מערכות לומדות תרגיל 6

שיר שבח 322407701

22 ביוני 2022

חלק תיאורטי

1. יהי $f_1,...,f_m : C \to \mathbb{R}$ קבוצה של פונקציות קמורות ו־ $\gamma_1,..,\gamma_m \in \mathbb{R}_+$. הוכיחו מההגדרה ש־ $g(u)=\sum_{i=1}^{m}\gamma_i f_i(u)$ זאת פונקציה קמורה.

עבור $u,v \in C$ ועבור $\alpha \in [0,1]$:

$$\begin{aligned} g(\alpha v+(1-\alpha)u) &= \sum_{i=1}^{m}\gamma_i f_i(\alpha v+(1-\alpha)u) \\ &\le \sum_{i=1}^{m}\gamma_i(\alpha f_i(v)+(1-\alpha)f(u)) \\ &= \alpha\sum_{i=1}^{m}\gamma_i f_i(v)+(1-\alpha)\sum_{i=1}^{m}\gamma_i f(u) \\ &= \alpha g(v)+(1-\alpha)g(u) \end{aligned}$$

קיבלנו:

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

ולכן g פונקציה קמורה.

2. תנו דוגמה נגדית לטענה הבאה: בהינתן שתי פונקציות $f,g : \mathbb{R} \to \mathbb{R}$, הגדירו את הפונקציה $h : \mathbb{R} \to \mathbb{R}$ ע"י $h = f \circ g$. אם f ו־g קמורות אז h קמורה גם כן.

ניקח את הפונקציה $f(x)=g(x)=x^2-1$. מתקיים:

$$\begin{aligned} f'(x) &= 2x \\ f''(x) &= 2>0 \end{aligned}$$

הנגזרת השנייה חיובית לכל x לכן הפונקציה קמורה.

אולם נשים לב כי עבור $h = f(g(x)) = (x^2-1)^2 - 1$ מתקיים:

$$\begin{aligned} h(x) &= (x^2-1)^2 - 1 \\ &= x^4 - 2x^2 + 1 - 1 \\ &= x^4 - 2x^2 \end{aligned}$$

$$\begin{aligned} h'(x) &= 4x^3 - 4x \\ h''(x) &= 12x^2 - 4 \end{aligned}$$

עבור $x = 0$ נקבל כי:

$$h''(0) = -4 < 0$$

וקיבלנו כי הנגזרת השנייה לא חיובית לכל x. לכן h לא פונקציה קמורה.

3. יהי $f : C \to \mathbb{R}$ פונקציה המוגדרת על קבוצה קמורה C. הוכיחו כי f קמורה אמ"מ האפיגרף שלה הוא קבוצה קמורה, כאשר $\mathbf{epi}(f) = \{(u,t) : f(u) \leq t\}$.

$\Leftarrow$:

נניח כי הפונקציה קמורה ונוכיח כי הפיאגרף שלה הוא קבוצה קמורה.

נניח בשלילה כי ההפיגרף לא קבוצה קמורה. אזי אם ניקח $\begin{pmatrix} u \\ t_2 \end{pmatrix}, \begin{pmatrix} v \\ t_1 \end{pmatrix} \in \text{epi}(f)$, נקבל כי $\alpha \begin{pmatrix} v \\ t_1 \end{pmatrix} + (1-\alpha) \begin{pmatrix} u \\ t_2 \end{pmatrix}$ לא ב$\text{epi}(f)$. פורמלית: $f(\alpha v + (1-\alpha) u) > \alpha t_1 + (1-\alpha) t_2$.

הנחנו כי הפונקציה קמורה לכן היא מקיימת כי לכל $u, v \in C$ ולכל $\alpha \in [0,1]$:

$$\begin{aligned} f(\alpha v + (1-\alpha) u) &\leq \alpha f(v) + (1-\alpha) f(u) \\ &\leq \alpha t_1 + (1-\alpha) t_2 \end{aligned}$$

סתירה.

$\Rightarrow$:

נניח כי הפיגרף של הפונקציה הוא קבוצה קמורה ונוכיח כי הפונקציה קמורה.

אם ההפיגרף זה קבוצה קמורה, אזי, עבור $\begin{pmatrix} u \\ t_2 \end{pmatrix}, \begin{pmatrix} v \\ t_1 \end{pmatrix} \in \text{epi}(f)$, גם $\alpha \begin{pmatrix} v \\ t_1 \end{pmatrix} + (1-\alpha) \begin{pmatrix} u \\ t_2 \end{pmatrix} \in \text{api}(f)$ לכל $\alpha \in [0,1]$.

לכן נובע מכך כי:

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

ומזה נובע כי הפונקציה f פונקציה קמורה.

4. יהי $f_i : V \to \mathbb{R}$, $i \in I$. יהי $f : V \to \mathbb{R}$ המוגדר כך: $f(u) = \sup_{i \in I} f_i(u)$. אם f_i קמורה לכל $i \in I$, אז f גם קמורה.

עבור $u, v \in V$ מתקיים כי לכל $i \in I$:

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

לכן:

$$\begin{aligned} f(\alpha u + (1-\alpha) v) &= \sup_{i \in I} f_i(\alpha u + (1-\alpha) v) \\ &\le \sup_{i \in I} (\alpha f_i(u) + (1-\alpha) f_i(v)) \end{aligned}$$

לכן קיבלנו כי f קמורה.

5. בהינתן $x \in \mathbb{R}^d$ ו־$y \in \{\pm 1\}$. הראו כי הhinge loss קמור בw, b. כלומר, הגדירו:

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

הראו כי f קמור בw, b.

נשים לב כי הפונקציה f היא מקסימום של פונקציות לינאריות.
למדנו כי פונקציה לינארית היא פונקציה קמורה, ולמדנו כי מקסימום של פונקציות קמורות הוא גם קמור, לכן f פונקציה קמורה.

