

(67822) מבוא ללמידה عمוקה | תרגיל 1

שם: עמית חן (308162502) , נדב אללי (313549206) | בהצלחה :

7 בנובמבר 2021

חלק פרקיי - דו"ח:

הסבר על **בנייה הרשות**:

הערות על הדאטה:

1. קיבלנו ייצוג של 9 תווים (מתוך 20 תווים אפשריים) שמייצג חומצת אmino כלשהי.
2. כמות דגימות חיוביות - 2991. ואילו כמות הדגימות השליליות הוא - 24492 (כלומר היחס הוא כמעט 100%-90%).
3. טיפלנו מראש ב מקרה של אותיות קטנות ברכפים שלנו (*lower – case*) על ידי שינוי כל התווים בקובץ *upper – case*.
רשות הנירונים שאנו צריכים לבנות נדרש האם רצף התווים, מזוהה (*detect*) על ידי ה- *HLA* או אינה מזוהה.
בחרנו לתרגם את רצף 9 התווים לוקטור, בייצוג אשר למדנו בכיתה (*one – hot – representation*).

הציג המודלים והתוצאות שלהם:

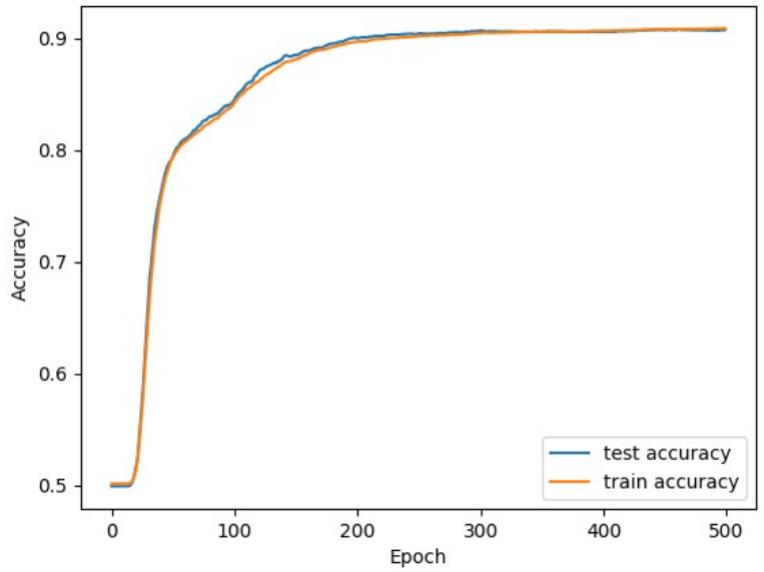
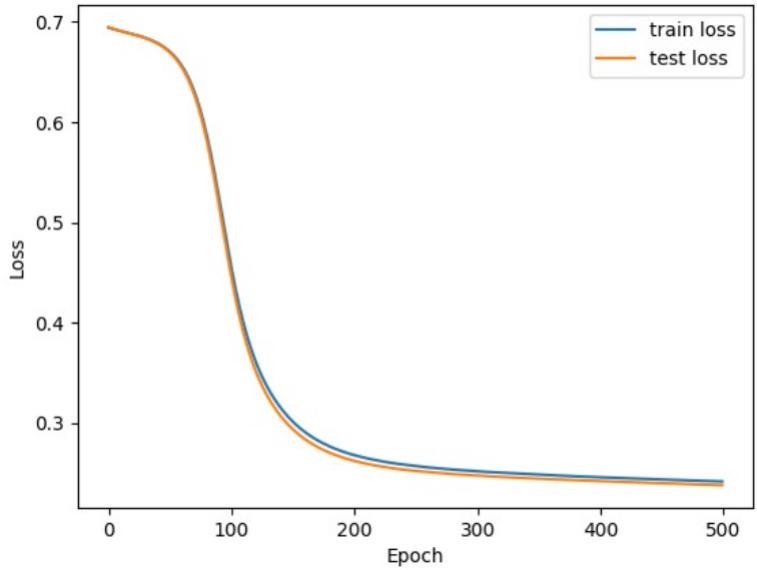
החליטנו למש סוגים שונים של רשותות שונות ומדדנו את ביצועיה לפי הדרכים שלמדו בכיתה, קרי בעזרת פונקציית *Loss* ובעזרת מדריך *Accuracy*.

א. **חלוקת הדאטה (Batch – size)** - ביצענו חלוקה של הדאטה לגזרים שונים, זאת כחלק מהניסויים שלנו לשיפור יכולות הרשות. כלומר חלוקה של הדאטה הכללי לקבוצות בגודלים הבאים:

$Batch – size = 24, 48$. כאשר גדים אלו יאמנו את הרשות בכל איטרציה שלה.

פונקציית ה-*Loss*: ביצענו ניסויים עם *BCE* (*Binary – Cross – Entropy*) *Loss* :

אופטימיזר - בחנו תחילת את *SGD* ברשות של 3 שכבות, אולם שמננו לב שבניסוי שביבענו קצב ירידת והתקנסות *h-Loss* מאוד איטי, כלומר רק לאחר ערך *Epoches* : 100 –



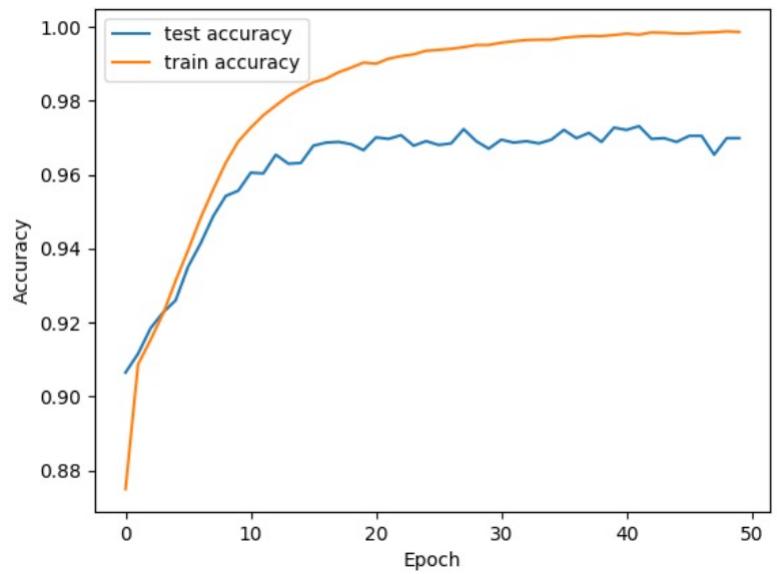
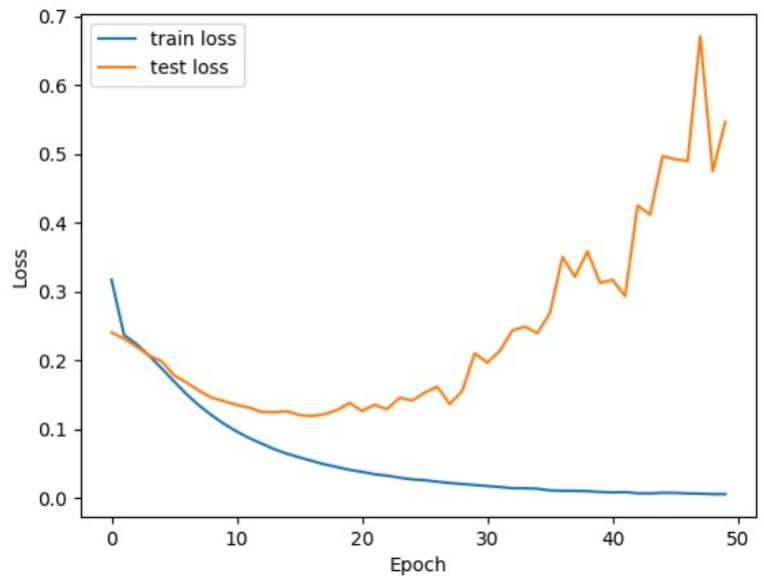
לכן בחרנו לבחון את ביצועי האופטימיזטור *Adam* בהמשך הניסויים שלנו.
כמויות ה- Adam Epochs : 50 בכל ניסוי שערךנו
ביצענו מספר בדיקות עליהן נפרט בהמשך. *learning – rate*

מבנה הרשתות שהרכנו:

אתחלנו את היפר הפרמטר: *OverSampling* **ו** *Batch – size = 24* (בגלל ההטייה שקיים בדאטה (כלומר דגימות שליליות רבות יותר מאשר חיוביות) בחרנו להוסיף דגימות חיוביות באופן מלאכותי על ידי פרוצדורה של שכפול דגימות רנדומלי כך שכמויות הדגימות החיוביות לאחר התחלת הוא באותו סדר גודל של הדגימות השליליות):
1 . השתמשנו ב-3 שכבות: (*Linear(180, 80)*, *Linear(80, 24)*, *Linear(24, 1)*), כאשר השתמשנו לאחר כל שכבה בפונקציית האקטיבציה *ReLU*, למעט בשכבה האחורונה בה השתמשנו ב-*Sigmoid*.

א. נצפה בתוצאות עבור $LR = 0.0003$

התוצאות:

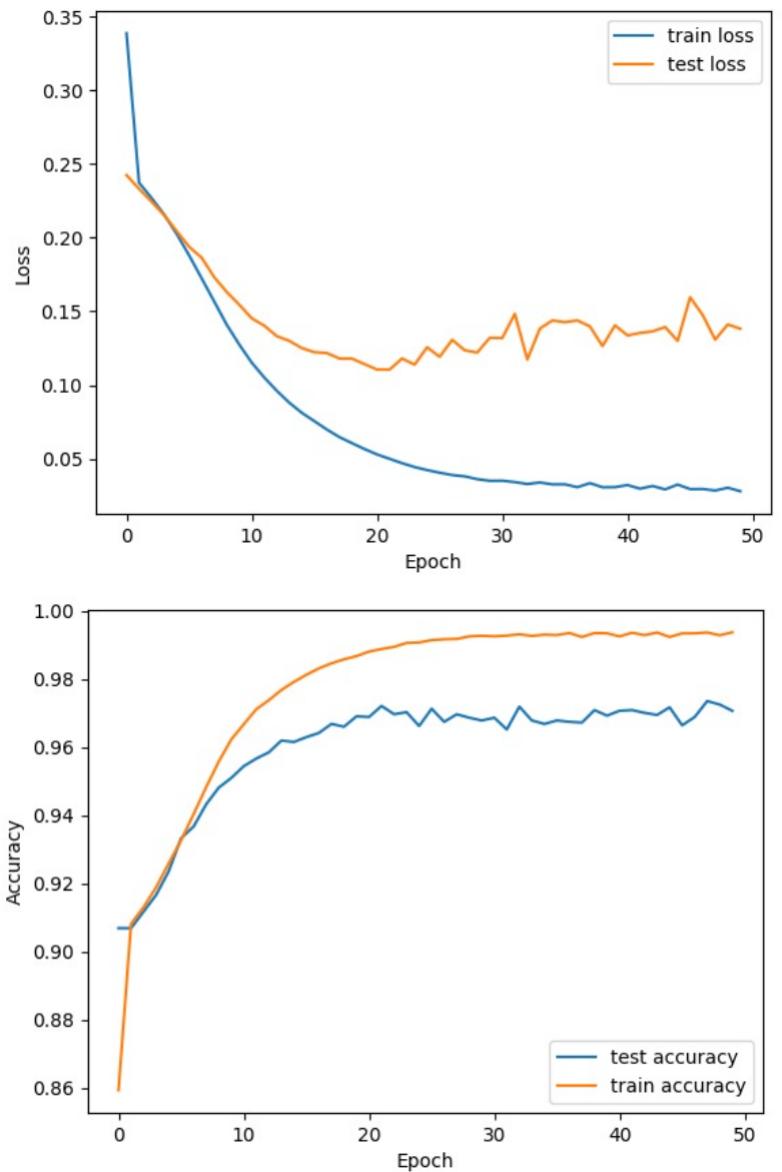


ניתן לראות שהשגיאה באימון יורדת באופן תקין, אך השגיאה בבדיקה יורדת וזו עולה באופן לא פרופורציוני, לכן להבנתנו ניתן להסיק שמדובר ב-*Overfit*.

