הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ג – 16 ביולי 2023

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד א' – פתרון חלקי

שימו לב: הפתרונות המופיעים כאן הם חלקיים בלבד ומובאים בשביל לעזור לכם בתהליך הלמידה.

ייתכנו כאן חוסרים / ליקויים / טעויות של ממש.

בהצלחה!

שאלה 1: Feature Selection [8 נק']

סמנו את התשובות המתאימות (לפי ההוראות). אין צורך לכתוב הסברים. סימונים לא ברורים יובילו לפסילת התשובה.

נתון ה-dataset הבא.

sample/feat.	$x[1]$	$x[2]$	$x[3]$	y
#1	1	-1	1	-1
#2	-1	1	1	-1
#3	1	-1	1	-1
#4	-1	1	1	-1
#5	1	1	-1	$+1$
#6	1	1	1	$+1$

נפעיל sequential feature selection (כפי שהגדרנו בתרגיל הבית) בעזרת אלגוריתם בסיס $\mathcal{A}$ שפותר ERM ומוצא מסווג ליניארי לא-הומוגני עם שגיאת 0-1 מינימלית על הנתונים. כאן – נמצא שני פיצ'רים.

אילו פיצ'רים ייבחרו בתהליך Forward Selection?

a. $x[1], x[2]$.

b. $x[1], x[3]$.

c. $x[2], x[3]$.

d. מהנתונים הקיימים, ניתן לענות רק לגבי פיצ'ר יחיד.

e. מהנתונים הקיימים, לא ניתן לענות כלל.

אילו פיצ'רים ייבחרו בתהליך Backward Selection?

a. $x[1], x[2]$.

b. $x[1], x[3]$.

c. $x[2], x[3]$.

d. מהנתונים הקיימים, ניתן לענות רק לגבי פיצ'ר יחיד.

e. מהנתונים הקיימים, לא ניתן לענות כלל.

שאלה 2: Capacity של מחלקות שונות [21 נק']

נתונה התפלגות $\mathcal{D}$ על דאטה עם $d \geq 10$ פיצ'רים בינאריים ($\mathcal{X} = \{0,1\}^d$) ותיוגים בינאריים ($\mathcal{Y} = \{-1,+1\}$).

ידוע שההתפלגות $\mathcal{D}$ נותנת לכל דוגמה אפשרית $\mathbf{x} \in \mathcal{X}$ הסתברות כלשהי סופית גדולה ממש מאפס.

ידוע שהתיוג האמיתי של דוגמה כלשהי הוא חיובי אם ורק אם בדוגמה יש לפחות שני פיצ'רים כלשהם "דולקים" (שערכם 1).

לדוגמה: כאשר $d = 5$, לדוגמאות $(0,1,1,0,1)$ ו-$(1,1,0,0,0)$ יש בהכרח תיוג חיובי, ולדוגמה $(0,0,1,0,0)$ בהכרח תיוג שלילי.

עבור כל אחת מהמחלקות הבאות, נבדוק האם ניתן ליצור בעזרתה מסווג שיגיע לשגיאת **הכללה** אפס.

שימו לב: זו שאלה על מחלקות היפותזות, ולא על אלגוריתמי למידה ספציפיים.

א. האם קיים עץ החלטה בינארי (צומת שאינו עלה יכול להתפצל רק לשני צמתים) שמגיע לשגיאת הכללה אפס?

סמנו: כן / לא

אם קיים – מהו עומק העץ המינימלי הנדרש (בסדר גודל, למשל $\Theta(1), \Theta(\ln d), \Theta(d), \Theta(2^d)$ וכו')? הסבירו בקצרה.
אחרת – הסבירו באופן מפורט למה לא.

תשובה (לרשותכם דפי טיוטה בסוף הגיליון):

נדרש עץ עם $\Theta(d)$ רמות.

בשורש שואלים לגבי פיצ'ר 1. אם דלוק – ברמה הבאה שואלים אם פיצ'ר 2 דלוק. אם כן – חוזים +1.

אחרת, שואלים אם פיצ'ר 3 דלוק וכן הלאה.

נדרשות לכל היותר d שאלות כדי להבין אם התחזית היא חיובית או שלילית.

