1. A function f has a local maximum or relative maximum at a point x_o if the values $f(x)$ of f for x near x_o are all less than $f(x_o)$. Thus, the graph of f near x_o has a peak at x_o. A function f has a local minimum or relative minimum at a point x_o if the values $f(x)$ of f for x near x_o are all greater than $f(x_o)$. Thus, the graph of f near x_o has a trough at x_o. (To make the distinction clear, sometimes the plain maximum and minimum are called absolute maximum and minimum.)

To find the local maxima and minima of a function f on an interval $[a, b]$:

Solve $f'(x) = 0$ to find critical points of f.

Drop from the list any critical points that aren't in the interval $[a, b]$.

Add to the list the endpoints (and any points of discontinuity or non-differentiability): we have an ordered list of special points in the interval:

$$a = x_o < x_1 < \cdots < x_n = b$$

Between each pair $x_i < x_{i+1}$ of points in the list, choose an auxiliary point t_{i+1}. Evaluate the derivative f' at all the auxiliary points. For each critical point x_i, we have the auxiliary points to each side of it: $t_i < x_i < t_{i+1}$. There are four cases best remembered by drawing a picture!:

if $f(t_i) > 0$ and $f'(t_{i+1}) < 0$ (so f is increasing to the left of x_i and decreasing to the right of x_i, then f has a local maximum at x_o).

if $f'(t_i) < 0$ and $f'(t_{i+1}) > 0$ (so f is decreasing to the left of x_i and increasing to the right of x_i, then f has a local minimum at x_o).

if $f'(t_i) < 0$ and $f'(t_{i+1}) < 0$ (so f is decreasing to the left of x_i and also decreasing to the right of x_i, then f has neither a local maximum nor a local minimum at x_o).

if $f'(t_i) > 0$ and $f'(t_{i+1}) > 0$ (so f is increasing to the left of x_i and also increasing to the right of x_i, then f has neither a local maximum nor a local minimum at x_o).

The endpoints require separate treatment: There is the auxiliary point t_o just to the right of the left endpoint a, and the auxiliary point t_n just to the left of the right endpoint b:

At the left endpoint a, if $f'(t_o) < 0$ (so f' is decreasing to the right of a) then a is a local maximum.

At the left endpoint a, if $f'(t_o) > 0$ (so f' is increasing to the right of a) then a is a local minimum.

At the right endpoint b, if $f'(t_n) < 0$ (so f' is decreasing as b is approached from the left) then b is a local minimum.

At the right endpoint b, if $f'(t_n) > 0$ (so f' is increasing as b is approached from the left) then b is a local maximum.

2. Finding the derivative of $f(x)$ we get, $f'(x) = 3x^2 - 6x$. Solving the equation $f'(x) = 0$ that is, $3x^2 - 6x = 0$, we find the critical points at $x = 0$ and $x = 2$. Now, evaluating the function at these points, we find:

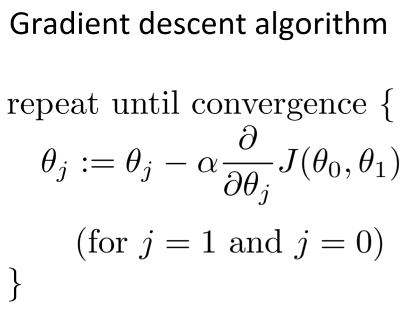
We also note that $f''(x) < 0$ at $x = 0$ which means there is a local maxima at $x = 0$

Similarly, $f''(x) > 0$ at $x = 2$ which means there is a local minima at $x = 2$

3. Gradient Descent is an optimization algorithm for finding a local minimum of a differentiable function. Gradient descent is simply used to find the values of a function's parameters (coefficients) that minimize a cost function as far as possible. It is an optimization algorithm that's used when training a machine learning model. It's based on a convex function and tweaks its parameters iteratively to minimize a given function to its local minimum. Optimization is a big part of machine learning. Almost every machine learning algorithm has an optimization algorithm at its core. We start with a random point on the function and move in the **negative direction** of the gradient of the function to reach the local/global minima.

Vanilla gradient descent, also known as batch gradient descent, computes the gradient of the cost function w.r.t. to the parameters θ for the entire training dataset:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{1}$$



Gradient descent algorithm

```
repeat until convergence {
    $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$
        (for $j = 1$ and $j = 0$)
}
```

As we need to calculate the gradients for the whole dataset to perform just one update, batch gradient descent can be very slow and is intractable for datasets that don't fit in memory.

As we need to calculate the gradients for the whole dataset to perform just one update, batch gradient descent can be very slow and is intractable for datasets that don't fit in memory.

4. $f'(x) = 4x^3$, so the only critical point is $x^* = 0$. The Hessian is $f''(x) = 12x^2$. Evaluated at x^*, the Hessian is $f''(x^*) = 0$, so the second order condition does not tell us whether x^* is a maximum or a minimum. However, looking at the Hessian for all points in the domain gives us more information: $f''(x) = 12x^2 \geq 0$ for all x, so the Hessian is positive semi-definite on the whole domain and f is convex. Therefore, x^* has to be a global minimum.

