הטכניון-מכון טכנולוגי לישראל **אביב תשפ"א**
הפקולטה להנדסת חשמל ומחשבים **18.10.2021**

למידה עמוקה - 046211
בחינה סופית - מועד ב'

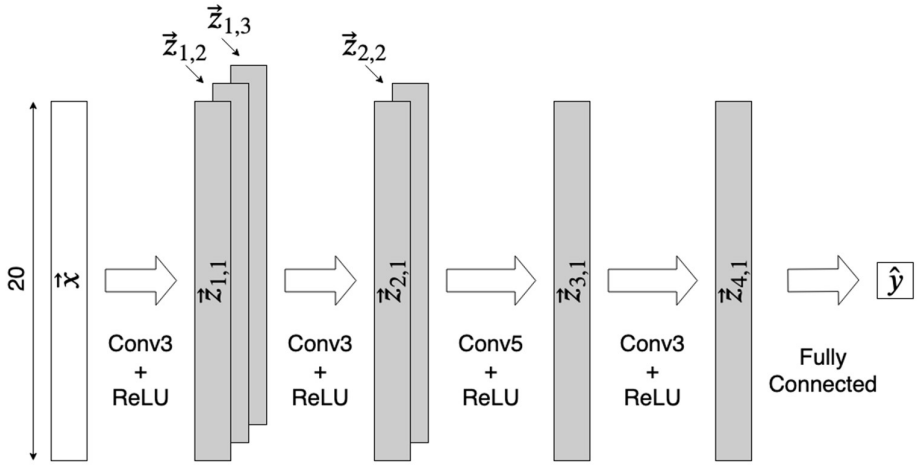
משך המבחן: *3* שעות

1. ניתן להשתמש בדפי נוסחאות ומחשבון.
2. יש לענות על כל השאלות.
3. משקל כל שאלה מצוין בטופס הבחינה. סך הנקודות הוא 100.
4. נא לכתוב בצורה ברורה ומסודרת.
5. יש לכלול פירוט מלא של דרך הפתרון. תשובות לא מנומקות לא יזכו בניקוד, אם לא נאמר אחרת.

ב ה צ ל ח ה !!!

שאלה מס׳ 1: קונבולוציה (35%)

שאלה זו עוסקת ברשת הקונבולוציה **החד-מימדית** הבאה:



הכניסה לרשת (שכבת הקלט) הינה וקטור באורך 20, והמוצא הוא סקלר. הרשת בנויה מהשכבות הבאות:

- שכבת קונבולוציה עם גרעין באורך 3 ושלושה ערוצים במוצא
- שכבת אקטיבציה מסוג ReLU
- שכבת קונבולוציה עם גרעין באורך 3 ושני ערוצים במוצא
- שכבת אקטיבציה מסוג ReLU
- שכבת קונבולוציה עם גרעין באורך 5 וערוץ יחיד במוצא
- שכבת אקטיבציה מסוג ReLU
- שכבת קונבולוציה עם גרעין באורך 3 וערוץ יחיד במוצא
- שכבת אקטיבציה מסוג ReLU
- שכבת Fully Connected

בנוסף:

- לשם הפשטות אין ברשת זו איברי היסט (bias).
- הקונובולוציות נעשות **ללא ריפוד (padding)**
- שאר הפרמטרים של השכבות הינם סטנדרטיים: stride=1, dilation=1.

3% א. מהו האורך של הוקטורים $\vec{z}_{1,i}$, $\vec{z}_{2,i}$, $\vec{z}_{3,1}$ ו $\vec{z}_{4,1}$? נסמן אורכים אלה כ-d_1, d_2 d_3 ו d_4.

בשני הסעיפים הבאים מומלץ תחילה לבטא את התשובה באמצעות באמצעות d_1, d_2 d_3 ו d_4.

4% ב. מהי כמות הפרמטרים ברשת?

6% ג. מהי כמות המכפלות (Multiplications) המתבצעות ברשת ב- forward pass יחיד?

המשך בעמוד הבא

8% ד. מהו ה- receptive field של $z_{2,1,5}$ (האיבר החמישי בוקטור $\vec{z}_{2,1}$)?

רשמו את תשובתכם כ- $\{x_?, x_?, \dots, x_?\}$.

<u>תזכורת</u>: ה- receptive field של משתנה ברשת, הוא אוסף האיברים בכניסה אשר משפיע עליו.

הערה: שימו לב שבשכבות הקונבולוציה אין ריפוד (padding).

בעבור מדגם נתון $\{\vec{x}_i, y_i\}$, נרצה לאמן את הרשת לבצע רגרסיה ולמזער את שגיאת ה- MSE (פונקציית סיכון עם l_2 loss).

בנוסף נרצה להשתמש ברגולריציית l_2 עם פרמטר λ על פרמטרי הרשת $\vec{w}$.

5% ה. נרצה למצוא את המקדמים האופטימליים של הרשת בהינתן המדגם.

רשמו את בעיית האופטימיזציה המתאימה. השתמשו ב- $g(\vec{x}; \vec{w})$ על מנת לייצג את הפונקציה שאותה מממשת הרשת, כאשר $\vec{w}$ ההוא הווקטור של כל הפרמטרים ברשת.

על מנת למצוא את הפרמטרים של הרשת נשתמש ב- Stochastic Gradient Descent (GD בגרסה הסדרתית עם דגימה אחת אחת בכל צעד עדכון) עם גודל צעד α.

