Formal Computational Skills

Week 4: Differentiation

Overview

Will talk about differentiation, how derivatives are calculated and its uses (especially for optimisation). Will end with how it is used to train neural networks

By the end you should:

- Know what dy/dx means
- Know how partial differentiation works
- Know how derivatives are calculated
- Be able to differentiate simple functions
- Intuitively understand the chain rule
- Understand the role of differentiation in optimisation
- Be able to use the gradient descent algorithm

Derivatives

Derivative of a function: y = f(x) is dy/dx (or df/dx or df(x)/dx

Read as: “dy by dx”

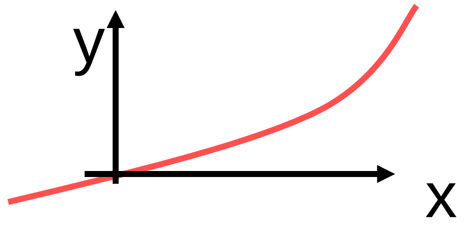
How a variable changes with respect to (w.r.t) other variables ie the rate of change y wrt x

It is a function of x and for any given value of x calculates the gradient at that point ie how much/fast y changes as x changes

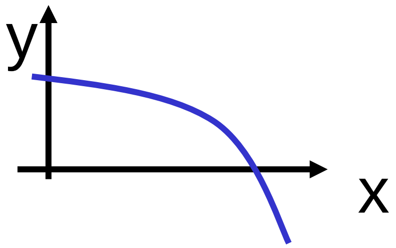
Examples: If x is space/distance, dy/dx tells us the slope or steepness, If x time, dy/dx tells us speeds and accelerations etc

