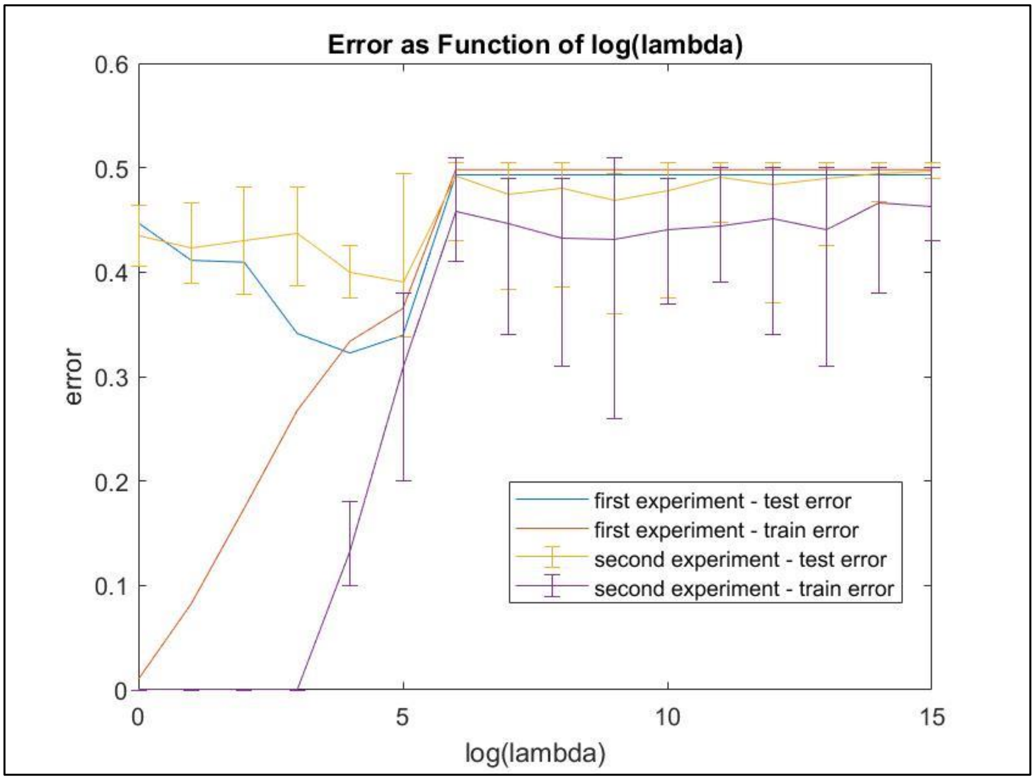
שאלה 2

א. הניסוי בוצע בגרסתו הראשונה עבור $0 \le \log(\lambda) \le 15$, כאשר הניסוי הראשון בוצע על כלל מדגם האימון, והניסוי השני בוצע על ממוצע של עשר ריצות על מדגם אימון אקראי מגודל 100 :

