Machine learning & deep learning basics

Lecture 3

Changho Suh

January 22, 2024

Outline

1. Study an efficient way of implementing gradient descent:

   Backpropagation

2. Study a practical variant of gradient descent:

   Adam optimizer

Backpropagation

Gradient descent for DNN

$$\min_{w=(W^{[1]},W^{[2]})} \underbrace{\boxed{\frac{1}{m}\sum_{i=1}^{m} -y^{(i)}\log\hat{y}^{(i)} - (1-y^{(i)})\log(1-\hat{y}^{(i)})}}_{=:\,J(w)}$$

$$\hat{y}^{(i)} = \sigma\left(W^{[2]}\max\left(0, W^{[1]}x^{(i)}\right)\right)$$

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha^{(t)}\nabla_w J(w^{(t)})$$

$$W^{[2],(t+1)} \leftarrow W^{[2],(t)} - \alpha^{(t)}\boxed{\nabla_{W^{[2]}} J(w^{(t)})}$$

$$W^{[1],(t+1)} \leftarrow W^{[1],(t)} - \alpha^{(t)}\boxed{\nabla_{W^{[1]}} J(w^{(t)})}$$

An efficient way of computing the two gradients:

Backpropagation!

Backpropagation

Idea: Successively compute gradients in a backward manner by using a chain rule for derivatives.

Consider: $y = f(x) \quad z = g(y)$

$$z = g(f(x)), \quad \frac{dz}{dx}?$$

Chain rule: $$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

Chain rule offers an easy way to compute.

Will provide backprop in a simple context: for *m*=1.

Backpropagation: *m*=1

Recall the forward path:

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

Backpropagation: *m*=1

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

Start from backward: $\frac{dJ(w)}{d\hat{y}}$

$$J(w) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

Easy to compute $\frac{dJ(w)}{d\hat{y}}$

Backpropagation: *m*=1

from an earlier stage

Chain rule: $\frac{dJ(w)}{dz^{[2]}} = \boxed{\frac{dJ(w)}{d\hat{y}}} \boxed{\frac{d\hat{y}}{dz^{[2]}}}$

compute from

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

Next consider: $\frac{dJ(w)}{dz^{[2]}}$ $\frac{dJ(w)}{d\hat{y}}$

Backpropagation: *m*=1

from an earlier stage

Chain rule: $$\frac{dJ(w)}{dW^{[2]}} = \boxed{\frac{dJ(w)}{dz^{[2]}}}\boxed{\frac{dz^{[2]}}{dW^{[2]}}}$$

compute from

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{\underline{z^{[2]} = W^{[2]}a^{[1]}}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

$$\frac{dJ(w)}{dz^{[2]}} \longleftarrow \frac{dJ(w)}{d\hat{y}}$$

Next consider: $\frac{dJ(w)}{dW^{[2]}}$

Backpropagation: *m*=1

from an earlier stage

Chain rule: $$\frac{dJ(w)}{da^{[1]}} = \boxed{\frac{dJ(w)}{dz^{[2]}}}\boxed{\frac{dz^{[2]}}{da^{[1]}}}$$

compute from

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{\underline{z^{[2]} = W^{[2]}a^{[1]}}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

Next consider:
$$\begin{tikzcd}
\frac{dJ(w)}{da^{[1]}} & \frac{dJ(w)}{dz^{[2]}} \arrow[l] \arrow[d] & \frac{dJ(w)}{d\hat{y}} \arrow[l] \\
& \frac{dJ(w)}{dW^{[2]}} &
\end{tikzcd}$$

