

## 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.

הוכחה הינה ש  $f$  ו- $g$  קווינט.  $g(u)$  ו- $f(g(u))$  קווינט.

זהו דבר על טוונקיות קמורות.

הטענה. טוונק. קמורה  $f : C \rightarrow \mathbb{R}$  מתקיימת  $\forall v, u \in C, \forall \alpha \in [0, 1]$

$$f(\alpha v + (1-\alpha)u) \leq \alpha f(v) + (1-\alpha)f(u)$$

הוכחה:  $v, w \in C$   $\alpha \in [0, 1]$

$$f(\alpha v + (1-\alpha)w) = \sum_{i=1}^m \gamma_i f_i(\alpha v + (1-\alpha)w) \stackrel{?}{\leq} \sum_{i=1}^m \gamma_i (\alpha f_i(v) + (1-\alpha)f_i(w))$$

$$\stackrel{2}{=} \alpha \sum_{i=1}^m \gamma_i f_i(v) + (1-\alpha) \sum_{i=1}^m \gamma_i f_i(w)$$

$$= \alpha f(v) + (1-\alpha) f(w)$$

הוכחה  $f$  קוינט, אז  $f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$

הוכחה:  $f$  קוינט  $\Leftrightarrow$   $\forall v, w \in C, \forall \alpha \in [0, 1] f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$

הוכחה:

ריבועון וגההה.

convex  $\Leftrightarrow$   $\forall v, w \in C, \forall \alpha \in [0, 1] f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$  (2)

$f(x) = x^2$  קוינט  $\Leftrightarrow$   $\forall v, w \in C, \forall \alpha \in [0, 1] f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$

convex  $\Leftrightarrow$   $f(-x) = -x^2$  קוינט  $\Leftrightarrow$   $\forall v, w \in C, \forall \alpha \in [0, 1] f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$

$f(x) = e^x$  קוינט  $\Leftrightarrow$   $\forall v, w \in C, \forall \alpha \in [0, 1] f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$

$f(x) = -e^{-x}$  קוינט  $\Leftrightarrow$   $\forall v, w \in C, \forall \alpha \in [0, 1] f(\alpha v + (1-\alpha)w) \leq \alpha f(v) + (1-\alpha) f(w)$

ל-ב-ג כונקס קס' ר-ה-ה-ה

## 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

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

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

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

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  $\sum_k \mathbf{g}_k \in \partial \sum_k f_k(\mathbf{x})$ .

ל-ב-ג כונקס קס' ר-ה-ה-ה  
 $g(x) = \mathbf{w}^\top \mathbf{x} + b$   
 $f(\mathbf{w}, b) = \max(0, 1 - y g(\mathbf{x}))$

ל-ב-ג כונקס קס' ר-ה-ה-ה  
 $\max_{\mathbf{w}, b} \min_{\mathbf{x}} \max_{y \in \{-1, 1\}} (1 - y) g(\mathbf{x})$

ל-ב-ג כונקס קס' ר-ה-ה-ה  
 $f(\mathbf{w}, b)$

convex כונקס  $f(\mathbf{w}, b) \geq g(\mathbf{x})$

כונקס כונקס כונקס

$$g_i = \begin{cases} (0) & \text{if } \ell_{\mathbf{x}, y_i}(\mathbf{w}, b) = 0 \\ (-y_i) & \text{else} \end{cases}$$

ל-ב-ג כונקס קס' ר-ה-ה-ה  
 $f(\mathbf{x}) = \sum_{i=1}^m f_i(\mathbf{x})$ ,  $\|f\|$

$\mathbf{g}_k \in \partial f_k(\mathbf{x})$   
 $f$   $\mathbf{g}_k$  gradient  $\mathbf{x}$

$$f_i(w) \triangleq f_i(x) + \langle g_i, w - x \rangle \quad \forall w \in \mathbb{R}^d$$

$$\Rightarrow f(w) = \sum_{i=1}^m f_i(w) \geq \sum_{i=1}^m (f_i(x) + \langle g_i, w - x \rangle)$$

$$= \sum_{i=1}^m f_i(x) + \sum_{i=1}^m \langle g_i, w - x \rangle$$

$$= \sum_{i=1}^m f_i(x) + \left\langle \sum_{i=1}^m g_i, w - x \right\rangle$$

$$\Rightarrow f(w) \geq f(x) + \left\langle \sum_{i=1}^m g_i, w - x \right\rangle$$

