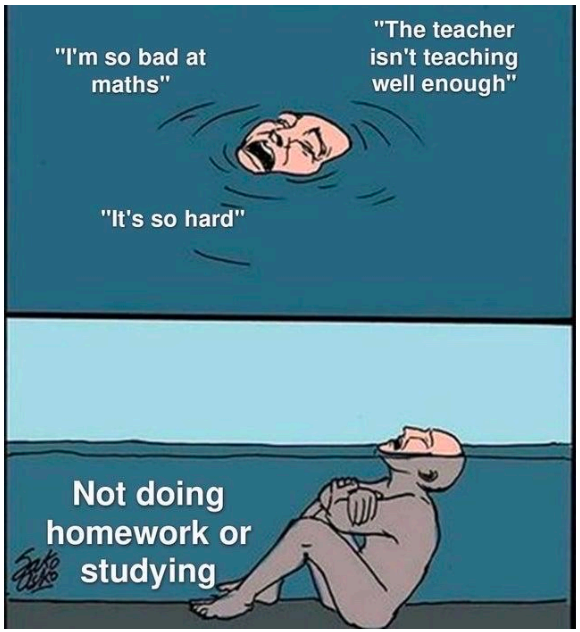
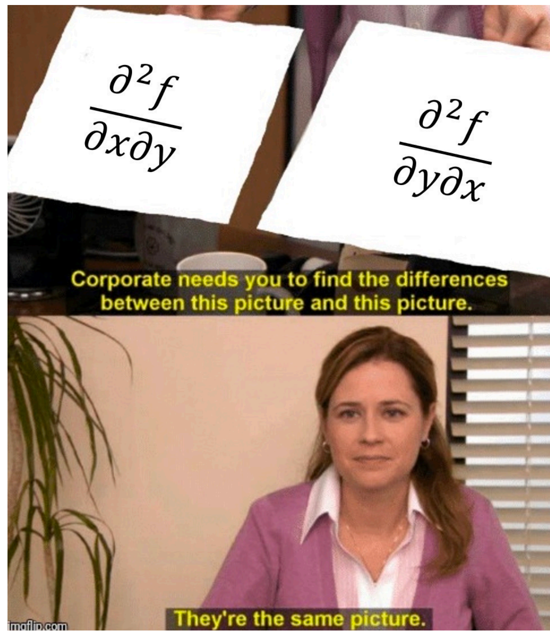
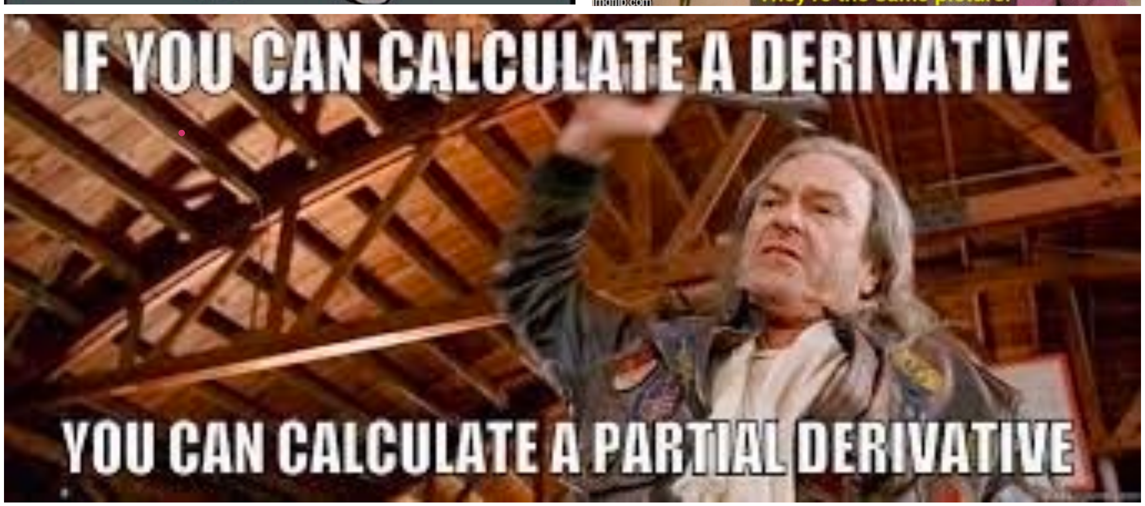
March 7, 2023 DSML : Math for ML.

Optimization 4: Constrained Optimization.