Essential to any calculation in a dynamic world

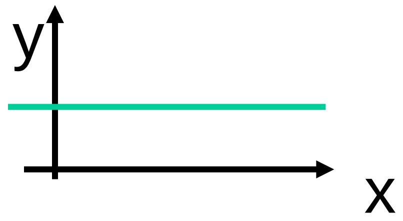
If dy/dx is >0 function is increasing



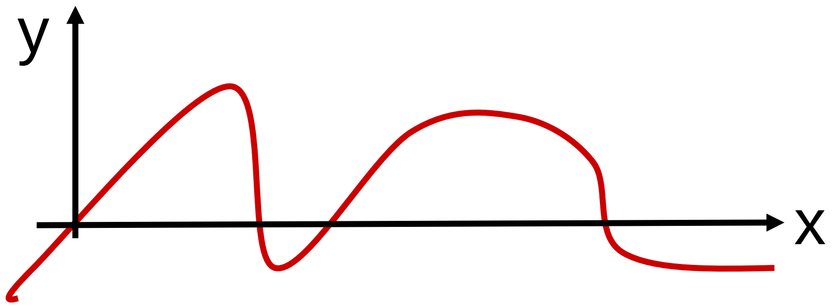
If dy/dx is <0 function is decreasing



If dy/dx = 0 function is constant



In general, dy/dx changes as x changes

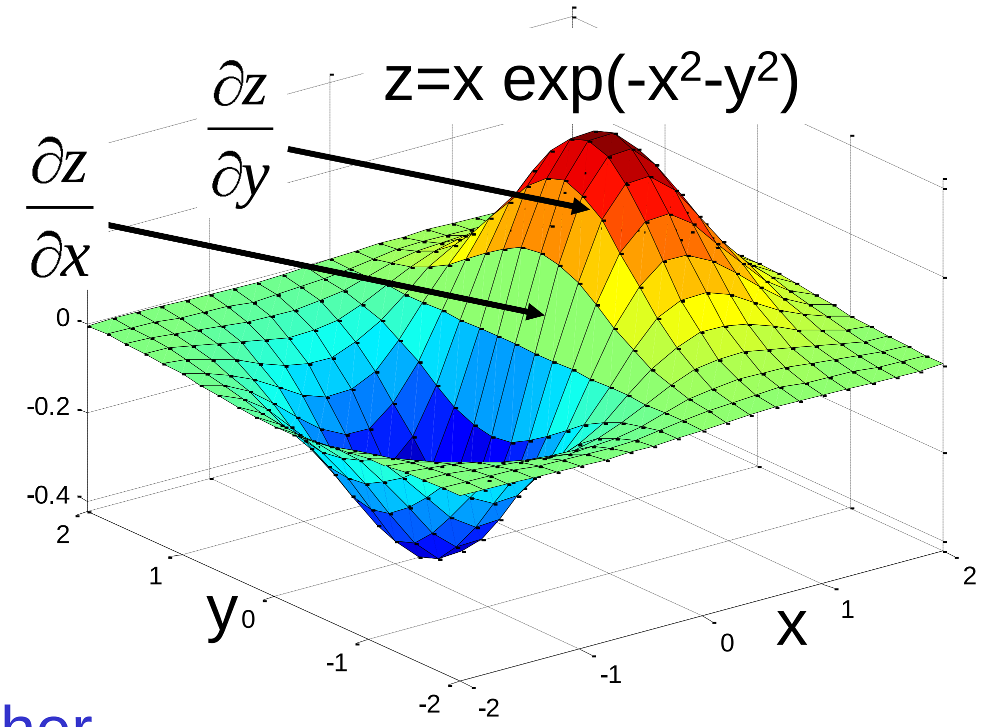


The bigger the absolute size of dy/dx (written as |dy/dx|) the faster change is happening and y is assumed to be more complex/difficult to work with

Partial Differentiation

For functions of 2 variables ($z=x\exp(-x^2-y^2)$), the partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ tell us how z varies wrt each parameter

Thus partial derivatives go along the lines



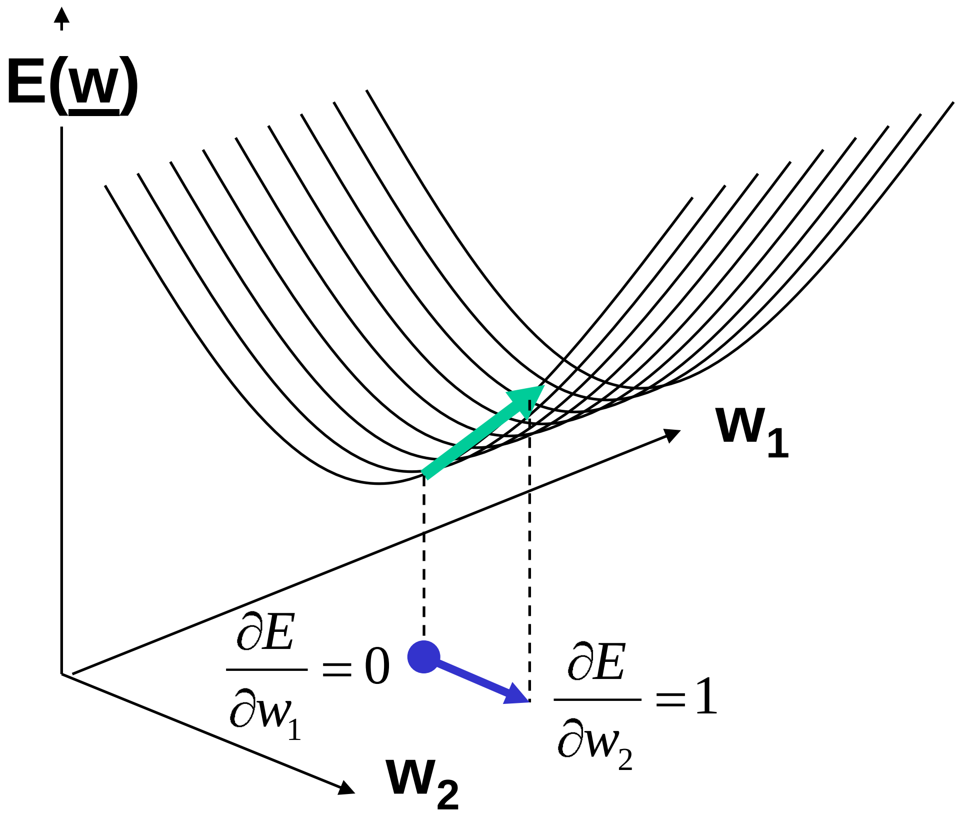
And for a function of n variables $y=f(x_1,x_2, \ldots,x_n)$,