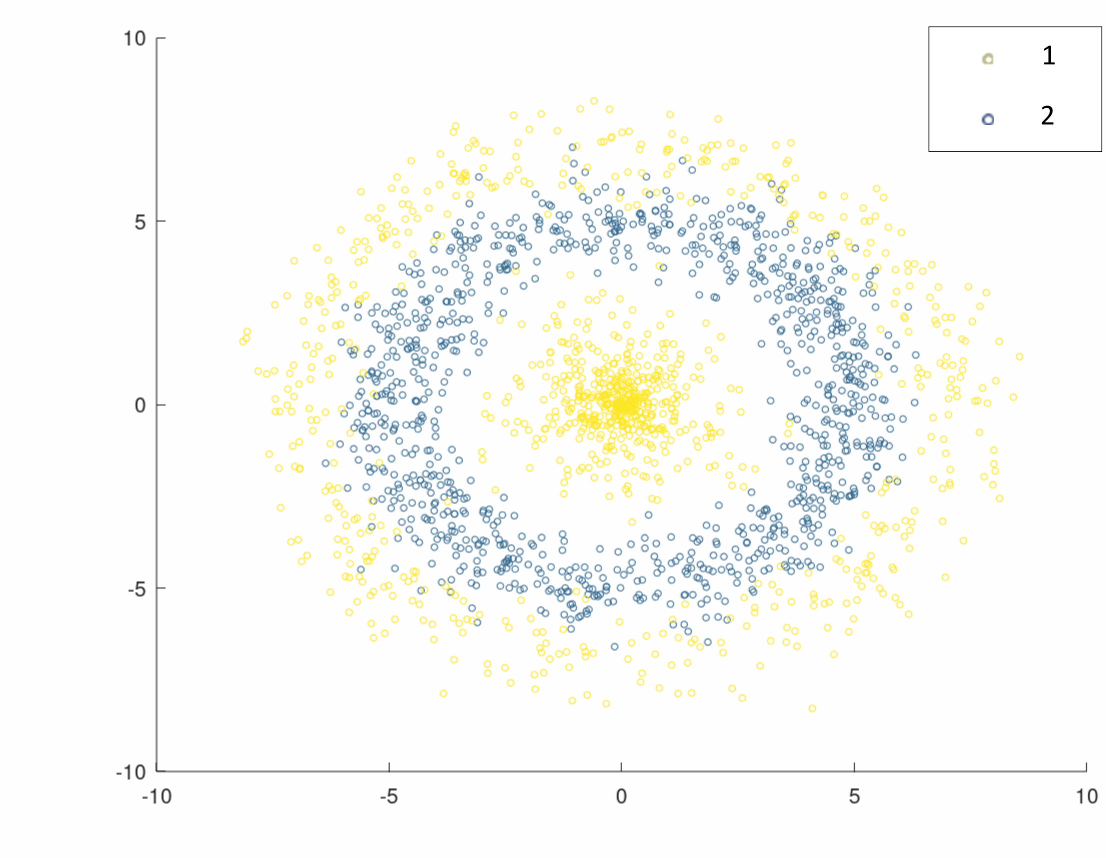


ב. נסמן S_1=מדגם האימון, S_2=תת קבוצה מגודל 100 של מדגם האימון.

- <u>שגיאת אימון:</u> מכיוון ש $S_2 \subseteq S_1$, ועי״פ הגדרת פונקציית ה hinge lost, ניתן להבחין כי לכל $w \in \mathbb{R}^d$ מתקיים כי $\ell^h(w, S_2) \le \ell^h(w, S_1)$. כלומר, נצפה כי אלגוריתם ה soft SVM יחזיר מפריד בעל hinge lost קטן יותר על S_2 מאשר על S_1, ומכיוון שערך השגיאה קטן יותר מערך פונקציית ה hinge lost, נצפה כי אלגוריתם ה soft SVM יחזיר מפריד בעל שגיאת אימון קטנה יותר על S_2 מאשר על S_1. ואכן, ניתן לראות כי שגיאת האימון בניסוי השני (סגול), קטנה יותר משגיאת האימון בניסוי הראשון (כתום).
<u>שגיאת מבחן:</u> ככל שגודל מדגם האימון גדול יותר, כך הוא מתאר טוב יותר את ההתפלגות. לכן, אם w_1, w_2 הינם המפרידים שחזרו ע״י אלגוריתם ה Soft SVM על S_1, S_2 בהתאמה, נצפה כי $\ell^h(w_1, D) \le \ell^h(w_2, D)$. מכיוון שערך השגיאה קטן יותר מערך פונקציית ה hinge lost, נצפה כי $err(h_{w_1}, D) \le err(h_{w_2}, D)$. ואכן, ניתן לראות כי שגיאת המבחן בניסוי הראשון (כחול), קטנה יותר משגיאת המבחן בניסוי השני (צהוב).
- ככל ש λ גדלה, אלגוריתם ה soft SVM משלם יותר על החלק של הנורמה ב optimization objective, ולכן יחזיר מפריד בעל hinge lost גבוהה יותר על S_2. מכיוון שהגדלת ערך ה hinge lost עשויה רק להגדיל את ערך השגיאה, נצפה כי ככל ש λ גדלה, כך שגיאת האימון גדלה. ואכן, ניתן לראות כי ככל ש λ גדלה כך שגיאת האימון בניסיו השני (סגול) גדלה.
- ככל ש λ גדלה, כך מצטמצמת מחלקת ההיפותזות. לכן, נצפה כי ככל ש λ קטנה ביחס לערכה האופטימלי נתקדם בכיוון של over fitting והשגיאה תגדל, וככל ש λ גדלה ביחס לערכה האופטימלי נתקדם בכיוון של under fitting והשגיאה תגדל. ואכן, ניתן להבחין כי ככל ש λ מתרחקת (צהוב) מערכה האופטימלי (5), כך השגיאה גדלה.