שימו לב: הסעיף הבא עוסק במסווגים ליניאריים ומתקיים בהם $\mathrm{sign}(\mathbf{w}^\top\mathbf{x}) = \begin{cases} -1, & \mathbf{w}^\top\mathbf{x} < 0 \\ 0, & \mathbf{w}^\top\mathbf{x} = 0 \\ +1, & \mathbf{w}^\top\mathbf{x} > 0 \end{cases}$.
משמע, margin של אפס בהכרח גורר שגיאת סיווג.

ב. האם קיים מפריד ליניארי לא הומוגני שמגיע לשגיאת הכללה אפס?   **סמנו: כן / לא**

אם קיים – הציעו וקטור $\mathbf{w} \in \mathbb{R}^d$ וסקלר $b \in \mathbb{R}$ שמגיעים לשגיאה אפס (במקרה כזה לא נדרש הסבר נוסף).

אחרת – הסבירו באופן מפורט למה לא.

תשובה:

$$\mathbf{w} = [1,1,\dots,1], \qquad b = -1.1$$

ג. האם ניתן ליצור מסווג 1-Nearest-Neighbor (עם מטריקת ℓ_2 האוקלידית) שמגיע לשגיאת הכללה אפס?
סמנו: כן / לא

(שימו לב, "ליצור מסווג 1NN" דורש למעשה ליצור אוסף מתאים של דוגמאות אימון.)

אם ניתן – מה הגודל המינימלי של סט אימון שנדרש ליצירת מסווג כזה (בסדר גודל, למשל $\Theta(1), \Theta(\ln d)$ וכו')?
הסבירו בקצרה.

אחרת – הסבירו באופן מפורט למה לא.

תשובה:

צריך את כל הווקטורים הסטנדרטיים $\{\mathbf{e}_i\}_{i\in[d]}$ שמתוייגים שלילית.

אח"כ צריך את כל הדוגמאות שבהן בדיוק 2 פיצ'רים דולקים (מתוייגות חיובית). יש $\binom{d}{2}$ כאלה.

כל דוגמה עם 2 פיצ'רים דולקים ויותר תהיה יותר קרובה לדוגמה חיובית עם 2 פיצ'רים דולקים

מאשר לדוגמאות השליליות. כל דוגמה עם פיצ'ר דולק ומטה, תהיה יותר קרובה לדוגמה שלילית.

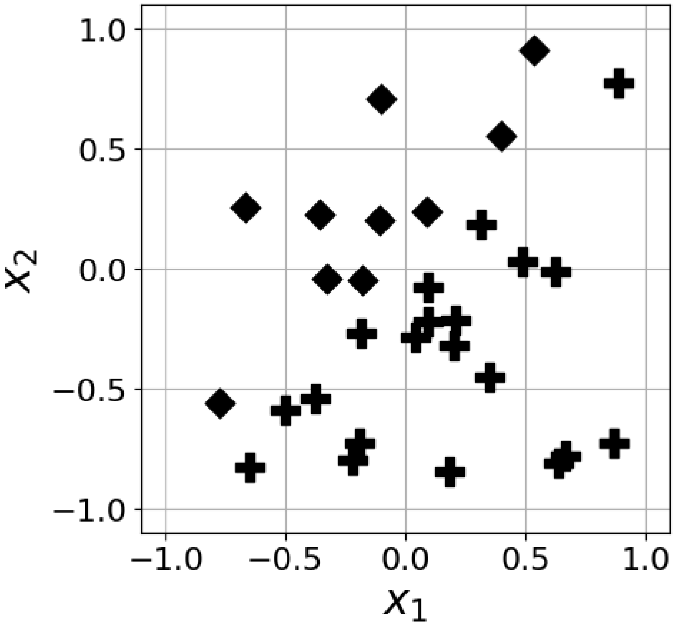
בסה"כ צריך $\Theta(d^2)$ דוגמאות באוסף.

הערה: זה לא משנה אם דוגמה "נחשבת שכנה של עצמה" או לא. השאלה עוסקת בשגיאת הכללה ואילו העניין שנק' היא שכנה של עצמה או לא רלוונטי רק ל-training set.

שאלה 3: AdaBoost ופרספטרון [21 נק']

משמאל מופיע אלגוריתם AdaBoost.

מימין סט אימון עם $m = 30$ דוגמאות בדו-ממד ($\mathcal{X} = \mathcal{R}^2$) עם תיוגים בינאריים ($+1$ מסומן ב-✚ ו-(-1) מסומן ב-◆).



