הטכניון-מכון טכנולוגי לישראל **חורף תשפ"ד**

הפקולטה להנדסת חשמל ומחשבים **18.9.2024**

למידה עמוקה - 046211

בחינה סופית - מועד א'

משך המבחן: ***3 שעות***

1. ניתן להשתמש ב-2 דפי נוסחאות (4 עמודים) 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.
6. אם קיבלת אישור מיוחד (עקב שירות מילואים), ניתן לבחור לענות רק על 2 שאלות מתוך 3. במקרה זה, הניקוד הכולל ינורמל חזרה ל-100.
7. כשסיימת את הבחינה, נא להעלות ל-Gradescope את הבחינה. בנוסף, יש להגיש את מחברת הבחינה (עליה כתוב הפתרון).

ב ה צ ל ח ה !!!

שאלה מס' 1 – אתחול רשתות (30%)

נתונה רשת עם חיבורי קפיצה (skip connections):

$$\boldsymbol{u}_{l+1} = \boldsymbol{u}_l + \boldsymbol{F}_l(\boldsymbol{u}_l)$$

במקרה שבו $\boldsymbol{F}_l(\boldsymbol{u}_l) = \boldsymbol{W}_l'\varphi(\boldsymbol{W}_l\boldsymbol{u}_l)$, $\boldsymbol{W}_l$ ו-$\boldsymbol{W}_l'$ מטריצת משקולות בגודל $d \times d$, ו-φ פונקציית אקטיבציה מסוג ReLU. הכניסה לרשת היא $\boldsymbol{u}_0 = \boldsymbol{x}$, עבורה כל רכיב בעל ממוצע אפס ושונות 1.

א. כזכור, שיטות האתחול הסטנדרטית של $\boldsymbol{W}_l$ ו-$\boldsymbol{W}_l'$ של רשת ללא חיבורי קפיצה צריכה לקיים

(1) $E(\boldsymbol{F}_l(\boldsymbol{u}_l)\,|\boldsymbol{u}_l) = 0$ וגם (2) $E\mathrm{Var}(\boldsymbol{F}_l(\boldsymbol{u}_l)\,|\boldsymbol{u}_l) = \mathrm{Var}(\boldsymbol{u}_l)$

איזה אתחולים של $\boldsymbol{W}_l$ ו-$\boldsymbol{W}_l'$ יקיימו זאת?

ב. הראו כי השונות $\boldsymbol{u}_l$ מתפוצצת ככל שמתקדמים ברשת, כלומר $\lim_{l\to\infty}\mathrm{Var}(\boldsymbol{u}_l) = \infty$ (השונות כאן מופעלת רכיב-רכיב), עבור האתחול הסטנדרטי מסעיף קודם, שמקיים את משוואות 1 ו-2.

ג. כדי למנוע את ההתפוצצות מוצע לשנות את האתחול הסטנדרטי בהתאם לשכבה בצורה הבאה

(3) $E(\boldsymbol{F}_l(\boldsymbol{u}_l)\,|\boldsymbol{u}_l) = 0$ וגם (4) $E\mathrm{Var}(\boldsymbol{F}_l(\boldsymbol{u}_l)\,|\boldsymbol{u}_l) = c_l\mathrm{Var}(\boldsymbol{u}_l)$

כאשר $c_l > 0$. כיצד נשנה את אתחול של $\boldsymbol{W}_l$ ו-$\boldsymbol{W}_l'$ לקבל זאת? בנוסף, הציעו בחירה של $\{c_l\}_{l=1}^{\infty}$ חיוביים המונעת התפוצצות של השונות, כלומר $\lim_{l\to\infty}\mathrm{Var}(\boldsymbol{u}_l) < \infty$.

ד. מהו הגבול $\lim_{l\to\infty}\mathrm{Var}(\boldsymbol{u}_l)$ עבור אתחול אפס $\boldsymbol{W}_l' = \boldsymbol{W}_l = \boldsymbol{0}$? מהי הבעיה העיקרית עם אתחול זה?

<u>הנחיה:</u> עבור האתחול הנתון חשבו בצורה מפורשת את הגרדיאנט של פונקציית ההפסד הכללית ℓ על מוצא הרשת ביחס למשקולות $\boldsymbol{W}_l$ ו-$\boldsymbol{W}_l'$.

ה. איזה פתרון נפוץ ברשתות סטנדרטיות עם חיבורי קפיצה כדי למנוע את התפוצצות השונות $\mathrm{Var}(\boldsymbol{u}_l)$? ציינו יתרון וחסרון של פתרון שכזה.

