

1 Theoretical Part

1.1 Convex optimization

Based on Lecture 9 and Recitations 2,11

Here we will see a nice property that will help see some property of convexity

- Let $f_1, \dots, f_m : C \rightarrow \mathbb{R}$ be a set of convex functions and $\gamma_1, \dots, \gamma_m \in \mathbb{R}_+$. Prove from definition that $g(\mathbf{u}) = \sum_{i=1}^m \gamma_i f_i(\mathbf{u})$ is a convex function.
- Give a counterexample for the following claim: Given two functions $f, g : \mathbb{R} \rightarrow \mathbb{R}$, define a new function $h : \mathbb{R} \rightarrow \mathbb{R}$ by $h = f \circ g$. If f and g are convex then h is convex as well.

$\forall \mathbf{v}, \mathbf{u} \in C$ $\exists \alpha \in [0, 1]$ $\text{such that } \mathbf{v} = \alpha \mathbf{u} + (1-\alpha) \mathbf{u}$

$$g(\alpha \mathbf{v} + (1-\alpha) \mathbf{u}) \leq \alpha g(\mathbf{v}) + (1-\alpha) g(\mathbf{u}) \quad \alpha \in [0, 1] \quad \text{(to prove)}$$

הוכחה:

אנו צריכים להוכיח $\forall \mathbf{v}, \mathbf{u} \in C \quad g(\alpha \mathbf{v} + (1-\alpha) \mathbf{u}) \leq \alpha g(\mathbf{v}) + (1-\alpha) g(\mathbf{u})$

אנו יוכיח שכל f_i היא פונקציית גודלה במרחב C .

$$\text{dom}(g) = C$$

$$\therefore \alpha \in [0, 1] \quad \mid \quad \mathbf{u}, \mathbf{v} \in C \quad , \quad \mathbf{u}, \mathbf{v}$$

$$g(\alpha \mathbf{v} + (1-\alpha) \mathbf{u}) = \sum_{i=1}^m \delta_i f_i(\alpha \mathbf{v} + (1-\alpha) \mathbf{u}) \leq \sum_{i=1}^m \delta_i (\alpha f_i(\mathbf{v}) + (1-\alpha) f_i(\mathbf{u}))$$

$\forall i \quad \delta_i \leq 1$

$$= \sum_{i=1}^m \alpha \cdot \delta_i f_i(\mathbf{v}) + \sum_{i=1}^m (1-\alpha) \delta_i f_i(\mathbf{u}) = \alpha \sum_{i=1}^m \delta_i f_i(\mathbf{v}) +$$

$$(1-\alpha) \sum_{i=1}^m \delta_i f_i(\mathbf{u}) = \alpha g(\mathbf{v}) + (1-\alpha) g(\mathbf{u})$$

done

$$f, g : \mathbb{R} \rightarrow \mathbb{R}, \quad g(x) = |x|, \quad f(x) = -x \quad \text{ר'ג} \quad (2)$$

ר'ג גדרה יאה, נסן, סעיפים 1.2 ו-1.3 מוכיחים ש f ו- g הן פולינומים.

$$\alpha \in [0,1] \quad \text{ולא} \quad x_1, x_2 \in \mathbb{R} \quad \text{אנו, } f_\alpha$$

$$g(\alpha x_1 + (1-\alpha)x_2) = |\alpha x_1 + (1-\alpha)x_2| \leq |\alpha x_1| + |(1-\alpha)x_2| = \alpha g(x_1) + (1-\alpha)g(x_2)$$