Initialize $D^{(1)} = \left(\frac{1}{m}, \dots, \frac{1}{m}\right)$

For $t = 1, \dots, T$:

$$h_t = \mathcal{A}\left(S, D^{(t)}\right)$$

$$\epsilon_t = \sum_i D_i^{(t)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i}$$

$$\alpha_t = \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right)$$

$$D_i^{(t+1)} = \frac{1}{Z_t} D_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right)$$

$$h_s(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

א. [5 נק'] כאשר ניתן למקבל הרצות נפרדות של $\mathcal{A}$ (עד T הרצות במקביל), מה פקטור השיפור שניתן לקבל בזמן ההרצה של AdaBoost? הסבירו בקצרה את תשובתכם.

תשובה: המיקבול לא יעזור.

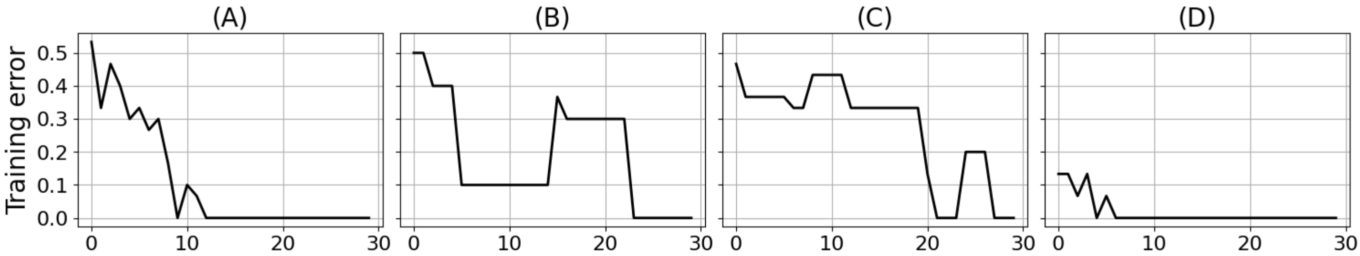
זה אלגוריתם סדרתי. בכל איטרציה צריך להשתמש בהתפלגות שחושבה באיטרציה הקודמת.

ב. [5 נק'] מצאו $\mathbf{w} \in \mathbb{R}^2$ שיוצר מפריד ליניארי הומוגני עם שגיאת אימון 0 על הדאטא (משמע, $\text{sign}(\mathbf{w}^\top \mathbf{x}_i) = y_i$).

כתבו את שני הערכים במסגרת: $\mathbf{w} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

(ניתן להבין את התשובה על סמך התרשים; בבדיקת התשובה יינתן מרווח שגיאה סביר בערכים המספריים עצמם.)

ג. [11 נק'] לפניכם ארבע עקומות התכנסות שונות של שגיאת ה-0-1 על הנתונים שבתרשים הקודם.



צירי y בתרשימים מראים את שגיאת האימון על כל 30 דוגמאות האימון.

נתאים **שניים** מהתרשימים לאלגוריתמי הלמידה הבאים:

i. אלגוריתם פרספטרון עם גודל צעד $\eta = 1$ ואתחול $\mathbf{w}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

ציר x בתרשים המתאים מראה 30 איטרציות של האלגוריתם (בכל איטרציה האלגוריתם רואה דוגמה בודדת).

הקיפו את האות של התרשים ששייך לאלגוריתם זה. **(A)** ***(B)*** **(C)** **(D)**

הסבירו בקצרה את תשובתכם.

הסבר: ברגע שפרספטרון מגיע לשגיאת אימון אפס הוא לא מעדכן יותר את הפיתרון.

כמו כן אנחנו מצפים שיהיו הרבה איטרציות שבהן אין עדכון (ויתקבלו איזורים שטוחים בעקומה).

ii. אלגוריתם AdaBoost כאשר $\mathcal{A}$ הוא אלגוריתם ERM ללמידת Decision Stump (עץ החלטה בעומק 1) ביחס לדוגמאות ולמשקלים הנתונים.

ציר x בתרשים המתאים מראה 30 איטרציות של האלגוריתם (כפי שמופיע בתחילת השאלה).

