	327	334	341	348	355	362	369	376
71	136	143	150	157	165	172	179	186	193	200	208	215	222	229	236	243	250	257	265	272	279	286	293	301	308	315	322	329	338	343	351	358	365	372	379	386
72	140	147	154	162	169	177	184	191	199	206	213	221	228	235	242	250	258	265	272	279	287	294	302	309	316	324	331	338	346	353	361	368	375	383	390	397
73	144	151	159	166	174	182	189	197	204	212	219	227	235	242	250	257	265	272	280	288	295	302	310	318	325	333	340	348	355	363	371	378	386	393	401	408
74	148	155	163	171	179	186	194	202	210	218	225	233	241	249	256	264	272	280	287	295	303	311	319	326	334	342	350	358	365	373	381	389	396	404	412	420
75	152	160	168	176	184	192	200	208	216	224	232	240	248	256	264	272	279	287	295	303	311	319	327	335	343	351	359	367	375	383	391	399	407	415	423	431
76	156	164	172	180	189	197	205	213	221	230	238	246	254	263	271	279	287	295	304	312	320	328	336	344	353	361	369	377	385	394	402	410	418	426	435	443

Source: Adapted from *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report.*

Source: https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi_tbl.htm

Partial Derivative (Mathematical Example)

Let us understand the partial derivative with an example. If f is a function of x and y such that

$$f(x,y) = x^2 + xy + y^2$$

The **partial derivatives** of f(x,y) is given by ,

$$\partial f/\partial x = 2x + y$$

$$\partial f/\partial y = x + 2y$$

What is the point to study **derivative** of a function ?

There are many applications of derivative in real life , few are listed below :

- Approximation or finding approximate value
- Rate of change of Quantity
- Maxima and minima of a function
- Determining increasing and decreasing of a function

For our topic of the talk, we will discuss about maxima and minima of a function and approximation.

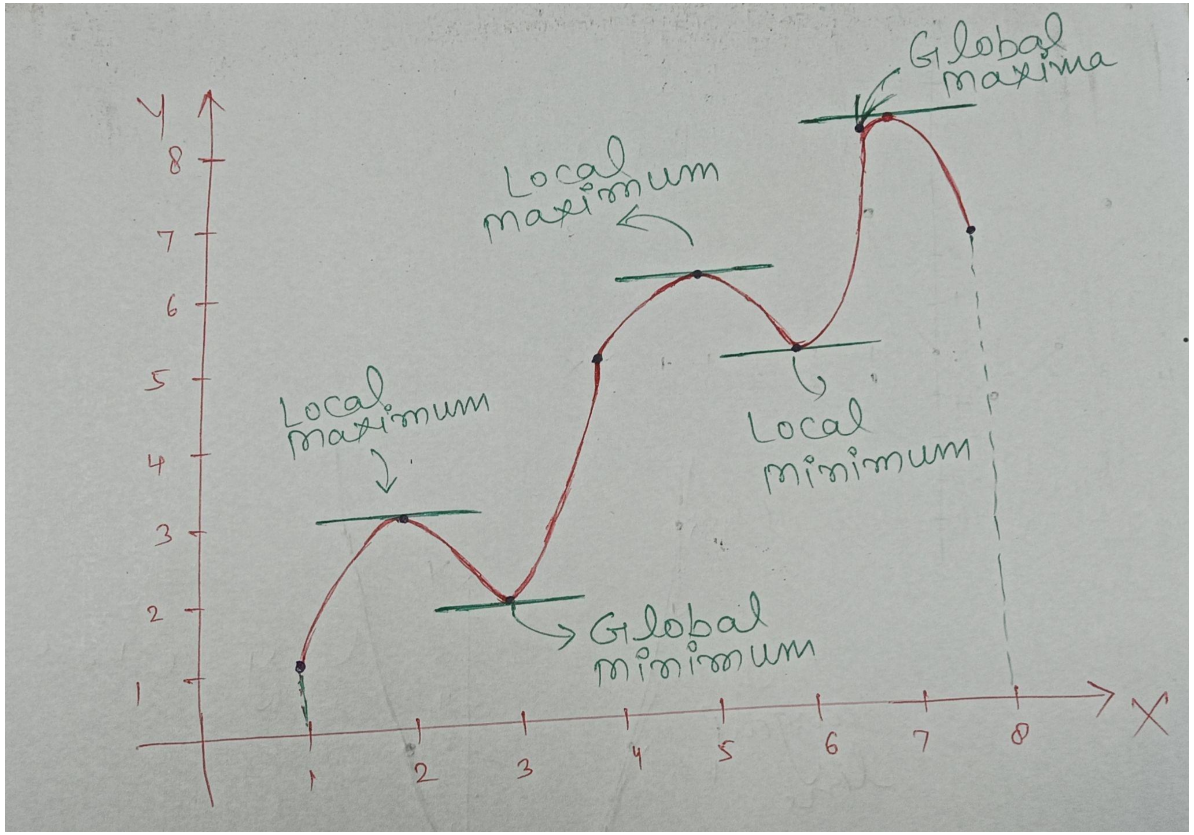
What is the maxima and minima of a function?

Maxima and minima of a function are the largest and smallest value of the function respectively either within a given range or on the entire domain.

Source:https://testbook.com/learn/maths-applications-of-derivatives

Pictorial representation of maxima and minima :

In this figure (on the right side) ,we can see the difference between Global minimum or global maximum and local maximum or local minimum of a function.



What is approximation or approximate value of a function?

If there is a very small change in one variable correspond to the other variable then we use the differentiation to find the **approximate value**.

What is the physical significance of approximation?

If your height is **5’11.5’’**, don’t you call yourself **approximately 6’** tall whenever asked? Similarly, if you scored **97.75%** in your examinations, don’t you boast about it by chanting that you scored **approximately 98%**? Such is the intuitive nature of the topic of approximations that it doesn’t even need an explanation!

