Constrained Optimization

[Optimisation]

Recap

$$\min_{\vec{w}, w_0} \; -\sum_{i=1}^{n} y_i \frac{w^T x_i + w_0}{\|w\|}$$

Since this is a multivariate function with d+1 variable (for 'd' features), we cannot directly optimise it using calculus, hence we will use gradient descent

Gradient Descent:

$$\theta_i^{t+1} = \theta_i^t - \eta \frac{\partial f}{\partial \theta_i}$$

- t: iteration #
- η: learning rate
- f: loss function
- θ_i: parameter

For our loss function, we need

$$\frac{\partial f}{\partial w_{1,2\ldots n}} \quad , \quad \frac{\partial f}{\partial w_0}$$

where,

$$f(w_0, w_1, \ldots, w_n) = -\sum_{i=1}^{\hat{}} y_i \frac{w^T x_i + w_0}{\|w\|}$$

Let's try to find derivatives !!

1st: principal way → always works !!

$$\frac{\partial f}{\partial w_1} = \frac{f(w_1 + \Delta, w_2 \ldots w_n) - f(w_1, w_2 \ldots w_n)}{\Delta}$$

$\Delta \sim 0$

However, this requires 2 computations of $f()$, which we know is expensive.

So for our problem, let's simplify the derivative before coding it.

2nd: manual differentiation

$$\frac{\partial f}{\partial w_1} \rightarrow -\sum y_i \left(\frac{w_1 x_1 + w_2 x_2 - w_0}{\sqrt{w_1^2 + w_2^2 - \dots - w_n^2}} \right)$$

$$= -\sum y_i \frac{f(w_1)}{g(w_1)}$$

We can use quotient rule here, but that will be complicated & compute expensive.

Idea:

Consider equation of a hyperplane

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

if we divide it by some number $d = \sqrt{w_1^2 + w_2^2} = \|\vec{w}\|$, we get

$$w_1' x_1 + w_2' x_2 + w_0' = 0$$

where

$$w_1' = w_1 / \|\vec{w}\|, \quad w_2' = w_2 / \|w\|$$

Q: is this the same hyperplane?

→ Desmos

Hence our equation becomes

$$w^T x + w_0 = 0 \longrightarrow \hat{w}^T x + \underbrace{\frac{w_0}{\|w\|}}_{w_0'} = 0$$

Going back,

$$\min_{w, w_0} \; -\sum y_i (w^T x + w_0)$$

such that $\underbrace{\|w\| = 1}_{\text{condition}}$

This is called constrained optimisation

Now,

$$\frac{\partial f}{\partial w_1} = -\sum y_i \left[\left(\underbrace{w_1 x_1}_{x_i} + \cancel{w_2 x^0} \cdots \cancel{0} \right) + \cancel{w_0^0} \right]$$

$$= -\sum y_i x_{1i}$$

$$\therefore w_j^{(t+1)} = w_j^{(t)} - \eta\left(-\sum_{i=1}^{\hat{}} y_i x_{ji}\right)$$

$$j \neq 0$$

$$\frac{\partial f}{\partial w_0} = -\sum y_i \left(\cancel{w_1 x_1^{0}} \cancel{+ w_2 x_2^{0}} \cdots - \frac{w_0}{1}\right)$$

$$\therefore \frac{\partial f}{\partial w_0} = -\sum y_i$$

$$\therefore w_0^{(t+1)} = w_0^{t} - \eta\left(-\sum y_i\right)$$

Hence we can see that f' is very easy to compute w.r.t all parameters.

General Form of Optimisation Problems

min / max

operation

Optimisation Function → $f(\theta)$

θ ← params

such that / subject to:

constraint functions → $g_1(\theta)$, $g_2(\theta)$ ..

Eg:

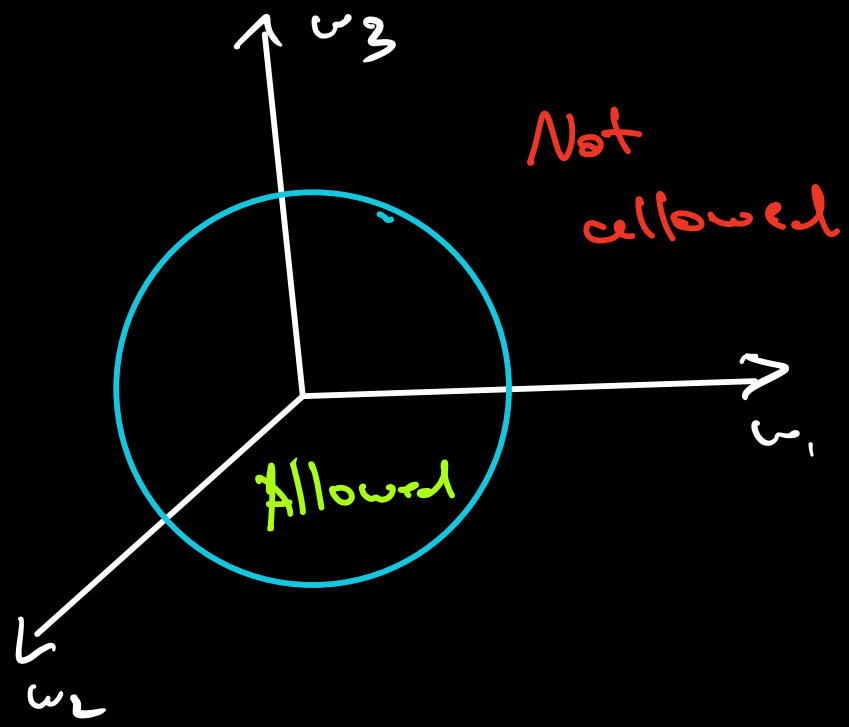
$$\min_{\theta_1, \theta_2} \quad f(\theta_1, \theta_2)$$

subject to $g_1(\theta_1)$, $g_2(\theta_2)$, $g_3(\theta_1, \theta_2)$ ..

We have to solve

$$\min_{\vec{w}, w_0} \quad -\sum y_i \left(w^T x_i + w_0\right); \text{ such that } \|w\| - 1 = 0$$

Essentially we are restricting our search space to a hyper-sphere of radius = 1



Q: Can this be solved with simple (vanilla) gradient descent?

→ No, because when we update, we dont take care of the constraint,

$$w_{new} = w_{old} - \eta \nabla_w f$$

↑

there is no gaurantee that $\|w_{new}\| = 1$

We need to convert this to an unconstrained problem again. To do this we use a new technique called

Lagrange Multipliers

Lagrange Multipliers

$$\min_{\theta} \; f(\theta) \; ; \text{ such that } \; g_1(\theta) = 0, \; g_2(\theta) = 0, \; \dots, \; g_n(\theta) = 0$$

$$\downarrow$$

$$\min_{\theta} \; f(\theta) + \lambda_1 g_1(\theta) + \lambda_2 g_2(\theta) \dots$$

λ_1: Lagrange mult

λ_2: Lagrange mult.

minimising this unconstrained problem gives the same result as above.

Example

$$\min_{x,y} : \rightarrow x^2 + y^2$$

such that, $\rightarrow x + 2y - 1 = 0$

→ Plot online

Lagrangian $\rightarrow L(x, y, \lambda)$

$$= x^2 + y^2 + \lambda(x + 2y - 1) = 0$$

$$\min_{x,y,\lambda} \rightarrow x^2 + y^2 + \lambda(x + 2y - 1) = 0$$

$$\frac{\partial L}{\partial x} = 2x + \lambda \quad , \frac{\partial L}{\partial y} = 2y + 2\lambda$$