הקיפו את האות של התרשים ששייך לאלגוריתם זה. **(A)** **(B)** **(C)** ***(D)***

הסבירו בקצרה את תשובתכם.

הסבר: רואים על הדאטה שבאיטרציה הראשונה ייבחר dec. stump שעושה 4-5 טעויות מתוך 30.

שאלה 4: מסווגים ליניאריים ואופטימיזציה [25 נק']

בכל השאלה נתון מדגם $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ (כאשר $\mathcal{X}=\mathbb{R}^d$ ו-$\mathcal{Y}=\{-1,+1\}$) ונניח שהוא <u>פריד ליניארית הומוגנית</u>.

נגדיר בעיית למידה קמורה בעזרת ה-exponential loss:

$$\operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d} \underbrace{\sum_{i=1}^m \exp(-y_i\mathbf{w}^\top\mathbf{x}_i)}_{\triangleq\mathcal{L}(\mathbf{w})}.$$

א. [4 נק'] השתמשו במטריצת ההסיאן (שהיא $\nabla^2_{\mathbf{w}}\mathcal{L}(\mathbf{w})=\sum_{i=1}^m \exp(-y_i\mathbf{w}^\top\mathbf{x}_i)\,\mathbf{x}_i\mathbf{x}_i^\top$) כדי להוכיח שהבעיה קמורה.

הוכחה: כל $\mathbf{x}_i\mathbf{x}_i^\top$ היא מטריצת גראם ולכן PSD. המקדמים $\exp(-y_i\mathbf{w}^\top\mathbf{x}_i)$ הם בהכרח חיוביים.

מכאן שההסיאן הוא צירוף ליניארי עם מקדמים חיוביים של מטריצות PSD ולכן הוא גם PSD.

ב. [8 נק'] למרות הקמירות, הוכיחו שתחת הנחת הפרידות, לבעיה <u>אין מינימום</u>.

הוכחה: הבעיה פרידה ולכן קיים $\overline{\mathbf{w}}\in\mathbb{R}^d$ כך שלכל $i\in[m]$ מתקיים $y_i\overline{\mathbf{w}}^\top\boldsymbol{x}_i>0$

לכל $\alpha>0$ הפיתרון $\alpha\overline{\mathbf{w}}$ גם הוא מקיים $y_i(\alpha\overline{\mathbf{w}})^\top\boldsymbol{x}_i>0$ לכל i.

משמע, לכל $i\in[m]$, הפונק' $y_i(\alpha\overline{\mathbf{w}})^\top\boldsymbol{x}_i$ חיובית ועולה ב-α ולכן הפונק' $\exp(-y_i\boldsymbol{w}^\top\boldsymbol{x}_i)$ יורדת ב-α.

בסה"כ בעיית האופטימיזציה היא סכום של פונקציות חיוביות יורדות (ממש) ב-α ולכן אין לה מינימום.

<u>הערה</u>: בניגוד לתרגיל הבית, לא הנחנו כאן שהדוגמאות בת"ל. במצב כזה לא ניתן לטעון בפשטות שהגרדיינט $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})=-\sum_{i=1}^m \underbrace{\exp(-y_i\mathbf{w}^\top\mathbf{x}_i)\,y_i}_{\text{יכול להיות חיובי או שלילי}}\mathbf{x}_i$ לא מתאפס. שימו לב גם שהגרדיינט לא היה נתון בשלב הזה בשאלה.

<u>טענה (אין צורך להוכיח)</u>: מתקיים $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})=-\sum_{i=1}^m \exp(-y_i\mathbf{w}^\top\mathbf{x}_i)\,y_i\mathbf{x}_i$.

<u>נסמן</u>: יהיו $\mathbf{w}(1),\dots,\mathbf{w}(t)\in\mathbb{R}^d$ הפתרונות המתקבלים מהרצת GD (מלא, לא סטוכסטי) על ה-exponential loss החל מהווקטור $\mathbf{w}(0)=\mathbf{0}_d$, עם גודל צעד $\eta>0$.

המדגם בנוי מג'ימי אותר שביו +1 ו -1

ג. [5 נק'] בסעיף זה <u>בלבד</u> נניח שיש רק דוגמה אחת $(\mathbf{x},y)$ במדגם, שהתיוג שלה הוא $y=1$ ושהיא מנורמלת ($\|\mathbf{x}\|=1$).

