תרגיל 4 - חלק תיאורטי

1.1

① יהי $a \in [0,1]$

$$a\,g(u) + (1-a)\,g(v) \underset{\text{מהגדרת } g}{=} a\sum_{i=1}^{m} r_i f_i(u) + (1-a)\sum_{i=1}^{m} r_i f_i(v)$$

$$= \sum_{i=1}^{m}\left[a\, r_i f_i(u) + (1-a)\, r_i f_i(v)\right]$$

$$\geq \sum_{i=1}^{m} r_i f_i(au + (1-a)v) \underset{\text{מהגדרת } g}{=} g(au + (1-a)v)$$

לכל $i \in [m]$ מתקיים מהגדרתן $a f_i(u) + (1-a) f(v) \geq f(au + (1-a))$

ולכן מתקיים $r_i > 0$, ולכן

$$a r_i f_i(u) + (1-a) r_i f(v) = r_i\left(a f_i(u) + (1-a) f_i(v)\right) \geq r_i f_i(au + (1-a)v)$$

② נקח לדוג- $f(x) = -x,\ g(x) = x^2$

ואז $f \circ g(x) = -x^2$ אינה המורה (אלא הקעורה)

דוג נוספת- $f = g = e^{-x}$, ואז $e^{-e^{-x}}$ אינה המורה (אלא הקעורה)

1.2

③ הפונ שאנו מגדירים הינ המקסימום בין שתי פונקציות אפינאריות ב-w, בפרט המורות

וכפי שראינו בכיתה (שקפים 4ע הגרדיאנט g של ה"ר בבריאל), המקסימום של פונ המורות

הינ פונ המורה בעצמה, בפרט f שהגדרנו

④ בתרגול שהבאתי נכון שגרדיאנט הינ אך ורק להגדיר פונ sub-gradient לכל x, y

$$g = \begin{cases} 0 & \text{אם } x,y \in \mathcal{L}_{x,y}(a,b) \\ (-yx, -y) & \text{אחרת} \end{cases}$$

⑤ יהי $y \in \mathbb{R}^d$ לכל $k \in [m]$ מתקיים מהגדרת sub-gradient ($g_k \in \partial f_k(x)$) ש- $f_k(y) \geq f_k(x) + \langle g_k(x), y-x\rangle$

$$\underset{\text{מהגדרת } f}{f(y)} = \sum_{i=1}^{m} \xi_k f_k(y) \geq \sum_{i=1}^{m} \xi_k\left[f_k(x) + \langle g_k(x), y-x\rangle\right] = \sum \xi_k f_k(x) + \sum \xi_k \langle g_k(x), y-x\rangle$$

$$= f(x) + \langle \sum \xi_k g_k(x), y-x\rangle = f(x) + \langle g(x), y-x\rangle$$

מהגדרת f ומלינאריות הגל

$g(x) := \sum \xi_k g_k(x)$

ומהגדרת sub-gradient- $\sum \xi_k g_k(x) = g(x) \in \partial f(x) = \partial \sum \xi_k f_k(x)$

בס"ד

⑥ שילוב של שאלות 5-4, נגדיר g_i sub-gradient מאיים כמו שעשינו בשאלה 4 לכל $\left(\ell^{hinge}_{x_i,y_i}(w,b)\right)$

