Gra

Gradient Problem

Forward

$$h^l = f_l\left(h^{l-1}\right) \text{ and } y = l \circ f_d \circ \ldots \circ f_1(x)$$

Gradient Exploding

1. If w is initialized with > 1,
2. get too many layers

Weights get too big for float16

Parameter Initializati

Target:

$$\mathbb{E}\left[\frac{\partial l}{\partial h^t}\right] = 0$$

$$Var\left[\frac{\partial l}{\partial h^t}\right] = b$$

dient Problem

Backward

$$\frac{\partial l}{\partial W^t} = \frac{\partial l}{\partial h^d}\frac{\partial h^d}{\partial h^{d-1}} \dots \frac{\partial h^{t+1}}{\partial h^t}\frac{\partial h^t}{\partial W^t}$$
$$= M^L \dots M^{l+1} v^l$$

Gradient Vanishing

1. If w is initialized with < 1,
2. get too many layers

Weights get too small for float16

Symmetry Problem

1. If w is initialized with same constant c

Weights are updated consistently

Target

Let gradients lie in proper interval

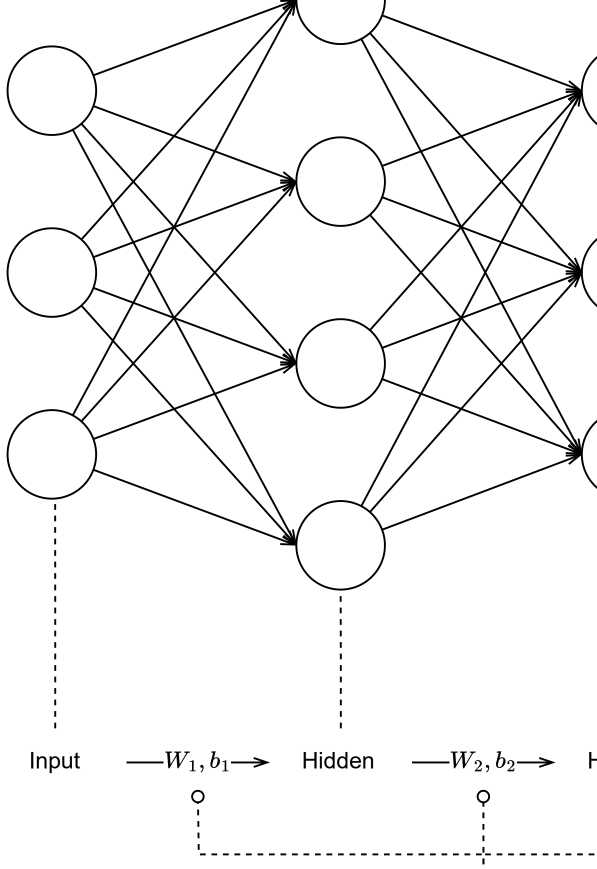
on

Others

Multplication $\rightarrow$ Add

Standardization

Solution



MLP example (h = hidden layer weights)

Condition:

1. $w_{i,j}^t$ is of i.i.d, **we get:** $\mathbb{E}\left[\frac{\partial l}{\partial h^t}\right] = 0 \;\; Var\left[\frac{\partial l}{\partial h^t}\right] = b$
2. h_i^{t-1} is independent to $w_{i,j}^t$
3. Dont consider activation function

Expectation and Variance of every layers' weights:

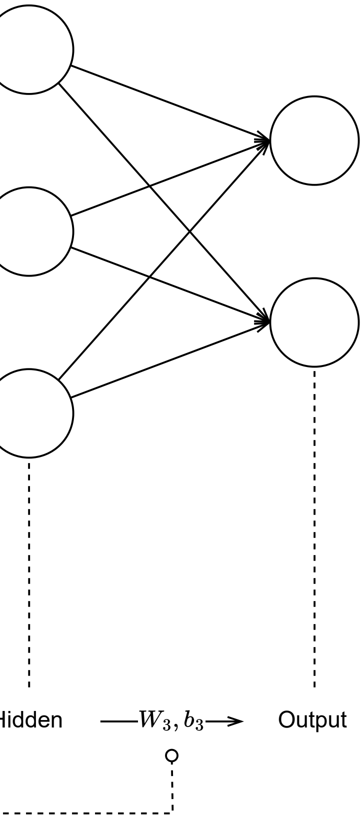
1. Forward $\rightarrow$ $\mathbb{E}\left[h_i^t\right] = 0$, $Var\left[h_i^t\right] = n_{t-1}\gamma_t Var\left[h_i^{t-1}\right]$ $\rightarrow$ $n_{t-1}\gamma_t =$
2. Backward $\rightarrow$ $\mathbb{E}\left[\frac{\partial l}{\partial h_i^{t-1}}\right] = 0$, $Var\left[\frac{\partial l}{\partial h_i^{t-1}}\right] = n_t\gamma_t Var\left[\frac{\partial l}{\partial h_i^t}\right]$ $\rightarrow$

Hard to fullfill two of them

Solution: **Xavier Initialization**

1. **Let** $\gamma_t \frac{n_{t-1} + n_t}{2} = 1 \rightarrow \gamma_t = \frac{2}{n_{t-1} + n_t}$
2. **We get** $\mathscr{N}\left(0, \sqrt{\frac{2}{n_{t-1} + nt}}\right)$, $\mathscr{U}\left(-\sqrt{\frac{6}{n_{t-1} + n_t}}, \sqrt{\frac{6}{n_{t-1} + n_t}}\right)$

What if to consider activation function?



1

$n_t \gamma_t = 1$

Activation function selection

require linear part

tanh: OK
ReLU: OK
Sigmoid: Not OK
Scaled Sigmoid: OK

Scaled Sigmoid: 4xsigmoid(x) - 2