כמו כן, נניח שגודל הצעד בהרצת GD הוא $\eta=1$.

קל לראות שתחת תנאי הסעיף מתקיים: $\mathbf{w}(t)=\underbrace{\|\mathbf{w}(t)\|}_{\in\mathbb{R}}\underbrace{\mathbf{x}}_{\in\mathbb{R}^d}$ (ניתן להשתמש בכך מבלי להוכיח זאת).

<u>הוכיחו</u> שתחת תנאי הסעיף מתקיים: $\|\mathbf{w}(t)\|=\|\mathbf{w}(t-1)\|+\mathrm{e}^{-\|\mathbf{w}(t-1)\|}$ (אין צורך להשתמש באינדוקציה).

הוכחה: $\mathbf{w}(t)=\mathbf{w}(t-1)+\exp(-\mathbf{w}_{t-1}^\top\mathbf{x})\,\mathbf{x}=\|\mathbf{w}(t-1)\|\mathbf{x}-\exp(-\|\mathbf{w}(t-1)\|\mathbf{x}^\top\mathbf{x})\,\mathbf{x}$

$$=\left(\|\mathbf{w}(t-1)\|-\mathrm{e}^{-\|\mathbf{w}(t-1)\|}\right)\mathbf{x}$$

$\{1,-1\}$

ו-$\mathbf{x}$ וקטור מנורמל.

$w_t = w_{t-1} + \sum e^{(\ )} y_i x_i = w_{t-1} + e$

תחת תנאי הסעיף האחרון ובהמשך לנוסחה הרקורסיבית שהוכחתם, ניתן להראות שמתקיים $\|\mathbf{w}(t)\| \approx \ln(t)$.

כעת נדון בתופעה כללית יותר (לדאטה פריד ליניארית עם $m \ge 1$ דוגמאות).

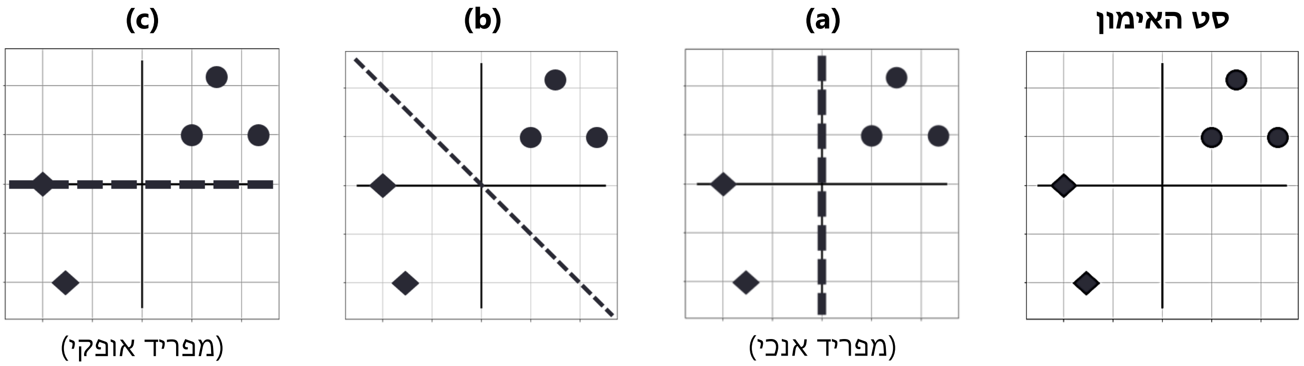
<u>נסמן</u>: $\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{\|\mathbf{w}\|}\ \text{s.t.}\ \left(\left(\min_{i\in[m]}|\mathbf{w}^\top \mathbf{x}_i| = 1\right) \text{ and } \left(\forall i\in[m]\colon\ y_i \mathbf{w}^\top \mathbf{x}_i > 0\right)\right)$.

<u>משפט</u>: לכל $t \ge 3$, הפתרונות $\mathbf{w}(1),\dots,\mathbf{w}(t) \in \mathbb{R}^d$ שמתקבלים מהרצת GD על ה-exponential loss הם מהצורה:

$$\mathbf{w}(t) = \underbrace{\ln(t)}_{\in\mathbb{R}} \cdot \underbrace{\hat{\mathbf{w}}}_{\in\mathbb{R}^d} + \underbrace{\tilde{\mathbf{r}}(t)}_{\in\mathbb{R}^d}$$

עבור וקטורים $\tilde{\mathbf{r}}(1),\dots,\tilde{\mathbf{r}}(t)$ לא ידועים שמקיימים $\|\tilde{\mathbf{r}}(t)\| = \mathcal{O}(\ln\ln(t))$.