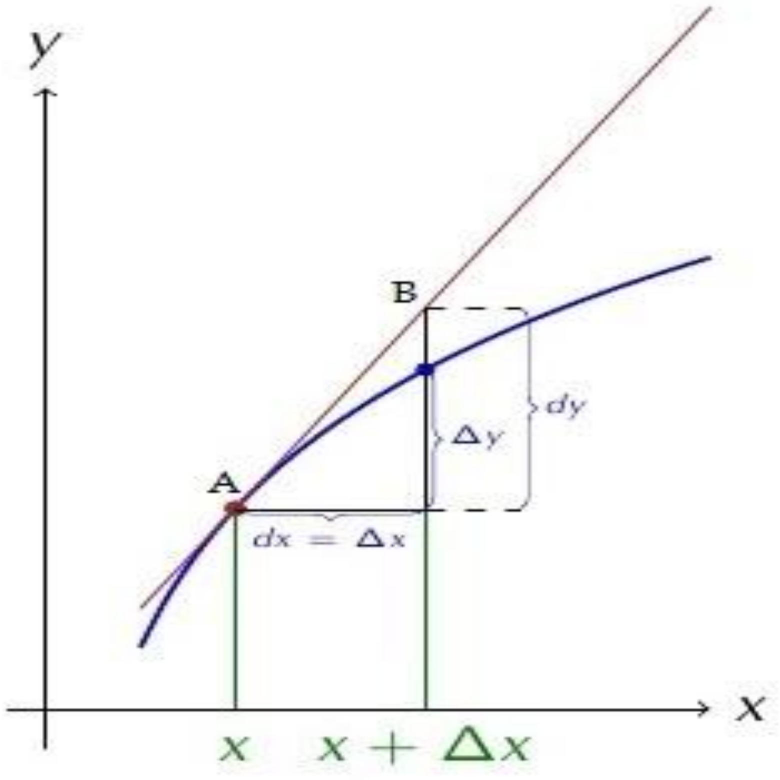
Approximation by Derivatives :

The general form of result obtained is:

$$\mathbf{f(x + \Delta x) = f(x) + f'(x)\,\Delta x}$$

Approximation by Derivatives (Continued) :

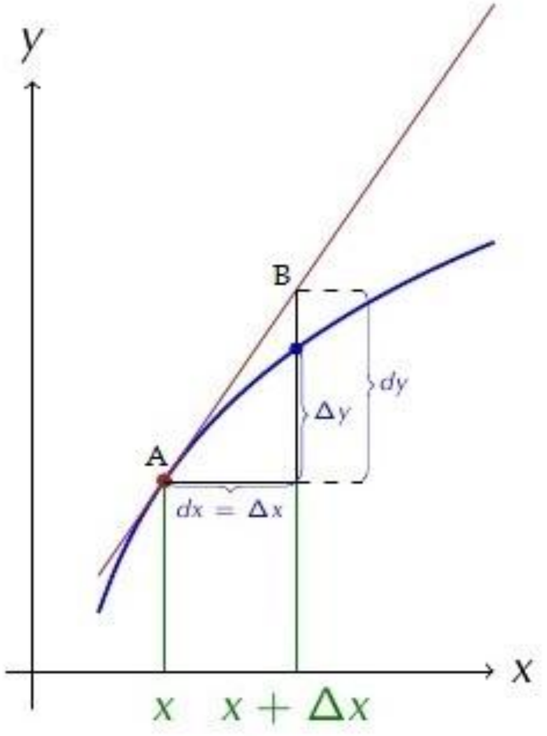
Which enables one to get the value of the function at a point near x. In connection with this formula, look at the figure below:



Source: https://www.toppr.com/guides/maths/application-of-derivatives/approximations/

How do we use approximation by derivatives?

Notation-wise let us define $y(x = x') = y(x = x_0) + \Delta y$; implying that Δy is the change in the value of the function y when the change in x is given by $\Delta x = x' - x_0$. Then we proceed as follows,



1)Find a point x_0 near the point x′, at which the value of the function is known.
2)Differentiate the function with respect to x ,

$$dy/\,dx = d/dx\,(f(x))$$
$$dy = f'(x)\,dx$$

3)Use the approximations i.e. the value of the change in x i.e. $dx = \Delta x = x' - x_0$ and calculate the derivative at $x = x_0$ to get dy, which is approximated as Δy:

$$\Delta y = f'(x_0)\,\Delta x$$
$$\Delta y = f'(x_0)\,(x' - x_0)$$

4)This would be the change in the value of the function y as x changes from x_0 to x′.
Thus, we have ,

$$f(x') = f(x_0) + \Delta y$$
$$f(x') = f(x_0) + f'(x_0)\,(x' - x_0)$$

Source: https://www.toppr.com/guides/maths/application-of-derivatives/approximations/

What is Gradient Descent ?

In mathematics, gradient descent (also often called steepest descent) is a first-order iterative optimization algorithm for finding a local minima of a differentiable function.

Let us take a simple example to understand the intuition behind **Gradient Descent**.

