תרגיל בית 3 – מבוא ללמידה חישובית

מגיש: נתן בלוך, 316130707

Theory Questions.

שאלה 1. Max of Convex Functions.

נתונות m פונקציות קמורות, $f_1(x),\dots,f_m(x)$, כאשר $f_i\colon \mathbb{R}^d \to \mathbb{R}$. נגדיר פונק' חדשה – $g(x) = max_i\ f_i(x)$.

סעיף א. נדרש להוכיח כי $g(x)$ קמורה. g מוגדרת מעל $\mathbb{R}^d$ שהינה קמורה. נדרש להוכיח כי לכל $w_1, w_2 \in \mathbb{R}^d$ ולכל $\lambda \in [0,1]$, מתקיים–

$$g(\lambda w_1 + (1-\lambda)w_2) \le \lambda g(w_1) + (1-\lambda)g(w_2)$$

ראשית, נתון כי $f_1(x),\dots,f_m(x)$ הינן קמורות, ולכן לכל $i \in \{1,\dots m\}$ מתקיים אי-השיוויון לכל $w_1, w_2 \in \mathbb{R}^d$ ולכל $\lambda \in [0,1]$ –

$$f_i(\lambda w_1 + (1-\lambda)w_2) \le \lambda f_i(w_1) + (1-\lambda)f_i(w_2)$$

ולפיכך ניתן לקבל את אי-השיוויון הבא –

$$(\mathbf{1})\quad max_i\ f_i(\lambda w_1 + (1-\lambda)w_2) \le max_i\ \lambda f_i(w_1) + (1-\lambda)f_i(w_2)$$

נוכיח כעת כי $g(x)$ קמורה. יהי $w_1, w_2 \in \mathbb{R}^d$ ויהי $\lambda \in [0,1]$. מתקיים –

$$g(\lambda w_1 + (1-\lambda)w_2) \underbrace{=}_{\substack{def\\ of\ g}} max_i\ f_i(\lambda w_1 + (1-\lambda)w_2) \underbrace{\le}_{\substack{by\\ (\mathbf{1})}} max_i\ \lambda f_i(w_1) + (1-\lambda)f_i(w_2) \le$$

$$\underbrace{\le}_{(2)} max_i\ \lambda f_i(w_1) + max_i\ (1-\lambda)f_i(w_2) = \lambda \cdot \underbrace{max_i\ f_i(w_1)}_{=g(w_1)} + (1-\lambda)\cdot \underbrace{max_i\ f_i(w_2)}_{=g(w_2)} = \lambda g(w_1) + (1-\lambda)g(w_2)$$

נכונות מעבר (2) נובע מתכונות של פונק' המקסימום(max). הראנו בעצם **שלכל** $w_1, w_2 \in \mathbb{R}^d$ **ולכל** $\lambda \in [0,1]$ מתקיים –

$$g(\lambda w_1 + (1-\lambda)w_2) \le \lambda g(w_1) + (1-\lambda)g(w_2)$$

כנדרש, על מנת להוכיח קמירות של $g(x)$. ■

סעיף ב. נדרש להוכיח כי $\nabla f_i(x)$, עבור $f_i(x) = max\{f_1(x),\dots,f_m(x)\}$, הינו סאב-גרדיינט($sub\ gradient$) של g בנקודה x.
נניח כי f_i דיפרנציאבילית לכל $i \in \{1,\dots m\}$. ראינו בהרצאה את טענה 5.3.3 לפיה, עבור פונק' דיפרנציאבלית וקמורה f בנקודה כלשהי w, מתקיים $\partial f(w) = \{\nabla f(w)\}$, ובמילים אחרות, ה-$sub\ gradient$ היחיד שלה בנקודה w הינו הגרדיאנט בנקודה.
יהי $x \in \mathbb{R}^d$. נסמן $f_i(x) = max\{f_1(x),\dots,f_m(x)\}$. נדרש להוכיח כי $z = \nabla f_i(x)$ הינו $sub\ gradient$ של g בנקודה x,
ולכן נרצה להוכיח שלכל $u \in \mathbb{R}^d$ מתקיים –

$$g(u) \ge g(x) + \langle u - x, \nabla f_i(x) \rangle$$

ובאופן שקול, **נרצה להוכיח** שלכל $u \in \mathbb{R}^d$ מתקיים –

$$g(u) - g(x) \ \ge\ \langle u - x, \nabla f_i(x) \rangle$$

ואכן, כיוון ש-$f_i(x)$ קמורה ודיפרנציאבילית בכל נקודה(ובפרט בנקודה $x \in \mathbb{R}^d$), נקבל לפי משפט 5.3.3 מההרצאה, שלכל $u \in \mathbb{R}^d$ מתקיים –

$$f_i(u) \ge f_i(x) + \langle u - x, \nabla f_i(x) \rangle$$

ובאופן שקול, לכל $u \in \mathbb{R}^d$ מתקיים –

$$(\mathbf{1})\quad f_i(u) - f_i(x) \ \ge\ \langle u - x, \nabla f_i(x) \rangle$$

כמו כן, נזכור כי – $f_i(x) = max_j\ f_j(x)$, וכן לכל $u \in \mathbb{R}^d$ מתקיים – $g(u) = max_j\ f_j(u) \ge f_i(u)$.
מכאן לכל $u \in \mathbb{R}^d$ מתקיים –

$$g(u) - g(x) = \underbrace{max_j\ f_j(u)}_{\ge f_i(u)} - \underbrace{max_j\ f_j(x)}_{= f_i(x)} \ge f_i(u) - f_i(x) \underbrace{\ge}_{\substack{by\\ (\mathbf{1})}} \langle u - x, \nabla f_i(x) \rangle$$

ולכן, לפי ההגדרה מתקיים ש- $z = \nabla f_i(x)$ הינו $sub\ gradient$ של g בנקודה x, כאשר $f_i(x) = max_j\ f_j(x)$, וזה כאמור לכל $x \in \mathbb{R}^d$.
כנדרש. ■

שאלה 2. ℓ^2 penalty.