הוכחה שלution

$$f(\alpha x_1 + (1-\alpha)x_2) = -(\alpha x_1 + (1-\alpha)x_2) \leq -\alpha x_1 - (1-\alpha)x_2 = \alpha f(x_1) + (1-\alpha)f(x_2)$$

$$h = f \circ g : \mathbb{R} \rightarrow \mathbb{R} \quad h(x) = -|x|$$

$$\alpha = 0.5, \quad x_2 = -1, \quad x_1 = 1$$

$$h(\alpha x_1 + (1-\alpha)x_2) = h(0.5 \cdot 1 + 0.5 \cdot (-1)) = h(0) = -|0| = 0$$

$$\alpha h(x_1) + (1-\alpha)h(x_2) = 0.5 \cdot h(1) + 0.5 \cdot h(-1) = 0.5 + (-0.5) = -1$$

$$\text{לכן } h \text{ היא פולינום דדרון, } h(-1) < 0 < h(1), \quad 0 > -1$$

$$\text{אך } h(-1) \neq 0, \quad h(1) \neq 0, \quad \text{הוכחה סכירה}$$

$$(a,b) = \text{הו, } b-a = \text{הו, } a-b = \text{הו}$$

$$\text{אנו מוכיחים ש } h(a) = h(b)$$

1.2 Sub-gradients for Soft-SVM Objective

Based on Lecture 9 and Recitations 2,11

The Soft-SVM objective, though convex, is not differentiable in all of its domain due to the use of the hinge-loss. Therefore, to implement a sub-gradient descent solver for this problem we must first describe sub-gradients of the objective.

- Given $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{\pm 1\}$. Show that the hinge loss is convex in \mathbf{w}, b . That is, define

$$f(\mathbf{w}, b) := \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = \max \left(0, 1 - y(\mathbf{x}^\top \mathbf{w} + b) \right)$$

and show that f is convex in \mathbf{w}, b .

- Deduce some sub-gradient of the hinge loss function $g \in \partial \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b)$.
- Let $f_1, \dots, f_m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a set of convex functions and $\mathbf{g}_k \in \partial f_k(\mathbf{x})$ for all $k \in [m]$ be sub-gradients of these functions. Define $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$. Show that $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$.
- Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{\pm 1\}$ be a sample and define $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by:

$$f(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{x}_i, y_i}^{\text{hinge}}(\mathbf{w}, b) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Find a sub-gradient of f for any \mathbf{w} .

$$\mathbf{x}^\top \mathbf{w} + b \quad \text{ט. } \mathbf{w} \quad \text{הinge} \quad \text{הinge} \quad g_1(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b) \quad (3)$$

לפנינו מוגדרת $g_1(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b)$

ולפנינו $g_2(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b)$

$$g_1(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b) \quad \text{לפנינו} \quad \text{הinge} \quad \text{הinge}$$

$$f(\mathbf{w}, b) = \max_{x, y} \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = \max(g_1(\mathbf{w}, b), g_2(\mathbf{w}, b))$$

$$f(\mathbf{w}, b) = \max(g_1(\mathbf{w}, b), g_2(\mathbf{w}, b))$$

לפנינו מוגדרת $g_1(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b)$

$$g_2(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b)$$

לפנינו מוגדרת $g_1(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b)$

$$\text{hinge}(\mathbf{w}, b) = \max \{ g_1(\mathbf{w}, b), g_2(\mathbf{w}, b) \} = \max \{ 1 - y(\mathbf{x}^\top \mathbf{w} + b), 1 - y(\mathbf{x}^\top \mathbf{w} + b) \}$$

≥ 0

$$h(\mathbf{w}, b) = \max \{ g_1(\mathbf{w}, b), g_2(\mathbf{w}, b) \} = f(\mathbf{w}, b)$$

3. Given $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{\pm 1\}$. Show that the hinge loss is convex in \mathbf{w}, b . That is, define

$$f(\mathbf{w}, b) := \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = \max(0, 1 - y(\mathbf{x}^\top \mathbf{w} + b))$$

and show that f is convex in \mathbf{w}, b .

4. Deduce some sub-gradient of the hinge loss function $g \in \partial \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b)$.