שאלה 4

א.



ע״פ הגרף הנ״ל ניתן להבחין כי לא קיים ישר במישור שיכול להפריד את דוגמאות המדגם בצורה טובה. לכן, המפריד הליניארי שיתקבל ע״י הרצת linear soft SVM על מדגם זה יהיה בעל שגיאה גבוהה על המדגם. לעומת זאת, kernel soft SVM יוכל לעלות את מימד דוגמאות המדגם, ולהחזיר משטח שהינו מפריד בעל שגיאה נמוכה על המדגם.

ב.

λ	σ	$validation\ error$
1	0.01	0.0775
1	0.5	0.1115
1	1	0.2010
10	0.01	0.0775
10	0.5	0.1110
10	1	0.1970
100	0.01	0.0775
100	0.5	0.1110
100	1	0.1965

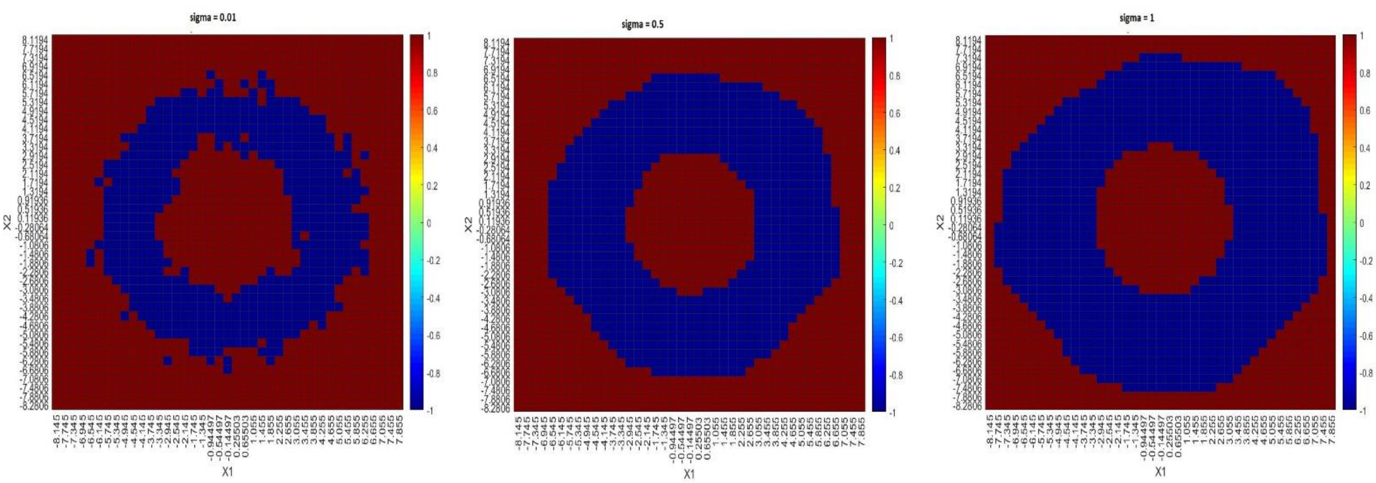
עבור הפרמטרים הנבחרים $\lambda = 1, \sigma = 0.01$, ריצת אלגוריתם ה kernel soft SVM על כל מדגם האימון החזירה מפריד בעל שגיאה של $\mathbf{0.06}$ על מדגם המבחן.

$\boldsymbol{\lambda}$	$\boldsymbol{validation\ error}$
1	0.4940
10	0.4940
100	0.4940

עבור הפרמטר הנבחר $\lambda = 1$, ריצת אלגוריתם ה linear soft SVM על כל מדגם האימון החזירה מפריד בעל שגיאה של $\mathbf{0.51}$ על מדגם המבחן.

