פתרון שיעורי בית 3 – רוני קוסיצקי 205817893

<u>שאלה 1</u>

מהגדרת הקמירות בהינתן שf קמורה מתקיים:

$$\forall \boldsymbol{w_1}, \boldsymbol{w_2} \in \mathbb{R}^n, \forall \alpha \in [0,1]:$$

$$f(\alpha \boldsymbol{w_1} + (1-\alpha)\boldsymbol{w_2}) \le \alpha f(\boldsymbol{w_1}) + (1-\alpha) f(\boldsymbol{w_2})$$

יהי $\alpha \in [0,1]$, $\boldsymbol{u_1}, \boldsymbol{u_2} \in \mathbb{R}^n$, ולכן מתקיים עבור g:

$$g(\alpha \boldsymbol{u_1} + (1-\alpha)\boldsymbol{u_2}) = f(A(\alpha \boldsymbol{u_1} + (1-\alpha)\boldsymbol{u_2}) + \boldsymbol{b}) = f(\alpha A\boldsymbol{u_1} + (1-\alpha)A\boldsymbol{u_2} + \boldsymbol{b})$$

נגדיר $\boldsymbol{w_1} = A\boldsymbol{u_1} + \boldsymbol{b}, \boldsymbol{w_2} = A\boldsymbol{u_2} + \boldsymbol{b}$ ולכן:

$$g(\alpha \boldsymbol{u_1} + (1-\alpha)\boldsymbol{u_2}) = f(\alpha A\boldsymbol{u_1} + (1-\alpha)A\boldsymbol{u_2} + \boldsymbol{b}) = f(\alpha A\boldsymbol{u_1} + \alpha\boldsymbol{b} + (1-\alpha)A\boldsymbol{u_2} + (1-\alpha)\boldsymbol{b})$$
$$= f(\alpha \boldsymbol{w_1} + (1-\alpha)\boldsymbol{w_2})$$

מהקמירות של f ומהתכונה הראשונה שציינו נובע ש:

$$g(\alpha \boldsymbol{u_1} + (1-\alpha)\boldsymbol{u_2}) = f(\alpha \boldsymbol{w_1} + (1-\alpha)\boldsymbol{w_2}) \le \alpha f(\boldsymbol{w_1}) + (1-\alpha) f(\boldsymbol{w_2}) = \alpha f(A\boldsymbol{u_1} + \boldsymbol{b}) + (1-\alpha) f(A\boldsymbol{u_2} + \boldsymbol{b})$$
$$= \alpha g(\boldsymbol{u_1}) + (1-\alpha) g(\boldsymbol{u_2})$$

ולכן קיבלנו שמתקיים עבורם u_1, u_2

$$g(\alpha \boldsymbol{u_1} + (1-\alpha)\boldsymbol{u_2}) \le \alpha g(\boldsymbol{u_1}) + (1-\alpha) g(\boldsymbol{u_2})$$

ולכן g קמורה על פי הגדרה.

שאלה 2

סעיף א

בהינתן שכל f_i קמורה מתקיים:

$$f_i(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f_i(x_1) + (1-\alpha)f_i x_2 \quad for\ \alpha \in [0,1]$$

ניקח מקסימום על שני האגפים ונקבל:

$$\max_i\{f_i(\alpha x_1 + (1-\alpha)x_2)\} \leq \max_i\{\alpha f_i(x_1) + (1-\alpha)f_i x_2\}$$

קל לראות שמתקיים:

$$\max_i\{\alpha f_i(x_1) + (1-\alpha)f_i(x_2)\} \leq \max_i\{\alpha f_i(x_1)\} + \max_i\{(1-\alpha)f_i(x_2)\}$$

ולכן:

$$\max_i\{f_i(\alpha x_1 + (1-\alpha)x_2)\} \leq \max_i\{\alpha f_i(x_1)\} + \max_i\{(1-\alpha)f_i(x_2)\}$$

ולכן נוכל לכתוב את $f(x)$ בצורה הבאה:

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2)$$

ולכן היא קמורה על פי הגדרה.

