<u>1.1.1</u>

נשים לב כי g היא פונקציה מ-C ל-R. כמו כן, יהיו $v,u\in C$, $\alpha\in[0,1]$

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

(f_i פונ' קמורה)

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

מכאן g פונקציה קמורה.

<u>1.1.2</u>

$f(x)=|x|,\ g(x)=x^2-1$

(מחזירות "1" הן זוגיות, כל פסל ייחודי במקרה הזה ⇐)

$$\Rightarrow h(x)=|x^2-1|$$

יהיו $\alpha\in[0,1]$, $u,v\in R$ נשים לב כי

$$f(\alpha v+(1-\alpha)u)=|\alpha v+(1-\alpha)u|\le\alpha|v|+(1-\alpha)|u|=\alpha f(v)+(1-\alpha)u\Rightarrow f \text{ קמורה}$$

(אי"ש המשולש)

בנוסף נשים לב כי $g''(x)=2>0$ ולכן מטענה 2.3 בהרצאה f, g קמורות.

כמו כן, יהיו $u,v=-1,1$ ו-$\alpha=\frac{1}{2}$.

$\Leftarrow$ $\alpha\cdot v=-\frac{1}{2}$, $(1-\alpha)u=\frac{1}{2}$, $h(v)=|1-1|=0$, $h(u)=|1-1|=0$.

$\Leftarrow$ $\alpha h(v)+(1-\alpha)h(u)=\underline{0}$, $h(\alpha v+(1-\alpha)u)=h(-\frac{1}{2}+\frac{1}{2})=h(0)=\underline{1}$

$\Leftarrow$ $h(\alpha v+(1-\alpha)u)=1>0=\alpha h(v)+(1-\alpha)h(u)$

$\Leftarrow$ h <u>לא</u> פונקציה קמורה.

1.2.3

נוכיח כי max של שתי פונקציות קמורות היא פונקציה קמורה.

יהיו $f,g: C\to R$ פונקציות קמורות. יהיו $u,v\in C$, $\alpha\in[0,1]$.

$$\max\{f(\alpha v+(1-\alpha)u),\ g(\alpha v+(1-\alpha)u)\}\le \max\{\alpha f(v)+(1-\alpha)f(u),\ \alpha g(v)+(1-\alpha)g(u)\}$$

$$\le \alpha\cdot\max\{f(v),g(v)\}+(1-\alpha)\max\{f(u),g(u)\}$$ ■

ונראה שהפונקציות $1-y(x^Tw+b)$ ו-0 הן קמורות באופן פשוט כי הן אפיניות.

<u>הוכחה</u>: תהי $f: C\to R$ פונ' אפינית, $u,v\in C$, $\alpha\in[0,1]$.

נרשום את f בצורה כללית ע"י $f(a)=Aa+b$:

$$f(\alpha v+(1-\alpha)u)=A(\alpha v+(1-\alpha)u)+b\le \alpha A v+b+(1-\alpha)Au+b=\alpha f(v)+(1-\alpha)f(u)$$

לכן כל פונ' אפינית היא קמורה.

סה"כ קיבלנו כי $\max\{0,1-y(x^Tw+b)\}$ הינה הפעלת max על שתי פונקציות קמורות, ולכן פונקציה קמורה כנדרש ■

1.2.4

$$\frac{d}{dw}\left(1-y(x^Tw+b)\right)=-yx,\quad \frac{d}{db}\left(1-(y(x^Tw+b))\right)=-y$$

נחלק לשלושה מקרים. לשם נוחות נסמן $h(w,b)=\max\{0,1-y(x^Tw+b)\}$.

1) $1-y(x^Tw+b)<0 \Rightarrow h(w,b)=0$ ← בסביבה זו h קבועה ב-0 ולכן $g=(0,0)$

2) $1-y(x^Tw+b)>0 \Rightarrow$ בסביבה זו $h(w,b)$ גזירה ולכן $g=(-yx,-y)$.

3) $1-y(x^Tw+b)=0 \Rightarrow$ h אינה גזירה בנקודה זו, אבל הסאב-גרדיאנטים שלה הם כל הפונקציות בתחום "בין" 0 — לבין הגזירה בסביבת הנקודה — ורק "בין" 1-, בהתאם לשיפוע של $h(w,b)$ בסביבת נקודה זו, בכל אחד מהכיוונים.

לכן: $g=(-\lambda yx,-\lambda y)$, $\lambda\in[0,1]$.

סה"כ נקבל ⇐

$$g=\begin{cases}0, & 1-y(x^Tw+b)<0\\ (-yx,-y), & 1-y(x^Tw+b)>0\\ (-\lambda yx,-\lambda y), & 1-y(x^Tw+b)=0\end{cases}$$

1.2.5

יהי $x,u\in\mathbb{R}^d$. לכל $i\in[m]$ מתקיים $g_i\in\partial f_i(x)$. לכן

$$\forall i\in[m]\quad f_i(u)\ge f_i(x)+\langle g_i,u-x\rangle \Rightarrow \underbrace{\sum_{i=1}^{m} f_i(u)}_{=f(u)} \ge \sum_{i=1}^{m}\left(f_i(x)+\langle g_i,u-x\rangle\right)$$

$$=\sum_{i=1}^{m} f_i(x)+\sum_{i=1}^{m}\langle g_i,u-x\rangle \underset{\text{לינאריות מכפלה פנימית}}{=} \sum_{i=1}^{m} f_i(x)+\left\langle \sum_{i=1}^{m} g_i,u-x\right\rangle \underset{\text{הגדרת } f(x)}{=} f(x)+\left\langle \sum_k g_k,u-x\right\rangle$$

$$\Rightarrow f(u)\ge f(x)+\left\langle \sum_k g_k,u-x\right\rangle \Rightarrow \sum_k g_k\in\partial\sum_k f_k(x) \quad \blacksquare$$

1.2.6

כאמור, נזכיר כי מתקיים מהרצאה ש- $\partial(f_1+f_2)=\partial f_1+\partial f_2$.

כמו כן גם כי $\partial(\alpha\cdot f)=\alpha\cdot\partial f$ לכל $\alpha>0$.

בנוסף ראינו כי אם $f:\mathbb{R}^d\to\mathbb{R}$ פונקציה גזירה ב- x, אזי $\partial f(x)=\{\nabla f(x)\}$.

מכיוון ש- $\frac{\lambda}{2}\|w\|^2$ גזירה לכל $w\in\mathbb{R}^d$, אזי $\partial\left(\frac{\lambda}{2}\|w\|^2\right)=\{\lambda(w,0)\}$.

← [illegible]

מצאנו בסעיף 1.2.4 תת-גרדיאנט של $\ell^{hinge}_{x,y}(w,b)$, סימנו אותו g.

באותה צורה נסמן את תת-גרדיאנטים המתאימים של $\ell^{hinge}_{x_i,y_i}(w,b)$ ב- g_i.

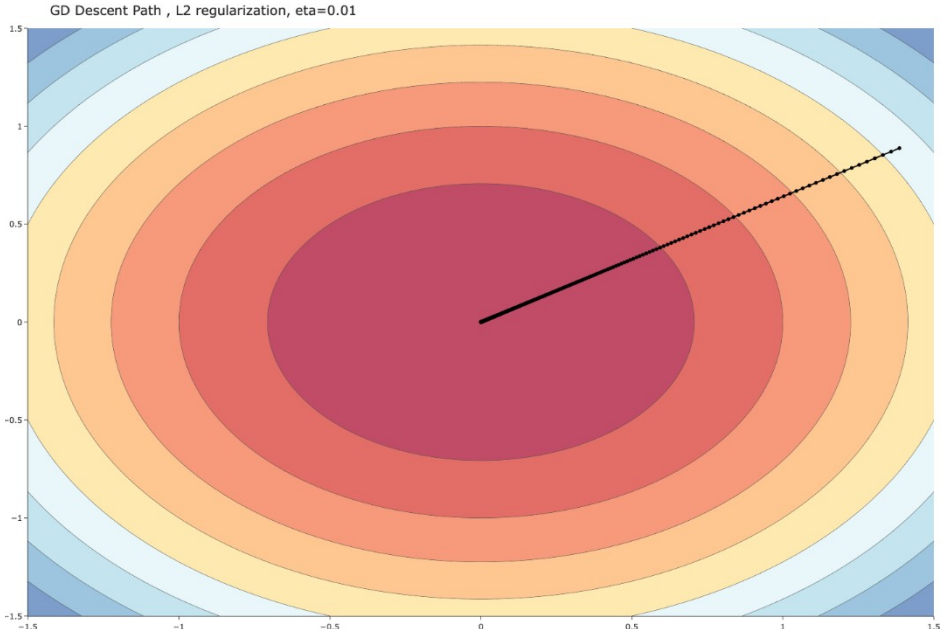
לפי סעיף 1.2.5 נקבל כי $\sum_{i=1}^{m} g_i\in\partial\sum_{i=1}^{m}\ell^{hinge}_{x_i,y_i}(w,b)$

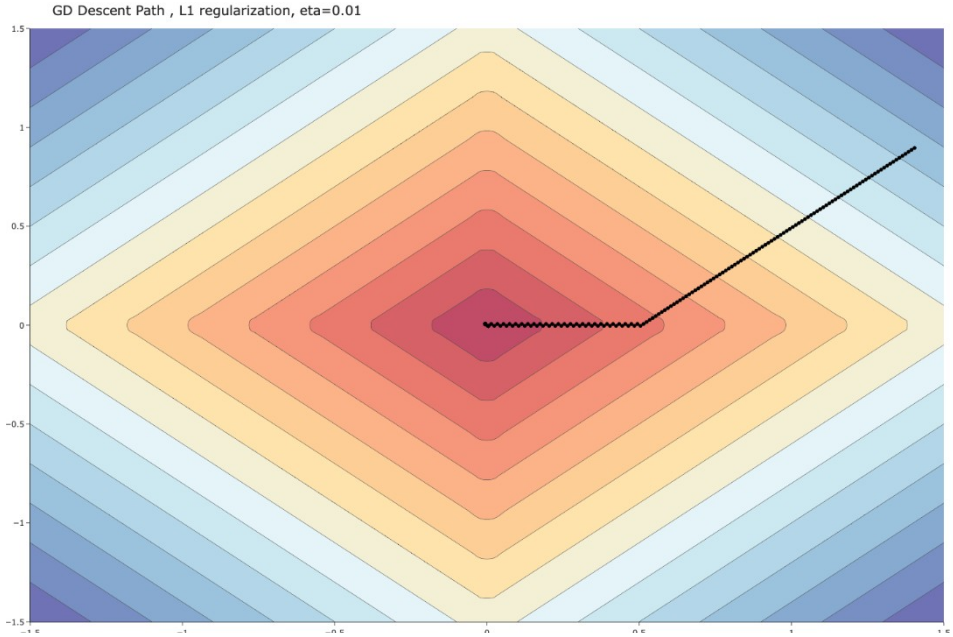
לכן סה"כ קיבלנו כי $\partial f(w,b) \ni \left\{\lambda(w,0)+\frac{1}{m}\sum_{i=1}^{m} g_i \,\middle|\, w\in\mathbb{R}^d,\ \forall i\in[m]\ g_i\in\partial\left(\ell^{hinge}_{x_i,y_i}(w,b)\right)\right\}$

Practical part

2.1.1

1



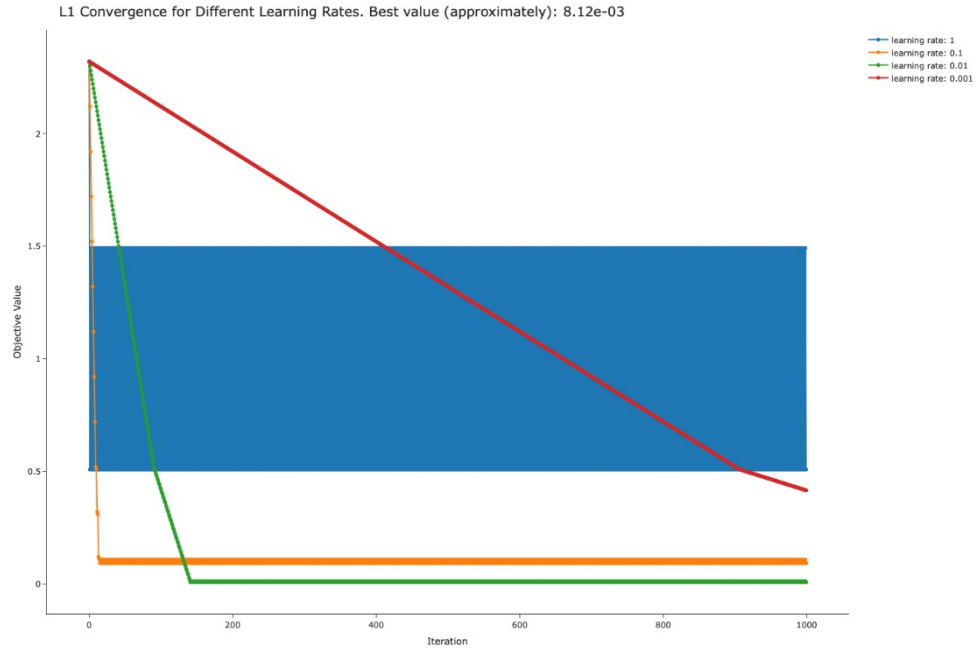


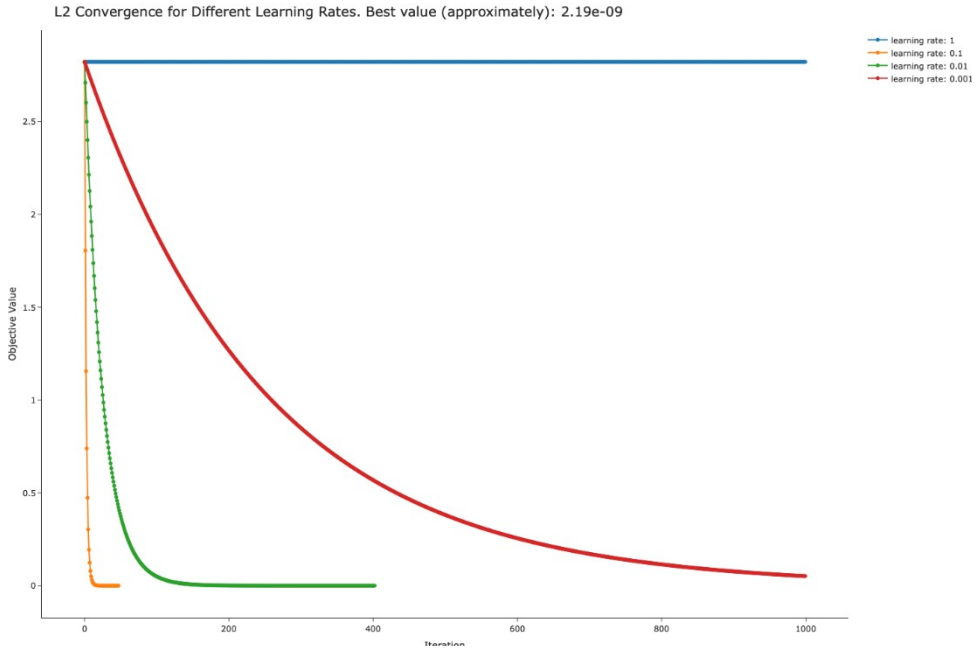
In the L1 objective, the descent path does not follow a straight line toward the minimum, whereas in the L2 objective it does. This difference arises because gradient descent moves in the direction of steepest descent at each step, which is influenced by the geometry of the level sets of the objective. For the L1 objective, level sets are diamond-shaped, so the distance to the minimum varies along the boundary of each level set. As a result, the direction of descent often shifts sharply, causing a more "zig-zag" path. In contrast, the L2 objective has circular level sets centered around the minimum, and the gradient at each point consistently points toward the center, resulting in a smoother and more .direct path

2

In the L1 objective, the descent path often changes direction sharply, switching from a diagonal to axis-aligned movement. Near convergence, the algorithm may oscillate between two points due to .the non-smoothness of the objective

3





L2 - The convergence curves are smooth across all learning rates, which is expected due to the differentiability and smooth curvature of the L2 objective. When using a learning rate of 1 with the given initialization, gradient descent fails to converge and oscillates between two points. In fact, for any learning rate greater than 1, the algorithm diverges entirely. For learning rates below 1, higher .values result in faster convergence

L1 – As with L2, a learning rate of 1 lead to divergence. This happens because the sub gradient at

non-differentiable points in the L1 objective can change direction abruptly, which combined with a large step size causes the algorithm to bounce or overshoot the minimum. The convergence plots for L1 also show “kinks” or sharp bends — a result of the non-smooth geometry of the L1 level sets. Despite this, when convergence does occur, higher learning rates still lead to faster convergence .overall

4

L1 objective best loss ≈ 8.12e-3
L2 objective best loss ≈ 2.19e-9
The large gap in final loss values is since the sub gradient of the L1 objective does not shrink as the weights approach the minimum. Combined with a fixed learning rate, this causes gradient descent to stop improving once it gets close enough. In contrast, the gradient of the L2 objective decreases with the weight size, allowing the algorithm to continue making smaller and more precise updates, even .with the same fixed step size

5+6+7