מבוא ללמידה חישובית – מטלה 3

1)נחשב את ההסיאן של $g(x) = f(Ax+b)$

נשתמש בכלל השרשרת:

$$\nabla g(x) = \nabla \mathrm{f}(Ax+b)A = A^T \nabla \mathrm{f}(Ax+b)$$

ושוב לי כלל השרשרת:

$$H_g(x) = A^T \mathrm{H}_{\mathrm{f}}(Ax+b)A$$

נתון ש f קמורה לכן $H_f(Ax+b)$ היא PSDו וסימטרית לכן קיימת Q הפיכה כך ש:

$$H_f(Ax+b) = Q^T D Q$$

וD אלכסונית לכן מתקיים

$$x^T A^T H_f(Ax+b)Ax = x^T A^T Q^T D Q A x = \sum_{i=1}^{n} (QAx)_i^2 D_{i,i}$$

$D_{i,i} \geq 0$ לכל i כי ההסיאן הוא PSD לכן

$$\forall x \in \mathbb{R}^n \ \sum_{i=1}^{n} (QAx)_i^2 D_{i,i} \geq 0 \Rightarrow g(x) \ is\ convex$$

(2

יהיה $x, y \in \mathbb{R}^d$ ו$\forall \lambda \in [0,1]$ נסמן $t = argmax_{i \in [m]} f_i(\lambda x + (1-\lambda)y)$

(a

$$\max_i f_i(\lambda x + (1-\lambda)y) = f_t(\lambda x + (1-\lambda)y) \leq \lambda f_t(x) + (1-\lambda) f_t(y) \leq \lambda \max_i f_i(x) + (1-\lambda)\max_i f_i(y) \Rightarrow \max_i f_i(x)\ is\ convex \Rightarrow g(x)\ is\ convex$$

(b

שנסמן $t = argmax_{i \in [m]} f_i(x)$ נתון שf_i דיפרנרצבלית לכן רציפות לכל i לכן קיימת סביבה קמורה(כדור עם רדיוס קטן מספיק) $U \subset \mathbb{R}^d$ כך ש:

$$\forall x \in U\ g(x) = f_t(x)$$

לכן מתקיים(מקמירות) :

$$\forall w, u \in U\ g(u) = f_t(u) \geq f_t(w) + < u - w, \nabla \mathrm{f}_t(u) >$$

לכן ∇f_t הוא סאב-גדיאנט של g

(3

a) נראה שℓ כפונקציה של המשקולות קמורה:

יהיה $w_1, \dots w_L$ לכן קיים $\hat{y}$ כך $w_{\hat{y}} x$ מקסימלי

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

נשים לב ש$\Delta_{zo}(\hat{y}, y)$ הוא קבוע ולכן קמור(כי פונקציה קבועה היא קמורה

$\max_{\hat{y}\in[L]} w_{\hat{y}}x$ קמורה משאלה 2 , $w_y x$ קמורה(פונקציה לינארית היא קמורה) ולכן $\ell(w_1,\dots,w_L,x,y)$ קמורה כפונקציה של $w_1,\dots,w_L$

(b

$$\max_{\hat{y}\in[L]} w_{\hat{y}}x - w_y x \geq 0 \Rightarrow$$

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

(c

$$\begin{aligned}\min_{w_1,\dots,w_L} \sum_i \ell(w_1,\dots,w_L,x,y) \\ &= \sum_i \ell\left(w_1^{opt},\dots,w_L^{opt},x_i,y_i\right) \Rightarrow 0 \\ &= \sum_i \ell(w_1^*,\dots,w_L^*,x_i,y_i) \geq \sum_i \ell\left(w_1^{opt},\dots,w_L^{opt},x_i,y_i\right) \geq 0\end{aligned}$$

לכן $\sum_i \ell\left(w_1^{opt},\dots,w_L^{opt},x_i,y_i\right) = 0$ ובסעיף b ראינו ש

$$\ell(w_1,\dots,w_L,x,y) \geq \Delta_{zo}(f(x;w_1,\dots,w_L),y) \in \{0,1\} \Rightarrow \Delta_{zo}(f(x;w_1^*,\dots,w_L^*),y) = 0$$

4

הערה: כדי נסמן את מספר הצעדים ב$\mathbb{T}$ כדי ששיחלוף ומספר הצעדים לא יסמנו באותו סימן בנוסף נניח כי $w \in \mathbb{R}^d$

$$f(w) = w^T A w \Rightarrow \nabla f(w) = (A + A^T)w = 2Aw$$

לכן

$$w_{t+1} = w_t - 2\eta A w_t = (1 - 2\eta A)w_t \Rightarrow w_{\mathbb{T}} = (1 - 2\eta A)^{\mathbb{T}-1} w_1$$

