רשתות נוירונים לתמונות, תרגיל 2

יואב שפירא | 312492838

22 באפריל 2023

חלק מעשי

הסבר לגבי המימוש שלי:
עשיתי שימוש בפיצ'ר של $Weight\&Biases$ שנקרא $Sweep$ על מנת לחקור את ההשפעות השונות של פרמטרים ברשת, ומבחינת קוד אולי הוא מבלבל. הדבר ממומש ע"י פונקצייה שנקראת סוכן $(agent)$ שמפעילה את ה$Sweep$ עם פרמטרים ספציפיים שעליו לחקור את ההשפעות שלהם (אם הפרמטר הוא יחיד, לדוגמא $batch\ size$ במימוש שלי, אז הוא לא נכלל ב$Sweep$ אלא שם רק לצורך בניית המודל), ועם פונקציית אימון שהגדרתי ובה מוגדרים גם המודלים שבנויים בהתבסס על הפרמטרים הספציפיים לאותה ההרצה.
בנוסף, במימוש שלי הפרמטרים הנ"ל מתייחסים רק אל הארכטיקטורה של ה$Encoder$. הארכטיקטורה של ה$Decoder$ נקבעת על ידי כך באופן יחידני, כך שהוא ישמש תמונת מראה של ה$Encoder$.
במהלך האימון השתמשתי ב$optimizer = Adam$, שראיתי כי השיג תוצאה טובה יותר מSGD.
הפרמטרים לארכיטקטורה הם:

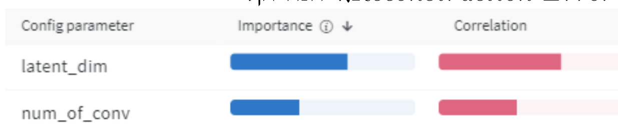
- $Num\ Of\ Conv$: מספר שכבות הקונבולוציה הרצויות. כל שכבות הקונבולוציה הן עם $stride = 2$, במקום פעולת ה$MaxPooling$ כמו שנדרש בתרגיל. (לשים לב שבקוד $stride$ הוא גם פרמטר, כי חקרתי גם את האפשרות ללא $stride$, אבל בפועל השימוש בו הוא רק עבור הערך 2)
- $C - Factor$: בכל שכבת קונבולוציה מספר הערוצים גדל ב$c - factor$. ביחד עם הפרמטר הקודם, הוא מעיד על גודל המודל (על מספר הפרמטרים).
- $Latent\ Dimension$: הערך d המוגדר בתרגיל, שהוא המימד האבסטרקטי.

בנוסף לשכבות הקונבולוציה, יש 2 שכבות של FC. לאחר כל שכבה (FC וקונב') יש שכבת אקטיבציה של $ReLU$, מלבד השכבה שלפני האאוטפוט. ב$Decoder$ יש גם $Sigmoid$ כמו שנדרש בתרגיל.

שאלה 1 $Auto - encoding$

הפרמטרים השונים שהרצתי:
הרצתי עבור מימד אבסטרקטי $d \in \{5, 10, 15, 20\}$, מספר שכבות קונבולוציה $Num\,Of\,Conv \in \{1, 2, 3\}$. הפרמטרים $c - factor = 8$ ו$kernel = 3$ היו קבועים, ונבחרו אחרי בחינה של

מספר ערכים שונים.
בעזרת $Sweep$ ניתן לחלץ קורלציות בין הפרמטרים השונים לבין פונקציית המטרה (במקרה זה - $Reconstruction\ Error$). הנה הן:

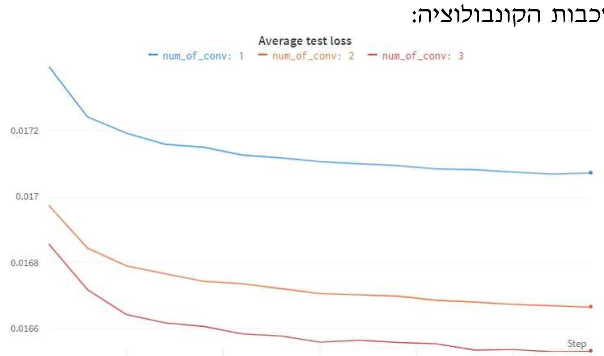


נשים לב שקורלציה שלילית (באדום), משמעותה שככל שהפרמטר גדל כך השגיאה קטנה, ומכאן שמשמעותה היא תרומה להצלחת המודל. ניתן להסיק שככל שגודל המימד האבסטרקטי גדל, כך השגיאה קטנה. באופן דומה אך פחות מובהק, גודל המודל משפיע על מזעור השגיאה.