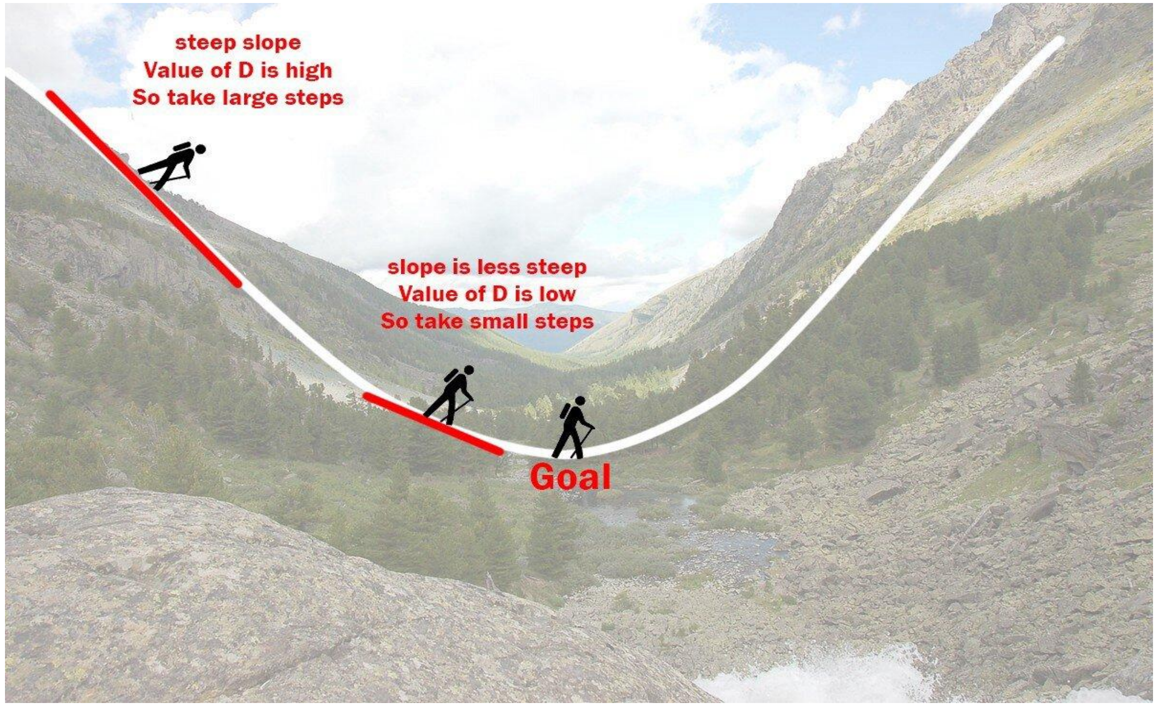
Imagine that you were in the hills, and had to find the lowest valley.

Source: https://towardsdatascience.com/complete-step-by-step-gradient-descent-algorithm-from-scratch-acba013e8420

How Gradient Descent Algorithm Works?

Do this repeatedly:

1) Start from any point on any hill
2) Look in all four directions (ahead, behind, left, right) to determine where you might be able to descend (rather than ascend)
3) Take a step in that direction
4) Return to step (2)



If you have reached a point where taking a step in any direction doesn't make a difference, you are at a minimum (you've reached the valley)

Source: https://towardsdatascience.com/linear-regression-using-gradient-descent-97a6c8700931

Gradient Descent Illustration :

Alert: When you are at a valley, from where you can see some other cavern or valley, you're likely to be at a "local minimum".

- Gradient descent algorithms do much the same things, but in an arbitrary number of dimensions.

- Gradient descent represents the opposite direction of gradient. Gradient of a function at any point represents direction of steepest ascent of the function at that point.
- If we have function of the form $Y = X^2 + 2X$. In a Cartesian coordinate system, this is an equation for a parabola and can be graphically represented as figure 1 below :

Gradient Descent Illustration (by example) :

To minimize the function above, we need to find the value of X that produces the lowest value of Y which is at the dot (-1,-1).