6. הסיקו את הsub gradient של פונקציית hinge loss $g \in \partial \ell^{hinge}_{x,y}(w,b)$.

עבור הנקודה הלא גזירה, $x = 0$, ה־sub gradient הוא פשוט 0. עבור שאר הנקודות שהן גזירות, נקבל כי:

$$\begin{aligned} \partial^w \left(1 - y\left(w^T x + b\right)\right) &= -yx \\ \partial^b \left(1 - y\left(w^T x + b\right)\right) &= -y \end{aligned}$$

ולכן:

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

7. יהי $f_1, ..., f_m : \mathbb{R}^d \to \mathbb{R}$ קבוצה של פונקציות קמורות ו־$g_k \in \partial_k(x)$ לכל $k \in [m]$ הסאב־גרדיאנט של הפונקציות האלו. הגדירו $f : \mathbb{R}^d \to \mathbb{R}$ ע"י $f(x) = \sum_{i=1}^{m} f_i(x)$. הראו כי $\sum_k g_k \in \partial \sum_k f_k(x)$.

לפי לינאריות הנגזרת, $\partial\left(\sum_{i=1}^{m} f_i(x)\right) = \sum_{i=1}^{m} \partial(f_i(x))$.

לכן

$$\partial\left(\sum_{i=1}^{m} f_i\left(x\right)\right)=\sum_{i=1}^{m} g_i$$

ולכן

$$\sum_{k} g_k\in\partial\sum_{k} f_k\left(x\right)$$

8. יהי $S=\{(x_i,y_i)\}_{i=1}^{m}\subseteq\mathbb{R}^d\times\{\pm1\}$ דגימה והגדירו $f:\mathbb{R}^d\to\mathbb{R}$ ע"י:

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

מצאו את הסאב־גרדיאנט של f עבור w כללי.

$$\begin{aligned}\partial^w f&=\partial^w\left(\frac{1}{m}\sum_{i=1}^{m}\ell_{x_i,y_i}^{hinge}\left(w,b\right)+\frac{\lambda}{2}\|w\|^2\right)\\&=\frac{1}{m}\left(-yx+\lambda w\right)\end{aligned}$$

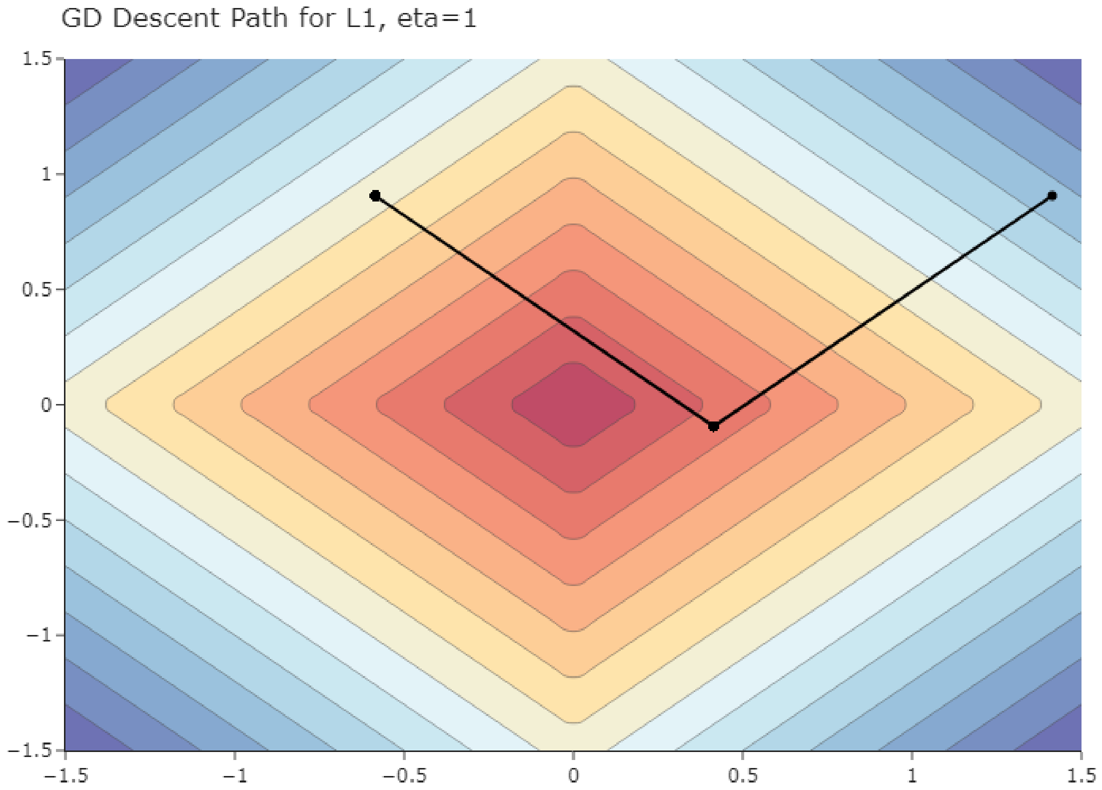
$$\begin{aligned}\partial^b f&=\partial^b\left(\frac{1}{m}\sum_{i=1}^{m}\ell_{x_i,y_i}^{hinge}\left(w,b\right)+\frac{\lambda}{2}\|w\|^2\right)\\&=-\frac{y}{m}\end{aligned}$$