ד. [8 נק'] לפניכם תרשימים של סט אימון פריד ליניארית עם 5 דוגמאות ובנוסף כמה גבולות החלטה (decision boundaries).



איזה תרשים מתאר את גבול ההחלטה שמתקבל ע"י $\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$?　　　(הניחו שאין שגיאות נומריות)

סמנו באופן <u>ברור</u> והסבירו בקצרה.　　　**(a)**　　　***(b)***　　　**(c)**　　　**(לא ניתן לדעת)**

הסבר:

כפי שלמדנו $\hat{\mathbf{w}}$ הוא הפיתרון של Hard-SVM (אפשר לראות זאת גם ע"י פעולות פשוטות על הגדרה).

$$\lim_{t\to\infty}\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \lim_{t\to\infty}\frac{\ln(t)\,\hat{\mathbf{w}} + \|\tilde{\mathbf{r}}(t)\|\frac{1}{\|\tilde{\mathbf{r}}(t)\|}\tilde{\mathbf{r}}(t)}{\|\mathbf{w}(t)\|} = \lim_{t\to\infty}\frac{\ln(t)\,\hat{\mathbf{w}} + c\ln\ln(t)\frac{1}{\|\tilde{\mathbf{r}}(t)\|}\tilde{\mathbf{r}}(t)}{\left\|\ln(t)\,\hat{\mathbf{w}} + c\ln\ln(t)\frac{1}{\|\tilde{\mathbf{r}}(t)\|}\tilde{\mathbf{r}}(t)\right\|}$$

$$= \lim_{t\to\infty}\frac{\hat{\mathbf{w}} + \overbrace{\frac{c\ln\ln(t)}{\ln(t)}}^{\to 0}\overbrace{\frac{1}{\|\tilde{\mathbf{r}}(t)\|}\tilde{\mathbf{r}}(t)}^{=O(1)}}{\left\|\hat{\mathbf{w}} + \frac{c\ln\ln(t)}{\ln(t)}\frac{1}{\|\tilde{\mathbf{r}}(t)\|}\tilde{\mathbf{r}}(t)\right\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}$$

ולכן יתקבל מפריד זהה לזה של ה-Hard-SVM.

שאלה 5: רגרסיה ורגולריזציה [25 נק']

עבור $\mathbf{y}\in\mathbb{R}^m$, $\mathbf{X}\in\mathbb{R}^{m\times d}$ וְ-$\lambda>0$, נסמן את הפתרונות של שלוש בעיות רגרסיה ליניארית שלמדנו:

$$\hat{\mathbf{w}}_{\ell 1}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left(\|\mathbf{Xw}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_1\right),\qquad \hat{\mathbf{w}}_{\ell 2}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left(\|\mathbf{Xw}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2\right),\qquad \hat{\mathbf{w}}_{\mathrm{LS}}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\|\mathbf{Xw}-\mathbf{y}\|_2^2,$$

<u>הנחה:</u> בשאלה זו נניח שהעמודות של $\mathbf{X}$ אורתונורמליות (משמע $\mathbf{X}^\top\mathbf{X}=\mathbf{I}_{d\times d}$).

<u>תזכורת:</u> הנגזרת של התבנית הריבועית היא $\nabla_{\mathbf{w}}\|\mathbf{Xw}-\mathbf{y}\|_2^2=2\mathbf{X}^\top(\mathbf{Xw}-\mathbf{y})$.