$\frac{\partial y}{\partial x_i}$ tells us how y changes if only x_i varies

Calculated by treating the other variables as if they are constant

eg z=xy: to get $\frac{\partial z}{\partial x}$ treat y as if it is a constant $\frac{\partial z}{\partial x} = y$ and $\frac{\partial z}{\partial y} = x$

Higher Dimensions



In more dimensions, the gradient is a vector called grad (represented by an upside down triangle)

$$\nabla E = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \cdots, \frac{\partial E}{\partial w_n} \right)$$

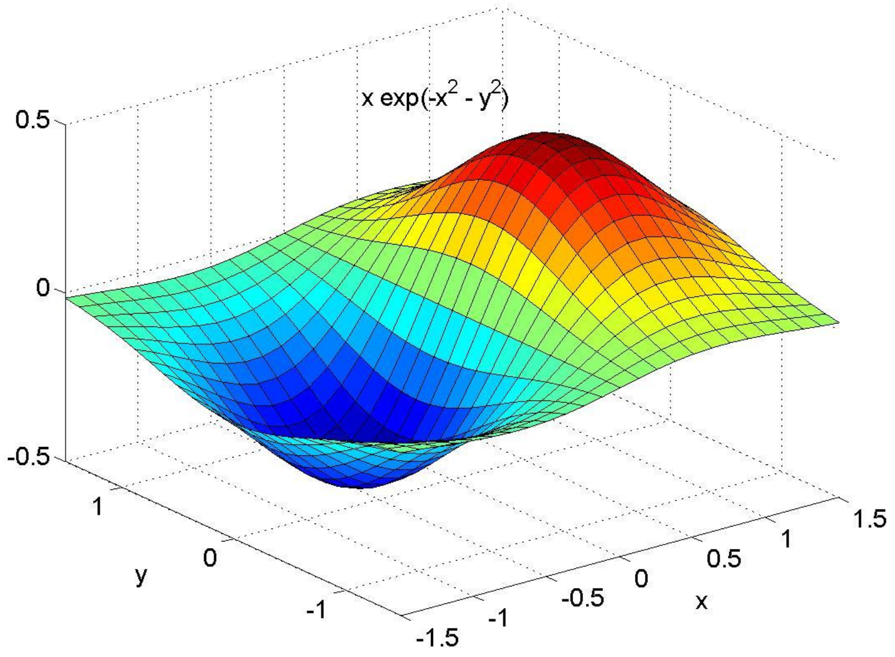
Grad tells us steepness <u>and</u> direction