ניתן לראות שהדיקוק באימון עולה בצורה תקינה וחלקה אך לעומת בדיקון הדיקוק לא מצליח להגיע לאותם ביצועים כמו באימונו.

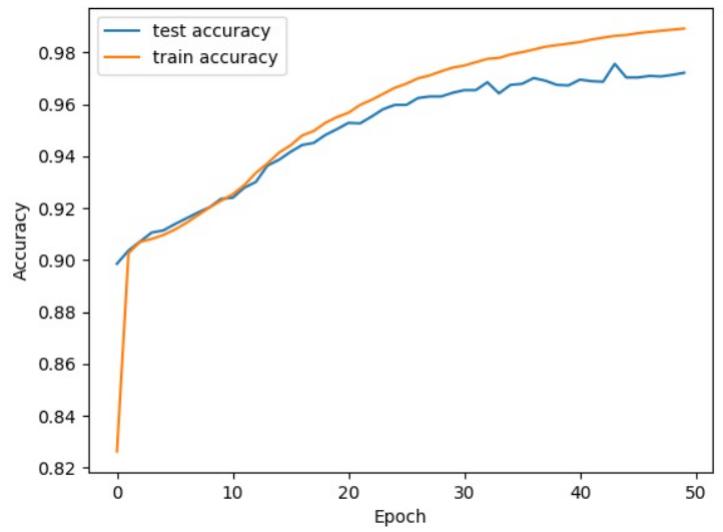
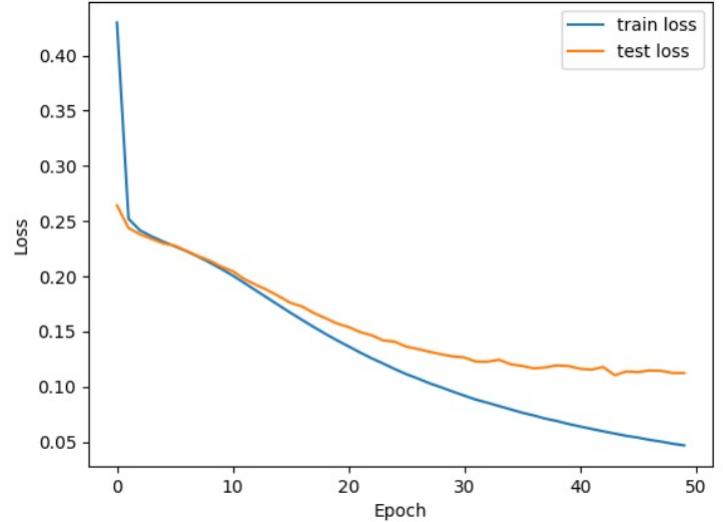
ב. נצפה בתוצאות עבור $LR = 0.0002$

התוצאות:



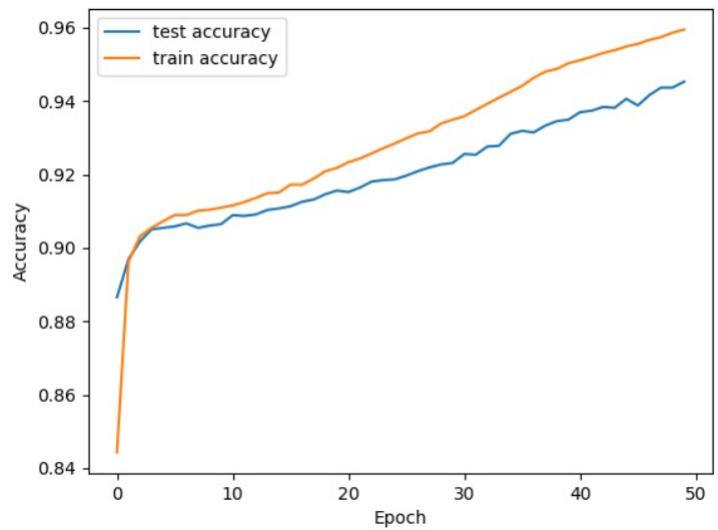
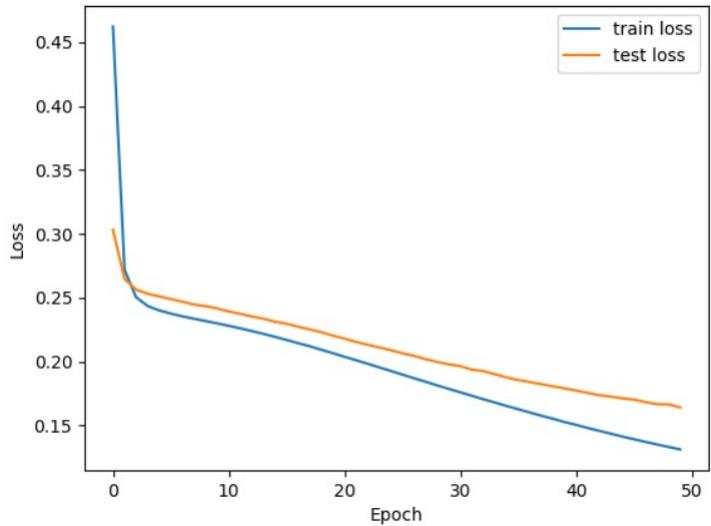
ניתן לראות שההטוצאות דומות למקירה של $LR = 0.0003$, וגם פה ניתן להבחין ב-*Overfit*

ג. נצפה בתוצאות עבור $LR = 0.0001$
התוצאות:



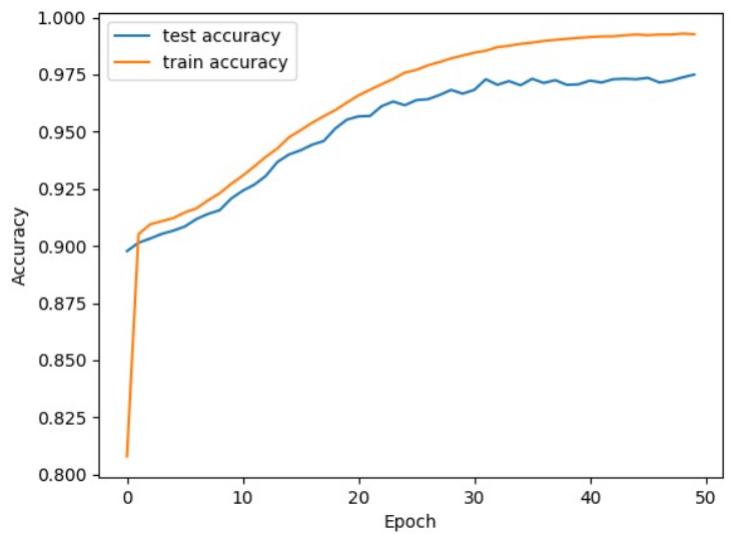
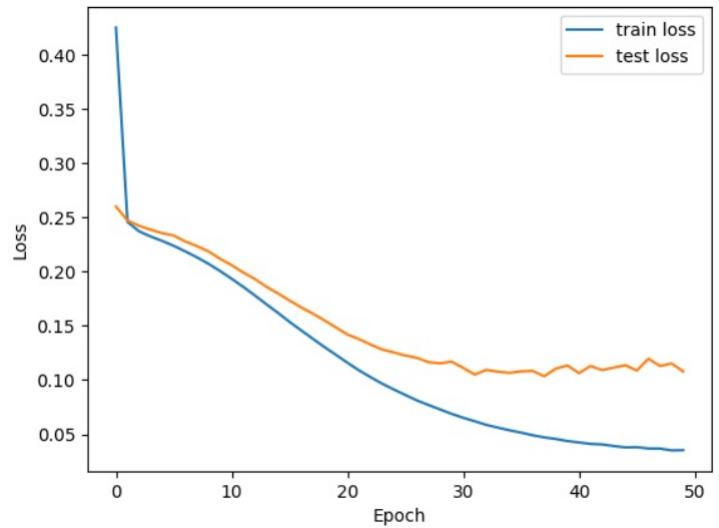
ניתן לראות שהשגיאה יורדת באופן תקין גם באימון וגם בבדיקה, אמנם לא באותו הקצב (להבנתנו זה סביר), ככלומר איננו רואים את תופעת *Overfit* שראינו בעבר $LR = 0.0002, LR = 0.0003$.
 ניתן לראות שהדיקות השתפים בבדיקה לעומת הבדיקות הקודמות, ככל הנראה הודות לכך $LR = 0.0001$.

2. השתמשנו ב-2 שכבות: $Linear(180, 80)$, $Linear(80, 1)$, כאשר השתמשנו לאחר השכבה הראשונה בפונקציית האקטיבציה $ReLU$, ובשכבה האחורונה השתמשנו ב- $Sigmoid$ וב- $LR = 0.0001$. התוצאות:



ניתן לראות שתוצאות שהשגיאה והדיקן הן סבירות, אם כי מעט פחות טובות לעומת הרשת בעלת 3 שכבות ("העומקה יותר").

