Theory Questions

Remark: Throughout this exercise, when we write a norm $\|\cdot\|$ we refer to the ℓ_2-norm.

1. **(15 points) Convex functions.**

   (a) Let $f : \mathbb{R}^n \to \mathbb{R}$ a convex function, $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Show that, $g(\mathbf{x}) = f(A\mathbf{x} + b)$ is convex.

יהיו $w_1, w_2 \in \mathbb{R}^n$, $\lambda \in [0,1]$.

נצטרך להראות שמתקיים: $g(\lambda w_1 + (1-\lambda) w_2) \leq \lambda \cdot g(w_1) + (1-\lambda) \cdot g(w_2)$

אכן:

$$g(\lambda w_1 + (1-\lambda) w_2) = f\left[A(\lambda w_1 + (1-\lambda) w_2) + b\right] = f\left(\lambda A w_1 + (1-\lambda) \cdot A w_2 + b\right)$$

$$= f(\lambda A w_1 + A w_2 - \lambda A w_2 + b) \overset{+\lambda b - \lambda b}{=} f(\lambda A w_1 + \lambda b + A w_2 + b - \lambda A w_2 - \lambda b)$$

$$= f\left(\lambda(A w_1 + b) + (1-\lambda)(A w_2 + b)\right) \overset{f \text{ is convex}}{\leq} \lambda \cdot f(A w_1 + b) + (1-\lambda) \cdot f(A w_2 + b)$$

$$\overset{\text{הגדרת } g}{=} \lambda \cdot g(w_1) + (1-\lambda) \cdot g(w_2) \quad \blacksquare$$

(b) Consider m convex functions $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$, where $f_i : \mathbb{R}^d \to \mathbb{R}$. Now define a new function $g(\mathbf{x}) = \max_i f_i(\mathbf{x})$. Prove that $g(\mathbf{x})$ is a convex function. (Note that from (a) and (b) you can conclude that the hinge loss over linear classifiers is convex.)

נתון באופן כללי פונקציה f, ויהיו $w_1, w_2 \in \mathbb{R}^n$, $\lambda \in [0,1]$.

צריך שמתקיים: $g(\lambda w_1 + (1-\lambda) w_2) \leq \lambda \cdot g(w_1) + (1-\lambda) g(w_2)$

נשים לב ש g מוגדרת להיות המקסימום על פני כל f_i המקבלת $\lambda w_1 + (1-\lambda) w_2$.

מתקיים:

$$g(\lambda w_1 + (1-\lambda) w_2) = f_k(\lambda w_1 + (1-\lambda) w_2) \overset{f_k \text{ is convex}}{\leq} \lambda f_k(w_1) + (1-\lambda) \cdot f_k(w_2) = *$$

נשים לב שמתקיים $g(w) \geq f_k(w)$ $\forall w \in \mathbb{R}^n$. לכן $* \leq \lambda \cdot g(w_1) + (1-\lambda) \cdot g(w_2)$. כנדרש. $\blacksquare$

(c) Let $\ell_{log} : \mathbb{R} \to \mathbb{R}$ be the log loss, defined by

$$\ell_{\log}(z) = \log_2\left(1 + e^{-z}\right)$$

Show that ℓ_{log} is convex, and conclude that the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(\mathbf{w}) = \ell_{log}(y\mathbf{w} \cdot \mathbf{x})$ is convex with respect to $\mathbf{w}$.

נראה שהנגזרת השנייה חיובית ולכן l_{log} קמורה:

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

$$\frac{\partial^2 l_{log}(z)}{\partial z} = \frac{e^{-z}(1+e^{-z})\ln(2) - (-\ln(2)\cdot e^{-z})\cdot(-e^{-z})}{\left[\ln(2)\cdot(1+e^{-z})\right]^2} =$$

$$= \frac{e^{-z}(1+e^{-z})\cdot\ln(2) - e^{-z}\cdot\ln(2)\cdot e^{-z}}{\left[\ln(2)\cdot(1+e^{-z})\right]^2} = \frac{\ln(2)\cdot e^{-z}\cdot\left[1+e^{-z}-e^{-z}\right]}{\left[\ln(2)\cdot(1+e^{-z})\right]^2}$$

$$= \frac{e^{-z}}{\ln(2)\cdot(1+e^{-z})^2} > 0 \quad \text{for all } z$$

כעת נסיק ש $f(w)=l_{log}(yw\cdot x)$ קמורה ביחס ל w:

בסעיף א' הראנו שהרכבת פונ' קמורה על פונ' אפינית היא פונ' קמורה.
מכיוון שמכפלה פנימית היא פונ' אפינית ומכיוון ש l_{log} היא פונ' קמורה
ניתן להסיק שאכן $f(w)$ קמורה ביחס ל w.