שאלה מס' 2 – גזירה (35%)

כדי למדל תהליכים בזמן רציף נשתמש ברשת נוירונים מסוג Neural ODE, המתארת מד"ר בזמן רציף

$$\forall t \in [0,T]: \quad \frac{d\mathbf{v}(t)}{dt} = \mathbf{f}(\mathbf{w}, \mathbf{v}(t), t) \qquad (*)$$

עם פרמטרים $\mathbf{w} \in \mathbb{R}^p$, מצב $\mathbf{v}(t) \in \mathbb{R}^d$, תנאי התחלה $\mathbf{v}(0) = \mathbf{v}_0$, פונקציה וקטורית $\mathbf{f}$, ופונקציית מחיר $C = \frac{1}{2}\|\mathbf{v}(T)\|^2$. שימו לב ש-$\mathbf{v}(t)$ היא פונקציה של $\mathbf{w}$ (מלבד ב-$t = 0$), ולכן גם $C = C(\mathbf{w})$. נרצה למצוא מערכת של משוואות דיפרנציאליות שפתרונן יהיה $\nabla C(\mathbf{w})$, כך שנוכל למצוא את $\mathbf{w}$ באמצעות אלגוריתם הגרדיאנט (GD). נבצע זאת בשתי השיטות שלמדנו: forward-mode ו-reverse-mode.

א. תחילה, נסמן $\boldsymbol{a}(t) \triangleq \frac{d\mathbf{v}(t)}{d\mathbf{w}}$. מצאו מד"ר עבור $\boldsymbol{a}(t)$ מהצורה

$$, \ \forall t \in [0,T]: \ \frac{d\boldsymbol{a}(t)}{dt} = \mathbf{g}(\mathbf{w}, \mathbf{v}(t), \boldsymbol{a}(t), t)$$

כאשר $\mathbf{g}$ יכולה להיות תלויה ב-$\mathbf{f}$ ונגזרותיה החלקיות $\frac{\partial \mathbf{f}(\mathbf{w},\mathbf{v}(t),t)}{\partial \mathbf{w}}$ ו-$\frac{\partial \mathbf{f}(\mathbf{w},\mathbf{v}(t),t)}{\partial \mathbf{v}(t)}$.

<u>רמז:</u> נגזרות מתחלפות וכלל השרשרת. שימו לב להבדל בין נגזרת מלאה וחלקית.

ב. מה יהיה תנאי ההתחלה של המד"ר ב-$t = 0$? כיצד נחשב את $\nabla C(\mathbf{w})$ באמצעות פתרון המד"ר $(*)$ והמד"ר מסעיף א'? <u>רמז:</u> כלל השרשרת.

ג. כעת, נפתח שיטה אחרת. תחילה, נכתוב את הלגרנז'יאן $\mathcal{L}(\mathbf{w}, \mathbf{v}, \mathbf{b}, \boldsymbol{\lambda})$ של הבעיה

$$\frac{1}{2}\|\mathbf{v}(T)\|^2 + \int_0^T \mathbf{b}(t)^\top \left(\mathbf{f}(\mathbf{w}, \mathbf{v}(t), t) - \frac{d\mathbf{v}(t)}{dt}\right) dt + \boldsymbol{\lambda}^\top \big(\mathbf{v}_0 - \mathbf{v}(0)\big)$$

הסבירו מאיפה מגיע כל אחד מהאיברים בלגרנז'יאן.

ד. הראו כי ניתן להפוך את האינטגרל בלגרנז'יאן לביטוי הבא ע"י אינטגרציה בחלקים

$$\int_0^T \left(\mathbf{b}(t)^\top \mathbf{f}(\mathbf{w}, \mathbf{v}(t), t) + \frac{d\mathbf{b}(t)^\top}{dt}\mathbf{v}(t)\right) dt - \mathbf{b}(T)^\top \mathbf{v}(T) + \mathbf{b}(0)^\top \mathbf{v}(0)$$

ה. מתנאי הסטציונריות של הלגרנז'יאן $\forall t \in (0,T): \frac{\partial \mathcal{L}}{\partial \mathbf{v}(t)} = 0$, מצאו מד"ר עבור $\boldsymbol{b}(t)$ מהצורה

$$, \ \forall t \in (0,T): \ \frac{d\boldsymbol{b}(t)}{dt} = \mathbf{h}(\mathbf{w}, \mathbf{v}(t), \boldsymbol{b}(t), t)$$

כאשר $\mathbf{h}$ יכולה להיות תלויה ב-$\mathbf{f}$ ונגזרותיה החלקיות $\frac{\partial \mathbf{f}(\mathbf{w},\mathbf{v}(t),t)}{\partial \mathbf{w}}$ ו-$\frac{\partial \mathbf{f}(\mathbf{w},\mathbf{v}(t),t)}{\partial \mathbf{v}(t)}$.

ו. מתנאי הסטציונריות $\frac{\partial \mathcal{L}}{\partial \mathbf{v}(T)} = 0$ מצאו את תנאי ההתחלה של המד"ר מסעיף קודם ב-$t = T$ (הניחו כי המד"ר מתקיימת לכל $t \in [0,T]$)?

ז. כיצד נחשב את $\nabla C(\mathbf{w})$ באמצעות פתרון המד"ר $(*)$ והמד"ר מסעיף ה'? אין צורך לנמק.

<u>רמז:</u> איזה תנאי סטציונריות נשאר?