רלוונטיים למבחן

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = 0 \iff 1 - y(\mathbf{x}^\top \mathbf{w} + b) \leq 0$$

לפניהם (ב- \mathbb{R}^{d+1})

הנחתה נסובב בפונקציית (\mathbf{w}, b) מינימום של $f(\mathbf{w}, b)$

זה הינו אמצעי-כדרון של פונקציית $f(\mathbf{w}, b)$

$$\partial \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = \nabla f(\mathbf{w}, b) = \{0\}$$

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = 1 - y(\mathbf{x}^\top \mathbf{w} + b) \iff 1 - y(\mathbf{x}^\top \mathbf{w} + b) \geq 0$$

$$\frac{\partial f(\mathbf{w}, b)}{\partial \mathbf{w}} = -y \mathbf{x} \quad \text{בנוסף לכך}$$

$$\frac{\partial f(\mathbf{w}, b)}{\partial b} = -y$$

לפניהם, (y, \mathbf{x}) מינימום של $f(\mathbf{w}, b)$ אם $\{(-y, \mathbf{x})\}$

$$f(\mathbf{w}, b) = \ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) \quad \text{לפניהם לא}$$

$$g = \begin{cases} (0, 0) \\ (-y, -\mathbf{x}) \end{cases}$$

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) = 0$$

$$\ell_{\mathbf{x}, y}^{\text{hinge}}(\mathbf{w}, b) \neq 0$$

זהו (\mathbf{w}, b) ?

5. Let $f_1, \dots, f_m : \mathbb{R}^d \rightarrow \mathbb{R}$ be a set of convex functions and $\mathbf{g}_k \in \partial f_k(\mathbf{x})$ for all $k \in [m]$ be sub-gradients of these functions. Define $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$. Show that $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$.

$\text{dom } f \supseteq \text{dom } f_k \quad \forall k \in [m] \quad \text{and} \quad \mathbf{g}_k \in \partial f_k(\mathbf{x}) \quad \forall k \in [m]$

$$\forall \mathbf{v} \in \text{dom } f_k \quad f_k(\mathbf{v}) \geq f_k(\mathbf{x}) + \langle \mathbf{g}_k, \mathbf{v} - \mathbf{x} \rangle$$

$$\sum_{k=1}^m f_k(\mathbf{v}) \geq \sum_{k=1}^m (f_k(\mathbf{x}) + \langle \mathbf{g}_k, \mathbf{v} - \mathbf{x} \rangle) = \sum_{k=1}^m f_k(\mathbf{x}) + \sum_{k=1}^m \langle \mathbf{g}_k, \mathbf{v} - \mathbf{x} \rangle$$

$$\sum_{k=1}^m f_k(\mathbf{v}) \geq \sum_{k=1}^m (f_k(\mathbf{x}) + \langle \mathbf{g}_k, \mathbf{v} - \mathbf{x} \rangle) = \sum_{k=1}^m f_k(\mathbf{x}) + \sum_{k=1}^m \langle \mathbf{g}_k, \mathbf{v} - \mathbf{x} \rangle$$

$$f(\mathbf{v}) = \sum_{k=1}^m f_k(\mathbf{v}) \quad f : \mathbb{R}^d \rightarrow \mathbb{R} \quad \forall i \in [m] \quad \text{and} \quad f_i \in \text{dom } f_k$$

לפיכך $f(\mathbf{v}) \geq f(\mathbf{x}) + \langle \sum_k \mathbf{g}_k, \mathbf{v} - \mathbf{x} \rangle$ (הנורמה הינה $\|\cdot\|$)

$$f(\mathbf{v}) = \sum_{k=1}^m f_k(\mathbf{v}) \quad \mathbf{v} \in \text{dom } f \Rightarrow f \in \partial f$$

וונר חישובית f_1, \dots, f_n \leftarrow $\mathbf{v} \in \text{dom } f$

$$\partial \left(\sum_{k=1}^n f_k \right) = \sum_{k=1}^n \partial f_k$$

וונר חישובית $f_1, f_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2$$

$$\partial \left(\sum_{k=1}^{n+1} f_k \right) = \partial \left(\sum_{k=1}^n f_k + f_{n+1} \right) = \sum_{k=1}^n \partial f_k + \partial f_{n+1}$$

