# חלק תיאורטי

מגיש: רם אולגוב

ת.ז: 206867517

## שאלה 1: (Step-Size Perceptron)

פתרון: לפי הנתון הדאטא ניתן להפרדה ולכן קיים $w^*$ שמסווג אותו בצורה מושלמת. בה"כ נניח $\|w^*\|_2=1$

(זה עבור נורמליזציה, אחרת אפשר לעשות סקיילינג מבלי לשנות את המסווג).

מהנתון בשאלה נובע שלכל $1\leq t\leq M$ האלגוריתם טועה (כי הוא מבצע עדכון $M$ פעמים ומניחים שגם ו טועה בהתחלה).

נחלק את ההוכחה לשני חלקים:

1. נראה שמתקיים $w_{t+1}\cdot w^* \geq M\cdot\gamma$.

$$w_{t+1}\cdot w^* \underset{\text{הגדרת Step-Size Perceptron}}{=} (w_t+\eta_t y_t x_t)w^* \underset{\text{אדיטיביות}}{=} w_t\cdot w^* + \eta_t y_t x_t w^* \underset{\text{הומוגניות}}{=}$$

$$w_t w^* + \eta_t(y_t x_t\cdot w^*) \underset{\text{נתון בשאלה}}{=} w_t w^* + \frac{1}{\sqrt{t}}\overbrace{(y_t x_t\cdot w^*)}^{\geq\gamma} \underset{\text{מהנחת הספרביליות}}{\geq} w_t\cdot w^* + \frac{1}{\sqrt{t}}\gamma$$

ולכן לאחר M טעויות:

$$W_t \cdot W^* \ge \sum_{t=1}^{M} \frac{1}{\sqrt{t}} \cdot \gamma$$

$$= \gamma \sum_{t=1}^{M} \frac{1}{\sqrt{t}} \ge \gamma \cdot \cancel{M} \cdot \frac{1}{\sqrt{\cancel{M}}} = \boxed{\gamma\sqrt{M}}$$

2. נראה ש- $\|W_t\|_2$ חסום מלמעלה. לכל $t \in [M]$ מתקיים:

$$\|W_{t+1}\|_2^2 \underset{\substack{\uparrow \\ \text{עדכון אחרי טעות}}}{=} \|W_t + \eta_t y_t x_t\|_2^2 \underset{\substack{\downarrow \\ \text{הקשר בין נורמה ומכפלה פנימית ותכונות מכפלה פנימית}}}{=} \|W_t\|_2^2 + 2 W_t \cdot \eta_t x_t y_t + \|\eta_t x_t y_t\|_2^2$$

לפי הנתון, לכל $t \in [M]$ האלגו' טעה כלומר $y_t W_t x_t < 0$

מכיוון שנתון $0 < \eta_t = \frac{1}{\sqrt{t}}$ בפרט נובע $2 W_t \cdot \eta_t x_t y_t$ שלילי ולכן מהסה"כ:

$$\|W_{t+1}\|_2^2 = \|W_t\|_2^2 + 2 W_t \cdot \eta_t x_t y_t + \|\eta_t x_t y_t\|_2^2 \underset{\substack{\downarrow \\ y_i \in \{-1,+1\}}}{\le} \|W_t\|_2^2 + \frac{1}{t}\|x_t\|_2^2$$

$$\underset{\substack{\uparrow \\ \text{ע"פ הנתון} \\ \|x_t\|_2 = 1}}{=} \|W_t\|_2^2 + \frac{1}{t}$$

ולכן לאחר M שגיאות מתקיים:

$$\|W_t\|_2^2 \le \sum_{k=1}^{M} \frac{1}{k} \le \underbrace{\sum_{k=1}^{\infty} \frac{1}{k}}_{\substack{\lim_{n\to\infty} H_n \\ \uparrow \\ \shortparallel}} \le 1 + \ln M$$

ולכן:

$$\gamma\sqrt{M} \le W_t \cdot W^* \underset{\text{ק"ש}}{\le} \|W_t\|_2 \cdot \|W^*\|_2 = \|W_t\|_2 \le \sqrt{1 + \ln M} \underset{\substack{\downarrow \\ M \ge 3}}{\le} \sqrt{2 \ln M}$$

$$\Longrightarrow M \le \frac{2}{\gamma^2}\ln M \underset{\text{לפי הרמז}}{\Longrightarrow} M \le \frac{4}{\gamma^2}\ln\frac{2}{\gamma^2} = O\left(\frac{1}{\gamma^2}\cdot\ln\frac{1}{\gamma^2}\right)$$

**נימוק:**

אם $M=0$ , מכיון ש $\gamma>0$ ברור ש- $M \le \frac{4}{\gamma^2}\ln\left(\frac{1}{\gamma}\right)$

אם $M=1$ אזי, $\gamma \le \|w_t\|_2 \cdot \|w^*\|_2 \le 1$ ולכן

$$1 < 4\cdot\ln 2 \le \frac{4}{\gamma^2}\cdot\ln\frac{2}{\gamma^2}$$

אם $M=2$ אזי, $\|w_t\|_2 \le 1\frac{1}{2}$ ולכן

$$\gamma \le \frac{3}{2\cdot\sqrt{2}}$$

$$M=2 < \frac{4}{\frac{9}{8}}\cdot\ln\frac{2}{\frac{9}{8}} \le \frac{4}{\gamma^2}\ln\frac{2}{\gamma^2}$$

ולכן בסה"כ הוכחנו, לכל $M\ge 0$ מתקיים $M = O\left(\frac{1}{\gamma^2}\ln\frac{1}{\gamma^2}\right)$

■

## שאלה 2:

(a) נראה שהרכבת פונקציה אפינית משמרת קמירות:

כראשית נשים לב לתחום של $g$: $g: C \longrightarrow \mathbb{R}$

כאשר $\mathbb{R}^n \supseteq C := \{Ax+b \mid x\in\mathbb{R}^n\}$ ולכן עלינו קודם להוכיח

ש-$C$ קבוצה קמורה. יהי $\lambda\in(0,1)$ ויהיו $c_1,c_2\in C$.

מהגדרת הקבוצה $C$ נובע שקיימים $x_1,x_2 \in \mathbb{R}^n$ כך ש-

$$\begin{cases} c_1 = Ax_1+b \\ c_2 = Ax_2+b \end{cases}$$

ונראה שמתקיים:

$$\lambda c_1 + (1-\lambda)c_2 = \lambda(Ax_1+b) + (1-\lambda)(Ax_2+b)$$

$$= \underline{\lambda A x_1} + \lambda b + \underline{(1-\lambda)Ax_2} + (1-\lambda)b$$

$$= A(\lambda x_1 + (1-\lambda)x_2) + b$$

ולכן לפי הגדרה $C$ קמורה. נראה ש-$g$ קמורה:

יהי $\lambda\in(0,1)$ ו-$c_1,c_2\in C$ אזי:

$$\underbrace{g(\lambda c_1+(1-\lambda)c_2)}_{\text{הגדרת } g} = f(A(\lambda c_1+(1-\lambda)c_2) + \underbrace{b+\lambda b-\lambda b}_{b(1-\lambda)}) \underset{\text{פתיחת סוגריים}}{=}$$

$$f(\lambda(Ac_1+b)+(1-\lambda)(Ac_2+b)) \underset{f \text{ קמורה}}{\le} \lambda f(Ac_1+b)+(1-\lambda)(Ac_2+b)$$

$$\underset{\text{הגדרת } \varphi}{=} \lambda g(c_1) + (1-\lambda) g(c_2)$$

ולכן לפי ההגדרה $\varphi$ קמורה ■

(b) נניח $f_1, \dots, f_m$ פונ' קמורות כלשהן

ונגדיר $g(x) = \max_{i\in[m]} f_i(x)$, $g: \mathbb{R}^d \to \mathbb{R}$. צ"ל $g$ קמורה.