Recap:

(a) Classification - choosing $\bar{w}$ and w_0.

(b) Brute force: very inefficient.

(c) Alternative: gradient descent

(d) Functions, Limits, derivatives.

(e) Partial derivatives and gradients.

Today:

(a) Applying Gradient descent to classification.

(b) Lagrange Multipliers.

(c) Constrained to unconstrained optimization.

Detailed Recap:

(a) Derivative:
$$\frac{d}{dx} f(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}.$$

(b) Gradient descent update rule.

Problem: $\min_{x} f(x)$

$x^{(0)} \rightarrow$ Initial guess for x.

$$x^{(t+1)} = x^{(t)} - \eta \cdot \frac{\partial f(x)}{\partial x}.$$

(c) Partial derivative: $f(x, y, z)$.

$$\frac{d}{dx} f(x,y,z) \quad ; \quad \frac{\partial}{\partial x} f(x,y,z).$$

I cannot treat y, z as constants.

y, z are treated as constants.

(d) $f(\bar{w})$

$\equiv f(w_1, w_2, w_3 \ldots w_d)$

$$\bar{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$\boxed{\nabla_{\bar{w}} f(\bar{w})}$ → "gradient of $f(\bar{w})$ with respect to $\bar{w}$"

$$= \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \\ \frac{\partial f}{w_d} \end{bmatrix}$$

$$\nabla_{\bar{x}} \left[\bar{a}^T \bar{x}\right] = \underline{\underline{\bar{a}}}$$

$$\nabla_{\bar{x}} \left[\bar{x}^T \bar{x}\right] = 2\bar{x}.$$

$$\bar{x} = [\, x_1 \quad x_2 \quad \cdots \quad x_d \,]^T$$

$$f(\bar{x}) = \bar{x}^T \bar{x} = \sum_{i=1}^{d} x_i \cdot x_i = \underbrace{\sum_{i=1}^{d} x_i^2}_{x_1^2 + \underbrace{\sum_{i=2}^{d} x_i^2}_{x_1 \to 0.}}$$

$$\nabla_{\bar{x}} f(\bar{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} = 2x_1 \\ \frac{\partial f}{\partial x_2} = 2x_2 \\ \vdots \\ \frac{\partial f}{\partial x_d} = 2x_d \end{bmatrix}$$

$$= 2 \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = 2\bar{x}.$$

$\bar{x}$!

Optimization problem:

$$g(\underbrace{D}_{\text{Dataset}} ; \underbrace{\bar{w}, w_0}_{\text{parameters}}) = \sum_{i=1}^{n} \left(\frac{\bar{w}^T \bar{x}_i + w_0}{\|\bar{w}\|} \right) y_i$$

$$\max_{\bar{w}, w_0} \; g(D; \bar{w}, w_0)$$

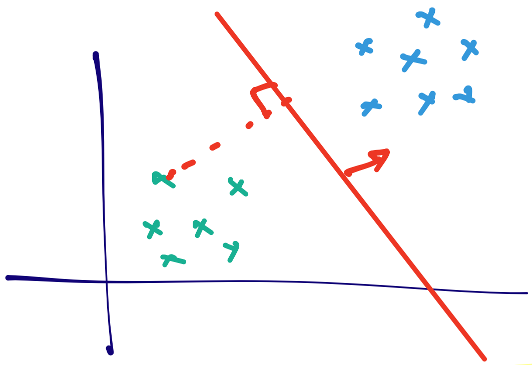
$$D = \{ (\bar{x}_i, y_i) \}_{i=1}^{n}$$

$$\min_{\bar{w}, w_0} \; \ell(D; \bar{w}, w_0)$$

where

$$\ell(D; \bar{w}, w_0) = -g(D; \bar{w}, w_0)$$

$\bar{x}_i$



$$h = \bar{w}^T \bar{x} + w_0$$

$$\boxed{\min_{\bar{w}, w_0} \ \mathcal{L}(D; \bar{w}, w_0)}$$

→ The problem we want to solve using Gradient Descent.

$$= \boxed{\min_{\bar{w}, \bar{w}_0} \ - \sum_{i=1}^{n} \left(\frac{\bar{w}^T \bar{x}_i + w_0}{\sqrt{w_1^2 + w_2^2 \dots}} \right) y_i}$$

It is difficult to calculate derivatives for this loss function because of the denominator.

Solution: Convert this into a constrained optimization problem!!

(New problem on the next page).