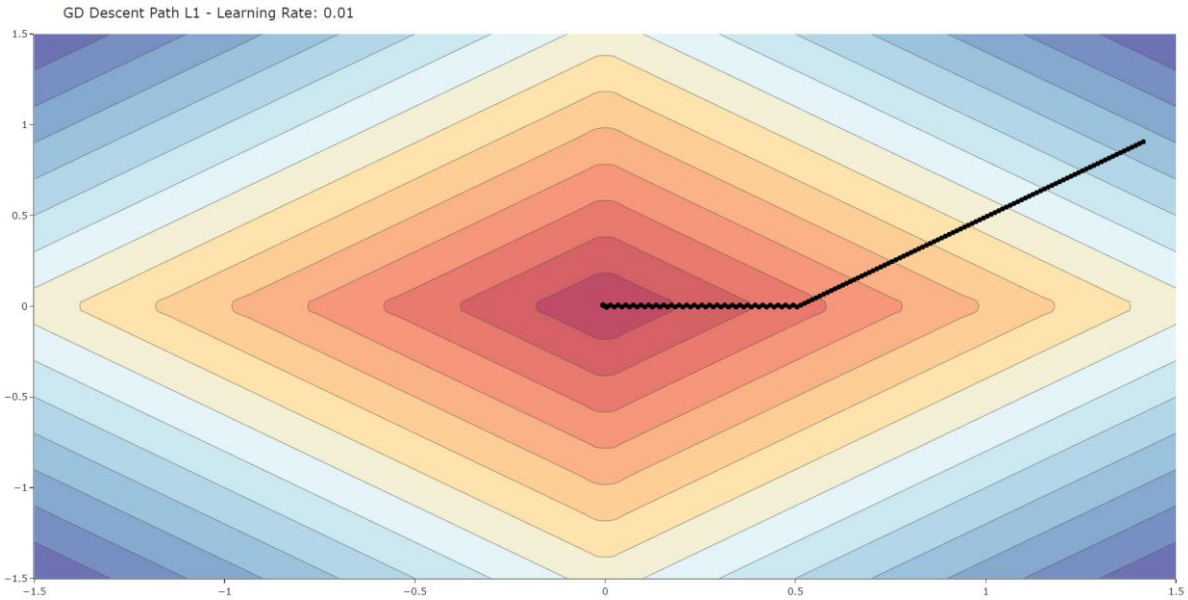
ואז מתקיים שהנחנו בשאלה 5 נובע כי -

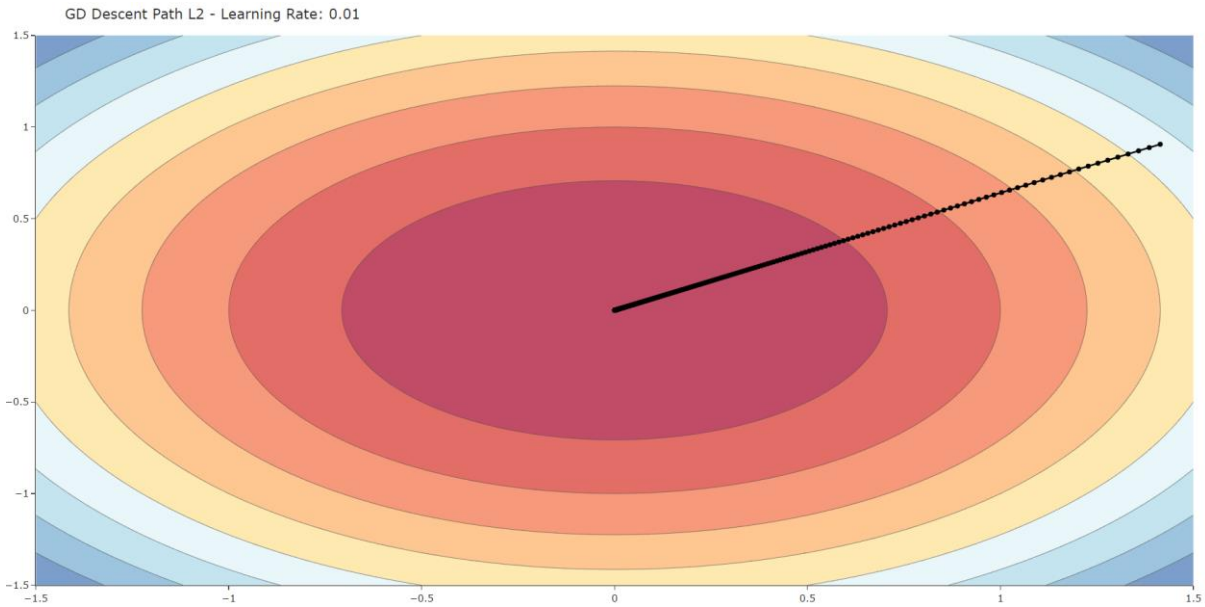
$$\frac{1}{m}\sum_i g_i + \frac{\lambda}{2}\|w\|^2 \in \partial\,\frac{1}{m}\sum_i \ell^{hinge}_{x_i,y_i}(w,b) + \frac{\lambda}{2}\|w\|^2$$

