Computing gradients

Logistic regression

x_1
x_2 $\longrightarrow$ σ $\longrightarrow$ $\hat{y} = a$
x_3

x_1
$x_2 \longrightarrow$ $\begin{matrix} 0 \\ 0 \\ 0 \end{matrix} \to 0 \to \hat{y}$
x_3

x
w ┌─────────────────┐ ┌──────────────┐ ┌──────────┐
b │ $z = w^T x + b$ │ →│ $a = \sigma(z)$ │ →│ $\mathcal{L}(a, y)$ │
     └─────────────────┘ └──────────────┘ └──────────┘

$dz = a - y$

$da = \dfrac{d}{da} \mathcal{L}(a,y) = -y \log a - (1-y) \log(1-a)$

$dw = dz \cdot x$

$db = dz$

$\boxed{dz = da \cdot g'(z)}$

$g(z) = \sigma(z)$

$= -\dfrac{y}{a} + \dfrac{1-y}{1-a}$

$\boxed{\dfrac{\partial \mathcal{L}}{\partial z} = \dfrac{\partial \mathcal{L}}{\partial a} \cdot \boxed{\dfrac{da}{dz}}}$

"dz" = "da" $\cdot \dfrac{d}{dz} g(z) = \underline{g'(z)}$

Andrew Ng
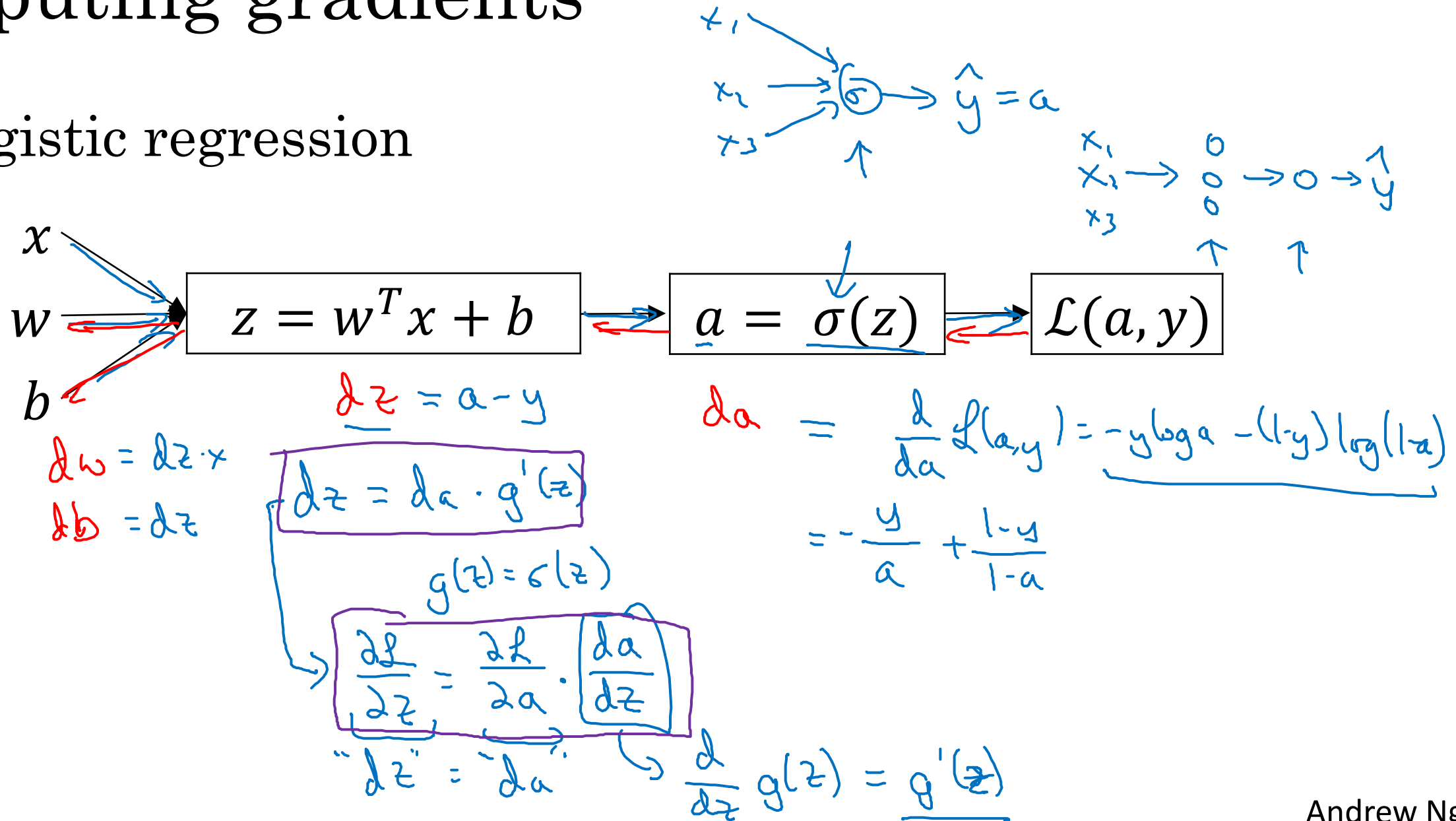
Neural network gradients

$x_1 \rightarrow \bigcirc$
$x_2 \rightarrow \bigcirc \rightarrow \bigcirc \rightarrow \hat{y}$
$x_3 \rightarrow \bigcirc$

$n_x = n^{[0]}$ $n^{[1]}$ $n^{[2]} = 1$