Backpropagation: *m*=1

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

compute from

Next: $$\frac{dJ(w)}{dz^{[1]}} \longleftarrow \frac{dJ(w)}{da^{[1]}} \longleftarrow \frac{dJ(w)}{dz^{[2]}} \longleftarrow \frac{dJ(w)}{d\hat{y}}$$

$$\frac{dJ(w)}{dz^{[2]}} \longrightarrow \frac{dJ(w)}{dW^{[2]}}$$

$$= \boxed{\frac{dJ(w)}{da^{[1]}}}\,\boxed{\frac{da^{[1]}}{dz^{[1]}}}$$

from an earlier stage

Backpropagation: *m*=1

$$x \rightarrow \boxed{\underline{z^{[1]} = W^{[1]}x}} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

compute from

$$\frac{dJ(w)}{dz^{[1]}} \longleftarrow \frac{dJ(w)}{da^{[1]}} \longleftarrow \frac{dJ(w)}{dz^{[2]}} \longleftarrow \frac{dJ(w)}{d\hat{y}}$$

$$\frac{dJ(w)}{dz^{[2]}} \downarrow \frac{dJ(w)}{dW^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} \downarrow \frac{dJ(w)}{dW^{[1]}}$$

Next: $\frac{dJ(w)}{dW^{[1]}} = \boxed{\frac{dJ(w)}{dz^{[1]}}}\boxed{\frac{dz^{[1]}}{dW^{[1]}}}$

from an earlier stage

Backpropagation: *m*=1

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

$$\begin{tikzcd}
\frac{dJ(w)}{dz^{[1]}} \arrow[d] & \frac{dJ(w)}{da^{[1]}} \arrow[l] & \frac{dJ(w)}{dz^{[2]}} \arrow[l] \arrow[d] & \frac{dJ(w)}{d\hat{y}} \arrow[l] \\
\boxed{\frac{dJ(w)}{dW^{[1]}}} & & \boxed{\frac{dJ(w)}{dW^{[2]}}} &
\end{tikzcd}$$

Mathematical formula: *m*=1

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} = \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dz^{[1]}} x^T$$

See Appendix 1 for detailed derivation.

Mathematical formula: General *m*

$$\frac{dJ(w)}{dZ^{[2]}} = \hat{Y} - Y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dZ^{[2]}} A^{[1]T}$$

$$\frac{dJ(w)}{dA^{[1]}} = W^{[2]T} \frac{dJ(w)}{dZ^{[2]}}$$

$$\frac{dJ(w)}{dZ^{[1]}} = \frac{dJ(w)}{dA^{[1]}} .* \mathbf{1}\{Z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dZ^{[1]}} X^T$$

$$Y := \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(m)} \end{bmatrix}$$

$$\hat{Y} := \begin{bmatrix} \hat{y}^{(1)} & \hat{y}^{(2)} & \cdots & \hat{y}^{(m)} \end{bmatrix}$$

$$A^{[i]} := \begin{bmatrix} a^{[i],(1)} & a^{[i],(2)} & \cdots & a^{[i],(m)} \end{bmatrix}$$

$$Z^{[i]} := \begin{bmatrix} z^{[i],(1)} & z^{[i],(2)} & \cdots & z^{[i],(m)} \end{bmatrix}$$

$$X := \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(m)} \end{bmatrix}$$

See Appendix 2 for detailed derivation.

Mathematical formula: *L*-layer DNN

$$\frac{dJ(w)}{dZ^{[L]}} = \hat{Y} - Y$$

$$\frac{dJ(w)}{dW^{[L]}} = \frac{dJ(w)}{dZ^{[L]}} A^{[L-1]T}$$

$$\frac{dJ(w)}{dA^{[L-1]}} = W^{[L]T} \frac{dJ(w)}{dZ^{[L]}}$$

$$\frac{dJ(w)}{dZ^{[L-1]}} = \frac{dJ(w)}{dA^{[L-1]}} .* \mathbf{1}\{Z^{[L-1]} \geq 0\}$$

$$\vdots$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dZ^{[2]}} A^{[1]T}$$

$$\frac{dJ(w)}{dA^{[1]}} = W^{[2]T} \frac{dJ(w)}{dZ^{[2]}}$$

$$\frac{dJ(w)}{dZ^{[1]}} = \frac{dJ(w)}{dA^{[1]}} .* \mathbf{1}\{Z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dZ^{[1]}} X^{T}$$

Adam optimizer

A challenge

Recall gradient descent:

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \nabla \boxed{J(w^{(t)})}$$

$$J(w^{(t)}) := \frac{1}{m} \sum_{i=1}^{m} -y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

Computationally heavy for a large *m*.