ג. כצפוי, RBF soft SVM השיג validation error הקטן במידה רבה ביחס ל linear soft SVM. זאת משום שכפי שתארנו בסעיף א׳, המדגם אינו פריד ליניארית כלל, ולכן האחרון ישיג שגיאה גבוהה, ואילו הראשון ישיג שגיאה נמוכה כאשר יגדיל את מימד המדגם.
באופן כללי, RBF soft SVM יכול להשיג validation error נמוך משום שהוא ״מעשיר״ את המחלקת ההיפותזות, וכך יכול לבחור כלל סיווג המתאים יותר להתפלגות ביחס ל linear soft SVM.
מצד שני, RBF יכול ״להעשיר מדי״ את מחלקת ההיפותזות וכך לבחור כלל סיווג אשר נותן שגיאה נמוכה על המדגם, ואילו שגיאה גבוהה על ההתפלגות, ביחס ל linear.

ד.



ה. נבחין במסווג אשר מתקבל ע״י אלגוריתם ה RBF:

$$h_w(\psi(x)) = sign\left(\sum_{i=1}^{m} \alpha(i) e^{-\frac{||x-x_i||^2}{2\sigma}}\right)$$

לכל $1 \le i \le m$, מתקיים כי משקל ה $\alpha(i)$ הינו $e^{-\frac{||x-x_i||^2}{2\sigma}}$. מכאן, ככל ש σ גדל, כך ההשפעה של המרחק בין הדוגמה ה i ל x על הסיווג של x קטנה. כלומר, דוגמאות שמרחקן גדול מ x משפיעות יותר על הסיווג שלה כאשר σ גדול יותר. ואכן, נבחין באיורים כי ככל ש σ גדל, כל ההשפעה של דוגמאות רחוקות על נקודה מסוימת גדלה, והציור הופך יותר ויותר ״חלק״.

שאלה 5

א. נניח בשלילה כי קיים ψ כך שפונקציית הקרנל שלו הינה $K(x,x') = -x(1)x'(1)$.
ע״פ הגדרת פונקציית קרנל נקבל כי:

$$-x(1)x'(1) = K(x,x') = \langle \psi(x), \psi(x') \rangle$$

יהי $x \in X$, כך ש $x(1) \neq 0$. נסמן $\psi(x) = y \in \mathbb{R}^t$. מכאן נקבל כי:

$$-x(1)x(1) = K(x,x) = \langle \psi(x), \psi(x) \rangle = \langle y, y \rangle = y(1)y(1) + \cdots + y(t)y(t)$$

$x(1) \neq 0$ אזי $-x(1)x(1) < 0$. בנוסף, $y(1)y(1) + \cdots + y(t)y(t) \geq 0$ וקיבלנו כי:

$$0 < -x(1)x(1) = y(1)y(1) + \cdots + y(t)y(t) \geq 0$$

סתירה.

ב. נניח בשלילה כי קיים ψ כך שפונקציית הקרנל שלו הינה $K(x,x') = x(2)x'(1)$.
ע״פ הגדרת פונקציית קרנל נקבל כי:

$$x(2)x'(1) = K(x,x') = \langle \psi(x), \psi(x') \rangle$$

מכאן נקבל כי:

$$x(2)x'(1) = K(x,x') = \langle \psi(x), \psi(x') \rangle = \langle \psi(x'), \psi(x) \rangle = K(x',x) = x'(2)x(1)$$

יהיו $x, x' \in X$ כך ש $x = (1, \dots, d), x' = (2, \dots, d+1)$. ע״פ השוויון הקודם נקבל כי:

$$2 \cdot 2 = 3 \cdot 1$$

מכאן נקבל כי:

$$4 = 3$$

סתירה.

