HW 2 - Page 1

14 April 2021

1

a)

$x = \{0; 1{,}5; 3; 4.5, 6\}$

$t = \{1, 0; -1, 0, 1\}$

$\phi_j(x) = x^j$

$y(x, w) = \sum_{j=0}^{3} w_j \cdot \phi_j(x) = w_0 + w_1 \cdot x + w_2\, x^2 + w_3\, x^3$

Design Matrix . $m \times (d+1)$ → 5×4

→ # input features

↳ # examples

→ rows

↳ cols

our regression uses 3 features !

As such, we have the matrix.

$\phi_1(x) = x^1$, $\phi_3(x) = x^3$

$$\Phi = \begin{bmatrix} 1 & \phi_1(x_1) & \phi_2(x_1) & \phi_3(x_1) \\ 1 & \phi_1(x_2) & \phi_2(x_2) & \phi_3(x_2) \\ 1 & \phi_1(x_3) & \phi_2(x_3) & \phi_3(x_3) \\ 1 & \phi_1(x_4) & \phi_2(x_4) & \phi_3(x_4) \\ 1 & \phi_1(x_5) & \phi_2(x_5) & \phi_3(x_5) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1{,}5 & 2.25 & 3.375 \\ 1 & 3 & 9 & 27 \\ 1 & 4{,}5 & 20.25 & 91.125 \\ 1 & 6 & 36 & 216 \end{bmatrix}$$

↳ $\phi_2(x) = x^2$

b)

With our Design matrix & our Target vector

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1{,}5 & 2.25 & 3.375 \\ 1 & 3 & 9 & 27 \\ 1 & 4{,}5 & 20.25 & 91.125 \\ 1 & 6 & 36 & 216 \end{bmatrix} \qquad Y = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

We want the weight vector $w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}$, which minimizes the sum of squared errors.

Knowing that $W = (X^T X)^{-1} X^T Y$, we have:

$$W = \left[\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1{,}5 & 3 & 4{,}5 & 6 \\ 0 & 2{,}25 & 9 & 20{,}25 & 36 \\ 0 & 3{,}375 & 27 & 91{,}125 & 216 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1{,}5 & 2.25 & 3.375 \\ 1 & 3 & 9 & 27 \\ 1 & 4{,}5 & 20.25 & 91.125 \\ 1 & 6 & 36 & 216 \end{bmatrix} \right]^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1{,}5 & 3 & 4{,}5 & 6 \\ 0 & 2{,}25 & 9 & 20{,}25 & 36 \\ 0 & 3{,}375 & 27 & 91{,}125 & 216 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 15 & 67{,}5 & 337{,}5 \\ 15 & 67{,}5 & 337{,}5 & 1792.125 \\ 67{,}5 & 337{,}5 & 1792{,}125 & 9871{,}835 \\ 337{,}5 & 1792{,}125 & 9871{,}835 & 55700.156 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1{,}5 & 3 & 4{,}5 & 6 \\ 0 & 2{,}25 & 9 & 20{,}25 & 36 \\ 0 & 3{,}375 & 27 & 91{,}125 & 216 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

HW 2 - Page 2

14 April 2021

1

b (cont.))

$$= \begin{bmatrix} 0.99 & 0.06 & -0.09 & 0.06 & -0.01 \\ -1.00 & 1.08 & 0.38 & -0.70 & 0.23 \\ 0.29 & -0.49 & -0.06 & 0.41 & -0.16 \\ -0.02 & 0.05 & 0 & -0.05 & 0.02 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

$$\Leftrightarrow \quad w = \begin{bmatrix} 1.057 \\ -1.143 \\ 0.190 \\ 0 \end{bmatrix} \approx \begin{bmatrix} 1.06 \\ -1.14 \\ 0.19 \\ 0 \end{bmatrix} //$$

c) To compute the gradient of $E(w) = \frac{1}{2} \sum_{k=1}^{M} (t_k - w^T \phi_k)^2 + \frac{\lambda}{2} \|w\|_2^2$ we have.

$$\frac{\delta E(w)}{\delta w} = \frac{\delta \left[\frac{1}{2} \sum_{k=1}^{M} (t_k - w^T \phi_k)^2 + \frac{\lambda}{2} \|w\|_2^2 \right]}{\delta w} = \frac{\delta \left[\frac{1}{2} \sum_{k=1}^{M} (t_k - w^T \phi_k)^2 \right]}{\delta w} + \frac{\delta \left[\frac{\lambda}{2} \|w\|_2^2 \right]}{\delta w}$$

$$= \frac{1}{2} \underbrace{\frac{\delta \left[\sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 \right]}{\delta w}}_{A} + \frac{\lambda}{2} \underbrace{\frac{\delta \left[\|w\|_2^2 \right]}{\delta w}}_{B}$$

Dealing with A & B individually we have.