Hence: Often use a part, called a *batch*.

batch1	batch2	batch3	$\cdots$	

$\underbrace{\hphantom{\text{batch1}}}_{m_{\mathcal{B}}}$ examples

Gradient descent with batch

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \nabla J(w^{(t)})$$

$$J(w^{(t)}) := \frac{1}{m_{\mathcal{B}}} \sum_{i=1}^{m_{\mathcal{B}}} -y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

batch1	batch2	batch3	$\cdots$	

$m_{\mathcal{B}}$ examples

$m_{\mathcal{B}} = 1$: Stochastic Gradient Descent (SGD)

Operation per batch is called “**step**”.

Operation per entire dataset is called “**epoch**”.

A challenge arises when $m_{\mathcal{B}}$ is small

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \boxed{\nabla J(w^{(t)})}$$

$$J(w^{(t)}) := \frac{1}{m_{\mathcal{B}}} \sum_{i=1}^{m_{\mathcal{B}}} -y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

Note: Relies only on the current gradient

The weight update may *oscillate too much*.

What we want is a "gradual (smooth) change".

To this end: Often use a variant of GD that exploits past gradients.

Momentum optimizer

$$w^{(t+1)} \leftarrow w^{(t)} + \alpha m^{(t)}$$

$$m^{(t)} \leftarrow \beta m^{(t-1)} + (1-\beta)(-\nabla J(w^{(t)}))$$

A typical choice: $\beta = 0.9$

Momentum optimizer

$$w^{(t+1)} \leftarrow w^{(t)} + \alpha \hat{m}^{(t)}$$

$$m^{(t)} \leftarrow \beta m^{(t-1)} - (1-\beta)\nabla J(w^{(t)})$$

If $\nabla J(w^{(t)})$ is too big or too small:

Yields quite different scalings

Motivate to normalize

Another variation

$$w^{(t+1)} \leftarrow w^{(t)} + \alpha \frac{\hat{m}^{(t)}}{\sqrt{\hat{s}^{(t)}}}$$

$$m^{(t)} \leftarrow \beta_1 m^{(t-1)} - (1-\beta_1)\nabla J(w^{(t)})$$

$$s^{(t)} \leftarrow \beta_2 s^{(t-1)} + (1-\beta_2)(\nabla J(w^{(t)}))^2$$

Called: Adam (Adaptive momentum) optimizer

A typical choice: $\beta_1 = 0.9, \quad \beta_2 = 0.999$

Look ahead

Will investigate advanced techniques.

Appendix 1:
Backpropagation (m=1)

Backpropagation: *m*=1

Recall the forward path:

$$x \longrightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \longrightarrow \hat{y}$$

Backpropagation: *m*=1

$$J(w) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

$$\frac{dJ(w)}{d\hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$$

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

Start from backward: $\frac{dJ(w)}{d\hat{y}}$

Backpropagation: *m*=1

$$\frac{dJ(w)}{d\hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} \qquad \frac{dJ(w)}{dz^{[2]}} = \frac{dJ(w)}{d\hat{y}}\frac{d\hat{y}}{dz^{[2]}} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)\hat{y}(1-\hat{y}) = \hat{y} - y$$

$$x \to \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \to \hat{y}$$

compute from $\hat{y}$

Next consider: $\frac{dJ(w)}{dz^{[2]}}$ $\qquad$ $\frac{dJ(w)}{d\hat{y}}$

Backpropagation: *m*=1

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y \qquad \frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} \frac{dz^{[2]}}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{\underline{z^{[2]} = W^{[2]}a^{[1]}}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