5. A Taylor series is a series expansion of a function about a point. A one-dimensional Taylor series is an expansion of a real function $f(x)$ about a point $x = a$ is given by

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \frac{(x-a)^3}{3!}f'''(a) \cdots + \frac{(x-a)^{(n-1)}}{(n-1)!}f^{(n-1)}(a) + \frac{(x-a)^n}{n!}f^n(a) \tag{2}$$

6. To find out the Taylor Series of $log(1+x)$ at $x = 0$:

$$f(x) = log(1+x) \quad f(0) = 1$$

$$f'(x) = \frac{1}{(1+x)} \quad f'(0) = 1$$

$$f''(x) = \frac{-1}{(1+x)^2} \quad f''(0) = -1$$

$$f'''(x) = \frac{2}{(1+x)^3} \quad f'''(0) = 2$$

$$f^{(4)}(x) = \frac{-2 \cdot 3}{(1+x)^4} \quad f^{(4)}(0) = -6$$

Using the Taylor Series in eq (2) above, it follows that, for $|x| < 1$ which is the necessary and sufficient condition for the convergence of the series we have:

$$log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \cdots$$

7(a). The gradient is just the vector of partial derivatives. The partial derivatives of f at the point $(x, y) = (3, 2)$ are:

$$\frac{\partial}{\partial x} f(x, y) = 2xy$$

$$\frac{\partial}{\partial x} f(3, 2) = 12$$

$$\frac{\partial}{\partial y} f(x, y) = x^2$$

$$\frac{\partial}{\partial x} f(3, 2) = 9$$

Therefore, the gradient is:

$$\nabla f(3, 2) = 12i + 9j = (12, 9)$$

7(b). Let $u = u_1 i + u_2 j$ be a unit vector. The directional derivative at (3, 2) in the direction of u is

$$D_u f(3, 2) = \nabla f(3, 2).u$$

$$= (12i + 9j).(u_1 i + u_2 j) = 12u_1 + 9u_2 \tag{3}$$

To find the directional derivative in the direction of the vector (1,2), we need to find a unit vector in the direction of the vector (1,2). We simply divide by the magnitude of (1,2).

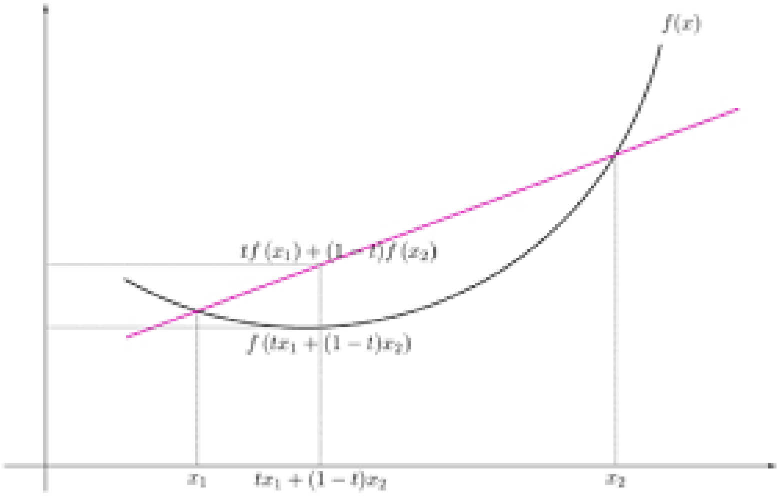
$$u = \frac{(1,2)}{||(1,2)||} = \frac{(1,2)}{\sqrt{(1^2+2^2)}} = \frac{(1,2)}{\sqrt{5}} = (\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}) \tag{4}$$

Plugging this expression for $u = (u1, u2)$ into equation (3) for the directional derivative, and we find that the directional derivative at the point (3,2) in the direction of (1,2) is

$$D_u f(3,2) = 12u_1 + 9u_2 = \frac{12}{\sqrt{5}} + \frac{18}{\sqrt{5}} = \frac{30}{\sqrt{5}}$$

8. Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex, if its domain is a convex set and for all x, y in its domain, and all $\theta \in [0,1]$, we have $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$.



In words, this means that if we take any two points x, y, then f evaluated at any convex combination of these two points should be no larger than the same convex combination of $f(x)$ and $f(y)$.

i. Geometrically, the line segment connecting $(x, f(x))$ to $(y, f(y))$ must sit above the graph of f.

ii. If f is continuous, then to ensure convexity it is enough to check the definition with $\theta = \frac{1}{2}$ (or any other fixed $\theta \in (0,1)$).

iii. We say that f is concave if $-f$ is convex.

Examples of Convex Functions

affine: $ax + b$ on $\mathbb{R}$, for any a, $b \in \mathbb{R}$

exponential: e^{ax}, for any $a \in \mathbb{R}$