4. **(15 points) Gradient Descent on Smooth Functions.** We say that a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

In words, β-smoothness of a function f means that at every point $\mathbf{x}$, f is upper bounded by a qaudratic function which coincides with f at $\mathbf{x}$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth and non-negative function (i.e., $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$). Consider the (non-stochastic) gradient descent algorithm applied on f with constant step size $\eta > 0$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$$

Assume that gradient descent is initialized at some point $\mathbf{x}_0$. Show that if $\eta < \frac{2}{\beta}$ then

$$\lim_{t\to\infty} \|\nabla f(\mathbf{x}_t)\| = 0$$

(Hint: Use the smoothness definition with points $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$ to show that $\sum_{t=0}^{\infty} \|\nabla f(\mathbf{x}_t)\|^2 < \infty$ and recall that for a sequence $a_n \geq 0$, $\sum_{n=1}^{\infty} a_n < \infty$ implies $\lim_{n\to\infty} a_n = 0$. Note that f is not assumed to be convex!)

יהי $\eta<\frac{2}{\beta}$ וגם f פונקציה β-חלקה, אזי מתקיים עבור $x,y\in\mathbb{R}^n$:

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|x-y\|^2$$

נסמן $x=x_t$, $y=x_{t+1}$ ונקבל:

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

כאמור, צעד ה GD הינו:

$$\eta\nabla f(x_t) = x_t - x_{t+1} \iff x_{t+1} - x_t = -\eta\nabla f(x_t) \iff x_{t+1} = x_t - \eta\nabla f(x_t)$$

נציב בצד ימין של אי השוויון ונקבל:

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

$$\iff$$

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

$$\iff$$

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

כעת מכיוון ש $\eta<\frac{2}{\beta}$ מתקיים:

$$\eta < \frac{2}{\beta} \iff \eta\frac{\beta}{2} < 0 \implies 1-\eta\frac{\beta}{2} > 0$$

לכן אם נחלק את שני האגפים ב $1-\frac{\beta}{2}\eta$ האי שוויון יישמר:

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

נסמן $C=\eta-\eta^2\frac{\beta}{2}$ ונקבל:

$$\|\nabla f(x_t)\|^2 \leq \frac{1}{C}\cdot\left(f(x_t)-f(x_{t+1})\right)$$

אם נסכום את שני האגפים על אותם ערכי t הא"ש יישמר, כלומר:

$$\sum_{t=0}^{T} \|\nabla f(x_t)\|^2 \leq \frac{1}{c} \cdot \sum_{t=0}^{T} f(x_t) - f(x_{t+1})$$

נבחין כי RHS מכיל טור טלסקופי ולכן:

$$\sum_{t=0}^{T} \|\nabla f(x_t)\|^2 \leq \frac{1}{c} \cdot \left(f(x_0) - f(x_{T+1})\right) \underset{\uparrow}{\leq} \frac{1}{c} \cdot f(x_0) < \infty$$

$$\forall x \in \mathbb{R}^n: f(x) \geq 0$$

הנ"ל נכון לכל $T \in \mathbb{N}$ ובפרט כאשר $T \to \infty$ נקבל: $\sum_{t=0}^{\infty} \|\nabla f(x_t)\|^2 < \infty$

מכיוון ש $\forall t \in \mathbb{N}: \|\nabla f(x_t)\|^2 \geq 0$ וגם הטור מתכנס נקבל שהאיבר הכללי שואף ל0:

$\lim_{t \to \infty} \|\nabla f(x_t)\|^2 = 0$ ומאריתמטיקה של גבולות כמובן שגם מתקיים: $\lim_{t \to \infty} \|\nabla f(x_t)\| \to \infty$ ■