$$\frac{dJ(w)}{dz^{[2]}} \qquad \frac{dJ(w)}{d\hat{y}}$$

Next consider: $\frac{dJ(w)}{dW^{[2]}}$

Backpropagation: *m*=1

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = \frac{dJ(w)}{dz^{[2]}} \frac{dz^{[2]}}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$x \to \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{\underline{z^{[2]} = W^{[2]}a^{[1]}}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \to \hat{y}$$

$$W^{[2]}$$

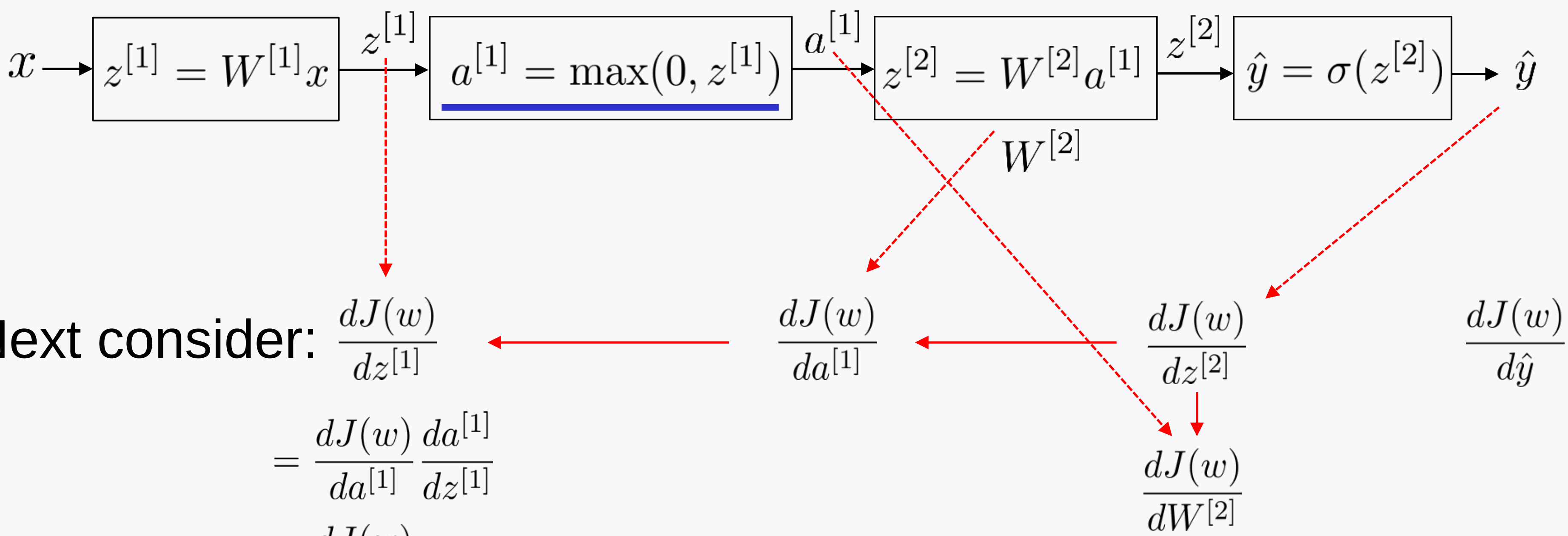
Next consider: $\frac{dJ(w)}{da^{[1]}}$ $\leftarrow$ $\frac{dJ(w)}{dz^{[2]}}$ $\quad \frac{dJ(w)}{d\hat{y}}$

$$\frac{dJ(w)}{dW^{[2]}}$$

Backpropagation: *m*=1

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y \qquad \frac{dJ(w)}{da^{[1]}} = W^{[2]T}\frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$