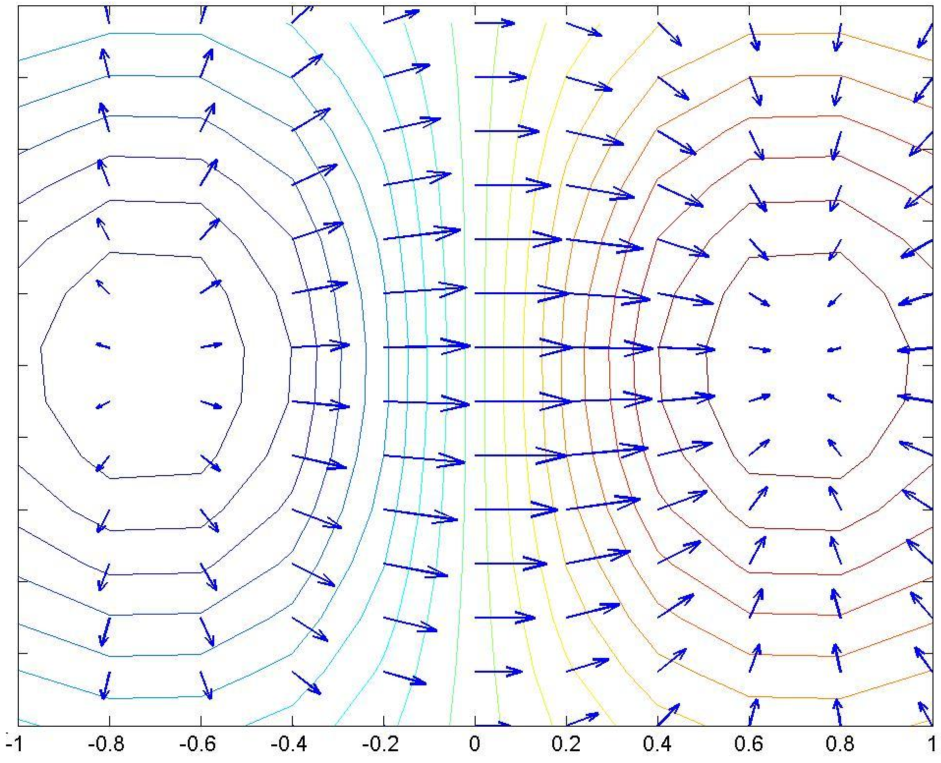
Elements are partial derivatives wrt each variable

Here $\frac{\partial E}{\partial w_1} = 0$ $\frac{\partial E}{\partial w_2} = 1$ so $\nabla E = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2} \right) = (0,1)$

Vector Fields

Since grad is a vector, often useful to view vectors as a field with vectors drawn as arrows





Note that grad points in the steepest direction up hill

Similarly minus grad points in the steepest direction downhill

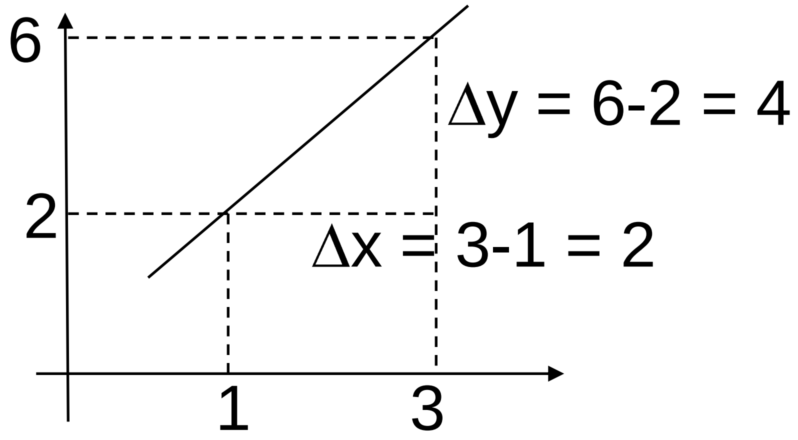
Calculating the Derivative

dy/dx = rate of change of y = change in y/change in x (if dy/dx remained constant)

For a straight line, y=mx + c it is easy

Since y remains constant, divide change in y by change in x

Eg y=2x



Aside: different sorts of d's

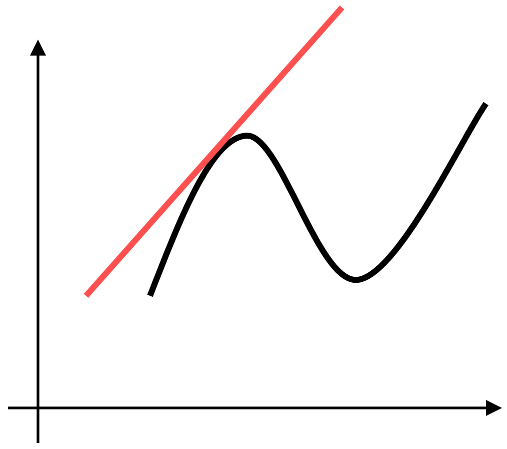
Δ Means "a change in"