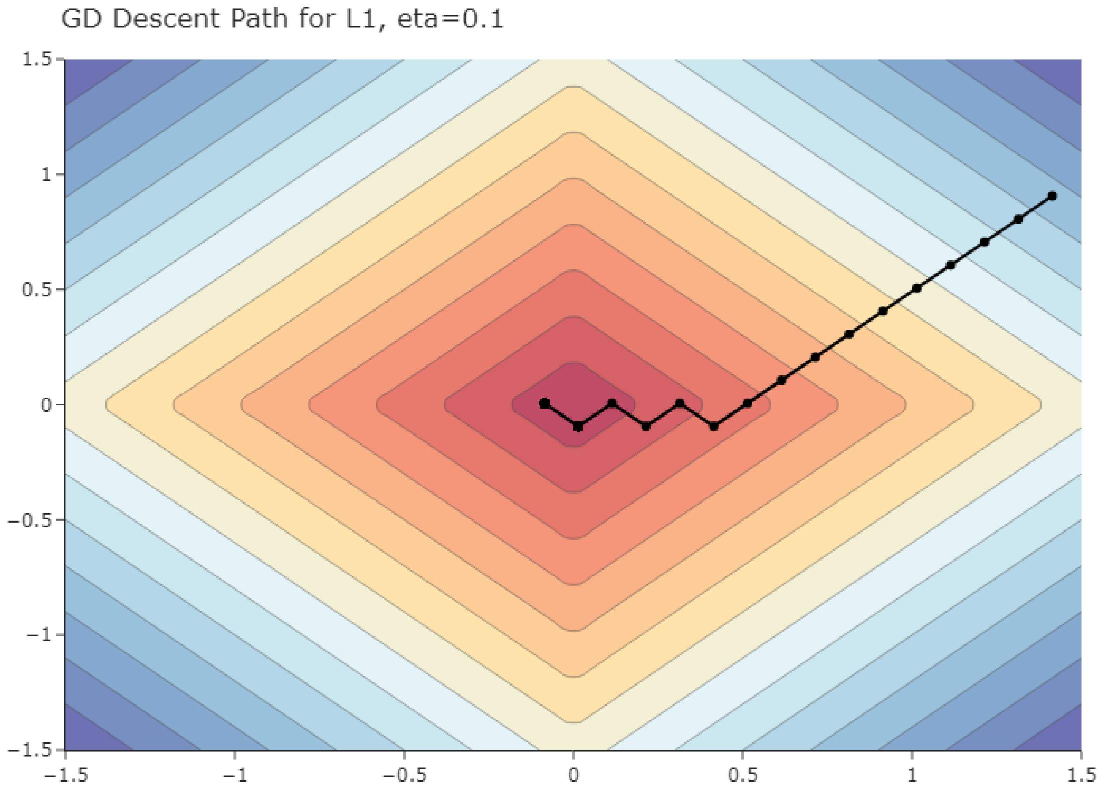
נמזג ביניהם, ונקבל:

$$\frac{1}{m}\sum_{i} g_i+\lambda\left(w,0\right)$$

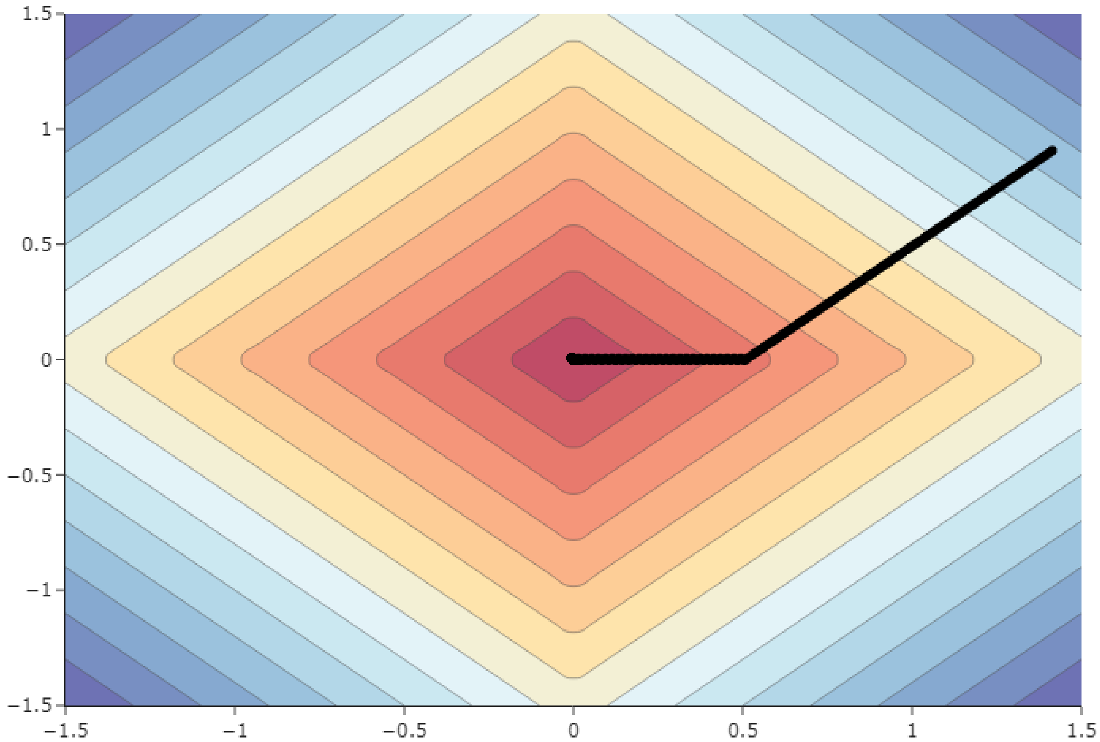
חלק פרקטי

.1

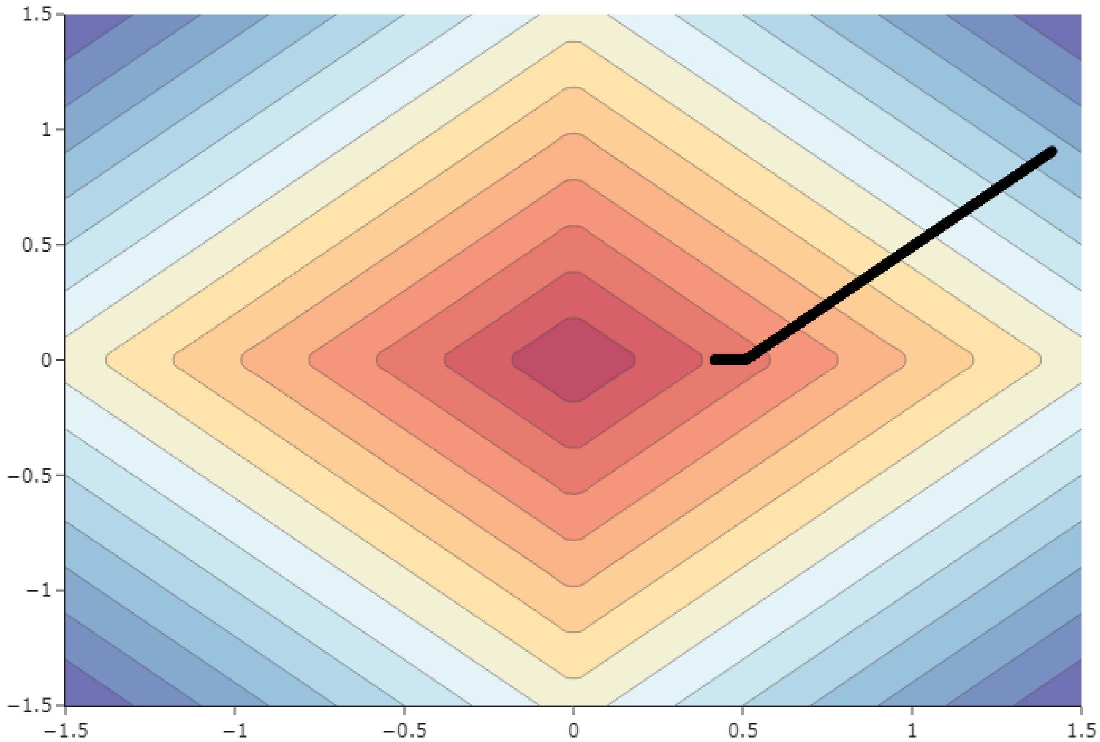

GD Descent Path for L1, eta=1
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5

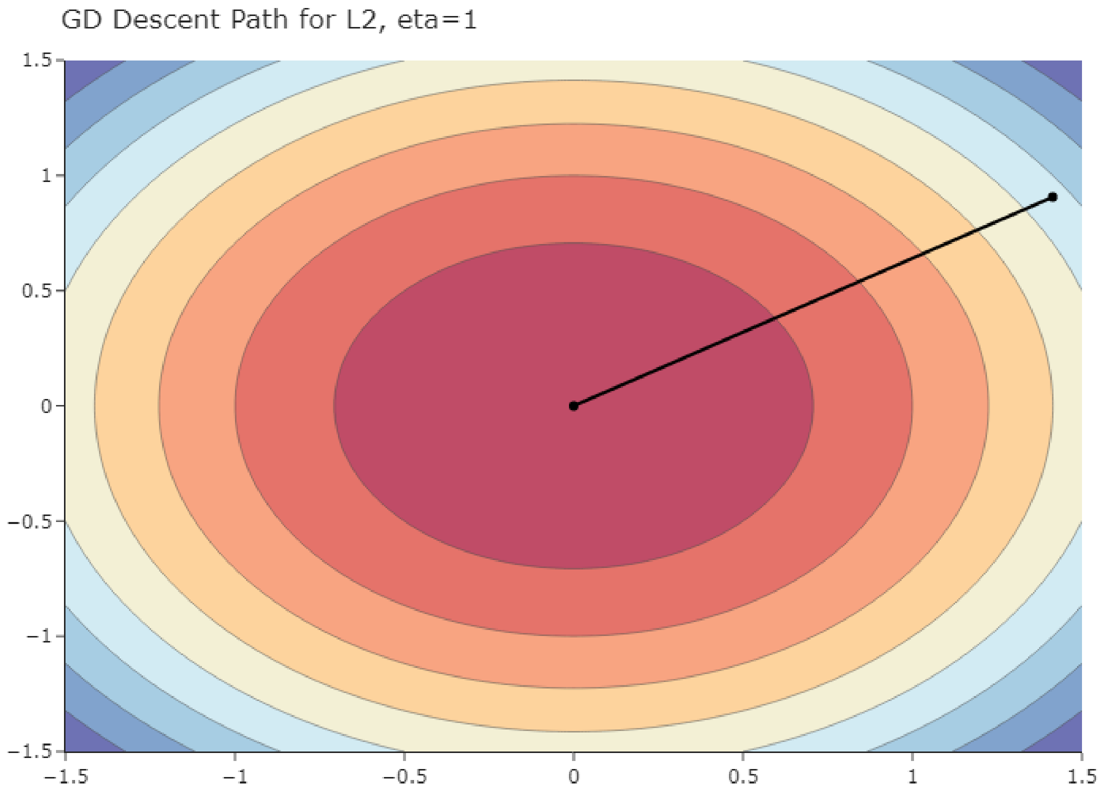

GD Descent Path for L1, eta=0.1
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5