$W^{[2]}$ $dW^{[2]}$

$b^{[2]}$ $db^{[2]}$

x

$W^{[1]}$

| $z^{[1]} = W^{[1]}x + b^{[1]}$ | $a^{[1]} = \sigma(z^{[1]})$ | $z^{[2]} = W^{[2]}x + b^{[2]}$ | $a^{[2]} = \sigma(z^{[2]})$ | $\mathcal{L}(a^{[2]}, y)$ |

$dW^{[1]}$

$b^{[1]}$

$db^{[1]}$

$\rightarrow dz^{[1]} = W^{[2]T} dz^{[2]}$ $da^{[1]} = \cdots$ $\rightarrow dz^{[2]} = a^{[2]} - y$ $da^{[2]}$

$* g^{[1]'}(z^{[1]})$
element wise product

$\rightarrow dW^{[2]} = dz^{[2]} a^{[1]T}$ $\rightarrow "dw = dz \cdot x"$

$\rightarrow db^{[2]} = dz^{[2]}$

$W^{[2]}$ $(n^{[2]}, n^{[1]})$

$W : [\underline{\qquad}]$

$\rightarrow z^{[2]}, dz^{[2]}$ $(n^{[2]}, 1) - (1,1)$

$\rightarrow z^{[1]}, dz^{[1]}$ $(n^{[1]}, 1)$

foo $W^{[1]}$ $\rightarrow dW^{[1]} = dz^{[1]} \cdot x^T$

$dfoo$ $dW^{[1]}$ $\rightarrow db^{[1]} = dz^{[1]}$ $a^{[0]T}$

$dz^{[1]} = \boxed{W^{[2]T} dz^{[2]}} * \boxed{g^{[1]'}(z^{[1]})}$

$\underline{(n^{[1]}, 1)} = \underline{(n^{[1]}, n^{[2]})} \ \underline{(n^{[2]}, 1)} * \underline{(n^{[1]}, 1)}$