δ Means "a small change in"

d Means "an infinitesimally small change in"

So: $dy/dx = \Delta y/\Delta x = 4/2 = 2$

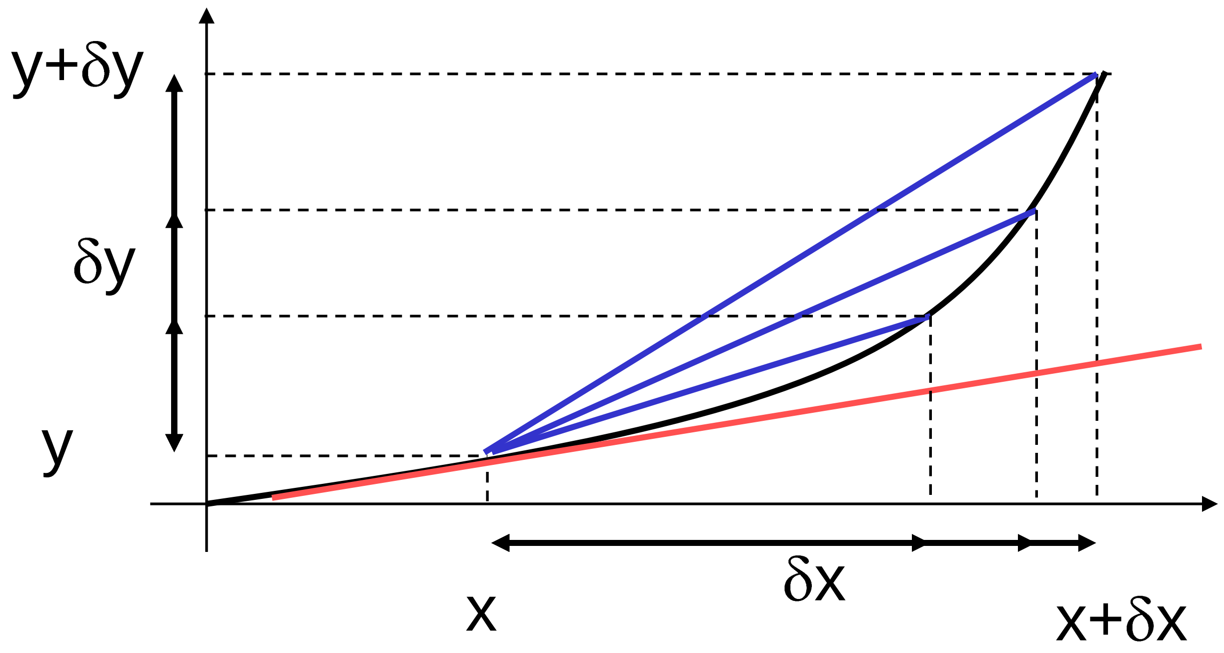
Thus dy/dx is constant and equals m.



However if derivative is not constant: we want gradient at a point ie gradient of a tangent to the curve at that point

A tangent is a line that touches the curve at that point

It is approximated by the gradient of a chord = δy/δx (a chord is a straight line from one point to another on a curve)