A $\frac{\delta \left[\sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 \right]}{\delta w} \rightarrow \sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 = (T - Xw)^T (T - Xw)$

$$= \frac{\delta (T - Xw)^T (T - Xw)}{\delta w}$$

$$= \left(\frac{\delta}{\delta w} (T - Xw)^T \right) (T - Xw) + (T - Xw)^T \left(\frac{\delta}{\delta w} (T - Xw) \right)$$

$$= (-X^T)(T - Xw) + (T - Xw)^T (-X)$$

$$= -2X^T (T - Xw)$$

HW 2 - Page 3

14 April 2021

$$\frac{\delta(w^T \cdot A w)}{\delta w} = (\underbrace{A + A^T}_{I})\, w$$

1

C (cont.))

$$\beta) \quad \frac{\delta\left[\|w\|_2^2\right]}{\delta w} = \frac{\delta\left[\overbrace{\sum^m w_i^2}^{\sum w_i^2 = w^T \cdot w}\right]}{\delta w} = \frac{\delta\left[(w^T w)\right]}{\delta w}$$

$$= \frac{\delta\left[(w^T \cdot \overbrace{I w}^{= w})\right]}{\delta w} = (I + I^T)\, w = 2 I w = 2w$$

Joining things together we have

$$\frac{\delta E(w)}{\delta(w)} = 1/2 \cdot \left(-2 X^T (T - X w)\right) + \frac{\lambda}{2}(2w)$$

$$= -X^T (T - X w) + \lambda w$$

d) $E(w) = 1/2 \sum_{m=1}^{N} (t_m - w^T \cdot x_m)^2 + \frac{\lambda}{2} \|w\|_2^2$

(Ridge Regression.

$$E(w) = 1/2 \sum_{m=1}^{N} (t_m - w^T \cdot x_m)^2 + \|\underbrace{\Gamma}_{I\frac{\lambda}{\sqrt{2}}} \cdot w\|_2^2$$

$$= 1/2 \sum_{m=1}^{N} (t_m - w^T \cdot x_m)^2 + I\left(\frac{\lambda}{\sqrt{2}}\right)^2 \|w\|_2^2$$

$$= 1/2 \sum_{m=1}^{N} (t_m - w^T \cdot x_m)^2 + I\frac{\lambda^2}{2} \|w\|_2^2 \quad)$$

We have $X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1{,}5 & 2{,}25 & 3\,375 \\ 1 & 3 & 9 & 27 \\ 1 & 4{,}5 & 20{,}25 & 91{,}125 \\ 1 & 6 & 36 & 216 \end{bmatrix}$ $\quad Y = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$

$$\log\left(\frac{\lambda}{2}\right) = 0 \Leftrightarrow \frac{\lambda}{2} = 10^0 \Leftrightarrow \frac{\lambda}{2} = 1 \Leftrightarrow \lambda = 2$$

from c)

$$\frac{\delta E(w)}{\delta w} = 0 \Leftrightarrow -X^T (T - X w) + \lambda w = 0 \Leftrightarrow -X^T T + X^T X w + \lambda w = 0$$

$$\Leftrightarrow X^T X w + \lambda w = X^T T \Leftrightarrow (X^T X + I\lambda) w = X^T T$$

$$\Leftrightarrow w = (X^T X + I\lambda)^{-1} X^T T$$

14 April 2021

1

d (cont.))