3. **(20 points) GD with projection.** In the context of convex optimization, sometimes we would like to limit our solution to a convex set $\mathcal{K} \subseteq \mathbb{R}^d$; that is,

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{s.t.} \quad \mathbf{x} \in \mathcal{K}$$

for a convex function f and a convex set $\mathcal{K}$. In this scenario, each step in the gradient descent algorithm might result in a point outside $\mathcal{K}$. Therefore, we add an additional projection step. The projection operator finds the closest point in the set, i.e.:

$$\Pi_{\mathcal{K}}(\mathbf{y}) := \arg\min_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x} - \mathbf{y}\|_2$$

A modified iteration in the *gradient descent with projection* therefore consists of:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1})$$

(a) Let $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} = \Pi_{\mathcal{K}}(\mathbf{y})$. Prove that for any $\mathbf{z} \in \mathcal{K}$, we have $\|\mathbf{y} - \mathbf{z}\|_2 \geq \|\mathbf{x} - \mathbf{z}\|_2$. (**Guidance:** use the projection definition and the fact that for any $\lambda \in (0,1)$, $(1-\lambda)\mathbf{x} + \lambda\mathbf{z} \in \mathcal{K}$ to show that $\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{z} \rangle \leq \lambda \|\mathbf{z} - \mathbf{x}\|_2^2$ for any $\lambda \in (0,1)$. Conclude that $\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{z} \rangle \leq 0$. Use that to show the claim in the question.)

יהיו $y \in \mathbb{R}^d$, $z \in K$, $\lambda \in (0,1)$, $x = \Pi_K(y)$ כאשר K קבוצה קמורה.

מהגדרתם, $x, z \in K$ ולכן: $(1-\lambda)x + \lambda z \in K$

נסמן $k = (1-\lambda)x + \lambda z$.

מהגדרת x אנו יודעים ש x הוא הנקודה הכי קרובה ב K ל y ולכן:

$$\|x - y\| \leq \|k - y\| \iff \|x - y\|^2 \leq \|k - y\|^2$$

נפתח את הביטוי בצד ימין של אי השוויון:

$$\|k - y\|^2 = \|(1-\lambda)x + \lambda z - y\|^2 = \|x - \lambda x + \lambda z - y\|^2 = \|x - \lambda(x - z) - y\|^2$$

$$= \|x - y - \lambda(x - z)\|^2 = \langle x - y - \lambda(x - z), x - y - \lambda(x - z) \rangle = *$$

כעת אם נסמן $a=x-y$, $b=\lambda(x-z)$ נקבל:

$$* = \langle a-b, a-b\rangle = \langle a,a\rangle - 2\langle a,b\rangle + \langle b,b\rangle = \|a\|^2 - 2\langle a,b\rangle + \|b\|^2$$

נציב בחזרה את הביטויים המקוריים:

$$* = \|x-y\|^2 - 2\langle x-y, \lambda(x-z)\rangle + \|\lambda(x-z)\|^2$$

במכפלה הפנימית סקלר בגורם הימני במכפלה צריך לצאת עם הצמדה אבל מכיוון ש $\lambda\in\mathbb{R}$ להצמדה אין משמעות:

$$-2\langle x-y, \lambda(x-z)\rangle = -2\cdot\bar{\lambda}\langle x-y, x-z\rangle = -2\lambda\langle x-y, x-z\rangle$$

נקבל שצד ימין שווה:

$$* = \|x-y\|^2 - 2\lambda\langle x-y, x-z\rangle + \lambda^2\|x-z\|^2$$

לכן סה"כ קיבלנו:

$$\|x-y\|^2 \le \|x-y\|^2 - 2\lambda\langle x-y, x-z\rangle + \lambda^2\|x-z\|^2$$

$$\iff$$

$$0 \le -2\lambda\langle x-y, x-z\rangle + \lambda^2\|x-z\|^2$$

$$\iff$$

$$\langle x-y, x-z\rangle \le \frac{\lambda}{2}\cdot\|x-z\|^2$$

הוי שיוויון שקיבלנו צריך להתקיים גם עבור $z=x$ שכן $z,x\in K$ וגם עבור $\lambda\to 0^+$ לכן מתקיים:

$$\langle x-y, x-z\rangle \le 0 \quad (*)$$

כעת נראה ש:

$$\|y-z\| \ge \|x-z\| \iff \|y-z\|^2 \ge \|x-z\|^2$$

$$\|y-z\|^2 = \|\underbrace{y-x}_{a} + \underbrace{x-z}_{b}\|^2 = \langle a+b, a+b\rangle = \|a\|^2 + 2\langle a,b\rangle + \|b\|^2$$

$$= \|y-x\|^2 + 2\langle y-x, x-z\rangle + \|x-z\|^2 = \|y-x\|^2 - 2\langle x-y, x-z\rangle + \|x-z\|^2$$

