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

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

2) נבחר $f(x)=-x$, $g(x)=|x|$, $f,g:\mathbb{R}\to\mathbb{R}$

$\mathbb{R}$ קבוצה קמורה, מסגירות המספרים הממשיים לחיבור ומכפלה.

כעת, לכל $x_1,x_2\in\mathbb{R}$ ולכל $\alpha\in[0,1]$

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

אי שוויון המשולש

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

$h(x)=-|x|$ $h=f\circ g:\mathbb{R}\to\mathbb{R}$ אינה קמורה.

נבחר $x_1=1$, $x_2=-1$, $\alpha=0.5$.

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

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

$0>-1$, h אינה קמורה. הטענה בשאלה 1 כי דרוש

לינאריות של פונקציות הקמורות, וכן g קמורה.

$g_2(a,b)=0$, פונקציה קמורה (אינה תלויה ב a,b)

וכן קמורה לכל בחירת a,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.

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

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

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

4. Deduce some sub-gradient of the hinge loss function $g \in \partial \ell^{hinge}_{\mathbf{x},y}(\mathbf{w}, b)$.
5. Let $f_1, \ldots, f_m : \mathbb{R}^d \to \mathbb{R}$ 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 $f : \mathbb{R}^d \to \mathbb{R}$ by $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$. Show that $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$.

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

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

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

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

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

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

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

5. Let $f_1, \ldots, f_m : \mathbb{R}^d \to \mathbb{R}$ 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 $f : \mathbb{R}^d \to \mathbb{R}$ by $f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x})$. Show that $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$.

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

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

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

$$\nabla_{w,b} f(w,b) = \nabla\left(\frac{1}{m}\sum_{i=1}^{m} \ell^{hinge}_{x^{(i)},y^{(i)}}(w,b)\right) + \nabla\left(\frac{1}{2}\|w\|^2\right) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{w,b}\,\ell^{hinge}_{x^{(i)},y^{(i)}}(w,b) + (w,0)$$

$$\nabla_{w,b}\,\ell^{hinge}_{x^{(i)},y^{(i)}}(w,b) = \begin{cases} 0 & y^{(i)}\langle w, x^{(i)}\rangle \geq 1 \\ -y^{(i)}x^{(i)} & \text{otherwise} \end{cases}$$

$$\nabla_{w,b} f(w,b) = (w,0) + \frac{1}{m}\sum_{i=1}^{m} \nabla_{w,b}\,\ell^{hinge}_{x^{(i)},y^{(i)}}(w,b)$$

$$= \begin{cases} (w,0) & y^{(i)}\langle w, x^{(i)}\rangle \geq 1 \\ (w,0) - \frac{1}{m}\left(y^{(i)}x^{(i)}, y^{(i)}\right) & \text{otherwise} \end{cases}$$

$$= \begin{cases} (w,0) & y^{(i)}\langle w, x^{(i)}\rangle \geq 1 \\ (w,0) - \left(y^{(i)}x^{(i)}, y^{(i)}\right)\eta & \text{otherwise} \end{cases}$$

