Q. Derive the backprop algorithm for the neural network + loss defined by

$$\tilde{J} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} -y_{ik} \log a^{(3)}_{ik} - (1-y_{ik}) \log (1-a^{(3)}_{ik}).$$

where $a_i^{(3)}$ is a k-dim vector of activations for the ith input. Omitting index i, recall that

$$a^{(3)} = \varphi(z^{(3)})$$
$$z^{(3)} = W^{(2)} a^{(2)} + b^{(2)}$$
$$a^{(2)} = \varphi(z^{(2)}).$$
$$z^{(2)} = W^{(1)} a^{(1)} + b^{(1)}.$$
$$a^{(1)} = X, \text{ input.}$$

We need to compute the gradients

$$\left(\frac{\partial \tilde{J}}{\partial W^{(2)}_{\alpha\beta}}, \frac{\partial \tilde{J}}{\partial b^{(2)}_{\alpha}}, \frac{\partial \tilde{J}}{\partial W^{(3)}_{\alpha\beta}}, \frac{\partial \tilde{J}}{\partial b^{(3)}_{\alpha}} \right) \quad \text{for every } \alpha, \beta.$$

For simplicity, we first compute the gradients wrt

$$J^i = \sum_{k=1}^{K} - y_{ik} \log a^{(3)}_{ik} - (1-y_{ik}) \log (1-a^{(3)}_{ik}).$$

for a single input i. From this we can retrieve the gradients wrt $\tilde{J}$ via

$$\frac{\partial \tilde{J}}{\partial W^{(2)}_{\alpha\beta}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial J^i}{\partial W^{(2)}_{\alpha\beta}} + \lambda W^{(2)}_{\alpha\beta} \quad \text{etc.} \quad (*)$$

(due to a.2d)

For simplicity we drop the index i for the computation of $\frac{\partial \hat{J}^{[i]}}{\partial W^{(2)}_{\alpha\beta}} = \frac{\partial \hat{J}^{i}}{\partial W^{(2)}_{\alpha\beta}}$ etc.

First compute

$$\eta^{(3)}_\alpha := \frac{\partial \hat{J}}{\partial a^{(3)}_\alpha} = -\frac{y_\alpha}{a^{(3)}_\alpha} + \frac{1-y_\alpha}{1-a^{(3)}_\alpha} \quad \text{(due to Q.2c)}.$$

$\forall \alpha$

Then, compute via chain rule.

$$\delta^{(3)}_\alpha := \frac{\partial \hat{J}}{\partial z^{(3)}_\alpha} = \frac{\partial \hat{J}}{\partial a^{(3)}_\alpha} \frac{\partial a^{(3)}_\alpha}{\partial z^{(3)}_\alpha}.$$

$$= \eta^{(3)}_\alpha \varphi(z^{(3)}_\alpha)(1-\varphi(z^{(3)}_\alpha)). \quad \text{(due to Q.2b)}$$

$\forall \alpha$

$\Rightarrow$ lines of Alg1.

In vectorial notation,

$$\delta^{(3)} = \eta^{(3)} \cdot \varphi(z^{(3)}) \cdot (1-\varphi(z^{(3)})).$$

pointwise multiplications. $\Rightarrow$ line 5 of Alg1.

We're ready to compute via multivar chain rule:

$$\partial W^{(2)}_{\alpha\beta} := \frac{\partial \hat{J}}{\partial W^{(2)}_{\alpha\beta}} = \sum_\gamma \frac{\partial \hat{J}}{\partial z^{(3)}_\gamma} \frac{\partial z^{(3)}_\gamma}{\partial W^{(2)}_{\alpha\beta}}$$

$$= \sum_\gamma \delta^{(3)}_\gamma \cdot \frac{\partial z^{(3)}_\gamma}{\partial W^{(2)}_{\alpha\beta}}.$$

Note that $z_\gamma^{(3)} = \sum_\tau W^{(2)}_{\gamma\tau} a^{(2)}_\tau + b^{(2)}_\gamma$ and so

$$\frac{\partial z^{(3)}_\gamma}{\partial W^{(2)}_{\alpha\beta}} = \sum_\tau \frac{\partial W^{(2)}_{\gamma\tau}}{\partial W^{(2)}_{\alpha\beta}} a^{(2)}_\tau$$

$$= \sum_\tau \delta_{\alpha\gamma} \delta_{\tau\beta} a^{(2)}_\tau$$

$$= \delta_{\alpha\gamma} a^{(2)}_\beta .$$

$$= \sum_\gamma \delta^{(3)}_\gamma \delta_{\alpha\gamma} a^{(2)}_\beta .$$

$$= \delta^{(3)}_\alpha a^{(2)}_\beta .$$

In vectorial notation,

$$\partial W^{(3)} = \delta^{(3)} a^{(2)T}, \text{ an outer product.}$$

⇐ line 11 of Alg 1.

and similarly,

$$\partial b_\alpha^{(3)} := \sum_\gamma \frac{\partial \hat{J}}{\partial z_\gamma^{(3)}} \frac{\partial z_\gamma^{(3)}}{\partial b_\alpha^{(2)}}$$

$$= \sum_\gamma \delta_\gamma^{(3)} \delta_{\alpha\gamma} .$$

$$= \delta_\alpha^{(3)} . \qquad \Leftarrow \text{line 13 of Algl.}$$

Now we compute $\eta^{(2)}$ and $\delta^{(2)}$ to recursively obtain $\partial W^{(1)}$ and $\partial b^{(1)}$.

$$\eta_\alpha^{(2)} := \frac{\partial \hat{J}}{\partial a_\alpha^{(2)}} = \sum_\gamma \frac{\partial \hat{J}}{\partial z_\gamma^{(3)}} \frac{\partial z_\gamma^{(3)}}{\partial a_\alpha^{(2)}} .$$

$$= \sum_\gamma \delta_\gamma^{(3)} \cdot \frac{\partial z_\gamma^{(3)}}{\partial a_\alpha^{(2)}} .$$

$$\left[\text{Since } z_\gamma^{(3)} = \sum_\tau W_{\gamma\tau}^{(2)} a_\tau^{(2)} + b_\gamma^{(2)} , \right.$$

$$\left. \frac{\partial z_\gamma^{(3)}}{\partial a_\alpha^{(2)}} = \sum_\tau W_{\gamma\tau}^{(2)} \delta_{\alpha\tau} = W_{\gamma\alpha}^{(2)} . \right]$$

$$= \sum_\gamma \delta_\gamma^{(3)} W_{\gamma\alpha}^{(2)}$$

In vectorial notations,

$$\eta^{(2)} = W^{(2)T} \delta^{(3)} . \qquad \Leftarrow \text{line 15 of Algl.}$$

It easily follows that

$$\delta^{(2)} = \eta^{(2)} \cdot \varphi(z^{(2)}) \cdot (1 - \varphi(z^{(2)})).$$

⇐ line 17 of Alg 1.

And analogous computations lead to

$$\partial W^{(1)} \text{ and } \partial b^{(1)}.$$

We have thus derived the derivative wrt $\hat{J}$ and from (*) follows the derivative wrt $\tilde{J}$. ∎

⇐ $\frac{1}{N}$ factors at lines 11, 13, line 26 of Alg 1.