

Layered architectures

Computer Vision (CSCI 5520G)

Faisal Z. Qureshi

<http://vclab.science.ontariotechu.ca>



Logistic regression and layers

- ▶ Lets look at how we can specify logistic regression as layers.
- ▶ The ability to specify such models as layers is key to designing neural networks.
- ▶ We will also discuss *backpropagation*.

2-class softmax classifier

2-class softmax classifier

- ▶ Define cost $C(\theta)$ that needs to be minimized to train this classifier.
- ▶ We set this cost equal to the negative log-likelihood for this 2-class softmax classifier

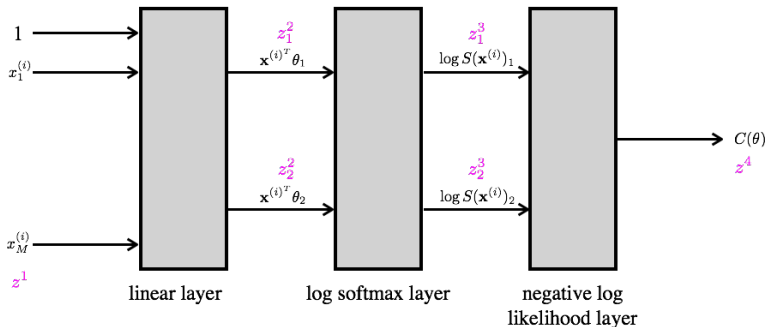
Negative log-likelihood

$$\begin{aligned} C(\theta) &= l(\theta) \\ &= - \sum_{i=1}^N \mathbb{I}_0(y^{(i)}) \log \frac{e^{\mathbf{x}^{(i)T} \theta_1}}{e^{\mathbf{x}^{(i)T} \theta_1} + e^{\mathbf{x}^{(i)T} \theta_2}} + \mathbb{I}_1(y^{(i)}) \log \frac{e^{\mathbf{x}^{(i)T} \theta_2}}{e^{\mathbf{x}^{(i)T} \theta_1} + e^{\mathbf{x}^{(i)T} \theta_2}} \end{aligned}$$

Layer representation

$$C(\theta) = - \sum_{i=1}^N \mathbb{I}_0(y^{(i)}) \log \frac{e^{\mathbf{x}^{(i)T} \theta_1}}{e^{\mathbf{x}^{(i)T} \theta_1} + e^{\mathbf{x}^{(i)T} \theta_2}} + \mathbb{I}_1(y^{(i)}) \log \frac{e^{\mathbf{x}^{(i)T} \theta_2}}{e^{\mathbf{x}^{(i)T} \theta_1} + e^{\mathbf{x}^{(i)T} \theta_2}}$$

Layers



Chain rule

$$\frac{\partial f(g(u, v), h(u, v))}{\partial u} =$$

Chain rule

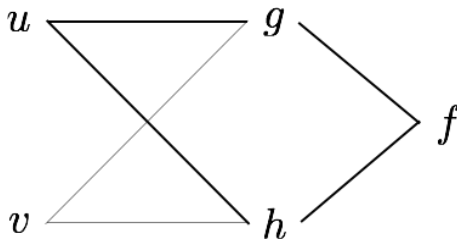
$$\frac{\partial f(g(u, v), h(u, v))}{\partial u} = \frac{\partial f}{\partial g} \frac{\partial g}{\partial u}$$

Chain rule

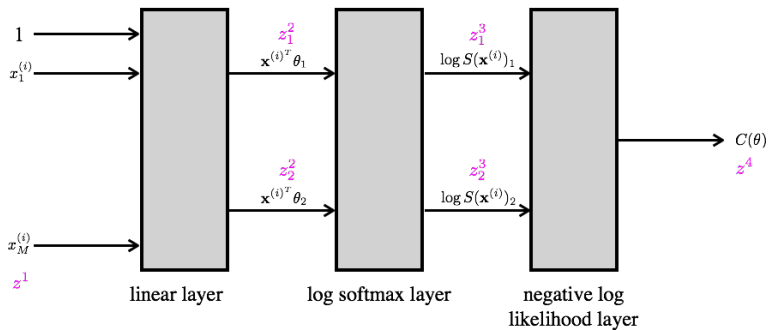
$$\frac{\partial f(g(u, v), h(u, v))}{\partial u} = \frac{\partial f}{\partial g} \frac{\partial g}{\partial u} + \frac{\partial f}{\partial h} \frac{\partial h}{\partial u}$$

Chain rule

$$\frac{\partial f(g(u, v), h(u, v))}{\partial u} = \frac{\partial f}{\partial g} \frac{\partial g}{\partial u} + \frac{\partial f}{\partial h} \frac{\partial h}{\partial u}$$

