Theoretical Part

1.1 Convex optimization

1. נתון ש- $f_1,\dots,f_m$ הן פונקציות קמורות אזי לכל $\alpha\in[0,1]$.
כך שלכל $i\in[m]$ ו- $x,y\in C$

$$f_i(\alpha x+(1-\alpha)y) \leq \alpha f_i(x)+(1-\alpha)y$$

כעת נראה כאות על g. יהינתן $x,y\in C$ ו- $\bar{\alpha}\in[0,1]$

$$g(\alpha x+(1-\alpha)y) = \sum_{i=1}^{m} \gamma_i f_i(\alpha x+(1-\alpha)y)$$

$$\leq \sum_{i=1}^{m} \gamma_i\, \alpha f_i(x)+(1-\alpha)f(y)$$

$$= \sum_{i=1}^{m} \gamma_i\, \alpha f_i(x) + \sum_{i=1}^{m} \gamma_i (1-\alpha) f_i(y)$$

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

$$= \alpha g(x) + (1-\alpha) g(y)$$

כנדרש.

2. דוגמה נגדית: נגדיר $g(x)=x^2$, $f(x)=-x$. שתי הפונקציות הן קמורות עפ"י הגדרה אבל מתקיים ש- $h=f\circ g=-(x^2)$ אינה קמורה.

1.2 Sub-gradients for soft-SVM objective

3. ראינו בכיתה ש- $w^T x+b$ (פונקציה אופינית) היא פונקציה קמורה ולכן נקבל ש- $1-y(w^T x+b)$ היא גם קמורה. 0 היא פונקציה קמורה באופן טריוויאלי. וראינו בכיתה כי פונקציית ה- max של פונקציות קמורות היא קמורה גם כן.
לכן סה"כ נקבל ש- hinge loss היא פונקציה קמורה.

4. בהינתן $\ell hinge_{x,y}(w,b)$ מתקיים:

$$\ell hinge_{x,y}(w,b) = \begin{cases} 0 & y(x^T w+b) \geq 1 \\ 1-y(x^T w+b) & y(x^T w+b) < 1 \end{cases}$$

ה- sub-gradient g מוגדר באופן הבא: $g\left(\frac{\partial hinge}{\partial w}, \frac{\partial hinge}{\partial b}\right)$
אם נפצל למקרים: $y(x^T w+b) \geq 1 \Leftarrow 0 = \ell hinge(w,b)$ אז:

$$\frac{\partial hinge}{\partial w} = 0, \quad \frac{\partial hinge}{\partial b} = 0 \Rightarrow g = 0$$

אם $y(w^T x+b) < 1 \Leftarrow 1-y(x^T w+b) = \ell hinge(w,b)$ אז:

$$\frac{\partial hinge}{\partial w} = -yx, \quad \frac{\partial hinge}{\partial b} = -y \Rightarrow g(-yx, -y)$$

סה"כ נקבל:

$$\ell_{hinge}(w,b)=0 \qquad g=\begin{cases}0\\(-yx,-y)\end{cases}$$

$$\ell_{hinge}(w,b)=1-y(x^Tw+b)$$

5. מהגדרה: $g_k \in \partial f_k(x)$ לכל $k\in[m]$. כלומר לכל $y\in\mathbb{R}^d$:

$f_k(y)\geq f_k(x)+g_k^T(x-y)$ אזי מתקיים:

$$\sum_{i=1}^m f_i(y)\geq\sum_{i=1}^m f_i(x)+g_i^T(x-y)=\sum_{i=1}^m f_i(x)+\sum_{i=1}^m g_i^T(x-y)$$

מהגדרה: $f(x)=\sum_{i=1}^m f_i(x)$ מתקיים:

$$f(y)\geq f(x)+\sum_{i=1}^m g_i^T(x-y)$$

$$\rightarrow f(y)\geq f(x)+\left(\sum_{i=1}^m g_i\right)^T(x-y)$$

כנדרש.

6. נשים לב כי משוואה 4 חישבנו וקיבלנו:

$$g_i=\begin{cases}0 & \ell_{hinge_{x_i,y_i}}(w,b)=0\\(-y_ix_i,-y_i) & \text{else}\end{cases}$$

בנוסף, משוואה 5, שני הפונקציות בסכום הן קמורות. וכפל בסקלר של פונקציה הוא קמור נקבל:

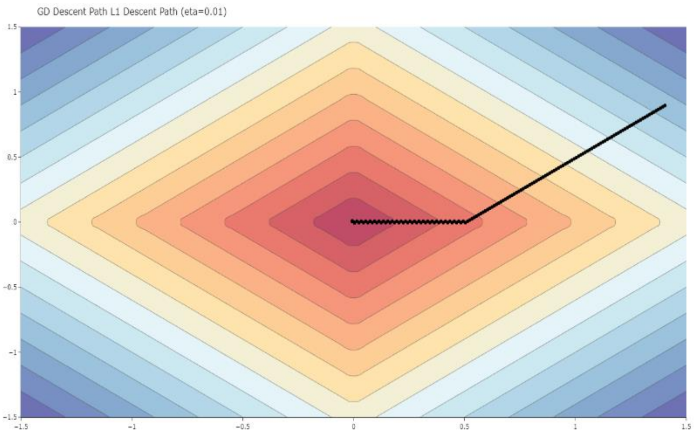
עבור $w\in\mathbb{R}^d$.

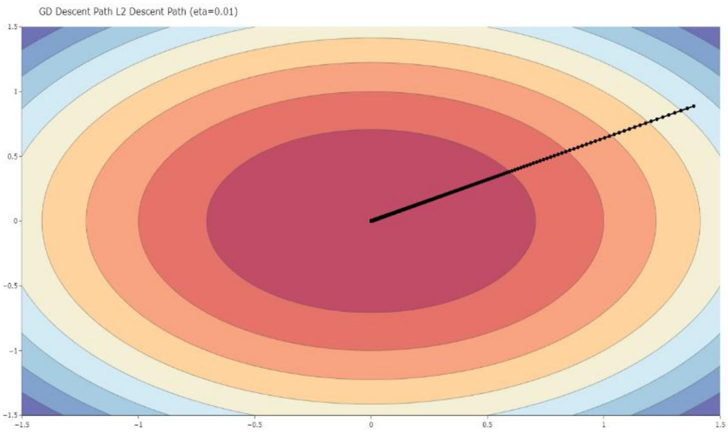
$$\frac{1}{m}\sum_{i=1}^m g_i+\lambda\|w\|\in\partial f(w,b)$$

<u>Practical part – IML</u>

2.1.1 Comparing Fixed learning rates

<u>1. explain the differences seen between the L1 and L2 modules:</u>





נשים לב להבדלים בין המודלים:

במודל L2 הכיוון של הגרדיאנט הוא בקו ישר אל נקודת המינימום בעוד במודל L1 הגרדיאנט יורד בקו ישר עד נקודה כלשהי (עד שערך ה Y מתאפס) ואז מבצע פנייה חדה ויורד בזיג-זג במקביל לציר ה-X עד נקודת המינימום בערך (0,0).

הדבר נובע מכך שבמודל L2 כל שתי נקודות שהן על אותו "טבעת" יש מרחק שווה מנקודת המינימום לכן ההתנהגות של הקו הוא לרדת במורד הטבעות ישירות אל נקודת המינימום בעוד שבמודל L1 לכל שתי נקודות שנמצאות על אותו "מסגרת" אין מרחק שווה מנקודת המינימום.

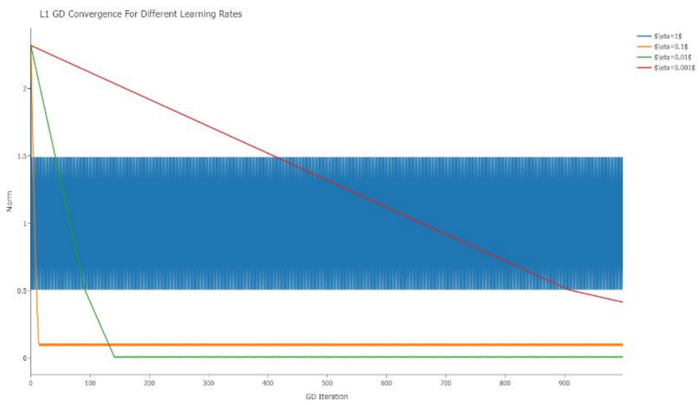
<u>2. following the previous question describe two phenomena that you have seen in the descent path of the ℓ1 objective when using GD and a fixed learning rate:</u>

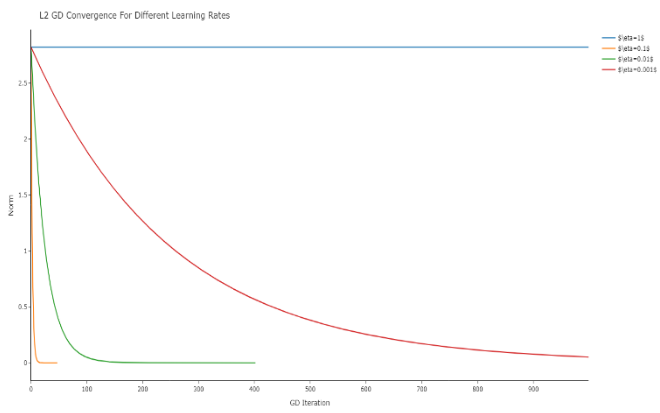
שתי תופעות שמתרחשות במודל L1 הן:

1. כיוון הגרדיאנט יורד באלכסון לכיוון נקודת המינימום עד שהוא מבצע "שבירה" כאשר ערך ה Y מתאפס ואז הכיוון הוא לאורך ציר ה xעד נקודת המינימום.