וונר חישובית $\sum_{k=1}^n f_k$

הדרוגים נספחים בפערם בפערם

$$= \sum_{k=1}^m \partial f_{i_k} + \partial f_{m+1} = \sum_{k=1}^{m+1} \partial f_{i_k}$$

הנגינה הדרוגים נספחים בפערם בפערם

$$\sum_{k=1}^m g_k \in \partial f = \sum_{k=1}^m \partial f_{i_k} \leftarrow g_k \in \partial f_{i_k}(x) \text{ ו } \forall i$$

. מגדיר.

6. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^d \times \{\pm 1\}$ be a sample and define $f : \mathbb{R}^d \rightarrow \mathbb{R}$ by:

$$f(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{x}_i, y_i}^{hinge}(\mathbf{w}, b) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Find a sub-gradient of f for any \mathbf{w} .

$$\text{הנגינה } f : \mathbb{R} \rightarrow \mathbb{R} \quad f(x) = x^2 \quad \text{הדרוגים נספחים בפערם בפערם} \quad \|\cdot\|_2^2$$

$$\text{הנגינה } f' : \mathbb{R} \rightarrow \mathbb{R}^+ \quad f'(x) = 2x \quad \|\cdot\|_2 : \mathbb{R} \rightarrow \mathbb{R}^+$$

$$\text{הנגינה } f'(\|\cdot\|_2) \text{ היא נון-נשענית (0,0)}$$

$$\text{הנגינה } f'(\|\cdot\|_2) \text{ היא נון-נשענית (0,0)} \quad \|\cdot\|_2 \text{ הוא } f \text{ הע}$$

$$\text{הנגינה } f'(\|\cdot\|_2) \quad \|\cdot\|_2$$

$$\text{הנגינה } f'(\|\cdot\|_2) \text{ היא נון-נשענית (0,0)}$$

$$\Rightarrow f'_i \text{ הוא } f'(\|\cdot\|_2) \text{ עבור } i \in \{1, \dots, m\}$$

$$g_i = \begin{cases} (0,0) & f'_{i,i}(\mathbf{w}, b) = 0 \\ -y_i x_i, -y_i & f'_{i,i}(\mathbf{w}, b) \neq 0 \end{cases}$$

$$(1.872) \rightarrow \text{הנורמליזציה} \rightarrow 1.872 \text{ הינה } h(w) = \|w\|^2$$

$$\nabla \left(\frac{\lambda}{2} h(w) \right) = \lambda w \quad \leftarrow \nabla h(w) = \lambda w$$

$$\partial f(w, b) = \partial \left(\frac{1}{m} \sum_{i=1}^m g_{\text{ hinge}}(w, b) + \frac{\lambda}{2} \|w\|^2 \right) = \frac{1}{m} \sum_{i=1}^m \partial(g_{\text{ hinge}}(w, b)) +$$

+ \frac{\lambda}{2} \partial(\|w\|^2)

+ \frac{\lambda}{2} \partial(b)

+ \frac{\lambda}{2} \partial(w)

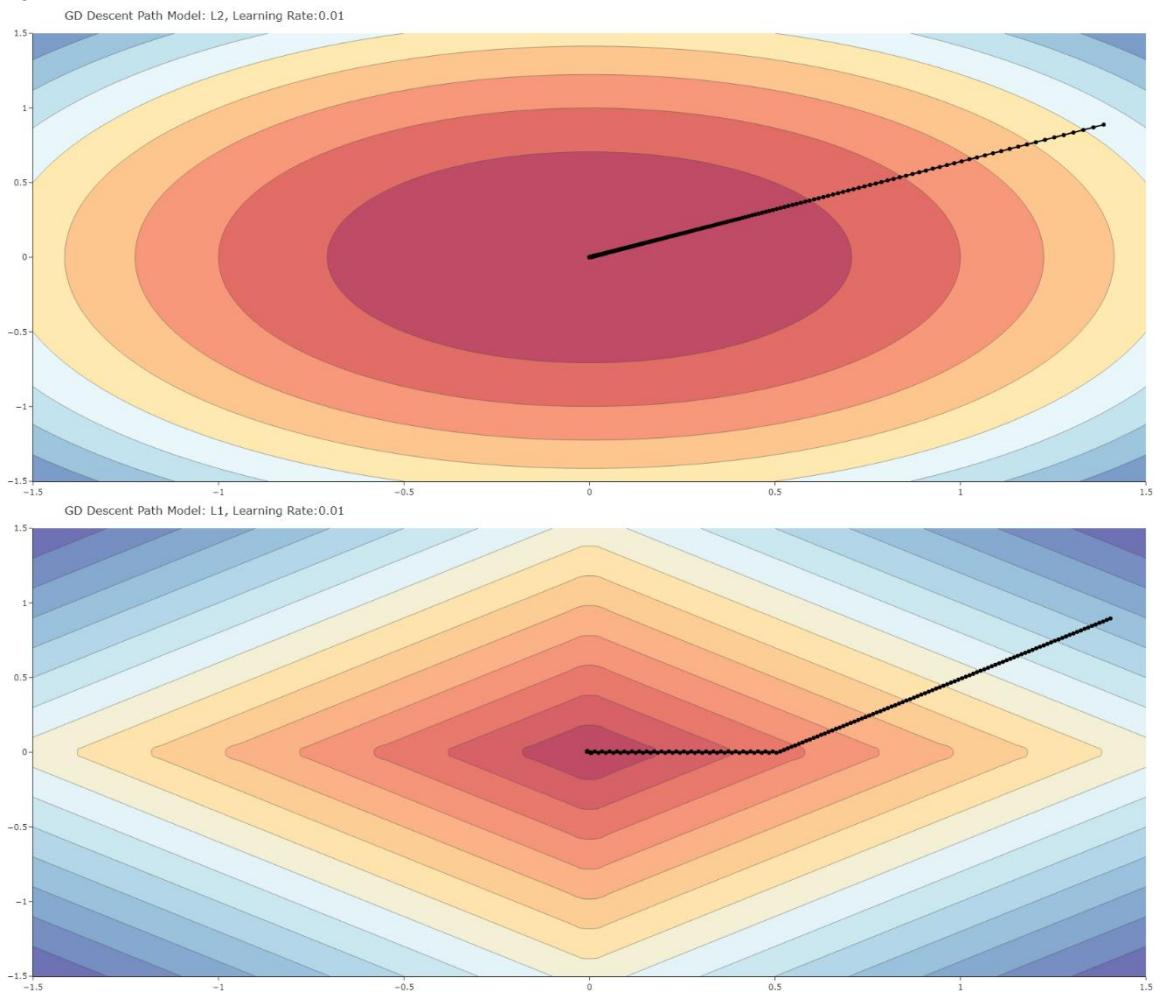
$$+ \partial \left(\frac{\lambda}{2} h(w) \right) = \frac{1}{m} \sum g_i + (\lambda w, 0)$$

הזר חואם בפער

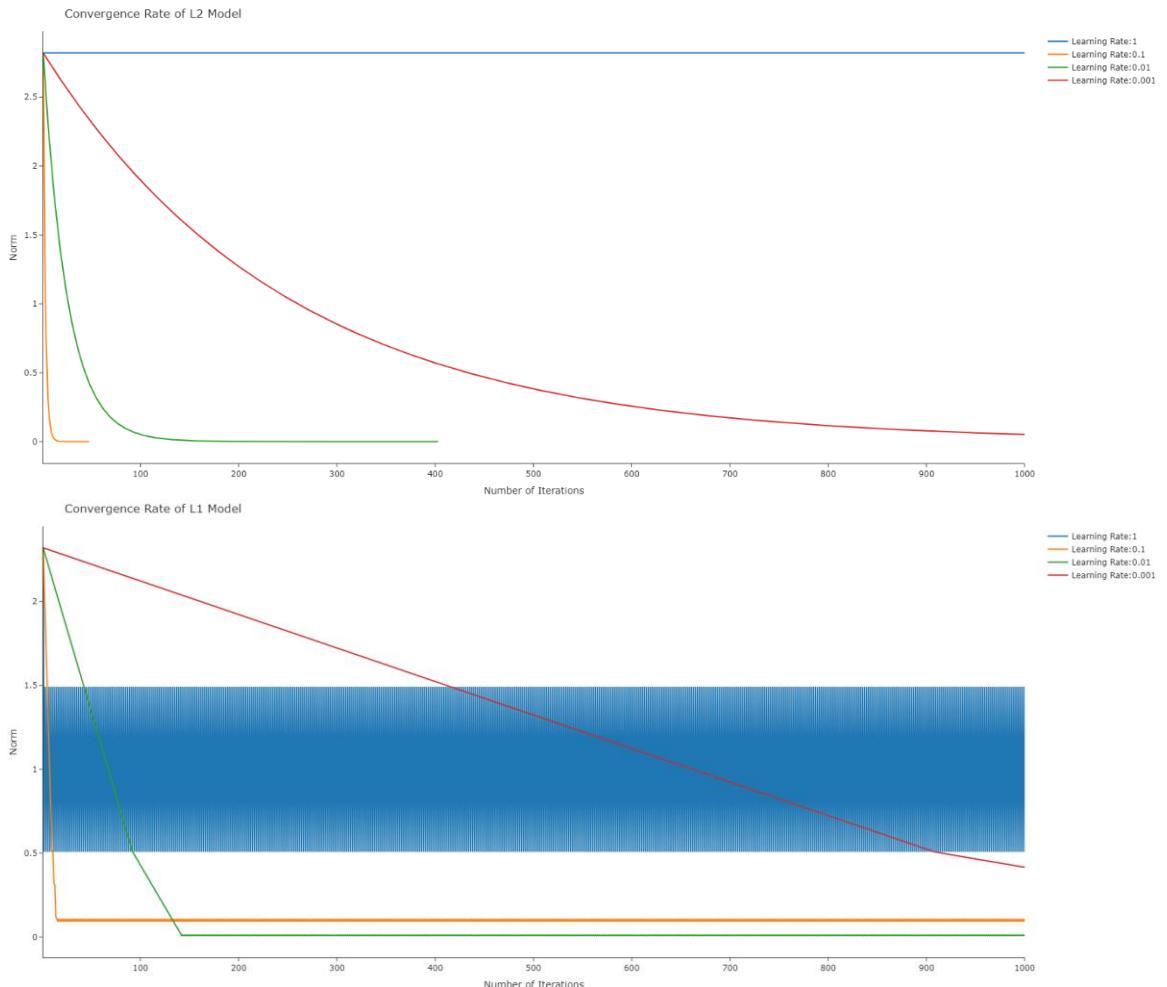
exercise 4 - IML , ron keter 316417427

practical part

1. Q1:



2. we can observe that the main differences between L1 and L2 is that the descent path for L2 is smooth and since this regularization penalize all coefficient equally. also we can see that L1 reaches around norm of 0.5 a "zig-zag" pattern indicating that the algorithm did not converge (from that point each iteration jumping back and forth between two values)
3. we can see that for learning rate 1 , the algorithm doesnt converge (norm remains high and unchanged). for the rest of the learnign rate the norma decreases as a function of number of iteration and stablizes indicating the algorithm convergence. we can also see that learning rate 0.1 the most effective out of the group.

