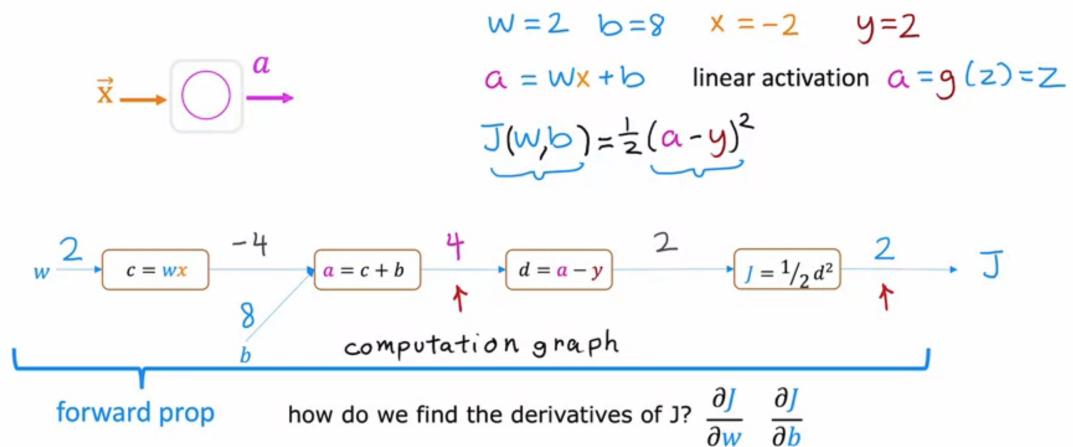


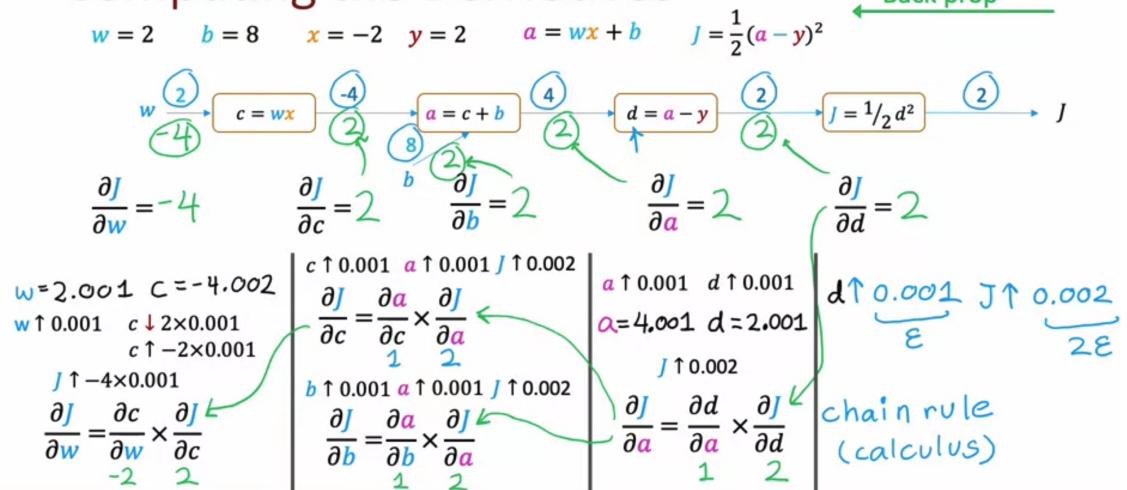
Optional Lab: Back propagation using a computation graph

Working through this lab will give you insight into a key algorithm used by most machine learning frameworks. Gradient descent requires the derivative of the cost with respect to each parameter in the network. Neural networks can have millions or even billions of parameters. The *back propagation* algorithm is used to compute those derivatives. *Computation graphs* are used to simplify the operation. Let's dig into this below.