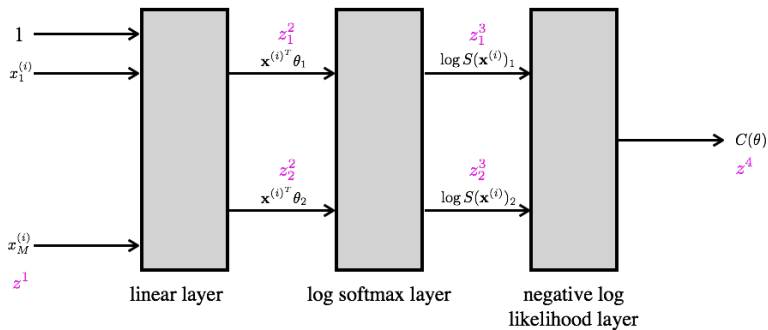


Layers



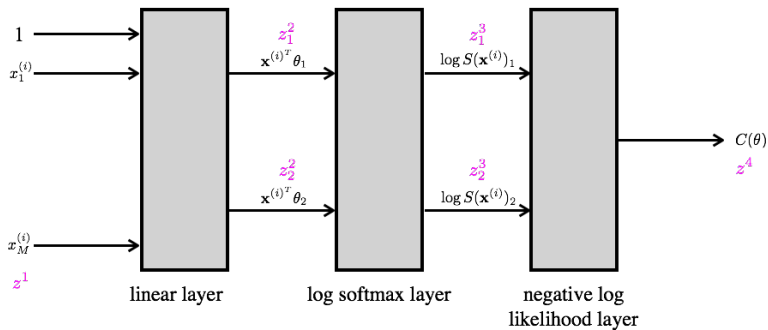
- What are our model parameters?

Layers



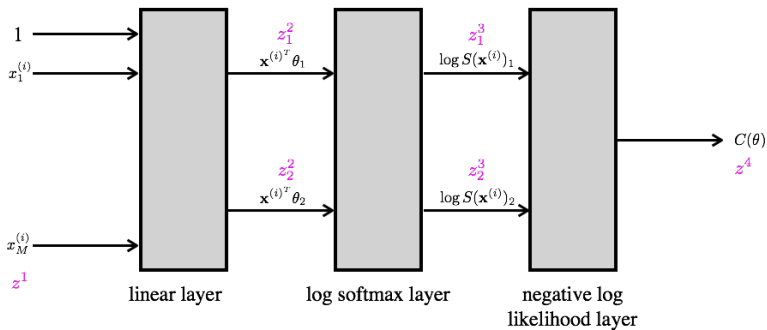
- ▶ What are our model parameters? (θ_1 and θ_2)
- ▶ We are interested in computing $\frac{\partial z^4}{\partial \theta_1}$ and $\left(\frac{\partial z^4}{\partial \theta_2}\right)$
 - ▶ Why?

Layers



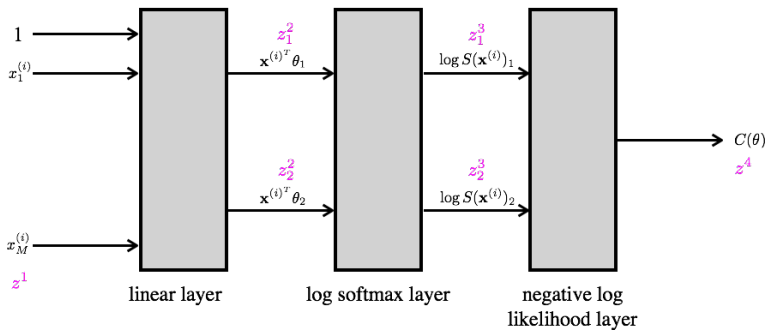
- ▶ What are our model parameters? (θ_1 and θ_2)
- ▶ We are interested in computing $\frac{\partial z^4}{\partial \theta_1}$ and $\left(\frac{\partial z^4}{\partial \theta_2}\right)$
 - ▶ Why? (z^4 is the cost that we can minimize using gradient descent using these gradients)

Layers



- ▶ What are our model parameters? (θ_1 and θ_2)
- ▶ We are interested in computing $\frac{\partial z^4}{\partial \theta_1}$ and $(\frac{\partial z^4}{\partial \theta_2})$
 - ▶ Why? (z^4 is the cost that we can minimize using gradient descent using these gradients)
 - ▶ How do we even compute $\frac{\partial z^4}{\partial \theta_1}$?

