Part A

סעיף 1

$$f(x) = x^T x = \sum_{i=1}^{d} x_i^2$$

$$\nabla_x f(x) = \frac{\partial}{\partial x}(x^T x) = \frac{\partial}{\partial x}\sum_{i=1}^{d} x_i^2$$

$$\frac{\partial f(x)}{\partial (x_i)} = 2x_i \quad \forall i \Rightarrow \nabla_x f(x) = 2x$$

---

2 a_i לכל עמודה i

במטריצה A

$$f(x) = A^T x \qquad f(x) = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_n^T x \end{bmatrix}$$

$$f_i(x) = a_i^T x = \sum_{k=1}^{d} a_{ki} x_k$$

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \cdots & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & & \vdots \end{bmatrix} \qquad \frac{\partial f_i}{\partial x_j} = a_{ji}$$

$$f(x) = A^T x \rightarrow \boxed{J = A^T}$$

3

$$f(g(x)) = (A^T x)^T (A^T x) = x^T A A^T x$$

$$\nabla_x (x^T M x) = 2Mx$$ וע״פ כלל· כנלמד

כאשר M היא מטריצה סימטרית

ובפרט לנו AA^T היא מטריצה סימטרית ולכן

בס"ד

הכפלה ולכן אנחנו יכולים לומר

$$\nabla_x f(g(x)) = 2AA^T x$$

---

(4)

$$f(g(A)) = \|Ax\|_2^2 \qquad f(y) = \|y\|_2^2 \qquad \frac{\partial f}{\partial y} = 2y$$

$$\frac{\partial f(g(A))}{\partial A} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial A}$$

$$\frac{\partial f}{\partial g} = 2g = 2Ax$$

$$\frac{\partial g}{\partial A} = \frac{\partial}{\partial A}(Ax) = x^T$$

$$\frac{\partial f(g(A))}{\partial A} = 2Axx^T$$

Part B

(1) $\hat{y} = \sigma(w_2 \cdot \sigma(w_1 x + b_1) + b_2)$

---

(2) $J = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2}$$

$$\frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2} \cdot \frac{\partial z_2}{\partial b_2}$$

$$\frac{\partial J}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = \hat{y} - y$$

$$\hat{y} = \sigma(z_2) = (1+e^{-z_2})^{-1}$$

$$\frac{\partial \hat{y}}{\partial z_2} = -(1+e^{-z_2})^{-2} \cdot -e^{-z_2} = (1+e^{-z_2})^{-2} \cdot e^{-z_2}$$

$$= \sigma(z_2)(1-\sigma(z_2)) = \hat{y}(1-\hat{y})$$

$$\frac{\partial z_2}{\partial w_2} \Rightarrow z_2 = w_2 a_1 + b_2 \Rightarrow \frac{\partial z_2}{\partial w_2} = a_1^T$$

$$\frac{\partial z_2}{\partial b_2} \Rightarrow z_2 = w_2 a_1 + b_2 \Rightarrow \frac{\partial z_2}{\partial b_2} = 1$$

$$\frac{\partial J}{\partial w_2} = (\hat{y}-y)\,\hat{y}(1-\hat{y}) \cdot a_1^T = (\hat{y}-y)\,a_1^T$$

$$\frac{\partial J}{\partial b_2} = (\hat{y}-y)\,\hat{y}(1-\hat{y}) \cdot 1 = \hat{y}-y$$

---

(3) $$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$\frac{\partial z_1}{\partial w_1} = \frac{\partial(w_1 x + b)}{\partial w_1} = x^T$$

$$\frac{\partial J}{\partial z_1} = \frac{\partial J}{\partial H_1} \cdot \sigma'(z_1) \qquad \frac{\partial J}{\partial H_1} = w_2^T \frac{\partial J}{\partial z_2} \qquad \frac{\partial J}{\partial z_2} = \hat{y} - y$$

$$\frac{\partial J}{\partial w_1} = (w_2^T(\hat{y}-y)) \cdot \sigma'(z_1) \cdot x^T$$

המשך

$$\frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1}$$

$$\frac{\partial z_1}{\partial b_1} = 1$$

$$\frac{\partial J}{\partial z_1} = \frac{\partial J}{\partial h_1} \cdot \sigma'(z_1) \qquad \frac{\partial J}{\partial h_1} = W_2^T(\hat{y} - y)$$

$$\frac{\partial J}{\partial b_1} = (W_2^T(\hat{y} - y)) \cdot \sigma'(z_1)$$

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial z_1} \cdot \frac{\partial z_1}{\partial x}$$

$$\frac{\partial J}{\partial z_1} = (W_2^T(\hat{y} - y)) \cdot \sigma'(z_1) \quad (\text{חישוב} \qquad \text{מקודם})$$

$$\frac{\partial z_1}{\partial x} = W_1$$

$$\frac{\partial J}{\partial x} = W_1^T(W_2^T(\hat{y} - y)) \cdot \sigma'(z_1)$$

---

④ צריכים לשמור את המשתנים הבאים:

X: צריך אותו בשביל W_1

z_1: בשביל שנוכל לחשב את $\sigma(z_1)$

h_1: מה שנכנס לשכבה השניה (נכנס ל W_2)

z_2: מגדיר את $\hat{y}$ שמחשב ל loss

$\hat{y}$: מגדיר את ה loss $(\hat{y} - y)$

y: מגדיר את ה loss $(\hat{y} - y)$