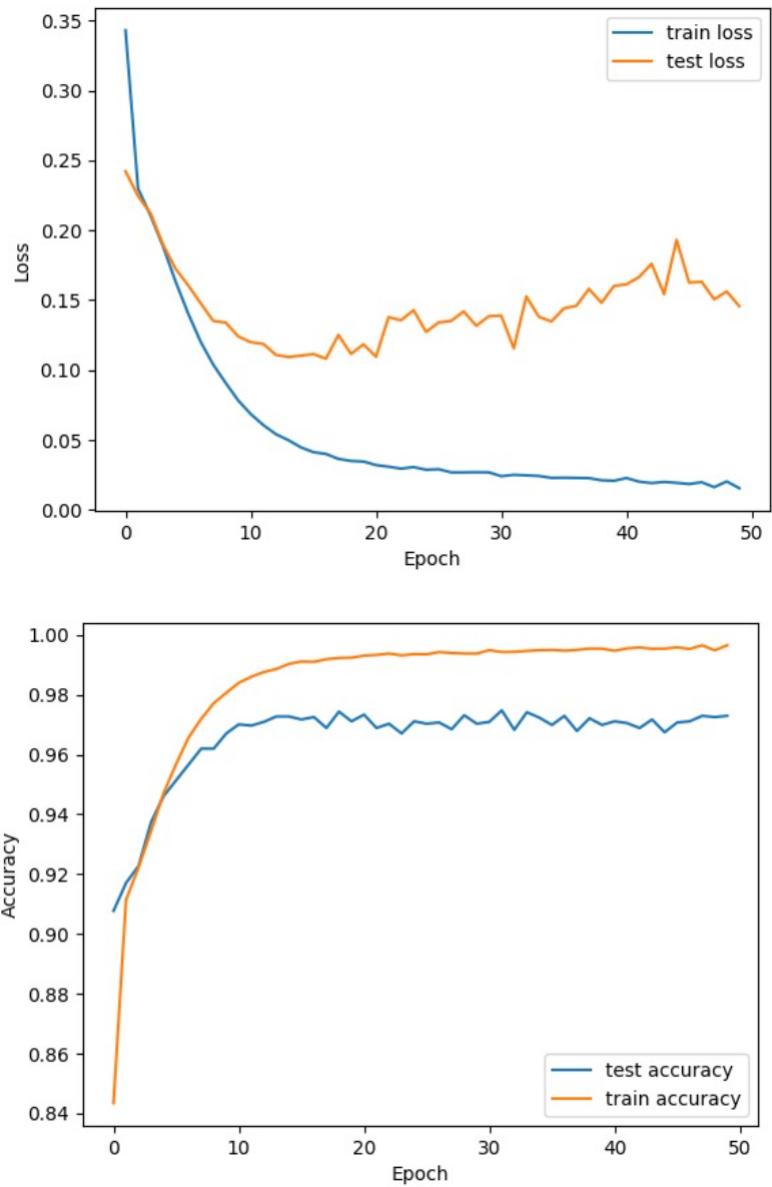
3. השתמשנו ב-4 שכבות: $\text{Linear}(180, 80)$, $\text{Linear}(80, 24)$, $\text{Linear}(24, 8)$, $\text{Linear}(8, 1)$, כאשר השתמשנו לאחר כל שכבה בפונקציית האקטיבציה ReLU , למעט בשכבה האחורונה בה השתמשנו ב- Sigmoid . נציין שבסוף לא בחרנו במודל זה מכיוון שלהבנתנו יש חשש ל- Overfit , שכן ניתן לראות שהשגיאה בבדיקה נשארת יחסית גבוהה ואינה יורדת כמו באימון. התוצאות:



כעת השתמשנו ב-5 שכבות: $Linear(180, 100)$, $Linear(100, 80)$, $Linear(80, 62)$, $Linear(62, 18)$, $Linear(18, 1)$ כאשר השתמשנו לאחר כל שכבה בפונקציית האקטיבציה $ReLU$, מיעט בשכבה האחורונה בה השתמשנו ב- $Sigmoid$. כמו כן

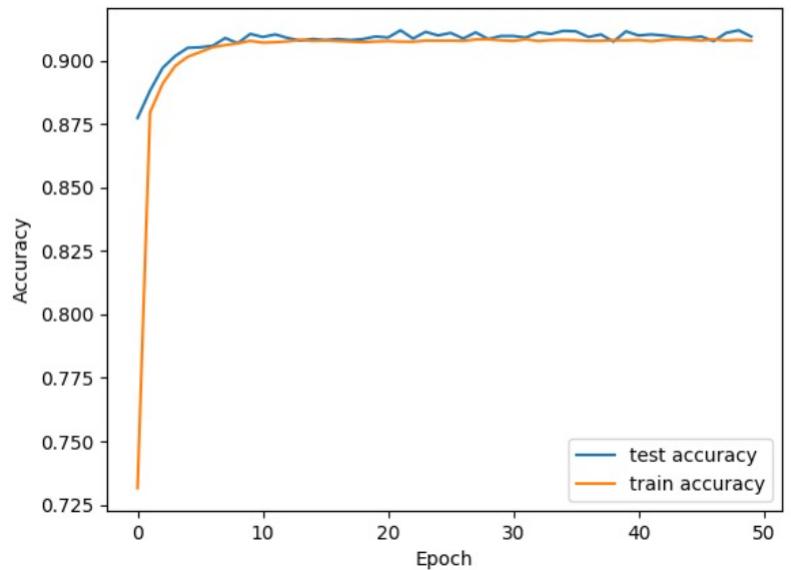
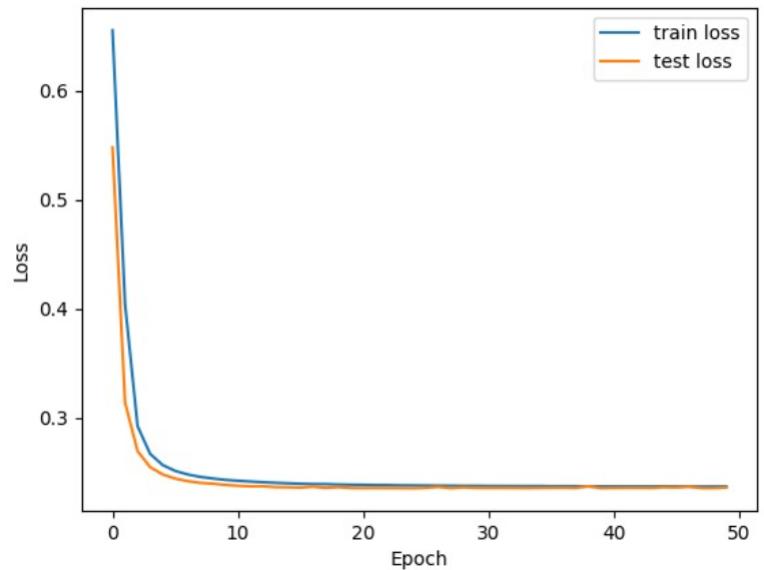
אתחלנו - $LR = 0.0001$

התוצאות:



ניתן לראות שהשגיאה בבדיקה גבוהה מאוד לעומת האימון ולכן להבנתנו מדובר בתופעת *Overfit*. כמו כן בדיקות ניתן לראות תופעה זהה.

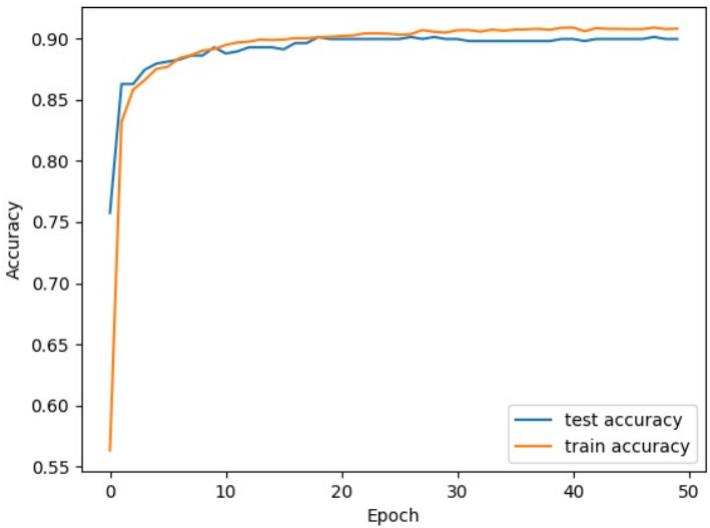
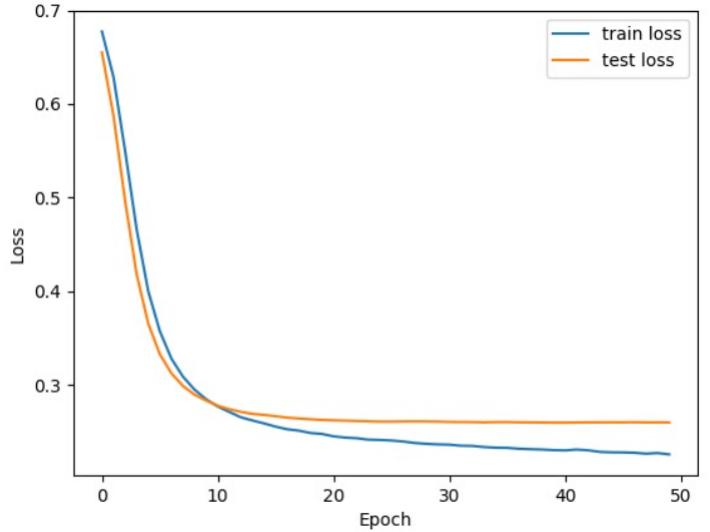
בדיקה מענינית נוספת - השתמשנו ב-3 שכבות: (1) כאשר השתמשנו לאחר כל שכבה בפונקציית האקטיבציה *Sigmoid*, למעט בשכבה. התוצאות:



ניתן לראות כי הרשת השיגה תוצאות יפות בסך הכל, אם כי לא טובות כמו ברשות שכוללת שימוש בפונקציית האקטיבציה *ReLU*.

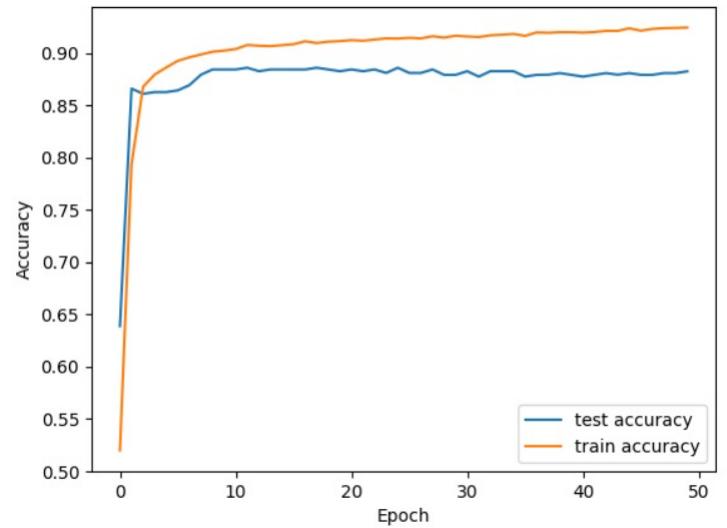
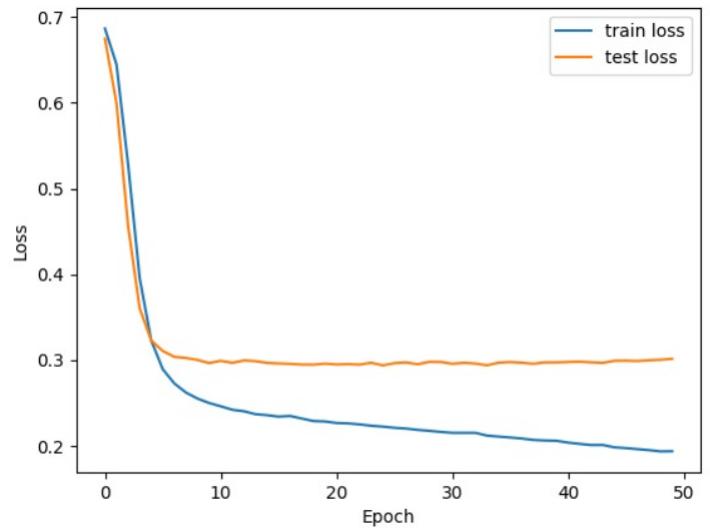
כעת נפעיל *Oversampling* במקומן *UnderSampling* שביצענו עד כה:

1. השתמשנו ב-2 שכבות: $Linear(180, 80)$, $Linear(80, 1)$, כאשר השתמשנו לאחר השכבה הראשונה בפונקציית האקטיבציה *ReLU*, ובשכבה האחורונה השתמשנו ב-*Sigmoid*. התוצאות:



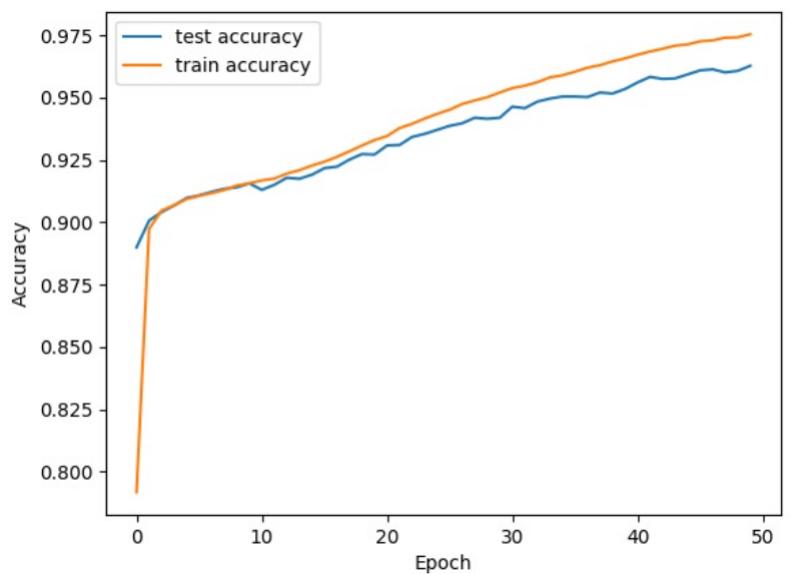
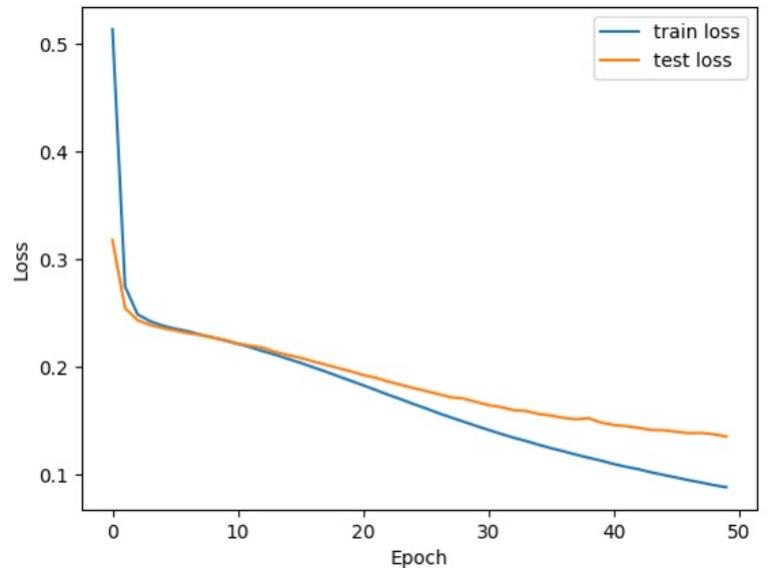
ניתן לראות שהתוצאות שהושגו עם שיטת *UnderSampling* בסך הכל טובות, אם כי פחות טובות מאשר שיטת *OverSampling*

2. השטמשנו ב-3 שכבות: $Linear(180, 80)$, $Linear(80, 24)$, $Linear(24, 1)$. כאשר השטמשנו לאחר כל שכבה בפונקציית *Sigmoid*, *ReLU*, למעט בשכבה האחרונה בה השטמשנו ב-*id*.
הaktivitzation: התוצאות:



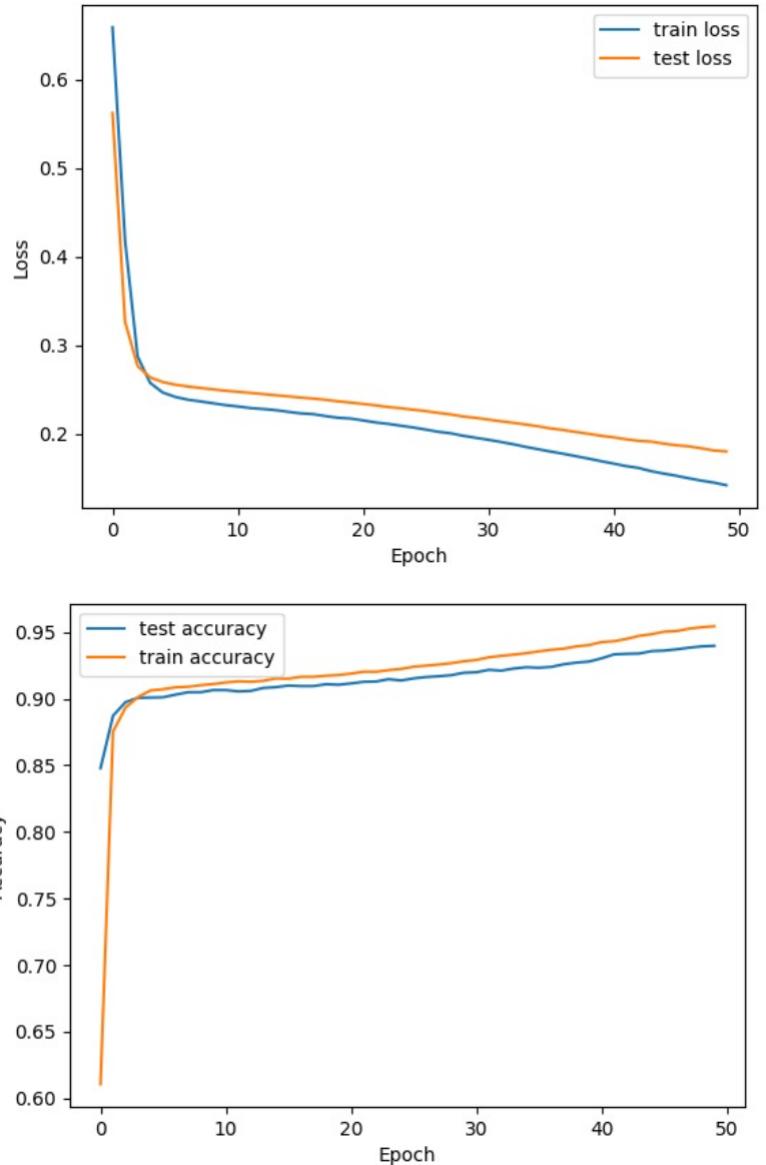
גם ניתן לראות שההתוצאות שהושגו עם שיטת *UnderSampling* בסך הכל טובות, אם כי פחות טובות מאשר שיטת *OverSampling*

cutת נונה את ההיפר פרמטר של גודל ה-*Batch* כך ש- $:Batch - Size = 48$ *Linear*(180, 80), *Linear*(80, 24), *Linear*(24, 1) *Sigmoid*, *ReLU* בעור 3 שכבות - בשכבה האחרונה בה השתמשנו ב-



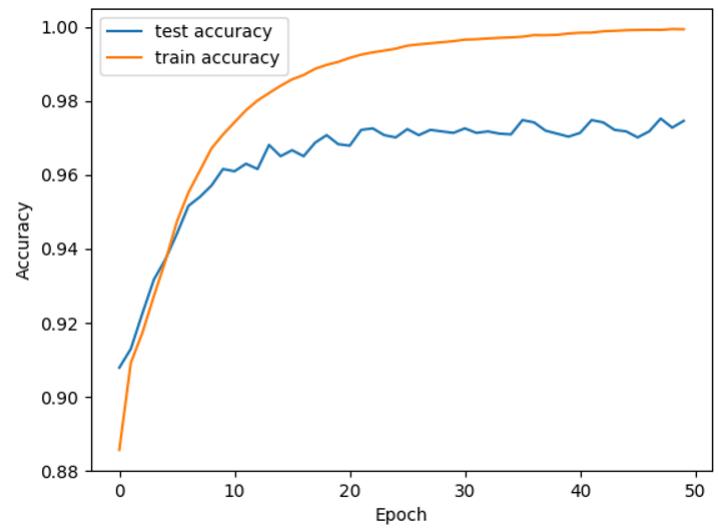
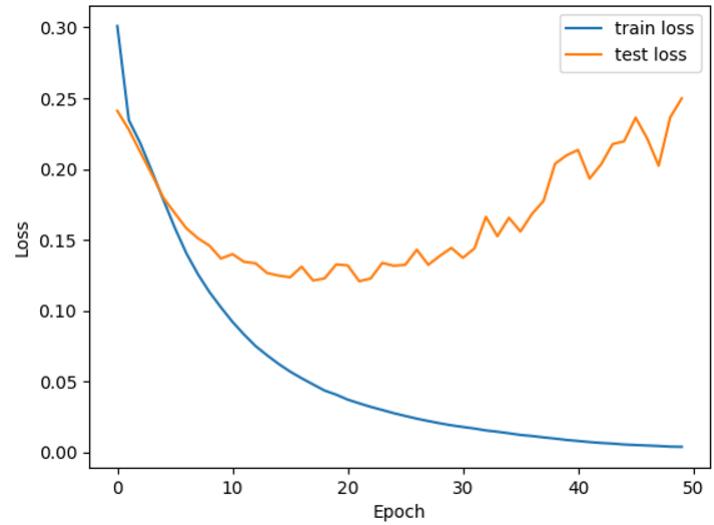
ניתן לראות ששינוי ה- $Batch - size$ הוביל לוצאות פחות מדויקות מאשר $Batch$ קטן יותר (24).