Since we are interested only in the direction of $\bar{w}$, we can add a constraint to our problem so that $\|\bar{w}\| = 1$.

$$\min_{\bar{w}, w_0} \; - \sum_{i=1}^{n} \bar{w}^T \bar{x}_i + w_0) \cdot y_i$$

Subject to $\quad ||\bar{w}|| = 1$

" Constrained Optimization problem "

Gradient descent does not work for these types of problems!

Q] How to change a constrained optimization problem into an unconstrained optimization problem?

"Lagrangian multipliers !!"

Constrained optimization problem. —— $\boxed{\text{Lagrangian multipliers}}$ —— Unconstrained optimization problem.

$$\min_{\bar{x}} f(\bar{x})$$

$$\text{s.t.} \quad g(\bar{x}) = 0$$

$$\min_{\bar{x}} f(x) + \lambda\, g(x)$$

Generic problem :

$$\min_{\bar{x}} f(x)$$

$$\text{s.t.} \quad \overset{\lambda_1}{\boxed{g_1(\bar{x}) = 0}}, \quad \overset{\lambda_2}{\boxed{g_2(\bar{x}) = 0}}, \ldots \quad \overset{\lambda_k}{\boxed{g_k(x)}} = 0.$$

Unconstrained problem : (using Lagrangian multipliers).

Step 1 : Count the number of constraint equations. (make sure it is in form shown above).

Step 2 : Introduce new variables to the problem (1 for each constraint equation).

Step 3 : Rewrite the problem as follows :

$$\min_{\bar{x}} f(\bar{x}) + \lambda_1 g_1(\bar{x}) + \lambda_2 g_2(x) + \ldots + \lambda_k g_k(x).$$

Q] Find the minima of

$$f(x) = \underbrace{x^2 - 3x - 3}_{f}$$

s.t. $\underbrace{-x^2 + 2x + 3}_{g} = 0.$

Ans: $x = 3$, $f(x) = -3$.

Apply Lagrange multipliers to solve this.

$$\min_x f(x) \quad \text{s.t. } g(x) = 0 \xrightarrow{\text{convert using L.M.}} \min_x f(x) + \lambda_1 \cdot g(x).$$

$$= \min_x \underbrace{x^2 - 3x - 3 - \lambda_1 x^2 + 2\lambda_1 x + 3\lambda_1}_{\boxed{h(x, \lambda)}}$$

$\boxed{h(x, \lambda)}$ → new function to minimize.

$$\bar{w} = \begin{bmatrix} x \\ \lambda_1 \end{bmatrix}$$

$$\nabla_{\bar{w}} h(\bar{w}) = \begin{bmatrix} \frac{\partial h}{\partial x} \\ \frac{\partial h}{\partial \lambda} \end{bmatrix} = \begin{bmatrix} 2x - 3 + 0 - 2\lambda_1 x + 2\lambda_1 + 0 \\ 0 + 0 + 0 - x^2 + 2x + 3 \end{bmatrix}$$

$$\begin{bmatrix} 2x - 3 + 0 - 2\lambda_1 x + 2\lambda_1 + 0 \\ 0 + 0 + 0 - x^2 + 2x + 3 \end{bmatrix} = \begin{bmatrix} ①\ 2x - 3 - 2\lambda_1 (x - 1) \\ ②\ -x^2 + 2x + 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

now, we have 2 linear equations in terms of λ & x.

① $2x - 3 - 2\lambda_1 (x-1) = 0$.

$= 2x - 3 = 2\lambda_1 (x-1)$

$= \dfrac{2x-3}{2(x-1)} = \lambda_1 \Big\}$ First equation

$x = 3$, $\lambda = \dfrac{2 \times 3 - 3}{2(3-1)} = \dfrac{3}{4}$

$x = -1$, $\lambda = \dfrac{-2 - 3}{2(-1-1)} = \dfrac{\cancel{+}5}{\cancel{+}2 \times 2} = \dfrac{5}{4}$.

$-x^2 + 2x + 3 = 0$.

$-x^2 + 3x - x + 3$

$-x(x-3) - 1(x-3) = 0$

$(x-3)(x+1) = 0$

$\Rightarrow x = 3$ or

$\Rightarrow x = -1$.