חלק פרקטי

2.1

.1

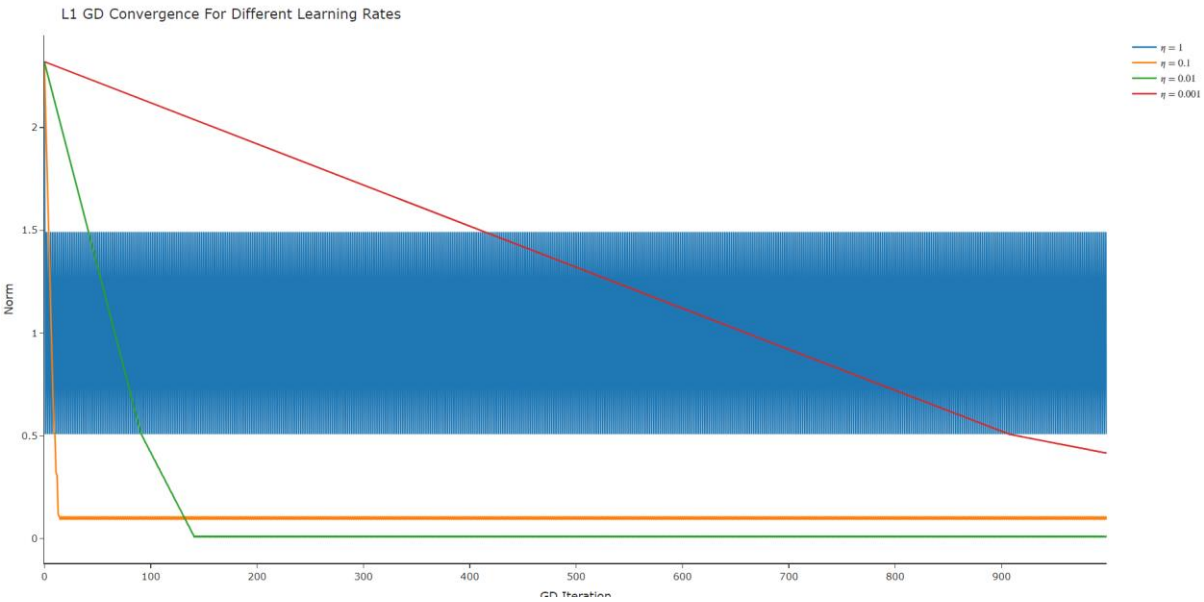




ההבדל המרכזי שניתן לזהות הוא שעבור l2 נקבל תנועה אחידה עבור כל רמה

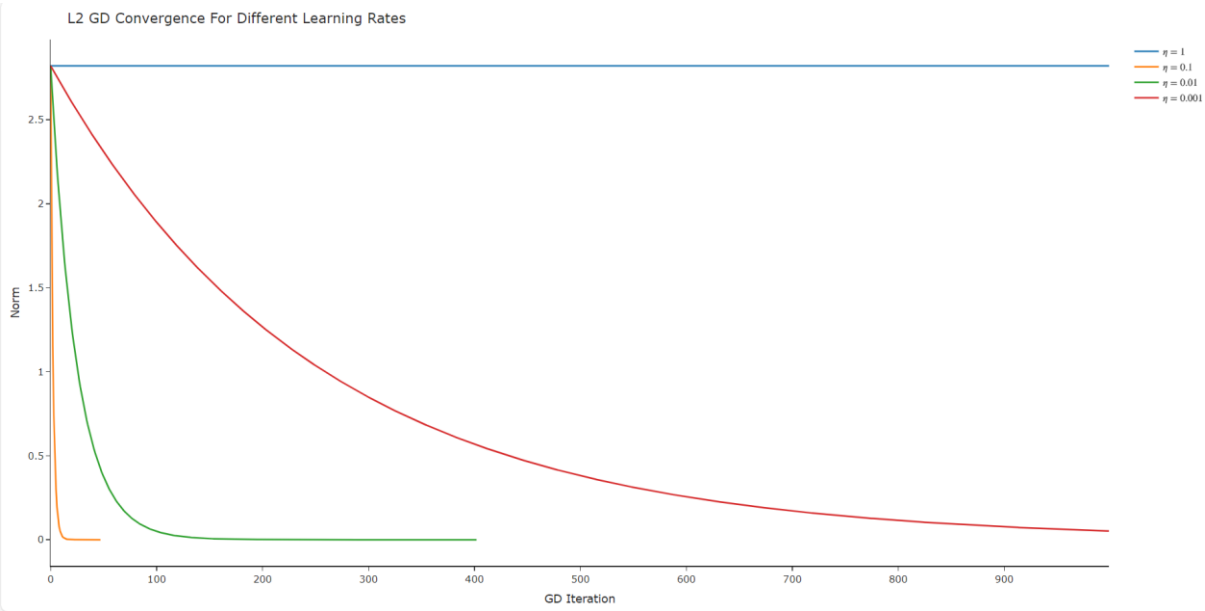
2. נשים לב ל-2 תופעות:
   a. שינוי הכיוון של הירידה מירידה אלכסונית להמשך על הציר המקביל.
   b. לאחר ההמשך בציר במקביל יש קפיצה בין 2 ערכים – כלומר, כל פעם מפספסים את ההתכנסות ומנסים לחזור, אך שינויי הכיוון הוא חד מדי. המשמעות היא שבפועל הגענו למקסימום ההתכנסות שנוכל.