נתונה הבעיה הבאה –

$$\min_{w,b,\xi}\ \frac{1}{2}w^Tw+\frac{C}{2}\sum_{i=1}^{m}{\xi_i}^2$$

$$s.t.\quad y_i(w^Tx_i+b)\ge 1-\xi_i\qquad \forall i\in\{1,\dots,m\}$$

סעיף א. נוכיח כי אילוצים מהצורה $\xi_i\ge 0$ לא ישנו את הבעיה. ובמילים אחרות, נוכיח כי כל פיתרון אופטימלי לבעיה בהכרח יקיים $\xi_i\ge 0$ לכל i.
פיתרון. יהי w^*,b^*,ξ^* פיתרון אופטימלי כלשהו לבעיה. כלומר מתקיים –

$$y_i((w^*)^Tx_i+b^*)\ge 1-{\xi^*}_i\qquad \forall i\in\{1,\dots,m\}$$

ובכן לכל w,b,ξ, מתקיים –

$$\frac{1}{2}(w^*)^T(w^*)+\frac{C}{2}\sum_{i=1}^{m}({\xi^*}_i)^2\le\frac{1}{2}w^Tw+\frac{C}{2}\sum_{i=1}^{m}{\xi_i}^2$$

שכן הנחנו כי w^*,b^*,ξ^* פיתרון אופטימלי. נוכיח כי מתקיים – ${\xi^*}_i\ge 0$ לכל i. נניח **<u>בשלילה</u>** כי קיים i עבורו ${\xi^*}_k<0$. עבור k זה מתקיים –

$$(\mathbf{1})\qquad y_k((w^*)^Tx_k+b^*)\ge 1-{\xi^*}_k\underbrace{\ge}_{-{\xi^*}_k>0}1$$

נגדיר את הפיתרון הבא – $w'=w^*,b'=b^*$ וכן נגדיר ξ' באופן הבא –

$${\xi'}_j=\begin{cases}{\xi^*}_j & if\ j\ne k\\ 0 & if\ j=k\end{cases}$$

נראה כי הפיתרון w',b',ξ' מקיים את האילוצים של הבעיה, ואכן לכל $j\ne k$, האילוץ מתקיים שכן ${\xi^*}_j={\xi'}_j$, $w'=w^*,b'=b^*$ ולכן –

$$y_j\big((w')^Tx_j+b'\big)=y_j\big((w^*)^Tx_j+b^*\big)\ge 1-{\xi^*}_j=1-{\xi'}_j$$

ועבור $j=k$, מתקיים –

$$y_k((w')^Tx_k+b')=y_k((w^*)^Tx_k+b^*)\underbrace{\ge}_{\substack{by\\(1)}}1=1-\underbrace{{\xi'}_k}_{=0}$$

ואכן מתקיימים כל האילוצים הנדרשים. כעת, מתקיים – ${\xi'}_k=0>{\xi^*}_k$ ולכן – $({\xi'}_k)^2=0<({\xi^*}_k)^2$. לכן נבחין מכך שההבדל היחיד בין הפתרונות הינו ${\xi^*}_k$, שמתקיים מכך ש – $w'=w^*,b'=b^*$ וכן –

$$\underbrace{\frac{1}{2}(w^*)^T(w^*)+\frac{C}{2}\sum_{i=1}^{m}({\xi^*}_i)^2}_{\substack{result\ for\\ w^*,b^*,\xi^*}}=\frac{1}{2}(w^*)^T(w^*)+\frac{C}{2}\sum_{\substack{i=1\\i\ne k}}^{m}({\xi^*}_i)^2+\frac{C}{2}({\xi^*}_k)^2>\frac{1}{2}(w')^T(w')+\frac{C}{2}\sum_{\substack{i=1\\i\ne k}}^{m}({\xi'}_i)^2+\frac{C}{2}({\xi'}_k)^2$$

$$=\underbrace{\frac{1}{2}(w')^T(w')+\frac{C}{2}\sum_{i=1}^{m}({\xi'}_i)^2}_{\substack{result\ for\\ w',b',\xi'}}$$

וזו <u>סתירה</u> באמור, שכן קיבלנו כי הפיתרון עבור w^*,b^*,ξ^* אינו פיתרון אופטימלי. על כן, נסיק את הנדרש – מתקיים – ${\xi^*}_i\ge 0$ לכל i, ולכן אילוצים מהצורה $\xi_i\ge 0$ לא ישנו את הבעיה.

סעיף ב. הלגרנג'יאן של הבעיה.
פיתרון. נבחין כי ניתן לכתוב את האילוצים באופן הבא –

$$1-\xi_i-y_i(w^Tx_i+b)\le 0\qquad \forall i\in\{1,\dots,m\}$$

ולכן הלגרנג'יאן במקרה זה הינו –

$$\boxed{\mathcal{L}(w,b,\xi,\alpha)=\frac{1}{2}w^Tw+\frac{C}{2}\sum_{i=1}^{m}{\xi_i}^2+\sum_{i=1}^{m}\alpha_i\big(1-\xi_i-y_i(w^Tx_i+b)\big)}$$

סעיף ג. נביא את הלגרנג'יאן למינימום ביחס ל-w, b, ξ , ע"י גזירה ביחס למשתנים הללו והשוואת הנגזרות החלקיות ל-אפס.

פיתרון. מתקיים –

$$\frac{\partial \mathcal{L}(w,b,\xi,\alpha)}{\partial w_k} = w_k - \sum_{i=1}^{m} \alpha_i y_i x_i^{(k)}$$

כאשר $x_j^{(k)}$ מייצג את האינדקס ה-k של הוקטור x_i. ע"י השוויון $\frac{\partial \mathcal{L}(w,b,\xi,\alpha)}{\partial w_k} = 0$, נקבל – $\boxed{w_k = \sum_{i=1}^{m} \alpha_i y_i x_i^{(k)}}$, ונסמן שוויון זה ב-(*).

כמו כן, מתקיים –

$$\frac{\partial \mathcal{L}(w,b,\xi,\alpha)}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i$$

ע"י השוויון $\frac{\partial \mathcal{L}(w,b,\xi,\alpha)}{\partial b} = 0$, נקבל – $\sum_{i=1}^{m} \alpha_i y_i = 0$, ונסמן שוויון זה ב-(**).

כמו כן, מתקיים –

$$\frac{\partial \mathcal{L}(w,b,\xi,\alpha)}{\partial \xi_k} = C\xi_k - \alpha_k$$

ע"י השוויון $\frac{\partial \mathcal{L}(w,b,\xi,\alpha)}{\partial \xi_k} = 0$, נקבל – $C\xi_k = \alpha_k$, לכל k, ונסמן שוויונות אלה ב-(***).

נבחין כי מתקיים –

$$\sum_{i=1}^{m} \alpha_i y_i x_i = \alpha_1 y_1 x_1 + \cdots + \alpha_m y_m x_m = \begin{pmatrix} \alpha_1 y_1 x_1^{(1)} \\ \vdots \\ \alpha_1 y_1 x_1^{(n)} \end{pmatrix} + \ldots + \begin{pmatrix} \alpha_m y_m x_m^{(1)} \\ \vdots \\ \alpha_m y_m x_n^{(n)} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} \alpha_i y_i x_i^{(1)} \\ \vdots \\ \sum_{i=1}^{m} \alpha_i y_i x_i^{(n)} \end{pmatrix} \underbrace{=}_{(*)} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = w$$

