מבוא מערכות לומדות - 67577

רועי זהבי 208648154

Convex optimization:

1. יהיו $f_1, \dots, f_m$ פונק' קמורות, נוכיח כי $g(u) = \sum_{i=1}^{m} \gamma_i f_i(u)$ קמורה גם כן.

הוכחה: לפי הגדרה אנו יודעים כי

$$\forall_i \quad f_i(\alpha u + (1-\alpha) v) \le \alpha f(u) + (1-\alpha) f(v)$$

$u, v \in C$

$$g(\alpha u + (1-\alpha) v) = \quad \Leftarrow$$

$$= \sum_{i=1}^{m} \gamma_i f_i(\alpha u + (1-\alpha) v) \le \sum_{i=1}^{m} \gamma_i (\alpha f_i(u) + (1-\alpha) f_i(v)) =$$

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

$$= \alpha g(u) + (1-\alpha) g(v)$$

$\Leftarrow$ קיבלנו כי $g(\alpha u + (1-\alpha) v) \le \alpha g(u) + (1-\alpha) g(v)$

$\Leftarrow$ פונק' קמורה לפי הגדרה, כנדרש!

$\boxtimes$

2.) יהיו $f(x)=x^2$ ו- $g(x)=-x$

בהרצאה ראינו כי $f(x)$ קמורה (פרבולה חיובית קמורה בצורה ישירה) וראינו כי g פונק' לינארית קמורה (מתקיים שוויון בהגדרה).

$$\Leftarrow \quad g(f(x)) = -x^2$$

⇐ זאת פונק' לא קמורה, אלא קעורה (נגזרת שניה שלילית).

3.) נראה כי פונק' היא קמורה אמ"מ האפיגרף קמור

הוכחה:

⇐ נניח כי f פונק' קמורה.

יהיו $(x_1,y_1),(x_2,y_2)$ נקודות באפיגרף של f

לפי הגדרה ראינו כי $\forall\alpha\in[0,1]$

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

$$\overset{\text{מהגדרת } epi(f)}{\le} \alpha y_1+(1-\alpha)y_2$$

⇐ קיבלנו כי לכל נקודות u,v הצירוף הקמור שייך לאפיגרף ולכן הוא קבוצה קמורה כנדרש.

⇒: כעת נניח כי $epi(f)$ קבוצה קמורה.

יהיו $(u,f(u))$ ו- $(v,f(v))$ נקודות ב-$epi(f)$

⇐ מהגדרת epi נקבל כי

$$f(\alpha u+(1-\alpha)v) \le \alpha f(u)+(1-\alpha)f(v) \in epi(f)$$

מכיוון שזו קבוצה קמורה!

⇐ קרי קיבלנו מן ההגדרה כי f קמורה, כנדרש!
■

4) יהיו $f_i : V \to \mathbb{R}$, $i \in I$ ו- $f : V \to \mathbb{R}$ המוגדרת

$$f(u) = \sup_{i\in I} f_i(u)$$

נראה f קמורה.

הוכחה: יהיו $u, v \in V$ ו- $\alpha \in [0,1]$

יהי $i \in I$

$$f_i(\alpha u + (1-\alpha) v) \le \alpha f_i(u) + (1-\alpha) f(u) \le$$

הגדרת f סד

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

מכיוון שקיים $i \in I$ כך שהנקודה $\alpha u + (1-\alpha) v$ מקיים $f = f_i$ והיא קיבלנו לכל הנקודות

נקבל כי

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

⇐ קיבלנו כי f קמורה לפי הגדרה, כנדרש!
■

5.) נראה כי $f(w,b) = \ell^{hinge}_{x,y}(w,b)$ פונק' קמורה.

<u>פתרון</u>: נתבונן בפונק'

$$\ell^{hinge}_{x,y}(w,b) = \max\{0,\ 1-y(x^Tw+b)\}$$

⇐ בהרצאה ראינו כי פונק' אפיניות הן קמורות, והרי $f(w,b)=0$ וגם $f(w,b) = 1-y(x^Tw+b)$ פונק' <u>אפיניות</u>!

כמו כן ראינו כי פונק' max של פונק' קמורות קמורה גם כן (ובפרט תוצר פונק' <u>קמורה</u>!)

⇐ באופן זה נקבל כי $\ell^{hinge}_{x,y}$ היא פונק' קמורה

מכיוון שזו מורכבת מפונק' קמורות <u>אפיניות</u>!

Fa

6.) נוכיח g sub-gradient הוא $g \in \partial \ell^{hinge}_{x,y}$

⇐ ראינו בתרגול כי בנקודה בה פונק' היא גזירה דיפרנציאבילית, אז נוכל להסתפק על ידי קמור מדדך הפונק' באמת גרדיאנט) ובהמשך ה-0 יוקן מכיוון שאנו עובדים על פונק' שאין דיפרנציאביליות בכמות סופית נקודות

⇐ <u>נגדיר</u>:

$$g = \begin{cases} 0 & \ell^{hinge}_{x,y}(w,b)=0 \\ (-yx,\,-y) & else \end{cases}$$

בחרנו באמת גרדיאנט 0 בנקודה בה הפונק' לא רציפה

אחרת נשאר עם הגדרות הפונקציה!

$\boxtimes$

7) יהיו $f_1,\dots,f_m$ פונקציות קמורות ויהיו $g_i \in \partial f_i(x)$ כאשר $i\in[m]$ תת-גרדיאנט כלשהו של הפונקציה.

נגדיר $f=\sum_{i=1}^{m} f_i$ ונראה כי $\sum_{i=1}^{m} g_i \in \partial \sum_{i=1}^{m} f_i$

הוכחה: ראינו כי פונקציה f גם כן קמורה.

ניזכר בהגדרת תת-הגרדיאנט:

כאשר לכל $u\in dom(f)$ וקטור v הוא תת-גרדיאנט

אם $f(u)\ge f(x)+\langle v,u-x\rangle$

⇐ נקבל כי לכל $i\in[m]$

$$f_i(u)\ge f_i(x)+\langle g_i,u-x\rangle$$

$$f(u)=\sum_{i=1}^{m} f_i(u) \ge \sum_{i=1}^{m} \underbrace{f_i(x)+\langle g_i,u-x\rangle}_{\text{סכום תתי גרדיאנט}} \Leftarrow$$

$$\sum_{i=1}^{m} g_i \in \partial f=\partial \sum_{i=1}^{m} f_i \Leftarrow$$

לפי הגדרת תת-הגרדיאנט כנדרש!

$\boxtimes$

ט) יהי $S=\{(x_i,y_i)\}_{i=1}^{m}$ סט דגימות ותהי $f:\mathbb{R}^d\to\mathbb{R}$

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

<= נמצא תת-גרדיאנט ל-f עם w וקטור משקולות.

פתרון: בדומה לתרגיל 6, כי הדרוש לפונקציה סכומית.

יהי g_i תת-גרדיאנט לפונקציה ה-i המוגדר

$$g_i=\begin{cases}0, & \ell^{hinge}_{x_i,y_i}(w,b)=0\\ (-y_i x_i,-y_i) & else\end{cases}$$

<= ראינו בתרגיל כי הנ"ל סגור תחת סכום ולכן

$$\frac{1}{m}\sum_{i=1}^{m}g_i+\left(\frac{\lambda}{2}\|w\|^2\right)'\in\partial f$$

ובאופן מפורש:

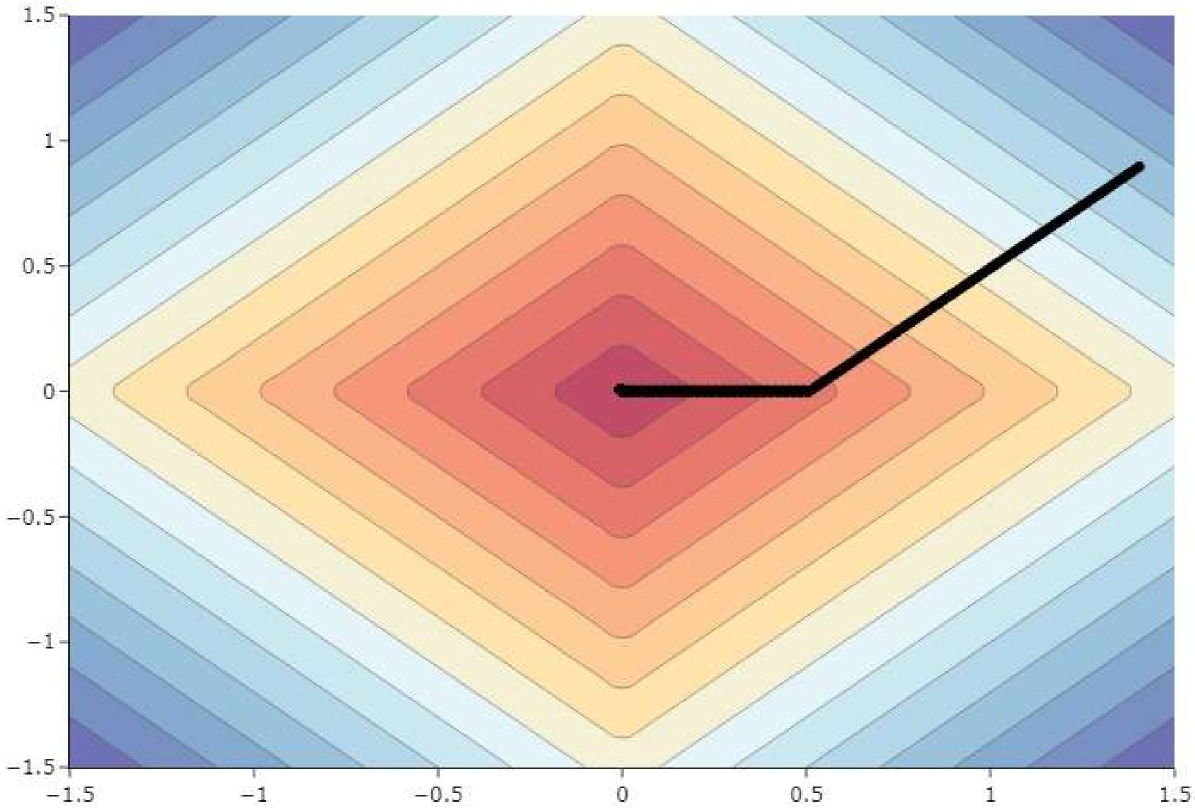
$$=\frac{1}{m}\sum_{i=1}^{m}g_i+\lambda w\in\partial\, \frac{1}{m}\sum_{i=1}^{m}\ell^{hinge}_{x_i,y_i}(w,b)+\frac{\lambda}{2}\|w\|^2$$

כנדרש!

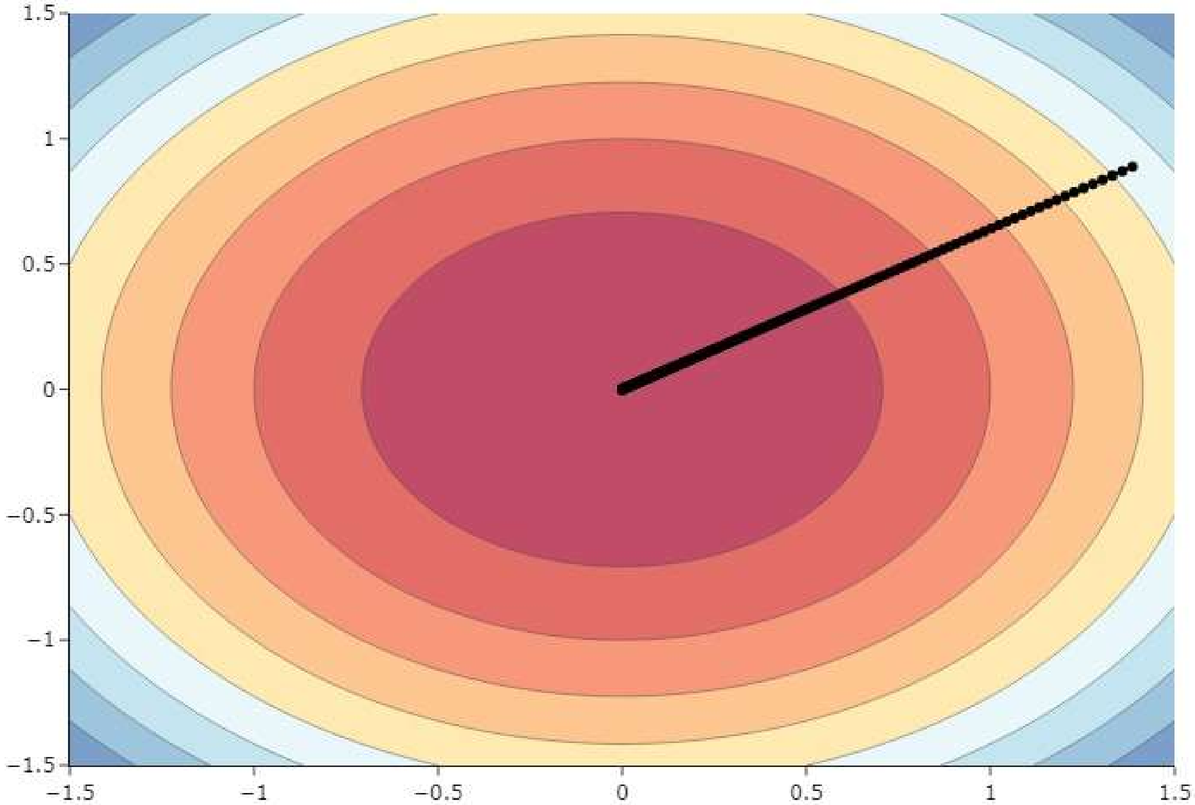
$\blacksquare$

1. להלן הפלטים הנדרשים

GD Descent Path decent path for L1 module with eta=0.01



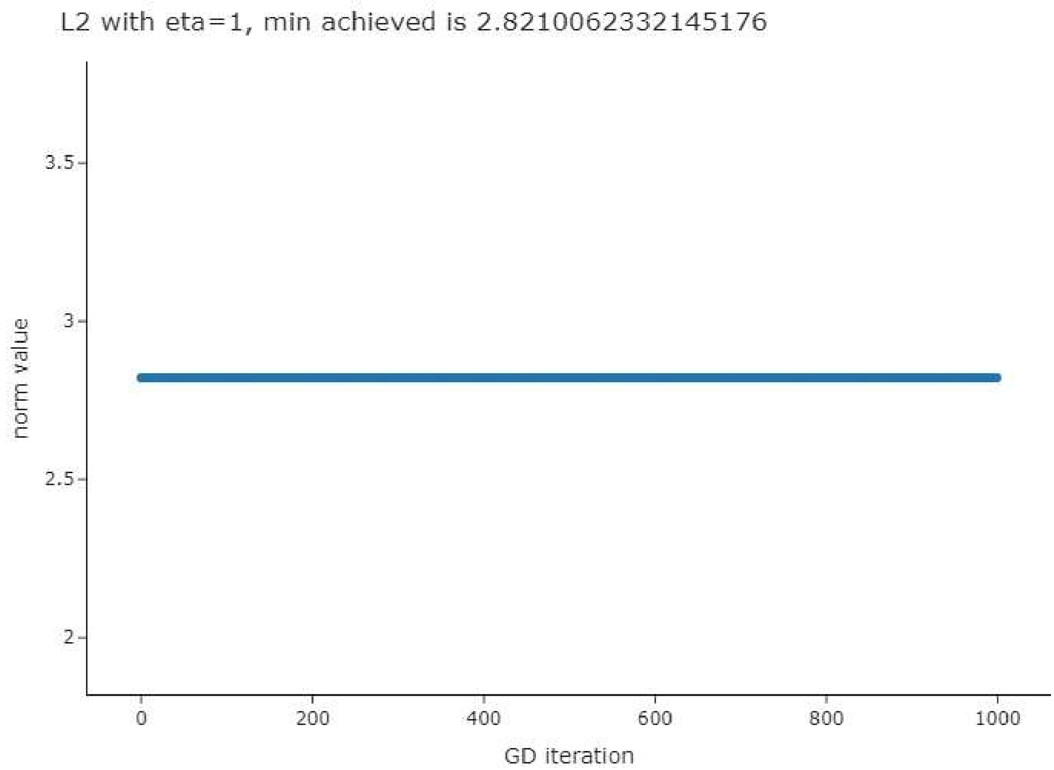
GD Descent Path decent path for L2 module with eta=0.01



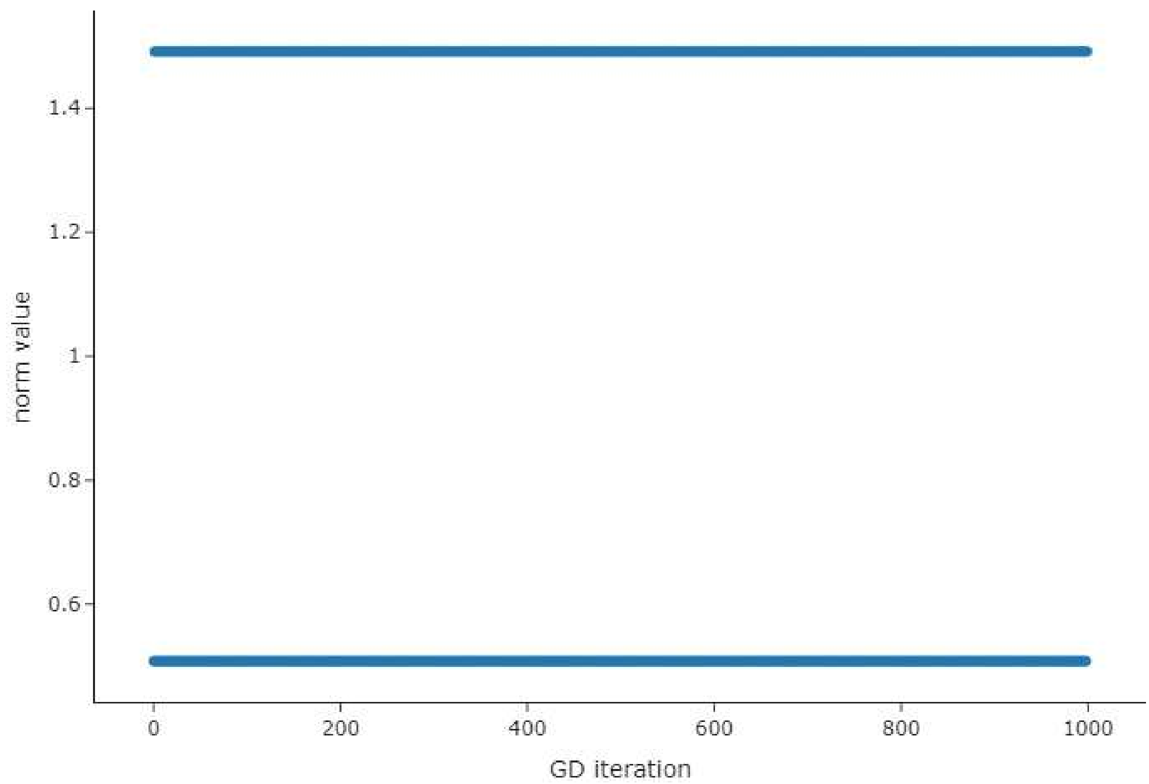
נתבונן בהבדלים בין המטריקות, ניתן לראות כי עבור $L1$ כיווני הגרדיאנט יביאו אותנו לאחד מן הצירים ומשם נרד על גבי הציר עד שנגיע לראשית – זו הדרך המהירה ביותר במקרה זה, את הנ״ל ניתן לראות באופן ברור כאשר גרף הירידה מגיע לציר הקרוב ביותר וממשיך בכיוונו לראשית. זאת בניגוד ל-$L2$ אשר אין משמעות לירידה דרך ציר ולכן אנו מתכוונים בכיוון הגרדיאנט החל מהנקודה בה התחלנו עד הראשית.

2. ניתן לראות שתי תופעות בביצוע GD על נורמה $l1$ עבור צעד קבוע בפלט הנתון.
   a. ראשית ניתן לראות כי אין אנו מתכנסים בראשית ביגוד ל-$l2$, אשר במקרה זה הגרדיאנט שואף ממש ל-0 ולכן הצעד הקבוע אינו משנה. זאת ביחס ל-$l1$ אשר בנקודה זו היא אינה גזירה ועל כן לא מתכנסת במקרה זה.
   b. שינוי הכיוון החד בגרף הירידה – כפי שציינתי זו תופעה שמתקיימת רק בנורמה $l1$, הנ״ל מתקיים מכיוון שאנו מכיוון כלשהו מגיעים ל-0 ושם מתקיימת נקודה אי הגזירות של הנורמה בכל נגזרת חלקית, לכן הדרך בה התמודדנו היא ע״י בחירת תת-גרדיאנט כלשהו (לדוגמא איפוס הנגזרת החלקית הנ״ל). הפונקציה מבצעת תהליך מינימיזציה אשר מביא אותה לאחד מן הצירים מכיוון שהגרדיאנטים בכל נקודה שלא על הגרדיאנטים שווים ואז משם אנו נעים בכיוון אחד עד לאיפוס של שני הערכים בוקטור הנגזרות.

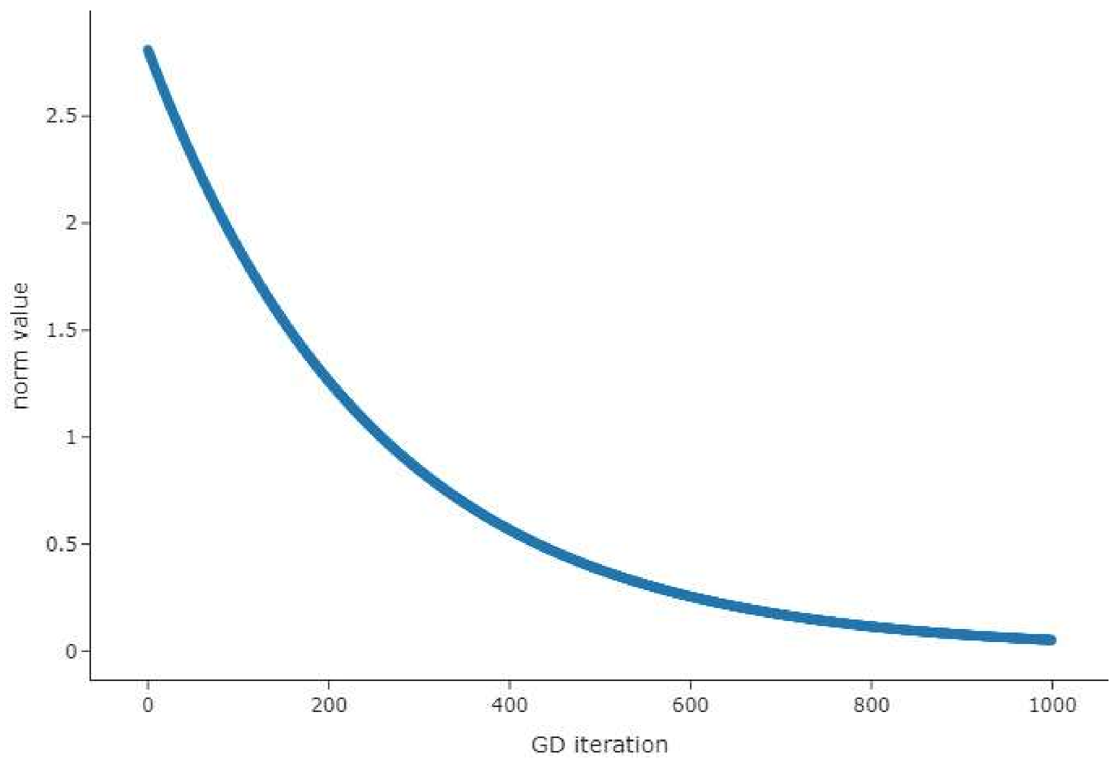
3. מצב הפלטים לכל אחד מן הנורמות וערכי הη השונים, הטעות המינימלית מופיע בכותרת הפלט.



L1 with eta=1, min achieved is 0.5081196195534134

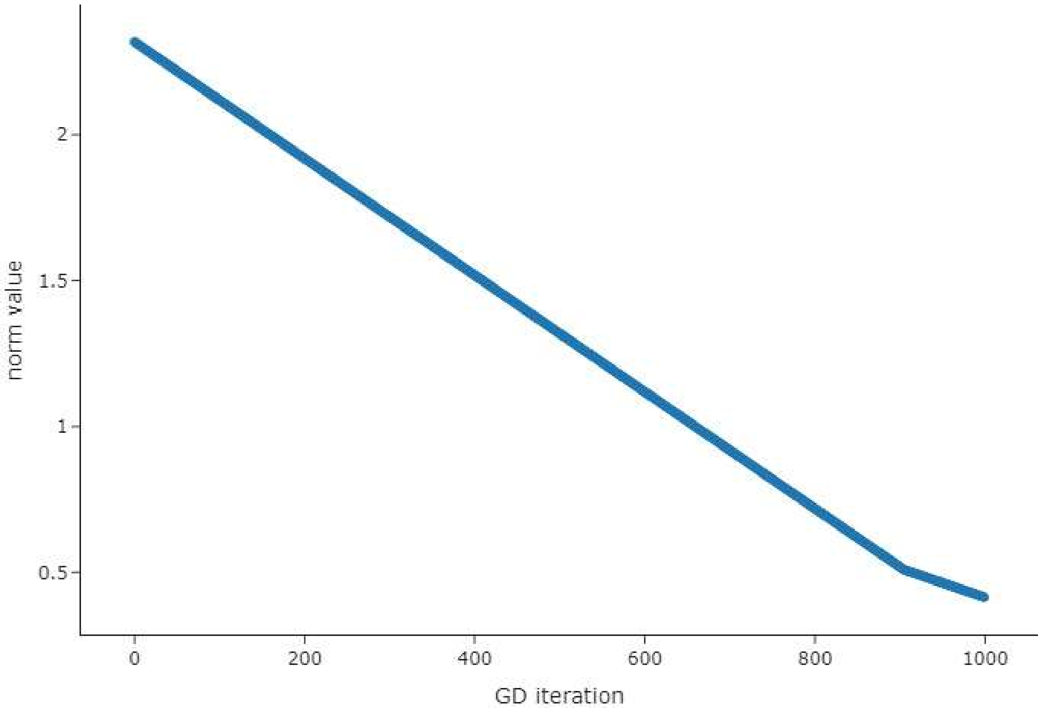


L2 with eta=0.001, min achieved is 0.05146199526515047

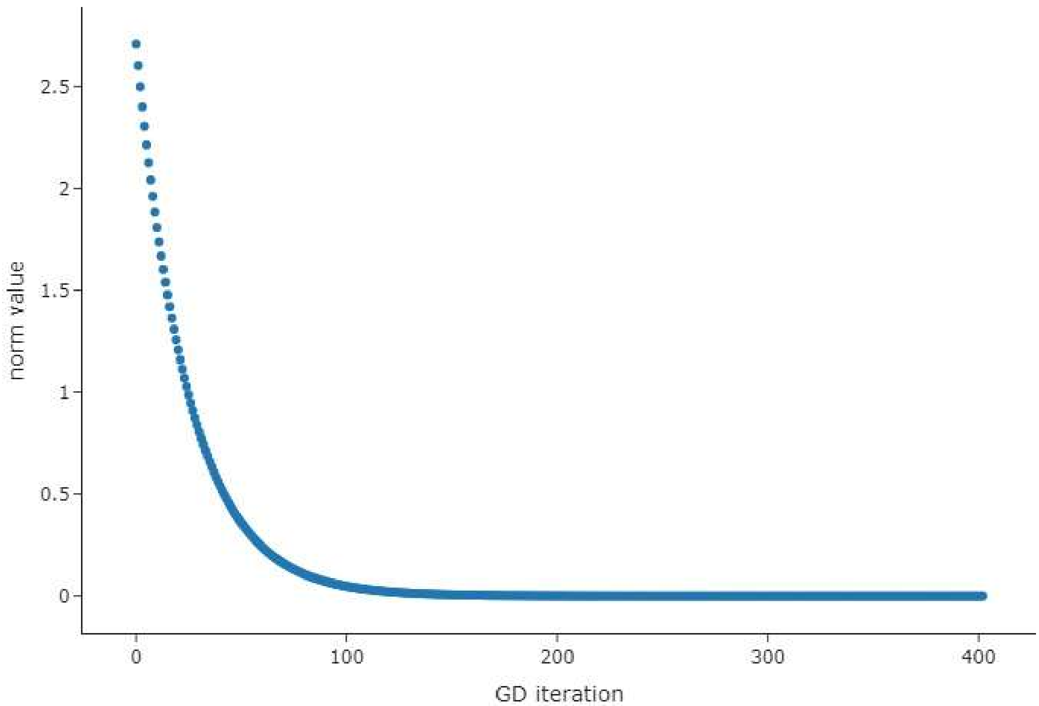


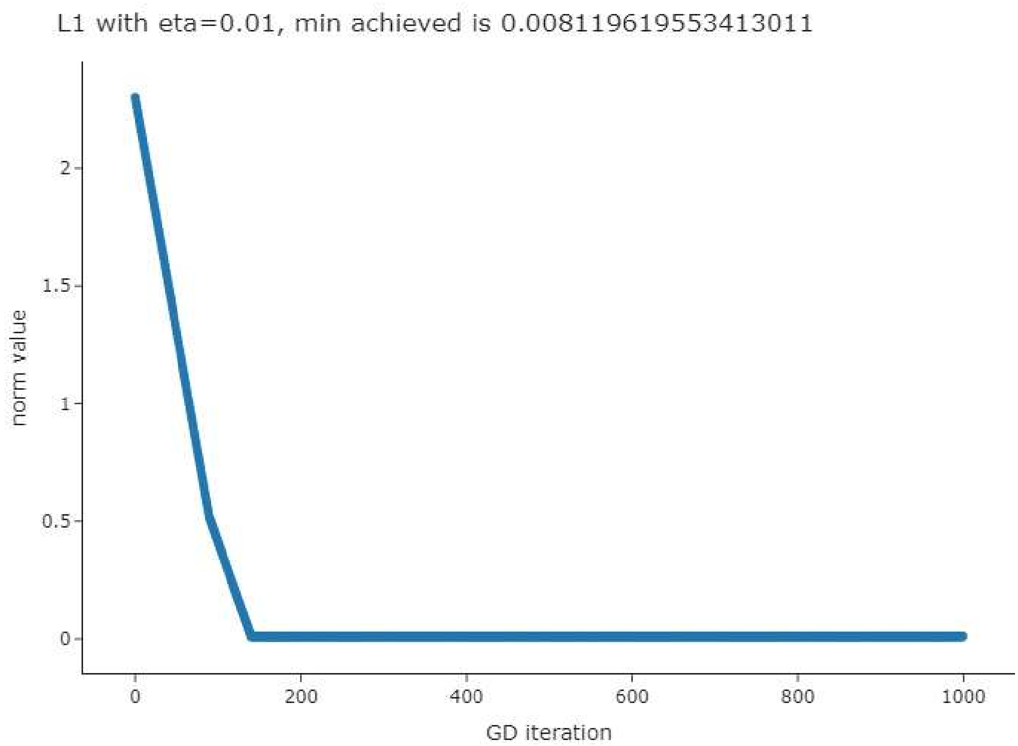
מבוא למערכות לומדות 67577
רועי זהבי 208648154

L1 with eta=0.001, min achieved is 0.4143075051928211

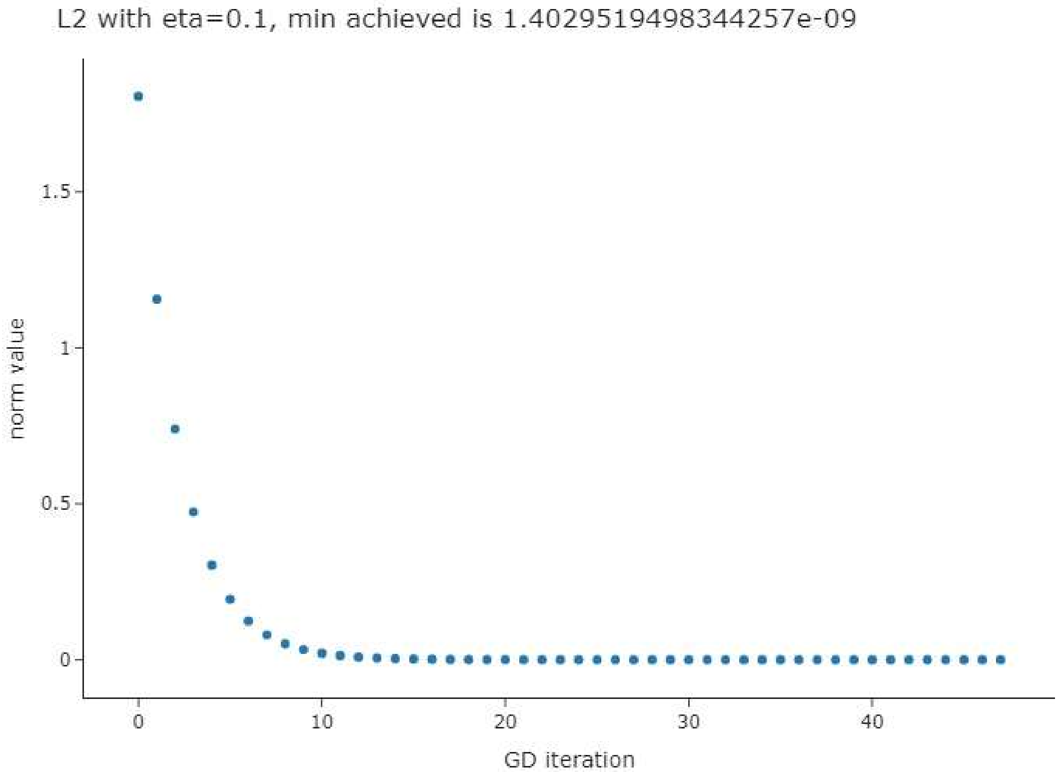


L2 with eta=0.01, min achieved is 2.3912283661218465e-07

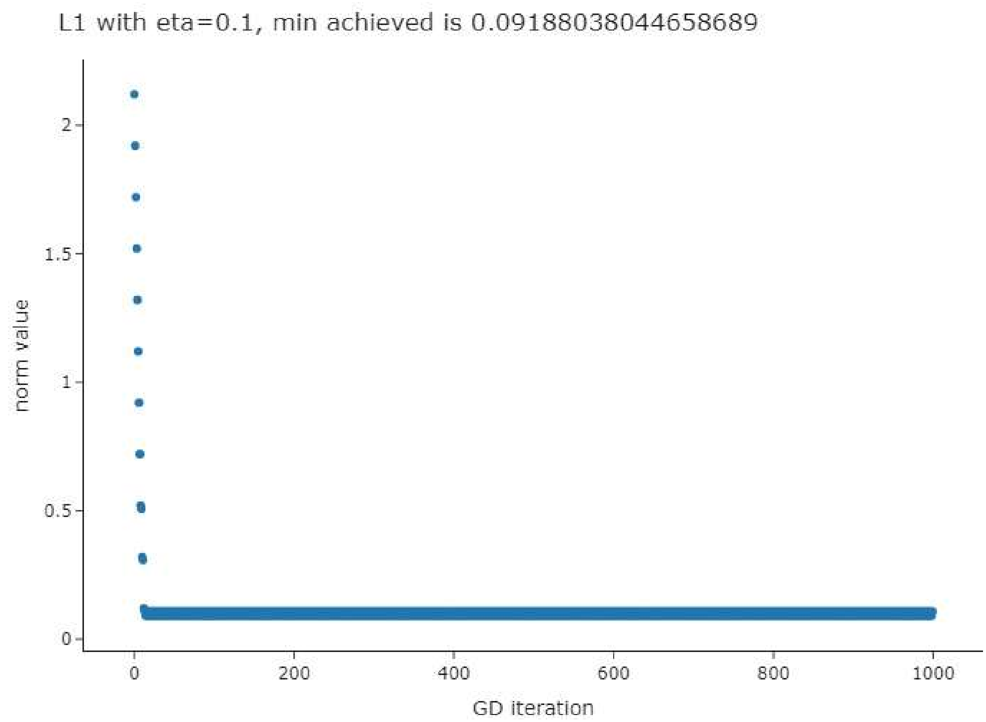




L1 with eta=0.01, min achieved is 0.008119619553413011
norm value
GD iteration
0
0.5
1
1.5
2
200
400
600
800
1000

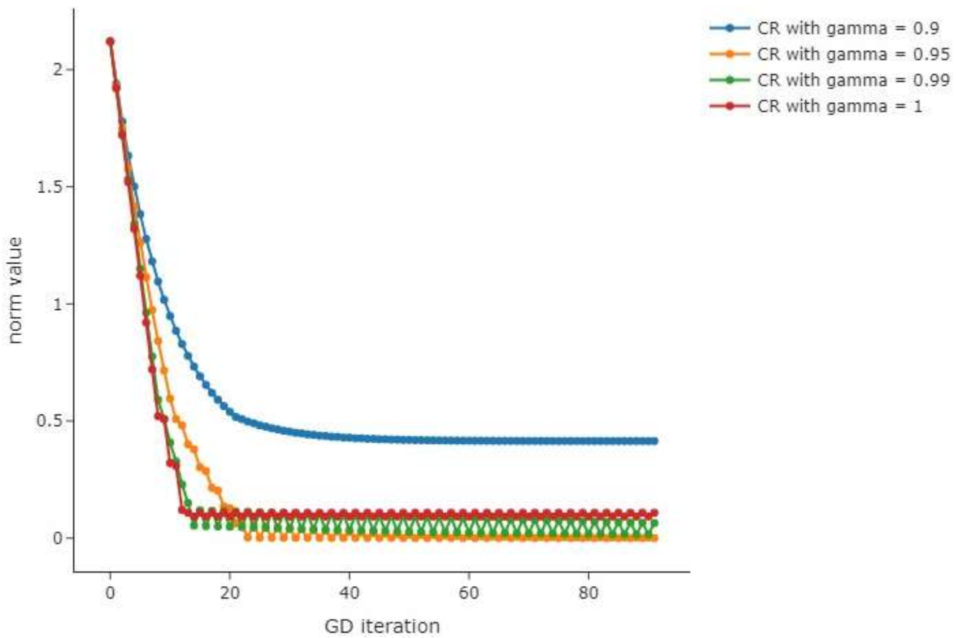


L2 with eta=0.1, min achieved is 1.4029519498344257e-09
norm value
GD iteration
0
0.5
1
1.5
10
20
30
40



4. ניתן לראות כי בהינתן צעדים קטנים מאוד הפלטים מתכנסים ועל כן תוצאות הנמוכות ביותר מאוד קטנות ושאופות ל-0. כאשר הצעד קבוע יחסית גדול האלגוריתם בכל שלב קופץ בין אזורים שונים סביב המינימום ועל כן הוא לא מצליח לבצע תהליך מינמום טוב – קרי הערך הנמוך ביותר יחסית גדול ביחס לגדלי צעד אחרים.

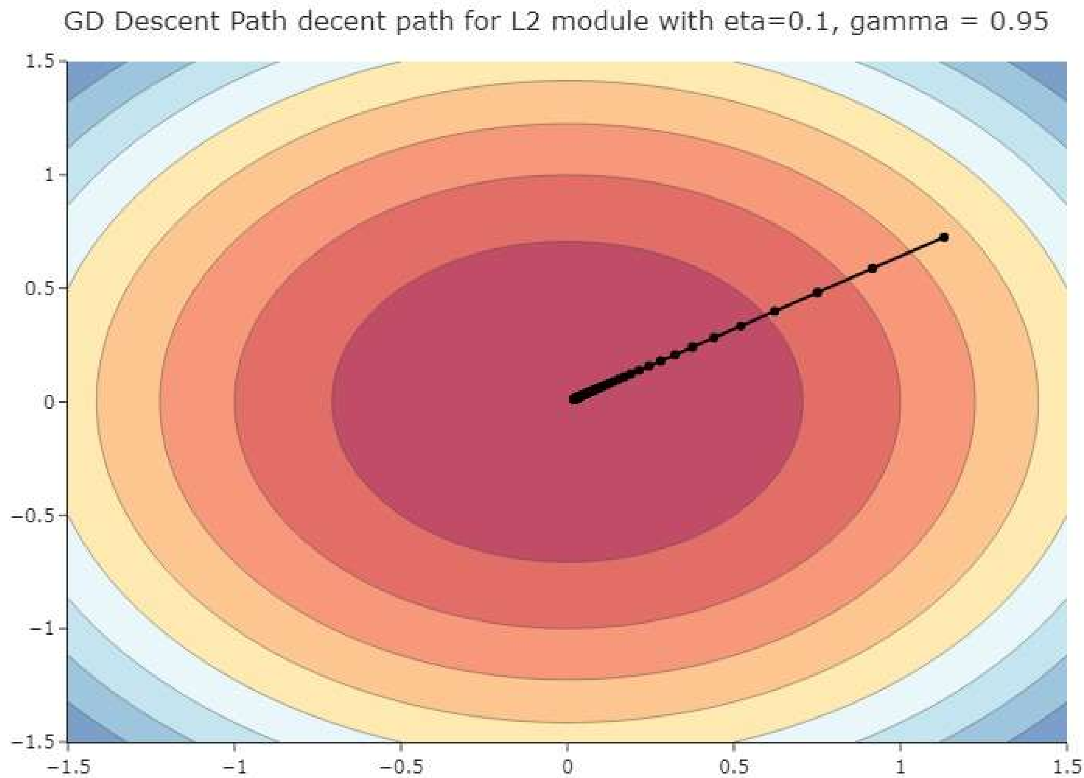
5. להלן הפלט

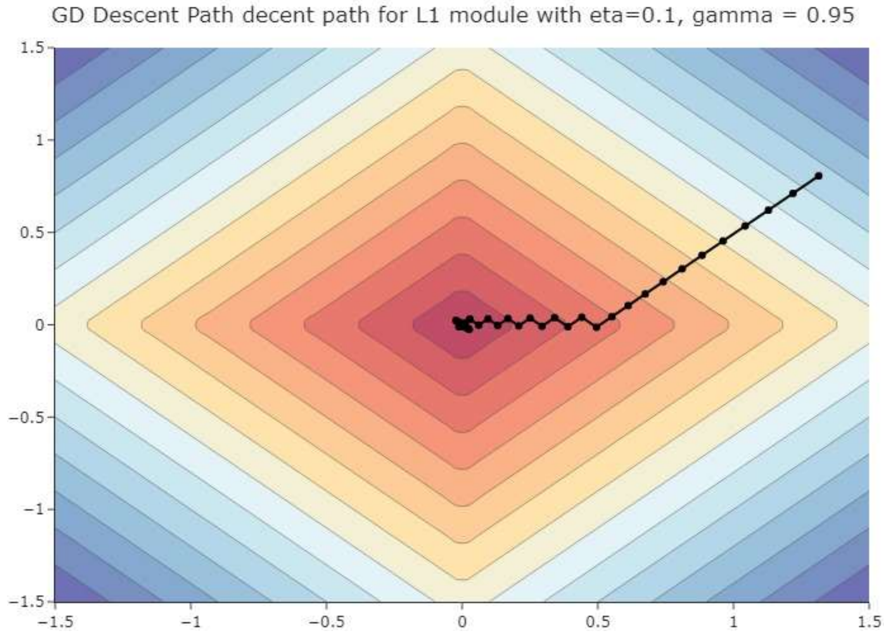


ניתן לראות כי עבור ערכי γ קטנים צעדים שואפים ל-0 מהר מאוד ועל כן אנו נתקעים בשגיאה יחסית גדולה אשר תעלה מספר גדול מאוד של איטרציות ע״מ לתקן. מנגד על γ גדולה אז הגרף מצבע מעין ״זיגזוג״, הצעדים עדין אל מתכנסים סביב המינימום ועל כן התוצאה אינה מדויקת – על כן בערכי הביניים אנו יכולים לראות כי תוך מספר לא גבוהה של איטרציות האלגוריתם מתכנס למינימום – על כן γ הינו היפר פרמטר שעלינו לוודא כדי לראות כי אנו מקבלים ערכים טובים למודל שלנו, תוך מצב שבו אנו נמנעים מצעדים גדולים שעלולים לגרור אי דיוק, או קטנים שלא בטוח שניתנים לחישוב מפאת מספר האיטרציות הגובה.

6. ניתן לראות כי במקרה של גודל צעד משתנה האלגוריתם השיג תוצאות משמעותיות טובות יותר עבור נורמה $l1$ – עובדה כי אין אנו מופתעים ממנה, בניגוד ל$l2$ בה הנגזרת בנקודת המינימום מתאפסת אין זה הדבר במקרה של $l1$ על כן לעולם לא נוכל לבצע תהליך מינימזציה איכותי ומדויק עם גודל צעד קבוע ולכן במקרה בו הפונקציה אינה גזירה בנקודות המינימום דרך לבצע GD תהיה ע״י שימוש בגודל צעד משתנה.
זאת ניתן לראות בערך המינימום:
עבור צעד קבוע הערך הקטן ביותר היה 0.00811-
ועבור צעד משתנה בדעיכה אקספוננציאלית - $\sim 6.573 * e^{-11}$.

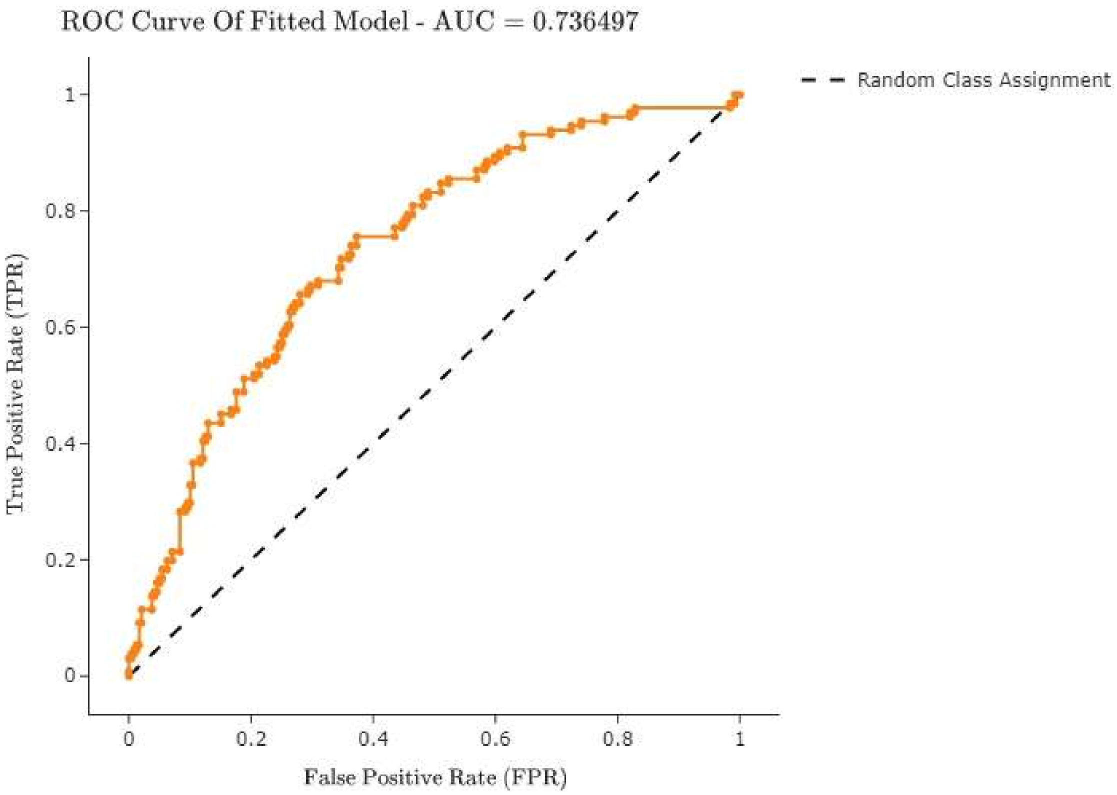
7. להלן הפלטים





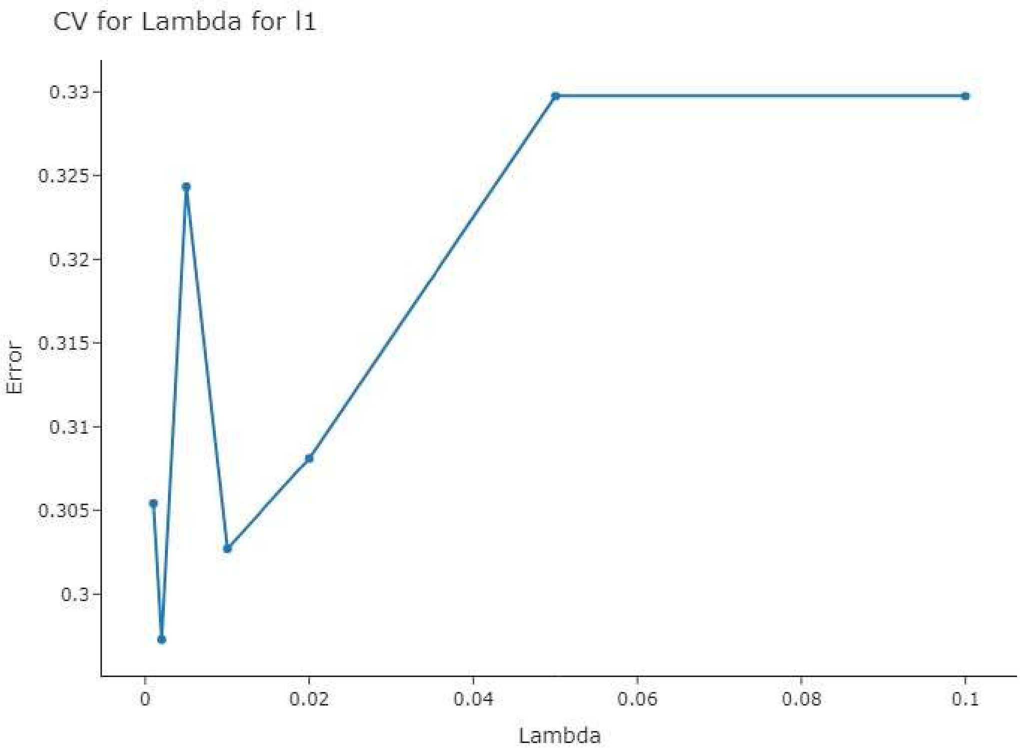
ניתן לראות את השינוי בהתכנסות של $l1$ סביב המינימום האמיתי, הצעדים הולכים וקטנים ועל כן אנו מסוגלים לבצע תהליך מינימזציה. במקרה זה אנו רואים גם התכנסות מהירה יותר ב-$l2$ מכיוון שכעת ישנם שני ערכים אשר מקטינים את גודל הצעד - הגרדיאנט וגודל הצעד.

8. להלן פלט ה-ROC



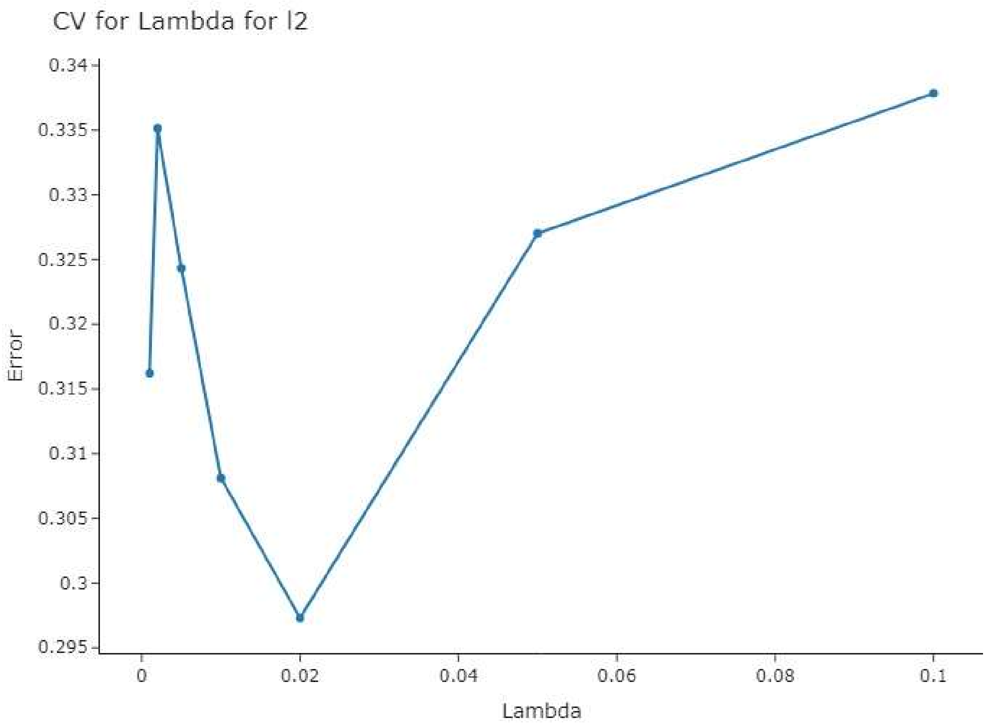
9. נשים לב כי ערך האלפא הקטן ביותר שהתקבל הינו 0.31 כאשר שגיאת המודל עם אלפא זו הינה 0.315.

10. להלן פלט הולידציה למציאת ערך הלמדא לרוגלריזציה תחת נורמת $l1$.



נשים כי במקרה זה נבחרה $\lambda = 0.002$ והמודל השיג בעזרתה שגיאה של 0.283

11. עבור $l2$ להלן פלט הולידציה



במקרה זה נבחרה $\lambda = 0.02$ ובעזרתה המודל השיג שגיאה של 0.24.