א. [3 נק'] תחת ההנחה (אורתונורמליות), הוכיחו שמתקיים $\hat{\mathbf{w}}_{\mathrm{LS}}=\mathbf{X}^\top\mathbf{y}$.

$$\nabla_{\mathbf{w}}\|\mathbf{Xw}-\mathbf{y}\|_2^2=2\mathbf{X}^\top(\mathbf{Xw}-\mathbf{y})=2\mathbf{X}^\top\mathbf{Xw}-2\mathbf{X}^\top\mathbf{y}=2\mathbf{w}-2\mathbf{X}^\top\mathbf{y}=\mathbf{0}$$

$$\hat{\mathbf{w}}_{\mathrm{LS}}=\mathbf{X}^\top\mathbf{y}$$

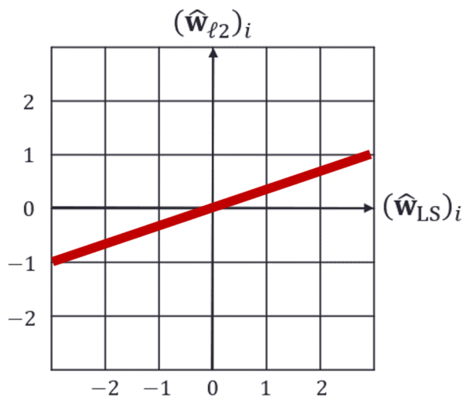
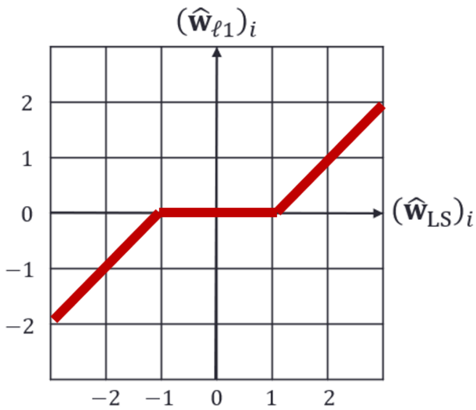
ב. [3 נק'] תחת ההנחה, מצאו ביטוי סגור ל-$\hat{\mathbf{w}}_{\ell 2}$ כפונקציה של $\lambda,\hat{\mathbf{w}}_{\mathrm{LS}}$. צרפו לתשובתכם פיתוח <u>קצר</u> מתאים.

$$\nabla_{\mathbf{w}}\left(\|\mathbf{Xw}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2\right)=2\mathbf{X}^\top(\mathbf{Xw}-\mathbf{y})+2\lambda\mathbf{w}=2\mathbf{w}-2\mathbf{X}^\top\mathbf{y}+2\lambda\mathbf{w}=\mathbf{0}$$

$$\hat{\mathbf{w}}_{\ell 2}=\left(\frac{1}{1+\lambda}\right)\mathbf{X}^\top\mathbf{y}=\left(\frac{1}{1+\lambda}\right)\hat{\mathbf{w}}_{\mathrm{LS}}$$

<u>נתונה טענה 1:</u> תחת ההנחה, מתקיים $(\hat{\mathbf{w}}_{\ell 1})_i=\operatorname{sign}((\hat{\mathbf{w}}_{\mathrm{LS}})_i)\cdot\max\left(0,\ |(\hat{\mathbf{w}}_{\mathrm{LS}})_i|-\frac{\lambda}{2}\right)$.

ג. [4 נק'] עבור כניסה i שרירותית וערך $\lambda=2$, השתמשו בביטוי שמצאתם בסעיף הקודם ובטענה 1 וציירו <u>באופן ברור</u> על גבי התרשימים הבאים את העקומות של $(\hat{\mathbf{w}}_{\ell 2})_i$ וְ-$(\hat{\mathbf{w}}_{\ell 1})_i$ כפונקציה של $(\hat{\mathbf{w}}_{\mathrm{LS}})_i$.

$(\hat{\mathbf{w}}_{\ell 2})_i$ כפונקציה של $(\hat{\mathbf{w}}_{\mathrm{LS}})_i$ כאשר $\lambda=2$	$(\hat{\mathbf{w}}_{\ell 1})_i$ כפונקציה של $(\hat{\mathbf{w}}_{\mathrm{LS}})_i$ כאשר $\lambda=2$
	

ד. [8 נק'] הסבירו כיצד הסעיפים הקודמים ממחישים את התכונות של Ridge regression ו-Lasso וההבדלים ביניהם.

Ridge מכווץ את המשקלים ו-LASSO מאפס משקלים שקרובים לאפס.

כעת נוכיח את טענה 1 במקרה חד-ממדי פשוט, משמע: $d = 1$.