Layers

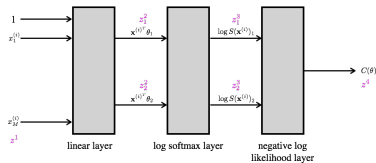


- ▶ What are our model parameters? (θ_1 and θ_2)
- ▶ We are interested in computing $\frac{\partial z^4}{\partial \theta_1}$ and $\left(\frac{\partial z^4}{\partial \theta_2}\right)$
 - ▶ Why? (z^4 is the cost that we can minimize using gradient descent using these gradients)
 - ▶ How do we even compute $\frac{\partial z^4}{\partial \theta_1}$? (or $\frac{\partial z^4}{\partial \theta_2}$)

Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

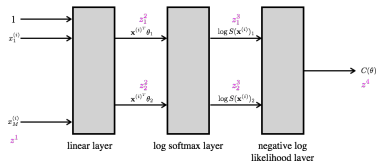
$$\frac{\partial z^4}{\partial \theta_1} =$$



Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

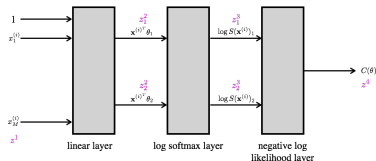
$$\frac{\partial z^4}{\partial \theta_1} = \frac{\partial z^4}{\partial z_1^3} \frac{\partial z_1^3}{\partial \theta_1}$$



Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

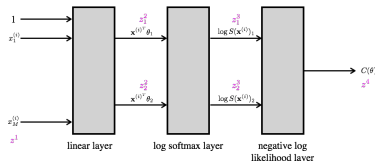
$$\begin{aligned}\frac{\partial z^4}{\partial \theta_1} &= \frac{\partial z^4}{\partial z_1^3} \frac{\partial z_1^3}{\partial \theta_1} + \frac{\partial z^4}{\partial z_2^3} \frac{\partial z_2^3}{\partial \theta_1} \\ &= \frac{\partial z^4}{\partial z_1^3}\end{aligned}$$



Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

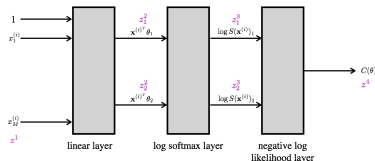
$$\begin{aligned}\frac{\partial z^4}{\partial \theta_1} &= \frac{\partial z^4}{\partial z_1^3} \frac{\partial z_1^3}{\partial \theta_1} + \frac{\partial z^4}{\partial z_2^3} \frac{\partial z_2^3}{\partial \theta_1} \\ &= \frac{\partial z^4}{\partial z_1^3} \left(\frac{\partial z_1^3}{\partial z_1^2} \frac{\partial z_1^2}{\partial \theta_1} + \frac{\partial z_1^3}{\partial z_2^2} \frac{\partial z_2^2}{\partial \theta_1} \right)\end{aligned}$$



Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

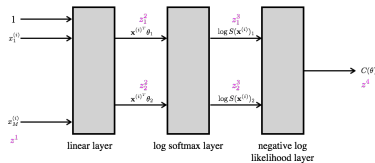
$$\begin{aligned}\frac{\partial z^4}{\partial \theta_1} &= \frac{\partial z^4}{\partial z_1^3} \frac{\partial z_1^3}{\partial \theta_1} + \frac{\partial z^4}{\partial z_2^3} \frac{\partial z_2^3}{\partial \theta_1} \\ &= \frac{\partial z^4}{\partial z_1^3} \left(\frac{\partial z_1^3}{\partial z_1^2} \frac{\partial z_1^2}{\partial \theta_1} + \frac{\partial z_1^3}{\partial z_2^2} \frac{\partial z_2^2}{\partial \theta_1} \right) \\ &\quad + \frac{\partial z^4}{\partial z_2^3}\end{aligned}$$



Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

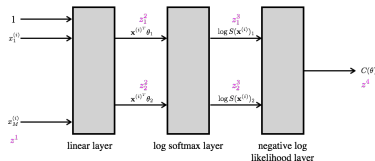
$$\begin{aligned}\frac{\partial z^4}{\partial \theta_1} &= \frac{\partial z^4}{\partial z_1^3} \frac{\partial z_1^3}{\partial \theta_1} + \frac{\partial z^4}{\partial z_2^3} \frac{\partial z_2^3}{\partial \theta_1} \\ &= \frac{\partial z^4}{\partial z_1^3} \left(\frac{\partial z_1^3}{\partial z_1^2} \frac{\partial z_1^2}{\partial \theta_1} + \frac{\partial z_1^3}{\partial z_2^2} \frac{\partial z_2^2}{\partial \theta_1} \right) \\ &\quad + \frac{\partial z^4}{\partial z_2^3} \left(\frac{\partial z_2^3}{\partial z_1^2} \frac{\partial z_1^2}{\partial \theta_1} + \frac{\partial z_2^3}{\partial z_2^2} \frac{\partial z_2^2}{\partial \theta_1} \right)\end{aligned}$$



Computing $\frac{\partial z^4}{\partial \theta_1}$

We can use the *chain rule* to compute $\frac{\partial z^4}{\partial \theta_1}$.