ברור כי התחום של $g$ – $\mathbb{R}^d$ הינו קבוצה קמורה

(נובע ישירות מתרגול של שבוע)

כעת, יהיו $x_1, x_2 \in \mathbb{R}^d$ ו- $\lambda \in (0,1)$. אזי:

$$g(\lambda x_1 + (1-\lambda) x_2) \underset{\text{הגדרת } g}{=} \max_{i\in[m]} f_i(\lambda x_1 + (1-\lambda) x_2) \underset{\text{קמורה } f}{\leq} \max_{i\in[m]} (\lambda f_i(x_1) + (1-\lambda) f_i(x_2))$$

$$\underset{\substack{\text{תכונות} \\ \text{של מקס'} \\ \text{ומכך שאי-שלילי}}}{\leq} \lambda \max_{i\in[m]} f_i(x_1) + (1-\lambda) \max_{i\in[m]} f_i(x_2) \underset{\text{הגדרת } g}{=} \lambda g(x_1) + (1-\lambda) g(x_2)$$

ולכן לפי ההגדרה גם $g$ קמורה. ■

(C) ברור כי $\mathbb{R}$ קבו' קמורה מהיותה של $\mathbb{R}$ שלמה.

נשתמש במשפט מתרגו"ל על קמירות של פונ'

גזירות פעמיים במשתנה יחיד:

$$\ell_{log}(z)' = \frac{-e^{-z}}{(1+e^{-z})\ln(2)}$$

נשים לב שגם הנגזרת הראשונה גזירה בכל התחום

ונחשב נגזרת שניה:

$$l_{log}(z)'' = \frac{e^{-z}\cdot(1+e^{-z})\cancel{\ln 2} - e^{-z}\cdot e^{-z}\cancel{\ln 2}}{(1+e^{-z})^2 \ln^{\cancel{2}} 2}$$

$$= \frac{e^{-z}(1+\cancel{e^{-z}}-\cancel{e^{-z}})}{(1+e^{-z})^2 \ln 2} = \frac{e^{-z}}{(1+e^{-z})\ln 2} > 0$$

ולכן לפי המשפט $l_{log}$ קמורה.

כעת מסעיף $(a)$ עם $b=\bar{0}$ ו$_i$- $A$ מטריצה

אלכסונית שכל איברי האלכסון שלה טהורים ושווים ל$_i$- $xy$

נקבל שגם $l_{log}(ywx)$ קמורה ביחס ל$_i$- $w$.

∎

## שאלה 3:

(א) בשאלה הקודמת הוכחנו שמקסימום של פונ' קמורות הינה פונ' קמורה. לכן מספיק להוכיח שסכום של קמורות נותן פונ' קמורה, וש- $1-\text{sgn}(y_j-y_r)w(x_j-x_r)$ פונ' קמורה ונקבל שה- hinge loss קמורה, גם כן.

נסמן $f(w)=1-\text{sgn}(y_j-y_r)w(x_j-x_r)$ זו פונ' שתחומה $\mathbb{R}^d$ ונראה כמו שצו' קב' קמורה. יה"י $\lambda\in(0,1)$ ו- $w_1,w_2\in\mathbb{R}^d$

אזי:

$$f(\lambda w_1+(1-\lambda)w_2) \underset{\text{ליניאריות } f}{=} 1-\text{sgn}(y_j-y_r)(\lambda w_1+(1-\lambda)w_2)(x_j-x_r)$$

$$= 1-\text{sgn}(y_j-y_r)[\lambda w_1(x_j-x_r)+(1-\lambda)w_2\cdot(x_j-x_r)] =$$

$$= 1-\lambda\cdot\text{sgn}(y_j-y_r)(w_1\cdot(x_j-x_r))+(1-\lambda)\text{sgn}(y_j-y_r)(w_2\cdot(x_j-x_r))$$

$$= \lambda+1-\lambda-\lambda\cdot\text{sgn}(y_j-y_r)(w_1(x_j-x_r))-(1-\lambda)\text{sgn}(y_j-y_r)(w_2(x_j-x_r))$$

$$= \lambda(1-\text{sgn}(y_j-y_r)(w_1(x_j-x_r)))+(1-\lambda)(1-\text{sgn}(y_j-y_r)(w_2(x_j-x_r)))$$

$$= \lambda f(w_1)+(1-\lambda)f(w_2)$$

ולכן לפי הגדרה $f$ קמורה ב-$w$.