.3



עבור eta=1 אנחנו מקבלים over-shoot חוזר של המודל, שכל פעם קופץ בצורה קיצונית מעל המינימום של פונקציית הLoss מה שמוביל לאוסילציות אין סופיות.

מעבר לכך אנחנו רואים שעבור eta=0.1 ההתכנסות היא מאוד מהירה, ואילו ככל שהוא קטן ההתכנסות היא הרבה יותר איטית (נקווה שב0.001 הוא אכן מתכנס בסוף 😊...)



עבור eta=1 קבלנו התכנסות לפתרון לא טוב (נורמה מאוד גבוהה שלא השתפרה כלל). לשאר האטות נקבל אותו דבר כמו קודם, רק שכאן הלמיד הרבה יותר מהירה.

.4

```
for learning rate: 1 - Lowest loss achieved for L1 module: 0.5081196195534134
for learning rate: 0.1 - Lowest loss achieved for L1 module: 0.09188038044658689
for learning rate: 0.01 - Lowest loss achieved for L1 module: 0.008119619553413011
for learning rate: 0.001 - Lowest loss achieved for L1 module: 0.41611961955345933
for learning rate: 1 - Lowest loss achieved for L2 module: 2.8210062332145176
for learning rate: 0.1 - Lowest loss achieved for L2 module: 2.19211242161629e-09
for learning rate: 0.01 - Lowest loss achieved for L2 module: 2.4898254541043804e-07
for learning rate: 0.001 - Lowest loss achieved for L2 module: 0.05166846244106496
```

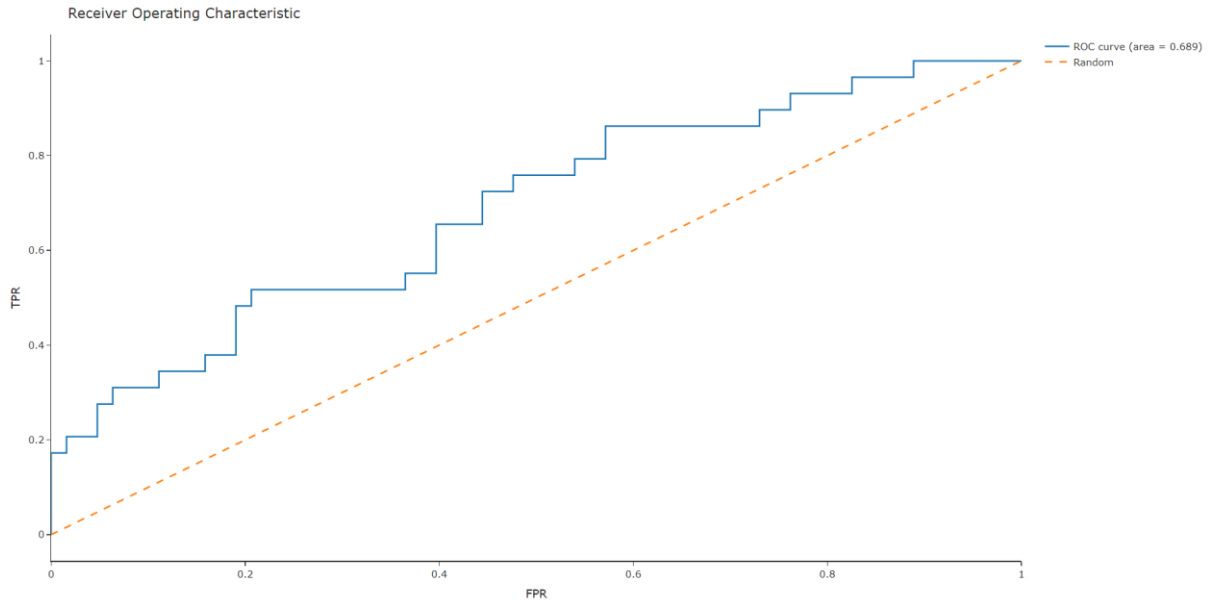
למה l1 פחות טוב מl2?