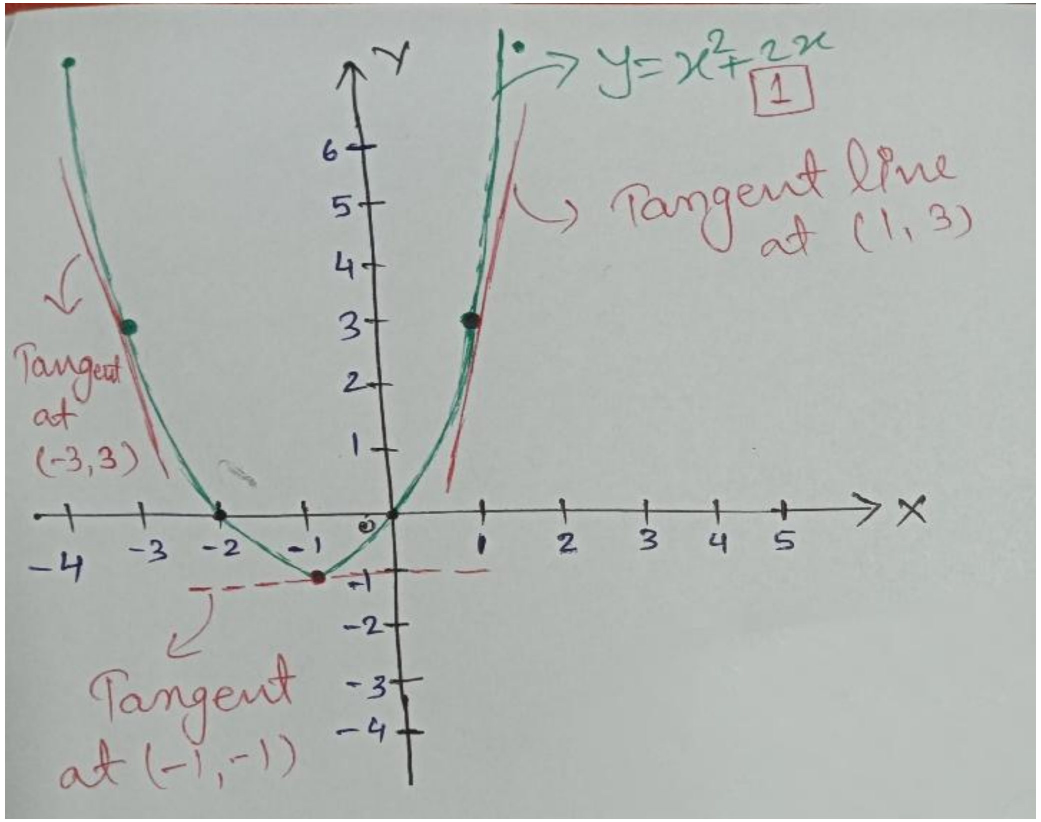


Figure 1: Red line shows the tangent at any particular point

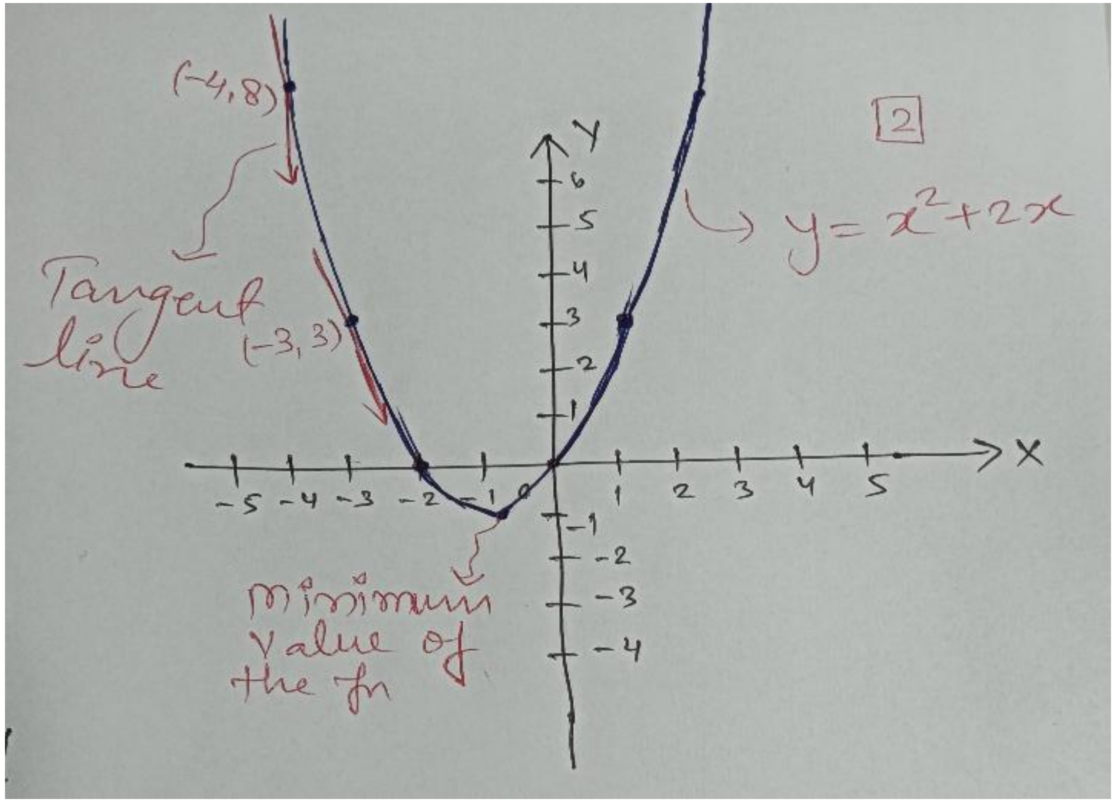
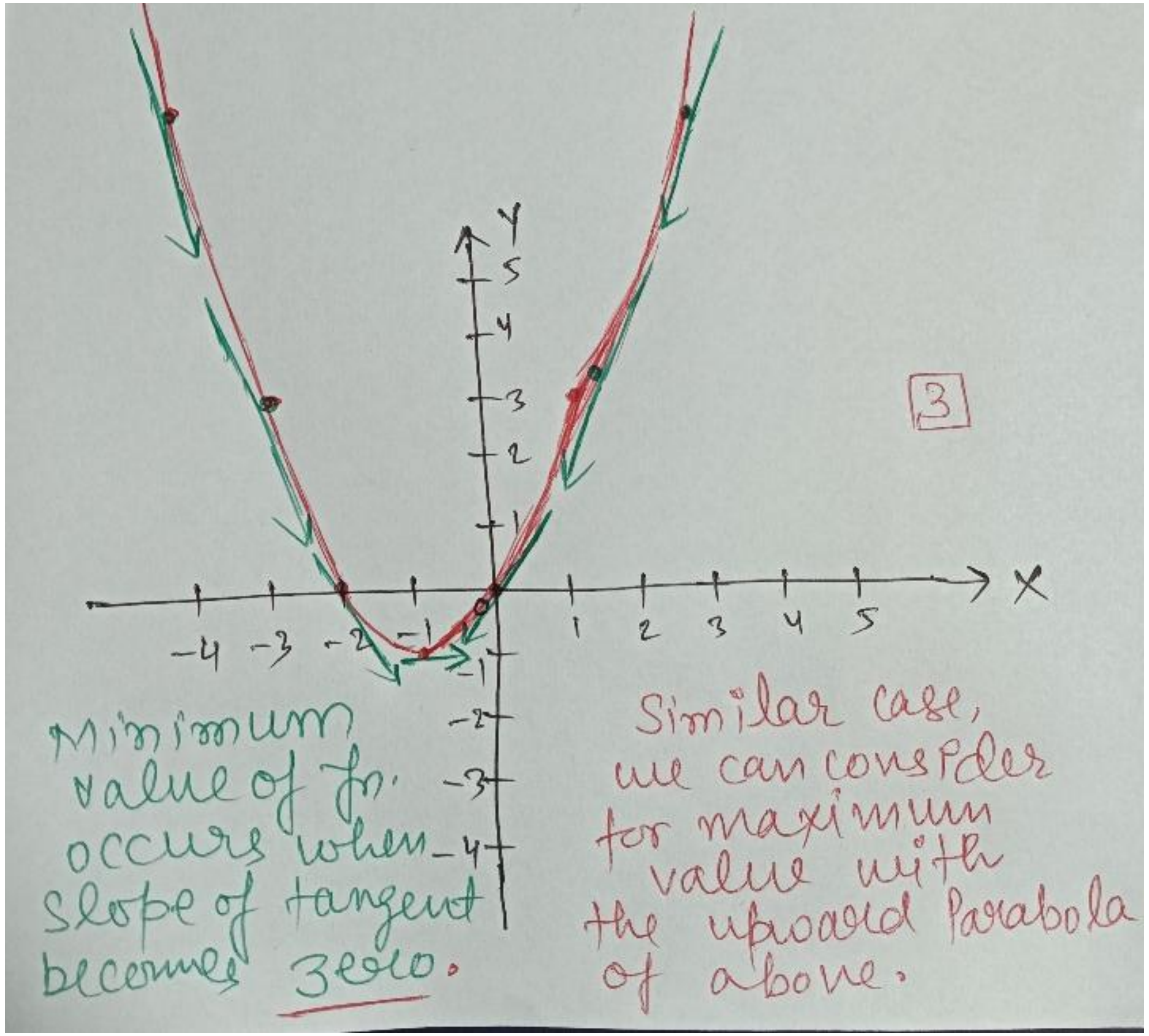