1. **גודל מודל קבוע עם d משתנה:**
הערכים של d הורצו עם כל הפרמטרים האפשריים לגודל המודל, וכאן אציג את התוצאות כפרמטר של d. הגרף הבא מציג את ה$Reconstrction\ Error$ של כל המודלים לפי פילוח של הפרמטר d. לכל ערך של d יש צבע אחר, והגרף הוא ממוצע השגיאות של המודלים עם d כלשהו:

ניתן לראות שאכן שככל שd גדל, כך השגיאה קטנה. בנוסף אפשר לראות שעבור הd שקטן מ10, השגיאה רחוקה מלהיות טובה, וזה הגיוני - כי הדאטא מורכב מ10 סוגים שונים של תמונות, והמודל אם כך צריך לפחות 10 'מקומות שונים' כדי לקודד אותם.

2. **גודל מימד קבוע וגודל רשת משתנה:**
מוצג כאן גרף כאמו בסעיף הקודם, אך כעת לפי פילוח של גודל המודל - מספר שכבות הקונבולוציה:



גם כאן ניתן לראות שככל שהמודל גדל, כך השגיאה קטנה.

שאלה 2: אינטרפולציה

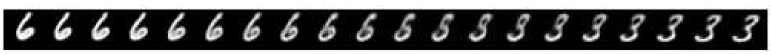
אשווה פה בין תוצאות שמתקבלות ממודלים עם $d = 20$ ו$d = 10$. אפשר לראות בכל הדוגמאות, שהמודל של $d = 20$ משיג תמונה חדה יותר בקצוות של האינטרפולציה (כלומר כש$\alpha = 0$ או $\alpha = 1$, ואז זהו למעשה רק שחזור של תמונה אחת), והמודל עם $d = 10$ משיג תוצאה 'ברורה יותר' - כלומר דומה יותר למספר אמיתי - כשα יותר קרוב ל0.5. התוצאה לגבי $d = 20$ ברורה, בהתחשב בסעיף הקודם - המודל טוב יותר בשחזור התמונה המקורית. התוצאה לגבי $d = 10$, נובעת לדעתי מכך שבמודל הזה כל כניסה ב$latent\ vector$ מכילה מידע מרחבי די גדול, שמשותף לכל הספרות, לעומת מידע מרחבי קטן בכל כניסה בוקטור בגודל 20. אפשר להסתכל על זה כטרייד שבו המודל הקטן 'נזהר' לא לתת פרטים משונים או ספציפיים מאוד, על חשבון חדות התמונה (למעשה זה מאוד דומה לטרייד של $Bias - Variance$):

1. בין הספרות 1 - 5:

$d = 10$:

$d = 20$:

2. בין הספרות 3 - 6:



$d = 10$:

$d = 20$:

3. בין הספרות 0 - 4:

$d = 10$:

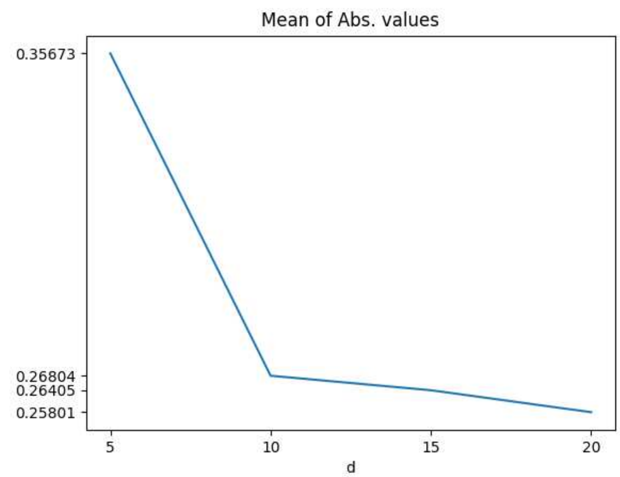
$d = 20$:

שאלה 3: קורלציות

בשאלה זו חקרנו את הקורלציות בין הפיצ'רים השונים ב$latent\ vector$ שה$Encoder$ מייצר, כתלות בגודל המימד d.