נסתכל כעת על הביטוי $\sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i(w^T x_i + b)\right)$ המופיע כגורם בלגרנג'יאן ונציב בו את האילוצים שקיבלנו כך ש – $\nabla \partial \mathcal{L}(w,b,\xi,\alpha) = 0$, ביחס למשתנים w, b, ξ. נקבל –

$$\sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i(w^T x_i + b)\right) = \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \alpha_i y_i (w^T x_i + b) = \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \alpha_i y_i w^T x_i - \sum_{i=1}^{m} \alpha_i y_i b =$$

$$= \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \xi_i - w^T \underbrace{\sum_{i=1}^{m} \alpha_i y_i x_i}_{=w} - b \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{\substack{=0 \\ by (**)}} \underbrace{=}_{\substack{by\ (***) \\ and (**)}} \sum_{i=1}^{m} \alpha_i - \frac{1}{C} \sum_{i=1}^{m} \alpha_i \alpha_i - \left(\underbrace{\sum_{i=1}^{m} \alpha_i y_i x_i}_{=w} \right)^T \left(\underbrace{\sum_{i=1}^{m} \alpha_i y_i x_i}_{=w} \right)$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{C} \sum_{i=1}^{m} \alpha_i \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i y_i x_i \alpha_j y_j x_j$$

נציב בלגרנגי'יאן את האילוצים –

$$\mathcal{L}(w,b,\xi,\alpha) = \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{m} {\xi_i}^2 + \sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i(w^T x_i + b)\right) =$$

$$= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i y_i x_i \alpha_j y_j x_j + \frac{C}{2} \sum_{i=1}^{m} \frac{{\alpha_i}^2}{C^2} + \sum_{i=1}^{m} \alpha_i - \frac{1}{C} \sum_{i=1}^{m} \alpha_i \alpha_i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i y_i x_i \alpha_j y_j x_j =$$

$$= \boxed{-\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i y_i x_i \alpha_j y_j x_j + \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \cdot \frac{1}{C} \sum_{i=1}^{m} {\alpha_i}^2}$$

סעיף ד. התוכנית הדואלית.

ראינו בהרצאה כי התוכנית הפרימאלית, שקולה לבעיית ה-min-max הבאה –

$$\underset{w,b,\xi}{min}\max_{\alpha}\mathcal{L}(w,b,\xi,\alpha)$$

ובן ראינו כי מתקיים האי-שיוויון –

$$\underset{w,b,\xi}{min}\max_{\alpha}\mathcal{L}(w,b,\xi,\alpha) \;\geq \max_{\alpha}\;\underset{w,b,\xi}{min}\,\mathcal{L}(w,b,\xi,\alpha)$$

ואם נגדיר – $g(\alpha) = \underset{w,b,\xi}{min}\,\mathcal{L}(w,b,\xi,\alpha)$, ופונק' המטרה של התוכנית הדואלית תהיה מקסימיזציה של החסם התחתון לפונק' המטרה המקורית, ולכן –

$$\max_{\alpha} g(\alpha) \;= \max_{\alpha}\;\underset{w,b,\xi}{min}\,\mathcal{L}(w,b,\xi,\alpha)$$

וכיוון שהפונק' $\mathcal{L}(w,b,\xi,\alpha)$ הינה קמורה(כסכום של פונק' קמורות כל אחת), נקבל כי המינימום שלה מתקבל בנקודה בה הגרדיאנט שלה לפי w,b,ξ מקיים $\nabla\mathcal{L}(w,b,\xi,\alpha) = 0$. ראינו כי במקרה זה, המינימום מקיים –

$$\mathcal{L}(w,b,\xi,\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i x_i \alpha_j y_j x_j \;+\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\cdot\frac{1}{C}\sum_{i=1}^{m}{\alpha_i}^2$$

ולכן פונק' המטרה של התוכנית הדואלית הינה –

$$\max_{\alpha} g(\alpha) \;=\; \max_{\alpha}\;\; -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i x_i \alpha_j y_j x_j \;+\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\cdot\frac{1}{C}\sum_{i=1}^{m}{\alpha_i}^2$$

נבחין כי אספנו בדרך אילוץ נוסף לפיו –

$$\sum_{i=1}^{m}\alpha_i y_i = \;0$$

אילוץ זה נובע מכך שבלגרנג'יאן קיבלנו –

$$\mathcal{L}(w,b,\xi,\alpha) = \frac{1}{2}w^T w + \frac{C}{2}\sum_{i=1}^{m}{\xi_i}^2 \;+\; \sum_{i=1}^{m}\alpha_i\big(1-\xi_i - y_i(w^T x_i + b)\big) =$$

$$= \frac{1}{2}w^T w + \frac{C}{2}\sum_{i=1}^{m}{\xi_i}^2 + \sum_{i=1}^{m}\alpha_i \;- \sum_{i=1}^{m}\alpha_i\xi_i \;- \sum_{i=1}^{m}\alpha_i y_i w^T x_i - \boldsymbol{b}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}}\boldsymbol{\alpha_i y_i}$$

ונבחין כי אם $\sum_{i=1}^{m}\alpha_i y_i \;\neq\; 0$, אזי נקבל תמיד כי – $\underset{w,b,\xi}{min}\,\mathcal{L}(w,b,\xi,\alpha) = -\infty$.

כמו כן, כיוון ש-α_i הינם כופלים של אילוצי אי-שיוויון, נקבל את אילוצי אי-שליליות לגביהם – $\alpha_i \geq 0$.

בסה"כ, התוכנית הדואלית הינה –

$$\boxed{\begin{array}{c} \max_{\alpha}\;\; -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i y_i x_i \alpha_j y_j x_j \;+\sum_{i=1}^{m}\alpha_i \;-\; \frac{1}{2}\cdot\frac{1}{C}\sum_{i=1}^{m}{\alpha_i}^2 \\ \boldsymbol{s.t.}\;\; \sum_{i=1}^{m}\alpha_i y_i = 0 \\ \alpha_i \geq 0 \qquad\qquad \forall i \in \{1,..,m\} \end{array}}$$

שאלה 3. The Multi-Class Hinge-Loss.

נגדיר את פונק' ההפסד($multi\ class\ hinge\ loss\ function$) באופן הבא –

$$\ell(w_1,\dots w_L,x,y)=\max_{\hat{y}\in[L]} w_{\hat{y}}\cdot x-w_y\cdot x+\Delta_{zo}(\hat{y},y)$$