$$w = (X^T X + I\lambda)^{-1} X^T T = \left[\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1.5 & 3 & 4.5 & 6 \\ 0 & 2.25 & 9 & 20.25 & 36 \\ 0 & 3.375 & 27 & 91.125 & 216 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1.5 & 2.25 & 3.375 \\ 1 & 3 & 9 & 27 \\ 1 & 4.5 & 20.25 & 91.125 \\ 1 & 6 & 36 & 216 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot 2 \right]^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1.5 & 3 & 4.5 & 6 \\ 0 & 2.25 & 9 & 20.25 & 36 \\ 0 & 3.375 & 27 & 91.125 & 216 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \left[\begin{bmatrix} 5 & 15 & 67.5 & 337.5 \\ 15 & 67.5 & 337.5 & 1792.125 \\ 67.5 & 337.5 & 1792.125 & 9871.835 \\ 337.5 & 1792.125 & 9871.835 & 55700.156 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \right]^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1.5 & 3 & 4.5 & 6 \\ 0 & 2.25 & 9 & 20.25 & 36 \\ 0 & 3.375 & 27 & 91.125 & 216 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.29 & -0.07 & -0.02 & 0.00 \\ -0.07 & 0.35 & -0.15 & 0.01 \\ -0.02 & -0.15 & 0.09 & -0.01 \\ 0.00 & 0.01 & -0.01 & 0.00 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1.5 & 3 & 4.5 & 6 \\ 0 & 2.25 & 9 & 20.25 & 36 \\ 0 & 3.375 & 27 & 91.125 & 216 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \Leftrightarrow w = \begin{bmatrix} 0.295 \\ -0.132 \\ -0.138 \\ 0.030 \end{bmatrix}$$

e)

To compute the gradient of $E(w) = \frac{1}{2} \sum_{k=1}^{M} (t_k - w^T \phi_k)^2 + \lambda \|w\|_1$ we have.

$$\frac{\delta E(w)}{\delta w} = \frac{\delta \left[\frac{1}{2} \sum_{k=1}^{M} (t_k - w^T \phi_k)^2 + \lambda \|w\|_1 \right]}{\delta w} = \frac{\delta \left[\frac{1}{2} \sum_{k=1}^{M} (t_k - w^T \phi_k)^2 \right]}{\delta w} + \frac{\delta [\lambda \cdot \|w\|_1]}{\delta w}$$

$$= \frac{1}{2} \underbrace{\frac{\delta \left[\sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 \right]}{\delta w}}_{A} + \lambda \underbrace{\frac{\delta [\|w\|_1]}{\delta w}}_{B}$$

Dealing with A & B individually we have.

(A) $\dfrac{\delta \left[\sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 \right]}{\delta w} \rightarrow \sum_{k=1}^{M} (t_k - w^T \cdot \phi_k)^2 = (T - Xw)^T (T - Xw)$

$$= \frac{\delta (T - Xw)^T (T - Xw)}{\delta w}$$

$$= \left(\frac{\delta}{\delta w} (T - Xw)^T \right) (T - Xw) + (T - Xw)^T \left(\frac{\delta}{\delta w} (T - Xw) \right)$$

$$= (-X^T)(T - Xw) + (T - Xw)^T (-X)$$

$$= -2X^T (T - Xw)$$

HW 2 - Page 5

19 April 2021

1

e (cont.))

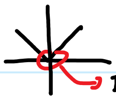
(β) $$\frac{\delta[\|w\|]}{\delta w} = \frac{\delta\left[\sum_{i}^{m} |w_i|\right]}{\delta w} \rightarrow \frac{\delta[1^T |w|]}{\delta w}$$

row vector of 1's

$$= 1^T \cdot \frac{\delta[|w|]}{\delta w}$$

which is not differentiable!

$f(w) = |w| \rightarrow$



Joining things together we have

$$\frac{\delta E(w)}{\delta(w)} = 1/2 \cdot (-2X^T(T - Xw)) + \lambda \cdot 1^T \frac{\delta[|w|]}{\delta w} //$$

f) The LASSO regression lacks a closed form solution. By definition, a closed form solution is one in which there are no limit, **differentiation** or integration. To perform this regularization we would have to perform $\frac{\delta E(w)}{\delta(w)} = 0$ where $E(w) = 1/2 \sum_{n=1}^{N} (t_n - w_t \cdot \phi_n)^2 + \lambda \|w\|_1$. From exercise 1.e), we've already seen that when computing the gradient of E(w), we get a formula that still maintains a differentiation (due to $\|w\|_1$ not being differentiable). As such, LASSO regression, does not possess a closed form solution.