נראה שסכום של קמורות נותן קמורה:

נוכיח באינדוקציה שלמה על מס' הפונ' בסכום.

מסק' - אם f היא פונ' קמורה יש רק אות הסכום
סיימני.

שלב - נניח נכונות לסכום של $n-1>1$ פונ' ונראה ל-$n$

תהיינה $f_1,\dots,f_n$ קמורות על הקבוצה $C$ שהיא קבו' קמורה.

אזי: הסכום $\sum_{i=1}^{n} f_i = S$ שווה ל- $\underbrace{\sum_{i=1}^{n-1} f_i}_{\text{הנ"ל}} + \underbrace{f_n}_{\text{קמורה עפ"י הנתון}}$

ומהנחה מה.ג. סכום של 2 פונ' קמורות היא פונ' קמורה.

ולכן מההנ"ל הסכום $S$ היא פונ' קמורה.

ולכן מההנ"ל הראנו את כל מה שצריך וכעת נותר שה hinge loss

הנתונה היא פונ' קמורה ∎

(b) כעת נראה, $\Delta(h_w(\bar{x}),y)\le \ell(h_w(\bar{x}),y)$.

יהיו $w\in\mathbb{R}^d$ ו- $\bar{x}\in\mathcal{X}^k$ ו- $y\in\mathbb{R}^k$. מתקיים:

$$\Delta(h_w(\bar{x}^i),y^i) = \max\{0,\ 1-\overbrace{sgn(y_j-y_r)\cdot w(x_j-x_r)}^{\#}\} = \begin{cases}\alpha(y), & \#<1\\ 0, & else\end{cases}$$

$$= \begin{cases}\alpha(y), & sgn(y_j-y_r)\cdot sgn(y'_j-y'_r)|y'_j-y'_r|<1\\ 0, & else\end{cases} = \begin{cases}\beta(y)+1, & sgn(y_j-y_r)\neq sgn(y'_j-y'_r)\\ \delta(y), & sgn(y_j-y_r)=sgn(y'_j-y'_r) \text{ and } |y'_j-y'_r|<1\\ 0, & else\end{cases}$$

$$\ge \begin{cases} 1, & sgn(y'_j - y'_r) \neq sgn(y_j - y_r) \\ 0, & else \end{cases} = \mathbb{1}\{sgn(y'_j - y'_r) \neq sgn(y_j - y_r)\}$$

הגדרת אינדיקטור

כי $y$ ... $\ge 0$

ולכן מהגדרות שני הלוסים קיבלנו

$\Delta(h_w(\bar{x}), y) \le \ell(h_w(\bar{x}), y)$ כמבוקש. ■

(c) נראה שאם מביאים למינימום את ה-hinge loss נקבל היפותזה שתביא את ה-Kendal-Tau loss למינימום גם כן.

מהספרביליות עם שוליים $\gamma$ נובע שקיים $W^* \in \mathbb{R}^d$ כך ש-:

$\gamma \le sgn(y_j - y_r) W^* \cdot (X_j - X_r)$. נבחר $\tilde{W} = \frac{W^*}{\gamma}$ ונקבל:

$$sgn(y_j - y_r) \cdot \tilde{W} \cdot (X_j - X_r) = sgn(y_j - y_r) \cdot \frac{W^*}{\gamma} \cdot (X_j - X_r) \ge 1$$