סעיף א. נדרש להוכיח כי ℓ פונקצייה קמורה של המשתנים $w_1,\dots,w_L$. **ידוע כי פונק' המקסימום של פונק' קמורות כל אחת הינה גם כן פונק' קמורה**, ולכן מספיק להראות שלכל $\hat{y}\in[L]$ הפונק' $w_{\hat{y}}\cdot x-w_y\cdot x+\Delta_{zo}(\hat{y},y)$ הינה פונק' קמורה של המשתנים $w_1,\dots,w_L$.
כמו כן, **ידוע כי סכום של פונק' קמורות הינה פונק' קמורה** ולכן מספיק להראות כי כל גורם בפונק' $w_{\hat{y}}\cdot x-w_y\cdot x+\Delta_{zo}(\hat{y},y)$ הינו פונק' קמורה של המשתנים $w_1,\dots,w_L$. ואכן, יהיו $x,y,\hat{y}$ קבועים כלשהם ונוכיח קמירות. הגורם $\Delta_{zo}(\hat{y},y)$ הינה פונק' קבועה של המשתנים $w_1,\dots,w_L$ ולכן זו פונק' קמורה. הגורמים של $w_{\hat{y}}\cdot x$ ו- $-w_y\cdot x$ הינם פונק' לינאריות של המשתנים $w_1,\dots,w_L$, ולכן מדובר בפונק' קמורות גם כן.
ובסה"כ, כל הגורמים מהווים פונק' קמורות ביחס למשתנים $w_1,\dots,w_L$, ולכל לכל $\hat{y}\in[L]$ הפונק' $w_{\hat{y}}\cdot x-w_y\cdot x+\Delta_{zo}(\hat{y},y)$ הינה פונק' קמורה של המשתנים $w_1,\dots,w_L$. **מכאן, נסיק כי הפונק' $\ell(w_1,\dots w_L,x,y)$ הינה קמורה במקסימום של פונק' קמורות כל אחת, כנדרש.**

סעיף ב. נדרש להוכיח שמתקיים $\ell(w_1,\dots w_L,x,y)\geq\Delta_{zo}(f(x;w_1,\dots w_L),y)$ לכל w,x,y.
יהי w,x,y כלשהם. נסמן כך – $\boldsymbol{i}=:\operatorname{argmax}_{y\in[L]} w_y\cdot x=f(x;w_1,\dots w_L)$. מתקיים לכן – $i\in[L]$, וכן $w_i\cdot x\geq w_y\cdot x$ לכל $y\in[L]$.
כיוון שלפי ההגדרה $i\in[L]$, לפי הגדרת מקסימום מתקיים –

$$\max_{\hat{y}\in[L]} w_{\hat{y}}\cdot x-w_y\cdot x+\Delta_{zo}(\hat{y},y)\ \geq w_i\cdot x-w_y\cdot x+\Delta_{zo}(i,y)\qquad(\mathbf{1})$$

שבן הגורם $w_i\cdot x-w_y\cdot x+\Delta_{zo}(i,y)$ הינו גורם בפונק' ה-$\max_{\hat{y}\in[L]}$ הנ"ל. לכן,

$$\ell(w_1,\dots w_L,x,y)=\max_{\hat{y}\in[L]} w_{\hat{y}}\cdot x-w_y\cdot x+\Delta_{zo}(\hat{y},y)\underbrace{\geq}_{(1)}\underbrace{w_i\cdot x-w_y\cdot x}_{\substack{\geq 0\ as\\ \forall y\in[L].\\ w_i\cdot x\geq w_y\cdot x}}+\Delta_{zo}(i,y)\geq\Delta_{zo}(i,y)=\Delta_{zo}(f(x;w_1,\dots w_L),y)$$

כאשר השיוויון האחרון נובע מכך שסימנו – $i:=\operatorname{argmax}_{y\in[L]} w_y\cdot x=f(x;w_1,\dots w_L)$.
ובסה"כ הראנו בסעיף זה כי לכל w,x,y, מתקיים –

$$\ell(w_1,\dots w_L,x,y)\geq\Delta_{zo}(f(x;w_1,\dots w_L),y)$$

כנדרש.

סעיף ג. נתון כי קיימים $w_1^\star,\dots,w_L^\star$, שמשיגים טעות אמפירית(טעות על גבי סט האימון) של אפס. במילים אחרות, לכל $i\in[L]$, מתקיים –

$$\Delta_{zo}(f(x_i;w_1^\star,\dots,w_L^\star),y_i)=0$$

נוכיח כי גם בעבור $w_1^{opt},\dots,w_L^{opt}$, לכל $i\in[L]$, מתקיים – $\Delta_{zo}\left(f\left(x_i;w_1^{opt},\dots,w_L^{opt}\right),y_i\right)=0$.
פיתרון. נוכיח כי מתקיים שיוויון ע"י כך שנוכיח כי מתקיימים שני אי-שיוויונים – $0\leq\Delta_{zo}\left(f\left(x_i;w_1^{opt},\dots,w_L^{opt}\right),y_i\right)\leq 0$.

בכיוון הראשון, בו נוכיח כי $0\leq\Delta_{zo}\left(f\left(x_i;w_1^{opt},\dots,w_L^{opt}\right),y_i\right)$ הינו מיידי מהגדרת הפונק' $\Delta_{zo}(\cdot,\cdot)$, שכן לכל $y,\hat{y}\in[L]$, מתקיים – $\Delta_{zo}(y,\hat{y})\in\{0,1\}$, ובפרט אי-שלילי. ולכן $0\leq\Delta_{zo}\left(f\left(x_i;w_1^{opt},\dots,w_L^{opt}\right),y_i\right)$, כנדרש בכיוון זה.

בכיוון השני, נדרש להוכיח כי מתקיים $\Delta_{zo}\left(f\left(x_i;w_1^{opt},\dots,w_L^{opt}\right),y_i\right)\leq 0$. לפי ההנחה, לכל $i\in[L]$, מתקיים –

$$\Delta_{zo}(f(x_i;w_1^\star,\dots,w_L^\star),y_i)=0\quad\underbrace{\Longrightarrow}_{\substack{\boldsymbol{def\ of}\\ \Delta_{zo}}}\quad f(x_i;w_1^\star,\dots,w_L^\star)=\ y_i\quad\underbrace{\Longrightarrow}_{\substack{\boldsymbol{def\ of}\\ f}}\quad\operatorname{argmax}_{y\in[L]} w_y\cdot x=\ y_i$$

$$\Longrightarrow\ \max_{y\in[L]} w_y\cdot x\ \leq w_{y_i}\cdot x\quad\Longrightarrow\ \max_{\substack{y\in[L]\\ y\neq y_i}} w_y\cdot x\ <w_{y_i}\cdot x$$

כאשר המעבר האחרון נובע מכך ש- $f(x_i;w_1^\star,\dots,w_L^\star)=\max_{y\in[L]} w_y\cdot x$, ו- f פונק' ולכן חד-ערכית, ולכן לא ייתכנו שני ארגומנטים שהם $maximizers$.
לכל $i\in[L]$, נגדיר –

$$\varepsilon_i=\max_{\substack{y\in[L]\\ y\neq y_i}}\ w_y\cdot x\ -w_{y_i}\cdot x$$