נסמן את המשקלים בשכבת ה fully connected ב- $w_{5,1}, \dots, w_{5,d_4}$ (כאשר $w_{5,i}$ היא המשקולת שמכפילה את $z_{4,1,i}$)

8% ו. בצעד ה- t נבצע עדכון על פי הדגימה $\vec{x}_k, y_k$. רשמו את העדכון של $w_{5,5}$ בצעד זה.

בטאו את תשובתכם בעזרת α, λ, $x_{k,m}$, y_k, $z_{m,n,l}$ ו- $\hat{y} = g\left(\vec{x}_k; \vec{w}^{(t)}\right)$ (מוצא הרשת בעבור הכניסה $\vec{x}_k$ וסט פרמטרים $\vec{w}^{(t)}$), כאשר m, n ו l הם אינדקסים כלשהם.

שאלה מס׳ 2: עצירה מוקדמת בנוירון בודד (35%)

נבחן את ההשפעה של עצירה מוקדמת בנוירון בודד עם כניסה חד מימדית, אקטיבציה לינארית, ופונקציית מחיר ריבועית כלומר, רגרסיה לינארית ללא רגולריזציה:

$$\mathcal{L}(w)=\frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)}-wx^{(n)}\right)^2$$

נרצה למזער את השגיאה $\mathcal{L}(w)$ באמצעות GD מאתחול $w(0)=0$ וגודל צעד η. נניח וכי הדוגמאות מיוצרות ממודל לינארי $y^{(n)}=w_*x^{(n)}+\varepsilon^{(n)}$ כאשר w_* קבוע ו- $\varepsilon^{(n)}$ רעש גאוסי i.i.d. עם שונות σ^2.

א. הראו כי מתקיים

$$w(t)=\left(\frac{h_1}{h_0}+w_*\right)\left(1-e^{-rt}\right)$$

כאשר $h_0=\sum_{n=1}^{|N}\left(x^{(n)}\right)^2$, $h_1=\sum_{n=1}^{N}x^{(n)}\varepsilon^{(n)}$, ו- $r=-\ln\left(1-\eta h_0\right)$. הניחו כי $h_0>0$.

ב. הראו כי מתקיים $E\left[\left(w(t)-w_*\right)\right]^2=\left(Ew(t)-w_*\right)^2+E\left[\left(w(t)-Ew(t)\right)\right]^2$. איך קוראים לפירוק זה?

ג. הראו כי השגיאה הריבועית הממוצעת (על פני הרעש $\varepsilon^{(n)}$) היא

$$E\left(w(t)-w_*\right)^2=\exp\left(-2rt\right)w_*^2+\left(1-\exp\left(-rt\right)\right)^2\sigma^2h_0^{-1}$$

ד. מצאו את זמן העצירה האופטימלי t^* הממזער את השגיאה הריבועית הממוצעת.

ה. יחס האות לרעש בבעיה זו הוא $\mathrm{SNR}=w_*^2\sigma^{-2}$. כיצד זמן העצירה האופטימלי משתנה עם ה- SNR? האם זה דומה למה שלמדנו ברגולרזציית L2?

שאלה מס' 3: אופטימיזציה (30%)

הפונקציה $f:\mathbb{R}^d \to \mathbb{R}$ גזירה ברציפות אינסוף פעמים, ומקיימת $\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) = f_* > -\infty$. אנו רוצים למזער פונקציה זו באמצעות גרסא של gradient descent (GD) עם גודל צעד η, בו בכל איטרציה הגרדיאנטים מוכפלים במטריצה $\mathbf{A}$

$$(*)\ \mathbf{w}(t+1) = \mathbf{w}(t) - \eta\mathbf{A}\nabla f(\mathbf{w}(t))$$

נתון כי $\mathbf{A}$ סימטרית וחיובית ממש, כלומר $\lambda_{\min} \triangleq \lambda_{\min}(\mathbf{A}) > 0$, ונסמן גם $\lambda_{\max} \triangleq \lambda_{\max}(\mathbf{A})$.

א. בסעיף זה בלבד הניחו כי $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w}$ עם $\mathbf{H}$ חיובית ממש. איזה $\mathbf{A}$ ו- η, נבחר כדי לקבל התכנסות במספר מינימלי של צעדים באלגוריתם (*)? מדוע בחירה זו לא סבירה ב- d גדול? איזה קירוב מקובל ניתן לעשות?

ב. הוכיחו כי gradient flow (כלומר GD בגבול של $\eta \to 0$),

$$\dot{\mathbf{w}}(t) = -\mathbf{A}\nabla f(\mathbf{w}(t))$$

מתכנס לנקודה קריטית לכל f ו- $\mathbf{A}$ המקיימות את תנאי השאלה.

<u>רמז</u>: מתכונות הערכים העצמיים מתקיים $\forall v \in \mathbb{R}^d : \lambda_{\min}\|\mathbf{v}\|^2 \le \mathbf{v}^T\mathbf{A}\mathbf{v} \le \lambda_{\max}\|\mathbf{v}\|^2$

ג. בהנחה שהפונקציה f היא $\beta-\text{smooth}$, מצאו תנאי על גודל הצעד כך שנתכנס בהכרח לנקודה קריטית באלגוריתם (*). הוכיחו התכנסות בהינתן תנאי זה.

<u>רמז</u>: כזכור, עבור פונקציות $\beta-\text{smooth}$ ניתן לרשום

$$f(\mathbf{w}(t+1)) - f(\mathbf{w}(t)) \le (\mathbf{w}(t+1)-\mathbf{w}(t))^T \nabla f(\mathbf{w}(t)) + \frac{\beta}{2}\|\mathbf{w}(t+1)-\mathbf{w}(t)\|^2$$