כאשר בשיוויון האחרון הכפלנו ב-1- את הגורם השמאלי במכפלה הפנימית וכפל בסקלר של הגורם השמאלי במ"פ הוא לינארי:

נשים לב שכדי להראות את הדרוש מספיק להראות $\|y-z\|^2 - \|x-z\|^2 \ge 0$

ולכן:

$$\|y-z\|^2 - \|x-z\|^2 = \underbrace{\|y-x\|^2}_{\ge 0} \underbrace{- 2\underbrace{\langle x-y, x-z\rangle}_{\le 0 \text{ לפי } (*)}}_{\ge 0} + \|x-z\|^2 - \|x-z\|^2 \ge 0$$

■

(b) Prove that the convergence theorem for GD (as stated in lecture #5) still holds.

נסמן $\forall t: x_t := w_t$, $x^* := w^* \in \arg\min_w f(w)$

We now turn to prove convergence to global minimum for GD applied to a convex and differentiable function. Let $f : \mathbb{R}^d \to \mathbb{R}$ be such a function, with $w^* \in \arg\min_w f(w)$. Assume that $\|w^*\|_2 \le B$ and that $\|\nabla f(w)\|_2 \le G$ for every $w \in \mathbb{R}^d$. We will show that $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w_t$, the average point over all iterations of GD, converges to the optimal value of $f(\cdot)$.

Let $\epsilon > 0$. *For* $w_1 = w' = 0$, $\eta_t \equiv \frac{\epsilon}{G^2}$ *and* $T = \frac{B^2G^2}{\epsilon^2}$

נראה שאכן מתקיים: $f(\bar{w}) - f(w^*) \le \varepsilon$

jensen inequality

$$f(\bar{w}) = f\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) \le \frac{1}{T}\cdot\sum_{t=1}^{T} f(w_t) \iff f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*)$$

$$= \frac{1}{T}\sum_{t=1}^{T}\left(f(w_t) - f(w^*)\right)$$

מהגדרת קמירות מסדר ראשון נקבל:

$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(w_t), w_t - w^*\rangle$$

כעת אם נסמן $v := \nabla f(w_t)$, $u := w_t - w^*$ נקבל:

$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T}\langle u, v\rangle$$

נבחן את הביטוי: $\|u - \eta v\|^2$

$$\|u - \eta v\|^2 = \langle u - \eta v, u - \eta v\rangle = \|u\|^2 - 2\eta\langle u, v\rangle + \eta^2\|v\|^2$$

נבטא את $\langle u,v\rangle$ כתלות בבעיה המקורית:

$$\langle u, v\rangle = \frac{\|u\|^2 + \eta^2\|v\|^2 - \|u - \eta v\|^2}{2\eta} =$$

$$= \frac{\|w_t - w^*\|^2 + \eta^2\|\nabla f(w_t)\|^2 - \|\overbrace{w_t - w^* - \eta\nabla f(w_t)}^{y_{t+1} - w^*}\|^2}{2\eta}$$

נחזור לחישוב הראשון:

$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T}\langle u, v\rangle = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\|w_t - w^*\|^2 - \|y_{t+1} - w^*\|^2}{2\eta} + \frac{\eta}{2}\|\nabla f(w_t)\|^2\right)$$

מהסעיף הקודם אנו יודעים ש: $\|y_{t+1} - x^*\|^2 \ge \|x_{t+1} - x^*\|^2$ ולכן ניתן להחליף את $\|y_{t+1} - x^*\|^2$ ב $\|x_{t+1} - x^*\|^2$ (מחסרים משהו אז לא גדל קטן ולכן החסם מתקיים).

ב $\|x_{t+1}-x^*\|^2$ (מחסרים משהו אי שלילי יותר קטן ולכן הסכום מתקיים).