2. מטבע התנהגות הפונקציה L1, כאשר מתקרבים אל נקודת המינימום, הפונקציה מזגזגת אל המינימום (עולה ויורדת) ולא יורדת בקו ישר.

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:





נסתכל על הגרף של המודל L1 (צד שמאל)

η =1 : הנורמה נותרת תנודתית, וערכי הגרדיאנט נעים ללא התכנסות מ0.5 ל1.5. מה שמעיד על קצב למידה גבוה מדי שגורם לקפיצות יתר ולאי התכנסות.

η =0.1 : הנורמה יורדת במהירות לנקודה יציבה בתוך מספר קטן של איטרציות ונשארת קבועה, מה שמעיד על קצב למידה יחסית טוב.

η =0.01 : הנורמה יורדת בהדרגה ומתכנסת לנקודה יציבה לאחר כ- 140 איטרציות לערך.

η =0.001 : הנורמה יורדת מאוד לאט לאורך כל האיטרציות, מה שמעיד על קצב התכנסות איטי.

נסתכל על הגרף של המודל L2 (צד ימין)

η =1 : הנורמה לא יורדת, מה שמעיד על קצב למידה גבוה מדי וכישלון בהתכנסות.

η =0.1 : הנורמה יורדת במהירות לאפס בתוך 50 איטרציות, מה שמראה על התכנסות מהירה מדי.

η =0.01 : הנורמה יורדת בהדרגה ומגיעה כמעט לאפס לאחר כ-200 איטרציות בערך.

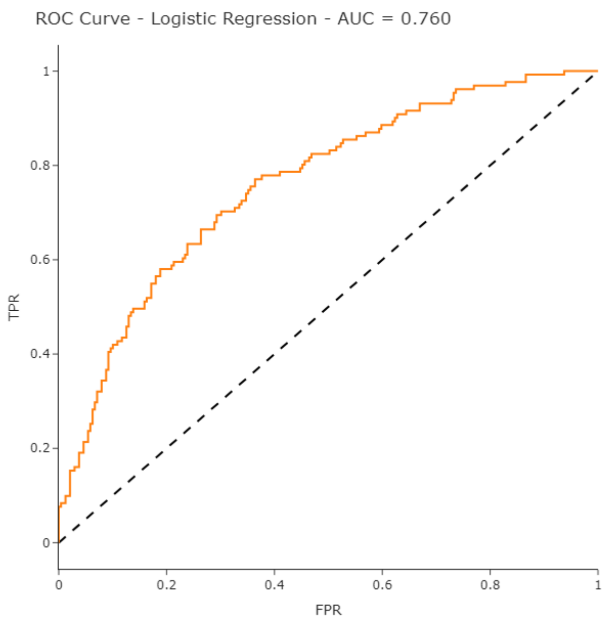
η =0.001 : הנורמה יורדת בהדרגה לאורך כל האיטרציות, מה שמעיד על קצב התכנסות איטי מדי.

נסיק מהגרפים שקצב הלמידה אופטימלי חשוב להתכנסות יעילה. קצב איטי מדי עלול לקחת זמן רב וקצב מהיר מדי עלול לגרום ל"קפיצות". קצב הלמידה משתנה בין שני המודלים L1 ל-L2 בשל השפעתם השונה על עדכוני המשקולות במהלך גרדיאנט דסנט.

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

2.2 Minimizing Regularized Logistic Regression

5. Using your implementation, fit a logistic regression model over the data. Use the predict_proba to plot an ROC curve:



6. Which value of α achieves the optimal ROC value according to the criterion below. Using this value of α ∗ what is the model's test error?:

```
pnina_ei@pond:~/IML/ex4 $ python3 gradient_descent_investigation.py
The value of α achieves the optimal ROC is:  0.3247731561147603
The model's test error is: 0.33695652173913043
```

קבלנו שהאלפא האופטימלית היא: **α = 0.324** ושגיאת הטסט היא: **0.336**.

7. What value of λ was selected and what is the model's test error?:

```
Optimal regularization parameter: 0.02
Model achieved test error of 0.28
```

קבלנו שהלמדא האופטימלית היא **λ = 0.02**. ושגיאת הטסט היא **0.28**.