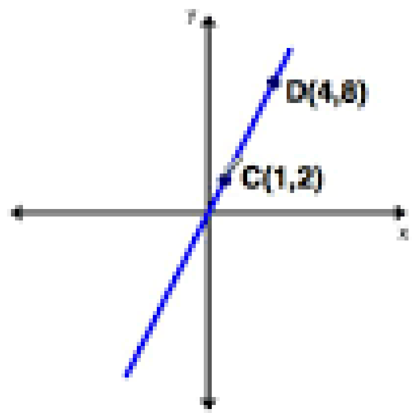
Introduction to Regression
Supplementary Material

Slope

Definition: A slope or gradient of a line represents the steepness/direction of a line.
How far away from the horizontal it is?
It is calculated as

$$slope = \frac{rise}{run} = \frac{\Delta y}{\Delta x} = \frac{change\ in\ y}{change\ in\ x} = \frac{y_2 - y_1}{x_2 - x_1}$$

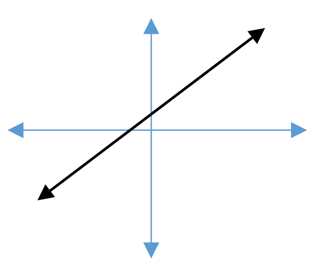
Example: Calculate the slope between the points C and D



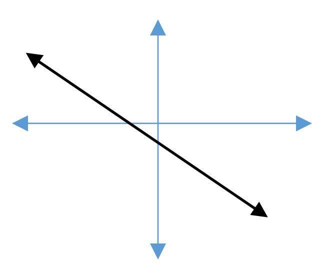
$$slope = \frac{rise}{run} = \frac{\Delta y}{\Delta x} = \frac{change\ in\ y}{change\ in\ x} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{8-2}{4-1} = \frac{6}{3} = 2$$

The slope of the line is positive.

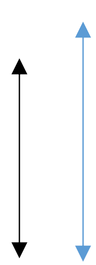
Below you will see graphs with slopes that are positive, negative, zero and undefined.



Line: y = 2x +1
Slope: positive



Line: y = -3x - 2
Slope: negative



Line: x = -5
Slope: undefined (extremely steep)

Line: y = 4
Slope: zero, no steepness

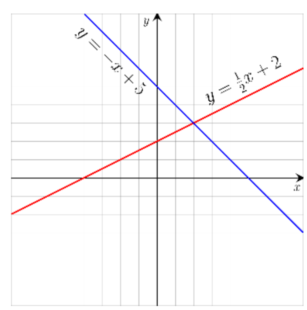
Slope-intercept equation of a line

The equation of a line is given by: $y = \underbrace{m}_{slope} x + \underbrace{b}_{y-\mathrm{int}}$

If m is positive, slope is positive.

If m is negative, slope is negative.
b represents the y-intercept which the point where the line crosses the y-axis. At this point, x = 0. So the coordinates of the y-intercept are (0,b)

Different types of polynomials and corresponding graphs

Polynomial	Equation	Graph
linear	y = ax + b ; a nonzero	 Lines

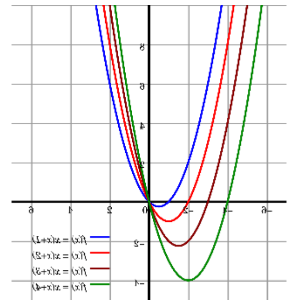
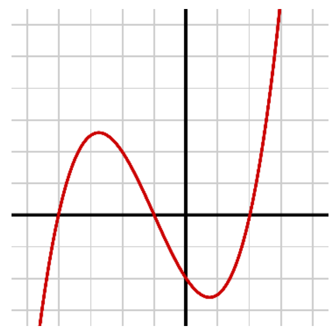
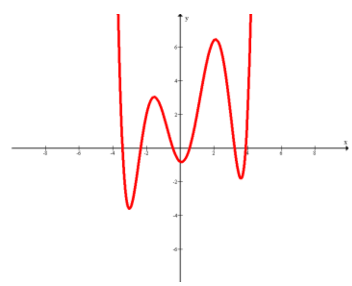
quadratic	$y = ax^2 + bx + c$; a nonzero	 Parabolas(open up or down)
cubic	$y = bx^2 + cx + d$; a nonzero	
polynomials	$Y = ax^n + bx^{n-1} + cx^{n-2} + ... z$	