$$\begin{aligned}\frac{\partial z^4}{\partial \theta_1} &= \frac{\partial z^4}{\partial z_1^3} \frac{\partial z_1^3}{\partial \theta_1} + \frac{\partial z^4}{\partial z_2^3} \frac{\partial z_2^3}{\partial \theta_1} \\ &= \frac{\partial z^4}{\partial z_1^3} \left(\frac{\partial z_1^3}{\partial z_1^2} \frac{\partial z_1^2}{\partial \theta_1} + \frac{\partial z_1^3}{\partial z_2^2} \frac{\partial z_2^2}{\partial \theta_1} \right) \\ &\quad + \frac{\partial z^4}{\partial z_2^3} \left(\frac{\partial z_2^3}{\partial z_1^2} \frac{\partial z_1^2}{\partial \theta_1} + \frac{\partial z_2^3}{\partial z_2^2} \frac{\partial z_2^2}{\partial \theta_1} \right)\end{aligned}$$



We can similarly compute $\frac{\partial z^4}{\partial \theta_2}$.

Notation

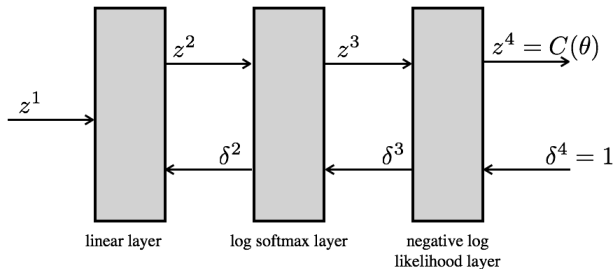
- ▶ z^l : input to layer l
- ▶ z^{l+1} : output of layer l
- ▶ $C(\theta)$: cost (or loss) to be minimized, e.g., negative log-likelihood for the case of the two-class softmax classifier
- ▶ Define

$$\delta^l = \frac{\partial C(\theta)}{\partial z^l}$$

- ▶ If L denotes the last layer, $\delta^L = 1$, since the output of the last layer is z^L which is equal to the cost $C(\theta)$, i.e.,

$$\delta^L = \frac{\partial C(\theta)}{\partial z^L} = \frac{\partial z^L}{\partial z^L} = 1$$

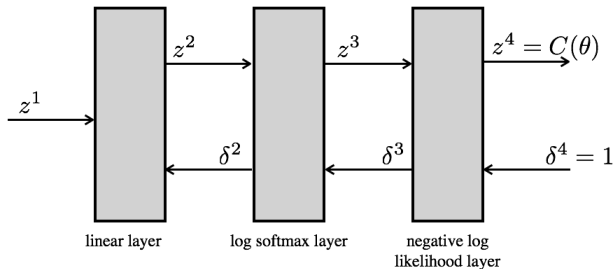
Forward pass and backward pass



Forward pass

Backward pass

Forward pass and backward pass

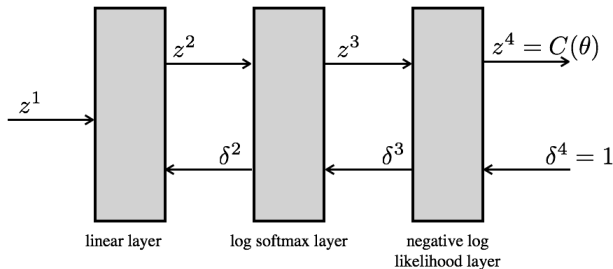


Forward pass

$z^1 = f(\mathbf{x})$ (*input data*)

Backward pass

Forward pass and backward pass



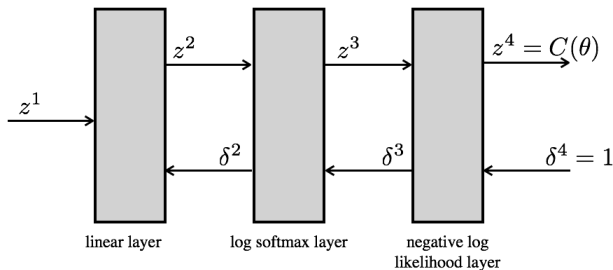
Forward pass

$z^1 = f(\mathbf{x})$ (*input data*)

$z^2 = f(z^1)$ (*linear function*)

Backward pass

Forward pass and backward pass



Forward pass

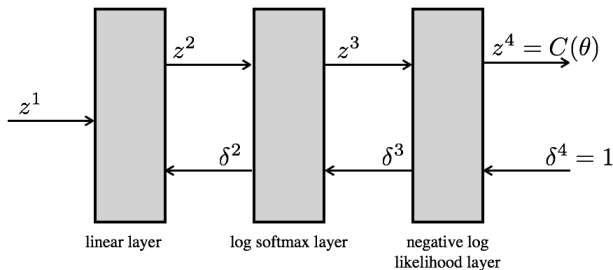
$z^1 = f(\mathbf{x})$ (*input data*)