2

a) $w^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ $b^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ // $w^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $b^{[2]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ // $w^{[3]} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ $b^{[3]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x) \qquad \eta = 0.1 \qquad x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \; t = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

1 Forward Propagation

- $z^{[1]} = w^{[1]} x^{[0]} + b^{[1]} \Leftrightarrow \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix}$

element wise

$$x^{[1]} = f(z^{[1]}) = \begin{bmatrix} \tanh(5) \\ \tanh(5) \\ \tanh(5) \end{bmatrix} \simeq \begin{bmatrix} 0.9999 \\ 0.9999 \\ 0.9999 \end{bmatrix}$$

HW 2 - Page 6

19 April 2021

2

a (cont.))

- $$z^{[2]} = W^{[2]} x^{[1]} + b^{[2]} \Leftrightarrow \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0\,9999 \\ 0\,9999 \\ 0\,9999 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2\,9997 \\ 2\,9997 \end{bmatrix}$$

$$x^{[2]} = f(z^{[2]}) = \begin{bmatrix} \tanh(2\,9997) \\ \tanh(2\,9997) \end{bmatrix} = \begin{bmatrix} 0\,9951 \\ 0\,9951 \end{bmatrix}$$

- $$z^{[3]} = W^{[3]} x^{[2]} + b^{[3]} \Leftrightarrow \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0,9951 \\ 0,9951 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2\,9902 \\ 2\,9902 \end{bmatrix}$$

$$x^{[3]} = f(z^{[3]}) = \begin{bmatrix} \tanh(2,9902) \\ \tanh(2\,9902) \end{bmatrix} = \begin{bmatrix} 0\,995 \\ 0\,995 \end{bmatrix}$$

2 Back Propagation

Squared Error Loss $E(t, x) = \frac{1}{2}(x - t)^2$

$$\rightarrow \frac{\delta E}{\delta x^{[L]}}(t, x^{[L]}) = \frac{\delta E}{\delta (x^{[L]}-t)^2} \cdot \frac{\delta (x^{[L]}-t)^2}{\delta (x^{[L]}-t)} \cdot \frac{\delta (x^{[L]}-t)}{\delta x^{[L]}} = \frac{1}{2}\left[2(x^{[L]}-t)\right] = x^{[L]} - t$$

$$\rightarrow \frac{\delta x^{[l]}}{\delta z^{[l]}}(z^{(l)}) = 1 - \tanh(z^{[l]})^2$$

$$\rightarrow \frac{\delta z^{[l]}}{\delta W^{[l]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = x^{[l-1]}$$

$$\rightarrow \frac{\delta z^{[l]}}{\delta b^{[l]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = 1$$

$$\rightarrow \frac{\delta z^{[l]}}{\delta x^{[l-1]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = W^{[l]}$$

Let us begin the recursion.

- $$\delta^{[3]} = \frac{\delta E}{\delta x^{[3]}} \circ \frac{\delta x^{[3]}}{\delta z^{[3]}} = (x^{[3]} - t) \circ (1 - \tanh(z^{[3]})^2) = \left(\begin{bmatrix} 0\,995 \\ 0\,995 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \end{bmatrix}\right) \circ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} \tanh(2,9902)^2 \\ \tanh(2,9902)^2 \end{bmatrix}\right)$$

$$= \begin{bmatrix} -0\,005 \\ 1\,995 \end{bmatrix} \circ \begin{bmatrix} 0\,01 \\ 0\,01 \end{bmatrix} = \begin{bmatrix} -0.00005 \\ 0\,01995 \end{bmatrix} \approx \begin{bmatrix} 0 \\ 0\,02 \end{bmatrix}$$