$$\eta = \frac{1}{m}$$

$$\eta = \frac{1}{\lambda m}$$

$$\eta = \frac{1}{\|w\|^2}$$

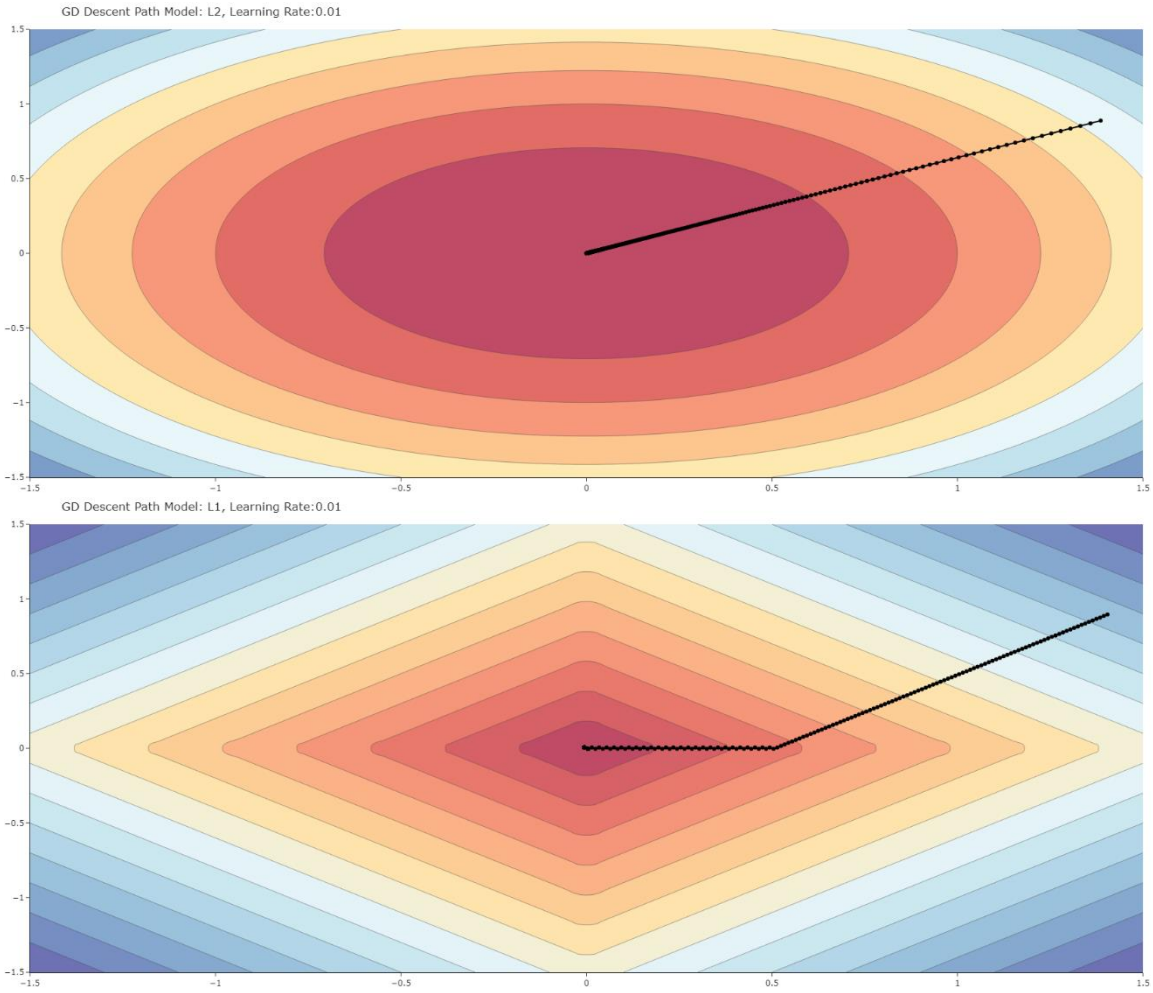
$$\eta = \frac{1}{\lambda\|w\|^2}$$

$$\eta = \frac{1}{\lambda m\|w\|^2}$$

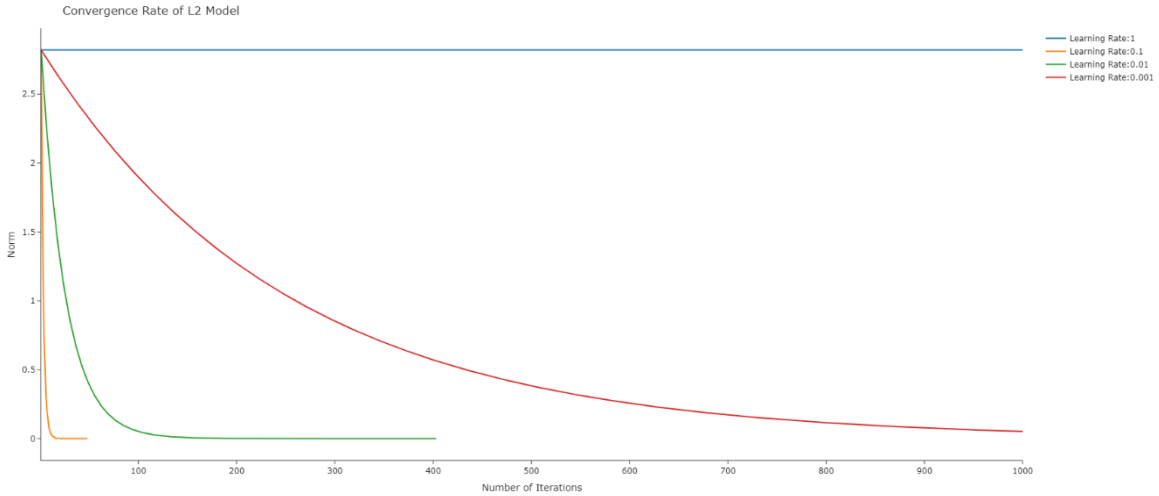
exercise 4 - IML , ron keter 316417427

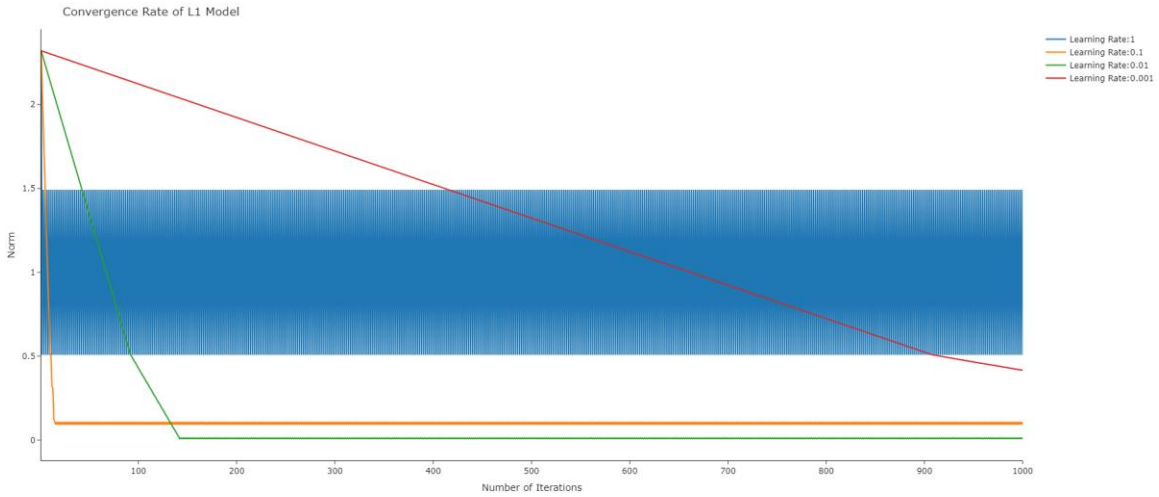
practical part

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 agorithm 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.

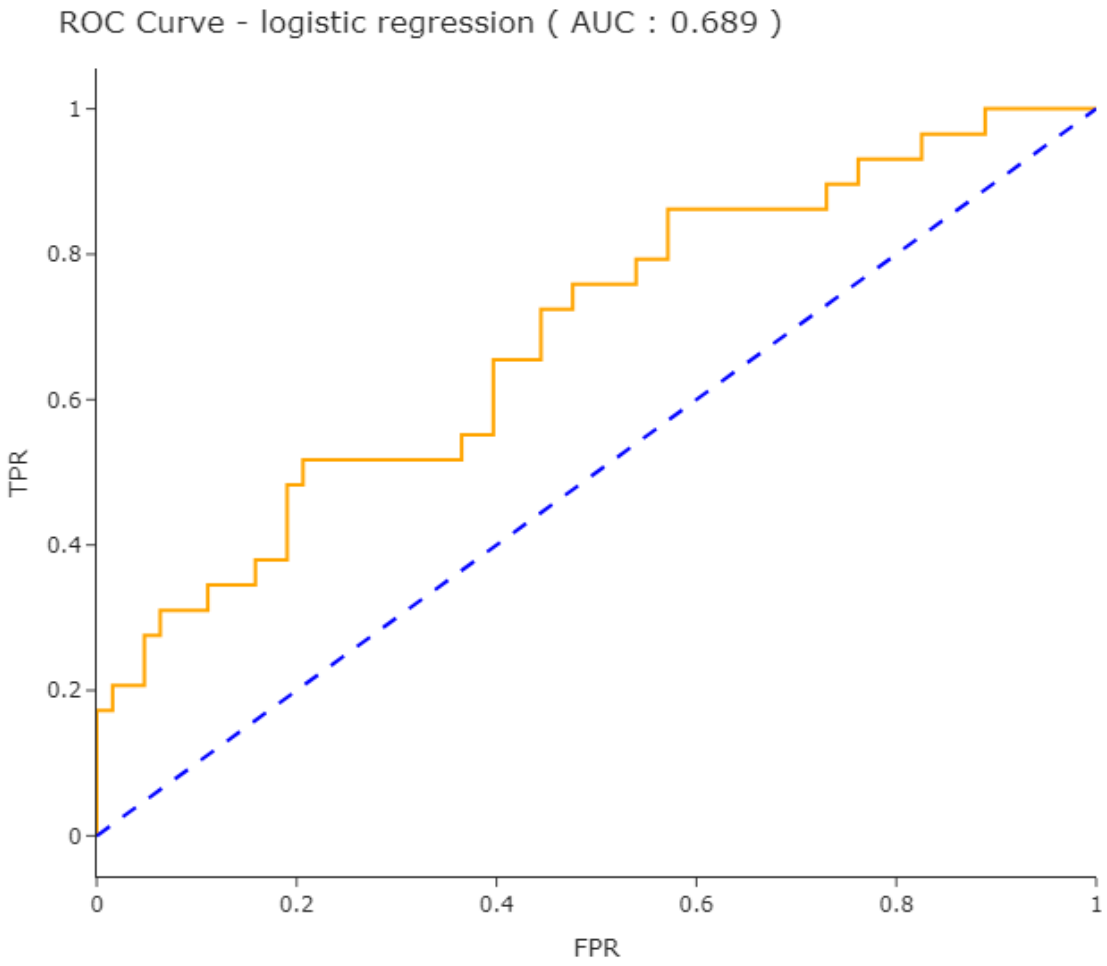




4. The L2 model achieves a significantly lower loss compared to the L1 , the reason for the difference is how their gradients behave.for L1 the gradient is not dependent on any changing paramters. and for L2 gradient it is dependent on X (2 times current X) so as X get smaller the gradient get smalller, 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.



6.

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

7.

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