כעת ננסה את ההיפר פרמטר של גודל ה- $Batch$ כך ש- $Batch - Size = 128$, כאשר השתמשנו לאחר כל שכבה בפונקציית האקטיבציה
עבור 3 שכבות - $Linear(180, 80)$, $Linear(80, 24)$, $Linear(24, 1)$, למעט בשכבה האחרונה בה השתמשנו ב- $Sigmoid$, $ReLU$

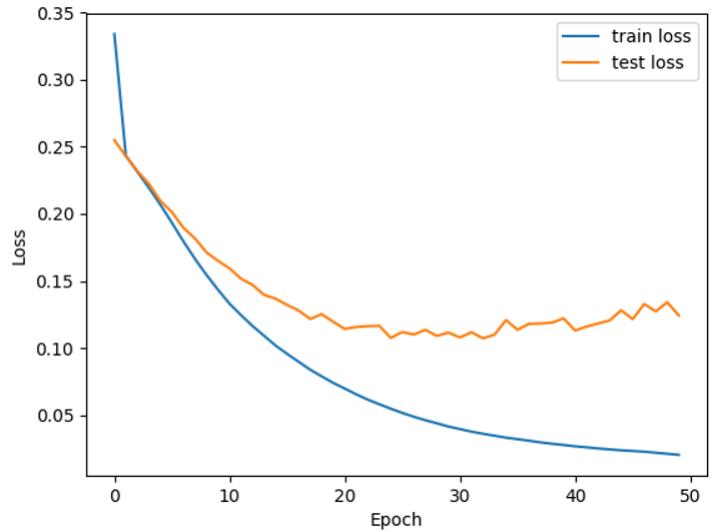


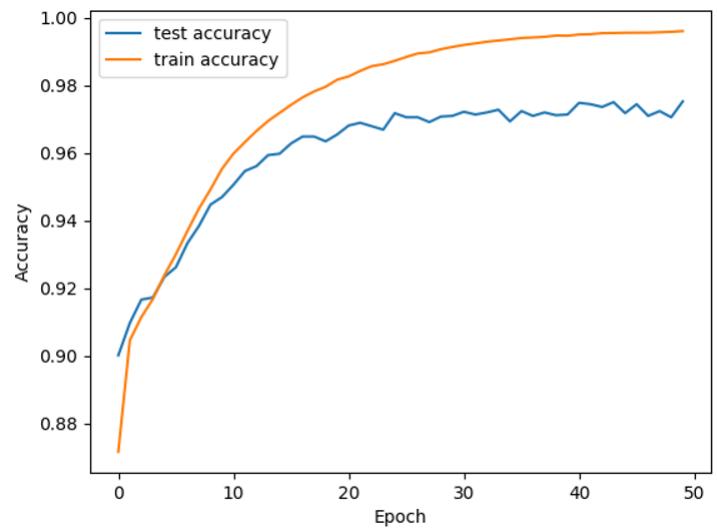
גם במקרה זה ניתן לראות שינוי ה-*Batch size* – *Batch* הוביל ל吒צאות פחות מדויקות מאשר קטן יותר.
חיפוש אחר רשת יציבה יותר תוך שימוש באופטימיזר *ADAMAX*:

רצינו לבחון באופן דומה למה שראינו לעילת את יכולות רשת דומה תוך שינוי האופטימיזר מ-*ADAM* ל-*ADAMAX*.
מדובר באופטימיזר אשר
להלן התוצאות, הפעם נזכיר בהסברים עבור כל תמונה כדי לא להכביר על הקורא, אם כי ניתן לראות שהמודלicut
יציב יותר מאשר המודל כאשר השתמשנו ב-*ADAM*. נציין שהגדרת המימדים בכל שכבה *Linear* דומה להגדרה שראינו
במודלים קודמים:
עומק הרשת = 3, $LR = 0.001$

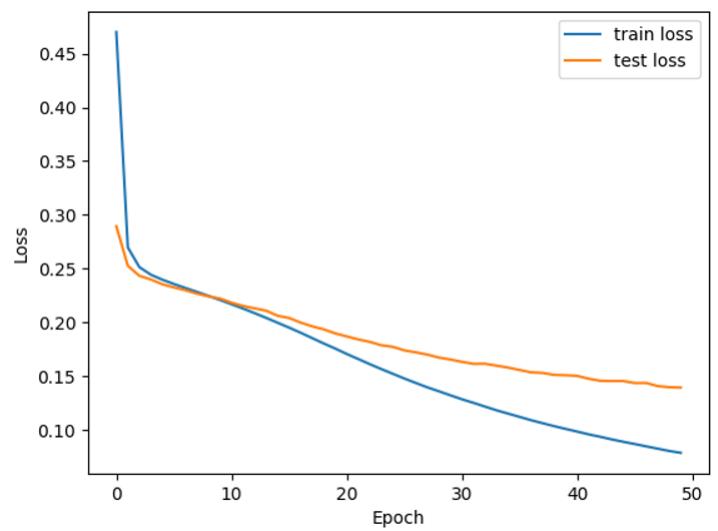


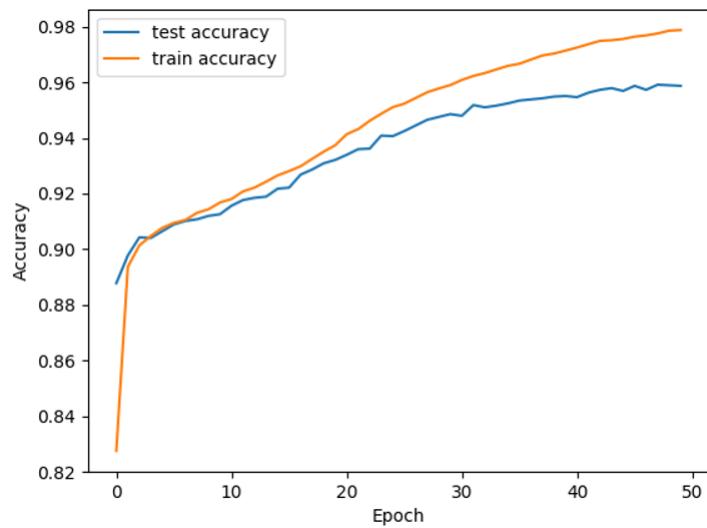
עומק הרשת = 3 , $LR = 0.0006$



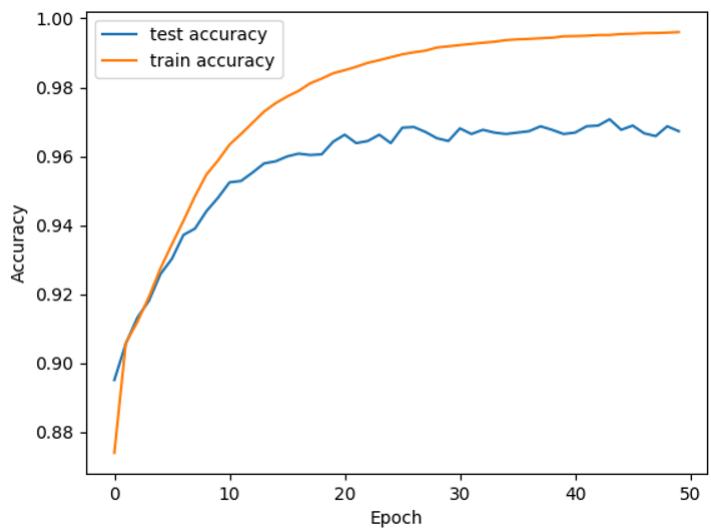
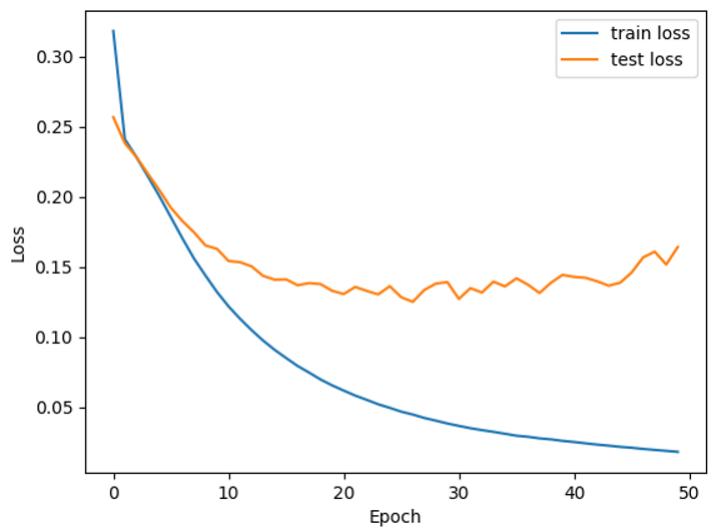


עומק הרשת 3 : $LR = 0.0002$

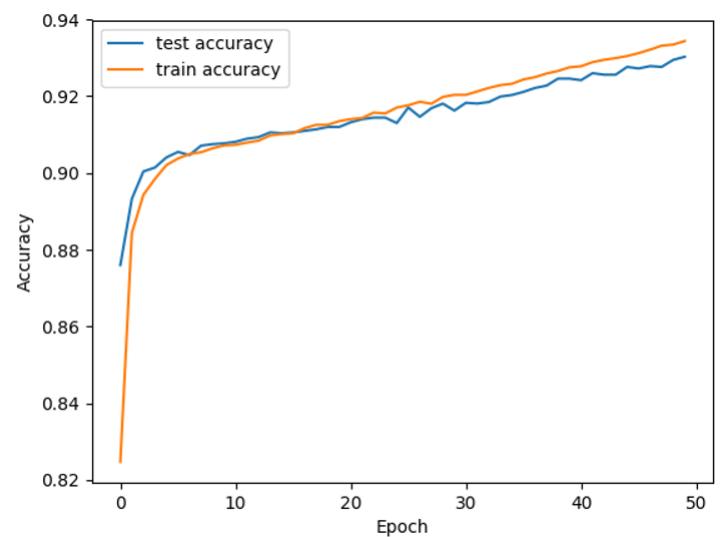
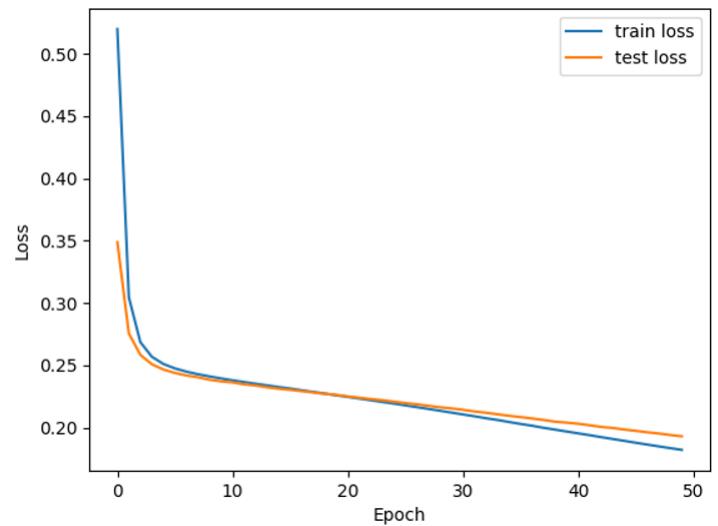