- $$\delta^{[2]} = \frac{\delta z^{[3]}}{\delta x^{[2]}}^T \cdot \delta^{[3]} \circ \frac{\delta x^{[2]}}{\delta z^{[2]}} = (W^{[3]})^T \cdot \delta^{[3]} \circ (1 - \tanh(z^{[2]})^2) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}^T \cdot \begin{bmatrix} 0 \\ 0\,02 \end{bmatrix} \circ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \tanh(z^{[2]})^2\right)$$

$$= \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{bmatrix} 0 \\ 0\,02 \end{bmatrix} \circ \left(\begin{bmatrix} 0\,005 \\ 0\,005 \end{bmatrix}\right) = \begin{bmatrix} 0\,0001 \\ 0\,0001 \end{bmatrix}$$

- $$\delta^{[1]} = \frac{\delta z^{[2]}}{\delta x^{[1]}}^T \cdot \delta^{[2]} \circ \frac{\delta x^{[1]}}{\delta z^{[1]}} = (W^{[2]})^T \delta^{[2]} \circ (1 - \tanh(z^{[1]})^2) = \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}^T \cdot \begin{bmatrix} 0\,0001 \\ 0\,0001 \end{bmatrix}\right) \circ \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \tanh(z^{[1]})^2\right)$$

$$= \begin{bmatrix} 0.0002 \\ 0\,0002 \\ 0\,0002 \end{bmatrix} \circ \begin{bmatrix} 0\,00018 \\ 0\,00018 \\ 0\,00018 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

HW 2 - Page 7

19 April 2021

2

a (cont.)) → Update

- $W^{[1]} = W^{[1]} - \eta \frac{\delta E}{\delta W^{[1]}} = W^{[1]} - 0.1 \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Leftrightarrow \boxed{W^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}}$ ✓

  $\frac{\delta E}{\delta W^{[1]}} = \delta^{[1]} \frac{\delta z^{[1]}}{\delta W^{[1]}}^T = \delta^{[1]} \cdot (x^{[0]})^T = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

- $b^{[1]} = b^{[1]} - \eta \frac{\delta E}{\delta b^{[1]}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \boxed{b^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}$ ✓

  $\frac{\delta E}{\delta b^{[1]}} = \delta^{[1]} \cdot \frac{\delta z^{[1]}}{\delta b^{[1]}}^T = \delta^{[1]}$

- $W^{[2]} = W^{[2]} - \eta \frac{\delta E}{\delta W^{[2]}} = W^{[2]} - \eta \left(\delta^{[2]} x^{[1]T} \right) = W^{[2]} - \eta \left(\begin{bmatrix} 0.0001 \\ 0.0001 \end{bmatrix} \cdot \begin{bmatrix} 0.9999 \\ 0.9999 \end{bmatrix}^T \right)$

  $= W^{[2]} - 0.1 \begin{bmatrix} 0.0001 & 0.0001 & 0.0001 \\ 0.0001 & 0.0001 & 0.0001 \end{bmatrix} \Leftrightarrow \boxed{W^{[2]} \approx \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}}$ ✓

- $b^{[2]} = b^{[2]} - \eta \frac{\delta E}{\delta b^{[2]}} = b^{[2]} - \eta \delta^{[2]} = b^{[2]} - 0.1 \begin{bmatrix} 0.0001 \\ 0.0001 \end{bmatrix} \Leftrightarrow \boxed{b^{[2]} \approx \begin{bmatrix} 0 \\ 0 \end{bmatrix}}$ ✓

- $W^{[3]} = W^{[3]} - \eta \frac{\delta E}{\delta W^{[3]}} = W^{[3]} - \eta \left(\delta^{[3]} x^{[2]T} \right) = W^{[3]} - \eta \left(\begin{bmatrix} 0 \\ 0.02 \end{bmatrix} \begin{bmatrix} 0.9951 \\ 0.9951 \end{bmatrix}^T \right)$

  $= W^{[3]} - 0.1 \begin{bmatrix} 0 & 0 \\ 0.02 & 0.02 \end{bmatrix} = \boxed{W^{[3]} = \begin{bmatrix} 1 & 1 \\ 0.998 & 0.998 \end{bmatrix}}$ ✓