סעיף ב

על פי הגדרה, בהינתן נקודה $x \in C$ כאשר C תחום קמורה ופתוחה(במקרה שלנו $\mathbb{R}^d \to \mathbb{R}$) ולכל $w \in C$ מתקיים שקיים $z \in \mathbb{R}^d$ כך ש $\forall u \in C\ f(u) \geq f(w) + < u-w, z >$, אזי הפונקציה הזו תהיה קמורה.

בנוסף, על פי הגדרה לכל $w \in C$ כל $z \in \mathbb{R}^d$ נקראת הסאב גרדיאנט של f. בהנתן שהפונקציה גזירה אזי הסאב גרדיאנט שלה בנקודה w כלשהיא תהיה הגרדיאנט שלה בנקודה.
הוכחנו בסעיף א', שמתקיים שהפונקציה g היא קמורה. באופן ספציפי בכל נקודה w היא תחזיר את הפונקציה המקסימלית בנקודה זו. מכיוון שכולן קמורות וגזירות לפי הנתון אזי באופן ספציפי היא תחזיר את הגרדיאנט של הפונקציה המקסימלית בנקודה זו.

בסך הכול נקבל שסאב הגרדיאנט תהיה קבוצה של כל הגרדיאנטים של הפונקציות המקסימליות המתקבלות לאורך התחום שלנו.

<u>שאלה 3</u>

<u>סעיף א</u>

ראינו בסעיף הקודם שמקסימום של פונקציות קמורות היא פונקציה קמורה. לכן, נשאר לנו להראות שhinge function היא פונקציה קמורה וסיימנו.

בהינתן $f(x) =< x, w >$ נרצה להראות שהפונקציה קמורה. עפ"י הגדרה פונקציה היא קמורה אם

$$f(\alpha w_1 + (1-\alpha)w_2) \leq \alpha f(w_1) + (1-\alpha)f(w_2)$$

ולכן צריך להוכיח שמתקיים:

$$< (\alpha w_1 + (1-\alpha)w_2), w > \leq < \alpha f(w_1), w > + < (1-\alpha)f(w_2), w >$$

ושוויון נובע ישירות מלינאריות הפונקציה ותכונות המכפלה הפנימית.

<u>סעיף ב</u>

נרצה להוכיח שפונקציית ההפסד חוסמת מלמעלה את פונקציית Δ_{zo}.

נגדיר

$$L(j) \coloneqq \Delta_{zo} + (w_j x - w_y x)$$

כאשר נציב $\hat{y}$ במשוואה נקבל ש

$$L(\hat{y}) = \Delta_{zo}(\hat{y}) + (w_{\hat{y}} x - w_{\hat{y}} x)$$

אבל $w_{\hat{y}} x \geq w_{\hat{y}} x$ מכיוון ש$\hat{y}$ הוא בהגדרה המקסימום של w_i עבור כל הi. לכן נקבל ש

$$\max_{j \in [L]} L(j) \geq L(\hat{y}) \geq \Delta_{zo}$$

<u>סעיף ג</u>

ידוע שלסט האימון שלנו קיים משקולות כך ששגיאת האימון תהיה 0. ראינו בסעיף הקודם שהשגיאה של שפונקציית ההפסד חוסמת את Δ_{zo} ולכן, אם נאמן את המודל שלנו בעזרת סט האימון נוכל להגיע לשגיאה אבסולוטית של 0 – אנו יודעים שקיימות משקולות כאלה שעבורן זה מתקיים, ובנוסף אנו יכולים להגיע עד כדי ϵ עבור כל טעות.

מהעובדה שנחסום את השגיאה ב0 אנו נחסום גם את השגיאה של Δ_{zo} ב0 ולכן נוכל להשיג שגיאה אופטימלית עבור מדגם האימון הנתון עבור פונקציית השגיאה הזו גם כן.