Next consider: $\frac{dJ(w)}{dz^{[1]}}$

$$= \frac{dJ(w)}{da^{[1]}} \frac{da^{[1]}}{dz^{[1]}}$$

$$= \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

component-wise multiplication

Backpropagation: *m*=1

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} = \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

$$W^{[2]}$$

$$\frac{dJ(w)}{dz^{[1]}} \longleftarrow \frac{dJ(w)}{da^{[1]}} \longleftarrow \frac{dJ(w)}{dz^{[2]}} \qquad \frac{dJ(w)}{d\hat{y}}$$

$$\frac{dJ(w)}{dW^{[2]}}$$

Next consider: $\frac{dJ(w)}{dW^{[1]}}$

$$= \frac{dJ(w)}{dz^{[1]}} \frac{dz^{[1]}}{dW^{[1]}} = \frac{dJ(w)}{dz^{[1]}} x^T$$

Backpropagation: *m*=1

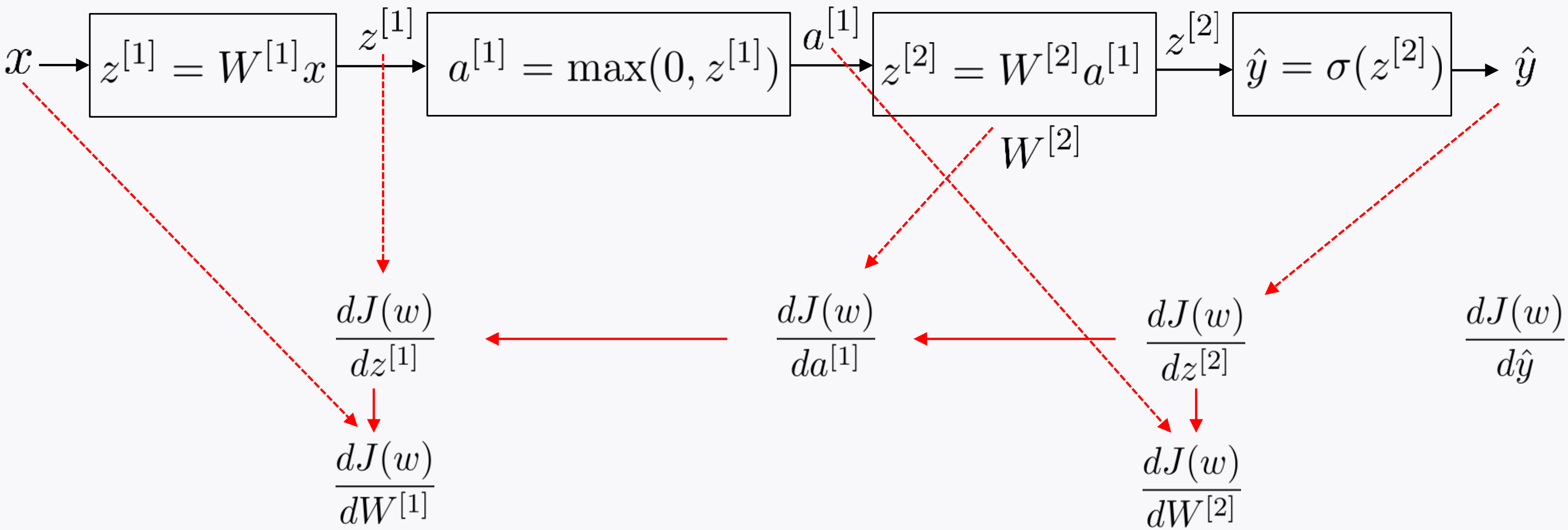
$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} = \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dz^{[1]}} x^T$$



Appendix 2:
Backpropagation (general *m*)

Backpropagation: General *m*

$$m = 1:$$

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} = \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dz^{[1]}} x^T$$

Matrix notation helps:

$$Y := \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(m)} \end{bmatrix}$$

$$\hat{Y} := \begin{bmatrix} \hat{y}^{(1)} & \hat{y}^{(2)} & \cdots & \hat{y}^{(m)} \end{bmatrix}$$

$$A^{[i]} := \begin{bmatrix} a^{[i],(1)} & a^{[i],(2)} & \cdots & a^{[i],(m)} \end{bmatrix}$$

$$Z^{[i]} := \begin{bmatrix} z^{[i],(1)} & z^{[i],(2)} & \cdots & z^{[i],(m)} \end{bmatrix}$$

$$X := \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(m)} \end{bmatrix}$$