$z^2 = f(z^1)$ (*linear function*)

$z^3 = f(z^2)$ (*log softmax*)

Backward pass

Forward pass and backward pass



Forward pass

$z^1 = f(\mathbf{x})$ (input data)

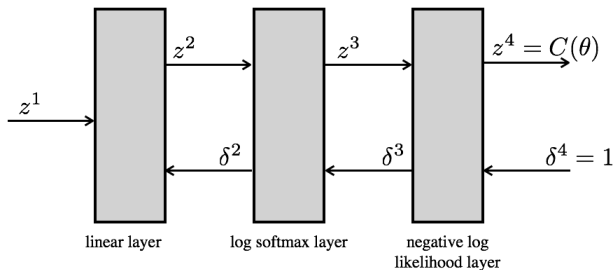
$z^2 = f(z^1)$ (linear function)

$z^3 = f(z^2)$ (log softmax)

$z^4 = f(z^3) = C(\theta)$ (negative log likelihood, cost)

Backward pass

Forward pass and backward pass



Forward pass

$z^1 = f(\mathbf{x})$ (input data)

$z^2 = f(z^1)$ (linear function)

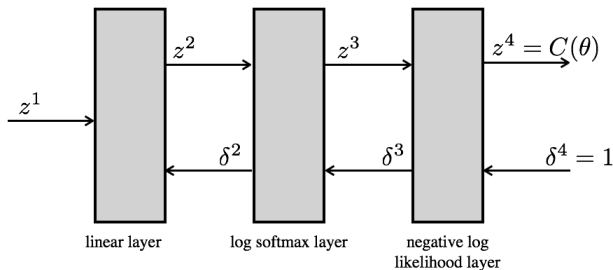
$z^3 = f(z^2)$ (log softmax)

$z^4 = f(z^3) = C(\theta)$ (negative log likelihood, cost)

Backward pass

$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = 1$$

Forward pass and backward pass



Forward pass

$z^1 = f(\mathbf{x})$ (input data)

$z^2 = f(z^1)$ (linear function)

$z^3 = f(z^2)$ (log softmax)

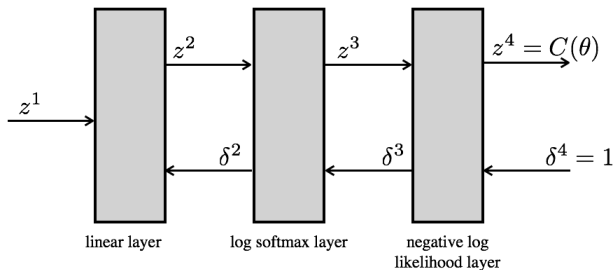
$z^4 = f(z^3) = C(\theta)$ (negative log likelihood, cost)

Backward pass

$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = 1$$

$$\delta^3 = \frac{\partial C(\theta)}{\partial z^3}$$

Forward pass and backward pass



Forward pass

$z^1 = f(\mathbf{x})$ (input data)

$z^2 = f(z^1)$ (linear function)

$z^3 = f(z^2)$ (log softmax)

$z^4 = f(z^3) = C(\theta)$ (negative log likelihood, cost)

Backward pass

$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = 1$$

$$\delta^3 = \frac{\partial C(\theta)}{\partial z^3}$$

$$\delta^2 = \frac{\partial C(\theta)}{\partial z^2}$$

Computing δ^l

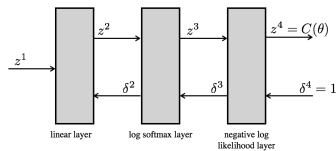
$$\delta^4 =$$

$$\delta_1^3 =$$

$$\delta_2^3 =$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

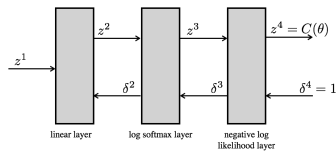
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4}$$

$$\delta_1^3 =$$

$$\delta_2^3 =$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

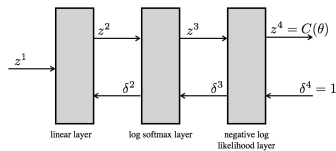
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 =$$

$$\delta_2^3 =$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

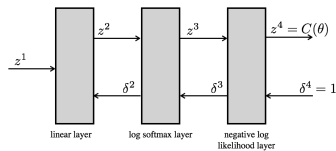
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} =$$

$$\delta_2^3 =$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

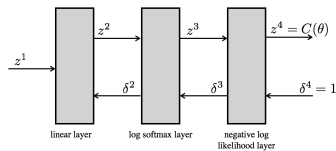
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 =$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

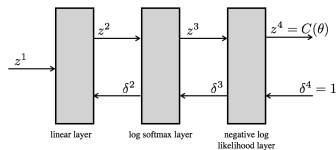
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 =$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