<u>שאלה 4</u>

<u>טענת עזר</u> – אם v ע"ע של A אזי v הוא ע"ע של $I - \eta A$

<u>הוכחה</u> – מהגדרת ערך עצמי מתקיים $\lambda v = Av$. עבור λ מתקיים ש

$$(I - \eta A)v = v - 2\eta Av = (1 - 2\eta\lambda)v$$

לכן קיבלנו שv וקטור עצמי עם ערך עצמי של $1 - 2\eta\lambda$

<u>טענת עזר</u> – אם v וקטור עצמי של A אזי הוא גם וקטור עצמי של A^t.

<u>הוכחה</u> – נעזר שוב בהגדרת וקטור עצמי ונראה ש:

$$A^t v = A^{t-1} \cdot Av = A^{t-1} \cdot \lambda v = \lambda \cdot A^{t-1} v = \cdots = \lambda^t v$$

<u>טענת עזר</u> – בהנחה שA סימטרית, אם v הוא וקטור עצמי של A עם ע"ע של λ אזי הוא גם וקטור עצמי של המטריצה $((I - 2\eta A)^t)^T A(I - 2\eta A)$ עם הע"ע $((I - 2\eta A)^t)^T A(I - 2\eta A)$.

<u>הוכחה</u> – משתי טענות העזר הקודמות אנו יודעים שv הוא וקטור עצמי של $(I - 2\eta A)^t$. מתקיים:

$$\begin{aligned} ((I - 2\eta A)^t)^T A(I - 2\eta A)^t \cdot v &= ((I - 2\eta A)^t)^T A(I - 2\eta\lambda)^t \cdot v \\ &= (I - 2\eta\lambda)^t \cdot ((I - 2\eta A)^t)^T A \cdot v = (I - 2\eta\lambda)^t \cdot ((I - 2\eta A)^t)^T \cdot \lambda v \\ &= \lambda(I - 2\eta\lambda)^t \cdot ((I - 2\eta A)^t)^T \cdot v \end{aligned}$$

מכיוון שA היא סימטרית אזי גם $(I - 2\eta A)^t$. לכן מתקיים $((I - 2\eta A)^t)^T = (I - 2\eta A)^t$ ובסך הכל נקבל:

$$\begin{aligned} ((I - 2\eta A)^t)^T A(I - 2\eta A)^t \cdot v &= \lambda(I - 2\eta\lambda)^t \cdot ((I - 2\eta A)^t)^T \cdot v = \lambda(I - 2\eta\lambda)^t \cdot (I - 2\eta A)^t v \\ &= \lambda(I - 2\eta\lambda)^{2t} \end{aligned}$$

<u>ההוכחה:</u>

מתקיים ש: $\nabla f(w) = (A + A^T)w$ ולכן עבור כל צעד של GD מתקיים:

$$w_{t+1} = w_t - \eta(A + A^T)w_t = \left(I - \eta(A + A^T)\right)w_t$$

נתון שהמטריצה היא PSD ולכן סימטרית. לכן $A + A^t = 2A$ ונקבל ש- $w_{t+1} = (I - 2\eta A)w_t$

לכן, אנו יודעים ש- $w_{t+1} = (I - 2\eta A)w_t$.

ניקח את הבסיס $\mathbb{R}^n$ כווקטורים העצמיים של המטריצה. נסמן את הבסיס $\{v_1, \ldots, v_n\}$. מכיוון שזה בסיס מתקיים:

$$w_1 = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = \sum_{i=1}^{n} \alpha_i v_i$$

ובסך הכול נקבל:

$$f(w_t) = f((I-2\eta A)^t w_1) = ((I-2\eta A)^t w_1)^T A (I-2\eta A)^t w_1 = w_1^T ((I-2\eta A)^t)^T A (I-2\eta A)^t w_1$$
$$= w_1^T ((I-2\eta A)^t)^T A (I-2\eta A) \cdot \sum_{i=1}^{n} \alpha_i v_i = \sum_{i=1}^{n} \alpha_i w_1^T ((I-2\eta A)^t)^T A (I-2\eta A)^t v_i$$