ג. נגדיר $R: \mathbb{R}_+ \to \mathbb{R}, f: \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ באופן הבא:

$$R(a) = \lambda a^4, f(a_1, \dots, a_m) = \sum_{i=1}^{m} exp^{-y_i|a_i|}$$

בתחום $a \geq 0$ נבחין כי R הינה פונקציה מונוטונית לא יורדת.
$Minimize_{w \in \mathbb{R}^d} \lambda \|w\|^4 + \sum_{i=1}^{m} exp^{-y_i|\langle w, x_i \rangle|} = Minimize_{w \in \mathbb{R}^d} R(\|w\|) + f(\langle w, x_1 \rangle, \dots, \langle w, x_m \rangle)$,
לכן ע״פ $representer\ theorem$ נקבל כי קיים פתרון $w \in \mathbb{R}^d$ המקיים $w = \sum_{i=1}^{m} \alpha_i x_i$ כך ש $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ כנדרש.

ד. נגדיר $\psi: \mathbb{R}^d \to \mathbb{R}$ באופן הבא:

$$\psi(x) = x(1) + x(2)$$

נבחין כי לכל $x, x' \in \mathbb{R}^d$ מתקיים כי:

$$\langle \psi(x), \psi(x') \rangle = \langle x(1) + x(2), x'(1) + x'(2) \rangle = \big(x(1) + x(2)\big)\big(x'(1) + x'(2)\big)$$
$$= x(1)x'(1) + x(2)x'(1) + x(1)x'(2) + x(2)x'(2) = K(x,x')$$

כנדרש.

שאלה 6

א. קיים $h_c \in \mathcal{H}_{10}$ שהרופאים מחליטים בדיוק על פיו, כלומר עבורו מתקיים כי $err(h_c, D) = 0$.
<u>$n < 10$</u>: כיוון ש $c \leq 10$, לא בהכרח מתקיים כי $h_c \in \mathcal{H}_n$ עבור $n < 10$. מכאן שאם נריץ אלגוריתם ERM עם מחלקת היפותזות $\mathcal{H}_n$, ייתכן כי האלגוריתם יחזיר $h_{c'} \in \mathcal{H}_n$ בעל $err(h_{c'}, D) > 0$. כלומר, מכיוון ש $\mathcal{H}_n$ פשוטה ביחס ל $\mathcal{H}_{10}$ עבור בעיה זו, האלגוריתם עלול להחזיר כלל בעל שגיאת אפרוקסימציה הגדולה מכלל אופטימלי ונקבל מצב של *under fitting*.
<u>$n > 10$</u>: כיוון ש $\mathcal{H}_{10} \subseteq \mathcal{H}_n$ עבור $n > 10$, יתכן כי $|A(\mathcal{H}_{10})| < |A(\mathcal{H}_n)|$ כאשר:
$A(\mathcal{H}) = \{h \in \mathcal{H}: err(h, S) = 0 \wedge err(h, D) > 0\}$.
משמע, יתכן כי ביחס ל $\mathcal{H}_{10}$, ERM עם $\mathcal{H}_n$ יחזיר $h_{c'}$ בעל $err(h_{c'}, D) > 0$ בסיכוי גדול יותר. כלומר, מכיוון ש $\mathcal{H}_n$ מורכבת ביחס ל $\mathcal{H}_{10}$ עבור בעיה זו, האלגוריתם עלול להחזיר כלל בעל שגיאת אסטימיזיציה הגדולה מכלל אופטימלי ונקבל מצב של *over fitting*.

ב. מכיוון ש $\mathcal{H}_{10}$ הינה מחלקת היפותזות בעלת גודל סופי, ו D הינה התפלגות *realizable* ע"י $\mathcal{H}_{10}$, נוכל להשתמש בחסם PAC עבור המקרה ה *realizable*:

$$m \geq \frac{\log(|\mathcal{H}_{10}|) + \log\left(\frac{1}{0.01}\right)}{0.1} = \frac{\log\left(\sum_{i=0}^{10} 128^i\right) + \log(100)}{0.1} \approx 230.75$$