נעביר אגף ונקבל:

$$1 - sgn(y_j - y_r) \cdot \tilde{W}(X_j - X_r) \le 0$$

$$\Longrightarrow \max\{0, 1 - sgn(y_j - y_r) \cdot \tilde{W}(X_j - X_r)\} = 0$$

ה-hinge loss היא פונקציה אי-שלילית ורציפה ולכן ערך המינימום שלה הוא אכן ה-0. מה שהראנו שקיבלנו בסעיף הקודם ומהעובדה שה-Kendal-Tau אי שלילי (כי אינדיקטור לכן אי-שלילי) נובע שהוקטור שמביא את ה-hinge loss למינימום (=0)

הוא גם מבטא את ה- KendallTau loss. ושוב מכיוון שלילי גורר
שהמינימום של הוא 0 ולכן שתי הפונ' "מסכימות"
על המינימום כנדרש. ∎

## שאלה 4:

תהיה $f:\mathbb{R}^n \to \mathbb{R}$ פונ' $\beta$-חלקה שמקיימת

$f(x) \geq 0$, $\forall x \in \mathbb{R}^n$. תחת הנחות השאלה, נראה

כי אם $\eta < \frac{2}{\beta}$ אזי $\lim_{t\to\infty} \|\nabla f(x_t)\|_2 = 0$.

מהגדרת $\beta$-חלקות מתקיים לכל $x_t, x_{t+1}$:

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^T (x_{t+1} - x_t) + \frac{\beta}{2}\|x_t - x_{t+1}\|_2^2$$

ומאופן פעולת אלג' GD נקבל: $x_{t+1} - x_t = -\eta \nabla f(x_t)$

ולכן: $f(x_{t+1}) - f(x_t) \leq \nabla f(x_t)^T(-\eta \nabla f(x_t)) + \frac{\beta}{2}\|\eta \nabla f(x_t)\|_2^2$

מתכונות מכפלה פנימית ונורמת $L_2$:

$$f(x_{t+1}) - f(x_t) \leq -\eta \|\nabla f(x_t)\|_2^2 + \frac{\beta}{2}\eta^2 \|\nabla f(x_t)\|_2^2$$

$$= \|\nabla f(x_t)\|_2^2 \left(\frac{\beta}{2}\eta^2 - \eta\right)$$

מהנחות השאלה $0 < \eta < \frac{2}{\beta}$ ולכן אס״ק

$\eta - \eta^2 \frac{\beta}{2} > 0$ אס״ק

$$\frac{f(x_{t+1}) - f(x_t)}{\frac{\beta}{2}\eta^2 - \eta} \geq \|\nabla f(x_t)\|_2^2 \geq 0 \quad (*)$$

פונ׳ כינוע׳

וזה מתקיים כמובן לכל $t$.

נתבונן בסכום החלקי $\sum_{t=0}^{T}\|\nabla f(x_t)\|_2^2$ עבור $T \ge 0$ כלשהו

מ- ⓐ נקבל ע"י סכום טלסקופי ש-:

$$\overbrace{f(x_0)\left(\eta-\frac{\beta}{2}\eta^2\right)}^{\text{הביטוי חסום ולא תלוי ב-}T} \underset{\substack{f\ge 0\\ \text{וגם } (\#)>0}}{\ge} \left(f(x_0)-f(x_{T+1})\right)\cdot\underbrace{\left(\eta-\frac{\beta}{2}\eta^2\right)}_{\#} \ge \sum_{t=0}^{T}\|\nabla f(x_t)\|_2^2$$

הסמני את הסכומים החלקיים של הטור החיובי הנ"ל חסומים

ולכן לפי משפט מחדו"א הטור מתכנס ומתנאי הכרחי

להתכנסות טור אינסופי נקבל $\lim_{t\to\infty}\|\nabla f(x_t)\|_2^2=0$

ומאריתמטיקה של גבולות נקבל $\lim_{t\to\infty}\|\nabla f(x_t)\|_2=0$

כנדרש. ∎

# חלק תכנותי

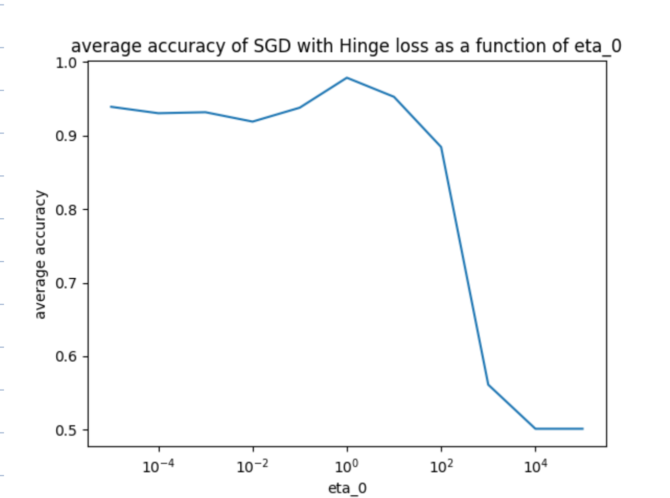
1. התוצאות שהתקבלו עבור הרצת הקוד:

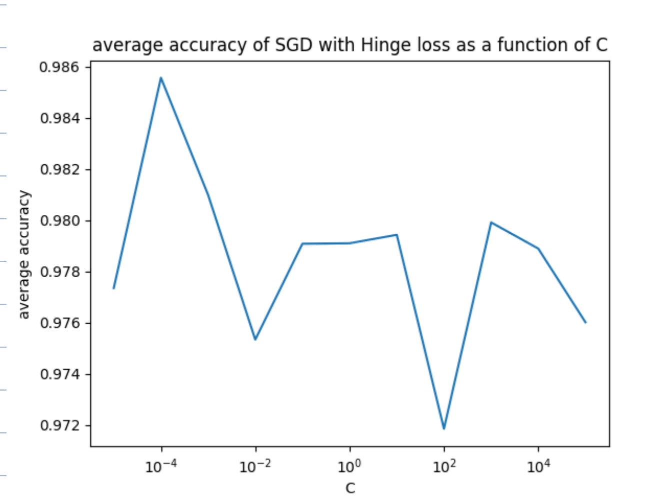
```
Q1 results:
optimal eta_0:  1
optimal C:  0.0001
accuracy of the best classifier:  0.9923
```

כלומר מה Cross validation קיבלנו ש learning rate

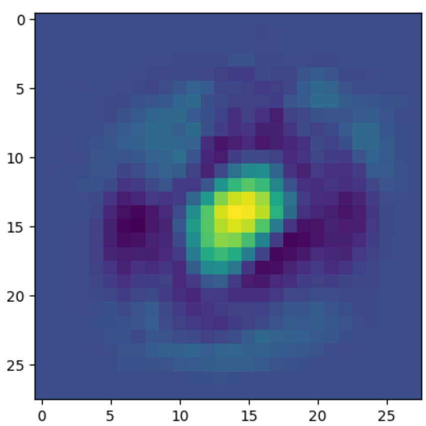
האופטימלי הוא $\eta_0 = 1$ . ובהינתן $\eta_0 = 1$

קיבלנו שהקבוע הטוב ביותר הוא $C = 10^{-4}$



best classifier Was image



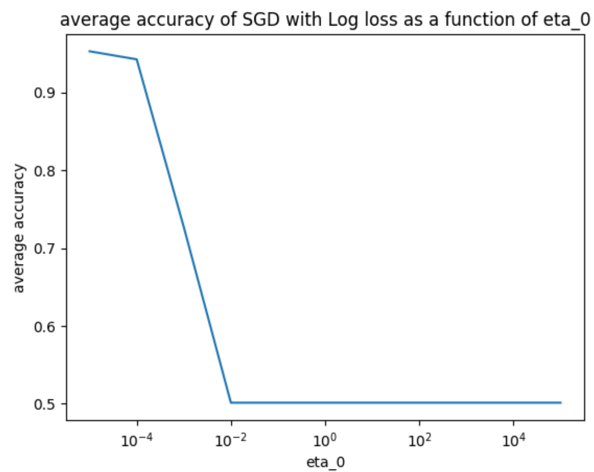
ניתן לראות כי הפיקסלים באמצע שכמעט
בהיר מעידים על הספרה 8 (על החלק
ה"אמצעי" של הספרה) ואילו הפיקסלים הכהים
מעידים על הספרה 0 .

