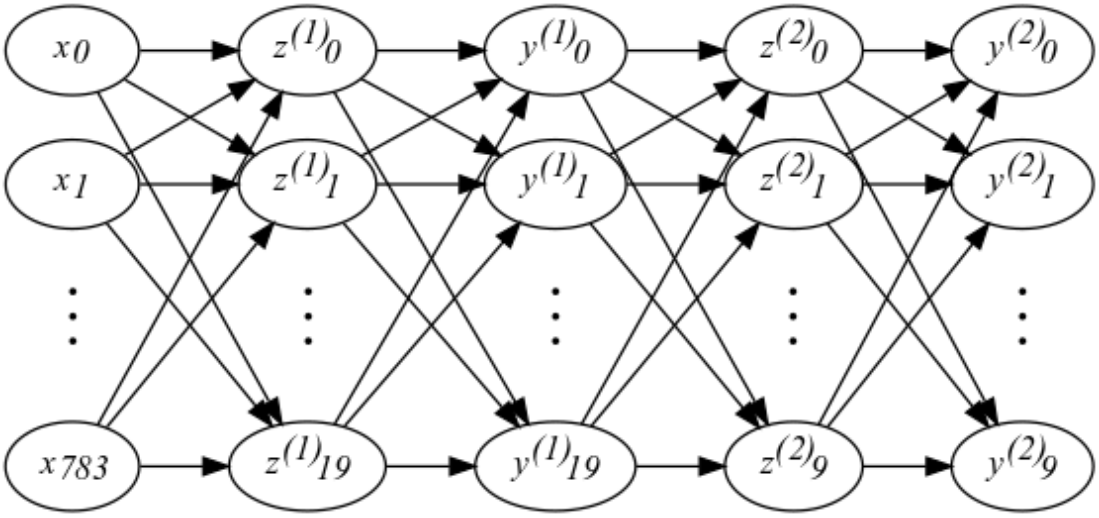
פעפוע לאחור

ביחידת הלימוד הקודמת ראינו כיצד לאמן רשת נוירונים עמוקה בעזרת אלגוריתם מורד הגרדיאנט, אך השארנו את הפרטים של חישוב הנגזרות למערכת הגזירה האוטומטית. כעת נביט בתהליך זה לעומק, ודרכו נבין את האלגוריתם העומד מאחורי חישוב הגרדיאנט עצמו בכל איטרציה, פעפוע לאחור (backpropagation).

נזכיר שהרשת בה אנו עוסקים הנתונה בציור הבא,



אליה מוזנת תמונת הקלט כוקטור משוטח, X , והחישוב אשר תבצע עליו הוא

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

תוצאה זו אנו מפרשים בתור הסתברויות הסיווג לעשר מחלקות פרטי הלבוש של אוסף הנתונים Fashion-MNIST, ובכדי להעריך את טיב החיזוי של הרשת עבור הקלט הנתון, אנו מחשבים את פונקציית המחיר האנטרופיה הצולבת,

$$H\left(X, Y_t\right) = -\sum_{n=0}^{9} y_{tn} \log\left(y_n^{(2)}\right).$$

בהמשך, את פונקציה זו אנו גוזרים לפי כל ערכי הפרמטרים המופיעים ברשת, כאשר במקרה הנוכחי הם מקדמי השכבות הליניאריות, המטריצות $W^{(1)}, W^{(2)}$ והוקטורים $b^{(1)}, b^{(2)}$. לאחר חישוב גרדיאנט זה (או למעשה בו זמנית), חוזרים על הפעולה עבור minibatch של נתונים, ולבסוף בשביל צעד עדכון יחיד של אלגוריתם SGD אנו מחשבים ממוצע של הגרדיאנטים,

$$\nabla C = \frac{1}{\#minibatch} \sum_{(X, Y_t) \in minibatch} \nabla H\left(X, Y_t\right).$$

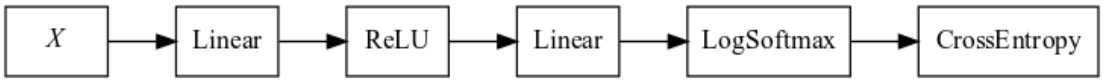
מטרתנו כעת היא לדון בפירוט בתהליך בו מחושב $\nabla H\left(X, Y_t\right)$. התובנות העיקריות מאחורי השיטה הן:

1. לצורך חישוב פונקציית המחיר H אנו מפעילים פונקציות פשוטות אחת על השניה, ואת הנגזרות של כל אחת מפונקציות אלו ביחס לקלט שלה קל לנו לחשב.
2. H תלויה בערכי הפרמטרים דרך **מספר רב** של פונקציות (מספר השכבות ברשת), ואת הנגזרות של H לפיהם ניתן לקבל מהנגזרות הפשוטות **בעזרת כלל השרשרת**.
3. כל הערכים הדרושים לחישוב נגזרות אלו חושבו במהלך הזנת הקלט ברשת וחישוב ההסתברויות החזויות. בהתאם, על מנת להדגיש את חשיבותו לאלגוריתם backpropagation, שלב הזנת הקלט נקרא גם forward propagation.
4. עלות חישוב ∇H היא כעלות החישוביות של צעד forward propagation יחיד.

ניגש למשימה אם כן, מהסוף אל ההתחלה, אך קודם לכן נזכיר שבעת מימוש הרשת בחרנו לחשב בשכבה האחרונה את הלוגריתם של פונקציית ה-Softmax במקום את הפונקציה עצמה. כעת נראה יתרון נוסף לבחירה זו – חישוב הנגזרות הופך קל משמעותית. נניח אם כן שפלט הרשת הוא

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

ונצייר את הרשת ברמת הפשטה גבוהה יותר, המדגישה את המבנה הרשת כפונקציה מקוננת.

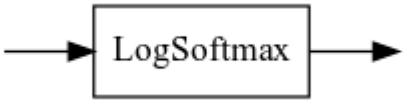


בשלב הראשון עלינו לחשב את הנגזרת של פונקציית המחיר H לקלט שלה, $\log\left(Y^{(2)}\right)$:

$$\frac{\partial H\left(X,Y_t\right)}{\partial\log\left(y_k^{(2)}\right)}=-\sum_{n=0}^{9}\frac{\partial}{\partial\log\left(y_k^{(2)}\right)}\left(y_{tn}\log\left(y_n^{(2)}\right)\right)=-y_{tk}$$

ויש לזכור ש-$y_{tk}=1$ אם ורק אם k היא המחלקה אליה מסווג הקלט X באוסף נתוני האימון.

כעת נמשיך לנוע אחורה ברשת, ונתמקד בשכבת הפלט של הרשת,



אם כן, עלינו לגזור את פונקציית ה-LogSoftmax. נזכור שזו פונקציה המקבלת כקלט וקטור k-מימדי U, ומחזירה וקטור k-מימדי V, אשר הגדרתה היא

$$V=\operatorname{LogSoftmax}U=\operatorname{LogSoftmax}\begin{pmatrix}u_0\\ \vdots\\ u_k\end{pmatrix}=\begin{pmatrix}u_0-\log\left(\sum_{n=0}^{k}e^{u_n}\right)\\ \vdots\\ u_k-\log\left(\sum_{k=0}^{k}e^{u_n}\right)\end{pmatrix}.$$

בהתאם, יש לחשב את הנגזרת של כל אחד מערכי וקטור הפלט, v_m , ביחס לכל אחד ממשתני הקלט u_n. למעשה, אנו רוצים לחשב את מטריצת היעקוביאן:

$$J_{\text{LogSoftmax}}=\begin{pmatrix}\frac{\partial v_0}{\partial u_0} & \cdots & \frac{\partial v_0}{\partial u_k}\\ \vdots & \ddots & \vdots\\ \frac{\partial v_k}{\partial u_0} & \cdots & \frac{\partial v_k}{\partial u_k}\end{pmatrix}.$$

חישוב ישיר יניב עבור איברי האלכסון של מטריצה זו,

$$\frac{\partial v_k}{\partial u_k}=1-e^{v_k}$$

ועבור שאר המטריצה:

$$\frac{\partial v_m}{\partial u_n}=-e^{v_n}.$$

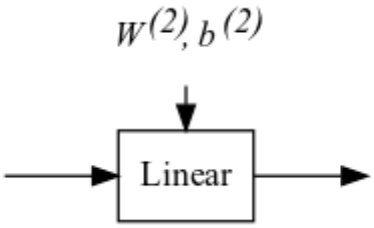
כעת יש לזכור שהקלט לשכבה זו ברשת שלנו הוא $Z^{(2)}$, וכן שפלט שכבת ה-Softmax, ללא הלוגריתם הוא $Y^{(2)}$, ולשים לב לכך ש-