- $b^{[3]} = b^{[3]} - \eta \frac{\delta E}{\delta b^{[3]}} = b^{[3]} - \eta \delta^{[3]} = b^{[3]} - 0.1 \begin{bmatrix} 0 \\ 0.02 \end{bmatrix} \Leftrightarrow \boxed{b^{[3]} = \begin{bmatrix} 1 \\ 0.998 \end{bmatrix}}$ ✓

b) Cross-Entropy loss measures the performance of a classification model whose output is a probability value between 0 and 1. It aims to minimize the distance between two **probability distributions**, making it a good loss function for classification problems. Looking at our target vector given in this exercise, we can see that it constitutes of [1, -1]^T, and as such, it clearly does not fall under this category. For Cross-Entropy Loss to be favourable, our outcome vector would have to have all its values between 0 and 1, and they would need to sum up to a total of 1 (making them probabilistic values).

c) $W^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad b^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad // \quad W^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad b^{[2]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad // \quad W^{[3]} = \begin{bmatrix} 1 & 1 \\ 0.998 & 0.998 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 0.998 \end{bmatrix}$

$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \tanh(x) \qquad \eta = 0.1 \qquad x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \; t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$\text{Softmax}(z) = x \rightarrow x_i = \frac{e^{z_i}}{\sum_{k=1}^{2} e^{z_k}}$

19 April 2021

2

C (cont.))

1 Find derivative of a given x_i with respect to z_i (SoftMax)

$$\frac{\delta x_i}{\delta z_j} = \frac{\delta}{\delta z_j} \frac{e^{z_i}}{\sum_{k=1}^{d} e^{z_k}}$$

- For $i=j$: $\frac{\delta x_i}{\delta z_j} = \frac{\delta}{\delta z_j} \frac{e^{z_i}}{\sum_{k=1}^{d} e^{z_k}} = \left(\frac{\delta}{\delta z_j} e^{z_i}\right)\left(\sum_{k=1}^{d} e^{z_k}\right) - \left(e^{z_i}\right)\left(\frac{\delta}{\delta z_j} \sum_{k=1}^{d} e^{z_k}\right)$

$$= \frac{e^{z_i} \sum_{k=1}^{d} e^{z_k} - e^{z_i} e^{z_j}}{\left(\sum_{k=1}^{d} e^{z_k}\right)^2} = \frac{e^{z_i}\left(\sum_{k=1}^{d} e^{z_k} - e^{z_j}\right)}{\left(\sum_{k=1}^{d} e^{z_k}\right)^2} = \frac{e^{z_i}}{\sum_{k=1}^{d} e^{z_k}} \left(\frac{\sum_{k=1}^{d} e^{z_k}}{\sum_{k=1}^{d} e^{z_k}} - \frac{e^{z_j}}{\sum_{k=1}^{d} e^{z_k}}\right)$$

$$= x_i(1-x_j) = x_i(1-x_i)$$

- For $i \neq j$: $\frac{\delta x_i}{\delta z_j} = \frac{\delta}{\delta z_j} \frac{e^{z_i}}{\sum_{k=1}^{d} e^{z_k}} = e^{z_i} \frac{\delta}{\delta z_j} \frac{1}{\sum_{k=1}^{d} e^{z_k}} = e^{z_i}\left(-\frac{1}{\left(\sum_{k=1}^{d} e^{z_k}\right)^2}\right) e^{z_j}$

$$= -x_i x_j$$

2 Forward Propagation

- $z^{[1]} = W^{[1]} x^{[0]} + b^{[1]} \Leftrightarrow \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix}$
- $x^{[1]} = \underbrace{f}_{\text{element wise}}(z^{[1]}) = \begin{bmatrix} \tanh(5) \\ \tanh(5) \\ \tanh(5) \end{bmatrix} \simeq \begin{bmatrix} 0.9999 \\ 0.9999 \\ 0.9999 \end{bmatrix}$
- $z^{[2]} = W^{[2]} x^{[1]} + b^{[2]} \Leftrightarrow \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.9999 \\ 0.9999 \\ 0.9999 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2.9997 \\ 2.9997 \end{bmatrix}$
- $x^{[2]} = f(z^{[2]}) = \begin{bmatrix} \tanh(2.9997) \\ \tanh(2.997) \end{bmatrix} = \begin{bmatrix} 0.9950 \\ 0.9950 \end{bmatrix}$
- $z^{[3]} = W^{[3]} x^{[2]} + b^{[3]} \Leftrightarrow \begin{bmatrix} 1 & 1 \\ 0.998 & 0.998 \end{bmatrix} \begin{bmatrix} 0.995 \\ 0.995 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.998 \end{bmatrix} = \begin{bmatrix} 2.990 \\ 2.984 \end{bmatrix}$
- $x^{[3]} = \text{Softmax}(z) = \begin{bmatrix} \frac{e^{2.990}}{e^{2.990} + e^{2.984}} \\ \frac{e^{2.984}}{e^{2.990} + e^{2.984}} \end{bmatrix} = \begin{bmatrix} 0.5015 \\ 0.4985 \end{bmatrix}$