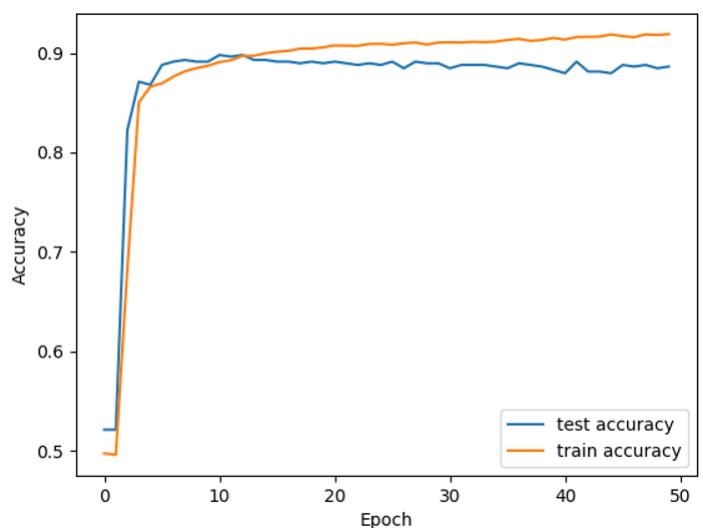
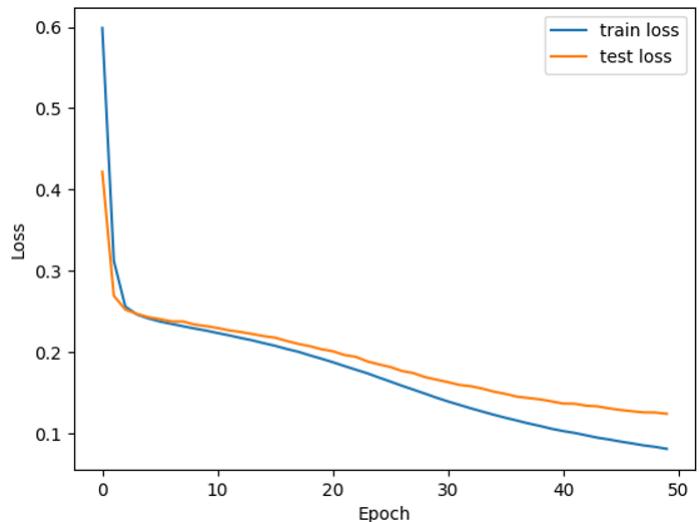
עומק הרשות $\beta = 0.0008$, $LR = 0.0008$



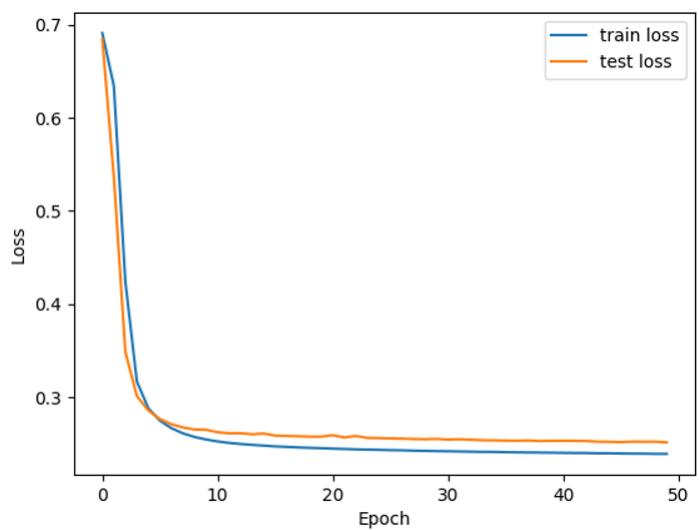
עומק הרשת = 2, $LR = 0.002$

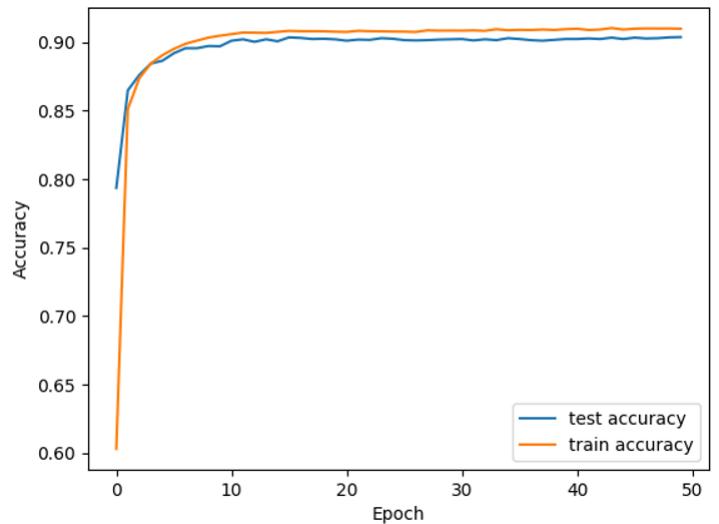


עומק הרשת = 4, $LR = 0.002$, תוק שימוש ביותר נוירונים (כלומר שינוי המימדים):

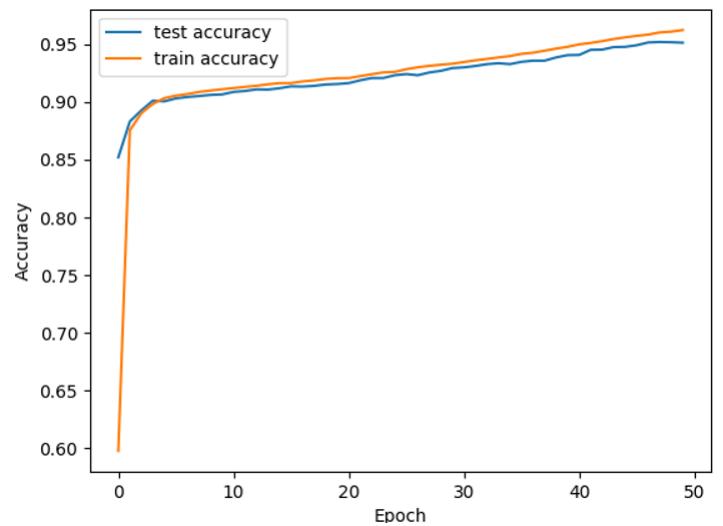
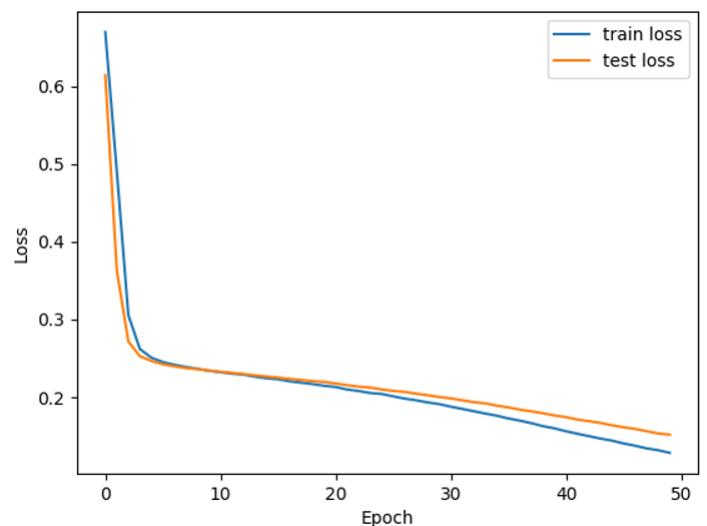


עומק הרשות = 4, $LR = 0.002$, כאשר כל האקטיבציות הן מסוג Sigmoid

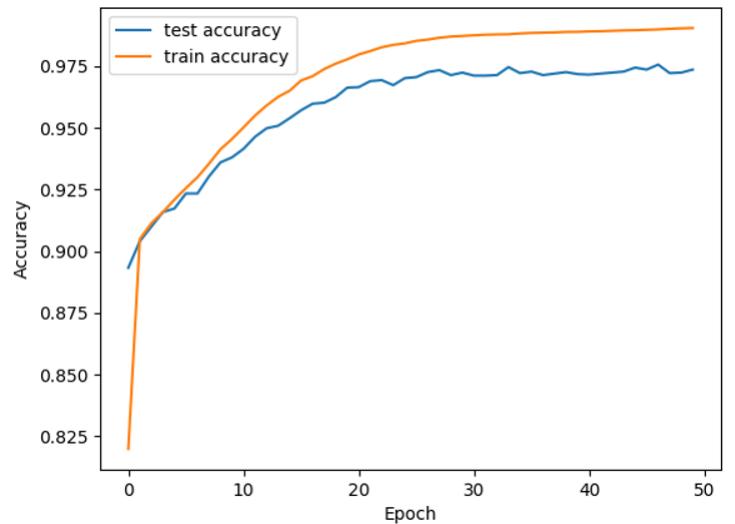
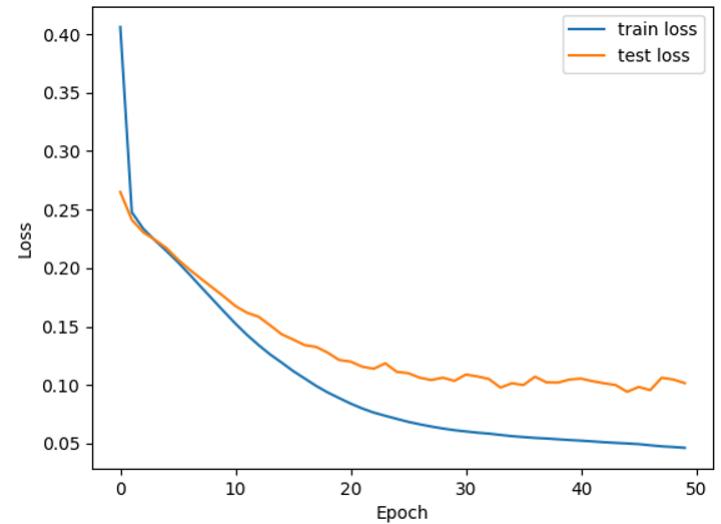




יוםק הרשות $batch-size = 128$ כאשר $LR = 0.002$, 4=



עומק הרשת = 5, $LR = 0.002$



נסכם בכך שהמודל האופטימלי (מתוך המודלים שבחנו) הוא:

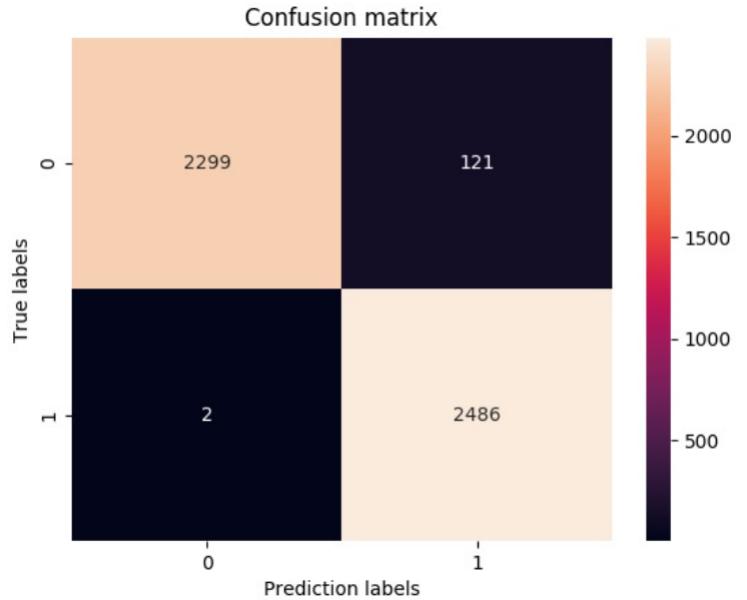
- .1 $Batch - Size = 24$

- .2 $Adamax$ with $LR = 0.0002$

- .3 ארבע שכבות - $Linear(180, 80), Linear(80, 24), Linear(24, 18), Linear(18, 1)$

- .4 *Oversampling*

רצינו לבדוק גם את ה- $Confusion - Matrix$ של תוצאות המבחן עבור המודל האופטימלי זהה:



ניתן לראות כי כמות ה- $FN = 2$, כלומר כמות זניחה (0.04%), ואילו ה- $FP = 121$, כמות קטנה ביחס לגודל הדadataה (2.4%). כמו כן, כמות החיוויים (בין אם חווינו אותם חיוביים או שליליים) הינה גבוהה (97.5% מכלל הדadataה).