Backpropagation: General *m*

$$m = 1:$$

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} = \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dz^{[1]}} x^T$$

Claim: $\text{general } m:$

$$\frac{dJ(w)}{dZ^{[2]}} = \hat{Y} - Y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dZ^{[2]}} A^{[1]T}$$

$$\frac{dJ(w)}{dA^{[1]}} = W^{[2]T} \frac{dJ(w)}{dZ^{[2]}}$$

$$\frac{dJ(w)}{dZ^{[1]}} = \frac{dJ(w)}{dA^{[1]}} .* \mathbf{1}\{Z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dZ^{[1]}} X^T$$

Proof

$$\hat{Y} := \begin{bmatrix} \hat{y}^{(1)} & \hat{y}^{(2)} & \cdots & \hat{y}^{(m)} \end{bmatrix}$$

$$\hat{y}^{(1)} = \sigma(z^{[2],(1)})$$

$$\boxed{\frac{dJ(w)}{dZ^{[2]}} = \hat{Y} - Y}$$

$$J(w) = \sum_{i=1}^{m} -y^{(i)} \log \hat{y}^{(i)} - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

$$\frac{dJ(w)}{d\hat{Y}} = \begin{bmatrix} -\frac{y^{(1)}}{\hat{y}^{(1)}} + \frac{1-y^{(1)}}{1-\hat{y}^{(1)}} & \cdots & -\frac{y^{(m)}}{\hat{y}^{(m)}} + \frac{1-y^{(m)}}{1-\hat{y}^{(m)}} \end{bmatrix}$$

$$\frac{dJ(w)}{dZ^{[2]}} = \frac{dJ(w)}{d\hat{Y}} \frac{d\hat{Y}}{dZ^{[2]}} = \begin{bmatrix} -\frac{y^{(1)}}{\hat{y}^{(1)}} + \frac{1-y^{(1)}}{1-\hat{y}^{(1)}} & \cdots & -\frac{y^{(m)}}{\hat{y}^{(m)}} + \frac{1-y^{(m)}}{1-\hat{y}^{(m)}} \end{bmatrix} \begin{bmatrix} \hat{y}^{(1)}(1-\hat{y}^{(1)}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{y}^{(m)}(1-\hat{y}^{(m)}) \end{bmatrix}$$

$$= \begin{bmatrix} \hat{y}^{(1)} - y^{(1)} & \cdots & \hat{y}^{(m)} - y^{(m)} \end{bmatrix}$$

$$= \hat{Y} - Y$$

Proof

$$Z^{[2]} = W^{[2]} A^{[1]}$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dZ^{[2]}} \frac{dZ^{[2]}}{dW^{[2]}}$$

$$= \frac{dJ(w)}{dZ^{[2]}} A^{[1]T}$$

$$\frac{dJ(w)}{dA^{[1]}} = \frac{dJ(w)}{dZ^{[2]}} \frac{dZ^{[2]}}{dA^{[1]}}$$

$$= W^{[2]T} \frac{dJ(w)}{dZ^{[2]}}$$

$$\boxed{\begin{aligned} \frac{dJ(w)}{dZ^{[2]}} &= \hat{Y} - Y \\ \frac{dJ(w)}{dW^{[2]}} &= \frac{dJ(w)}{dZ^{[2]}} A^{[1]T} \\ \frac{dJ(w)}{dA^{[1]}} &= W^{[2]T} \frac{dJ(w)}{dZ^{[2]}} \end{aligned}}$$