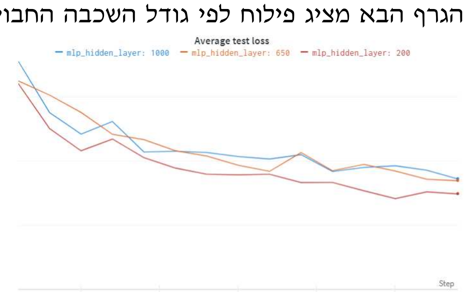
ביצעתי את ההרצה על 8000 תמונות שנבחרו באקראי, עליהם חישבתי מטריצת קורלציות של פירסון בעזרת $numpy$ ($np.corrcoef$). על המטריצה הזו ביצעתי ערך מוחלט ($element-wise$), ומשם חישבתי את התוחלת של הערכים. הרציונל הוא, שכל ערך שהוא לא 0 במטריצה מייצג תלות של פיצ'רים, ובפרט גם ערכים שליליים מייצגים תלות - פשוט קורלציה שלילית. לכן חישוב התוחלת בלי ערך מוחלט עלול להוביל לתוצאה לא חד משמעית. לאחר ערך מוחלט, ככל שהערך מתקרב ל0 כך הקורלציה חלשה יותר, ומכאן שאם נמדוד את הממוצע של הערכים המוחלטים במטריצה נדע האם התלויות בין הפיצ'רים הולכות ונהיות חזקות (לשלילה או לחיוב) או שלהפך. התוצאה היא שככל שd גדל, כך הקורלציה קטנה. אפשר להסביר את זה בכך שככל שהמימד גדל ביחס לכמות המידע שהוא צריך לבטא (למשל כאן, הכמות הזאת היא 10 קידודים שונים של ספרות), כך גם יהיה לו מקום לקודד את הספרות השונות וגם יהיה לו מקום 'עודף', שבו הוא יוכל לבטא אותם בצורה ייחודית ובלתי תלויה.

הגרף:

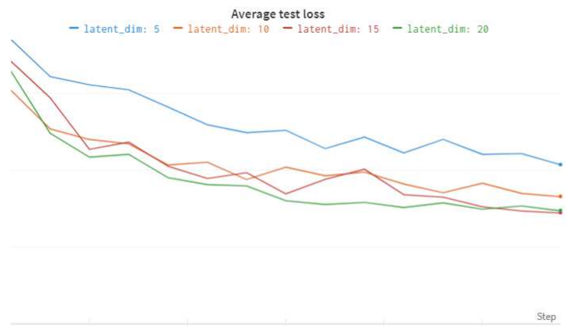


שאלה 4

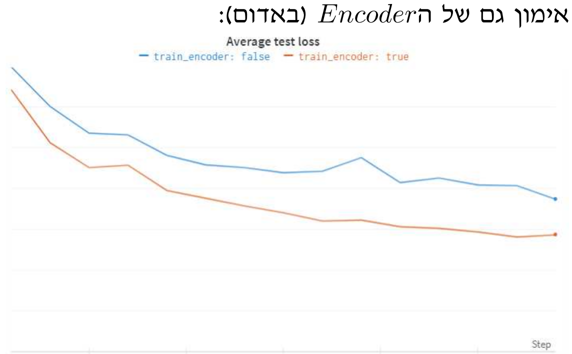
השתמשתי כאן באותם מודלים של $AutoEncoder$ מהשאלה הקודמת (עבור $d \in \{5, 10, 15, 20\}$) וחילצתי מהם את ה$Encoder$.
את הMLP בניתי מ3 שכבות, כשהמימד של המעבר הראשון ניתן כפרמטר, אחד מתוך: $\{200, 650, 1000\}$. השכבה השנייה מורידה את המימד ל100 והשלישית ל10 אל ההאוטפוט.
השתמשתי ב100 תמונות אקראיות מה$Train-set$ שמכילות את כל 10 הלייבלים האפשריים, וביצעתי $evaluation$ על כל הטסט סט.
את הניסוי הרצתי גם עבור סנריו בו מאמנים את ה$Encoder$ ביחד עם ה MLP, וגם עבור סנריו בו מאמנים רק את הMLP.
בסה"כ 24 ריצות, שמוצגות כאן בפילוחים שונים כשהקו מייצג ממוצע של הריצות.
הגרף הבא מציג פילוח לפי גודל השכבה החבויה הראשונה בMLP:



ניתן לראות שההבדלים לא מאוד גדולים (התוצאות די מעורבבות אחת בתוך השנייה) - שזה כבר מעיד על כך שהMLP הוא לאו דווקא הרכיב החשוב בארכיטקטורה. בכל אופן, יש עדיפות כלשהי לMLP קטן יחסית.
הגרף הבא מציג פילוח לפי המימד:



באופן שמתיישב עם הסעיפים הקודמים, ישנה עדיפות למימד גדול יותר שיכול לקודד את המידע יותר באקספרסיביות.
הגרף הבא מציג פילוח לפי שני הניסויים: פעם רק עם אימון של הMLP (בכחול) ופעם עם אימון גם של ה$Encoder$ (באדום):



הניסוי שבו מאמנים רק את הMLP השיג תוצאות פחות טובות. זה הגיוני בהתחשב בכך שלא היה לו הרבה סט אימון, ואין לו ידע מקדים על הדאטא. בניסוי שכולל אימון של ה$Encoder$, המודל שלנו כבר יודע כל מיני דברים על הדאטא כי הוא $Pre-trained$, וההרצה הנוכחית הייתה כמעין $Fine-tuning$ שלו אל בעיית קלסיפיקציה (שאותה הוא מבצע ביחד עם הקלסיפייר שהוא הMLP).

חלק תאורטי

שאלה 1

1. **ליניאריות:** יהיו f,g פונקציות ליניאריות. נראה שההרכבה היא ליניארית:

$$
\begin{aligned}
(f\circ g)(\alpha x+y) &= f(g(\alpha x+y))\\
&{}_{1}= f(\alpha g(x)+g(y))\\
&{}_{2}= \alpha f(g(x))+f(g(y))\\
&{}_{3}= \alpha (f\circ g)(x)+(f\circ g)(y)
\end{aligned}
$$

כאשר השורה הראשונה היא הגדרת הרכבת פונקציות, 1 נובע מליניאריות של g, 2 נובע מליניאריות של f, ו3 שוב הגדרת פונקציות.

2. **אפיניות:** יהיו הפונקציות האפיניות:

$$f:\mathbb{R}^m\to\mathbb{R}^d \quad f(\bar{x})=A\bar{x}+B \quad (\bar{x}\in\mathbb{R}^m A\in\mathbb{R}^{d\times m}, B\in\mathbb{R}^d)$$

$$g:\mathbb{R}^n\to\mathbb{R}^m \quad f(\bar{y})=M\bar{y}+C \quad (\bar{y}\in\mathbb{R}^n, M\in\mathbb{R}^{m\times n}, C\in\mathbb{R}^m)$$

נראה שההרכבה שלהן היא אפינית. לצורך נוחות מעתה נתעלם מהסימן קו שמעל הקלט, ונניח כי מדובר בוקטור (נראה כי המימדים מתאימים אחר כך). יהי $x\in\mathbb{R}^n$:

$$\begin{aligned}(f\circ g)(x) &= f(g(x))\\ {}_1 &= f(Mx+C)\\ {}_2 &= A(Mx+C)+B\\ {}_3 &= AMx+AC+B\end{aligned}$$

הכל פה ישירות מההגדרות. נבחן כעת את המימדים כדי להראות שהכל מוגדר היטב. נבחין ש:$Mx\in\mathbb{R}^m$, ולכן $AMx\in\mathbb{R}^d$. כמו כן גם $AC\in\mathbb{R}^d$ וכמובן גם B, כלומר הכל תקין ואכן נחתנו ב$\mathbb{R}^d$. כעת נגדיר:

$$P=AM,\ Q=AC+B$$

ונקבל ש

$$f\circ g:\mathbb{R}^n\to\mathbb{R}^d,\quad (f\circ g)(x)=Px+Q$$

הגדרה של פונקצייה אפינית.