ח. מבין שתי השיטות לעיל (השיטה בסעיפים א'-ב' מול השיטה בסעיפים ג'-ז'), איזו שיטה היא forward-mode ואיזה reverse-mode?, ציינו חסרון של כל שיטה לעומת השיטה האחרת.

שאלה מס' 3 – רשתות עם משוב (35%)

נתון ש-$\{\mathbf{x}_t\}_{t=1}^{T}$ הם סדרה (sequence) של וקטורים (tokens) המקודדים מילים בטקסט בייצוג one-hot עם מילון בגודל d. נאמן רשת עם משוב על אוסף של N סדרות $\{\mathbf{x}_t\}_{t=1}^{T}$, כאשר כל סדרה מייצגת טקסט אחר.

$$(*)\ \ \forall t = 1, \dots, T-1:\ \ \mathbf{v}_t = \varphi(\mathbf{W}\mathbf{v}_{t-1} + \mathbf{B}\mathbf{x}_t)\ \ ;\ \ \hat{\mathbf{y}}_t = \mathbf{C}\mathbf{v}_t$$

כאשר $\mathbf{v}_t \in \mathbb{R}^k$ וקטור מצב, $\Theta = \{\mathbf{W}, \mathbf{B}, \mathbf{C}\}$ מטריצות פרמטרים, φ פונקציית אקטיביציה ו-$\hat{\mathbf{y}}_t$ חזאי.

תחילה נבחן את הבעיה של שערוך אוטו-רגרסיבי, בו בכל צעד אנו חוזים את המילה הבאה $\hat{\mathbf{y}}_t = \hat{\mathbf{x}}_{t+1}$.

א. נרצה לאמן את הרשת ע"י מזעור פונקציית מחיר שתעודד חיזוי נכון על סדרה אחת ($N = 1$). מהי פונקציית המחיר הסטנדרטית לכך? רשמו את שגיאת האימון על הסדרה כפונקציה של $\{\mathbf{x}_t, \hat{\mathbf{x}}_t\}_{t=1}^{T}$.

ב. הסבירו מהו ההבדל בחישוב פונקציית המחיר בזמן האימון כאשר משתמשים ב-teacher forcing לעומת כאשר אין teacher forcing?

ג. מה הבעיה ש-teacher forcing מנסה לפתור? מתי יותר חשוב להשתמש ב-teacher forcing : בתחילת האימון או בסופו? כיצד נחשב את החיזוי בזמן ההסקה, כלומר לאחר שתהליך האימון נגמר?

ד. הגדירו מתמטית מודל RNN רב-שכבתי בעל L שכבות המכליל את $(*)$ <u>כפי שראינו בהרצאה</u>. בשביל מה מודל כזה טוב? כיצד נבצע חישוב ע"י wavefront במודל שכזה? למה זה טוב?

כעת נרצה לחזות מילים חסרות במשפט באופן כללי (לא רק המילה הבאה), באמצעות bi-directional RNNs. ברשת זו, על כל סדרה, אנחנו מריצים את הרשת על הקלט בשני כיוונים מנוגדים: בכיוון הרגיל

$$\forall t = 1, \dots, T-1:\ \ \mathbf{v}_t = \varphi(\mathbf{W}\mathbf{v}_{t-1} + \mathbf{B}\mathbf{x}_t)$$

ובכיוון ההפוך:

$$\forall t = T-1, \dots, 1:\ \ \bar{\mathbf{v}}_t = \varphi(\bar{\mathbf{W}}\bar{\mathbf{v}}_{t+1} + \bar{\mathbf{B}}\bar{\mathbf{x}}_t)$$

כאשר המוצא הוא שילוב משני הכיוונים: $\hat{\mathbf{y}}_t = \mathbf{C}\mathbf{v}_t + \bar{\mathbf{C}}\bar{\mathbf{v}}_t$

ה. מתי לא ניתן להשתמש ב-bi-directional RNNs ?

ו. הסבירו מה היתרון של bi-directional RNNs על פני $(*)$? תנו דוגמא למקרה שלא ניתן לפתור באמצעות הגישה הרגילה $(*)$, אבל כן ניתן לפתור עם bi-directional RNNs.

ז. קיבלנו סט אימון שמורכב מ-N משפטים מהצורה "נולדתי ב-[ישראל/ סין/...]. המאכל שאני הכי אוהב הוא ****. נולדתי ב-[.../2000/1990]." וצריך להשלים את המילה החסרה (****). בנוסף, נניח כי בסט יש $K > 2$ ארצות שונות, $M > 2$ שנות לידה שונות, ו-$N = KM$ תשובות שונות (מילה אחת לכל קומבינציה של ארץ ושנה). הסבירו מתמטית מדוע לא ניתן להתאים בצורה מושלמת לדוגמאות האימון עם מודל bi-directional RNN. <u>תזכורת:</u> כל מילה מקודדת בייצוג one-hot.

ח. הגדירו מתמטית כיצד נראה מודל bi-directional RNNs רב שכבתי שכן יכול להתאים בצורה מושלמת לסט האימון מסעיף קודם.