וכפי שראינו לעיל רמת הדיוק של המסווג
טובה מאוד : 0.9923

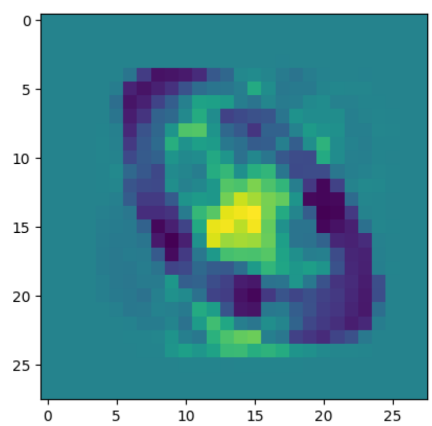
```
Q2 results:
optimal eta_0:  1e-05
accuracy of the best classifier:  0.9678
Ram
```

2. לאחר הרצת הקוד התקבלו:

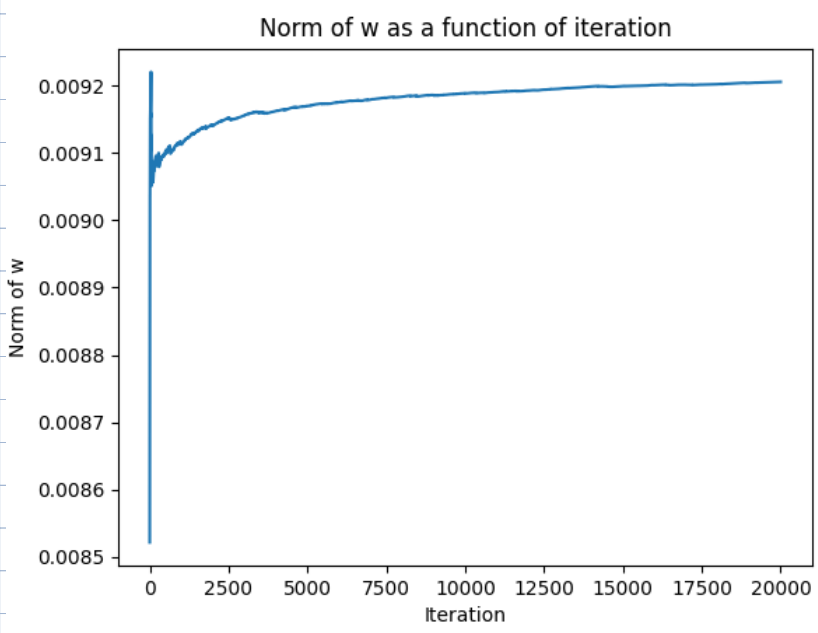
כלומר ה learning rate האופטימלי שהתקבל
מה Cross Validation הוא $\eta_0 = 10^{-5}$ ורמת הדיוק
גם כאן יחסית גבוהה: 0.9678 אך פחות
מזו שהתקבלה עם ה Hinge loss.




best classifier w as image



כאן ניתן לראות בבירור שהפיקסלים הבהירים
משפיעים על הסיווג 8 ואילו הכהים על 0,
לבסוף עבור הנורמה של w קיבלנו:




הסבר: SGD ממזער את ה-log-loss לאחר
מספר מועט של איטרציות ולכן הנורמה מפסיקה
לשנות את גודלה באופן משמעותי לאחר מעט צעדים.
כמו כן, בכל צעד ה-learning rate קטן ביחס
הפוך ל-t ולכן השינוי בנורמה (שתלויה ב-η)
וקטן ככל שמגדילים את t.