כלומר, עבור $\mathbf{X}, \boldsymbol{y} \in \mathbb{R}^m$ וְ-$\lambda > 0$, נגדיר: $\widehat{w}_{\ell 1} = \underset{w\in\mathbb{R}}{\operatorname{argmin}} \underbrace{\left(\|\mathbf{X}w - \mathbf{y}\|_2^2 + \lambda|w|\right)}_{\triangleq \mathcal{L}(w)}$, $\widehat{w}_{\text{LS}} = \underset{w\in\mathbb{R}}{\operatorname{argmin}} \|\mathbf{X}w - \mathbf{y}\|_2^2$.

<u>נמשיך להניח</u>: מתקיים $\mathbf{X}^\top\mathbf{X} = 1$ (ולכן עדיין מתקיים $\widehat{w}_{\text{LS}} = \mathbf{X}^\top\mathbf{y}$).

לשימושכם תזכורת לתתי-נגזרות והכללה של תנאי האופטימליות למקרה של תת-נגזרת:

<u>תזכורת</u>: תהא $f: \mathbb{R} \to \mathbb{R}$ קמורה. נסמן ב-$\partial f(w)$ את קבוצת תתי-הנגזרות (sub-derivatives) שלה בנקודה $w \in \mathbb{R}$.

למשל, עבור $f(z) = |z|$, מתקיים $\partial f(-2) = \{-1\}$ וגם $\partial f(0) = [-1,1]$.

<u>תנאי אופטימליות (אין צורך להוכיח)</u>: תהא $f: \mathbb{R} \to \mathbb{R}$ קמורה. $f(w) = \min_{z\in\mathbb{R}} f(z)$ אמ"מ $0 \in \partial f(w)$.

ה. [8 נק'] הוכיחו שבמקרה החד-ממדי, תחת הנחת האורתונורמליות, מתקיים $\widehat{w}_{\ell 1} = \operatorname{sign}(\widehat{w}_{\text{LS}}) \cdot \max\left(0,\ |\widehat{w}_{\text{LS}}| - \frac{\lambda}{2}\right)$.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

$$\mathcal{L}'(w) = \boldsymbol{2X}^\top(\boldsymbol{X}w - \boldsymbol{y}) + \lambda g(w) = 2w - 2\boldsymbol{X}^\top\boldsymbol{y} + \lambda g(w) = 2w - 2\widehat{w}_{LS} + \lambda g(w)$$

כאשר g תת-נגזרת כלשהי $g(w) \in \partial f(w)$ עבור $f(w) = |w|$. נשים לב שמתקיים $\mathcal{L}'(w) \in \partial\mathcal{L}(w)$

נראה שמתקיים $\mathcal{L}'(\widehat{w}_{\ell 1}) = 0$ וסיימנו (לפי תנאי האופטימליות).

מקרה א': מתקיים $\widehat{w}_{\ell 1} = |\widehat{w}_{LS}| - \frac{\lambda}{2} > 0$. הנגזרת היא בהכרח $g(\widehat{w}_{\ell 1}) = sign(\widehat{w}_{LS})$.

$$\mathcal{L}'(\widehat{w}_{\ell 1}) = 2\widehat{w}_{\ell 1} - 2\widehat{w}_{LS} + \lambda g(\widehat{w}_{\ell 1}) = 2\left(sign(\widehat{w}_{LS})\left(|\widehat{w}_{LS}| - \frac{\lambda}{2}\right)\right) - 2\widehat{w}_{LS} + sign(\widehat{w}_{LS})\lambda$$

$$= 2\widehat{w}_{LS} - sign(\widehat{w}_{LS})\lambda - 2\widehat{w}_{LS} + sign(\widehat{w}_{LS})\lambda = 0$$

מקרה ב': מתקיים $\widehat{w}_{\ell 1} = 0 \geq |\widehat{w}_{LS}| - \frac{\lambda}{2}$ ולכן $\lambda \geq 2|\widehat{w}_{LS}| > 0$.

<u>נבחר</u> תת-נגזרת $g(0) = \frac{2\widehat{w}_{LS}}{\lambda} \in [-1,1]$.

$$\mathcal{L}'(\widehat{w}_{\ell 1}) = 2\widehat{w}_{\ell 1} - 2\widehat{w}_{LS} + \lambda g(\widehat{w}_{\ell 1}) = 0 - 2\widehat{w}_{LS} + \lambda\left(\frac{2\widehat{w}_{LS}}{\lambda}\right) = 0$$