לנארו גיבובן גסן כ'ז  $\sum_{i=1}^m g_i$  | יג

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:

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

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

הנראה ש  $\nabla f(\mathbf{w}, b) \leq \mathbf{0}$

כ'ז  $\ell_{x_i, y_i}(\mathbf{w}, b)$  דק כירזת גנטה ופונקציית  $(-6)$

$$g_i = \begin{cases} (0) & \text{if } \ell_{x_i, y_i}(\mathbf{w}, b) = 0 \\ (-y, -1) & \text{else} \end{cases}$$

הנראה ש  $\nabla f(\mathbf{w}, b) \leq \mathbf{0}$

$$\nabla f(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \ell_{x_i, y_i}(\mathbf{w}, b) + \lambda \mathbf{w}$$

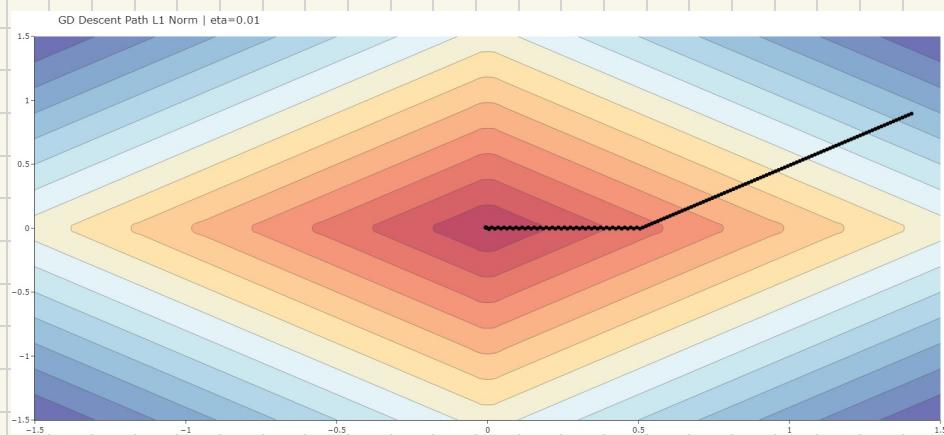
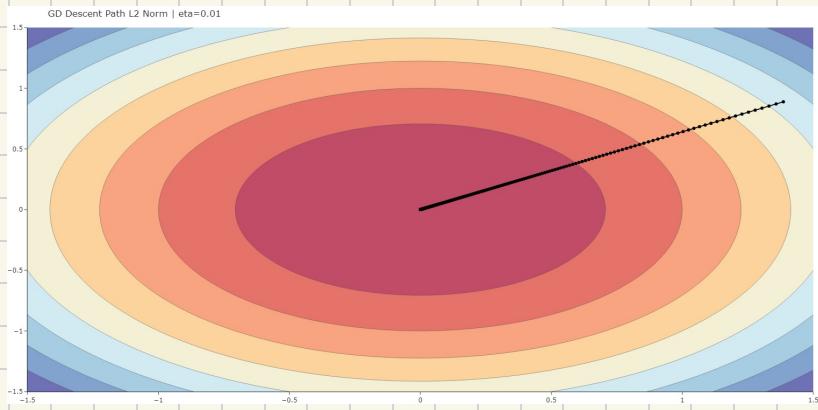
## 2 Practice Pan

Write the necessary code in the files specified in the questions.

• 1

new, answer the following questions.

- Plot the descent path for each of the settings described above (you can use the `plot_descent_path`). Add below the plots for  $\eta = 0.01$  and explain the differences seen between the L1 and L2 modules.



הנובט מינימיזציה בפונקציית ל2 מינימום גראדיאנט שמשתמש באלגוריתם ה- $\text{GD}$  ו- $\text{SGD}$ .  
בפונקציית ל1 מינימום גראדיאנט שמשתמש באלגוריתם ה- $\text{GD}$  ו- $\text{SGD}$ .  
הבדן בין התוצאות של מינימיזציה בפונקציית ל2 מינימום גראדיאנט שמשתמש באלגוריתם ה- $\text{GD}$  ו- $\text{SGD}$  לבין התוצאות של מינימיזציה בפונקציית ל1 מינימום גראדיאנט שמשתמש באלגוריתם ה- $\text{GD}$  ו- $\text{SGD}$ .

modules.

2. following the previous question describe two phenomena that you have seen in the descent path of the  $\ell_1$  objective when using GD and a fixed learning rate.