$$e^V=\exp\left(\operatorname{LogSoftmax}U\right)=\operatorname{Softmax}U.$$

על כן, היעקוביאן המתקבל במקרה הנוכחי הוא מטריצה 10X10,

$$\begin{pmatrix} \frac{\partial \log\left(y_0^{(2)}\right)}{\partial z_0^{(2)}} & \frac{\partial \log\left(y_0^{(2)}\right)}{\partial z_1^{(2)}} & \cdots & \frac{\partial \log\left(y_0^{(2)}\right)}{\partial z_9^{(2)}} \\ \frac{\partial \log\left(y_1^{(2)}\right)}{\partial z_0^{(2)}} & \frac{\partial \log\left(y_1^{(2)}\right)}{\partial z_1^{(2)}} & \cdots & \frac{\partial \log\left(y_1^{(2)}\right)}{\partial z_9^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \log\left(y_9^{(2)}\right)}{\partial z_0^{(2)}} & \frac{\partial \log\left(y_9^{(2)}\right)}{\partial z_1^{(2)}} & \cdots & \frac{\partial \log\left(y_9^{(2)}\right)}{\partial z_9^{(2)}} \end{pmatrix} = \begin{pmatrix} 1-y_0^{(2)} & -y_1^{(2)} & \cdots & -y_9^{(2)} \\ -y_0^{(2)} & 1-y_1^{(2)} & \cdots & -y_9^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ -y_0^{(2)} & -y_1^{(2)} & \cdots & 1-y_9^{(2)} \end{pmatrix}.$$

נמשיך לנוע אחורה ברשת ולחשב נגזרות, ונשים לב במיוחד לכך שלראשונה הגענו לשכבה בה יש פרמטרים אשר לפיהם, בין השאר, יש לחשב את גרדיאנט פונקציית המחיר. זו השכבה הליניארית השניה, ובאיור:



ראשית יש לחשב את הנגזרות של פלט שכבה זו ביחס **לפרמטרים**: כאן נשים לב שהפלט הוא וקטור 10 מימדי, $Z^{(2)}$, אשר כל איבר שלו יש לגזור לפי כל אחד מהפרמטרים – איברי המטריצה $W^{(2)}$ ואיברי הוקטור $b^{(2)}$. זכרו שהפעולה הליניארית אשר מתבצעת בשכבה זו היא

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

ולכן ערכי הנגזרות הם:

1. עבור פרמטרים ופלט מאותה שורה, $\frac{\partial z_k^{(2)}}{\partial b_k^{(2)}} = 1$, $\frac{\partial z_k^{(2)}}{\partial w_{k,m}^{(2)}} = y_m^{(1)}$.
2. עבור פרמטרים ופלט משורות שונות: $\frac{\partial z_k^{(2)}}{\partial b_n^{(2)}} = 0$, $\frac{\partial z_k^{(2)}}{\partial w_{n,m}^{(2)}} = 0$, כאשר $k \neq n$.

מטרתנו היא לחשב את הנגזרות של H ביחס לערכי הפרמטרים, וכעת יש בידינו כל הדרוש לכך עבור הפרמטרים $W^{(2)}$ ו- $b^{(2)}$: יש לחלץ ביטוי זה מהערכים שחושבו לעיל בעזרת שימוש חוזר בכלל השרשרת: לאחר התבוננות במבנה הרשת, נשים לב שניתן לבטא את התלות של H בפרמטרים דרך $Z^{(2)}$, ועל כן נקבל מכלל השרשרת, למשל עבור $w_{p,q}^{(2)}$:

$$\frac{\partial H}{\partial w_{p,q}^{(2)}} = \sum_{k=0}^{9} \frac{\partial H}{\partial z_k^{(2)}} \frac{\partial z_k^{(2)}}{\partial w_{p,q}^{(2)}}.$$

את הנגזרות מהצורה $\frac{\partial H}{\partial z_k^{(2)}}$ יש לחשב פעם אחת בלבד, ולהשתמש בהן עבור כל הפרמטרים. שוב כלל השרשרת יהיה שימושי,

$$\frac{\partial H}{\partial z_k^{(2)}} = \sum_{n=0}^{9} \frac{\partial H\left(X, Y_t\right)}{\partial \log\left(y_n^{(2)}\right)} \frac{\partial \log\left(y_n^{(2)}\right)}{\partial z_k^{(2)}}.$$