Figure 2: Minima of the function will be at point (-1,-1)

Gradient Descent Illustration (by example) :



It is quite easy to locate the minima here since it is a 2D graph but this may not always be the case especially in case of higher dimensions.

For those cases, we need to devise an algorithm to locate the minima, and that algorithm is called Gradient Descent.

Gradient Descent: Mathematical Formulation

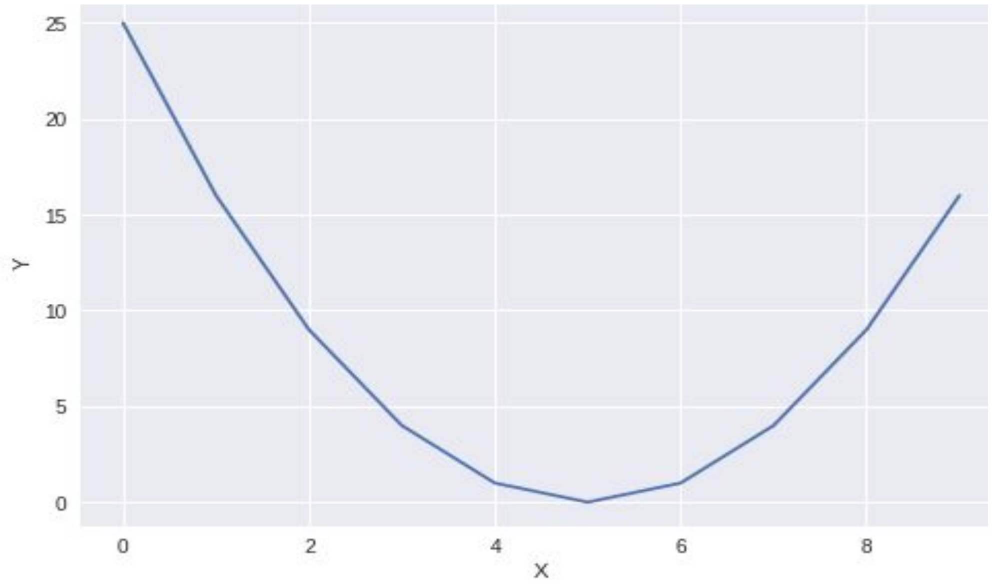
$$a_{t+1} = a_t - \gamma \nabla f(a_t)$$

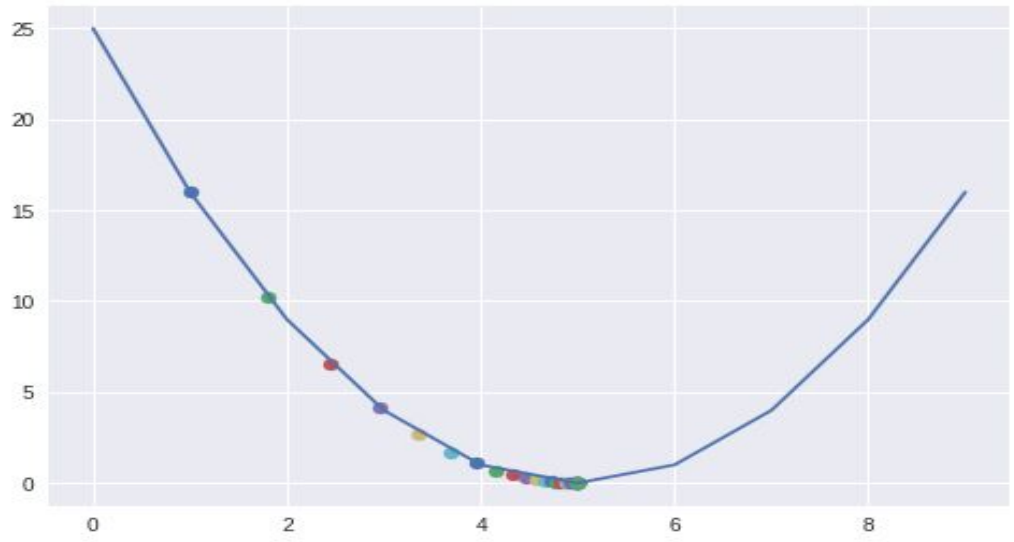
- ➢ a_{t+1} is the next position of our climber
- ➢ a_t represents his current position
- ➢ Minus sign refers to the minimization part of the gradient descent algorithm
- ➢ γ in the middle is a learning rate
- ➢ $\nabla f(a_t)$ is simply the direction of the steepest descent

Source: https://medium.com/@sunil.jangir07/the-outline-of-gradient-descent-da7763a0d66c

Illustration 1 : Gradient Descent for $y = (x-5)^2$

$y = (x-5)^2$ and the derivative of y w.r.t. x is $dy/dx = 2(x-5)$





This is the result of last five iterations:

- ❑ 4.999825775428136
- ❑ 4.999860620342509
- ❑ 4.999888496274007
- ❑ 4.999910797019206
- ❑ 4.999928637615365

```
x = 0 #initial iteration
learning_rate = 0.1
grad = 2 * (x-5)
x = x - learning_rate *
grad
```

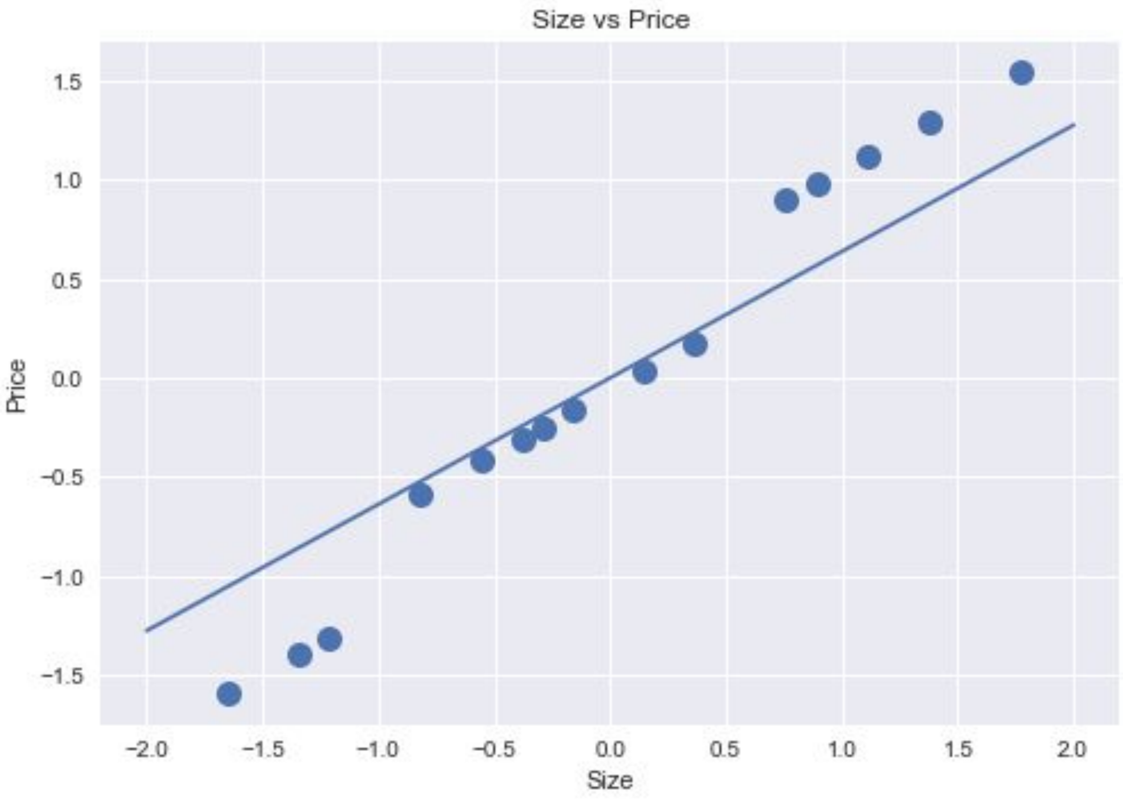
Illustration 2 : Gradient Descent

Size	Bedroom	Price
110	2	399900
129	3	416700
151	3	485600
179	4	586500
144	3	463500
120	2	403400
188	4	614500
141	3	457600
168	4	545600
135	3	434500
139	3	446500
156	3	502500
165	4	534500
173	4	567500
117	2	400000

- ❖ In this example, We can predict the house prices on the basis of size (in Square fit) and number of bedrooms.
- ❖ If we assume there is a linear relationship for the given data.
- ❖ We have used Gradient Descent algorithm for this problem.

Illustration 2 : Gradient Descent

Implementation after gradient descent algorithm



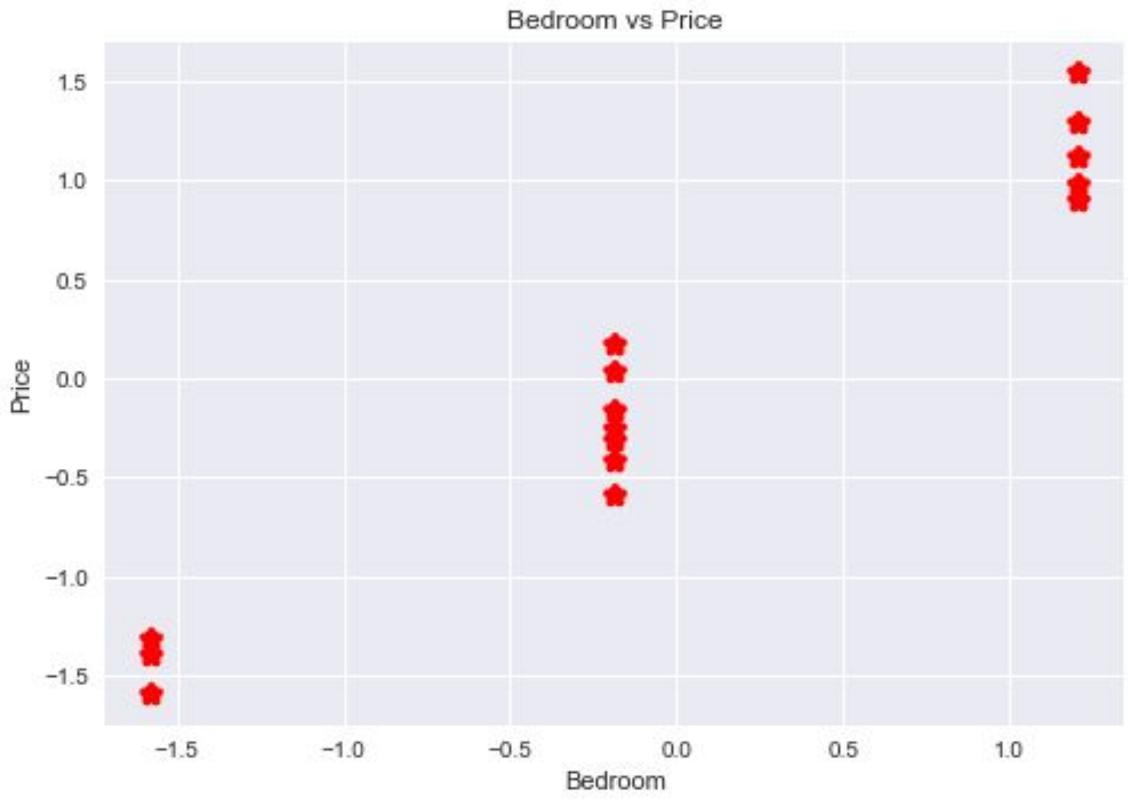
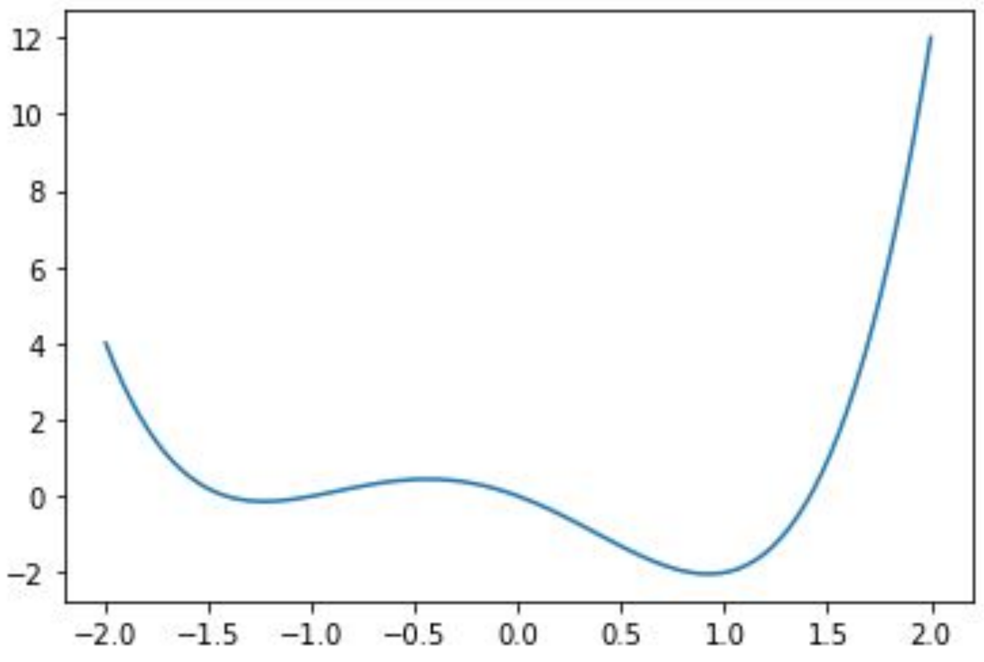
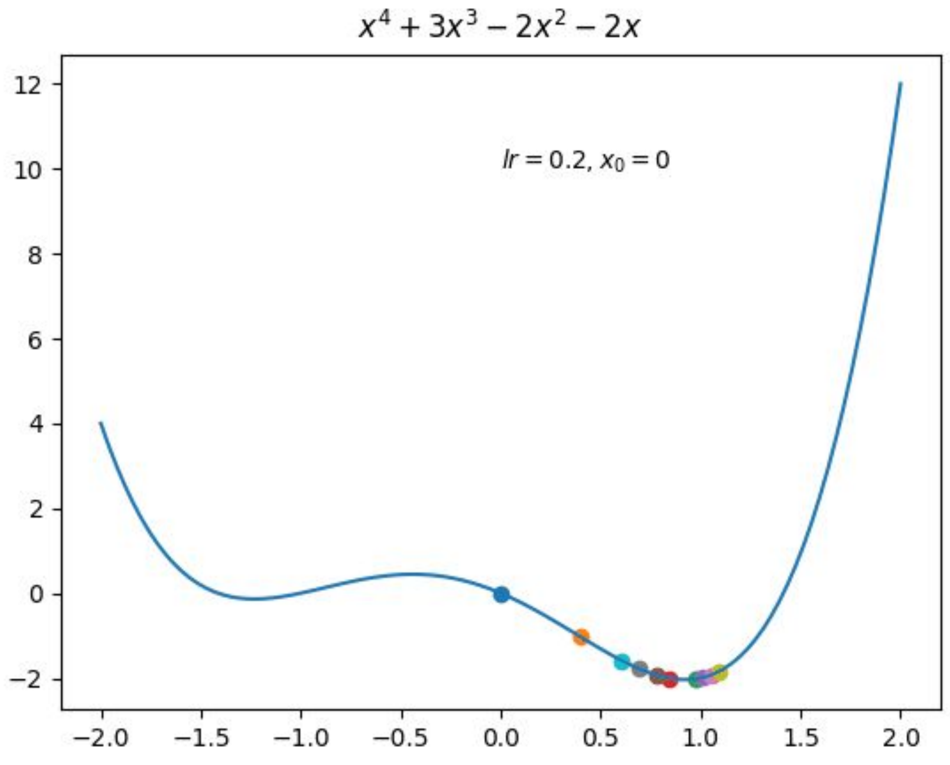