לפי ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

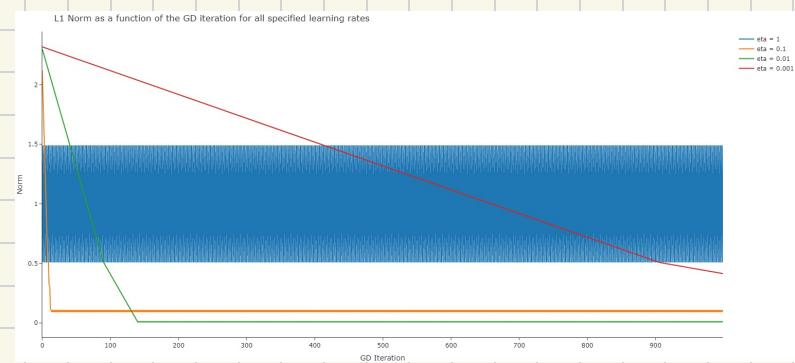
ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

לפי ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

3. For each of the modules, plot the convergence rate (i.e. the norm as a function of the GD iteration) for all specified learning rates. Explain your results

4. What is the lowest loss achieved when minimizing each of the modules? Explain the differences



לפי ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

תפקידו

כגון פותח נרמז שולחן סטטיסטיק.

הנורמליזציה של ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

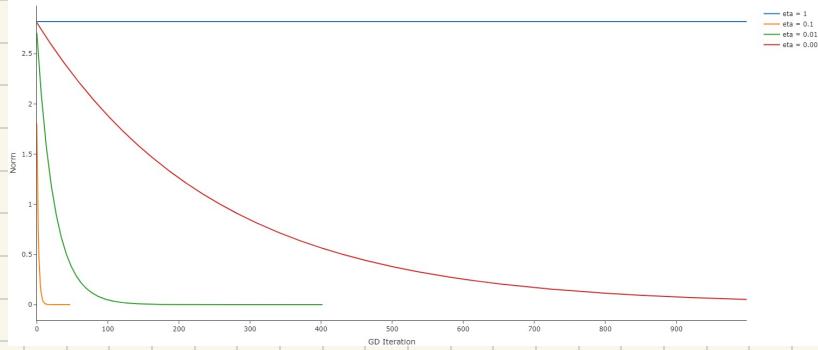
תפקידו

פוקוס פותח נרמז שולחן סטטיסטיק.

הנורמליזציה של ה  $\ell_1$  פותח נרמז שולחן סטטיסטיק.

כגון פותח נרמז שולחן סטטיסטיק.

פוקוס פותח נרמז שולחן סטטיסטיק.



לעומת זה, מבחן הראה שפונקציית האפסון מושפעת מהתא המבוקש. מבחן הראה שפונקציית האפסון מושפעת מהתא המבוקש.

המבחן מינימיזה פונקציית האפסון.

בנוסף, מבחן הראה שפונקציית האפסון מושפעת מהתא המבוקש.

מבחן הראה שפונקציית האפסון מושפעת מהתא המבוקש.

4. What is the lowest loss achieved when minimizing each of the modules? Explain the differences

eta: 1

L2 Norm, lowest error: [2.82100623]

L1 Norm, lowest error: [0.50811962]

eta: 0.1

L2 Norm, lowest error: [1.40295195e-09]

L1 Norm, lowest error: [0.09188038]

eta: 0.01

L2 Norm, lowest error: [2.39122837e-07]

L1 Norm, lowest error: [0.00811962]

eta: 0.001

L2 Norm, lowest error: [0.051462]

L1 Norm, lowest error: [0.41430751]

eta=0.01

תפקיד

הו גורם

$\ell_1$

הו גורם

$\ell_2$

הו גורם

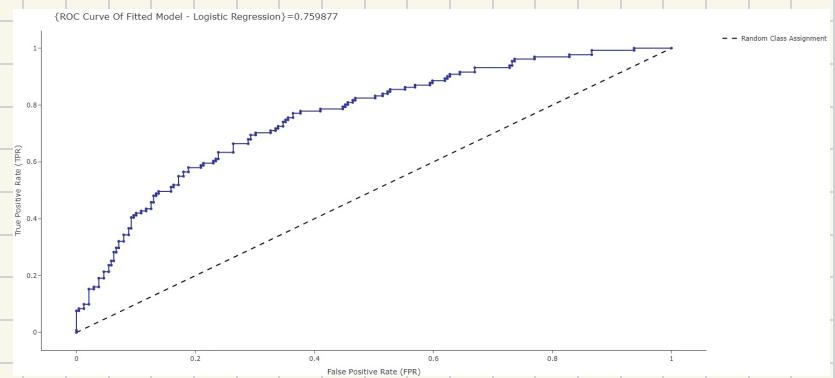
הו גורם

ככל שטפסת ב- $\ell_2$ , יתגלה שפונקציית האפסון מושפעת מהתא המבוקש.