שאלה 2

1. תנאי העצירה של הפונקצייה האיטרטיבית:

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

הוא למעשה המצב בו $\theta^{n+1}=\theta^n$. כלומר כשמתקיים

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

ומכאן שמתקיים

$$0=\alpha\nabla f_{\theta^n}(x)$$

בהינתן ש $\alpha\neq 0$, נקבל שתנאי העצירה הוא כש$\nabla f_{\theta^n}(x)=0$. כלומר כשהגרדיאנט של הפונקצייה f ביחס לθ^n מתאפס. זוהי ההגדרה של $stationaty\ point$, ובדיוק בהן מתקיים תנאי העצירה.

2. תהי נקודה x_0 כך ש $\nabla f(x_0)=0$ כלומר נקודה סטציונרית. נראה כעת שעבור dx קטן מספיק, נוכל להסיק האם הנקודה היא מינימום או מקסימום בעזרת תנאי על ההסיאן H.

מאחר שמסתכלים על dx קטן (נבחין ש dx הוא וקטור), אפשר להשתמש בפיתוח טיילור בנקודה x_0:

$$f(x_0+h) = f(x_0) + \nabla f(x_0)dx + (dx)^T H_{f(x_0)}(dx) + O\left(||dx||^3\right)$$

כדי להסיק שהנקודה x_0 **היא נקודת מינימום,** אנחנו נדרשים להראות שמתקיים לכל dx קטן התנאי:

$$f(x_0+h) > f(x_0)$$

נבחין תחילה שככל שdx נהיה קטן יותר ויותר, כך מתקיים היחס $(dx)^T H_{f(x_0)}(dx) >> ||dx||^3$ (זאת מכיוון שdx קטן מספיק הוא גם קטן בהרבה מ1) כלומר הביטוי $O\left(||dx||^3\right)$ זניח בפיתוח של טור טיילור לצרכים שלנו.
עכשיו נבחן את הביטוי שנשאר:

$$\begin{aligned} f(x_0+h) &= f(x_0) + \nabla f(x_0)dx + (dx)^T H_{f(x_0)}(dx) \\ &= f(x_0) + (dx)^T H_{f(x_0)}(dx) \end{aligned}$$

וזאת מאחר שידוע ש $\nabla f(x_0) = 0$. ולכן, על מנת שיתקיים $f(x_0+h) > f(x_0)$, לכל dx צריך להתקיים:

$$(dx)^T H_{f(x_0)}(dx) > 0$$

מכאן שהתנאי הוא שהסיאן $H_{f(x_0)}$ תהיה מטריצה $Positive-Definite$, כלומר שכל הע"ע שלה חיוביים.
בצורה זהה, כדי להסיק שהנקודה x_0 **היא נקודת מקסימום,** אנחנו נדרשים להראות שמתקיים לכל dx קטן התנאי:

$$f(x_0+h) < f(x_0)$$

ולכן, לכל dx צריך להתקיים:

$$(dx)^T H_{f(x_0)}(dx) < 0$$

מכאן שהתנאי הוא שההסיאן $H_{f(x_0)}$ תהיה מטריצה $Negative-Definite$, כלומר שכל הע"ע שלה שליליים.