Proof

$$Z^{[1]} = W^{[1]}X$$

$$A^{[1]} = \max(0, Z^{[1]})$$

$$\frac{dJ(w)}{dZ^{[1]}} = \frac{dJ(w)}{dA^{[1]}}\frac{dA^{[1]}}{dZ^{[1]}}$$

$$= \frac{dJ(w)}{dA^{[1]}} .* \mathbf{1}\{Z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dZ^{[1]}}\frac{dZ^{[1]}}{dW^{[1]}}$$

$$= \frac{dJ(w)}{dZ^{[1]}} X^T$$

$$\frac{dJ(w)}{dZ^{[2]}} = \hat{Y} - Y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dZ^{[2]}} A^{[1]T}$$

$$\frac{dJ(w)}{dA^{[1]}} = W^{[2]T}\frac{dJ(w)}{dZ^{[2]}}$$

$$\frac{dJ(w)}{dZ^{[1]}} = \frac{dJ(w)}{dA^{[1]}} .* \mathbf{1}\{Z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dZ^{[1]}} X^T$$

2-layer DNN with bias terms

$$x \rightarrow \boxed{z^{[1]} = W^{[1]}x + b^{[1]}} \xrightarrow{z^{[1]}} \boxed{a^{[1]} = \max(0, z^{[1]})} \xrightarrow{a^{[1]}} \boxed{z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}} \xrightarrow{z^{[2]}} \boxed{\hat{y} = \sigma(z^{[2]})} \rightarrow \hat{y}$$

$$m = 1:$$

$$\frac{dJ(w)}{dz^{[2]}} = \hat{y} - y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dz^{[2]}} a^{[1]T}$$

$$\frac{dJ(w)}{da^{[1]}} = W^{[2]T} \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{dz^{[1]}} = \frac{dJ(w)}{da^{[1]}} .* \mathbf{1}\{z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dz^{[1]}} x^T$$

$$\frac{dJ(w)}{db^{[2]}} = \frac{dJ(w)}{dz^{[2]}} \frac{dz^{[2]}}{db^{[2]}} = \frac{dJ(w)}{dz^{[2]}}$$

$$\frac{dJ(w)}{db^{[1]}} = \frac{dJ(w)}{dz^{[1]}}$$

2-layer DNN with bias terms

$$X \rightarrow \boxed{Z^{[1]} = W^{[1]}X + b^{[1]}} \xrightarrow{Z^{[1]}} \boxed{A^{[1]} = \max(0, Z^{[1]})} \xrightarrow{A^{[1]}} \boxed{Z^{[2]} = W^{[2]}A^{[1]} + b^{[2]}} \xrightarrow{Z^{[2]}} \boxed{\hat{Y} = \sigma(Z^{[2]})} \rightarrow \hat{Y}$$

general m :

$$\frac{dJ(w)}{dZ^{[2]}} = \hat{Y} - Y$$

$$\frac{dJ(w)}{dW^{[2]}} = \frac{dJ(w)}{dZ^{[2]}} A^{[1]T}$$

$$\frac{dJ(w)}{dA^{[1]}} = W^{[2]T} \frac{dJ(w)}{dZ^{[2]}}$$

$$\frac{dJ(w)}{dZ^{[1]}} = \frac{dJ(w)}{dA^{[1]}} .* \mathbf{1}\{Z^{[1]} \geq 0\}$$

$$\frac{dJ(w)}{dW^{[1]}} = \frac{dJ(w)}{dZ^{[1]}} X^T$$

$$\begin{aligned}
\frac{dJ(w)}{db^{[2]}} &= \frac{dJ(w)}{dZ^{[2]}} \frac{dZ^{[2]}}{db^{[2]}} \\
&= \left[\left[\frac{dJ(w)}{dZ^{[2]}}\right]_1 \quad \cdots \quad \left[\frac{dJ(w)}{dZ^{[2]}}\right]_m \right] \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \\
&= \sum_{i=1}^{m} \left[\frac{dJ(w)}{dZ^{[2]}}\right]_i
\end{aligned}$$

$$\frac{dJ(w)}{db^{[1]}} = \sum_{i=1}^{m} \left[\frac{dJ(w)}{dZ^{[1]}}\right]_i$$