By solving the equations together, we have

$$\left.\begin{array}{l}(3, 3/4) \\ (-1, 5/4)\end{array}\right\} \text{Candidate points.}$$

$$h(x,\lambda) = x^2 - 3x - 3 - \lambda_1 x^2 + 2\lambda_1 x + 3\lambda_1$$

$$h(3, 3/4) = \underbrace{3^2 - 3\times 3 - 3} - \underbrace{\frac{3}{4}(9 - 6 - 3)}_{0.}$$

$$= 9 - 9 - 3$$

$$= \underline{\underline{-3}} \leftarrow \text{minimum value.}$$

$$h(-1, 5/4) = \underbrace{1 + \cancel{3} - \cancel{3}} - \underbrace{\frac{5}{4}(1 + 2 - 3)}_{0}$$

$$= \underline{\underline{1}}$$

Hence, using Lagrange multipliers, our minimum point is

$$\underline{\underline{(3, -3)}}$$

Now, we use Lagrangian multipliers to make this problem easier!!

$$\min_{\bar{w}, w_0} \; - \sum_{i=1}^{n} (\bar{w}^T \bar{x}_i + w_0) \cdot y_i$$

subject to $\underbrace{\|\bar{w}\| - 1}_{} = 0$

$$\underbrace{\min_{\bar{w}, w_0} \; - \sum_{i=1}^{n} (\bar{w}^T \bar{x}_i + w_0) y_i + \lambda (\|\bar{w}\| - 1)}_{}$$

New problem to solve using G.D.!!

$$\min_{\bar{w}, w_0} \; \underbrace{-\sum_{i=1}^{n} (\bar{w}^T \bar{x}_i + w_0) y_i + \lambda (\|\bar{w}\| - 1)}_{h(\bar{w}, w_0, \lambda)}$$

We now apply Gradient Descent to solve the above problem.

Update rules:

(i) $$\bar{w}^{(t+1)} = \bar{w}^{(t)} - \eta \cdot \underbrace{\nabla_{\bar{w}} h(\bar{w}, w_0, \lambda)}_{\text{We will now compute this.}}$$

$$-\sum_{i=1}^{n} \Big[(\bar{w}^T x_i) y_i + \underbrace{w_0 y_i}_{0} \Big] + \lambda \Big[\sqrt{\bar{w}^T \bar{w}} - \underbrace{1}_{0} \Big].$$

$$y_i \nabla_{\bar{w}} \left[(\bar{w}^T \bar{x}_i) \right] = y_i \bar{x}_i$$

$$\nabla_{\bar{w}} \left[\lambda \sqrt{\bar{w}^T \bar{w}} \right]$$

$$= \lambda \nabla_{\bar{w}} \left[\sqrt{\bar{w}^T \bar{w}} \right]$$

$$f(x) = \sqrt{x}$$

$$h(x) = f(\bar{w}^T \bar{w}).$$

$$= \lambda \frac{1}{2 \cdot \sqrt{\bar{w}^T \bar{w}}} \times \nabla_{\bar{w}} \left[\bar{w}^T \bar{w} \right]$$

$$= \lambda \times \frac{1}{\cancel{2} \sqrt{\bar{w}^T \bar{w}}} \times \cancel{2} \cdot \bar{w}$$

$$= \lambda \cdot \frac{\bar{w}}{\|\bar{w}\|}$$

Putting it all together:

$$\nabla_{\bar{w}}\, h(\bar{w}, w_0, \lambda) = \underbrace{-\sum_{i=1}^{n} y_i \cdot \bar{x}_i + \lambda \frac{\bar{w}}{\|\bar{w}\|}}$$

$$\text{(ii)} \quad w_0^{(t+1)} = w_0^{(t)} - \eta \cdot \underbrace{\frac{\partial}{\partial w_0} h(\bar{w}, w_0, \lambda)}_{\text{Let's compute this}}.$$

$$-\sum_{i=1}^{n} \Big[\underbrace{(\bar{w}^T x_i) y_i}_{0} + \underline{w_0 y_i} \Big] + \underbrace{\lambda \left[\sqrt{\bar{w}^T \bar{w}} - 1 \right]}_{0}$$

$$\sum_{i=1}^{n} \frac{\partial (-w_0 y_i)}{\partial w_0} = \sum_{i=1}^{n} (-y_i)$$

Summary

* We wanted to build a classifier.
* We wanted to minimize the following:

$$\mathcal{L}(D; \bar{w}, w_0) = -\sum_{i=1}^{n} \left(\frac{\bar{w}^T \bar{x} + w_0}{\|\bar{w}\|} \right) y_i$$

* we converted the above problem to:

$$\min_{\bar{w}, w_0} \; -\sum_{i=1}^{n} (\bar{w}^T \bar{x} + w_0) y_i + \lambda \left(\sqrt{\bar{w}^T \bar{w}} - 1 \right)$$

using Lagrange multipliers.