ב- $\ell_1$  מבחן הראה שפונקציית האפסון מושפעת מהתא המבוקש.

ככל שטפסת ב- $\ell_1$ , מבחן הראה שפונקציית האפסון מושפעת מהתא המבוקש.

5. Using your implementation, fit a logistic regression model over the data. Use the `predict_proba` to plot an ROC curve. You can use `sklearn's metrics.roc_curve` function and the code provided in Lab 04.



6. Which value of  $\alpha$  achieves the optimal ROC value according to the criterion below. Using this value of  $\alpha^*$  what is the model's test error?

$$\alpha^* = \operatorname{argmax}_\alpha \{\text{TPR}_\alpha - \text{FPR}_\alpha\}$$

$\alpha^* = 0.32$  (using logit loss) leads to a test error of ~0.336

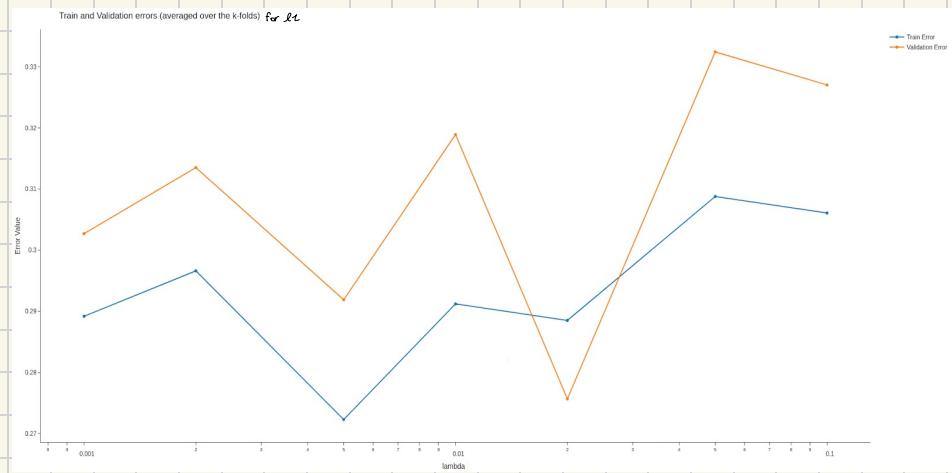
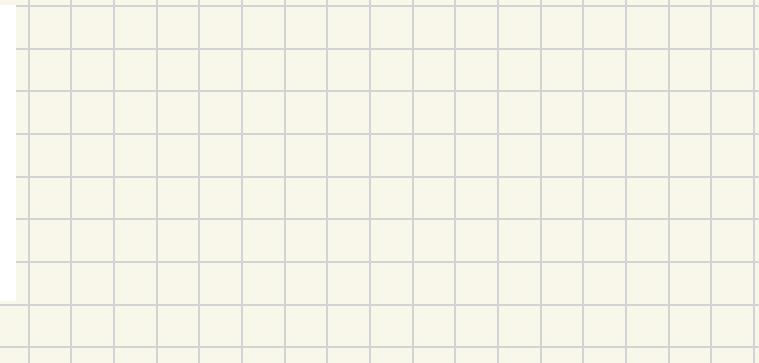
7. Fit an  $\ell_1$ -regularized logistic regression by passing `penalty="l1"` when instantiating a logistic regression estimator

- Set  $\alpha = 0.5$
- Use your previously implemented cross-validation procedure to choose  $\lambda$
- After selecting  $\lambda$  repeat fitting with the chosen  $\lambda$  and  $\alpha = 0.5$  over the entire train portion.

What value of  $\lambda$  was selected and what is the model's test error?

When fitting the model you can (but don't have to) set the parameters as follows:

- Use `max_iter=20,000` and `l1=1e-4`.
- When searching for the optimal  $\lambda$ :
  - Search the following values  $\lambda \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1\}$ .
  - Use  $\alpha = 0.5$  as the cutoff.



Using cross-validation → ensures that the model generalizes well across different folds of the data

module L1 | lambda chosen: 0.02  
model test error: 0.28