$$\frac{\partial L}{\partial \lambda} = x + 2y - 1$$

$\therefore \quad 2x + \lambda = 0 \quad \Rightarrow \quad \lambda = -2x$

$2y + 2\lambda = 0 \quad \Rightarrow \quad \lambda = -y$

$x + 2y - 1 = 0 \qquad \therefore \quad y = 2x$

$\therefore \quad x + 2(2x) - 1 = 0$

$\therefore \quad 5x - 1 = 0$

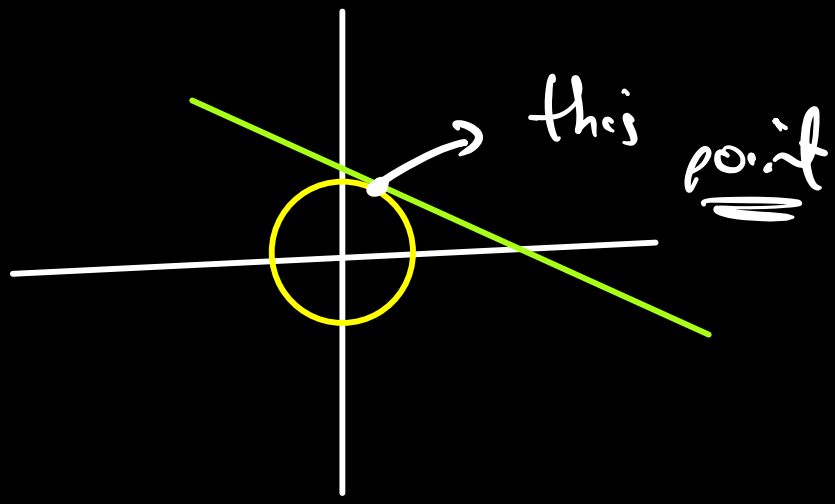
$\therefore \quad x = 1/5$

$\therefore \quad y = 2/5$

$\therefore \quad \lambda = -2/5$

$\Rightarrow$ desired result

$\frac{1}{5} + \frac{2}{5} = \frac{3}{5}$



Intuition (Very high level)

→ Why does this work?

$$\min_{x, y, \lambda} L = f(x, y) + \lambda\, g(x, y)$$

if λ is large, and $g(x, y)$ is > 0

$\lambda\, g(x, y)$ would be large,

hence L would not be min

The algorithm is incentivised to make $g(x, y) \sim 0$, hence the constraint is forced to be satisfied.

Exact proofs of why this works out are beyond our scope.

↳ Note that there are conditions which need to be satisfied for this to work, but we won't discuss them here.

GD with Lagrangian

$$\min_{\vec{w}, w_0, \lambda} \quad L(w, w_0, \lambda)$$

update

$$w_i^{+1} = w - \eta \frac{\partial L}{\partial w_i};$$

$$\lambda^{+1} = \lambda + \eta \frac{\partial L}{\partial \lambda}$$

[max Lambda to force $g(\vec{w}, w_0) \to 0$]