ולפי מה שהוכחנו ב-$(*)$, מתקיים לכל $i\in[L]$, $\varepsilon_i<0$. נגדיר גם כן – $\boxed{\boldsymbol{\varepsilon}\ :=\min_i\ |\varepsilon_i|>0}$. כעת, נגדיר את הוקטורים $w_1',\dots,w_L'$, באופן הבא –

$$w_i'=\frac{1}{\varepsilon}\cdot w_i^\star$$

כמו כן, נסתכל על הביטוי – $\max_{y\in[L]} w'_y \cdot x_i - w'_{y_i}\cdot x_i + \Delta_{zo}(y,y_i)$, ונבחין שעבור $y=y_i$, נקבל – $\underbrace{w'_{y_i}\cdot x_i - w'_{y_i}\cdot x_i}_{=0} + \underbrace{\Delta_{zo}(y_i,y_i)}_{=0} = 0$, ולכן ניתן להציג זאת באופן הבא **השקול** –

$$\max_{y\in[L]} w'_y \cdot x_i - w'_{y_i}\cdot x_i + \Delta_{zo}(y,y_i) = max\{0, \max_{\substack{y\in[L]\\ y\neq y_i}} w'_y \cdot x_i - w'_{y_i}\cdot x_i + \Delta_{zo}(y,y_i)\} = max\{0, \max_{\substack{y\in[L]\\ y\neq y_i}} w'_y \cdot x_i - w'_{y_i}\cdot x_i + 1\} \quad (*)$$

ובנוסף, נבחין כי במקרה זה, המקסימום נלקח על גבי איברי הקבוצה $[L]$ **השונים** מ-y_i, ולכן מתקיים – $\Delta_{zo}(y,y_i)=1$.

מתקיים –

$$\sum_i \Delta_{zo}\left(f\left(x_i; w_1^{opt},\dots,w_L^{opt}\right), y_i\right) \underbrace{\le}_{(b)} \sum_i \ell\left(w_1^{opt},\dots,w_L^{opt}, x_i, y_i\right) \underbrace{\le}_{\substack{w^{opt}\\ optimizer}} \sum_i \ell\left(w'_1,\dots,w'_L, x_i, y_i\right) =$$

$$= \sum_i \max_{y\in[L]} w'_y \cdot x_i - w'_{y_i}\cdot x_i + \Delta_{zo}(y,y_i) \underbrace{=}_{(*)} \sum_i max\{0, \max_{\substack{y\in[L]\\ y\neq y_i}} w'_y \cdot x_i - w'_{y_i}\cdot x_i + 1\} =$$

$$= \sum_i max\{0, 1 + \max_{\substack{y\in[L]\\ y\neq y_i}} \frac{1}{\varepsilon}\cdot w^\star_y \cdot x_i - \frac{1}{\varepsilon}\cdot w^\star_{y_i}\cdot x_i\} = \sum_i max\{0, 1 + \frac{1}{\varepsilon}\cdot \underbrace{\max_{\substack{y\in[L]\\ y\neq y_i}} w^\star_y \cdot x_i - w^\star_{y_i}\cdot x_i}_{=\varepsilon_i}\} =$$

$$= \sum_i max\{0, 1 + \frac{1}{\varepsilon}\cdot \varepsilon_i\} \underbrace{=}_{(**)} \sum_i max\{0, 1 - \frac{1}{\varepsilon}\cdot |\varepsilon_i|\} \underbrace{\le}_{(***)} \sum_i max\{0,0\} = \sum_i 0 = 0$$

כאשר השיוויון המסומן ב-$(**)$, נובע מכך ש-$\varepsilon_i < 0$, ולכן $\varepsilon_i = -|\varepsilon_i|$. אי-השיוויון המסומן ב-$(***)$, נובע מהטיעונים הבאים –

$$\varepsilon = \min_j |\varepsilon_j| \le |\varepsilon_i| \quad \Rightarrow \quad \frac{1}{\varepsilon} \ge \frac{1}{|\varepsilon_i|} \quad \Rightarrow \quad -\frac{1}{\varepsilon} \le -\frac{1}{|\varepsilon_i|} \quad \Rightarrow -\frac{1}{\varepsilon}\cdot|\varepsilon_i| \le -\frac{1}{|\varepsilon_i|}\cdot|\varepsilon_i| \quad \Rightarrow \mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{\varepsilon}}\cdot|\boldsymbol{\varepsilon_i}| \le 1 - \underbrace{\frac{1}{|\varepsilon_i|}\cdot|\varepsilon_i|}_{=1} = 0$$

כלומר, מצאנו כי –

$$\sum_i \Delta_{zo}\left(f\left(x_i; w_1^{opt},\dots,w_L^{opt}\right), y_i\right) \le 0$$

וכמו כן, הוכחנו קודם לכן כי – $0 \le \Delta_{zo}\left(f\left(x_i; w_1^{opt},\dots,w_L^{opt}\right), y_i\right)$, כלומר שמדובר בסכום של גורמים אי-שליליים החסום ע"י 0, ולכן בפרט ניתן להסיק את הרצוי –

$$\Delta_{zo}\left(f\left(x_i; w_1^{opt},\dots,w_L^{opt}\right), y_i\right) = 0$$

לכל $i\in[L]$, כנדרש. ■

שאלה 4. Growth Function of Composition.

נתונות $F_2 \subseteq \mathcal{Y}_2^{\mathcal{Y}_1}$, $F_1 \subseteq \mathcal{Y}_1^{\mathcal{X}}$, ונגדיר $F = F_2 \circ F_1$, המוגדרת ע"י הרכבה של פונק' מ-F_1 ומ-F_2, כך ש-

$$F = F_2 \circ F_1 = \{f_2 \circ f_1 \mid f_1 \in F_1, f_2 \in F_2\}$$

נדרש להוכיח כי מתקיים – $\pi_F(m) \le \pi_{F_1}(m) \cdot \pi_{F_2}(m)$.

פיתרון. תהי $C = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$ תת-קבוצה בגודל m. נגדיר בנוסף $f_1^C = \{f_1(x_1), \dots, f_1(x_m)\} = \{y_1, \dots, y_m \mid \ s.t. \ \ y_i = f(x_i)\} \subseteq \mathcal{Y}_1$, תת-קבוצה של$\mathcal{Y}_1$ בגודל m, שהינה תלויה כאמור בבחירת C וכן תלויה ב-$f_1 \in F_1$.
לפי ההגדרה, מתקיים –

$$|F_C| = |\{[f(x_1), \dots, f(x_m)]: f \in F, s.t. C = \{x_1, \dots, x_m\}\}| =$$

$$= \left|\{\left[f_2\left(f_1(x_1)\right), \dots, f_2\left(f_1(x_m)\right)\right]: f_1 \in F_1, f_2 \in F_2, s.t. \ \ C = \{x_1, \dots, x_m\}\}\right| =$$

$$= |\{[f_2(y_1), \dots, f_2(y_m)]: f_1 \in F_1, f_2 \in F_2, s.t. C = \{x_1, \dots, x_m\} \ and \ \ y_i = f_1(x_i)\}| =$$

$$\underbrace{=}_{\substack{def \\ of \\ F_{1_C}}} \left| \bigcup_{[y_1, \dots, y_m] \in F_{1_C}} \{[f_2(y_1), \dots, f_2(y_m)]: f_2 \in F_2\} \right| \le$$

$$\underbrace{\le}_{\substack{Union \\ Bound}} \sum_{[y_1, \dots, y_m] \in F_{1_C}} |\{[f_2(y_1), \dots, f_2(y_m)]: f_2 \in F_2\}| \ =$$

$$= \sum_{[y_1, \dots, y_m] \in F_{1_C}} \left|F_{2_{f_1^C}}\right| \ \le \sum_{[y_1, \dots, y_m] \in F_{1_C}} \max_{\substack{C' \subseteq \mathcal{Y}_1 \\ s.t. \ |C'|=m}} \left|F_{2_{C'}}\right| \ = \sum_{[y_1, \dots, y_m] \in F_{1_C}} \pi_{F_2}(m) \ =$$

$$= \left|F_{1_C}\right| \cdot \pi_{F_2}(m) \le \max_{\substack{C' \subseteq \mathcal{X} \\ s.t. \ |C'|=m}} \left|F_{1_{C'}}\right| \cdot \pi_{F_2}(m) = \pi_{F_1}(m) \cdot \pi_{F_2}(m)$$

