רשתות עמוקות

עד כה למדנו את יסודות תהליך האימון של רשתות נוירונים עמוקות, אך למעשה הפעלנו אותו על רשת "שטוחה" בלבד: רשת המכילה שכבה אחת של נוירונים. לאחר האימון ראינו שרשת מסוג זה מסוגלת לסווג את פריטי הלבוש של אוסף הנתונים Fashion-MNIST בדיוק של כ-80%. אם נרצה להגיע לדיוק גבוה יותר, עלינו לבחור במודל מורכב יותר – **רשת עמוקה**. זו תהיה רשת המורכבות משכבות של נוירונים המחוברות **לנוירונים נוספים** ורק אחרי מספר שכבות הפלט יועבר לפונקציית ה-Softmax למטרת חישוב ההסתברויות. כאן בא לידי ביטוי כוחה של הספרייה PyTorch: בכדי לאמן רשת עמוקה, עלינו לשנות את הגדרת המודל בלבד, בעוד ששאר הקוד הדרוש לאימון הרשת נשאר זהה. למשל בקטע הקוד הבא אנו מגדירים רשת בעלת שתי שכבות.

```
model = nn.Sequential(
          nn.Linear(784,20),
          nn.ReLU(),
          nn.Linear(20,10),
          nn.LogSoftmax(dim=1)
          )
```

חישוב הסתברויות החיזוי ברשת זו מתבצע בדרך הבאה:

1. תמונות פרטי הלבוש (לאחר שעברו שיטוח) מועברת לשכבה הראשונה, בה ראשית מופעלת פונקציה ליניארית על הקלט:

$$Z^{(1)}=\begin{pmatrix} z_0^{(1)} \\ z_1^{(1)} \\ \vdots \\ z_{19}^{(1)} \end{pmatrix}=\begin{pmatrix} w_{0,0}^{(1)}x_0+w_{0,1}^{(1)}x_1+\ldots+w_{0,783}^{(1)}x_{783}+b_0^{(1)} \\ w_{1,0}^{(1)}x_0+w_{1,1}^{(1)}x_1+\ldots+w_{1,783}^{(1)}x_{783}+b_1^{(1)} \\ \vdots \\ z_{19}^{(1)}=w_{19,0}^{(1)}x_0+w_{19,1}^{(1)}x_1+\ldots+w_{19,783}^{(1)}x_{783}+b_{19}^{(1)} \end{pmatrix}$$

ושנית, על הפונקציה הליניארית מופעלת אקטיבציה לא ליניארית מסוג ReLU:

$$,\mathrm{ReLU}(x)=\begin{cases} 0 & x<0 \\ x & x\geq 0 \end{cases}$$

ממנה מתקבל הפלט של השכבה הראשונה,

$$Y^{(1)}=\begin{pmatrix} y_0^{(1)} \\ y_1^{(1)} \\ \vdots \\ y_{19}^{(1)} \end{pmatrix}=\mathrm{ReLU}\left(Z^{(1)}\right)=\begin{pmatrix} \mathrm{ReLU}\left(z_0^{(1)}\right) \\ \mathrm{ReLU}\left(z_1^{(1)}\right) \\ \vdots \\ \mathrm{ReLU}\left(z_{19}^{(1)}\right) \end{pmatrix}$$

2. פלט זה מועבר לשכבה השניה, אשר בה שוב מופעלת פונקציה ליניארית, **שונה מהקודמת**:

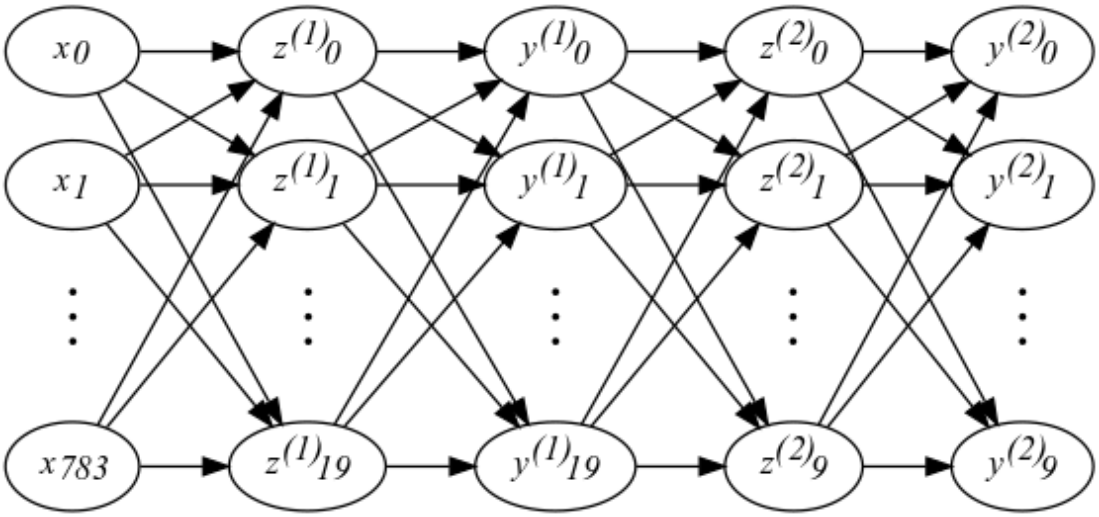
$$Z^{(2)}=\begin{pmatrix} z_0^{(2)} \\ z_1^{(2)} \\ \vdots \\ z_9^{(2)} \end{pmatrix}=\begin{pmatrix} w_{0,0}^{(2)}y_0^{(1)}+w_{0,1}^{(2)}y_1^{(1)}+\ldots+w_{0,19}^{(2)}y_{19}^{(1)}+b_0^{(2)} \\ w_{1,0}^{(2)}y_0^{(1)}+w_{1,1}^{(2)}y_1^{(1)}+\ldots+w_{1,19}^{(2)}y_{19}^{(1)}+b_1^{(2)} \\ \vdots \\ w_{9,0}^{(2)}y_0^{(1)}+w_{9,1}^{(2)}y_1^{(1)}+\ldots+w_{9,19}^{(2)}y_{19}^{(1)}+b_9^{(2)} \end{pmatrix}$$

3. לבסוף, תוצאת חישוב זה תועבר לפונקציית ה-Softmax,

$$Y^{(2)} = \begin{pmatrix} y_0^{(2)} \\ y_1^{(2)} \\ \vdots \\ y_9^{(2)} \end{pmatrix} = \mathrm{softmax}\left(Z^{(2)}\right) = \frac{1}{\sum_{n=0}^{9} e^{z_n^{(2)}}} \begin{pmatrix} e^{z_0^{(2)}} \\ \vdots \\ e^{z_9^{(2)}} \end{pmatrix}$$

אשר הפלט שלה הוא הסתברויות הסיווג הסופיות – פלט המודל.

באיור סכמתי,



בכתיב וקטורי, פעולת החישוב המבוצעת ברשת היא

$$Y^{(2)} = \mathrm{softmax}\left(W^{(2)}\mathrm{ReLU}\left(W^{(1)}X + b^{(1)}\right) + b^{(2)}\right),$$

וכאן ניתן לראות שפרמטרי המודל הם המטריצות $W^{(1)}, W^{(2)}$ והוקטורים $b^{(1)}, b^{(2)}$, אשר קובעים את ערך האגרגציות הליניאריות בכל שכבה.

למודל זה מספר יתרונות על פני הרשת השטוחה, הראשון שבהם הוא היותו לא מונוטוני כפונקציה של הקלט. זכרו שערכי וקטור הקלט X הם הפיקסלים של תמונת פריט לבוש כלשהו – ערך קרוב ל1 מייצג פיקסל כהה וערך קרוב ל0 מייצג פיקסל בהיר. זכרו שהחישוב המבוצע ברשת השטוחה הוא $Y = \mathrm{softmax}\left(W \cdot X + b\right)$.