שאלת המשך בחלק הפרקטי

במשימה האחורונה של החלק הפרקטי קיבלנו רצף 1,273 חומצות אמינו מחלבון מסווג *Spike*. ערכנו את מסמך הטקסט שהורדנו, כך שכל חומצות האמינו "עומדות" בשורה אחת. לאחר מכן עברנו על רצף זה באופן איטרטיבי על סמך 9 בסיסים, ותרגמנו את הרצף הזה לייצוג של *one-hot representation* כפי שעשינו בחלק הראשון של התרג'il. התוצאות שקיבלו (ההסתברויות לקבלת כל רצף, אשר ממיניות מהגובה לנמוכה) הן:

בדיקה זהה תוך שימוש באופטימיזר *ADAMAX* :
VIRGDEVRQ: 0.9997456669807434
ALLAGTITS: 0.9997407793998718
NLREFVFKN: 0.9996246099472046
KCTLKSFTV: 0.9990960359573364
LCPFGEVFN: 0.9988735318183899

שאלה 1 - מקרה לינארי

$$\begin{aligned}
 & xy \in F \quad \rightarrow \text{defn} \quad f, g : F \rightarrow F \\
 & \therefore f \circ g \quad \text{is defn} \quad f(g(x)) = f(g(x) + g(y)) \\
 & f \circ g(x+y) = f(g(x+y)) = f(g(x) + g(y)) \\
 & = f(g(x)) + f(g(y)) = f \circ g(x) + f \circ g(y) \\
 & f \circ g(cx) = f(g(cx)) = f(c \cdot g(x)) = c \cdot f(g(x)) \\
 & = c \cdot f \circ g(x)
 \end{aligned}$$

שאלה 1 - מקרה אפיני

נזכיר כי פונקציה אפינית היא מהצורה $f(x) = A \cdot x + b$:
 יהיו $f, g : \mathbb{F}^n \rightarrow \mathbb{F}^n$ העתקות אפיניות מעל אותו שדה. נראה ש- $g \circ f$ היא העתקה אפינית מ- \mathbb{F}^n ל- \mathbb{F}^n .

$$f^{\circ}g(x) = A_2(A_1 \cdot x + b_1) + b_2 = A_2A_1 \cdot x + A_2 \cdot b_1 + b_2$$

קיבלנו סכום של שני ביטויים: ביטוי התלי ב- x וביטוי נוסף, נוכל לסמן $b = A_2 \cdot b_1 + b_2$ ו- $A = A_1 A_2$ קלומר $Ax + b$ שווչורה האפינית, כנדרש.

שאלות סעיף 2

$$\theta^{n+1} = \theta^n - \alpha \nabla f_{\theta^n}(x)$$

אנו צריכים למצוא תנאי עצירה מספק לשאלת הנתונה:

נרצה שההפרש (בערך מוחלט) בין ערכי שני צעדים סמוכים יהיה קטן מס' מסוים. כלומר, $\epsilon > 0$, איזו:

$$|f(x^{n+1}) - f(x^n)| = |f(x^n - \alpha \nabla f_{\theta^n}(x^n)) - f(x^n)| < \epsilon$$

וזהו תנאי העצירה.

שאלה 2 סעיף ב

נשים לב ש- x היא נקודת סטציונრית, נניח בה”כ שמדובר בנקודת מינימום (המקרה של נקודת מקסימום הוא סימטרי), אזי מתקיים $f(x + dx) \geq f(x)$ לכל ”בשבביה קרובה” של x . לכן לפי פירוק טיילור מסדר שני:

$$f(x + dx) - f(x) = f(x) + \nabla f(x) \cdot dx + dx^T \cdot H(x) \cdot dx + \xi - f(x)$$

כאשר $\xi = O(||dx^3||)$ כיוון שפירוק טיילור הוא מסדר שני. natürlich ממנו בחישוב כיוון שמדובר בערך שניtinן לזמן מפאת גודלו.

לכן חישוב ההפרש, תוק ההנחה שמדובר בנקודת סטציונריה, גורר כי הגרדיינט מתאפס, כלומר:

$$= \nabla f(x) \cdot dx + dx^T \cdot H(x) \cdot dx + \xi = dx^T \cdot H(x) \cdot dx + \xi =$$

$$dx^T \cdot H(x) \cdot dx \geq 0$$

כאשר המעבר האחרון נובע מכך ש- $f(x + dx) \geq f(x)$ עבור נקודת מינימום כהנחה שהנחה. ראיינו בקורסי עבר שתנאי זה גורר כי הערכים העצמיים של מטריצת ההסיאן הם אי שליליים ולכן מטריצת ההסיאן אי שלילית: $H(x) \geq 0$. מכאן נסיק כי זהו התנאי לסיווג נקודת סטציונרית כמינימום (באופן שקול הוכחה סימטרית עבור הנחה כי מדובר בנקודת מקסימום).

שאלה 3

נניח שהרשת צריכה להזוז מעלות (בין 0 ל-360). נגדיר את פונקציית החפסד של החיזוי באופן הבא:

אלגוריתם 1 $(yPred, yTrue)$

1. נגדיר הפרש כ- $diff = torch.deg2rad(yTrue) - torch.deg2rad(yPred)$ (חישוב ההפרש ברדייאנים)
 2. נגדיר את $loss = 1 - torch.cos(diff)$
 3. נחזיר את $loss$
-

כאשר, ככל שהשגיאה תהיה קטנה יותר כך ה- $cos(diff)$ ישאף לאחד ולכן ה- $loss$ ישאף לאפס. באופן שמחזיר ערכים מצופה מהציגת הבעה.

שאלה 4 סעיף א

$$\frac{\partial}{\partial x} f(x+y, 2x, z)$$

לפי כלל השרשרת:

$$\frac{\partial}{\partial x} f(x+y, 2x, z) = \sum_i \left(\frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial x} \right) =$$

$$\frac{\partial f}{\partial x_1} + 2 \cdot \frac{\partial f}{\partial x_2} + 0 \cdot \frac{\partial f}{\partial x_3} = \frac{\partial f}{\partial(x+y)} + 2 \cdot \frac{\partial f}{\partial 2x}$$

שאלה 4 סעיף ב

$$f_1 \left(f_2 \left(\dots f_n(x) \right) \right)$$

נניח ש- $x \in R^m$ קלומר וקטור. לפי כלל השרשרת:

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f_1}{\partial f_2} \right)^T \left(\frac{\partial f_2}{\partial f_3} \right)^T \dots \left(\frac{\partial f_{n-1}}{\partial f_n} \right)^T \left(\frac{\partial f_n}{\partial x} \right)^T \nabla_x f_n$$

ראינו בכיתה שמדובר במכלה של מטריצות (יעקוביאן של הפונקציות). נשים לב שההרכבה של הפונקציות גוררת את המכפלה בין הנגזרות, מה שMOVEDIL לתאימות בכל הנוגע למידים של המכפלות ביונ הגזרות (ולכן המכפלה הנ"ל מוגדרת היטב).

שאלה 4 סעיף ג

$$f_1 \left(x, f_2 \left(x, f_3 \left(\dots f_{n-1} \left(x, f_n(x) \right) \right) \right) \right)$$

נשים לב כי מתקיים בשלב הראשון (קלומר עבור):

$$\frac{\partial f_{n-1}}{\partial x} = \frac{\partial f_{n-1}}{\partial x} + \frac{\partial f_{n-1}}{\partial f_n} \frac{\partial f_n}{\partial x}$$

נשים לב שכעת הפיתוח של ההרכבה כולל את כלל השרשרת על על כלל השרשרת:

$$\frac{\partial f_{n-2}}{\partial x} = \frac{\partial f_{n-2}}{\partial x} + \frac{\partial f_{n-2}}{\partial f_{n-1}} \frac{\partial f_{n-1}}{\partial x} = \frac{\partial f_{n-2}}{\partial x} + \frac{\partial f_{n-2}}{\partial x} \frac{\partial f_{n-2}}{\partial x} + \frac{\partial f_{n-2}}{\partial f_{n-1}} \frac{\partial f_{n-1}}{\partial f_n} \frac{\partial f_n}{\partial x}$$

ולכן:

$$\frac{\partial f_1}{\partial x} = \frac{\partial f_1}{\partial x} + \sum_{i=2}^n [\prod_{j=1}^i \frac{\partial f_{j-1}}{\partial f_j}] \frac{\partial f_i}{\partial x}$$

שאלה 4 סעיף ד

$$f(x + g(x + h(x)))$$

- נתו: $f(x + g(x + h(x)))$

- נפעיל את כלל השרשרת:

– נתחל בלהגדיר את הנגזר של $:h(x)$

$$\frac{\partial h}{\partial x} = t_1$$

– נמשיך עם הנגזרת של $:g(x + h(x))$

$$\frac{\partial g}{\partial x} = \frac{\partial g}{\partial(x + h(x))} \cdot \frac{\partial(x + h(x))}{\partial x} = \frac{\partial g}{\partial(x + h(x))} \cdot (1 + t_1)$$

– נסמן $\frac{\partial g}{\partial(x + h(x))}$ ונקבל:

$$\frac{\partial g}{\partial x} = d_2 \cdot (1 + t_1) = t_2 + t_2 \cdot t_1$$

– נמשיך לנגזרת הנדרשת:

$$\begin{aligned} \frac{\partial f}{\partial x} &= \frac{\partial f}{\partial(x + g(x + h(x)))} \cdot \frac{\partial(x + g(x + h(x)))}{\partial x} \\ &= \frac{\partial f}{\partial(x + g(x + h(x)))} \cdot (1 + t_2 + t_2 \cdot t_1) \end{aligned}$$

– נסמן $\frac{\partial f}{\partial(x+g(x+h(x)))} = t_3$ ונקבל את הנדרש:

$$\frac{\partial f}{\partial x} = t_3 \cdot (1 + t_2 + t_2 \cdot t_1) = t_3 + t_3 \cdot t_2 + t_3 \cdot t_2 \cdot t_1$$

שאלה 5

נוכיח ש $KL - divergence$ הוא אי שלילי:

$$D_{kl}(P||Q) = \sum_i p_i \cdot \log\left(\frac{p_i}{q_i}\right) = \sum_i p_i \cdot -\log\left(\left(\frac{p_i}{q_i}\right)^{-1}\right) = -\sum_i p_i \cdot \log\left(\frac{q_i}{p_i}\right) =$$

נזכיר שטענה שראינו בעבר – $-\log(x) \geq -x + 1$, כלומר $\log(x) \leq x - 1$, לכן:

$$\geq -\sum_i p_i \cdot \left(\frac{q_i}{p_i} - 1\right) = -\sum_i (q_i) + \sum_i (p_i)$$

נזכיר שהסכום של הסתברויות הוא 1 (המאורע השלם) לכן:

$$= -1 + 1 = 0$$

לסיכום קיבלנו :

$$D_{kl}(P||Q) \geq 0$$

ולכן מדובר בביטוי אי שלילי.