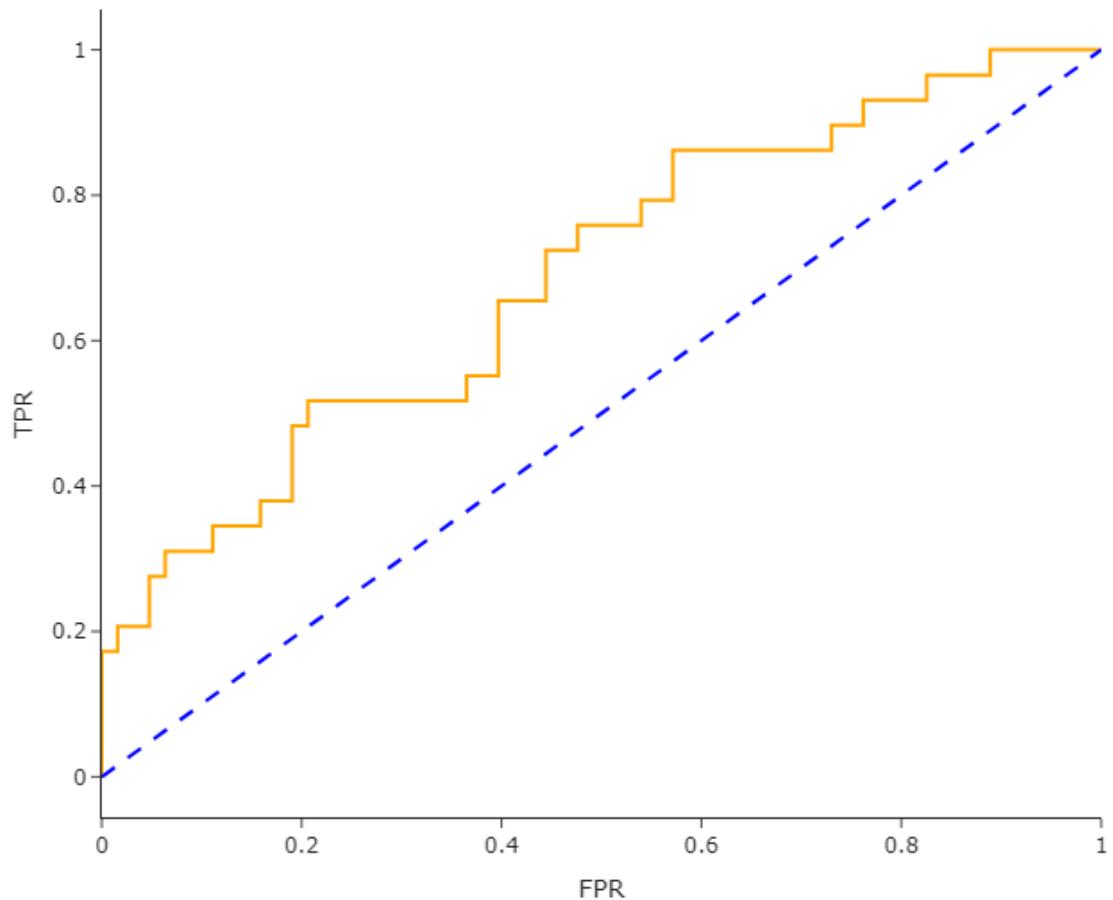


4. The L2 model achieves a significantly lower loss compared to the L1 , the reason for the difference is how their gradients behave. for L1 the gradient is not dependent on any changing paramters. and for L2 gradient it is dependent on X (2 times current X) so as X get smaller the gradient get smaller, therefore the norm. as both get smaller the algorithm could achieve more accurate results in each iteration and achieve much better convergence than L1.

```
Lowest loss achieved for L1 model: 0.008119619553413011 with learning rate: 0.01
Lowest loss achieved for L2 model: 2.19211242161629e-09 with learning rate: 0.1
```

5.

ROC Curve - logistic regression (AUC : 0.689)



6.

```
Optimal alpha: 0.47852547869060685, with test error of : 0.29347826086956524
```

7.

```
Best lambda: 0.002, with test error of : 0.25
```