19 April 2021

2

C (cont.) 2 Back Propagation

Cross-Entropy loss $E(t,x) = \sum_{i=1}^{d} t_i \log x_i$

$$\rightarrow \frac{\delta E}{\delta z_i}(t, x^{[3]}) = \delta_i^{[3]} = \frac{\delta}{\delta z_i}\left(-\sum_{k=1}^{d} t_k \log x_k^{[3]}\right) = -\sum_{k=1}^{d} t_k \frac{\delta}{\delta z_i} \log x_k^{[3]}$$

$$= -\sum_{k=1}^{d} t_k \frac{1}{x_k^{[3]}} \frac{\delta x_k^{[3]}}{\delta z_i} = -\sum_{k=1} t_k \frac{1}{x_k^{[3]}} \frac{\delta x_k^{[3]}}{\delta z_i} - \sum_{k \neq i} t_k \frac{1}{x_k^{[3]}} \frac{\delta x_k^{[3]}}{\delta z_i}$$

$$= -t_i \frac{1}{x_i^{[3]}}\left(x_i^{[3]}(1 - x_i^{[3]})\right) - \sum_{k \neq i} t_k \frac{1}{x_k^{[3]}}\left(-x_k^{[3]} x_i^{[3]}\right)$$

$$= -t_i + t_i x_i^{[3]} + \sum_{k \neq i} t_k x_i^{[3]} = -t_i + x_i^{[3]}\left(t_i + \sum_{k \neq i} t_k\right) = -t_i + x_i^{[3]}\left(\sum_{k=1}^{d} t_k\right)$$

$$= x_i^{[3]} - t_i$$

$$\rightarrow \frac{\delta x^{[l]}}{\delta z^{[l]}}(z^{(l)}) = 1 - \tanh(z^{[l]})^2$$

$$\rightarrow \frac{\delta z^{[l]}}{\delta W^{[l]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = x^{[l-1]}$$

$$\rightarrow \frac{\delta z^{[l]}}{\delta b^{[l]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = 1$$

$$\rightarrow \frac{\delta z^{[l]}}{\delta x^{[l-1]}}(W^{[l]}, b^{[l]}, x^{[l-1]}) = W^{[l]}$$

To start the recursion we need the $\delta^{[3]}$

$$\delta^{[3]} = \begin{bmatrix} \delta_1^{[3]} \\ \delta_2^{[3]} \end{bmatrix} = \begin{bmatrix} x_1^{[3]} - t_1 \\ x_2^{[3]} - t_2 \end{bmatrix} = \begin{bmatrix} 0.5015 \\ 0.4985 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.4985 \\ 0.4985 \end{bmatrix}$$

Now we can compute the remaining δ

- $$\delta^{[2]} = \frac{\delta z^{[3]}}{\delta x^{[2]}}^T \cdot \delta^{[3]} \circ \frac{\delta x^{[2]}}{\delta z^{[2]}} = (W^{[3]})^T \cdot \delta^{[3]} \circ (1 - \tanh(z^{[2]})^2)$$

$$= \begin{bmatrix} 1 & 0.998 \\ 1 & 0.998 \end{bmatrix} \begin{bmatrix} -0.4985 \\ 0.4985 \end{bmatrix} \circ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{pmatrix} \tanh(0.9950)^2 \\ \tanh(0.9950)^2 \end{pmatrix}\right) = \begin{bmatrix} -0.0004 \\ -0.0004 \end{bmatrix}$$