נקבל:
$$f(\bar{w})-f(w^*) \le \frac{1}{T}\cdot\frac{1}{2\eta}\sum_{t=1}^{T}\|w_t-w^*\|^2-\|x_{t+1}-w^*\|^2+\frac{\eta}{2}\|\nabla f(w_t)\|^2$$

טור טלסקופי.
$$\overset{\downarrow}{=} \frac{1}{2\eta T}\cdot\left(\|w_1-w^*\|^2-\|w_{T+1}-w^*\|^2\right)+\frac{\eta}{2}\cdot\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2$$

חיסור של אי שלילי מגדיל
$$\overset{\downarrow}{\le} \frac{\|w_1-w^*\|^2}{2\eta T}+\frac{\eta}{2}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2 \overset{w_1=0}{=} \frac{\|w^*\|^2}{2\eta T}+\frac{\eta}{2T}\sum_{t=1}^{T}\|\nabla f(w_t)\|^2$$

$$\le \frac{B^2}{2\eta T}+\frac{\eta}{2}\sigma^2 = \frac{B^2}{2\cdot\frac{\varepsilon}{\sigma^2}\cdot\frac{B^2\sigma^2}{\varepsilon^2}}+\frac{\varepsilon}{2\sigma^2}\sigma^2=\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon \quad \blacksquare$$

2. **(15 points) Hinge loss with linearly separable data.** Consider a setup of learning linear classifiers over a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1,+1\}$ for all i. Assume the data is linearly separable, i.e., given a training set there exists $\mathbf{w}^* \in \mathbb{R}^d$ such that $y_i\mathbf{w}^*\cdot\mathbf{x}_i > 0$ for all i (note the strict inequality; assume no bias for simplicity). Recall the definition of the *hinge loss* given in lecture #5:

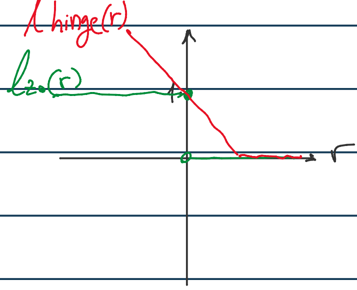
$$\ell_{hinge}(r) = \max\{0, 1-r\}.$$

We would like to show that in the linearly separable case, minimizing the hinge loss over the training data will yield a classifier with optimal zero-one loss. Formally, let

$$\mathbf{w}^*_{hinge} = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left\{\sum_{i=1}^n \ell_{hinge}(y_i\mathbf{w}\cdot\mathbf{x}_i)\right\}.$$

Show that $sign(\mathbf{w}^*_{hinge}\cdot\mathbf{x}_i) = y_i$ for all i, meaning $\mathbf{w}^*_{hinge}$ achieves optimal zero-one loss.
(**Hint:** First show that the hinge loss upper bounds the zero-one loss. Then consider the hinge loss of $c\mathbf{w}^*$ for some constant $c > 0$. What happens to the hinge loss as $c \to \infty$?)

בראשית נראה ש hinge-loss חוסם מלמעלה את zo-loss,
ואכן קל לראות על פי הגרפים של שתי הפונקציות שלכל r: $\ell_{hinge}(r) \ge \ell_{zo}(r)$



כעת מכיוון שהמדגם הוא linearly separable קיים $w^* \in \mathbb{R}^d$ כך ש $0 < y_i\langle w^*, x_i\rangle$ $\forall i$.

נראה שעבור אותו w^* ℓ_{hinge} משיג לו אותו ערך אופטימלי שמשיג ℓ_{0-1}.

עבור קבוע $c>0$ מתקיים ש ℓ_{hinge} על cw^* הוא:

$$\ell_{hinge}(y_i\langle cw^*, x\rangle) = \max\{0, 1 - y_i\langle cw^*, x_i\rangle\} = \max\{0, 1 - c\cdot y_i\langle w^*, x_i\rangle\}$$

כעת, מכיוון שעבור אותו w^* מתקיים ש $0 < y_i\langle w^*, x_i\rangle$ לכל i,

אם נשאיף את c לאינסוף נקבל:

$$\lim_{c\to\infty} \ell_{hinge}(y_i\langle cw^*, x_i\rangle) = \lim_{c\to\infty} \max\{0, 1 - c\cdot y_i\langle w^*, x_i\rangle\} = \max\{0, -\infty\} = 0$$