* We got the Gradient Descent update rules as follows:

G.D. Algo. →

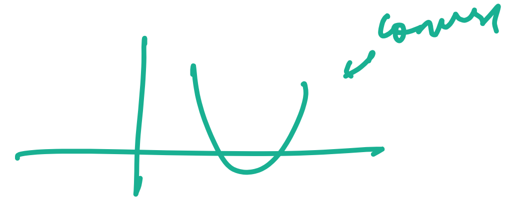
$$\text{(i)} \quad \bar{w}^{(t+1)} = \bar{w}^{(t)} - \eta \left[-\sum_{i=1}^{n} y_i \cdot \bar{x}_i + \lambda \frac{\bar{w}}{\|\bar{w}\|} \right]$$

$$\text{(ii)} \quad w_0^{(t+1)} = w_0^{(t)} - \eta \cdot \left[-\sum_{i=1}^{n} y_i \right]$$

* Convex Functions

(i) "Cup" like functions.

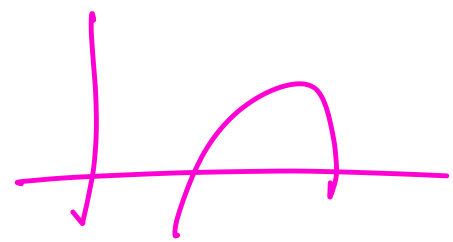
(ii) Have the unique property of single global minima.



* Concave functions :

(i) "Cap" like functions.

(ii) Have single unique maxima.

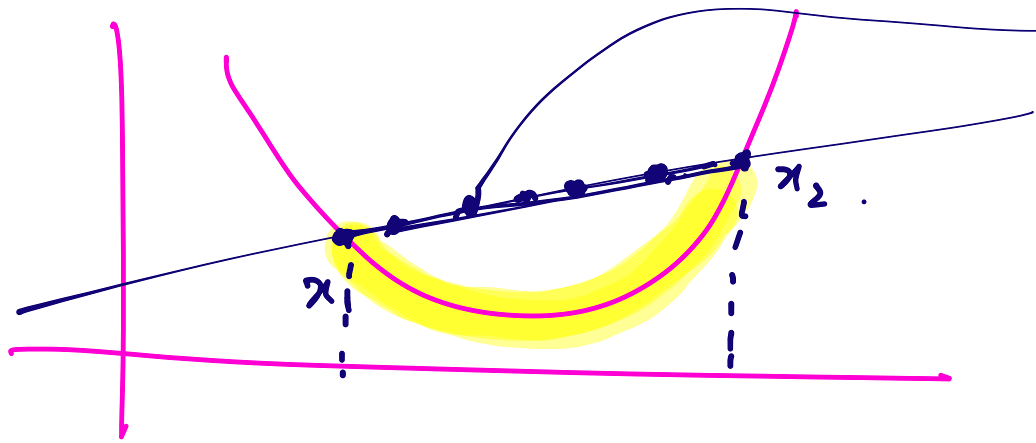


Math definitions.

(i) Convex function.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $\bar{x}_1, \bar{x}_2 \in \mathbb{R}^n$, $\lambda \in [0, 1]$, the following holds:

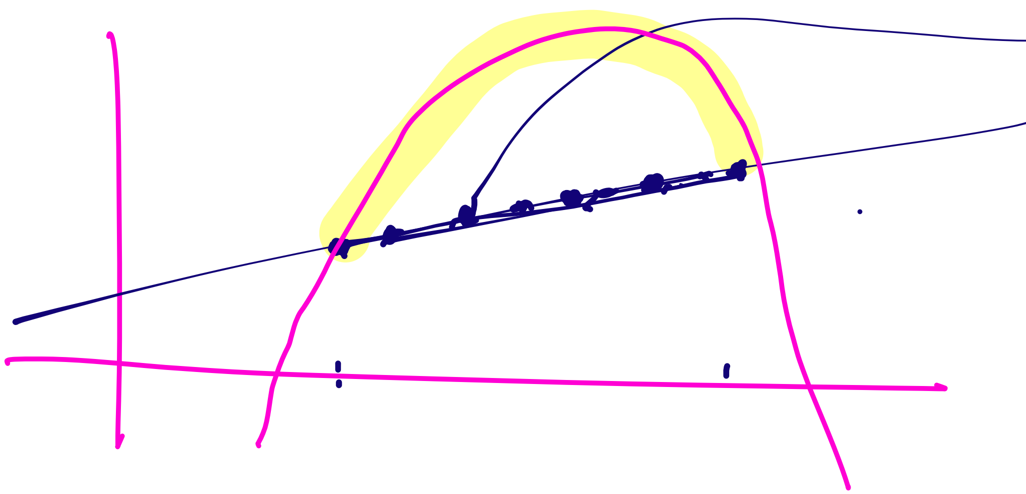
$$f\left(\lambda \bar{x}_1 + (1-\lambda)\bar{x}_2\right) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