Illustration 3 : Gradient Descent

$y = x^4 + x^3 - 2x^2 - 2x$ and $dy/dx = 4x^3 + 3x^2 - 4x - 2$



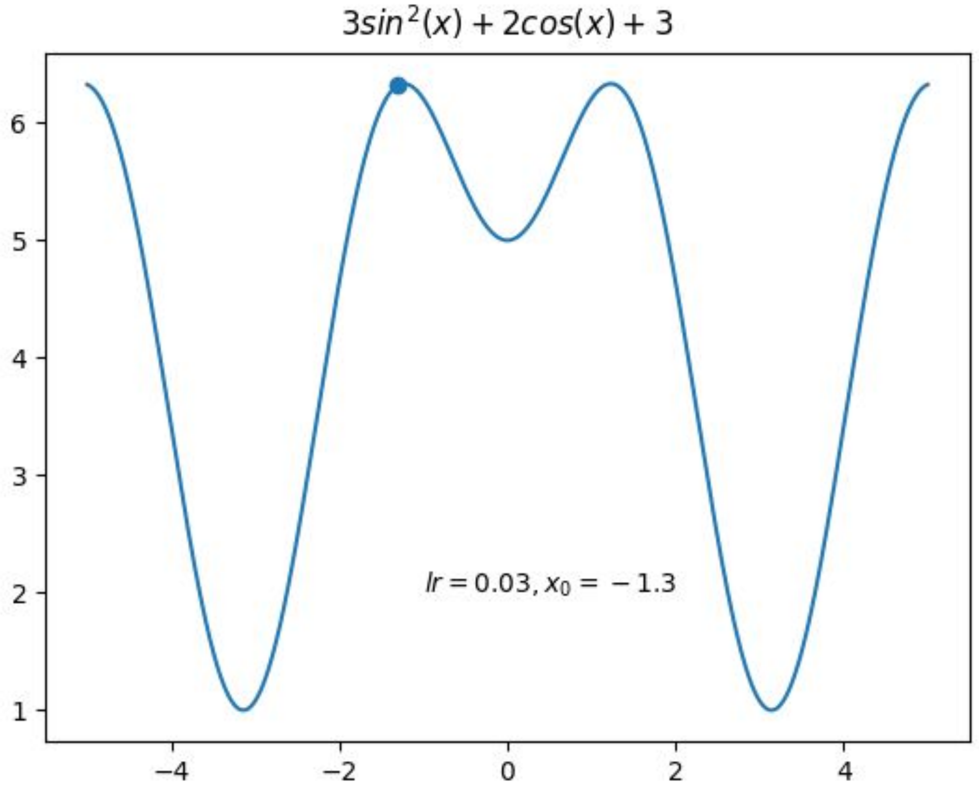
The above is graph of function.



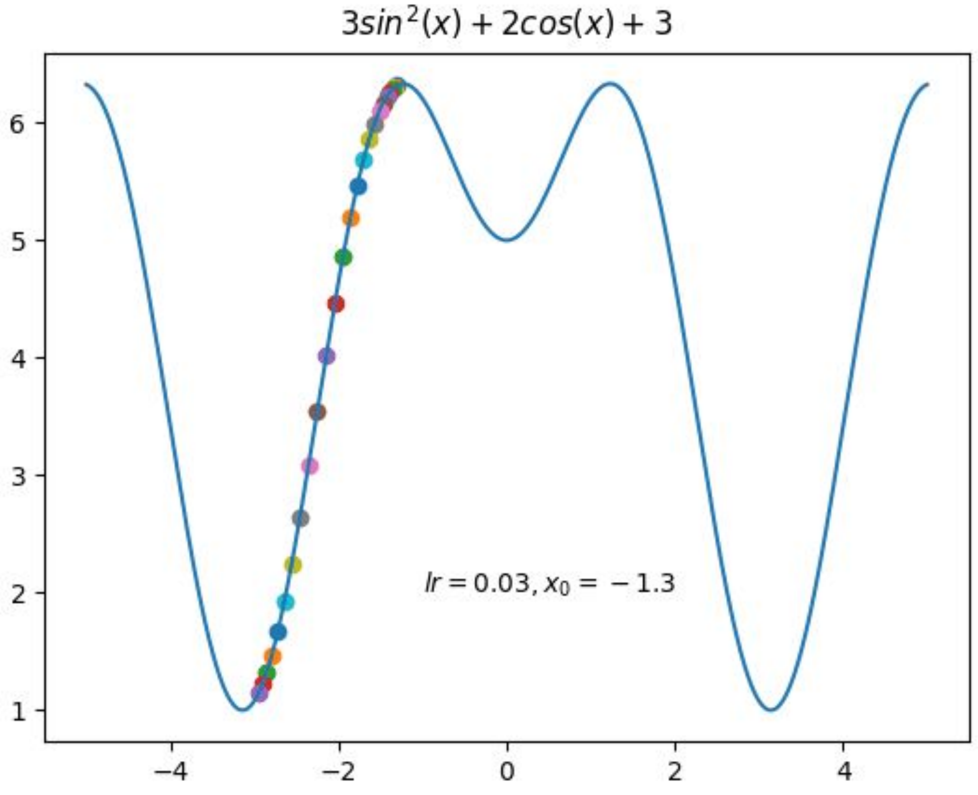
Implementation of Gradient Descent

Illustration 4: Gradient Descent :

$y = 3\ \mathrm{Sin}^2(x) + 2\ \mathrm{Cos}(x) + 3$ and dy/dx is $6\ \mathrm{Cos}(x) - 2\ \mathrm{Sin}(x)$



This is the graph of function.



Implementation of Gradient Descent.

Difference between Gradient (Steepest) Descent and Stochastic Gradient Descent :

Say we have 10,000 data points and 10 features. The sum of squared residuals consists of as many terms as there are data points, so 10000 terms in our case. We need to compute the derivative of this function with respect to each of the features, so in effect we will be doing 10000 * 10 = 100,000 computations per iteration.

It is common to take 1000 iterations, in effect we have 100,000 * 1000 = 100000000 computations to complete the algorithm. That is pretty much an overhead and hence gradient descent is slow on huge data.

What is Stochastic Gradient Descent?

It is while selecting data points at each step to calculate the derivatives. Stochastic Gradient Descent randomly picks one data point from the whole data set at each iteration to reduce the computations enormously.

Source: https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31

Thank You !