Small Neural Network Example

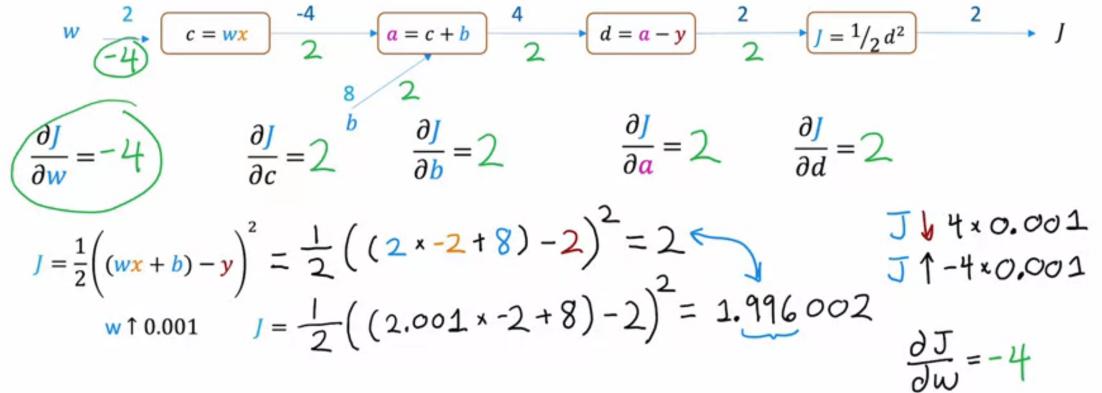


Computing the Derivatives

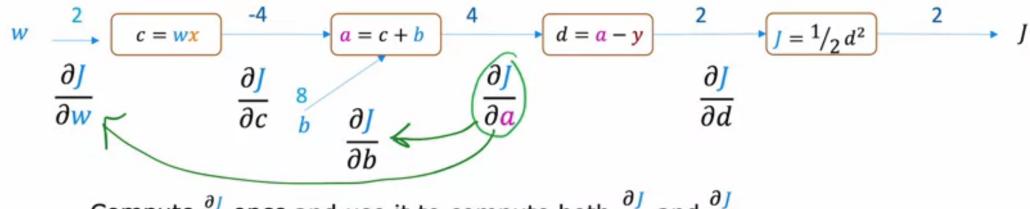


Computing the Derivatives

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



Backprop is an efficient way to compute derivatives

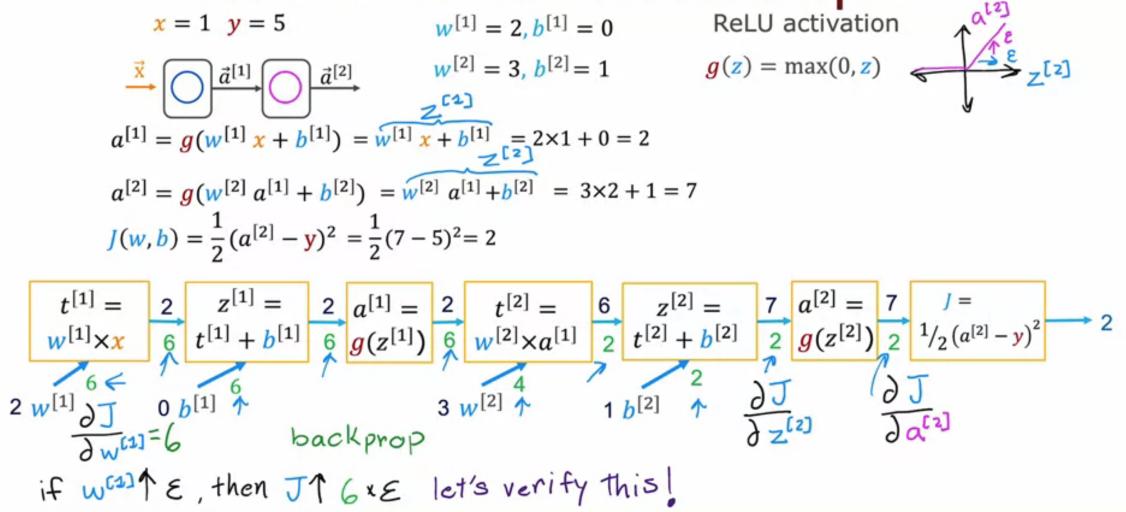


If N nodes and P parameters, compute derivatives

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

Neural Network Example



Neural Network Example

$x = 1 \quad y = 5$ $w^{[1]} = 2, b^{[1]} = 0$ ReLU activation
 $w^{[2]} = 3, b^{[2]} = 1$ $g(z) = \max(0, z)$

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

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

$J(w, b) = \frac{1}{2}(a^{[2]} - y)^2 = \frac{1}{2}(7 - 5)^2 = 2.006005$

$\frac{\partial J}{\partial w^{[1]}} \quad \frac{\partial J}{\partial b^{[1]}}$ N nodes $\square \rightarrow \square \rightarrow \square$
 $\frac{\partial J}{\partial w^{[2]}} \quad \frac{\partial J}{\partial b^{[2]}}$ P parameters $w_1, b_1, w_2, b_2 \dots$

$w^{[1]} \uparrow 0.001 \quad J \uparrow 6 \times 0.001 \quad \frac{\partial J}{\partial w^{[1]}} = 6$

inefficient way
 $N \times P$
efficient way (backprop)
 $N + P$

```
In [1]: from sympy import *
import numpy as np
import re
%matplotlib widget
import matplotlib.pyplot as plt
from matplotlib.widgets import TextBox
from matplotlib.widgets import Button
import ipywidgets as widgets
from lab_utils_backprop import *
```

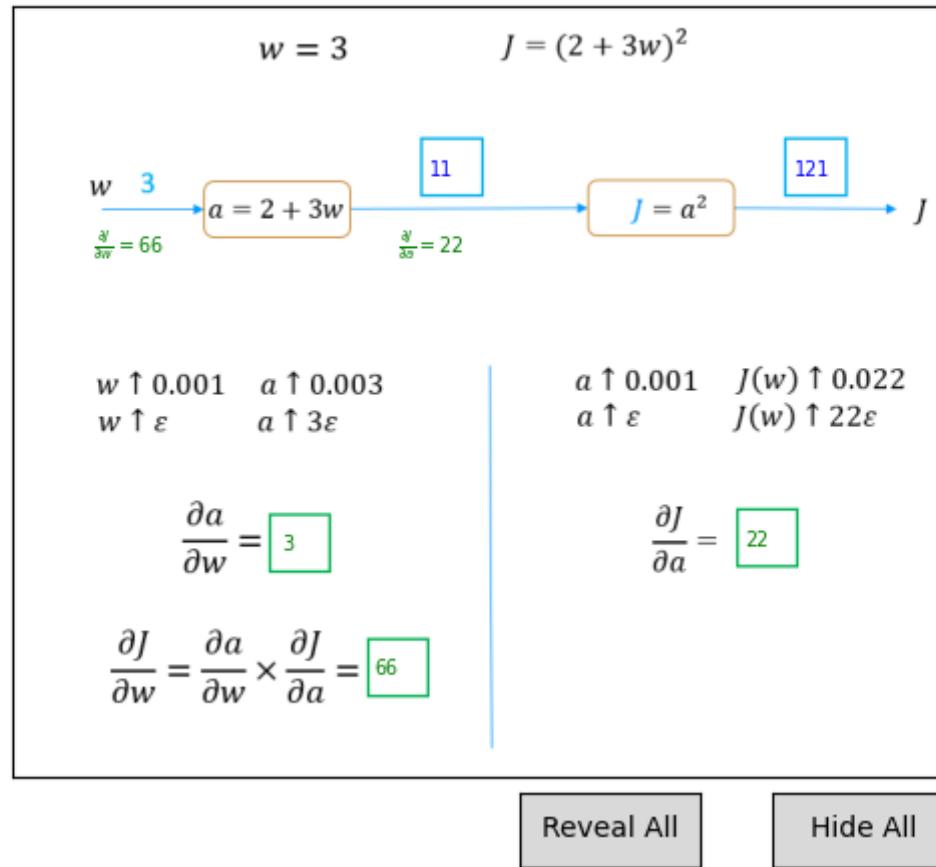
Computation Graph

A computation graph simplifies the computation of complex derivatives by breaking them into smaller steps. Let's see how this works.

Let's calculate the derivative of this slightly complex expression, $J = (2 + 3w)^2$. We would like to find the derivative of J with respect to w or $\frac{\partial J}{\partial w}$.

```
In [2]: plt.close("all")
plt_network(config_nw0, "./images/C2_W2_BP_network0.PNG")
```

Click in boxes to fill in values.



[Reveal All](#)

[Hide All](#)

Out [2]: <lab_utils_backprop=plt_network at 0x7f61685e0710>

Above, you can see we broke the expression into two nodes which we can work on independently. If you already have a good understanding of the process from the lecture, you can go ahead and fill in the boxes in the diagram above. You will want to first fill in the blue boxes going left to right and then fill in the green boxes starting on the right and moving to the left. If you have the correct values, the values will show as green or blue. If the value is incorrect, it will be red. Note, the interactive graphic is not particularly robust. If you run into trouble with the interface, run the cell above again to restart.

If you are unsure of the process, we will work this example step by step below.

Forward Propagation

Let's calculate the values in the forward direction.

Just a note about this section. It uses global variables and reuses them as the calculation progresses. If you run cells out of order, you may get funny results. If you do, go back to this point and run them in order.

```
In [3]: w = 3
a = 2+3*w
J = a**2
print(f"a = {a}, J = {J}")
```

a = 11, J = 121

You can fill these values in the blue boxes above.

Backprop

$$J = a^2$$

Backprop is the algorithm we use to calculate derivatives. As described in the lectures, backprop starts at the right and moves to the left. The first node to consider is $J = a^2$ and the first step is to find $\frac{\partial J}{\partial a}$

$$\frac{\partial J}{\partial a}$$

Arithmetically

Find $\frac{\partial J}{\partial a}$ by finding how J changes as a result of a little change in a . This is described in detail in the derivatives optional lab.

```
In [4]: a_epsilon = a + 0.001      # a epsilon
J_epsilon = a_epsilon**2      # J_epsilon
k = (J_epsilon - J)/0.001    # difference divided by epsilon
print(f"J = {J}, J_epsilon = {J_epsilon}, dJ_da ~ k = {k} ")
```

J = 121, J_epsilon = 121.0220009999999, dJ_da ~ k = 22.00099999999988835

$\frac{\partial J}{\partial a}$ is 22 which is $2 \times a$. Our result is not exactly $2 \times a$ because our epsilon value is not infinitesimally small.

Symbolically

Now, let's use SymPy to calculate derivatives symbolically as we did in the derivatives optional lab. We will prefix the name of the variable with an 's' to indicate this is a *symbolic* variable.

```
In [5]: sw,sJ,sa = symbols('w,J,a')
sJ = sa**2
sJ
```

Out [5]: a^2

```
In [6]: sJ.subs([(sa,a)])
```

Out [6]: 121

In [7]: `dJ_da = diff(sJ, sa)`

Out[7]: $2a$

So, $\frac{\partial J}{\partial a} = 2a$. When $a = 11$, $\frac{\partial J}{\partial a} = 22$. This matches our arithmetic calculation above. If you have not already done so, you can go back to the diagram above and fill in the value for $\frac{\partial J}{\partial a}$.

$$\frac{\partial J}{\partial w}$$

$$a = 2 + 3w$$

Moving from right to left, the next value we would like to compute is $\frac{\partial J}{\partial w}$.

To do this, we first need to calculate $\frac{\partial a}{\partial w}$ which describes how the output of this node, a , changes when the input w changes a little bit.

Arithmetically

Find $\frac{\partial a}{\partial w}$ by finding how a changes as a result of a little change in w .

In [8]: `w_epsilon = w + 0.001 # a plus a small value, epsilon
a_epsilon = 2 + 3*w_epsilon
k = (a_epsilon - a)/0.001 # difference divided by epsilon
print(f"a = {a}, a_epsilon = {a_epsilon}, da_dw ~ k = {k} ")`

$a = 11, a_{\text{epsilon}} = 11.003, da_{\text{dw}} \approx k = 3.0000000000001137$

Calculated arithmetically, $\frac{\partial a}{\partial w} \approx 3$. Let's try it with SymPy.

In [9]: `sa = 2 + 3*sw
sa`

Out[9]: $3w + 2$

In [10]: `da_dw = diff(sa, sw)
da_dw`

Out[10]: 3

The next step is the interesting part:

- We know that a small change in w will cause a to change by 3 times that amount.
- We know that a small change in a will cause J to change by $2 \times a$ times that amount. ($a=11$ in this example)
so, putting these together,
- We know that a small change in w will cause J to change by $3 \times 2 \times a$ times that amount.

These cascading changes go by the name of *the chain rule*. It can be written like this:

$$\frac{\partial J}{\partial w} = \frac{\partial a}{\partial w} \frac{\partial J}{\partial a}$$

It's worth spending some time thinking this through if it is not clear. This is a key take-away.

Let's try calculating it:

In [11]: `dJ_dw = da_dw * dJ_da
dJ_dw`

Out[11]: `6a`

And a is 11 in this example so $\frac{\partial J}{\partial w} = 66$. We can check this arithmetically:

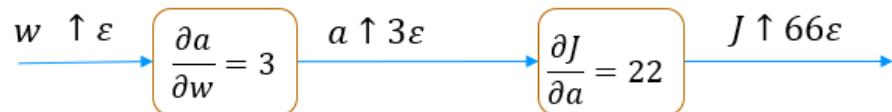
In [12]: `w_epsilon = w + 0.001
a_epsilon = 2 + 3*w_epsilon
J_epsilon = a_epsilon**2
k = (J_epsilon - J)/0.001 # difference divided by epsilon
print(f"J = {J}, J_epsilon = {J_epsilon}, dJ_dw ~ k = {k} ")`

`J = 121, J_epsilon = 121.0660090000001, dJ_dw ~ k = 66.0090000000082`

OK! You can now fill the values for $\frac{\partial a}{\partial w}$ and $\frac{\partial J}{\partial w}$ in the diagram if you have not already done so.

Another view

One could visualize these cascading changes this way:



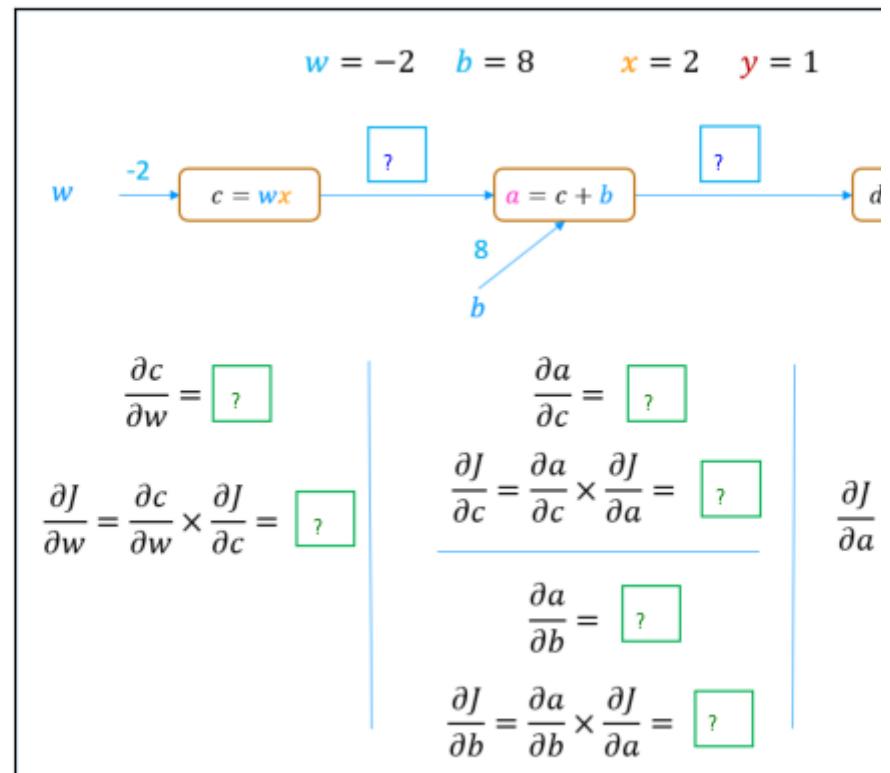
A small change in w is multiplied by $\frac{\partial a}{\partial w}$ resulting in a change that is 3 times as large. This larger change is then multiplied by $\frac{\partial J}{\partial a}$ resulting in a change that is now $3 \times 22 = 66$ times larger.

Computation Graph of a Simple Neural Network

Below is a graph of the neural network used in the lecture with different values. Try and fill in the values in the boxes. Note, the interactive graphic is not particularly robust. If you run into trouble with the interface, run the cell below again to restart.

```
In [13]: plt.close("all")
plt_network(config_nw1, "./images/C2_W2_BP_network1.PNG")
```

Click in boxes to fill in values.



```
Out[13]: <lab_utils_backprop=plt_network at 0x7f61497ffdd0>
```

Below, we will go through the computations required to fill in the above computation graph in detail. We start with the forward path.

Forward propagation

The calculations in the forward path are the ones you have recently learned for neural networks. You can compare the values below to those you calculated for the diagram above.

```
In [14]: # Inputs and parameters
x = 2
w = -2
b = 8
y = 1
# calculate per step values
c = w * x
a = c + b
d = a - y
J = d**2/2
print(f"J={J}, d={d}, a={a}, c={c}")
```

J=4.5, d=3, a=4, c=-4

Backward propagation (Backprop)

$J = \frac{1}{2}d^2$ As described in the lectures, backprop starts at the right and moves to the left. The first node to consider is $J = \frac{1}{2}d^2$ and the first step is to find $\frac{\partial J}{\partial d}$

$$\frac{\partial J}{\partial d}$$

Arithmetically

Find $\frac{\partial J}{\partial d}$ by finding how J changes as a result of a little change in d .

```
In [15]: d_epsilon = d + 0.001
J_epsilon = d_epsilon**2/2
k = (J_epsilon - J)/0.001  # difference divided by epsilon
print(f"J = {J}, J_epsilon = {J_epsilon}, dJ_dd ~ k = {k} ")
```

J = 4.5, J_epsilon = 4.5030005, dJ_dd ~ k = 3.000499999997395

$\frac{\partial J}{\partial d}$ is 3, which is the value of d . Our result is not exactly d because our epsilon value is not infinitesimally small.

Symbolically

Now, let's use SymPy to calculate derivatives symbolically, as we did in the derivatives optional lab. We will prefix the name of the variable with an 's' to indicate this is a *symbolic* variable.

```
In [16]: sx, sw, sb, sy, sJ = symbols('x,w,b,y,J')
sa, sc, sd = symbols('a,c,d')
sJ = sd**2/2
sJ
```

```
Out[16]:  $\frac{d^2}{2}$ 
```

```
In [17]: sJ.subs([(sd,d)])
```

```
Out[17]:  $\frac{9}{2}$ 
```

```
In [18]: dJ_dd = diff(sJ, sd)
dJ_dd
```

```
Out[18]: d
```

So, $\frac{\partial J}{\partial d} = d$. When $d = 3$, $\frac{\partial J}{\partial d} = 3$. This matches our arithmetic calculation above. If you have not already done so, you can go back to the diagram above and fill in the value for $\frac{\partial J}{\partial d}$.

$$\frac{\partial J}{\partial a}$$

$d = a - y$ Moving from right to left, the next value we would like to compute is $\frac{\partial J}{\partial a}$.

To do this, we first need to calculate $\frac{\partial d}{\partial a}$ which describes how the output of this node changes when the input a changes a little bit. (Note, we are not interested in how the output changes when y changes since y is not a parameter.)

Arithmetically

Find $\frac{\partial d}{\partial a}$ by finding how d changes as a result of a little change in a .

```
In [19]: a_epsilon = a + 0.001      # a plus a small value
d_epsilon = a_epsilon - y
k = (d_epsilon - d)/0.001    # difference divided by epsilon
print(f"d = {d}, d_epsilon = {d_epsilon}, dd_da ~≈ k = {k} ")
```

```
d = 3, d_epsilon = 3.001000000000003, dd_da ≈ k = 1.000000000000003
34
```

Calculated arithmetically, $\frac{\partial d}{\partial a} \approx 1$. Let's try it with SymPy.

Symbolically

```
In [20]: sd = sa - sy
sd
```

```
Out[20]: a - y
```

In [21]: `dd_da = diff(sd,sa)`
`dd_da`

Out[21]: 1

Calculated arithmetically, $\frac{\partial d}{\partial a}$ also equals 1.

The next step is the interesting part, repeated again in this example:

- We know that a small change in a will cause d to change by 1 times that amount.
- We know that a small change in d will cause J to change by d times that amount. ($d=3$ in this example)
so, putting these together,
- We know that a small change in a will cause J to change by $1 \times d$ times that amount.

This is again *the chain rule*. It can be written like this:

$$\frac{\partial J}{\partial a} = \frac{\partial d}{\partial a} \frac{\partial J}{\partial d}$$

Let's try calculating it:

In [22]: `dJ_da = dd_da * dJ_dd`
`dJ_da`

Out[22]: `d`

And d is 3 in this example so $\frac{\partial J}{\partial a} = 3$. We can check this arithmetically:

In [24]: `a_epsilon = a + 0.001`
`d_epsilon = a_epsilon - y`
`J_epsilon = d_epsilon**2/2`
`k = (J_epsilon - J)/0.001`
`print(f"J = {J}, J_epsilon = {J_epsilon}, dJ_da ~ k = {k} ")`

`J = 4.5, J_epsilon = 4.503000500000001, dJ_da ~ k = 3.0005000000000000
6277`

OK, they match! You can now fill the values for $\frac{\partial d}{\partial a}$ and $\frac{\partial J}{\partial a}$ in the diagram if you have not already done so.

The steps in backprop

Now that you have worked through several nodes, we can write down the basic method:

working right to left, for each node:

- calculate the local derivative(s) of the node
- using the chain rule, combine with the derivative of the cost with respect to the node to the right.

The 'local derivative(s)' are the derivative(s) of the output of the current node with respect to all inputs or parameters.

Let's continue the job. We'll be a bit less verbose now that you are familiar with the method.

$$\frac{\partial J}{\partial c}, \frac{\partial J}{\partial b}$$

$$a = c + b$$

The next node has two derivatives of interest. We need to calculate $\frac{\partial J}{\partial c}$ so we can propagate to the left. We also want to calculate $\frac{\partial J}{\partial b}$. Finding the derivative of the cost with respect to the parameters w and b is the object of backprop. We will find the local derivatives, $\frac{\partial a}{\partial c}$ and $\frac{\partial a}{\partial b}$ first and then combine those with the derivative coming from the right, $\frac{\partial J}{\partial a}$.

```
In [25]: # calculate the local derivatives da_dc, da_db
sa = sc + sb
sa
```

Out[25]: $b + c$

```
In [26]: da_dc = diff(sa,sc)
da_db = diff(sa,sb)
print(da_dc, da_db)
```

1 1

```
In [27]: dJ_dc = da_dc * dJ_da
dJ_db = da_db * dJ_da
print(f"dJ_dc = {dJ_dc}, dJ_db = {dJ_db}")
```

$dJ_{dc} = d, dJ_{db} = d$

And in our example, $d = 3$

$$\frac{\partial J}{\partial w}$$

$$c = wx$$

The last node in this example calculates c . Here, we are interested in how J changes with respect to the parameter w . We will not back propagate to the input x , so we are not interested in $\frac{\partial J}{\partial x}$. Let's start by calculating $\frac{\partial c}{\partial w}$.

```
In [28]: # calculate the local derivative
sc = sw * sx
sc
```

Out[28]: wx

```
In [29]: dc_dw = diff(sc,sw)
dc_dw
```

Out[29]: x

This derivative is a bit more exciting than the last one. This will vary depending on the value of x . This is 2 in our example.

Combine this with $\frac{\partial J}{\partial c}$ to find $\frac{\partial J}{\partial w}$.

```
In [30]: dJ_dw = dc_dw * dJ_dc
dJ_dw
```

Out[30]: dx

```
In [31]: print(f"dJ_dw = {dJ_dw.subs([(sd,d),(sx,x)])}")
```

$dJ_{dw} = 2*d$

$d = 3$, so $\frac{\partial J}{\partial w} = 6$ for our example.

Let's test this arithmetically:

```
In [32]: J_epsilon = ((w+0.001)*x+b - y)**2/2
k = (J_epsilon - J)/0.001
print(f"J = {J}, J_epsilon = {J_epsilon}, dJ_dw ~ k = {k} ")
```

$J = 4.5$, $J_{epsilon} = 4.506002$, $dJ_{dw} \sim k = 6.00199999999619$

They match! Great. You can add $\frac{\partial J}{\partial w}$ to the diagram above and our analysis is complete.

Congratulations!

You've worked through an example of back propagation using a computation graph. You can apply this to larger examples by following the same node by node approach.