מטענת העזר נקבל שאנו יודעים לכל i שהווקטור v_i הוא וקטור עצמי של $((I-2\eta A)^t)^T A (I-2\eta A)^t$ עם הערך העצמי שמצאנו. לכן בסך הכול אנו מקבלים:

$$f(w_t) = \sum_{i=1}^{n} \lambda_i (1-2\eta\lambda_i)^{2t} w_1^T \alpha_i v_i = \|w_1\|^2 \sum_{i=1}^{n} \lambda_i (1-2\eta\lambda_i)^{2t}$$

אם נרצה את החסם $f(w_t)$ אנו צריכים שעבור $t = O(\log\left(\frac{1}{\epsilon}\right))$ נקבל:

$$f(w_t) = \|w_1\|^2 \sum_{i=1}^{n} \lambda_i (1-2\eta\lambda_i)^{2t} \le \epsilon \to \sum_{i=1}^{n} \lambda_i (1-2\eta\lambda_i)^{2t} \le \frac{\epsilon}{\|w_1\|^2}$$

נחסום את $\sum_{i=1}^{n} \lambda_i (1-2\eta\lambda_i)^{2t}$ בעזרת $\lambda_{low} = \min_i \lambda_i \,, \lambda_{high} = \max_i \lambda_i$.

$$\sum_{i=1}^{n} \lambda_i (1-2\eta\lambda_i)^{2t} \le \lambda_{high} \sum_{i=1}^{n} (1-2\eta\lambda_i)^{2t} \le \lambda_{high} \cdot n \cdot (1-2\eta\lambda_{low})^{2t}$$

נדרוש –

$$(1-2\eta\lambda_{low})^{2t} \le \frac{\epsilon}{\lambda_{high} n \|w_1\|^2} \Rightarrow 2t \le \frac{\ln\left(\frac{\epsilon}{\lambda_{high} n \|w_1\|^2}\right)}{\ln\,(1-2\eta\lambda_{low})} \Rightarrow$$

$$t \le \frac{\ln\left(\frac{\epsilon}{\lambda_{high} n \|w_1\|^2}\right)}{2\ln\,(1-2\eta\lambda_{low})}$$

נבחר את η להיות $\eta = \frac{1-\frac{1}{e}}{2\lambda_{low}}$ ונקבל:

$$t \le -\frac{1}{2}\ln\left(\frac{\epsilon}{\lambda_{high} n \|w_1\|^2}\right) = \frac{1}{2}\ln\left(\frac{\lambda_{high} n \|w_1\|^2}{\epsilon}\right) = \frac{1}{2}\ln\left(\frac{1}{\epsilon} \cdot \lambda_{high} n \|w_1\|^2\right)$$
$$= \frac{1}{2}\ln\left(\frac{1}{\epsilon}\right) + \frac{1}{2}\ln\left(\lambda_{high} n \|w_1\|^2\right)$$

$\lambda_{high}, n, \|w_1\|^2$ הם קבועים. לכן קיים η עבורו אנו יכולים לקבל $f(w_t) \le \epsilon$ עבור $t \le O\left(\log\left(\frac{1}{\epsilon}\right)\right)$ כרצוי.

שאלה 5

יהיה k איטרציה כלשהיא במהלך ריצת הGD ולכן עפ"י הגדרת האלגוריתם אנו יודעים ש

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

לכן, עבור הפונקציה שלנו, נקבל ש:

$$f(x_{k+1}) - f(x_k) = f\big(x_k - \eta \nabla f(x_k)\big) - f(x_k) + \eta \left||\nabla f(x_k)|\right|^2 \ \leq \ + \frac{\beta \eta^2}{2} \left||\nabla f(x_k)|\right|^2$$

אם נבחר $\eta = \frac{1}{\beta}$ נקבל ש:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta} \left||\nabla f(x_k)|\right|^2$$