שאלה 3

אנחנו נדרשים למצוא פונקציית $Loss$ שמקיימות את התכונות:

1. אדישות לסימן מינוס, כלומר לכל x, y מתקיים $L(x-y) = L(y-x)$. זאת כמובן בלי להשתמש בערך מוחלט על מנת לשמור על גזירות.
2. אדישות למעגליות, כלומר לכל x, α, β כך שאם המרחקים על המעגל של α וβ מx הם זהים, הם יקבלו את אותו ה$loss$ כלומר מתקיים: $L(x-\alpha) = L(x-\beta)$. למעשה, אנחנו רוצים לקחת את הזווית המינימלית מבין שתי הזוויות האפשריות: $L\left(min\left\{|x-\alpha|, |x-\beta|\right\}\right)$, ונבחין שזו זווית שכמובן אינה גדולה יותר מ π בגלל המעגליות.

נבחין תחילה שהפונקצייה cos מקיימת את התנאי הראשון, $cos(x) = cos(-x)$. כעת נסתכל על הבעיה כבעיה גאומטרית: יהיו זוויות α, β, ונניח בה"כ כי $\alpha > \beta$. נסתכל על קוסינוס ההפרש בין הזוויות, $cos(\theta) = cos(\alpha - \beta) = cos(\beta - \alpha)$, ונסתכל על המשולש על מעגל שמוגדר על ידי θ (מצורף ציור להמחשה). הזווית θ יכולה להיות מבוטאת בצורה אחרת, ע"י הצלע שנמצאת מולה: אם נסמן את הצלע מולה בT, אז $\theta = arccos(T)$. למה זה טוב: נסתכל עכשיו על הזווית $\theta' = \beta - \alpha < 0$. זו זווית שלילית, ובפרט מתקיים $\theta = |\theta'|$, וכאמור מתקיים גם $cos(\theta) = cos(\theta')$. **עכשיו** נסתכל במשולש המוגדר ע"י θ', ונסמן את הצלע שנמצאת מול θ' בT'. נבחין ששני המשולשים הם חופפים: שתי צלעות באורך 1 (רדיוס המעגל) וזווית בגודל θ ביניהן. לכן, $T = T'$, וזה השורש של פיתוח המשוואה שלנו: נגדיר את T בצורה אדישה לסדר של α ושל β, ואז נפעיל $arccos(T)$ כדי לקבל את הגודל של הזווית המינימלית, שאינו עולה מעל π. מכיוון ש$arccos$ היא הופכית של cos, נקבל עבור θ ועבור θ' את אותו הערך. נסמן את T בהתחלה פשוט על ידי מרחק בין נקודות ($P1, P2$ בציור):

$$T = \sqrt{\left(cos(\alpha) - cos(\beta)\right)^2 + \left(sin(\alpha) - sin(\beta)\right)^2}$$

נפתח את המשוואה:

$$\begin{aligned} T &= \sqrt{cos(\alpha)^2 - 2cos(\alpha)cos(\beta) + cos(\beta)^2 + sin(\alpha)^2 - 2sin(\alpha)sin(\beta) + sin(\beta)^2} \\ {}_1 &= \sqrt{1 - 2cos(\alpha)cos(\beta) + 1 - 2sin(\alpha)sin(\beta)} \\ {}_2 &= \sqrt{2\left(1 - \left(cos(\alpha)cos(\beta) - sin(\alpha)sin(\beta)\right)\right)} \\ {}_3 &= \sqrt{2}\sqrt{1 - cos(\alpha - \beta)} \end{aligned}$$