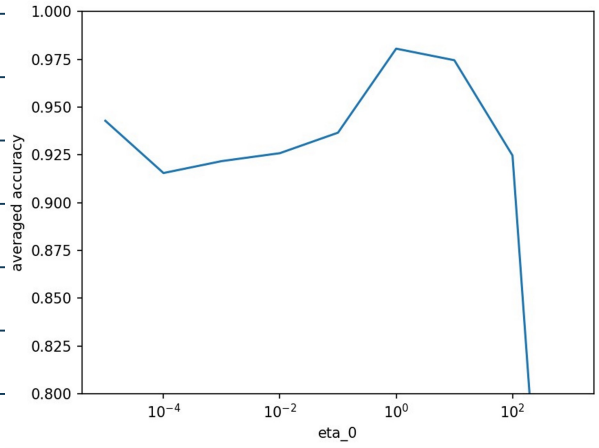
כלומר קיים מסווג w^* שייתן הפסד 0 עבור hinge-loss ולכן אימון על פי ההפסד הזה ימצא את w^*.

סה"כ הראנו ש ℓ_{hinge} חוסם מלמעלה את ℓ_{0-1} וגם עבור w^* ההפסד לפי hinge יהיה 0 כלומר ש w^*_{hinge} אכן משיג ℓ_{0-1} אופטימלי. ומכאן נקבל מתקיים לכל i: $sign(\langle w^*_{hinge}, x_i\rangle) = y_i$ כפי שרצינו להראות. ∎

מטלה תכנות

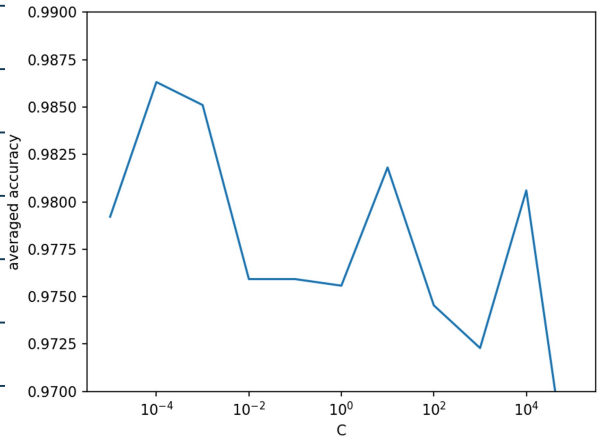
שאלה 1 - SGD for hinge loss

(a)



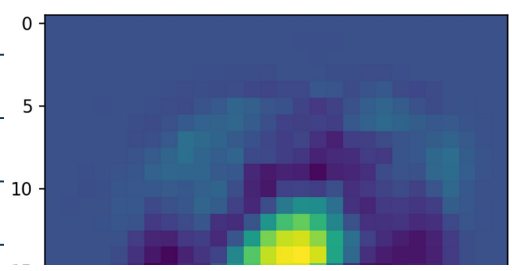
קיבלנו ש η_0 הטוב ביותר הוא 1.

(b)

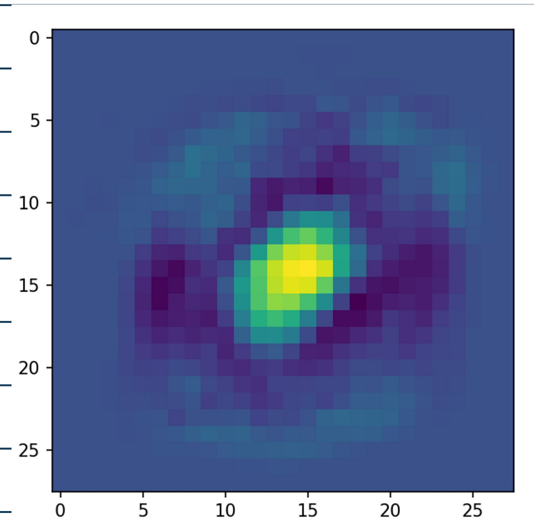


קיבלנו שה C הטוב ביותר הוא 10^{-4}.