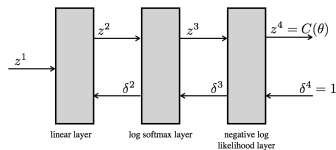
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3}$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

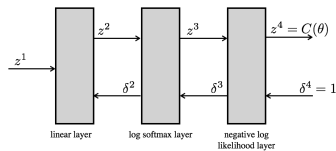
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

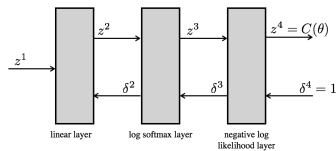
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 =$$

$$\delta_2^2 =$$



Computing δ^l

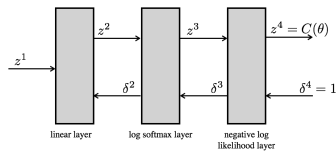
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 = \frac{\partial C(\theta)}{\partial z_1^2}$$

$$\delta_2^2 =$$



Computing δ^l

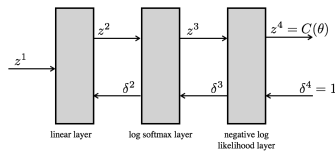
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 = \frac{\partial C(\theta)}{\partial z_1^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_1^2}$$

$$\delta_2^2 =$$



Computing δ^l

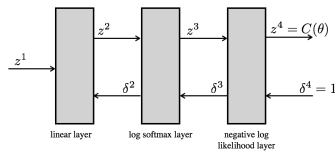
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 = \frac{\partial C(\theta)}{\partial z_1^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_1^2} = \sum_k \delta_k^3 \frac{\partial z_k^3}{\partial z_1^2}$$

$$\delta_2^2 =$$



Computing δ^l

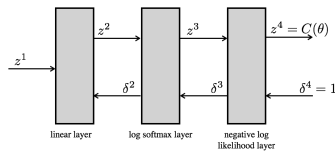
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 = \frac{\partial C(\theta)}{\partial z_1^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_1^2} = \sum_k \delta_k^3 \frac{\partial z_k^3}{\partial z_1^2}$$

$$\delta_2^2 = \frac{\partial C(\theta)}{\partial z_2^2}$$



Computing δ^l

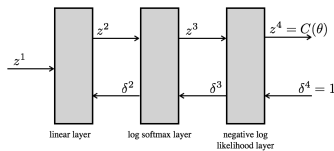
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 = \frac{\partial C(\theta)}{\partial z_1^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_1^2} = \sum_k \delta_k^3 \frac{\partial z_k^3}{\partial z_1^2}$$

$$\delta_2^2 = \frac{\partial C(\theta)}{\partial z_2^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_2^2}$$



Computing δ^l

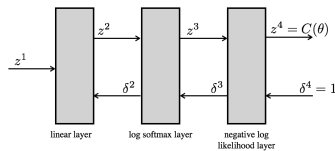
$$\delta^4 = \frac{\partial C(\theta)}{\partial z^4} = \frac{\partial z^4}{\partial z^4} = 1$$

$$\delta_1^3 = \frac{\partial C(\theta)}{\partial z_1^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_1^3} = \delta^4 \frac{\partial z^4}{\partial z_1^3}$$

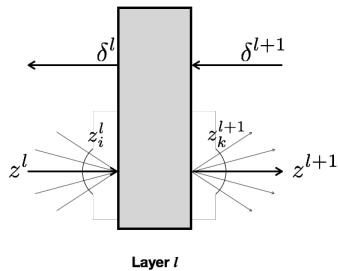
$$\delta_2^3 = \frac{\partial C(\theta)}{\partial z_2^3} = \frac{\partial C(\theta)}{\partial z^4} \frac{\partial z^4}{\partial z_2^3} = \delta^4 \frac{\partial z^4}{\partial z_2^3}$$

$$\delta_1^2 = \frac{\partial C(\theta)}{\partial z_1^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_1^2} = \sum_k \delta_k^3 \frac{\partial z_k^3}{\partial z_1^2}$$

$$\delta_2^2 = \frac{\partial C(\theta)}{\partial z_2^2} = \sum_k \frac{\partial C(\theta)}{\partial z_k^3} \frac{\partial z_k^3}{\partial z_2^2} = \sum_k \delta_k^3 \frac{\partial z_k^3}{\partial z_2^2}$$