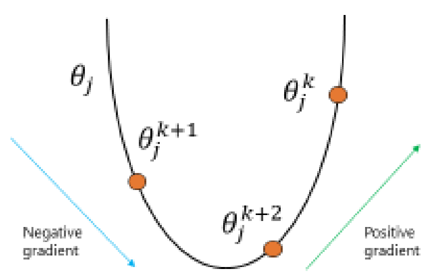
Table of derivatives

Table of derivatives

Functions	Derivative
$y = x^n$	$\frac{dy}{dx} = y' = n \cdot x^{n-1}$
$y = 5x$	$\frac{dy}{dx} = 5x^0 = 5(1) = 5$
$y = x^2 + 7x + 3$	$\frac{dy}{dx} = 2x + 7$
$y = 4x^3$	$\frac{dy}{dx} = (3)(4)x^{3-1} = 12x^2$
$y = 5x^4 - 2x^2$	$\frac{dy}{dx} = 20x^3 - 4x$

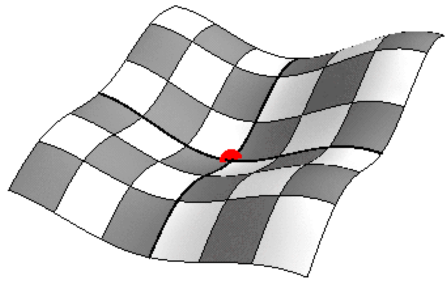
Gradient

Definition: It is a basically the derivative, or the rate of change of a function. IT is basically used for functions with multiple variables. The gradient is zero at a local maximum or minimum.

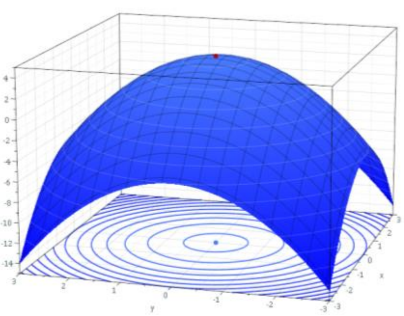


Minimum and maximum values.

Definition: The **minimum** value of a function is the **lowest** point of a graph. For a quadratic function, the minimum value is the vertex.



Definition: The **maximum** of a function is the **highest** point of a graph. For a quadratic function, the maximum value is the vertex.



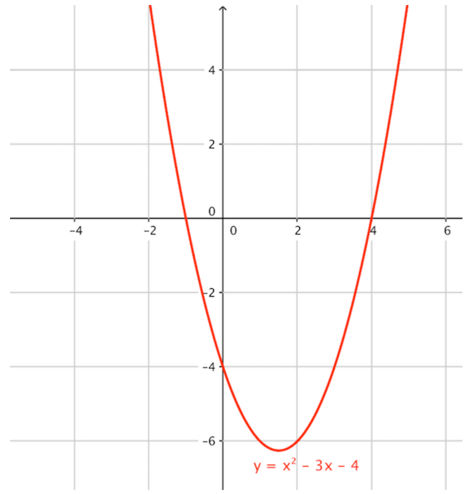
Minima/maxima can be determined graphically by visually looking at a graph (turning points) or can be computed mathematically. There is one important characteristic of minima and maxima. The slope of the tangent line is zero! That is, if you were to draw a line through the minimum or maximum value of a function, that was tangent to the function, the line would be horizontal, with slope 0. In other words, the derivative of the function at that point is zero.

$$\frac{df}{dx} = f'(x) = 0$$

<u>Example</u>: For the following function, identify if it has a maximum or minimum, and calculate the coordinates?

$$f(x) = x^2 + 4x + 4$$

Solution: This is a quadratic function, whose shape is a parabola It is open upward. That means the there is a minimum at the vertex. The graph looks like:



We can easily see that the graph has a minimum at (1.5, 6.25). Notice that the slope changes from negative to zero to positive.

To calculate the minimum mathematically,
Step 1: Find derivative of f(x) and set to 0. Then solve for x.

$$f(x) = x^2 - 3x - 4$$

$$f'(x) = \frac{df}{dx} = 2x - 3$$

$$2x - 3 = 0$$

$$x = 3/2 = 1.5$$

This means the minimum value occurs at x =1.5. To find the actual minimum value, we need to find the y value at this point. We can plug it back into the equation:

$$f(x) = x^2 - 3x - 4$$

$$f(1.5) = 1.5^2 - 3(1.5) - 4$$

$$f(1.5) = -6.25$$

The minimum value is -6.25

<u>Example:</u> Find two non-negative numbers whose sum is nine, and the product of one number and the square of the other number is a maximum.

Solution: Let x and y represent the two non-negative numbers. Let P represent the product of the two numbers. Our goal is to maximize P.

$$x + y = 9$$
$$P = xy^2$$

Solve for y in the first equation.

$$x + y = 9$$
$$y = 9 - x$$

Now substitute this into the second equation:

$$P = x(9 - x)^2$$
$$P = x(9 - x)(9 - x)$$
$$P = x(81 - 18x + x^2)$$
$$P = 81x - 18x^2 + x^3$$

This is a cubic function. In order to find the maximum, we must find the derivative (using chain rule) and set it to zero. Once we solve for x, we will need to put it back into the equation, to get y which represents our maximum value.