נרצה לבחור η כך ש

$$f(w_{\mathbb{T}}) = f((1-2\eta A)^{\mathbb{T}-1}w_1) = \left((1-2\eta A)^{\mathbb{T}-1}w_1\right)^T A\left((1-2\eta A)^{\mathbb{T}-1}w_1\right) \leq \epsilon$$

נסמן λ_{max} את הערך העצמי המקסימלי של A וλ_{min} את הערך העצמי המינימלי של A בנוסף בגלל שA PCD מתקיים:

$$0 > \lambda_{min} > \lambda_{max} , \forall v \in \mathbb{R}^d \; \lambda_{min}||v|| \leq v^T A v \leq \lambda_{max}||v||$$

$$\begin{aligned}\left((1-2\eta A)^{\mathbb{T}-1}w_1\right)^T A\left((1-2\eta A)^{\mathbb{T}-1}w_1\right) &\leq \lambda_{max}\left||(1-2\eta A)^{\mathbb{T}-1}w_1\right|| \\ &= \lambda_{max}\left((1-2\eta A)^{\mathbb{T}-1}w_1\right)^T\left((1-2\eta A)^{\mathbb{T}-1}w_1\right) \leq \lambda_{max}(1-2\eta\lambda_{min})^{2\mathbb{T}-2}||w_1||\end{aligned}$$

נבחר $\eta = \frac{1 - \frac{\epsilon}{\lambda_{max}||w_1||}^{\frac{1}{2\mathbb{T}-2}}}{2\lambda_{min}}$

ונקבל

$$f(w_{\mathbb{T}}) \leq \epsilon$$

לכן

$$(1-2\eta\lambda_{min})^{2\mathbb{T}-2} \le \frac{\epsilon}{\lambda_{max}||w_1||} \Rightarrow 2\mathbb{T}-2 \le \frac{\log\left(\frac{\epsilon}{\lambda_{max}||w_1||}\right)}{\log\left((1-2\eta\lambda_{min}\right)} = \frac{\log\left(\frac{\lambda_{max}||w_1||}{\epsilon}\right)}{-\log\left((1-2\eta\lambda_{min}\right)}$$

כאשר $1-2\eta\lambda_{min}$ חיובי קטן מאחד עבור אפסילון קטן מספיק לכן

$$\mathbb{T} \le \frac{\log\left(\frac{\lambda_{max}||w_1||}{\epsilon}\right)}{-2\log\left((1-2\eta\lambda_{min}\right)} + 2 \Rightarrow \mathbb{T} = O\left(\log\left(\frac{1}{\epsilon}\right)\right)$$

5)

1)

יהי $0 < \eta < \frac{2}{\beta}$. ((*) נשים לב $\eta\frac{\beta}{2} < 1 \Rightarrow 1-\frac{\beta}{2}\eta > 0$)

ℓ היא בטא-חלקה אז עבור x_t ו$x_{t+1} = x_t - \eta\nabla\ell(x_t)$ מתקיים:

$$\ell(x_{t+1}) \le \ell(x_t) + \nabla\ell(x_t)^T(x_{t+1}-x_t) + \frac{\beta}{2}\|x_t - x_{t+1}\|^2$$
$$= \ell(x_t) + \nabla\ell(x_t)^T\left(-\eta\nabla\ell(x_t)\right) + \frac{\beta}{2}\|\eta\nabla\ell(x_t)\|^2$$
$$= \ell(x_t) - \eta\|\nabla\ell(x_t)\|^2 + \frac{\beta}{2}\eta^2\|\nabla\ell(x_t)\|^2$$
$$= \ell(x_t) - \|\nabla\ell(x_t)\|^2\eta\left(1-\frac{\beta}{2}\eta\right)$$
$$\Rightarrow \|\nabla\ell(x_t)\|^2\eta\left(1-\frac{\beta}{2}\eta\right) \le \ell(x_t) - \ell(x_{t+1}) \overset{(*)}{\Rightarrow} \|\nabla\ell(x_t)\|^2 \le \frac{\left(\ell(x_t)-\ell(x_{t+1})\right)}{\eta\left(1-\frac{\beta}{2}\eta\right)}$$

נסמן $q = \frac{1}{\eta\left(1-\frac{\beta}{2}\eta\right)}$ אז מתקיים לכל $n \in \mathbb{N}$:

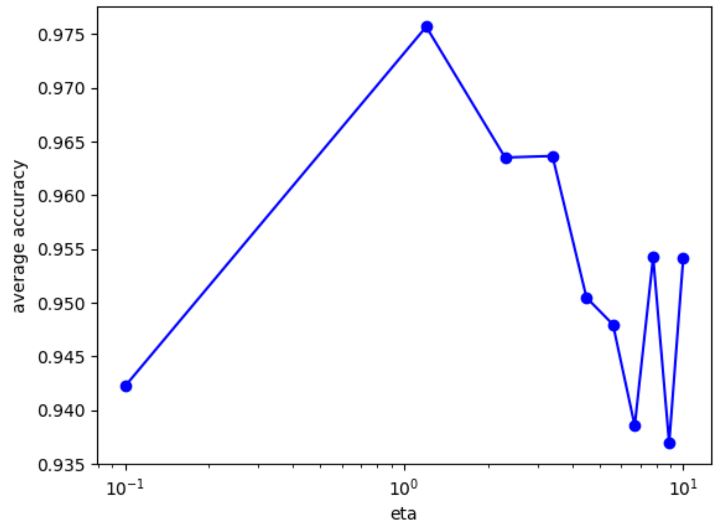
$$\sum_{t=0}^{n}\|\nabla\ell(x_t)\|^2 \le q\sum_{i=0}^{n}\left(\ell(x_t)-\ell(x_{t+1})\right) = q(\left(\ell(x_0)-\ell(x_n)\right)$$

קיבלנו שסדרת הסכומים החלקיים של הטור האינסופי $\sum_{t=0}^{n}\|\nabla\ell(x_t)\|^2$ חסומה כלומר הטור מתכנס ומתקיים

$$\lim_{t\to\infty}\|\nabla\ell(x_t)\|^2 = 0$$

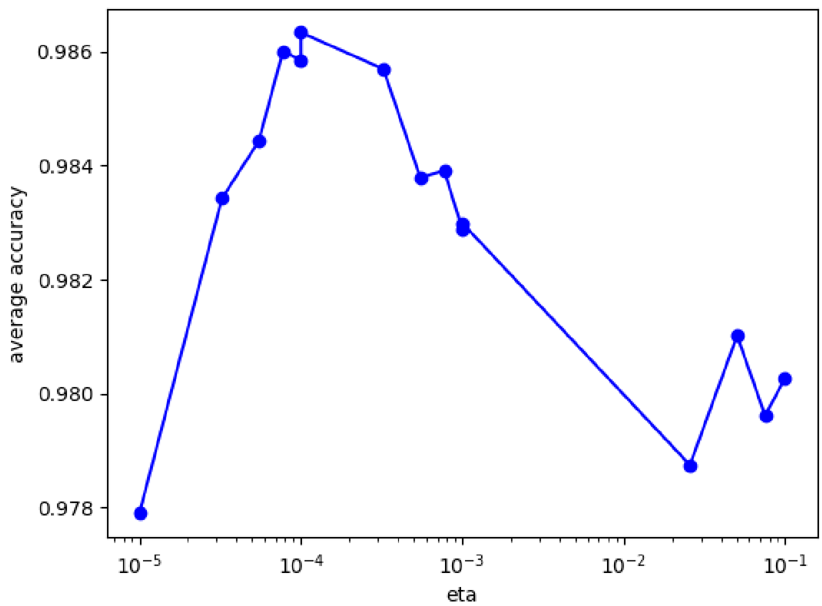
<u>חלק מעשי</u>

a)התחום המצומצם שמצאתי הוא $[10^{-1}, 10]$ קיבלנו את הגרף



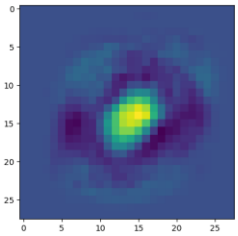
וη_0 האופטימלי שמצאנו הוא 0.85 בקירוב

b) בהינתן $\eta_0 = 0.85$ התחום המצומצם שבו C אופטימלי הוא $[10^{-5}, 10^{-1}]$ והגרף