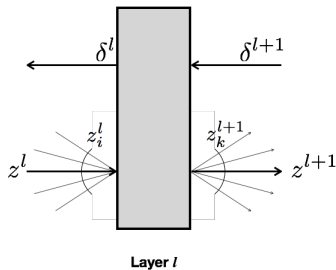
For any differentiable layer l



For any differentiable layer l

For a given layer l , with inputs z_i^l and outputs z_k^{l+1}

$$\delta_i^l = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_i^l}$$



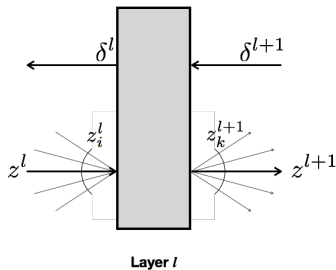
For any differentiable layer l

For a given layer l , with inputs z_i^l and outputs z_k^{l+1}

$$\delta_i^l = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_i^l}$$

Similarly, for layer l that depends upon parameters θ^l ,

$$\frac{\partial C(\theta)}{\partial \theta^l} =$$



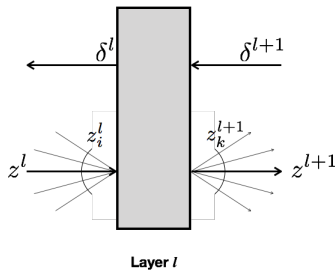
For any differentiable layer l

For a given layer l , with inputs z_i^l and outputs z_k^{l+1}

$$\delta_i^l = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_i^l}$$

Similarly, for layer l that depends upon parameters θ^l ,

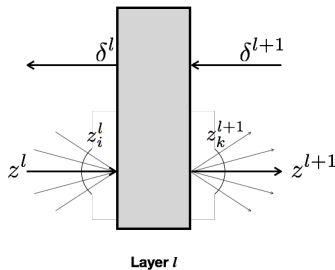
$$\frac{\partial C(\theta)}{\partial \theta^l} = \sum_k \frac{\partial C(\theta)}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial \theta^l} =$$



For any differentiable layer l

For a given layer l , with inputs z_i^l and outputs z_k^{l+1}

$$\delta_i^l = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_i^l}$$



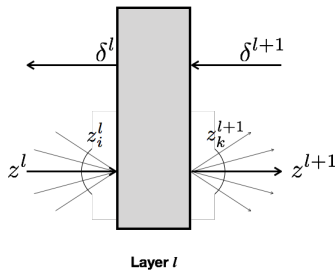
Similarly, for layer l that depends upon parameters θ^l ,

$$\frac{\partial C(\theta)}{\partial \theta^l} = \sum_k \frac{\partial C(\theta)}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial \theta^l} = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial \theta^l}$$

For any differentiable layer l

For a given layer l , with inputs z_i^l and outputs z_k^{l+1}

$$\delta_i^l = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_i^l}$$



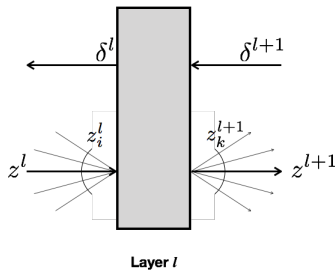
Similarly, for layer l that depends upon parameters θ^l ,

$$\frac{\partial C(\theta)}{\partial \theta^l} = \sum_k \frac{\partial C(\theta)}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial \theta^l} = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial \theta^l}$$

For any differentiable layer l

For a given layer l , with inputs z_i^l and outputs z_k^{l+1}

$$\delta_i^l = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_i^l}$$



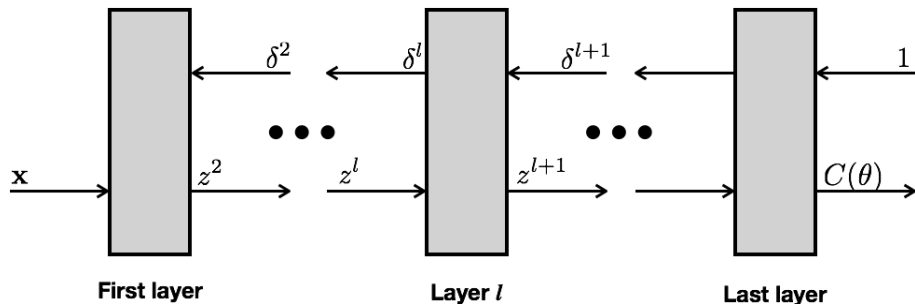
Similarly, for layer l that depends upon parameters θ^l ,

$$\frac{\partial C(\theta)}{\partial \theta^l} = \sum_k \frac{\partial C(\theta)}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial \theta^l} = \sum_k \delta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial \theta^l}$$

In our 2-class softmax classifier only layer 1 has parameters (θ_0 and θ_1)

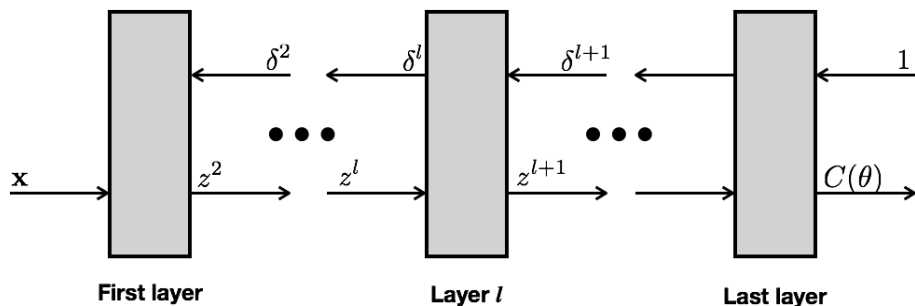
Layered architectures

As long as we have differentiable layers, i.e., we can compute $\frac{\partial z_k^{l+1}}{\partial z_i^l}$, we can use *backpropagation* to update the parameters θ to minimize the cost $C(\theta)$.