- $$\delta^{[1]} = \frac{\delta z^{[2]}}{\delta x^{[1]}}^T \delta^{[2]} \circ \frac{\delta x^{[1]}}{\delta z^{[1]}} = (W^{[2]})^T \cdot \delta^{[2]} \circ (1 - \tanh(z^{[1]})^2)$$

$$= \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}^T \cdot \begin{bmatrix} -0.0004 \\ -0.0004 \end{bmatrix} \circ \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} \tanh(s)^2 \\ \tanh(s)^2 \\ \tanh(s)^2 \end{bmatrix}\right) \simeq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

HW 2 - Page 9

19 April 2021

2

C (cont.)) 4 Update

- $$w^{[1]} = w^{[1]} - \eta \frac{\delta E}{\delta w^{[1]}} = w^{[1]} - \eta \left(\delta^{[1]} \frac{\delta z^{[1]}}{\delta w^{[1]}}^{T} \right)$$

$$= w^{[1]} - \eta \left(\delta^{[1]} x^{[0]^{T}} \right) = w^{[1]} - \eta \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \right)$$

$$= w^{[1]} - \eta \left(\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right) \Leftrightarrow \boxed{w^{[1]} = w^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}}$$

- $$b^{[1]} = b^{[1]} - \eta \frac{\delta E}{\delta b^{[1]}} = b^{[1]} - \eta \, \delta^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Leftrightarrow \boxed{b^{[1]} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}$$

- $$w^{[2]} = w^{[2]} - \eta \frac{\delta E}{\delta w^{[2]}} = w^{[2]} - \eta \left(\delta^{[2]} \frac{\delta z^{[2]}}{\delta w^{[2]}}^{T} \right)$$

$$= w^{[2]} - \eta \left(\delta^{[2]} x^{[1]^{T}} \right) = w^{[2]} - \eta \left(\begin{bmatrix} -0.0004 \\ -0.0004 \end{bmatrix} \begin{bmatrix} 0.9999 \\ 0.9999 \\ 0.9999 \end{bmatrix}^{T} \right)$$

$$= w^{[2]} - 0.1 \begin{bmatrix} -0.0004 & -0.0004 & -0.0004 \\ -0.0004 & -0.0004 & -0.0004 \end{bmatrix} \Leftrightarrow \boxed{w^{[2]} = \begin{bmatrix} 1.0004 & 1.0004 & 1.0004 \\ 1.0004 & 1.0004 & 1.0004 \end{bmatrix}}$$

- $$b^{[2]} = b^{[2]} - \eta \frac{\delta E}{\delta b^{[2]}} = b^{[2]} - \eta \, \delta^{[2]} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -0.0004 \\ -0.0004 \end{bmatrix} \Leftrightarrow \boxed{b^{[2]} = \begin{bmatrix} 0.00004 \\ 0.00004 \end{bmatrix}}$$

- $$w^{[3]} = w^{[3]} - \eta \frac{\delta E}{\delta w^{[3]}} = w^{[3]} - \eta \left(\delta^{[3]} \frac{\delta z^{[3]}}{\delta w^{[3]}}^{T} \right) = w^{[3]} - \eta \left(\delta^{[3]} \cdot x^{[2]^{T}} \right)$$

$$= w^{[3]} - \eta \left(\begin{bmatrix} -0.4985 \\ 0.4985 \end{bmatrix} \cdot \begin{bmatrix} 0.9950 \\ 0.9950 \end{bmatrix}^{T} \right) = \begin{bmatrix} 1 & 1 \\ 0.998 & 0.998 \end{bmatrix} - 0.1 \begin{bmatrix} -0.4960 & -0.4960 \\ 0.4960 & 0.4960 \end{bmatrix}$$

$$\Leftrightarrow \boxed{w^{[3]} = \begin{bmatrix} 1.050 & 1.050 \\ 0.9484 & 0.9484 \end{bmatrix}}$$

HW 2 - Page 10

19 April 2021

2

C (cont.)

- $b^{[3]} = b^{[3]} - \eta \frac{\delta E}{\delta b^{[3]}} = b^{[3]} - \eta\ \delta^{[3]} = \begin{bmatrix} 1 \\ 0.998 \end{bmatrix} - 0.1 \begin{bmatrix} -0.4985 \\ 0.4965 \end{bmatrix} \Leftrightarrow \boxed{b^{[3]} = \begin{bmatrix} 1.04985 \\ 0.94815 \end{bmatrix}}$