בשלב זה הגענו לביטויים אשר את ערכיהם חישבנו לעיל, ולכן כל שנשאר הוא להציב:

$$\frac{\partial H}{\partial z_k^{(2)}} = \sum_{n=0}^{9} -y_{tn}\left(1\{n=k\} - y_n^{(2)}\right)$$

כאשר $1\{n=k\} = \begin{cases} 1 & n=k \\ 0 & n \neq k \end{cases}$. לבסוף נקבל, למשל עבור $w_{0,1}^{(2)}$:

$$\frac{\partial H}{\partial w_{0,1}^{(2)}} = \sum_{k=0}^{9} \frac{\partial H}{\partial z_k^{(2)}} \frac{\partial z_k^{(2)}}{\partial w_{0,1}^{(2)}} = \frac{\partial H}{\partial z_0^{(2)}} \frac{\partial z_0^{(2)}}{\partial w_{0,1}^{(2)}} =$$

$$= \sum_{n=0}^{9} -y_{tn}\left(1\{n=0\} - y_n^{(2)}\right) \cdot y_1^{(1)}$$

ובערך זה נשתמש לעדכון הפרמטר בצעד הבא של אלגוריתם האופטימיזציה.

מכיוון שהרשת עמוקה, התהליך אינו נגמר בשלב זה, שכן יש עוד פרמטרים אותם יש לעדכן - נמשיך לגזור דרך השכבות עד אשר נגיע אליהם. ראשית, עלינו לגזור את השכבה הליניארית הנ"ל לפי הקלט שלה, $\frac{\partial z_k^{(2)}}{\partial y_m^{(1)}} = w_{k,m}^{(2)}$. ושנית, עלינו לפעפע את הנגזרות של H אל מעבר לשכבה זו, כלומר לחשב את הנגזרות ביחס **לקלט השכבה**, $\frac{\partial H}{\partial Y^{(1)}}$. זאת נעשה שוב בעזרת כלל השרשרת וכאן נשים לב שאנו משתמשים ב- $\frac{\partial H}{\partial Z^{(2)}}$, אשר חישבנו לעיל בשנית.

$$\frac{\partial H}{\partial y_m^{(1)}} = \sum_{k=0}^{19} \frac{\partial H}{\partial z_k^{(2)}} \frac{\partial z_k^{(2)}}{\partial y_m^{(1)}} = \sum_{k=0}^{19} \sum_{n=0}^{9} -y_{tn}\left(1\{n=k\} - y_n^{(2)}\right) w_{k,m}^{(2)}.$$

בערכים אלו נשתמש בכדי להמשיך ולפעפע את הנגזרות מעבר לאקטיבציית ה-ReLU. נזכור ש-

$$\begin{pmatrix} y_0^{(1)} \\ \vdots \\ y_0^{(1)} \end{pmatrix} = Y^{(1)} = \mathrm{ReLU}\left(Z^{(1)}\right) = \begin{pmatrix} \mathrm{ReLU}\left(z_0^{(1)}\right) \\ \vdots \\ \mathrm{ReLU}\left(z_0^{(1)}\right) \end{pmatrix},$$

וכן ש-

$$\mathrm{ReLU}(x) = x \cdot 1\{x \geq 0\} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

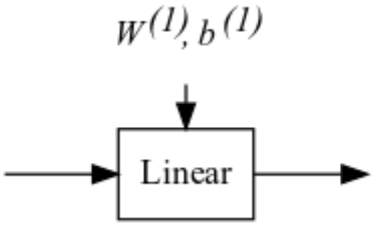
נתעלם מכך שפונקציה זו אינה גזירה בנקודה 0, שכן ההסתברות שהקלט שלה יהיה בדיוק 0 היא זניחה, ונקבל שכאשר $\frac{\partial y_m^{(1)}}{\partial z_n^{(1)}} = 0, m \neq n$ וכאשר יש שוויון,

$$\frac{\partial y_m^{(1)}}{\partial z_m^{(1)}} = 1\left\{z_m^{(1)} \geq 0\right\} = \begin{cases} 1 & z_m^{(1)} \geq 0 \\ 0 & z_m^{(1)} < 0 \end{cases}.$$