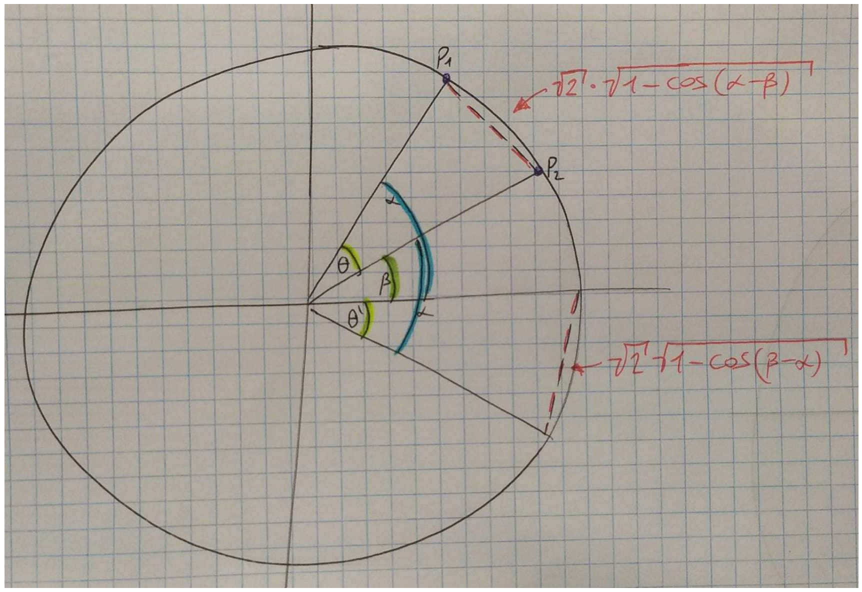
כש1 פתיחת סוגריים, 2 זהות $cos^2 + sin^2 = 1$, 3 קיבוץ ו4 פתיחה לשורש ומהזהות טריגונומטרית.
מאחר שאנחנו מעוניינים למצוא $Loss$ מונוטוני (כלומר שגדל ככל שהזווית מתרחקת מהמטרה), ומאחר ואנחנו יודעים ש$arccos$ הוא מונוטוני, אנחנו יכולים להשתמש במשוואה פשוטה יותר לביטוי T , ועדיין מונוטונית, על מנת להגדיר על ידיה את θ. נפשט כך:

$$T_{loss} = 1 - cos(\alpha - \beta)$$

הורדנו כפל בקבוע ושורש ושמרנו על מונוטוניות של T. ומכאן נגדיר לבסוף:

$$L\left(\alpha, \beta\right) = arccos\left(1 - cos\left(\alpha - \beta\right)\right)$$

זו פונקצייה מונוטונית בT, שהיא מונוטונית בהפרש בין α לβ בקטע $[0, \pi]$ כלומר לוקחת בחשבון את המעגליות. בנוסף היא מוגדרת על כל הישר ובפרט על הקטע $[0, 2\pi]$.
ובנוסף היא גזירה בכל נקודה!



שאלה 4

.1

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

.2

$$\left(f_1\left(f_2\left(f_3\left(...f_n\left(x\right)\right)\right)\right)\right)'=f_1'\left(f_2\left(f_3\left(...f_n\left(x\right)\right)\right)\right)\cdot f_2'\left(f_3\left(...f_n\left(x\right)\right)\right)$$
$$\cdot f_3'\left(...f_n\left(x\right)\right)\cdot\ .\ .\ .\ .\ \cdot f_{n-1}'\left(f_n(x)\right)\cdot f_n'\left(x\right)$$

ובצורה מפושטת יותר:

$$\prod_{i=1}^{n}f_i'\left(f_{i+1}\left(...f_n\left(x\right)\right)\right)$$

.3

$$\frac{\partial}{\partial x}f_1\left(x,f_2\left(x,f_3\left(...f_{n-1}\left(x,f_n(x)\right)\right)\right)\right)=\frac{\partial f_1}{\partial x}+\frac{\partial f_1}{\partial f_2}\frac{\partial f_2}{\partial x}$$

את הביטוי הזה מפתחים בצורה רקורסיבית:

$$=\frac{\partial f_1}{\partial x}+\frac{\partial f_1}{\partial f_2}\left(\frac{\partial f_2}{\partial x}+\frac{\partial f_2}{\partial f_3}\frac{\partial f_3}{\partial x}\right)=...=\frac{\partial f_1}{\partial x}+\frac{\partial f_1}{\partial f_2}\left(\frac{\partial f_2}{\partial x}+\frac{\partial f_2}{\partial f_3}\left(.....\frac{\partial f_{n-1}}{\partial x}+\left(\frac{\partial f_{n-1}}{\partial x}\frac{\partial f_n}{\partial x}\right)\right)\right)$$

.4

$$(f(x+g(x+h(x))))' = f'(x+g(x+h(x)))\cdot(1+g'(x+h(x)))\cdot(1+h'(x))$$