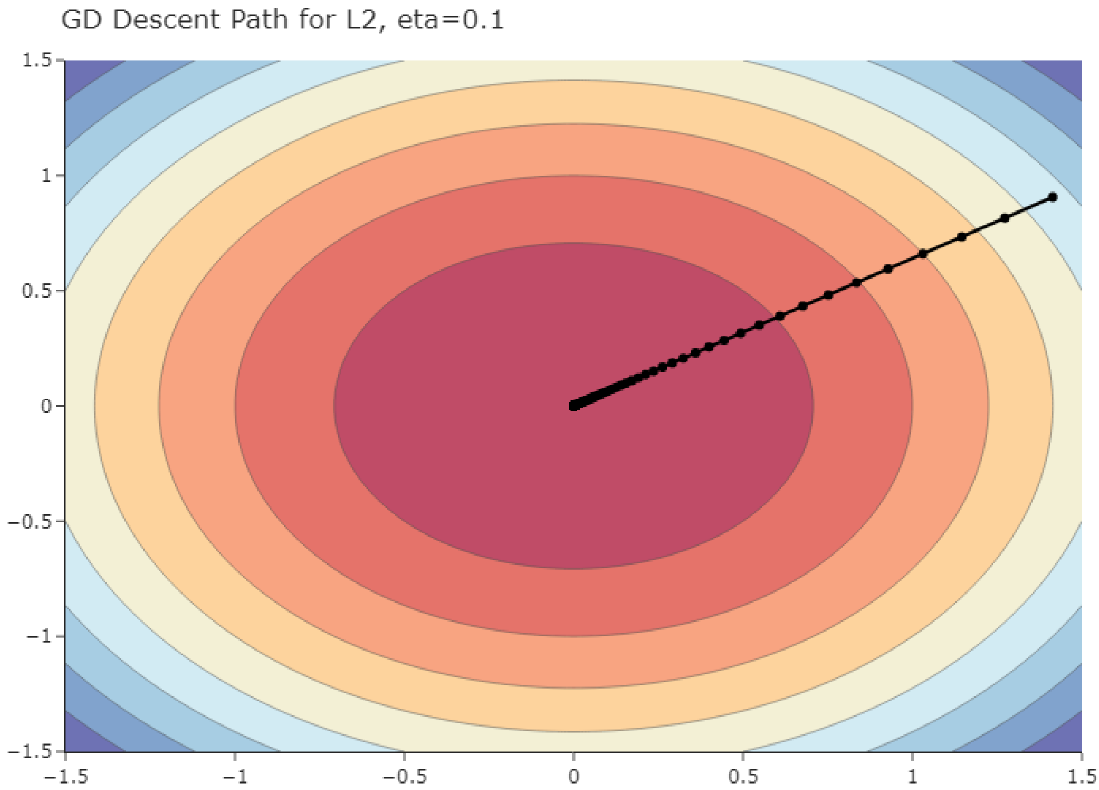
GD Descent Path for L1, eta=0.01



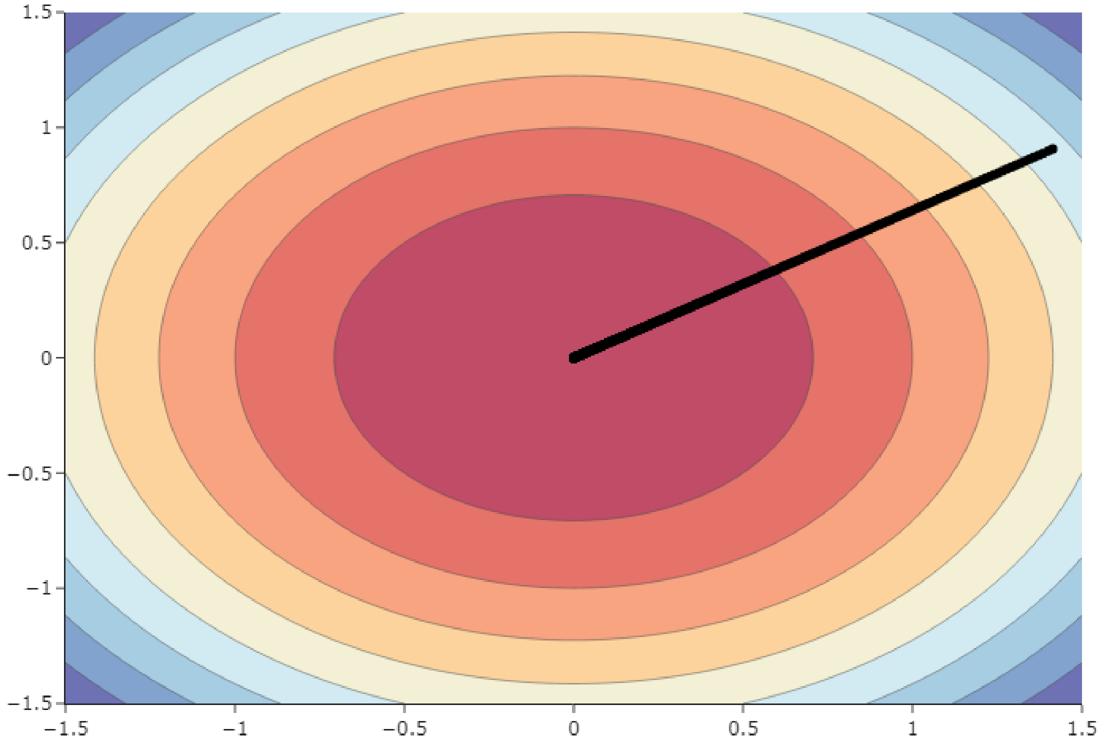
GD Descent Path for L1, eta=0.001



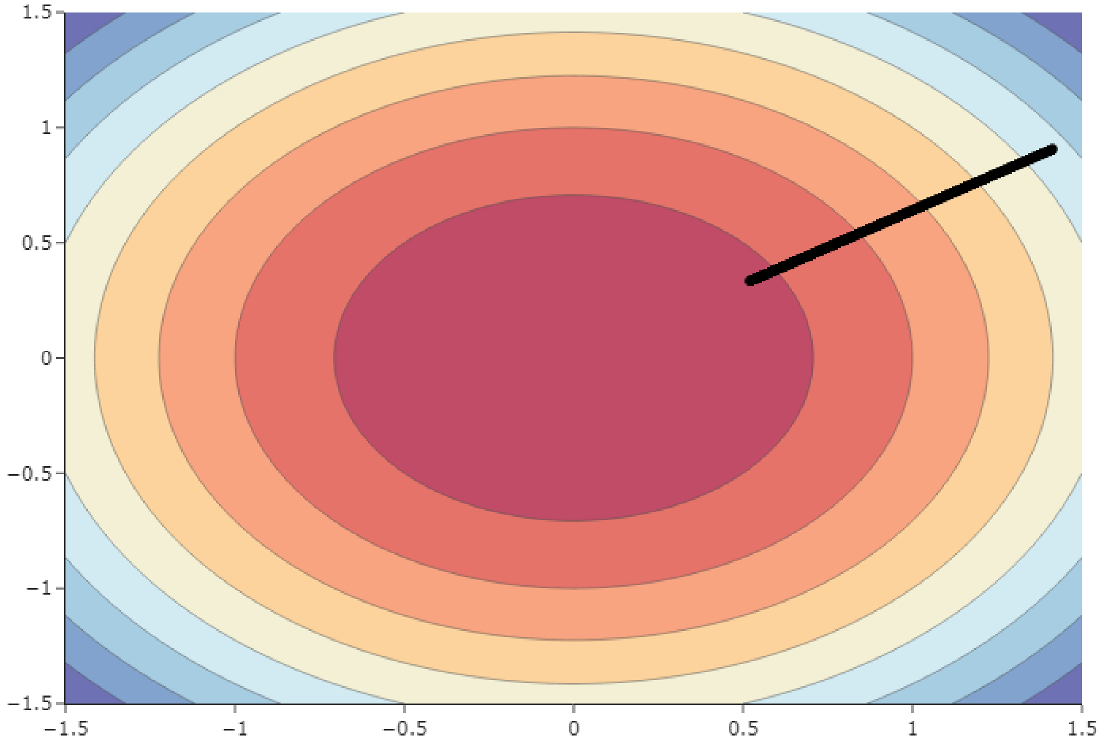

GD Descent Path for L2, eta=1
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5


GD Descent Path for L2, eta=0.1
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5