Andrew Ng
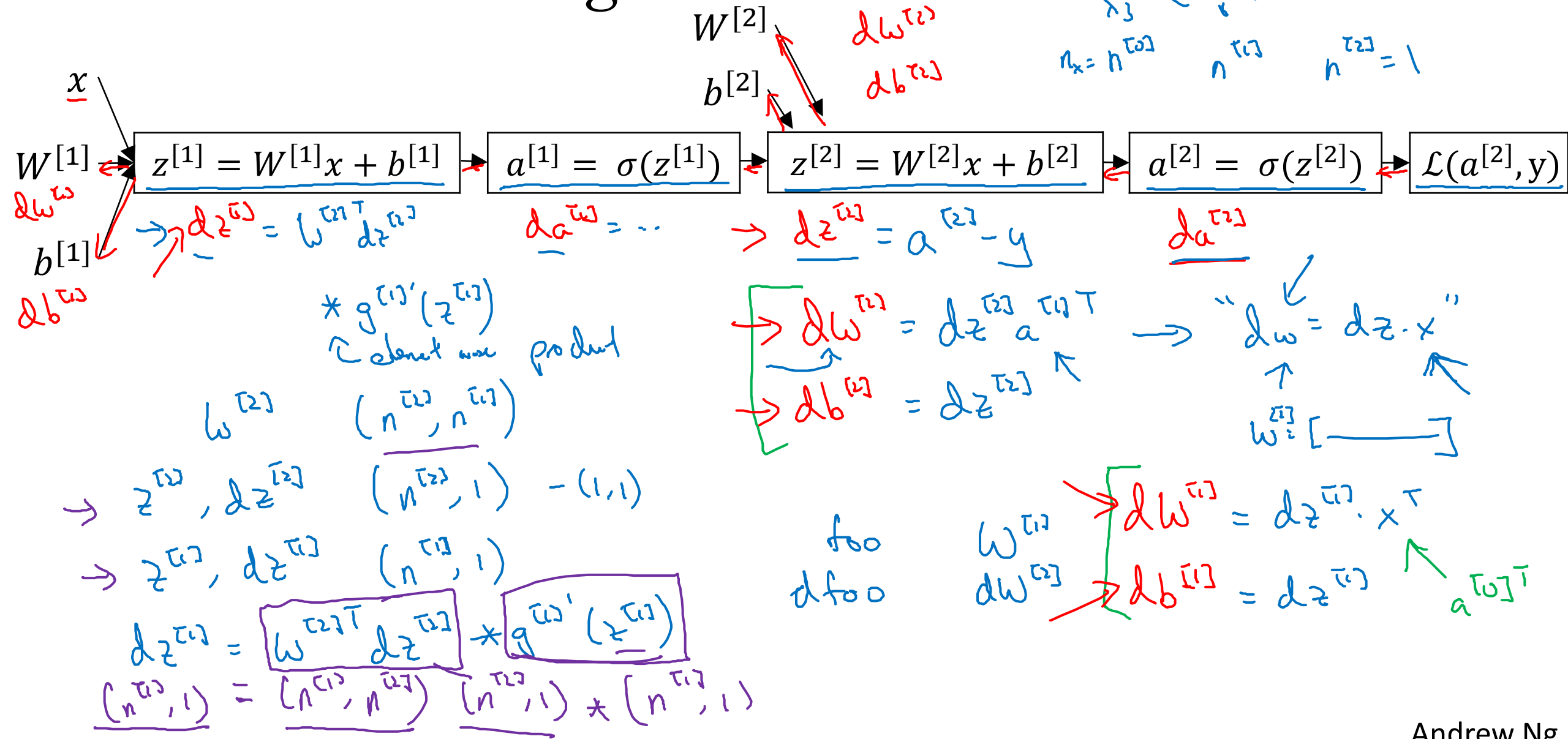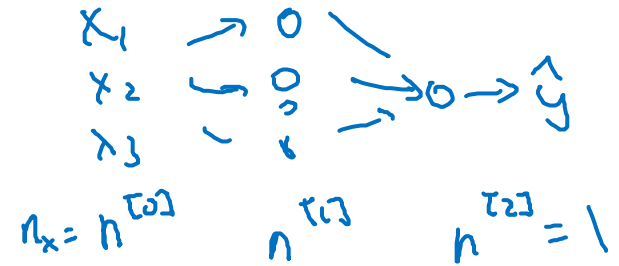
Summary of gradient descent

$$dz^{[2]} = a^{[2]} - y$$

$$dW^{[2]} = dz^{[2]} a^{[1]^T}$$

$$db^{[2]} = dz^{[2]}$$

$$dz^{[1]} = W^{[2]^T} dz^{[2]} * g^{[1]'}(z^{[1]})$$

$$dW^{[1]} = dz^{[1]} x^T$$

$$db^{[1]} = dz^{[1]}$$

Vectorized Implementation:

$$z^{[1]} = W^{[1]} x + b^{[1]}$$

$$a^{[1]} = g^{[1]}(z^{[1]})$$

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

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

$$A^{[1]} = g^{[1]}(Z^{[1]})$$

Andrew Ng

Summary of gradient descent

$$dz^{[2]} = a^{[2]} - y$$

$$dW^{[2]} = dz^{[2]} a^{[1]T}$$

$$db^{[2]} = dz^{[2]}$$

$$dz^{[1]} = W^{[2]T} dz^{[2]} * g^{[1]\prime}(z^{[1]})$$
$$(n^{[1]}, 1)$$

$$dW^{[1]} = dz^{[1]} x^T$$

$$db^{[1]} = dz^{[1]}$$

$$dZ^{[2]} = A^{[2]} - Y$$

$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]T}$$

$$db^{[2]} = \frac{1}{m} np.sum(dZ^{[2]}, axis = 1, keepdims = True)$$

$$J(\ldots) = \frac{1}{m} \sum_{i=1}^{n} \mathcal{L}(\hat{y}, y)$$

$$dZ^{[1]} = \underbrace{W^{[2]T} dZ^{[2]}}_{(n^{[1]}, m)} * \underbrace{g^{[1]\prime}(Z^{[1]})}_{(n^{[1]}, m)}$$
$$(n^{[1]}, m)$$

elementwise product

$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^T$$

$$db^{[1]} = \frac{1}{m} np.sum(dZ^{[1]}, axis = 1, keepdims = True)$$

Andrew Ng