קודם כל, l1 מנחית חלק מהמשקולות ל0, מה שגורם לפתרון יותר דליל שמאבד חלק מהיכולת להסיק דברים בעזרת פיצ'רים מסויימים (בפרט במקרים בו יש קשר ביניהם).

אך הסיבה העיקרית היא שהערך שהיא מאפטמת את החציון כלומר גרדיאנט קבוע (הנגזרת של פולינום ממעלה ראשונה הוא קבוע) ואילו l2 מאפטמת את הממוצע (משתנה, כיוון שהיא תלויה בx) ולכן הנורמה משתנה כתלות ב X ולכן למודל יש יכולת ללמוד ולהתאים את עצמו ל loss קטן יותר ולהתקרב יותר להתכנסות.

2.2

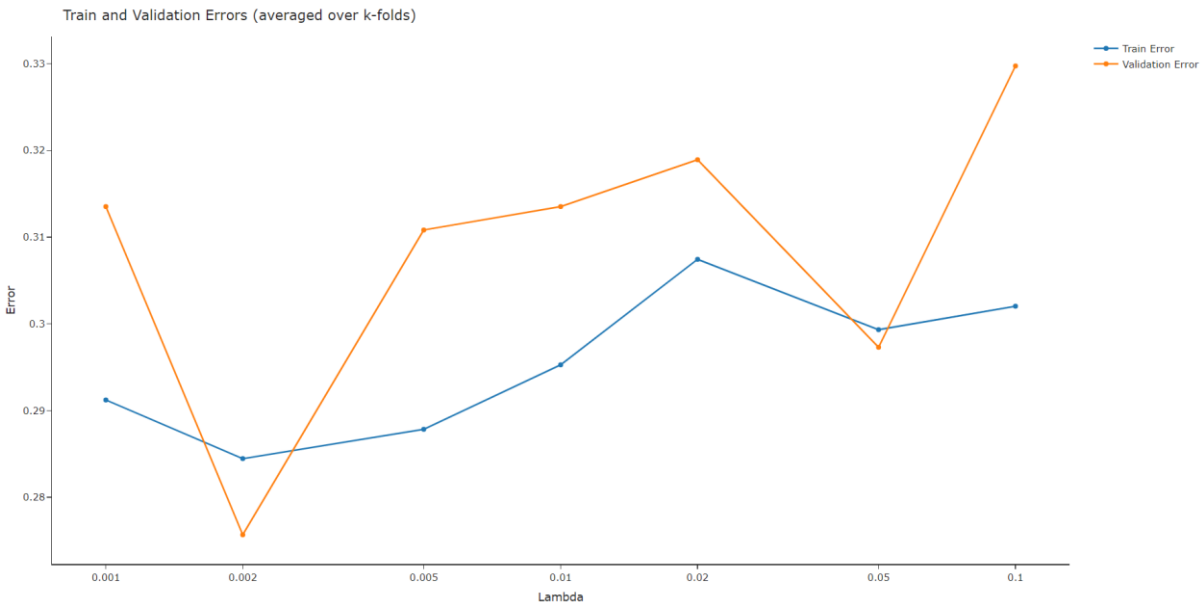
.5



.6

```
Optimal alpha: 0.4785254786906066
Test error with optimal alpha: 0.30434782608695654
```

.7




```
Optimal Regularization Parameter: 0.002
Model achieved test error of 0.25
```