(c)



(c)



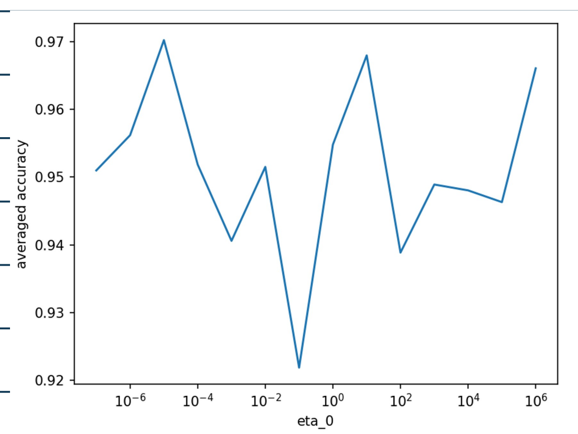
קיבלנו ש שמרכז׳ במובהק התמונה, שהו ההבדל העיקרי בין התמונה 8 לתמונה 0 .

(d)

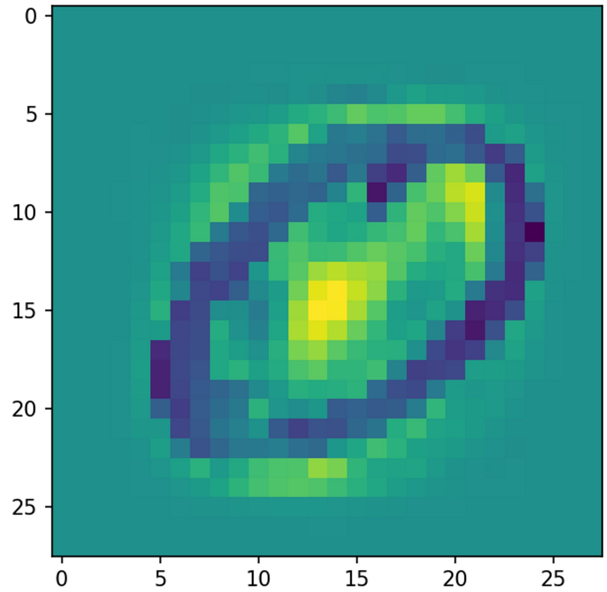
```
the accuracy of the best classifier on the test set is:  0.9928352098259979
```

שאלה 2 – SGD for log-loss

(a)

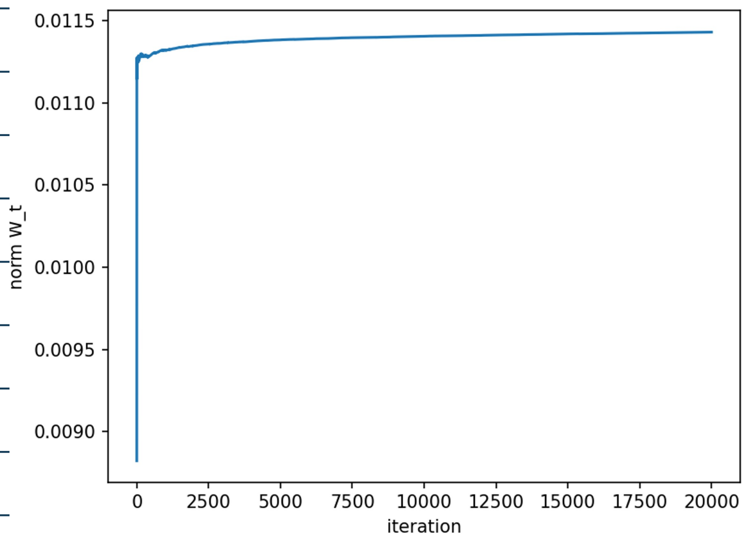


(ב)



```
the accuracy of the best classifier on the test set is:  0.9646878198567042
```

(ג)



ניתן לראות שבהתחלה הנורמה היא 0 וזה הגיוני שכן $w_1 = 0$.
לאחר כ 1000 איטרציות יש התייצבות של הנורמה וזה הגיוני כי למעשה האלגוריתם מתכנס ל w_t האופטימלי. בכל איטרציה עושים צעד קטן יותר ויותר וככה w_t לא משתנה בהרבה והנורמה שלו גם כן.