ולכן בגרדיאנט הk אנו מקבלים ש:

$$f_* - f(x_k) \leq f(x_k) - f(x_0) \leq \frac{1}{2\beta} \sum_{i=0}^{k-1} \left||\nabla f(x_k)|\right|^2$$

ולכן נקבל בסה"כ ש

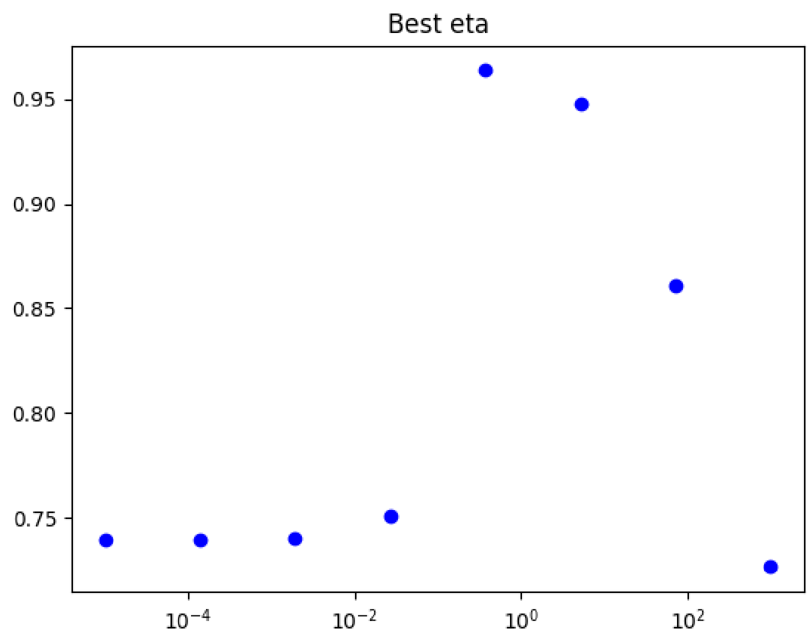
$$\sum_{i=0}^{k-1} \left||\nabla f(x_k)|\right|^2 \leq 2\beta\big(f(x_0) - f_*\big)$$

אנו רואים שהסכום הוא איבר סופי(כי הוא קטן מערך סופי כלשהוא). לכן, אם הסכום של סדרה כלשהיא הוא סופי אזי בפרט ממשפט בסיסי שלמדנו בחדוו"א מתקיים ש $0 \dashrightarrow \nabla f(x_k)$ כאשר k יישאף לאינסוף.

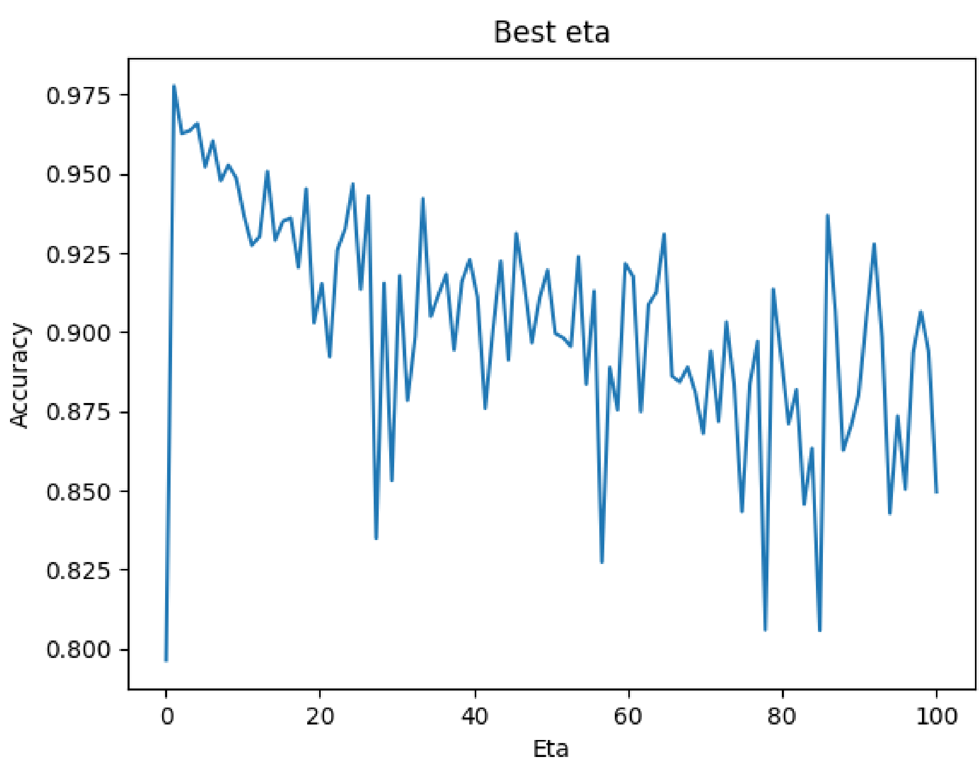
חלק תכנותי - Hinge

סעיף א

צמצמתי את טווח החיפוש המוצע כי נכנסו ל*Overflow*. נקבל את הגרף:

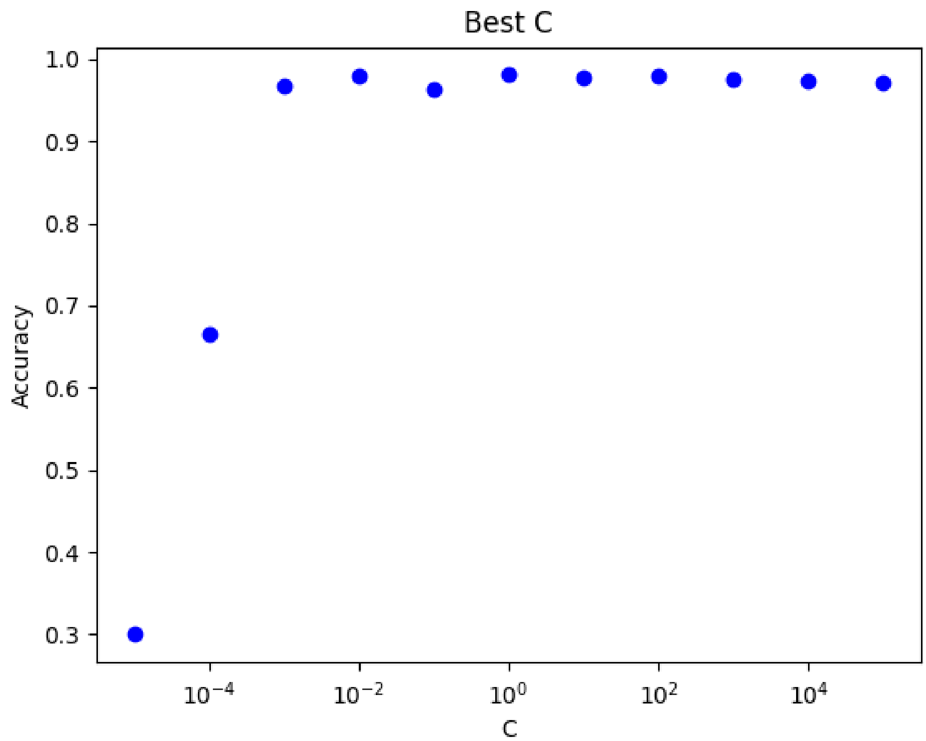


כעת, נתמקד באזורים הגבוהים כדי למצוא את הערך האופטימלי יהיה בערך 1.1:

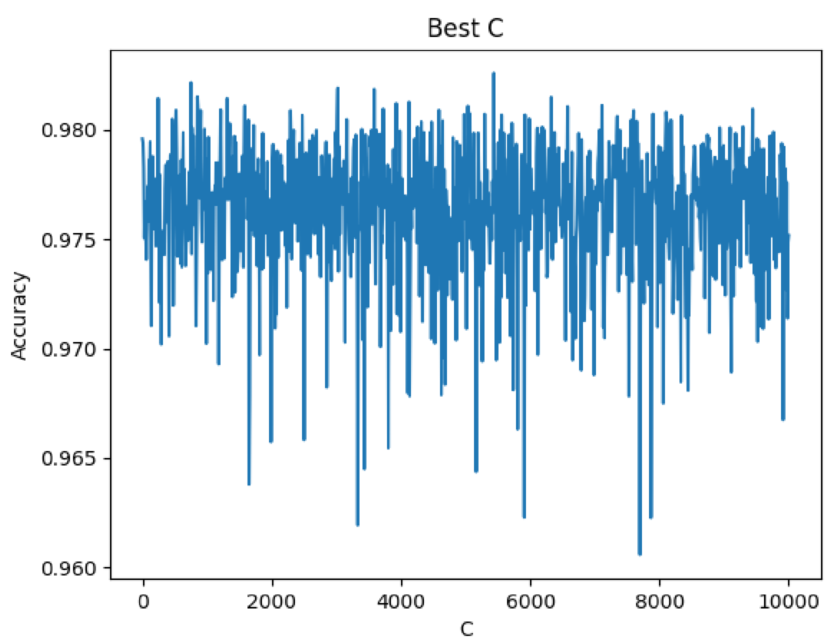


סעיף ב

נריץ את הערכים על ציר לוגריתמי ונקבל:



אנו רואים שהתחום האופטימלי הוא בין 1 ל10,000 ולכן נבצע חיפוש ממוקד שם ונקבל:



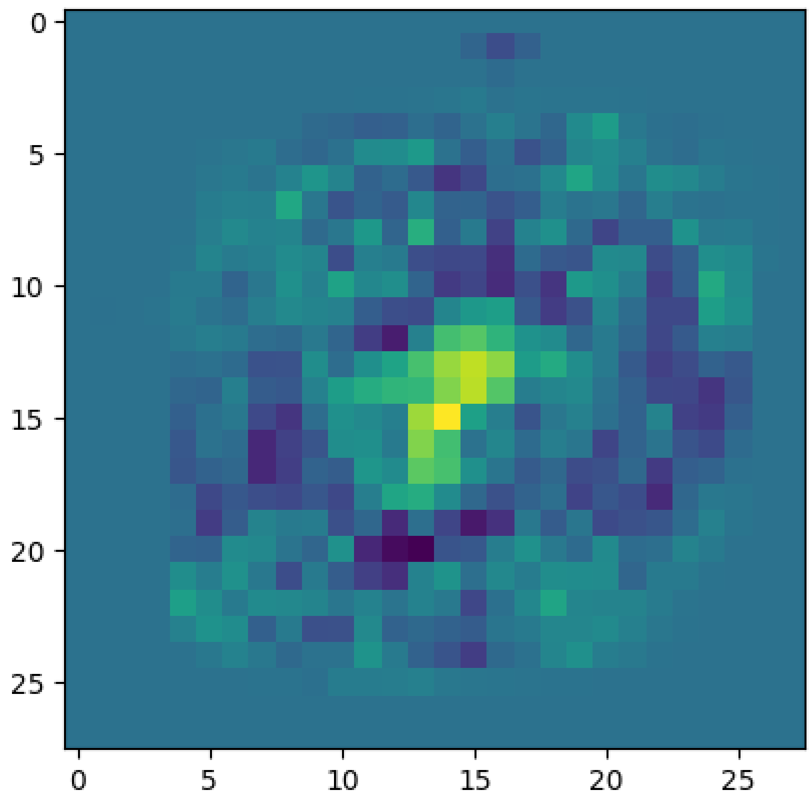
הערך המקסימלי מתקבל עבור 5435.48

סעיף ג

לאחר שתי הסעיפים הקודמים אנו מקבלים שהפרמטרים האופטימליים עם

$\eta = 1.1, C = 5435.48$

נאמן את המודל עם 20,000 נקודות ונוציא פלט של התמונה:



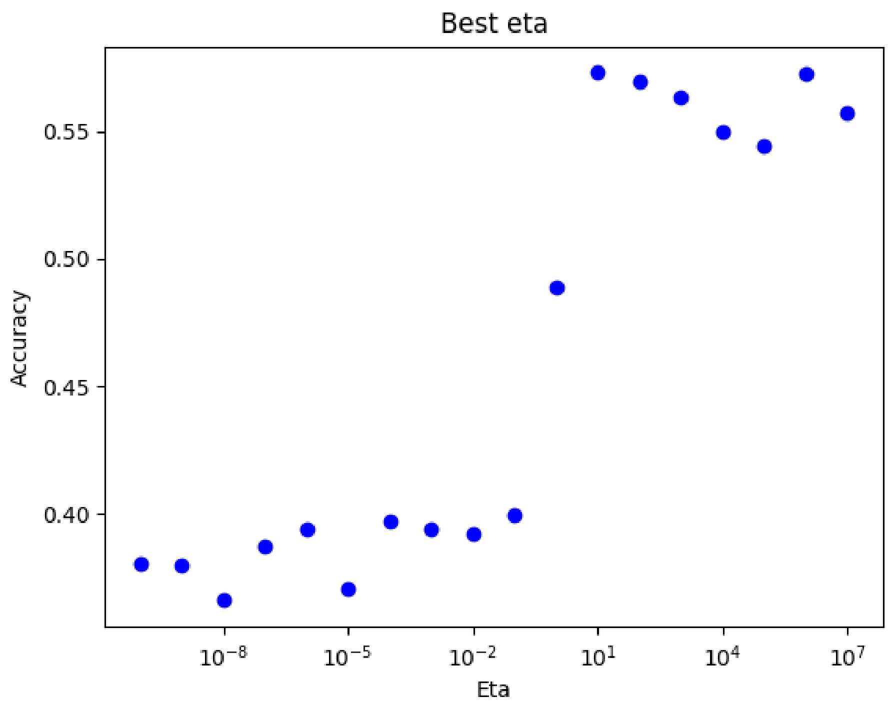
סעיף ד

נקבל שהדיוק יהיה 98.6%

חלק תכנותי - Entropy

סעיף א

נחזור על התהליך עם פונקציית ההפסד החדשה:

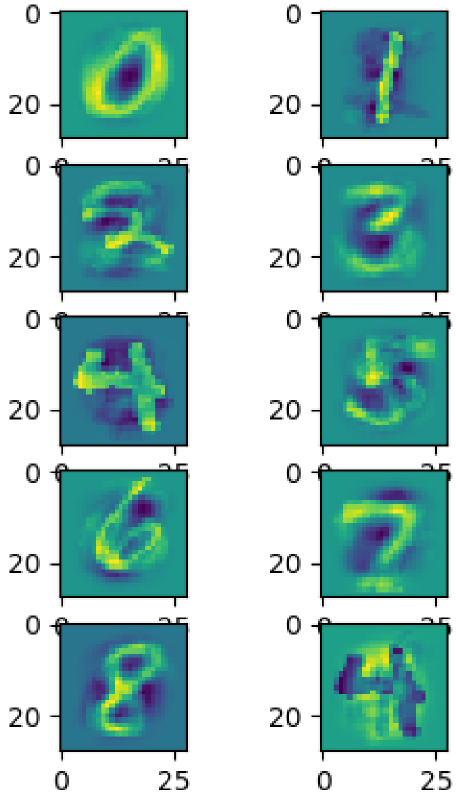


1

נרוץ על הטווח המצומצם ונקבל שהדיוק המקסימלי יתקבל עבור

$\eta \cong 82 * 10^4$

סעיף ב



סעיף ג

נקבל דיוק של בערך כ64%