1.2 (b)

| Layer | Input | Output |
|---|---|---|
| Linear 1 | $x$ | $W^{(1)}x + b^{(1)}$ | $= z_1$ |
| $f$ | $W^{(1)}x + b^{(1)}$ | $\max(0, W^{(1)}x + b^{(1)}) = y_1$ | $= z_2$ |
| Linear 2 | $y_1$ | $W^{(2)}y_1 + b^{(2)}$ | $= z_3$ |
| $g$ | $W^{(2)}y_1 + b^{(2)}$ | $W^{(2)}y_1 + b^{(2)} = \hat{y}$ | $= \hat{y}$ |
| Loss | $\hat{y}$ | $\|\hat{y} - y\|_2^2$ | |

(c)

$\frac{\partial l}{\partial z_3} = \frac{\partial l}{\partial \hat{y}}$ (identity act.)

$\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial z_3} \cdot \frac{\partial z_3}{\partial W^{(2)}} = \frac{\partial l}{\partial \hat{y}} \cdot (z_2)^T$

Assuming some dimension.

$x \in \mathbb{R}^{d_1 \times 1}$

$z_2, z_1, b^{(1)} \in \mathbb{R}^{d_2 \times 1}$

$W^{(1)} \in \mathbb{R}^{d_2 \times d_1}$

$z_3, b^{(2)}, \hat{y} \in \mathbb{R}^{d_3 \times 1}$

$W^{(2)} \in \mathbb{R}^{d_3 \times d_2}$

$\frac{\partial l}{\partial b^{(2)}} = \frac{\partial l}{\partial z_3} = \frac{\partial l}{\partial \hat{y}}$

Now, $\frac{\partial l}{\partial z_2} = \frac{\partial l}{\partial z_3} \cdot \frac{\partial z_3}{\partial z_2} = W_2^T \cdot \frac{\partial l}{\partial \hat{y}}$

$\frac{\partial l}{\partial z_1} = \frac{\partial l}{\partial z_2} \odot \frac{\partial z_2}{\partial z_1} = \left(W_2^T \cdot \frac{\partial l}{\partial \hat{y}}\right) \odot (z_1 > 0)$ ← Hadamard product.

$\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial z_1} \cdot \frac{\partial z_1}{\partial W_1} = \frac{\partial l}{\partial z_1} \cdot x^T$

$\frac{\partial l}{\partial b^{(1)}} = \frac{\partial l}{\partial z_1}$

Writing all these,

$\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_3} \cdot z_2^T = \frac{\partial l}{\partial \hat{y}} \cdot \frac{\partial z_3}{\partial z_2} \cdot \left[(W^{(1)}x + b^{(1)})^+\right]^T$

$\frac{\partial l}{\partial b^{(2)}} = \frac{\partial l}{\partial \hat{y}}$

$\frac{\partial l}{\partial W^{(1)}} = \left[W_2^T \frac{\partial l}{\partial \hat{y}} \odot (W^{(1)}x + b^{(1)} > 0)\right] \cdot x^T$

$\frac{\partial l}{\partial b^{(1)}} = \left[W_2^T \frac{\partial l}{\partial \hat{y}} \odot (W^{(1)}x + b^{(1)} > 0)\right]$

(d) $\frac{\partial l}{\partial \hat{y}} = \begin{bmatrix} \partial l/\partial \hat{y}_1 \\ \vdots \\ \partial l/\partial \hat{y}_{d_3} \end{bmatrix}$ $\frac{\partial \hat{y}}{\partial z_3} = I_{d_3 \times d_3}$ (diff identity)

$\frac{\partial z_2}{\partial z_1} = (z_1 > 0) = (W^{(1)}x + b^{(1)} > 0)_{d_2}$ Relu grad.

Ideally non differentiable

1.3. (a) with $f = g = \sigma$

→ the forward pass eqn$^s$ change accordingly.

→ $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ ← element wise product

↑ derivative changes

(b) $l(\hat{y}, y) = \frac{-1}{k} \sum_{i=1}^{k} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$

$\frac{\partial l}{\partial \hat{y}_i} = -\frac{1}{k}\left[y_i/\hat{y}_i - (1 - y_i)/(1 - \hat{y}_i)\right]$

→ make vectors, just derivative changes.

(c) Using a '$\sigma$' for intermediate activations results in a high probability that the gradients will explode or die.