תזכורת-1: נזכיר באידישיווין ינסן:

אם f פונקציה ממשית קעורה המוגדרת על הקטע ו- x_1, \dots, x_n נמצאים בתחום הגדרתה ובנוסף נתון לכל $a_i \in \mathbb{R}^+$ נקבע $\sum a_i = 1$ אז:

$$f\left(\frac{\sum a_i x_i}{\sum a_i}\right) \geq \frac{\sum a_i f(x_i)}{\sum a_i}$$

בנוסף, מתקיים שיוויון אס"ם f ליניארי בתחום של $x_1 = x_2 = \dots = x_n$ או $x_1 \dots x_n$

תזכורת-2: פונקציית ה- \log הינה פונקציה ממשית קעורה.

$$\text{צ"ל: } P = Q \text{ אם ו } D_{kl}(P \| Q) = 0$$

הוכחה:

- \Leftarrow : נתון $P = Q$ (להשלים)

- \Rightarrow : נתון $D_{kl}(P \| Q) = 0$ ונרצה להראות ש- $P = Q$

- נניח כי $\forall i \in \mathbb{N} p_i > 0$

- מתקיים:

$$D_{kl}(P \| Q) = \sum_i p_i \cdot \log\left(\frac{q_i}{p_i}\right)$$

- במקרה פרטי, מאי-שוויון ינסן לפונקציות קעורות נקבל:

$$\leq \log\left(\sum_i p_i \cdot \frac{q_i}{p_i}\right) = \log\left(\sum_i q_i\right) = 0$$

- לאחר מכן ש- $D_{kl}(P \| Q) = 0$ וכי שוויון באיד-שוויון ינסן מתקיים אם ו- $\frac{q_1}{p_1} = \dots = \frac{q_n}{p_n}$ או ש-

ליניאריות בתחום זה.

- אם סימנו כי מתקיים השוויון (הסתברות אחידה). בנוסף, \log אינה פונקציה ליניארית $\frac{q_1}{p_1} = \dots = \frac{q_n}{p_n}$ – ומכאן תהיה ליניארית רק אם עבור כל $i \in \mathbb{N}$ מתקיים $\frac{q_1}{p_1} = \dots = \frac{q_n}{p_n}$ – כנדרש.

שאלה 6

צ"ל: $D_{kl}(P \| Q)$ קמורה.

הוכחה:

• מתקיים:

$$D_{kl}(P \| Q) = \sum_i p_i \cdot \log\left(\frac{q_i}{p_i}\right)$$

- **תזכורת:** מוכיחי לוגריתמים נקבל:

$$D_{kl}(P \parallel Q) \geq \left(\sum_i p_i \right) \cdot \log \left(\frac{\sum_i q_i}{\sum_i p_i} \right)$$

- לכן נוכל להראות ש:

$$D_{kl}(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2)$$

$$= \sum_i \left((\lambda p_{1,i} + (1-\lambda)p_{2,i}) \cdot \log \left(\frac{\lambda q_{1,i} + (1-\lambda)q_{2,i}}{\lambda p_{1,i} + (1-\lambda)p_{2,i}} \right) \right)$$

$$\sum_i \left(\lambda p_{1,i} \cdot \log \left(\frac{\lambda q_{1,i} + (1-\lambda)q_{2,i}}{\lambda p_{1,i} + (1-\lambda)p_{2,i}} \right) + (1-\lambda)p_{2,i} \cdot \log \left(\frac{\lambda q_{1,i} + (1-\lambda)q_{2,i}}{\lambda p_{1,i} + (1-\lambda)p_{2,i}} \right) \right)$$

- נעזר בחוקי הלוגריתמים שהזכירנו קודם לכן ונקבל:

$$\leq \sum_i \left(\lambda p_{1,i} \cdot \log \left(\frac{\lambda q_{1,i}}{\lambda p_{1,i}} \right) + (1-\lambda)p_{2,i} \cdot \log \left(\frac{(1-\lambda)q_{2,i}}{(1-\lambda)p_{2,i}} \right) \right)$$

$$= \lambda \cdot \sum_i \left(p_{1,i} \cdot \log \left(\frac{\lambda q_{1,i}}{\lambda p_{1,i}} \right) \right) + (1-\lambda) \cdot \sum_i \left(p_{2,i} \cdot \log \left(\frac{(1-\lambda)q_{2,i}}{(1-\lambda)p_{2,i}} \right) \right)$$

$$= \lambda \cdot D_{kl}(p_1 \parallel q_1) + (1-\lambda) \cdot D_{kl}(p_2 \parallel q_2)$$

- כנדרש.

שאלה 7

משפט צ'בנקו גורס כי צירוף לינארי של “פונקציות סיגמוואידיות”, כולם אשר שייכות לקבוצה $(C[0, 1])$. משפט הורניק מרחיב את המשפט צ'בנקו עבור פונקציות מונוטוניות חסומות ולא קבועות (על פי תיקון של רענן בהרצאה 2). לכן נגדיר $\sigma(x) = \text{ReLU}(x) - \text{ReLU}(x - c)$, כאשר $c = \text{constant}$. כתוב ניוכח כי:

$$\sigma(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq c \\ c & x > c \end{cases}$$

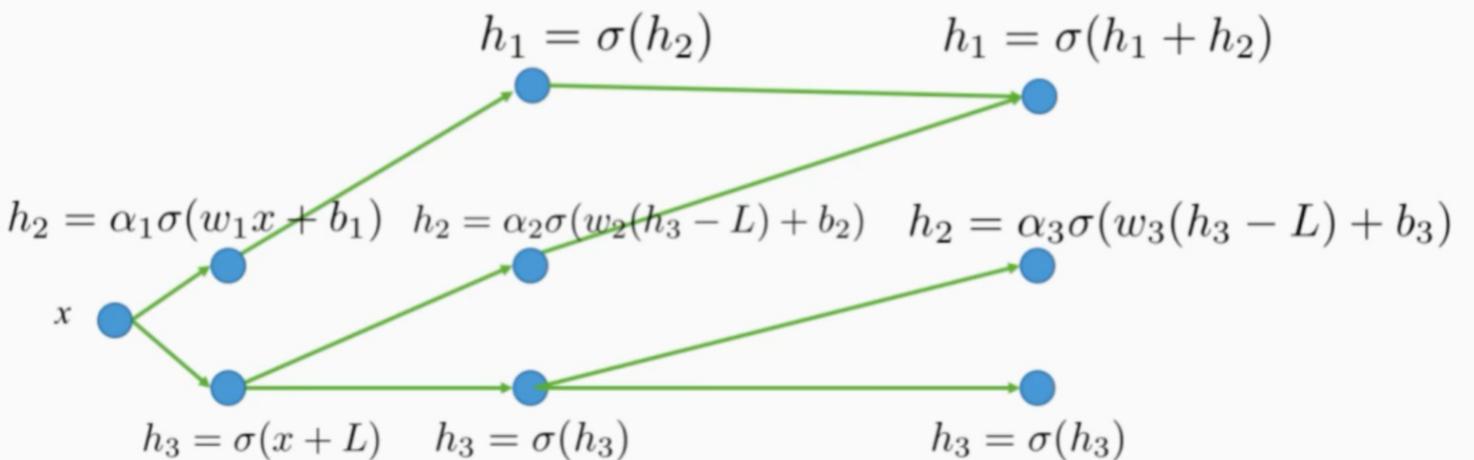
כולם $\sigma(x)$ היא פונקציה רציפה, חסומה ומונוטונית לא יורדת, לכן לפי משפט הורניק, שימוש בצירופים לינאריים של $\sigma(x)$ נוכל לקבל כל פונקציה אשר שייכת $C[0, 1]$. את הפונקציה בינוי מצירוף לינארי של ReLU ולכן ניתן לקרב כל פונקציה ב- $C[0, 1]$ בעזרת צירופים לינאריים אשר הם ReLU וגם ReLU של *dialeted translated*, כנדרש.

שאלה 8

נזכיר שראינו בכיתה:

$$f(x) = \sum_{i=1}^n \alpha_i \sigma(w_i x + b_i), \quad \alpha_i > 0$$

simplifying assumption



נשים לב כי:

h_3 “דווג” להרחבה ולהזזה
 h_2 “דווג” לסכימה

כמו כן נבחים כי מכיוון $\alpha_i > 0$ אז מתקיים כי $\text{ReLU}(\alpha_i \cdot x) = \alpha_i \cdot \text{ReLU}(x)$ כי $\text{ReLU}(\alpha_i \cdot x) = \alpha_i \cdot \text{ReLU}(x)$ עבור $\alpha_i \leq 0$.

בשאלה זו לא נוכל להניח כי $\alpha_i > 0$. כלומר, נרצה להראות מודל שיעבוד גם עבור $\alpha_i \leq 0$:

עבור $\alpha_i = 0$ קיבל $h_2 = 0$ ולכן המודל שהוצע לנו בשיעור מכסה את המקרה הנ"ל.
 עבור $0 < \alpha_i < \infty$ מתקיים כי $ReLU(-\alpha_i \cdot x) = -\alpha_i \cdot ReLU(x)$ ולכן נוכל להציג את המודל הבא:
 בהרצאה רأינו כי מכל קודקוד של h_2 נוכל "למתו" 2 חיצים במקומות ח' אחד (לדוגמא h_1 כמו שראויים בגרף כז' ש- $\hat{h}_1 = \sigma(-h_2)$ וגם $h_1 = \sigma(h_2)$)
 לכן נוכל "למתו" ח' נוסף אל הקודקוד הבא בתור. וכך נוכל לקבל את הסכימה של האיברים הקודמים בגרף:

