DeepLearning.AI
Stanford ONLINE

Backprop Intuition (Optional)

What is a derivative?

Informal Definition of Derivative

If $w \uparrow \varepsilon$ causes $J(w) \uparrow k \times \varepsilon$ then

$$\frac{\partial}{\partial w} J(w) = k \rightarrow 6$$

Gradient descent
```
repeat {
    $w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(\vec{w}, b)$    learning rate
}
```

If derivative is small, then this update step will make a small update to w_j

If the derivative is large, then this update step will make a large update to w_j

---

Derivative Example

Cost function $J(w) = w^2$

Say $w = 3$ $J(w) = 3^2 = 9$

If we increase w by a tiny amount $\varepsilon = 0.001$ ($\varepsilon = 0.002$) how does $J(w)$ change?

$w = 3 + \cancel{0.001}$ 0.002

$J(w) = w^2 = \cancel{9.006001}$ 9.012004

9.012

If $w \uparrow \cancel{0.001}$ 0.002 ε

$J(w) \uparrow 6 \times \cancel{0.001}$ 0.002 $6 \times \varepsilon$ $6 \times 0.002 = 0.012$

$\frac{\partial}{\partial w} J(w) = 6$

More Derivative Examples

$w = 3$ $J(w) = w^2 = 9$ $w \uparrow 0.001$ $J(w) = J(3.001) = 9.006001$ $\frac{\partial}{\partial w} J(w) = 6$

$J(w) \uparrow 6 \times 0.001$

$w = 2$ $J(w) = w^2 = 4$ $w \uparrow 0.001$ $J(w) = J(2.001) = 4.004001$ $\frac{\partial}{\partial w} J(w) = 4$

$J(w) \uparrow 4 \times 0.001$

$J(w) = w^2$

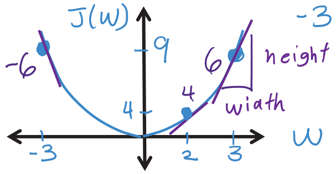
$\downarrow 0.006$

$w = -3$ $J(w) = w^2 = 9$ $w \uparrow 0.001$ $J(w) = J(-2.999) = 8.994001$ $\frac{\partial}{\partial w} J(w) = -6$

$-3 + 0.001$

$J(w) \downarrow 6 \times 0.001$

$J(w) \uparrow -6 \times 0.001$



Calculus

$\frac{\partial}{\partial w} J(w) = 2w$

w	$\partial J(w)/\partial w$
3	$2 \times 3 = 6$
2	$2 \times 2 = 4$
-3	$2 \times -3 = -6$

Even More Derivative Examples

$w = 2$

$J(w) = w^2 = 4$	$\frac{\partial}{\partial w}J(w) = 2w = 4$	$w \uparrow \underbrace{0.001}_{\varepsilon}$	$J(w) = 4.004001$ $J(w) \uparrow 4 \times \varepsilon$
$J(w) = w^3 = 8$	$\frac{\partial}{\partial w}J(w) = 3w^2 = 12$	$w \uparrow \varepsilon$	$J(w) = 8.012006$ $J(w) \uparrow 12 \times \varepsilon$
$J(w) = w = 2$	$\frac{\partial}{\partial w}J(w) = 1$	$w \uparrow \varepsilon$	$J(w) = 2.001$ $J(w) \uparrow 1 \times \varepsilon$
$J(w) = \frac{1}{w} = \frac{1}{2} = 0.5$	$\frac{\partial}{\partial w}J(w) = -\frac{1}{w^2} = -\frac{1}{4}$	$w \uparrow \varepsilon$ $w = \frac{1}{2.001}$	-0.25×0.001 $0.5 - 0.00025$ $J(w) = 0.49975$ $J(w) \uparrow -\frac{1}{4} \times \varepsilon$

$\frac{\partial}{\partial w}J(w)$ $\quad w \uparrow \varepsilon \quad$ $J(w) \uparrow k \times \varepsilon$

Backprop Intuition (Optional)

Computation Graph

A note on derivative notation

If $J(w)$ is a function of one variable (w),

$d \qquad \frac{d}{dw}J(w)$

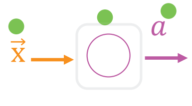
If $J(w_1, w_2, \ldots, w_n)$ is a function of more than one variable,

$\partial \qquad \frac{\partial}{\partial w_i}J(w_1, w_2, \ldots, w_n) \qquad \frac{\partial J}{\partial w_i}$ or $\frac{\partial}{\partial w_i}J$

"partial derivative"

notation used in these courses

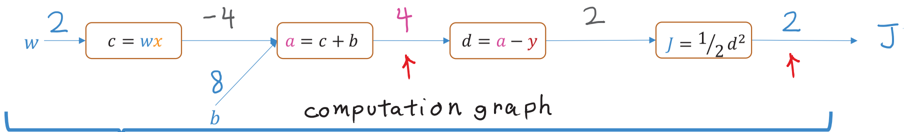
Small Neural Network Example



$w = 2 \quad b = 8 \quad x = -2 \quad y = 2$

$a = wx + b$ linear activation $a = g(z) = z$

$J(w,b) = \frac{1}{2}(a-y)^2$



computation graph

forward prop

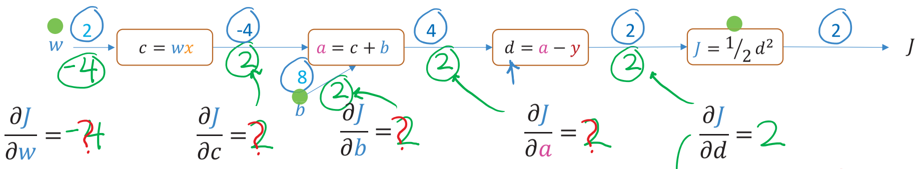
how do we find the derivatives of J? $\frac{\partial J}{\partial w} \quad \frac{\partial J}{\partial b}$

Computing the Derivatives

Forward prop →

← Back prop

$w = 2$ $b = 8$ $x = -2$ $y = 2$ $a = wx + b$ $J = \frac{1}{2}(a-y)^2$



$\frac{\partial J}{\partial w} = -4$? $\frac{\partial J}{\partial c} = 2$? $\frac{\partial J}{\partial b} = 2$? $\frac{\partial J}{\partial a} = 2$? $\frac{\partial J}{\partial d} = 2$

w=2.001 c=-4.002

w ↑ 0.001 c ↓ 2×0.001

c ↑ −2×0.001

J ↑ −4×0.001

$\frac{\partial J}{\partial w} = \underset{-2}{\frac{\partial c}{\partial w}} \times \underset{2}{\frac{\partial J}{\partial c}}$

c ↑ 0.001 a ↑ 0.001 J ↑ 0.002

$\frac{\partial J}{\partial c} = \underset{1}{\frac{\partial a}{\partial c}} \times \underset{2}{\frac{\partial J}{\partial a}}$

b ↑ 0.001 a ↑ 0.001 J ↑ 0.002

$\frac{\partial J}{\partial b} = \underset{1}{\frac{\partial a}{\partial b}} \times \underset{2}{\frac{\partial J}{\partial a}}$

a ↑ 0.001 d ↑ 0.001

a=4.001 d=2.001

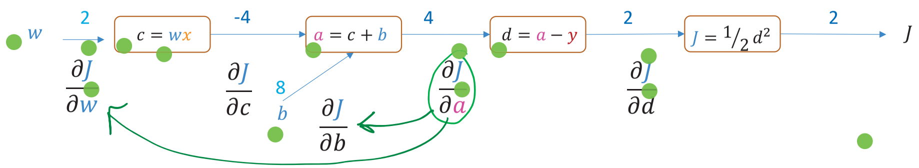
J ↑ 0.002

$\frac{\partial J}{\partial a} = \underset{1}{\frac{\partial d}{\partial a}} \times \underset{2}{\frac{\partial J}{\partial d}}$

$d \uparrow \underbrace{0.001}_{\varepsilon}$? $J \uparrow \underbrace{0.002}_{2\varepsilon}$

chain rule
(calculus)

Backprop is an efficient way to compute derivatives



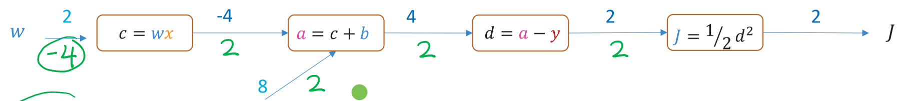
Compute $\frac{\partial J}{\partial a}$ once and use it to compute both $\frac{\partial J}{\partial w}$ and $\frac{\partial J}{\partial b}$.

If N nodes and P parameters, compute derivatives in roughly $N + P$ steps rather than $N \times P$ steps.

N	P	N+P	N×P
10,000	100,000	1.1×10^5	10^9

Computing the Derivatives

$w = 2$ $b = 8$ $x = -2$ $y = 2$ $a = wx + b$ $J = \frac{1}{2}(a-y)^2$



$\frac{\partial J}{\partial w} = -4$ $\frac{\partial J}{\partial c} = 2$ $\frac{\partial J}{\partial b} = 2$ $\frac{\partial J}{\partial a} = 2$ $\frac{\partial J}{\partial d} = 2$

$J = \frac{1}{2}\left((wx+b)-y\right)^2 = \frac{1}{2}\left((2\times -2+8)-2\right)^2 = 2$

w ↑ 0.001 $J = \frac{1}{2}\left((2.001\times -2+8)-2\right)^2 = 1.996002$

J ↓ 4×0.001

J ↑ −4×0.001

$\frac{\partial J}{\partial w} = -4$

Backprop Intuition (Optional)

Larger Neural Network Example

Neural Network Example

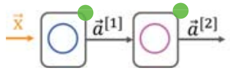
$x = 1 \quad y = 5$

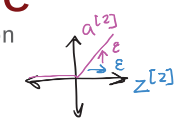
$w^{[1]} = 2, b^{[1]} = 0$

$w^{[2]} = 3, b^{[2]} = 1$

ReLU activation

$g(z) = \max(0, z)$

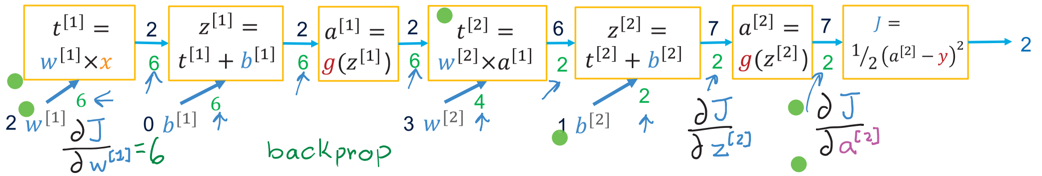




$$a^{[1]} = g(w^{[1]} x + b^{[1]}) = \overbrace{w^{[1]} x + b^{[1]}}^{z^{[1]}} = 2\times1 + 0 = 2$$

$$a^{[2]} = g(w^{[2]} a^{[1]} + b^{[2]}) = \overbrace{w^{[2]} a^{[1]} + b^{[2]}}^{z^{[2]}} = 3\times2 + 1 = 7$$

$$J(w,b) = \frac{1}{2}(a^{[2]} - y)^2 = \frac{1}{2}(7-5)^2 = 2$$



$\frac{\partial J}{\partial w^{[1]}} = 6$

backprop

if $w^{[1]}\uparrow \varepsilon$, then $J \uparrow 6 \times \varepsilon$ let's verify this!

Neural Network Example

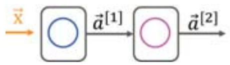
$x = 1 \quad y = 5$

$w^{[1]} = 2, b^{[1]} = 0$

$w^{[2]} = 3, b^{[2]} = 1$

ReLU activation

$g(z) = \max(0, z)$



$$a^{[1]} = g(w^{[1]} x + b^{[1]}) = w^{[1]} x + b^{[1]} = \overset{2.001}{\cancel{2}}\times1 + 0 = \overset{2.001}{\cancel{2}}$$

$$a^{[2]} = g(w^{[2]} a^{[1]} + b^{[2]}) = w^{[2]} a^{[1]} + b^{[2]} = 3\times\overset{2.001}{\cancel{2}} + 1 = \cancel{7}\ 7.003$$

$$J(w,b) = \frac{1}{2}(a^{[2]} - y)^2 = \frac{1}{2}(\overset{7.003}{\cancel{7}} - 5)^2 = \cancel{2}\ \ 2.\boxed{006}005$$

$$\frac{(2.003)^2}{2}$$

$w^{[1]}\uparrow 0.001$

$J \uparrow 6 \times 0.001$

$\frac{\partial J}{\partial w^{[1]}} = 6$

$\frac{\partial J}{\partial w^{[1]}} \quad \frac{\partial J}{\partial b^{[1]}}$

$\frac{\partial J}{\partial w^{[2]}} \quad \frac{\partial J}{\partial b^{[2]}}$

N nodes □→□→□

P parameters $w_1, b_1, w_2, b_2 \dots$

inefficient way

N × P

efficient way (backprop)

N + P