NET ID: sa857

gradients

Minimal neural net gradients:

Working:

$$h_1 = \text{relu}(w_1 x + b_1)$$

$$h_2 = \text{relu}(w_2 h_1 + b_2)$$

$$f = w_3 h_2 + b_3$$

$$L = \frac{1}{2}(y - f(x))^2$$

$$\frac{\partial f}{\partial w_3} = h_2^T \qquad \frac{\partial f}{\partial h_2} = w_3^T$$

$$\frac{\partial h_2}{\partial b_2} = 0 + 1 = 1$$

$$\frac{\partial h_2}{\partial w_2} = h_1 \qquad \frac{\partial h_2}{\partial h_1} = w_2$$

$$\frac{\partial h_1}{\partial b_1} = 0 + 1 = 1$$

$$\frac{\partial h_1}{\partial w_1} = x$$

LAYER 3

$$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial b_3} = (f-y) \cdot 1$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial w_3} = (f-y) \cdot h_2^T$$

$$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_2} = (f-y) \cdot w_3^T$$

LAYER 2

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_2} \cdot \frac{\partial h_2}{\partial b_2} = (f-y) \cdot w_3^T \cdot 1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_2} = (f-y) \cdot w_3^T \cdot h_1^T$$

$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} = (f-y) \cdot w_3^T \cdot w_2^T$$

LAYER 1

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial b_1}$$

$$= (f-y) \cdot w_3^T \cdot w_2^T \cdot 1$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_1}$$

$$= (f-y) \cdot w_3^T \cdot w_2^T \cdot x^T$$

$\frac{\partial L}{\partial x}$ == not required

Final answers in LaTeX (fully expanded except dL/df):

Layer 3

$$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial f}$$
$$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial f} h_2^T$$
$$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial f} W_3^T$$

Layer 2

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial f} W_3^T$$
$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial f} W_3^T h_1^T$$
$$\frac{\partial L}{\partial h_1} = \frac{\partial L}{\partial f} W_3^T W_2^T$$

Layer 1

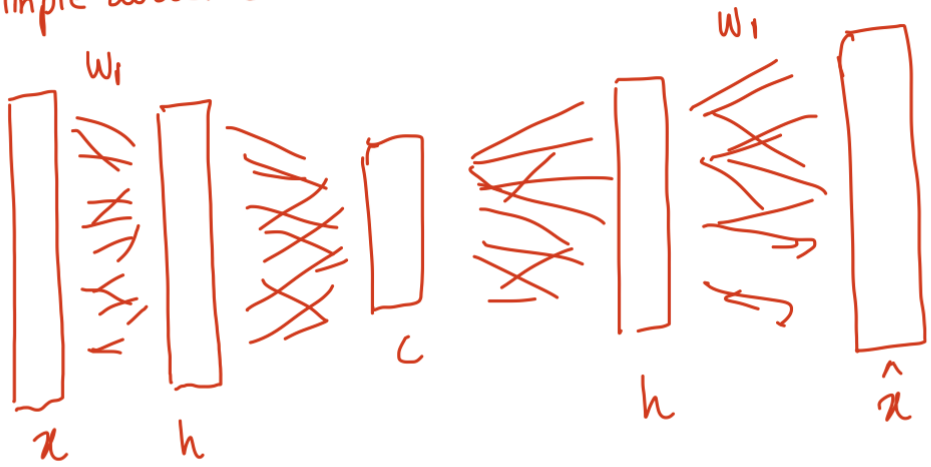
$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial f} W_3^T W_2^T$$
$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial f} W_3^T W_2^T X^T$$

Autoencoder gradients

gradients

Working:

Simple autoencoder



$h = \text{relu}(W_1 X + b_1)$ $\quad \hat{x} = \text{relu}(W_2 h + b_2)$

$X : (n, D)$

$W_1 : (H, D)$

$b_1 : (H,)$

$$\text{Loss } L = \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - \text{relu}\left(W_2 \cdot \text{relu}\left(W_1 x_i + b_1\right) + b_2\right) \right\|^2$$

$$L = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2$$

the true value is the data point itself

$$\frac{\partial L}{\partial \hat{x}} = \frac{2}{n} \sum_{i=1}^{n} \|-1\| = -2$$

$\hat{x} = \text{relu}(W_2 h + b_2)$ $\quad h = \text{relu}(W_1 X + b_1)$

$$\frac{\partial \hat{x}}{\partial b_2} = 0 + 1 = 1$$

Final answers in LaTeX (fully expanded except dL/dx^):

Layer 2

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{x}}$$
$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \hat{x}} h^T$$
$$\frac{\partial L}{\partial h} = \frac{\partial L}{\partial \hat{x}} W_2^T$$

Layer 1

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial \hat{x}} W_2^T$$
$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial \hat{x}} W_2^T X^T$$