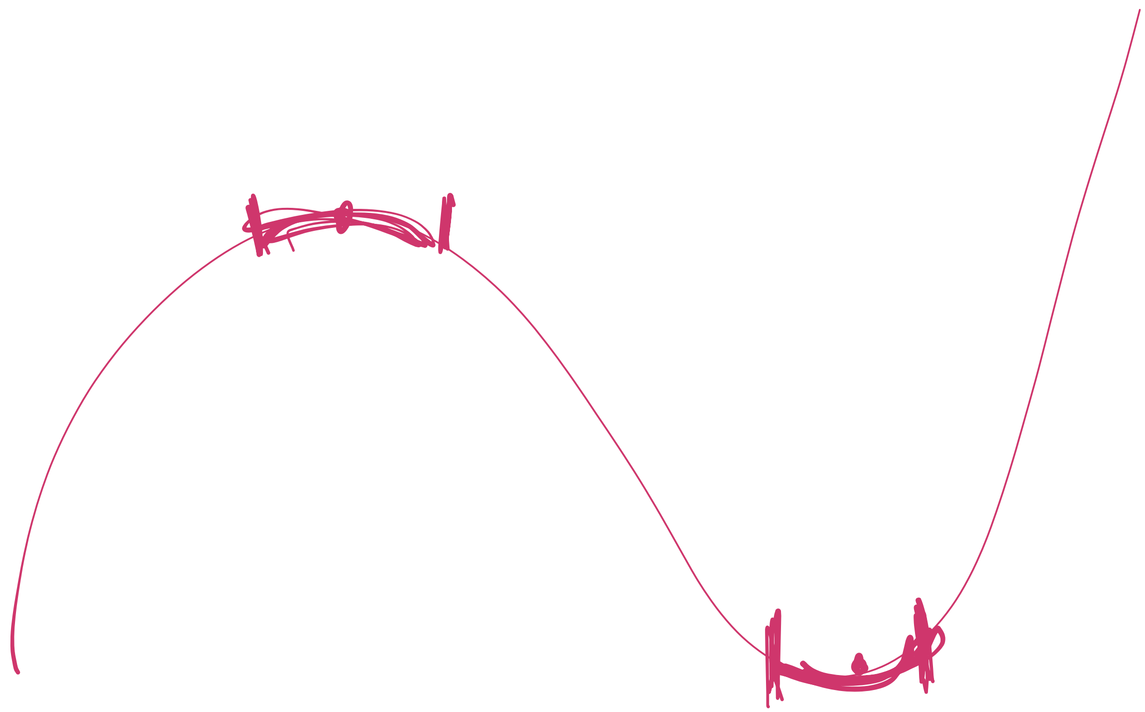


(ii) Concave function.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all $\bar{x}_1, \bar{x}_2 \in \mathbb{R}^n$, $\lambda \in [0, 1]$, the following holds:

$$f(\lambda \bar{x}_1 + (1-\lambda)\bar{x}_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$





$$\min_{x,y,z} \; f(x,y,z) = x^2 + y^2 - z$$

$$\text{s.t.} \quad g(x,y,z) = y + z - 1 = 0.$$

$$h(x,y,z,\lambda) = x^2 + y^2 - z + \lambda\,(y - z - 1)$$

Goal: Find classifier.

$(\bar{w}, \bar{w}_0)$

① Loss(   )

Create a loss function which uses the parameters of the classifier.

② Solve: $\bar{w}^*, w_0^* = \underset{\bar{w}, w_0}{\arg\min}\ \text{Loss}(\ \ )$

$$\bar{w}^T \bar{x} + w_0 = 0.$$

$$\underline{\min_A} \left[f\left(g\left(h\left(x \right) \right) \right) - y_i \right)^2$$