Assessment Doubt

Q.1 $x = (4, 3, 5)$

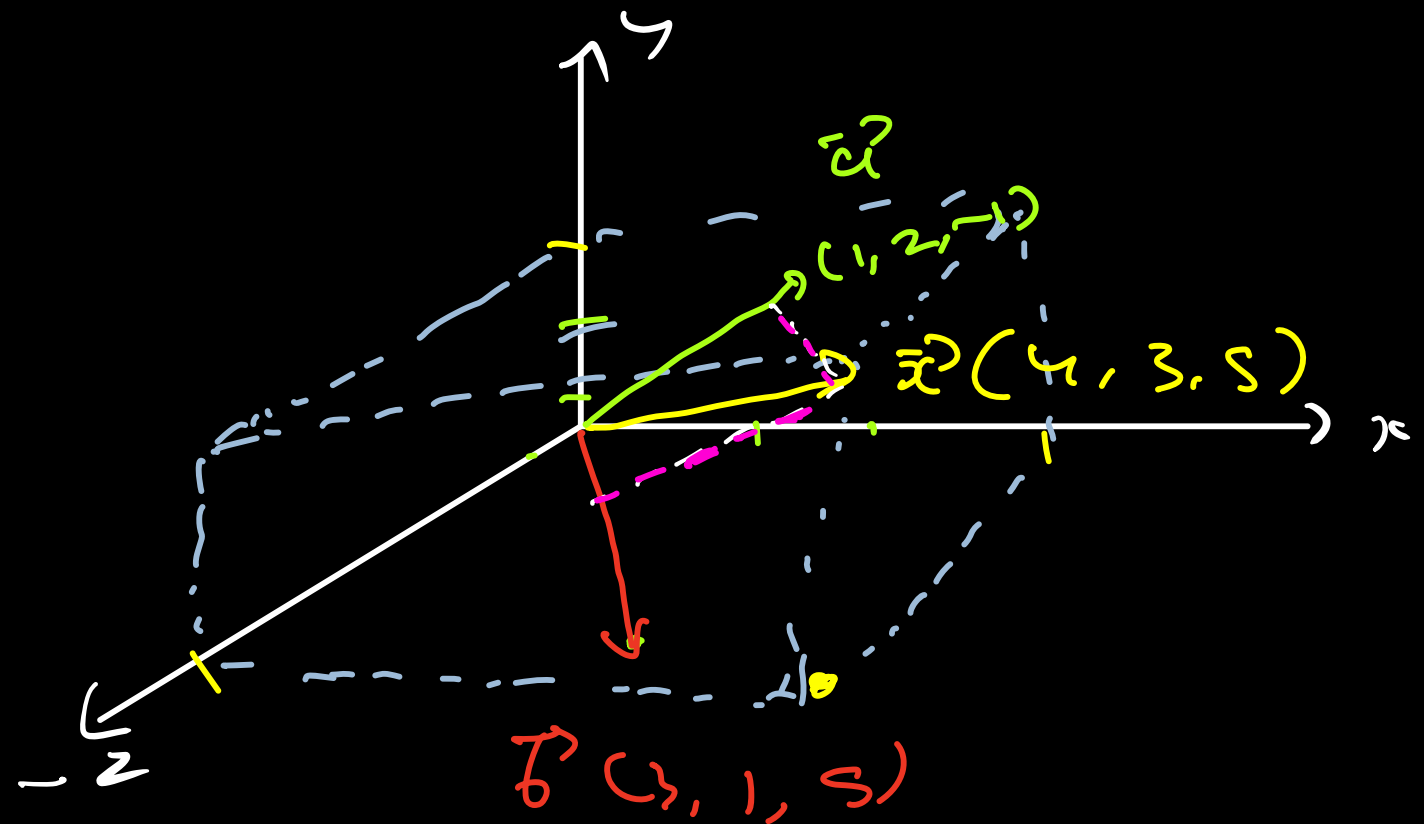
$a = (1, 2, -1)$ $\quad b = (3, 1, 5)$

$$\hat{a} = \left(\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{-1}{\sqrt{6}}\right) \qquad \hat{b} = \left(\frac{3}{\sqrt{35}}, \frac{1}{\sqrt{35}}, \frac{5}{\sqrt{35}}\right)$$

$$x = \frac{\vec{x}\cdot\vec{a}}{\|a\|}\cdot\hat{a} + \frac{\vec{x}\cdot\vec{b}}{\|b\|}\cdot\hat{b}$$

$$= \frac{5}{\sqrt{6}}\cdot\hat{a} + \frac{40}{\sqrt{35}}\cdot\hat{b}$$

also $\vec{a}\cdot\vec{b} = 0$, $\therefore$ $a \perp b$



Further ⇒ Find eqn of the plane containing $\vec{a}$, $\vec{b}$ and $\vec{0}$ → online calc

$11x - 8y - 5z + 0 = 0$

∴ put $\vec{x}$ in it → $11(4) - 8(3) - 5(5) + 0$

$= 44 - 24 - 25$

$= -5 \neq$ x is not on plane

This means that we cannot recreate the full point by projection on those 2 axis. Some information is lost, that's okay. This will form the foundation of our next topic → PCA

Q.2

Let $f(w, w_0) = y\,(w^T x + w_0)$. For a single datapoint $x = (1, 2, 1)$ and $y = -1$, which of the following are the correct update equations for w and w_0 if we are using gradient descent?

Options :

a) $w^{(t+1)} = w^t + \eta x, \quad w_0^{(t+1)} = w_0^{(t)} + \eta.1$

b) $w^{(t+1)} = w^t - \eta x, \quad w_0^{(t+1)} = w_0^{(t)} - \eta.1$

c) $w^{(t+1)} = w^t + \eta x, \quad w_0^{(t+1)} = w_0^{(t)} - \eta.1$

d) $w^{(t+1)} = w^t - \eta x, \quad w_0^{(t+1)} = w_0^{(t)} + \eta.1$

$$w^{(t+1)} = w^t - \eta \nabla f(\vec{w})$$

$$w_0^{(t+1)} = w^t - \eta \nabla f(w^0)$$

$$J = -y_i \left(\frac{w_1 x_1 + w_2 x_2 \cdots w_0}{\|w\|} \right) \rightarrow (1)$$

$$\therefore \frac{\partial f}{\partial w_i} = -y_j (x_1 + 0 + 0 \cdots 0)$$

$$= x_i y_j$$

$$\therefore \frac{\partial f}{\partial w_0} = -y_j (0 + 0 \cdots 1)$$

$$= -y_j$$

$$\therefore \quad \omega^{(t+1)} = \omega^t - \eta\left(-(-1)\cdot x\right) = \boxed{\omega^t - \eta x}$$

$$\omega_0^{(t+1)} = \omega_0^t - \eta\left(-(-1)\right) = \boxed{\omega_0^t - \eta}$$