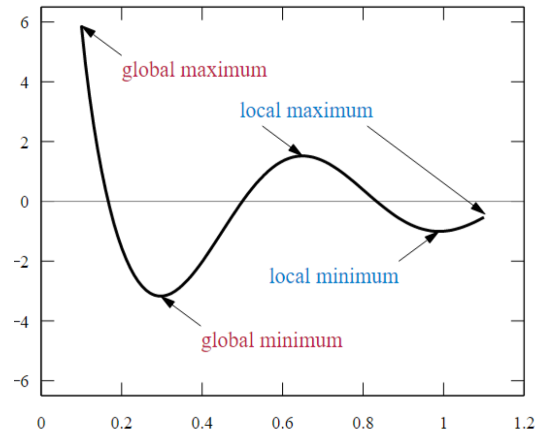
$$P = 81x - 18x^2 + x^3$$
$$\frac{dP}{dx} = P'(x) = 81 - 36x + 3x^2$$
$$P'(x) = 0$$
$$81 - 36x + 3x^2 = 0$$
$$3(27 - 12x + x^2) = 0$$
$$x^2 - 12x + 27 = 0$$
$$(x - 9)(x - 3) = 0$$
$$x - 9 = 0 \quad or \quad x - 3 = 0$$
$$x = 9 \qquad\qquad x = 3$$

The minimum value at x=9 is y=0, minimum value is 9(0^2) = 0
The maximum value at x=3 is y=6, maximum value is 3(6^2) = 108

Absolute versus local maximum/minimum

Functions can have "hills/peaks" or "valleys/troughs" and hence can have multiple minima and maxima. Therefore it is important to distinguish between local versus absolute (global) minima or maxima.

Absolute maxima and minima are referred to as extreme value, i.e. the highest or lowest point on the entire domain of a function.



Application – optimization problems

Real-world applications include finding the minimum/maximum height, cost, profit, power or area/volume. The process of finding the maximum/minimum is called optimization. Differentiation can be used to find these extreme values i.e. maxima/minima.

Minimize distance, maximize volume

<u>Example:</u> The daily profit of an oil company is given by:

$$P = 4x - 0.02x^2$$

Where x is the number of barrels of old refined. How many barrels will give the maximum profit and what is the maximum profit?

Solution: Maximum profit occurs when the derivative =0.

$$\frac{dP}{dx} = P'(x) = 0$$
$$P'(x) = 4 - 0.04x = 0$$

Solve for x

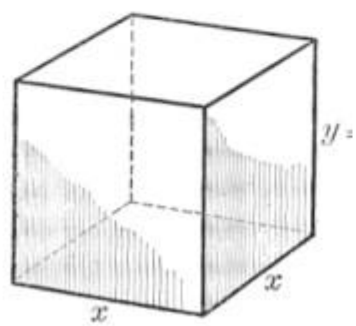
$$4-0.04x=0$$
$$x=\frac{4}{0.04}=100$$

Note: This is NOT the answer. The maximum value occurs at this value. The actual maximum Profit is

$$P(100)=4(100)-0.02(100)^2=\$200$$

Thus, if the company has 100 barrels of refined oil, the maximum profit will be \$200.

Example: A box having a square base and an open top is to contain 256ft^3. What should its dimensions be so that the material to make it will be a minimum? That is, what dimensions will cost the least?



Solution: Let x be the side of the square base, and let y be its height.

$$Area\ of\ base=x^2$$
$$Area\ of\ four\ sides=4xy$$
$$Surface\ Area=x^2+4xy$$

Now, express y in terms of x. Use Volume to do this.

$$Volume = (x^2)(y) = 256$$

$$y = \frac{256}{x^2}$$

$$Surface\,Area = x^2 + 4xy$$

$$Surface\,Area = x^2 + 4x\left(\frac{256}{x^2}\right)$$

$$Surface\,Area = x^2 + \frac{1024}{x}$$

Now, that we have the Surface Area function, in order to find he dimensions that will minimize it, we need to find the derivative and set to 0 and solve for x. We will then compute the value of the function at this x-value.

$$\frac{dS}{dx} = S'(x) = 2x - \frac{1024}{x^2}$$

$$2x - \frac{1024}{x^2} = 0$$

$$2x = \frac{1024}{x^2}$$

$$2x^3 = 1024$$

$$x^3 = 512$$

$$x = \sqrt[3]{512} = 8$$

So we need a length of 8 feet for the base, and we need to calculate the height y.