נשאר רק להפעיל שוב את כלל השרשרת,

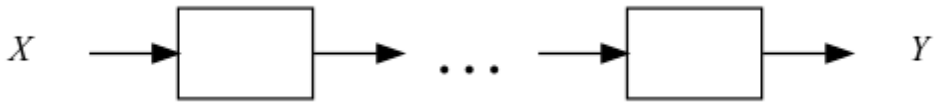
$$\frac{\partial H}{\partial z_m^{(1)}} = \sum_{p=0}^{19} \frac{\partial H}{\partial y_p^{(1)}} \frac{\partial y_p^{(1)}}{\partial z_m^{(1)}} = \frac{\partial H}{\partial y_m^{(1)}} \frac{\partial y_m^{(1)}}{\partial z_m^{(1)}} =$$
$$= \sum_{k=0}^{19} \sum_{n=0}^{9} -y_{tn} \left(1\{n=k\} - y_n^{(2)}\right) w_{k,m}^{(2)} 1\left\{z_m^{(1)} \geq 0\right\}.$$

בשלב זה הגענו שוב לשכבה ליניארית עם פרמטרים, כפי שניכר מהאיור הבא,



ובידינו הנגזרות של פונקציית המחיר ביחס לפלט השכבה, $\frac{\partial H}{\partial Z^{(1)}}$. החישוב החל מכאן אנלוגי לזה שנעשה לעיל עבור השכבה הליניארית הקודמת.

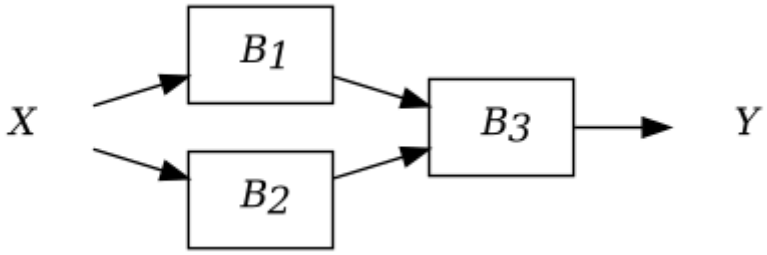
מדוגמה זו יש להבין את האלגוריתם הכללי, כפי שהוא פועל על רשת של **בלוקים** המחוברים זה לזה ברצף, כמו זו המופיעה באיור הבא.



בלוק יחיד יכול להיות מורכב ממספר שכבות, ולבצע חישוב מורכב, אך כל עוד ניתן לגזור דרכו: לגזור את פלט הבלוק לפי הקלט שלו, אלגוריתם הפעפוע לאחור יהיה שימושי. יש לזכור לפעפע אחורה את הנגזרת של פונקציית המחיר דרך הבלוק, וכן לחשב את הנגזרות של הפרמטרים של הבלוק עצמו.
שימו לב שהמחלקה `nn.Sequential` למעשה ממששת ארכיטקטורה זו – בעת הגדרת אובייקט, המשתמש מעביר את פרטי כל הבלוקים והסדר בהם הם מופיעים ברשת.

שאלות לתרגול

1. השלימו את חישוב ∇H: חשבו את הנגזרות לפי פרמטרי השכבה הליניארית הראשונה.
2. כתבו את אלגוריתם הפעפוע לאחור עבור רשת המורכבת מבלוקים ברצף בפסאודו קוד.
3. הסבירו כיצד יש להתאים את אלגוריתם הפעפוע לאחור בכדי להתמודד עם רשת המכילה את מבנה הבלוקים הבא.



כיצד יש לחשב את $\frac{\partial Y}{\partial X}$? **רמז:** שימו לב לקשר המתמטי בין הפלט לקלט:

$Y = B_3\left(B_2\left(X\right), B_1\left(X\right)\right)$ והשתמשו בכלל השרשרת עבור פונקציה רבת משתנים.

4. הריצו את קטע הקוד הבא,

```
from torch import nn
net = nn.Sequential(
          nn.Linear(2,1),
          nn.ReLU(),
          nn.Linear(1,2),
          nn.ReLU(),
          nn.Linear(2,1),
          nn.Sigmoid()
          )
```

א. ציירו באופן סכמטי את הרשת המתקבלת.

ב. הדפיסו את ערכי הפרמטרים של הרשת, וכתבו בשורה אחת את החישוב אשר הרשת מבצעת.

ג. הזינו לתוך הרשת את הקלט $X = \left(1,1\right)$ והניחו שהפלט הצפוי מהרשת הוא 1.

ד. חשבו באופן ידני ולפי אלגוריתם הפעפוע לאחור את ∇H עבור קלט זה, בהנחה שפונקציית המחיר היא האנטרופיה הצולבת.

ה. השוו את החישוב שלכם לזה המתקבל ממערכת הגזירה האוטומטית.