מכאן נקבל כי:

$$\boldsymbol{m \geq 231}$$

ג. מכיוון ש $\mathcal{H}_{10}$ הינה מחלקת היפותזות בעלת גודל סופי, ו D הינה התפלגות שאינה בהכרח *realizable* ע"י $\mathcal{H}_n$ עבור $n < 10$, נוכל להשתמש בחסם PAC עבור המקרה ה *agnostic*:

$$\begin{aligned} m &\geq \frac{2\log(|\mathcal{H}_n|) + 2\log\left(\frac{2}{0.01}\right)}{0.1^2} = \frac{2\log\left(\sum_{i=0}^{n} 128^i\right) + 2\log(200)}{0.01} \\ &= \frac{2\log\left(\frac{128^n - 1}{127}\right) + 2\log(200)}{0.01} \\ &= \frac{2(\log(128^n - 1) - \log 127) + 2\log(200)}{0.01} \\ &= 200\left(\log(128^n - 1) + \log\left(\frac{200}{127}\right)\right) \approx 421.45n + 39.45 \end{aligned}$$

מכאן נקבל כי:

$$\boldsymbol{m \geq \lceil 421.45n + 39.45 \rceil}$$

שאלה 7

נגדיר $\xi_i = \ell^h(w,(x_i,y_i))$.

עבור תוכנית ריבועית מהצורה: $minimize_{z\in\mathbb{R}^{d+m}} \frac{1}{2} z^T Hz + \langle u,z\rangle, subject\ to\ Az \geq v$ כך ש $z = \begin{bmatrix} w_1 \\ \vdots \\ w_d \\ \xi_1 \\ \vdots \\ \xi_m \end{bmatrix}$,

נגדיר את המשתנים הבאים:

$$H = \begin{bmatrix} 2\lambda\mathbb{I}_d & [0]_{d,m} \\ [0]_{m,d} & 2\mathbb{I}_m \end{bmatrix}, u = \vec{0} \in \mathbb{R}^{d+m}, A = \begin{bmatrix} [0]_{m,d} & \mathbb{I}_m \\ [x_i y_i]_{m,d} & \mathbb{I}_m \end{bmatrix}, v = \begin{bmatrix} [0]_{m,1} \\ [1]_{m,1} \end{bmatrix}$$

נסמן $P(w) = minimize\ \lambda\|w\|^2 + \sum_{i=1}^m \left(\ell^h\left(w,(x_i,y_i)\right)\right)^2$.

<u>טענה:</u> הפתרון לתוכנית הריבועית הנ"ל הינו הפתרון לבעיה $P(w)$.

<u>הוכחה:</u> האילוץ $Az \geq v$ גורר שלכל i כך ש $1 \leq i \leq m$ מתקיים כי $y_i\langle w,x_i\rangle \geq 1-\xi_i$ וגם $\xi_i \geq 0$.

אילוצים אלה שקולים לכך ש לכל i כך ש $1 \leq i \leq m$ מתקיים $\xi_i \geq \max\{0,1-y_i\langle w,x_i\rangle\}$. לכן, בפתרון האופטימלי מתקיים כי $\xi_i = \ell^h(w,(x_i,y_i))$.

בנוסף, $\frac{1}{2} z^T Hz + \langle u,z\rangle = \lambda\|w\|^2 + \sum_{i=1}^m \left(\ell^h\left(w,(x_i,y_i)\right)\right)^2$, ובסה"כ הראינו כי הפתרון לתוכנית הריבועית הנ"ל הינו הפתרון לבעיה $P(w)$.

שאלה 8

א. נוכיח את הטענה באינדוקציה על t:

<u>בסיס $t=1$:</u> ע"פ תיאור האלגוריתם $w^{(t)} = w^{(1)} = \vec{0}$, לכן $y_j\langle w^{(t)},x_j\rangle = 0$ לכל $1 \leq j \leq m$.
מכאן ש $w^{(t+1)} = w^{(2)} = w^{(1)} + y_l x_l = y_l x_l$, עבור $1 \leq l \leq m$ כלשהו.
ע"פ הגדרת המדגם, לכל $1 \leq i \leq d$ מתקיים כי $(y_l x_l)(i) \in \{-1,0,1\}$.
לכן, לכל $1 \leq i \leq d$, מתקיים כי $\left|w^{(t+1)}(i)\right| = \left|w^{(2)}(i)\right| = |(y_l x_l)(i)| \leq 1 = t$ כנדרש.
<u>הנחה:</u> נניח כי עבור $t > 1$ כלשהו מתקיים כי לכל $1 \leq i \leq d$ מתקיים $\left|w^{(t+1)}(i)\right| \leq t$.
<u>צעד:</u> צ"ל כי לכל $1 \leq i \leq d$ מתקיים $\left|w^{(t+2)}(i)\right| \leq t+1$.
נחלק למקרים:
I. לכל $1 \leq j \leq m$ מתקיים כי $y_j\langle w^{(t+1)},x_j\rangle > 0$ ומכאן ש $w^{(t+2)} = w^{(t+1)}$. לכן, ע"פ הנחת האינדוקציה, לכל $1 \leq i \leq d$ מתקיים כי $\left|w^{(t+2)}(i)\right| = \left|w^{(t+1)}(i)\right| \leq t < t+1$ כנדרש.
II. קיים $1 \leq l \leq m$ כך ש $y_l\langle w^{(t+1)},x_l\rangle \leq 0$, מכאן ש $w^{(t+2)} = w^{(t+1)} + y_l x_l$.
ע"פ הגדרת המדגם, לכל $1 \leq i \leq d$ מתקיים כי $(y_l x_l)(i) \in \{-1,0,1\}$.
לכן, לכל $1 \leq i \leq d$, מתקיים כי $|\left(w^{(t+1)} + y_l x_l\right)(i)| \leq |w^{(t+1)}(i) + 1|$.
מכאן ניתן להסיק כי לכל $1 \leq i \leq d$, מתקיים כי
$\left|w^{(t+2)}(i)\right| = |\left(w^{(t+1)} + y_l x_l\right)(i)|)| \leq \left|w^{(t+1)}(i) + 1\right| \leq \left|w^{(t+1)}(i)\right| + |1| \leq^{\text{ה"א}} t+1$ כנדרש.

ב. נוכיח את הטענה באינדוקציה על i :

<u>בסיס $i=1$:</u> לכל $1 \le j \le m$ מתקיים כי $y_j\langle w^{(T)}, x_j\rangle > 0$, ובפרט עבור $j=1$.

ע״פ הגדרת המדגם מתקיים כי $x_1 = (1,0,\ldots,0), y_1 = 1$.

מכאן ש $0 < y_j\langle w^{(T)}, x_j\rangle = 1 \cdot \left(w^{(T)}(1) \cdot 1 + w^{(T)}(2) \cdot 0 + \cdots + w^{(T)}(2) \cdot 0\right) = w^{(T)}(1)$.

קיבלנו כי $0 < w^{(T)}(1)$, ומכיוון שע״פ הגדרת האלגוריתם והמדגם מתקיים כי $w^{(T)} \in \mathbb{Z}^d$, ניתן להסיק כי $w^{(T)}(1) \ge 1 = 2^0 = 2^{i-1}$ כנדרש.

<u>הנחה :</u> נניח כי לכל j עבור $1 \le j < i$ מתקיים כי $w^{(T)}(j) \ge 2^{j-1}$.

<u>צעד :</u> צריך להוכיח כי $w^{(T)}(i) \ge 2^{i-1}$.

ע״פ הגדרת המדגם מתקיים כי $y_i\langle w^{(T)}, x_i\rangle = 1\,\langle w^{(T)}, (\overbrace{-1,\ldots,-1}^{i-1}, 1, \overbrace{0,\ldots,0}^{d-i})\rangle$ או

$y_i\langle w^{(T)}, x_i\rangle = (-1)\,\langle w^{(T)}, (\overbrace{1,\ldots,1}^{i-1}, -1, \overbrace{0,\ldots,0}^{d-i})\rangle$.

בכל אופן מתקיים כי $y_i\langle w^{(T)}, x_i\rangle = w^{(T)}(i) - \sum_{j=1}^{i-1} w^{(T)}(j)$.

מכיוון ש $y_i\langle w^{(T)}, x_i\rangle > 0$, נקבל כי $0 < y_i\langle w^{(T)}, x_i\rangle = w^{(T)}(i) - \sum_{j=1}^{i-1} w^{(T)}(j)$.

מכאן ניתן לראות ש $w^{(T)}(i) > \sum_{j=1}^{i-1} w^{(T)}(j) \ge \sum_{j=1}^{i-1} 2^{j-1} = 2^{i-1} - 1$.