$$y == \frac{256}{x^2} = \frac{256}{8^2} = \frac{256}{64} = 4$$

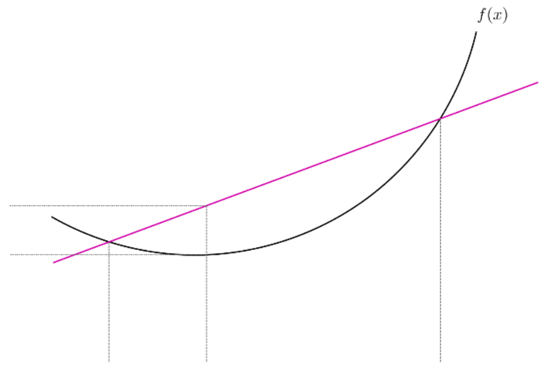
These are the dimensions that will minimize the cost.

<u>Example:</u>

There are 50 apple trees in an orchard. each tree produces 800 apples. for each additional tree planted in the orchard, the apple output per tree drops by 10 apples. how many additional trees should be planted in the existing orchard in order to maximize the apple output of the orchard?

Convex functions

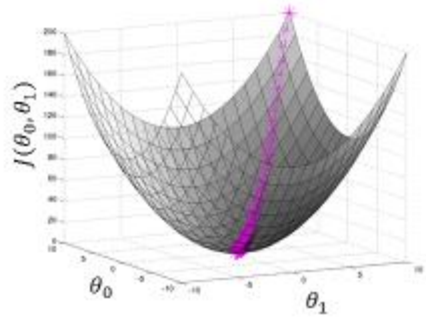
Definition: A function is called convex if the line segment between any two points on the graph of the function lies above or on the graph of the function. Examples include exponential functions, quadratic functions, etc.…
They are especially important in the study of optimization problems.



Why are convex functions good for minimization?
Gradient descent is used when you are trying to find the point where the gradient is zero. When your first-order derivative system is a linear system of equations, it is very easy to compute the maxima/minima; however, when your first-order derivative is NOT a linear system, gradient descent may be more appealing.
Advantages of gradient descent: It is extremely easy to implement and has a lower computational cost, which plays a big role with large datasets.
Pitfalls of gradient descent: It has a relatively poor rate of convergence and can potentially zigzag near the optimal value.

Convex functions in particular will yield a global minimum. If the function is non-convex, then the issue with gradient descent is that it will find a local minimum and not necessarily the absolute minimum. For convex functions, gradient descent will always eventually converge given a small enough step size and infinite time.

$$h_\theta(x) = \theta_0 + \theta_1 x \qquad J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta(x^i) - y^i\right)^2$$



Mean Absolute Error

The MAE measures the average magnitude of the errors in a set of predicted values. It measures the accuracy for continuous variables. It is the absolute average difference between the predicted and observed values. It represents a linear score, i.e. all individual difference are weighted in the average.

The Mean Absolute Error is a mean of residual values. It is the “pure” measure of error.

Formula

$$MAE(\theta) = \frac{\sum_{i=1}^{n}\left|h_\theta(x^i) - y^i\right|}{n}$$

Example:

$$y = \{36, 19, 34, 6, 1, 45\}$$
$$h_\theta(x) = \{27, -2.6, 13, -7.3, -2.6, 48\}$$
$$|h_\theta(x) - y| = \{9, 21.6, 21, 13.3, 3.6, 3\}$$
$$MAE(\theta) = \frac{71.5}{6} = 11.9$$

Root Mean Square Error (RMSE)

The RMSE is a quadratic coring rule which measure the average magnitude of the error. The difference between the predicted and observe values are each squared, and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE provides a relatively high weight to large errors. Hence it is most useful when large errors are undesirable.

The Root-Mean-Square-Error is the square root of the mean of squared residuals. It penalizes large errors more than small ones, which makes it good when large errors are particularly bad.

Formula

$$RMSE(\theta) = \sqrt{\frac{\sum_{i=1}^{n}(h_\theta(x^i) - y^i)^2}{n}}$$

Example

$$y = \{36, 19, 34, 6, 1, 45\}$$
$$h_\theta(x) = \{27, -2.6, 13, -7.3, -2.6, 48\}$$
$$(h_\theta(x) - y)^2 = \{81, 467, 441, 177, 13, 9\}$$
$$MAE(\theta) = \sqrt{\frac{1187}{6}} = 14.1$$

RMSE is very sensitive to outliers. So remove outliers before calculating the RMSE.
Both the RMSE and MAE can be used to understand the variation in errors. The RMSE will always be greater than or equal to the MAE. The greater the difference between them the greater the variance in the errors. If RMSE = MAE, then all the errors are of the same magnitude. If RMSE > MAE then there is variation in the magnitude of the errors.

They can range from 0 to infinity. Lower values are better.