GD Descent Path for L2, eta=0.01



GD Descent Path for L2, eta=0.001

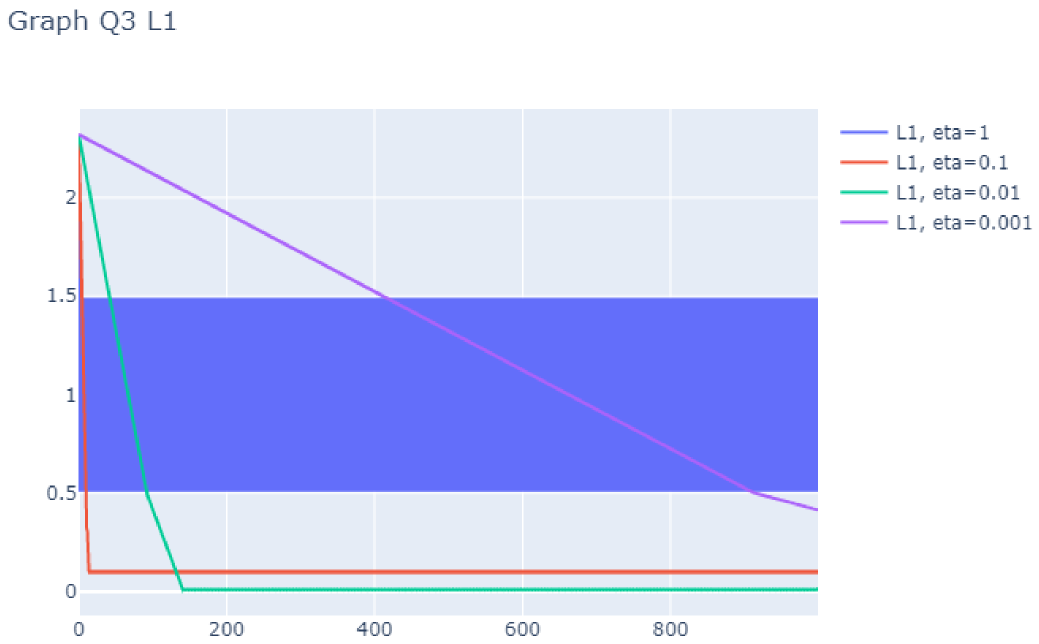


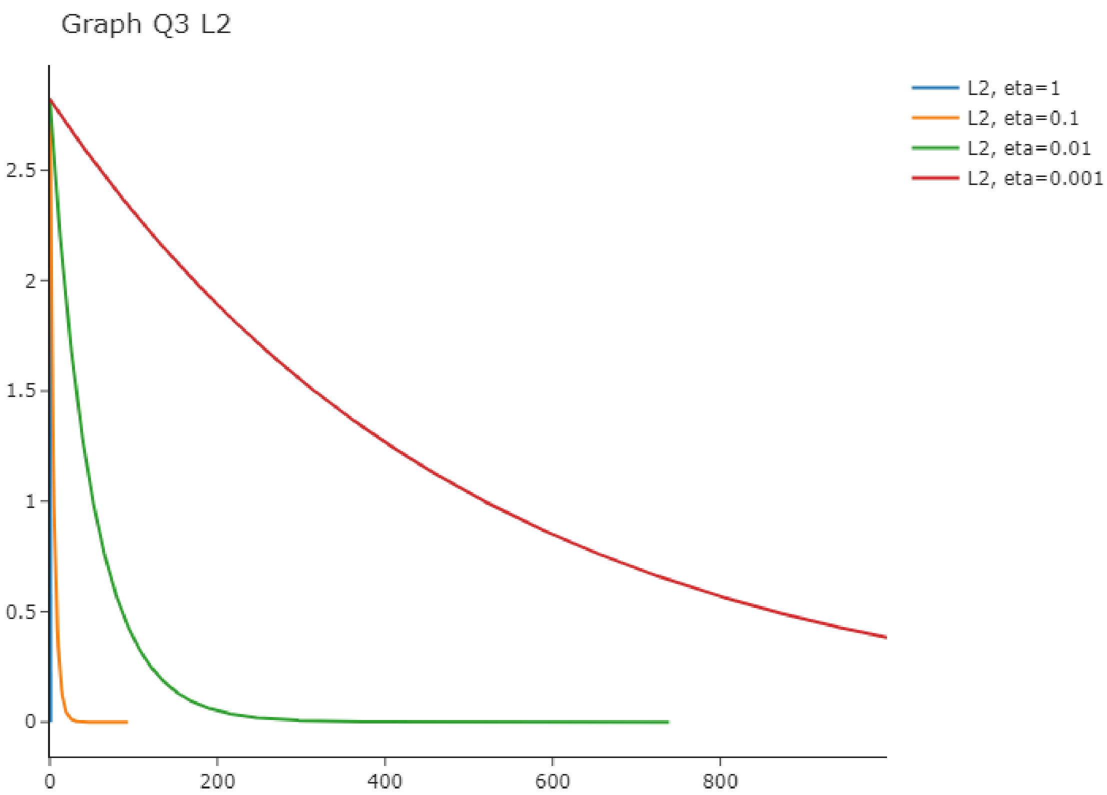
ניתן לשים לב כי הירידה של L1 בשלב מסוים ניהיית חדה יותר, בעוד שב L2 הירידה היא אלכסונית ללא עיקולים. ניתן להבין כי זה בדיוק כי נורמה 1 עובדת עם ערכים מוחלטים בעוד שנורמה 2 היא ריבועית לכן יותר חלקה.

