2(a)

$$J = CE(y, \hat{y})$$

נסמן

נניח כי $y_k = 1$

עבור $i = k$

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left(-\sum_j y_j \log \hat{y}_j\right)$$

$$= \frac{\partial}{\partial \theta_i}\left(-y_i \cdot \log \hat{y}_i\right) = \frac{\partial}{\partial \theta_i}\left(-\log \hat{y}_i\right)$$

$$= \frac{\partial}{\partial \theta_i}\left(-\log\left(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right)\right) =$$

$$\frac{\partial}{\partial \theta_i}\left(-\left(\log e^{\theta_i} - \log\left(\sum_j e^{\theta_j}\right)\right)\right) =$$

$$\frac{\partial}{\partial \theta_i}\left(-\left(\theta_i - \log\left(\sum_j e^{\theta_j}\right)\right)\right) =$$

$$-\left(\frac{\partial}{\partial \theta_i}\theta_i - \frac{\partial}{\partial \theta_i}\log \sum_j e^{\theta_j}\right)$$

$$-\left(-1-\frac{1}{\sum_j e^{\theta_j}}\cdot e^{\theta_i}\right)=$$

$$=-\left(1+\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right)=-1+\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}=$$

$$=\quad \text{softmax}(\theta)_i-1$$

עבור $i\neq k$

$$\frac{\partial J}{\partial\theta_i}=\frac{\partial(-\log\hat{y}_k)}{\partial\theta_i}=$$

$$\frac{\partial}{\partial\theta_i}\left(-\log\left(\frac{e^{\theta_k}}{\sum_j e^{\theta_j}}\right)\right)=\frac{\partial}{\partial\theta_i}\left(-\left(\theta_k-\log\sum_j e^{\theta_j}\right)\right)=$$

$$-\left(\frac{\partial}{\partial\theta_i}\theta_k - \frac{\partial}{\partial\theta_i}\log\sum_j e^{\theta_j}\right) =$$

$$= -\left(0 - \frac{1}{\sum_j e^{-\theta_j}} e^{\theta_i}\right) = \hat{y}_i$$

כאשר עבור $i \neq k$

לכן, אם נרכז את כל האינדקסים נקבל:

$$\frac{\partial J}{\partial\theta} = \hat{y} - y$$

2(b)

לפי ההרכבה:
$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial \theta} \cdot \frac{\partial \theta}{\partial h} \cdot \frac{\partial h}{\partial x}$$

כאשר
$$\begin{cases} \theta = h \cdot W_2 + b_2 \\ J = CE(y, \hat{y}) \end{cases}$$

הראינו בסעיף (a) כי
$$\frac{\partial J}{\partial \theta} = \hat{y} - y$$

$$\frac{\partial \theta}{\partial h} = \frac{\partial}{\partial h}(h \cdot W_2 + b_2) = W_2^T$$

מדוע? נסתכל על W_2 כאל העתקה f ממימד
D_h למימד D_y. במילים אחרות את f
אפשר לפרק ל- $f_1, \dots, f_{D_y}$ כאשר
כל f_i מעתיקה ממימד D_h למימד 1.
במונחים של W_2, כל f_i מתאימה לעמודה i ב- W_2

$$\frac{\partial f_i}{\partial h_j} = \frac{\partial}{\partial h_j} \sum_{k=1}^{D_h} h_k \cdot W_{2,k,i} = W_{2_{j,i}}$$

כל הרכובים מתבטלים פרט ל- $k=j$.

קיבלנו איפוא את היפוך של W_2.

נותר לחשב גם $\frac{\partial h}{\partial x}$

$$h = \sigma(xW_1 + b_1)$$

נסמן

$$z = xW_1 + b_1$$

$$\frac{\partial h}{\partial x} = \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial x}$$

$$\frac{\partial h}{\partial z} = \sigma'(z) = \sigma(z) \circ (1 - \sigma(z))$$

נגזרת של פונקצי sigmoid (הראינו בתרגיל הראשון)

$$\frac{\partial z}{\partial x} = W_1^T$$

בדומה לשלב הקודם

בסופו של דבר:

$$\frac{\partial J}{\partial x} = (\hat{y} - y) \cdot W_2^T \circ \sigma(x \cdot W_1 + b_1) \circ (1 - \sigma(x \cdot W_1 + b_1)) \circ W_1^T$$