נבחין כי הקבוצה $C = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$ הינה קבוצה כללית מגודל m, ולכן לכל קבוצה $C \subseteq \mathcal{X}$, מגודל m, מתקיים – $|F_C| \le \pi_{F_1}(m) \cdot \pi_{F_2}(m)$,
ולפיכך ניתן להסיק כי הא"ש מתקיים גם עבור תת-קבוצה C מגודל m שעבורה $|F_C|$ מקסימלי, כלומר –

$$\underbrace{\max_{\substack{C \subseteq \mathcal{X} \\ s.t. \ |C|=m}} \ \ |F_C|}_{=\pi_F(m)} \ \le \pi_{F_1}(m) \cdot \pi_{F_2}(m)$$

ואכן קיבלנו את הנדרש – $\pi_F(m) \le \pi_{F_1}(m) \cdot \pi_{F_2}(m)$, כרצוי. ■

שאלה 5. Gradient Descent on Smooth Function.

נתונה $l: \mathbb{R}^n \to \mathbb{R}$ פונק' שהיא β-חלקה($\beta - Smooth$) ואי-שלילית. נסתכל על אלגוריתם GD עבור הפונק' l עם גודל צעד קבוע $\eta > 0$ –

$$\boldsymbol{x_{t+1}} = \boldsymbol{x_t} - \eta \nabla l(\boldsymbol{x_t})$$

עם נק' התחלה של $\boldsymbol{x_0}$. נוכיח שאם $\eta < \frac{2}{\beta}$, אזי מתקיים –

$$\lim_{t\to\infty} ||\nabla l(\boldsymbol{x_t})|| = 0$$

פיתרון. נתון כי l פונק' שהיא β-חלקה, ולכן לכל $x, y \in \mathbb{R}^n$, מתקיים –

$$l(\boldsymbol{y}) \le l(\boldsymbol{x}) + \nabla l(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\beta}{2}||\boldsymbol{x} - \boldsymbol{y}||^2$$

ובפרט עבור הנקודות $\boldsymbol{x_{t+1}}, \boldsymbol{x_t}$ יתקיים – $\boldsymbol{x_{t+1}} - \boldsymbol{x_t} = -\eta\nabla l(\boldsymbol{x_t})$, וע"י שימוש בבך ש-$l$ פונק' שהיא β-חלקה, נקבל –

$$l(\boldsymbol{x_{t+1}}) \le l(\boldsymbol{x_t}) + \nabla l(\boldsymbol{x_t})^T \underbrace{(\boldsymbol{x_{t+1}} - \boldsymbol{x_t})}_{=-\eta\nabla l(\boldsymbol{x_t})} + \frac{\beta}{2}\underbrace{||\boldsymbol{x_t} - \boldsymbol{x_{t+1}}||^2}_{\substack{=||\boldsymbol{x_{t+1}}-\boldsymbol{x_t}||^2 \\ =||-\eta\nabla l(\boldsymbol{x_t})||^2}} = l(\boldsymbol{x_t}) - \eta \underbrace{\nabla l(\boldsymbol{x_t})^T \nabla l(\boldsymbol{x_t})}_{=||\nabla l(\boldsymbol{x_t})||^2} + \eta^2\frac{\beta}{2}||\nabla l(\boldsymbol{x_t})||^2 =$$

$$= l(\boldsymbol{x_t}) - \eta||\nabla l(\boldsymbol{x_t})||^2 + \eta^2\frac{\beta}{2}||\nabla l(\boldsymbol{x_t})||^2 = l(\boldsymbol{x_t}) + (\eta^2\frac{\beta}{2} - \eta)||\nabla l(\boldsymbol{x_t})||^2$$

ומכאן נקבל –

$$(\mathbf{1}) \quad -(\eta^2\frac{\beta}{2} - \eta)||\nabla l(\boldsymbol{x_t})||^2 \le l(\boldsymbol{x_t}) - l(\boldsymbol{x_{t+1}})$$

ונבחין כי

$$\eta^2\frac{\beta}{2} - \eta \underbrace{\le}_{\eta<\frac{2}{\beta}} \eta \cdot \underbrace{\frac{2}{\beta}\cdot\frac{\beta}{2}}_{=1} - \eta = \eta - \eta = 0 \qquad \Rightarrow -(\eta^2\frac{\beta}{2} - \eta) < 0$$

ולכן ע"י חלוקה של אי-השיוויון $(\mathbf{1})$ בגורם $-(\eta^2\frac{\beta}{2} - \eta)$, נקבל כי מתקיים –

$$||\nabla l(\boldsymbol{x_t})||^2 \le \underbrace{(-(\eta^2\frac{\beta}{2} - \eta))}_{const.} \cdot \left(l(\boldsymbol{x_t}) - l(\boldsymbol{x_{t+1}})\right)$$

ואי-שיוויון זה מתקיים באמור לכל $t \ge 0$ טבעי. נוכיח כי $\sum_{t=0}^{\infty}||\nabla l(\boldsymbol{x_t})||^2 < \infty$, כלומר מתכנסת, ולכן מכאן לפי ההתכנסות הטור הנ"ל, נסיק כי האיבר הכללי $||\nabla l(\boldsymbol{x_t})||^2 \xrightarrow[t\to\infty]{} 0$, ולפיכך בהכרח מתקיים $||\nabla l(\boldsymbol{x_t})|| \xrightarrow[t\to\infty]{} 0$, כפי שנדרש. על מנת להוכיח התכנסות של הטור, נוכיח התכנסות של סדרת הסכומים החלקיים. מתקיים –

$$\sum_{t=0}^{N}||\nabla l(\boldsymbol{x_t})||^2 \le \underbrace{\left(-\left(\eta^2\frac{\beta}{2} - \eta\right)\right)}_{const.} \cdot \sum_{t=0}^{N} l(\boldsymbol{x_t}) - l(\boldsymbol{x_{t+1}}) = \underbrace{\left(-\left(\eta^2\frac{\beta}{2} - \eta\right)\right)}_{const.} \cdot \left(l(\boldsymbol{x_0}) - l(\boldsymbol{x_{N+1}})\right) \underbrace{\le}_{l(\cdot)\ge 0} \underbrace{\left(-\left(\eta^2\frac{\beta}{2} - \eta\right)\right) \cdot l(\boldsymbol{x_0})}_{\substack{const. \\ \forall N\in\mathbb{N}}}$$