מכיוון שגם פונקציית ה-softmax וגם האגרגציה הליניארית מונוטוניות בקלט שלהן, נובע שהפלט גם הוא מונוטוני. משמעות הדבר היא שאם, לדוגמה, אחר האימון התקבל משקל חיובי $w_{k,m} > 0$ אזי ככל שהפיקסל ה- m בקלט כהה יותר, כך ההסתברות לסיווג פריט הלבוש למחלקה ה- k תהיה גבוהה יותר. מובן מאליו שתכונה זו אינה מתאימה לבעיה המורכבת של זיהוי פריט לבוש הדורשת מהמודל להבין את ההקשר בין הפיקסלים השונים. הוספת שכבה נוספת לרשת, עם אקטיבציה לא ליניארית כגון ReLU תאפשר את הגמישות הרצויה במקרה זה – למודל תהיה אפשרות להימנע מהתלות המונוטונית בקלט.

יש לשים לב שהאקטיבציה הלא ליניארית בין שכבה לשכבה **הינה הכרחית** לגמישות זו. לו היינו מפעילים פונקציה ליניארית בלבד במעבר משכבה לשכבה, קרי מחשבים את $Z^{(2)}$ ישירות מ- $Z^{(1)}$: $Z^{(2)} = W^{(2)}Z^{(1)} + b^{(2)}$, המודל המתקבל היה שקול לרשת השטוחה, כפי שתוכיחו בשאלות בהמשך. לעתים גמישות רצויה זו באה עם מחיר – אימון הרשת מאתגר יותר, וכן הרשת עלולה לשנן את אוסף הנתונים המשמש לאימונה, במקום ללמוד את הכללים הפשוטים העומדים מאחוריו. הכלים העומדים לרשותנו להתמודד עם אתגרים אלו הם נושאי הפרקים הבאים.

שאלות לתרגול

1. הוכיחו שלאחר אימון רשת שטוחה לזיהוי פרטי הלבוש, תמיד יתקבל מודל אשר מונוטוני בקלט. **הנחיה:** חשבו את הנגזרת $\frac{\partial y_k}{\partial x_m}$ בעזרת כלל השרשרת, והסתכלו על הסימן של ביטוי זה.
2. הראו שניתן לבחור את פרמטרי הרשת העמוקה שתוארה לעיל כך שהיא אינה מונוטונית.
הנחיות:
   - התבוננו בתלות של y_0 בפיקסל הקלט x_0 בלבד.
   - בחרו את $w_{0,0}^{(1)}$ ואת $w_{1,0}^{(1)}$ עם סימנים הפוכים: האחד חיובי, השני שלילי. שאר הפרמטרים בשכבה הראשונה יכולים להיות 0.
   - בחרו את $w_{0,0}^{(2)}$ ואת $w_{0,1}^{(2)}$ עם סימנים זהים. שאר הפרמטרים בשכבה השניה יכולים להיות 0.
3. הוכיחו שחישוב שתי שכבות של אגרגציה לינארית בזו אחר זו שקולה לחישוב אגרגציה לינארית אחת. במילים אחרות, הוכיחו שאם מחשבים את $Z^{(2)}$ בדרך הבאה,

$$Z^{(1)} = W^{(2)}X + b^{(1)}$$

$$Z^{(2)} = W^{(2)}Z^{(1)} + b^{(2)}$$

אז למעשה, ניתן לחשב ישירות $Z^{(2)} = Ax + c$. מצאו את המטריצה A והוקטור c המתאימים. הסיקו שרשת עמוקה ללא אקטיבציה לא-ליניארית בין שכבה לשכבה שקולה לרשת שטוחה עם שכבה ליניארית אחת.

4.
   א. צרו אוסף נתונים של נקודות שחורות ולבנות בעזרת הפונקציה `make_moons` מהספרייה `sklearn.datasets`.
   ב. אמנו את מודל הסיווג של הנוירון הבודד מתחילת יחידת הלימוד על אוסף נתונים זה.
   ג. האם מודל הסיווג מוצלח? נמקו תשובתכם בדגש על יכולות המודל ומבנה אוסף הנתונים.
   ד. הרחיבו את המודל – הוסיפו שכבות ביניים בין הקלט לבין נוירון הסיווג, ואמנו את המודל החדש.
   ה. האם ביצועי המודל טובים יותר כעת? האם תוכלו להסביר מדוע?