קיבלנו כי $2^{i-1} - 1 < w^{(T)}(i)$, ומכיוון שע״פ הגדרת האלגוריתם והמדגם מתקיים כי $w^{(T)} \in \mathbb{Z}^d$, ניתן להסיק כי $w^{(T)}(i) \ge 2^{i-1}$ כנדרש.

ג. נסתכל על הסדרה $\left\{w^{(t)}(d)\right\}_{t=1}^{T}$.

$w^{(1)}(d) = 0$, וע״פ סעיף א׳ לכל $2 \le t \le T$ מתקיים כי $w^{(t)}(d) \le t - 1$.

ע״פ סעיף ב׳ $w^{(T)}(d) \ge 2^{d-1}$.

כלומר, איברי הסדרה גדלים בלכל היותר 1 בכל איטרציה של האלגוריתם, האיבר האחרון של הסדרה הינו $O(2^d)$. משמע, מספר האיטרציות, שממנו נגזר זמן ריצת האלגוריתם, הינו $O(2^d)$.

<u>שאלה 9</u>

א. נגדיר

$f(w) = \lambda\|w\| + \sum_{i=1}^{m}(\langle w, x_i\rangle - y_i)^2 =$

$. = \lambda\sqrt{w(1)^2 + \cdots + w(d)^2} + \sum_{i=1}^{m}(x_i(1)w(1) + \cdots + x_i(d)w(d) - y_i)^2$

נבחין כי לכל $1 \le j \le d$ מתקיים כי

$\frac{\partial f(w)}{\partial w(j)} = \frac{\lambda w(j)}{\sqrt{w(1)^2+\cdots+w(d)^2}} + 2\sum_{i=1}^{m} x_i(j)\cdot(x_i(1)w(1) + \cdots + x_i(d)w(d) - y_i) =$

$$= \frac{\lambda w(j)}{\|w\|} + 2\sum_{i=1}^{m} x_i(j)\cdot(\langle w, x_i\rangle - y_i)$$

מכאן, ע״פ הגדרת אלגוריתם gradient decent נקבל כי

$\boldsymbol{w^{(t+1)}} = w^{(t)} - \eta\nabla f\left(w^{(t)}\right) =$

$\boldsymbol{w^{(t)} - \eta\left(\left(\frac{\lambda w^{(t)}(j)}{\|w^{(t)}\|} + 2\sum_{i=1}^{m} x_i(j)\cdot\left(\langle w^{(t)}, x_i\rangle - y_i\right)\right), \ldots, \left(\frac{\lambda w^{(t)}(j)}{\|w^{(t)}\|} + 2\sum_{i=1}^{m} x_i(j)\cdot\left(\langle w^{(t)}, x_i\rangle - y_i\right)\right)\right)}$

ב. נגדיר

$R(w) = \lambda\|w\|, \ell(\mathrm{w}, (\mathrm{x}_i, \mathrm{y_i})) = (\langle w, x_i\rangle - y_i)^2$

נבחין כי לכל $1 \le j \le d$ מתקיים כי

$$\frac{\partial R(w)}{\partial w(j)} = \frac{\lambda w(j)}{\|w\|}, \frac{\partial l\left(w, (x_l, y_l)\right)}{\partial w(j)} = 2x_i(j)\cdot(\langle w, x_i\rangle - y_i)$$

מכאן, ע״פ הגדרת אלגוריתם stochastic gradient decent נקבל כי

$\boldsymbol{w^{(t+1)}} = w^{(t)} - \eta\left(\nabla R\left(w^{(t)}\right) + \nabla l\left(w^{(t)}, (x_i, y_i)\right)\right) =$

$$\boldsymbol{w^{(t)} - \eta\left(\left(\frac{\lambda w^{(t)}(j)}{\|w^{(t)}\|} + 2x_i(j)\cdot\left(\langle w^{(t)}, x_i\rangle - y_i\right)\right), \ldots, \left(\frac{\lambda w^{(t)}(j)}{\|w^{(t)}\|} + 2x_i(j)\cdot\left(\langle w^{(t)}, x_i\rangle - y_i\right)\right)\right)}$$

עבור $1 \le i \le m$ שנבחר ע״י האלגוריתם באופן רנדומלי.