.2

ניתן לראות כי בקצב ירידה 1 הקפיצה היא גדולה מאוד, אין כמעט צעדים.

.3

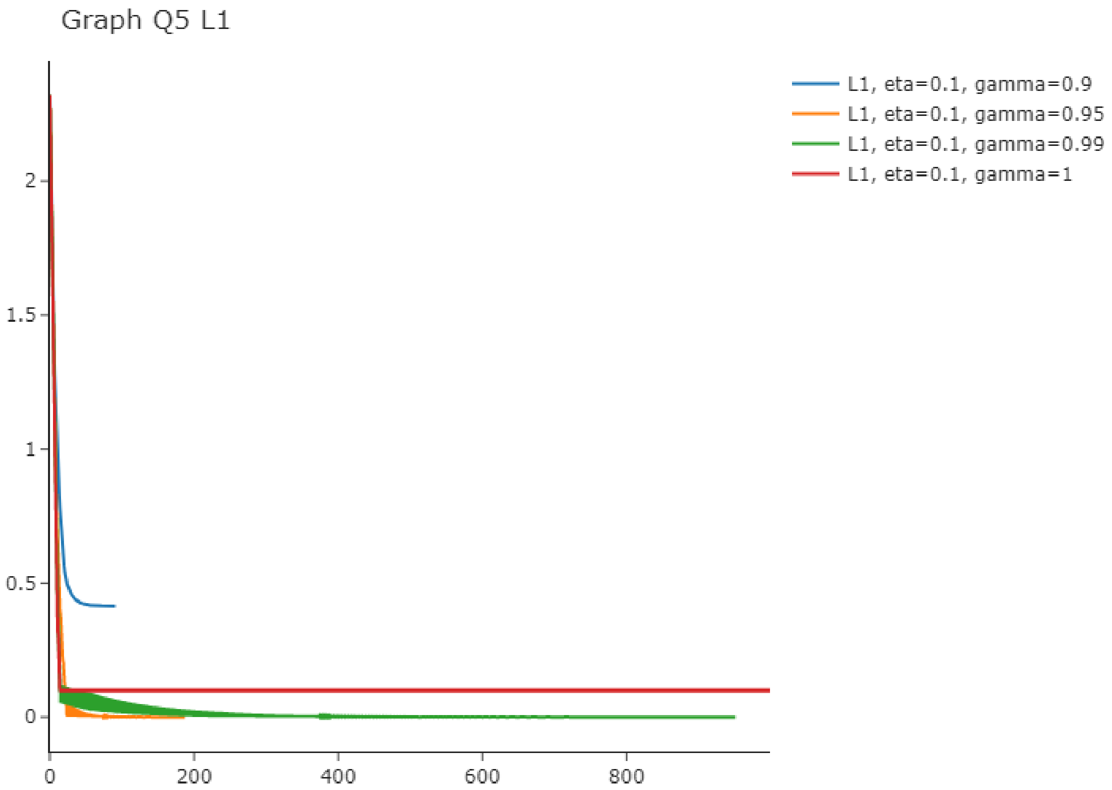




בגרף של נורמה 1 ניתן לראות כי ככל שמתקדמים לערך קטן יותר השיפוע יותר איטי ומגיע לערך קטן יותר, למעט הגרף של eta=0.001 שבה כנראה זה מספר קטן מידי ופחות מגיע לדיוק כי לוקח לו יותר זמן להגיע, לכן לא רואים בגרף.