כלומר, מצאנו כי סדרת הסכומים החלקיים חסומה ע"י קבוע. נבחין בנוסף כי סדרת הסכומים החלקיים הינה סדרה מונוטונית עולה, שכן $||\cdot|| \ge 0$, **ולפיכך ניתן להסיק כי סדרת הסכומים מתכנסת שכן היא מונו' עולה וחסומה מלעיל.**
לכן, נסיק כי –

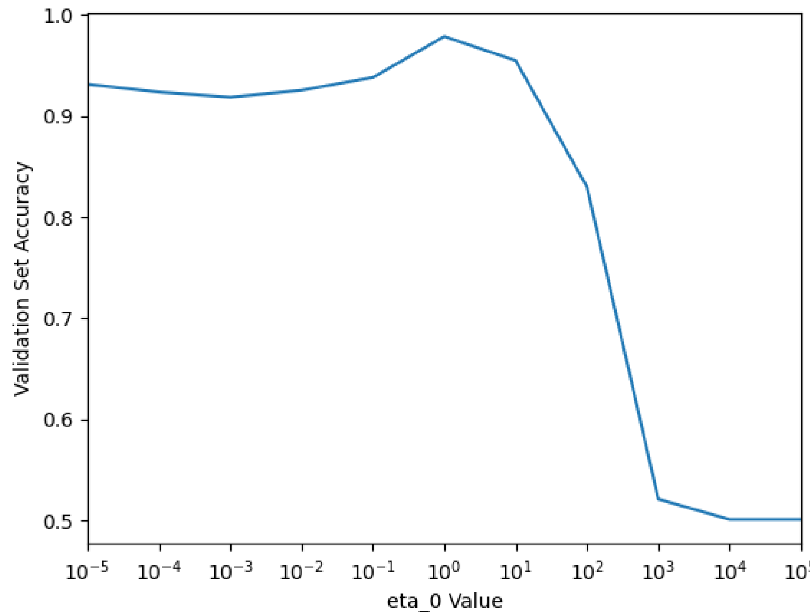
$$\sum_{t=0}^{\infty}||\nabla l(\boldsymbol{x_t})||^2 < \infty$$

ולכן האיבר הכללי שואף ל-אפס, כלומר $||\nabla l(\boldsymbol{x_t})||^2 \xrightarrow[t\to\infty]{} 0$, ולפיכך נסיק כי – $||\nabla l(\boldsymbol{x_t})|| \xrightarrow[t\to\infty]{} 0$, **כנדרש**. ■

Programming Assignment.

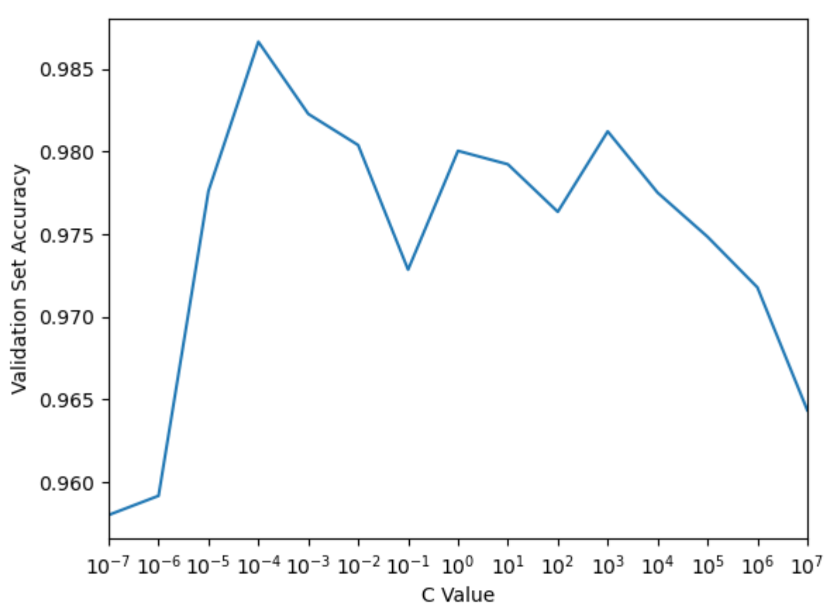
שאלה 1. SGD for Hinge Loss.

סעיף א. נגדיר $T = 1000, C = 1$, ונבצע SGD על גבי $\eta_0 \in \{10^{-5}, 10^{-4}, \dots, 10^4, 10^5\}$, ונבצע $cross - validation$, כדי לאמוד את טיב ההיפותזה שנלמדה, כדי לקבל את הגרף הבא של דיוק הממוצע על גבי סט הואלידציה($accuracy\ on\ validation\ set$), כפונקך של η_0. התוצאות –



נבחין כי η_0 הטוב ביותר התקבל עבור $\boxed{\eta_0 = 10^0 = 1}$, עפ"י הגרף שהתקבל.

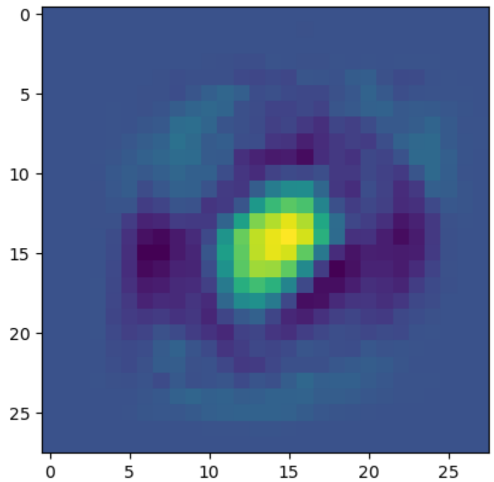
סעיף ב. **בסעיף זה נשתמש ב-η_0 הטוב ביותר שהתקבל בסעיף א', $\eta_0 = 1$.**
נגדיר גם $T = 1000$, ונבצע SGD על גבי $C \in \{10^{-7}, 10^{-6}, \dots, 10^6, 10^7\}$, ונבצע $cross - validation$, כדי לאמוד את טיב ההיפותזה שנלמדה, כדי לקבל את הגרף הבא של דיוק הממוצע על גבי סט הואלידציה($accuracy\ on\ validation\ set$), כפונקך של C. התוצאות –



נבחין כי C הטוב ביותר התקבל עבור $\boxed{C = 10^{-4}}$, עפ"י הגרף שהתקבל.

סעיף ג. בסעיף זה נשתמש ב-η_0 הטוב ביותר שהתקבל בסעיף א', $\eta_0 = 1$, וב-C הטוב ביותר שהתקבל בסעיף ב', $C = 10^{-4}$.