והC האופטימי הוא 0.00011

(c

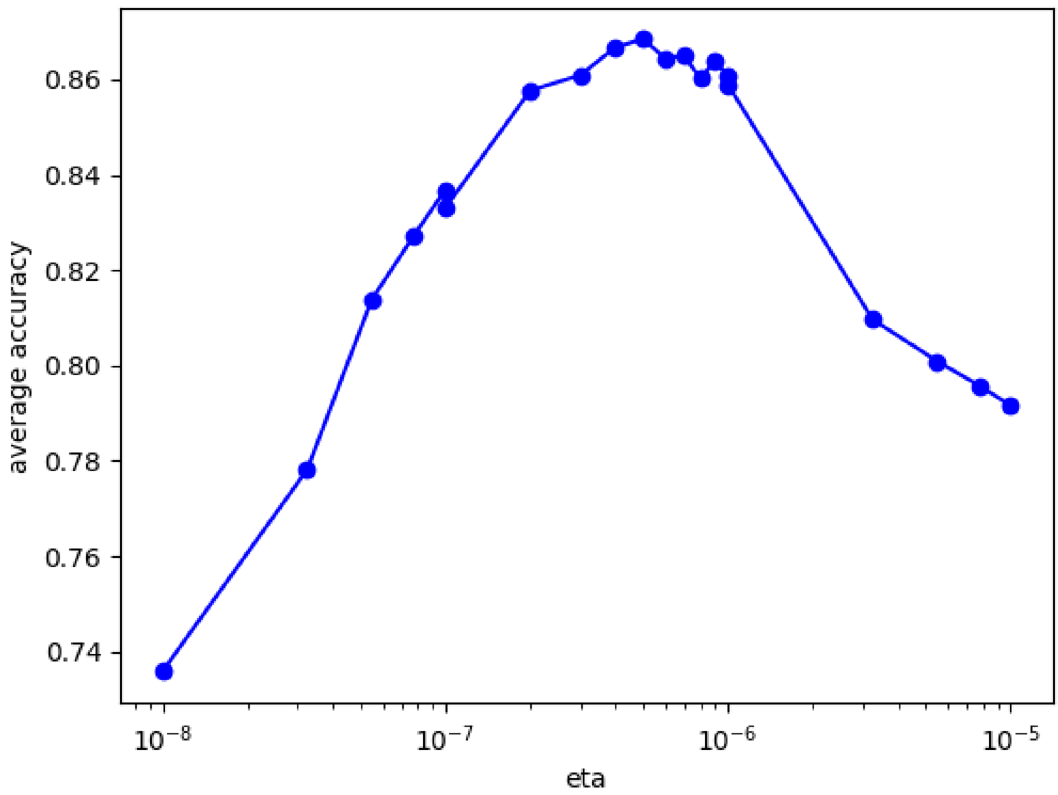


קיבלנו שרוב ה"משקל" הוא במרכז התמונה

(d לאחר הרצה על ה*test set* קיבלנו דיוק של 0.993858

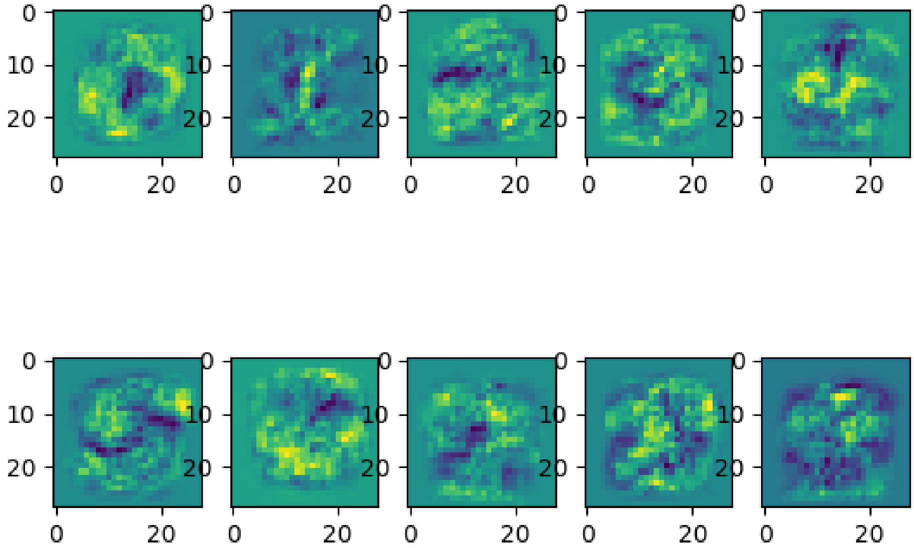
.2

(a לאחר בדיקה כל כמה תחומים התחום הרלוונטי שבו מקבלים הטא אופטימלי הוא $[10^{-8}, 10^{-5}]$ והגרף



והטא האופטימלית היא בערך $5 * 10^{-7}$

(b



בתמונות נראה שכל מסווג מקבל "משקל" גבוהה כך שבערך רואים במטושטש את המספר המתאים.

(c לאחר הרצה על ה*test set* קיבלנו דיוק של 0.8785