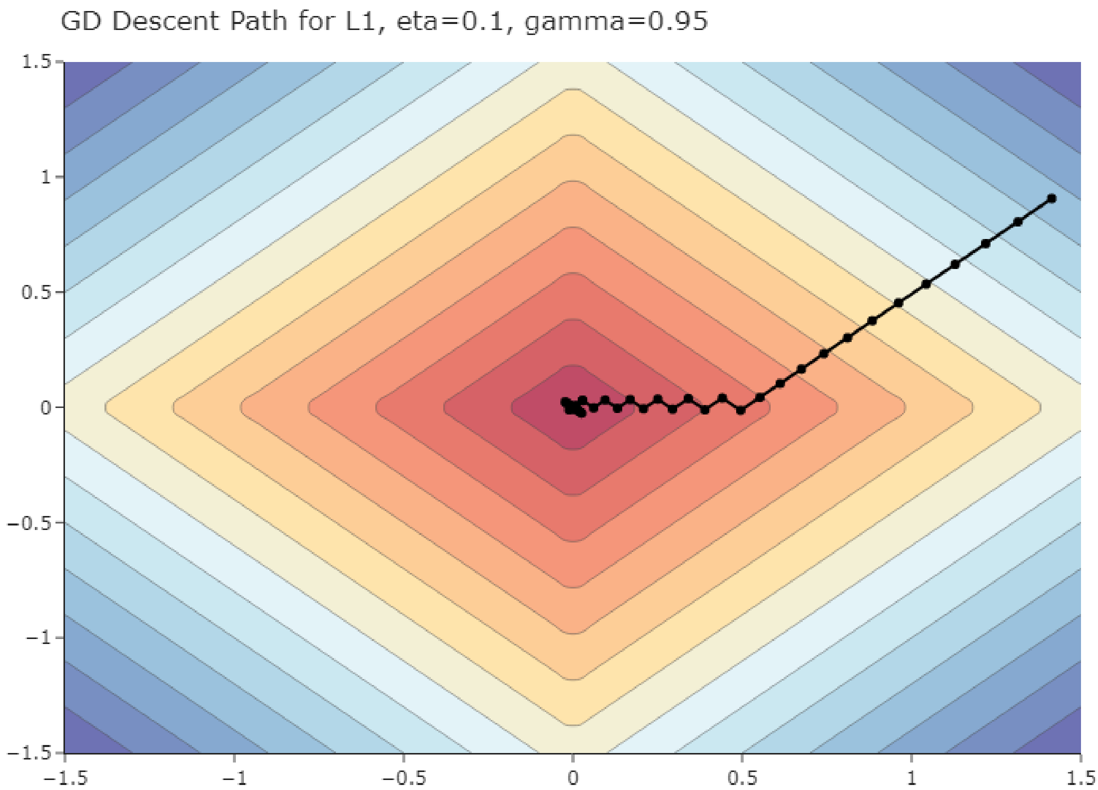
בגרף של נורמה 2 ניתן לראות מגמה דומה.

.5

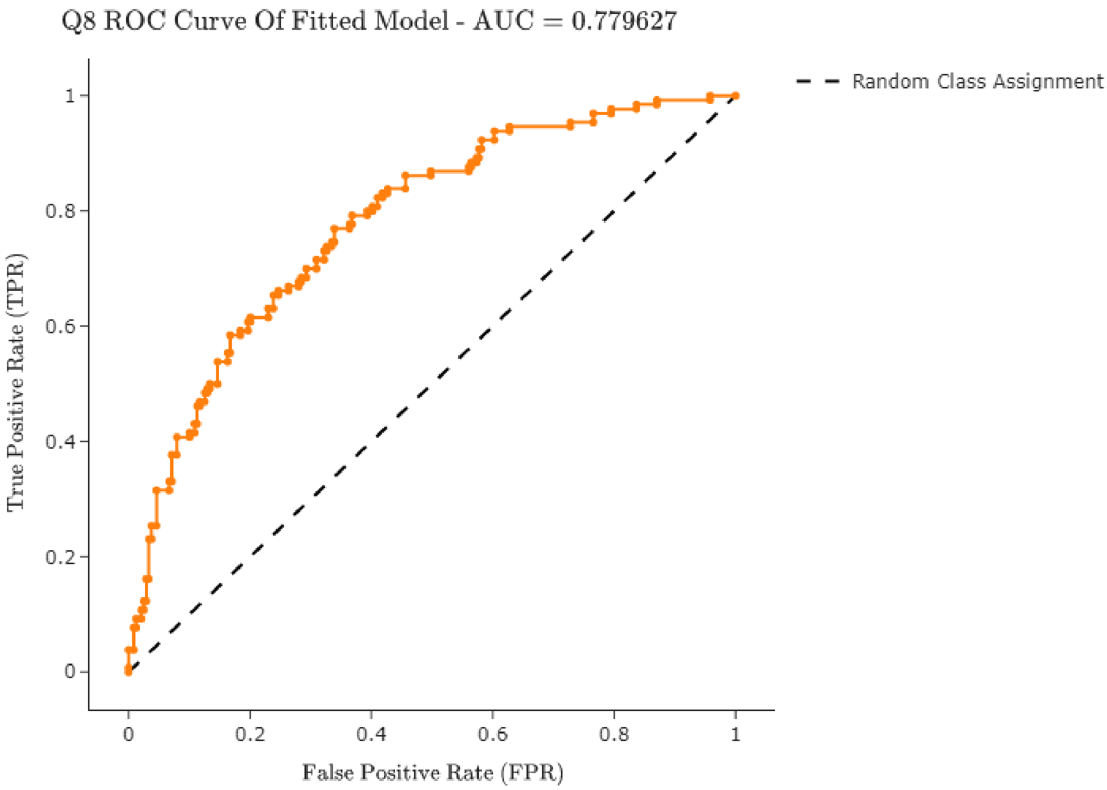


אנו רואים שככל שגמא גדל, אנו נהיים מדויקים, עד שיש bias משמעותי לא קטן מ-1, ואז פחות נהיים מדויקים.

.6



.8



.9

האלפא המקסימלי המתקבל בנוסחה הוא 0.867

.10

For L1, the minimum lambada is 0.001, its test error is 0.352

For L2, the minimum lambada is 0.001, its test error is 0.352