נאמן את המסווג עבור $T = 20000$, עם הפרמטרים שמצאנו בסעיפים הקודמים. נציג את w, וקטור המשקלים, המתקבל כתמונה –



ניתן לראות כי ישנו ריכוז במרכז התמונה, כלומר הפיקסלים במרכז התמונה הם הפיקסלים המרכזיים יותר באבחנה בין המספר 0 למספר 8.

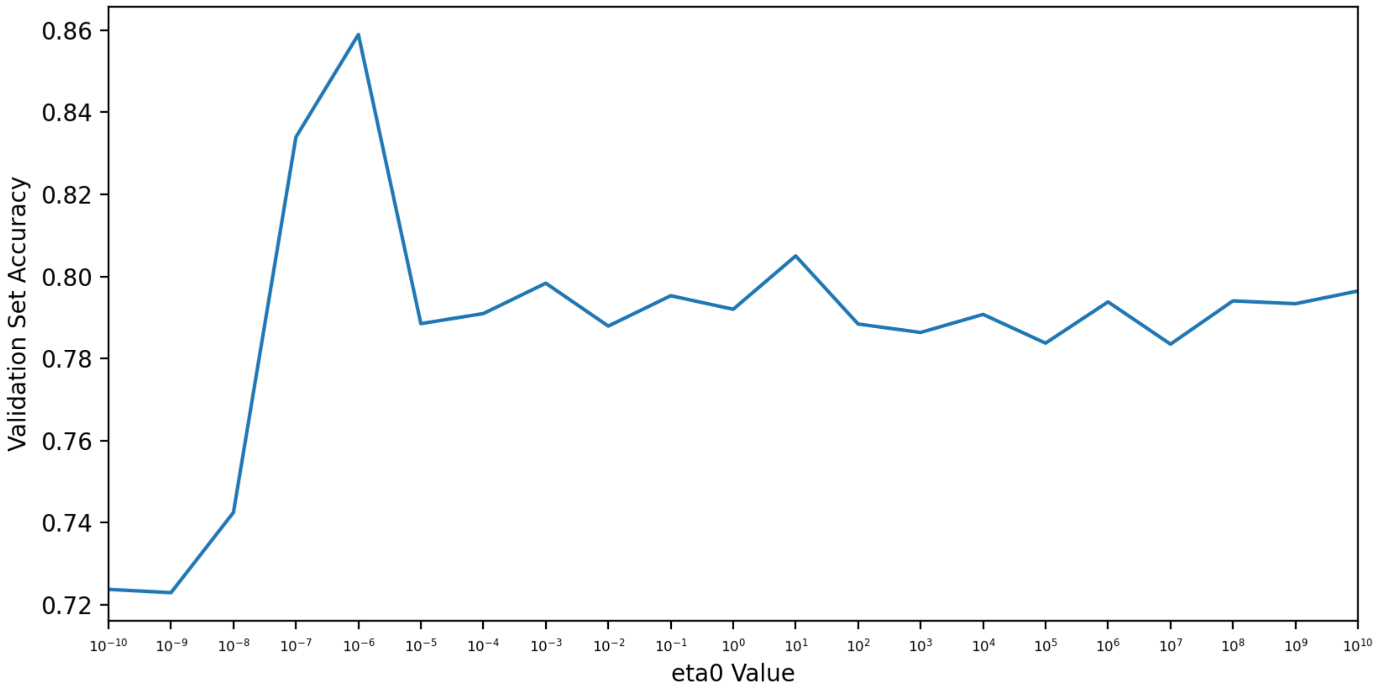
סעיף ד. נחשב את הדיוק של המסווג האופטימלי על גבי ה-*test set*.

```
The accuracy of the best classifier on the test set is 0.9923234390992836
```

הדיוק המתקבל הינו $\boxed{0.9923234390992836}$.

שאלה 2. SGD for Multi Class Cross Entropy.

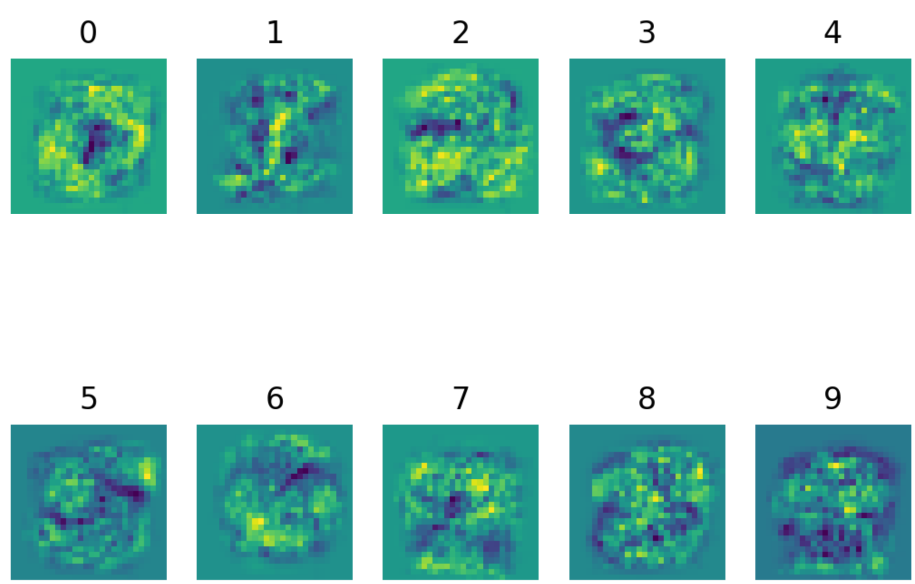
סעיף א. נגדיר $T = 1000$, ונבצע SGD על גבי $\eta_0 \in \{10^{-10}, 10^{-9}, \dots, 10^9, 10^{10}\}$, ונבצע $cross - validation$, כדי לאמוד את טיב ההיפותזה שנלמדה, כדי לקבל את הגרף הבא של דיוק הממוצע על גבי סט הואלידציה $(accuracy\ on\ validation\ set)$, כפונקציה של η_0. התוצאות –



נבחין כי η_0 הטוב ביותר התקבל עבור $\boxed{\eta_0 = 10^{-6}}$, עפ"י הגרף שהתקבל.

סעיף ב. בסעיף זה נשתמש ב-η_0 הטוב ביותר שהתקבל בסעיף א', $\eta_0 = 10^{-6}$. נאמן את המסווג עם $\eta_0 = 10^{-6}$ עבור $T = 20000$.

נציג את וקטורי המשקלים $w_0, w_1, \dots, w_9$ שהתקבלו בתמונות –



סעיף ג. נחשב את הדיוק של המסווג האופטימלי על גבי ה-$test\ set$.

```
The accuracy of the best classifier on the test set is 0.869
```

הדיוק המתקבל הינו $\boxed{0.869}$.