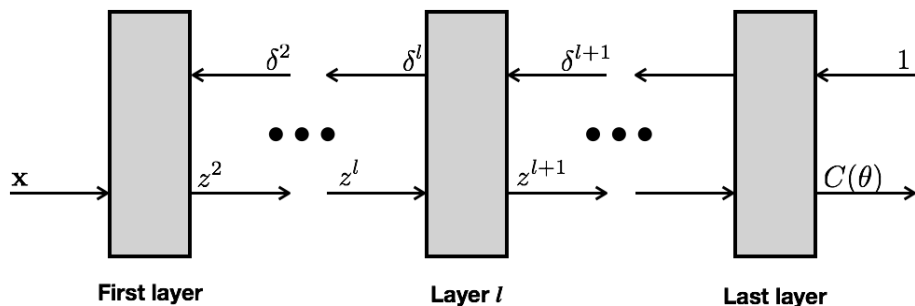
Layered architectures

As long as **we have differentiable layers**, i.e., we can compute $\frac{\partial z_k^{l+1}}{\partial z_i^l}$, **we can use *backpropagation* to update the parameters θ to minimize the cost $C(\theta)$.**



Layered architectures

As long as we have differentiable layers, i.e., we can compute $\frac{\partial z_k^{l+1}}{\partial z_i^l}$, we can use *backpropagation* to update the parameters θ to minimize the cost $C(\theta)$.



Backpropagation

Backpropagation

- ▶ Set z^1 equal to input x .

Backpropagation

- ▶ Set z^1 equal to input x .
- ▶ **Forward pass:**

Backpropagation

- ▶ Set z^1 equal to input x .
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations

Backpropagation

- ▶ Set z^1 equal to input x .
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations
 - ▶ Set δ at the last layer equal to 1

Backpropagation

- ▶ Set z^1 equal to input x .
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations
 - ▶ Set δ at the last layer equal to 1
- ▶ **Backward pass:**

Backpropagation

- ▶ Set z^1 equal to input $\mathbf{x}$.
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations
 - ▶ Set δ at the last layer equal to 1
- ▶ **Backward pass:**
 - ▶ Backpropagate $\delta^L, \delta^{L-1}, \dots, \delta^2$ (all the way to first layer)

Backpropagation

- ▶ Set z^1 equal to input $\mathbf{x}$.
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations
 - ▶ Set δ at the last layer equal to 1
- ▶ **Backward pass:**
 - ▶ Backpropagate $\delta^L, \delta^{L-1}, \dots, \delta^2$ (all the way to first layer)
 - ▶ Compute $\nabla_{\theta} C(\theta)$

Backpropagation

- ▶ Set z^1 equal to input $\mathbf{x}$.
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations
 - ▶ Set δ at the last layer equal to 1
- ▶ **Backward pass:**
 - ▶ Backpropagate $\delta^L, \delta^{L-1}, \dots, \delta^2$ (all the way to first layer)
 - ▶ Compute $\nabla_{\theta} C(\theta)$
- ▶ Update θ

Backpropagation

- ▶ Set z^1 equal to input $\mathbf{x}$.
- ▶ **Forward pass:**
 - ▶ Compute $z^2, z^3, \dots$ layers $1, 2, \dots$ activations
 - ▶ Set δ at the last layer equal to 1
- ▶ **Backward pass:**
 - ▶ Backpropagate $\delta^L, \delta^{L-1}, \dots, \delta^2$ (all the way to first layer)
 - ▶ Compute $\nabla_{\theta} C(\theta)$
- ▶ Update θ
- ▶ Repeat

Layered architecture: consequences

- ▶ Compositionality
- ▶ Reuse
- ▶ Ease of constructing **your own layers**

Layered architecture: consequences

- ▶ Compositionality
- ▶ Reuse
- ▶ Ease of constructing **your own layers**

A neural network can itself be treated as a layer within another neural network (recursion). This allows us to build new neural networks using existing (and sometimes pre-trained) models.

Summary

- ▶ Layered view of a 2-class softmax classifier
- ▶ Chain rule of differentiation
- ▶ Differentiable layers
- ▶ Backpropagation

Copyright and License

©Faisal Z. Qureshi



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.