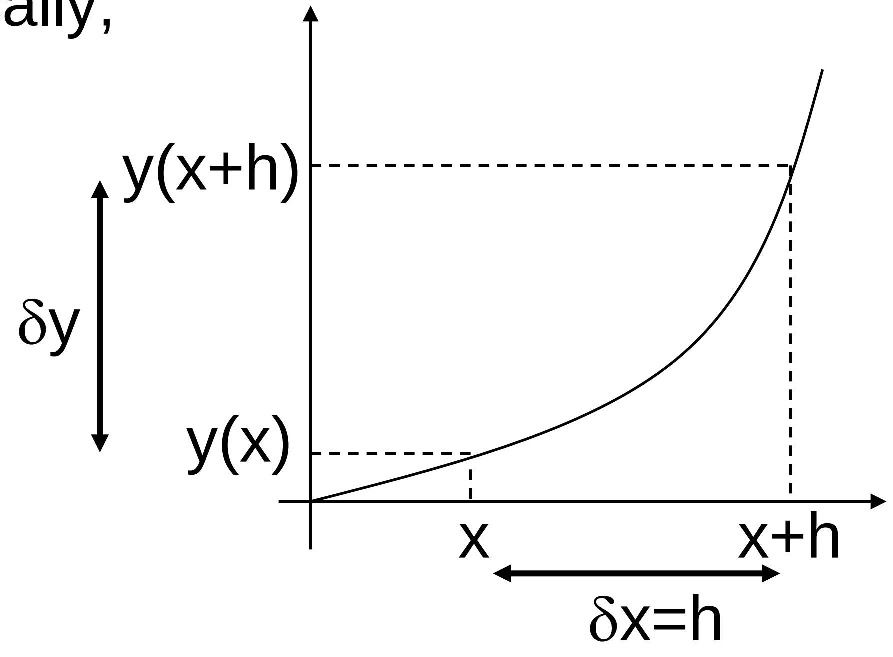
The smaller the chord, the better the approximation.

For infinitesimally small chord: δy/δx = dy/dx and the approximation is exact

Now do this process mathematically;

$$\frac{dy}{dx} \approx \frac{\delta y}{\delta x} = \frac{y(x+h)-y(x)}{x+h-x}$$

$$\boxed{\frac{dy}{dx} \approx \frac{y(x+h)-y(x)}{h}}$$



Eg y = x^2 $\quad y(x+h) = (x+h)^2 = x^2 + 2xh + h^2$

$$\frac{dy}{dx} \approx \frac{x^2+2xh+h^2-x^2}{h} = \frac{2xh+h^2}{h} = 2x+h$$

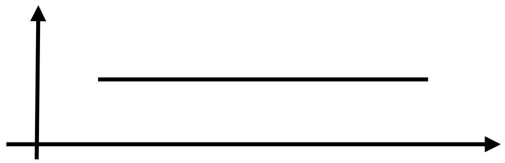
$$\frac{dy}{dx} = \lim_{h\to 0}(2x+h) = 2x$$

Can get derivatives in this way but have rules for most

Function	Derivative
$y = x^n,$	$\frac{dy}{dx} = nx^{n-1}$
$y = \sin x,$	$\frac{dy}{dx} = \cos(x)$
$y = \cos x,$	$\frac{dy}{dx} = -\sin(x)$
$y = \ln x,$	$\frac{dy}{dx} = \frac{1}{x}$
$y = e^x$	$\frac{dy}{dx} = e^x$

Other useful rules:

y=constant eg y=3, dy/dx = 0 as no change



y= k f(x) dy/dx = k df/dx

eg y = 3sin(x), dy/dx = 3cos(x)

y=f(x) + g(x), dy/dx = df/dx + dg/dx

y = x^3 + e^x, dy/dx = $3x^2$ + e^x

Can prove all eg why is dy/dx = e^x if y = e^x?

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \qquad \frac{dy}{dx} = \sum_{i=1}^{\infty} \frac{ix^{i-1}}{i!} = \sum_{i=1}^{\infty} \frac{x^{i-1}}{(i-1)!} = \sum_{i=0}^{\infty} \frac{x^i}{i!} = e^x$$

NB n! is n factorial and means eg 5! = 5x4x3x2x1

Product Rule

eg: $y = x\sin(x)$,

$$y = u(x)v(x)$$

$$\frac{dy}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$$

$u = x$, $v = \sin(x)$

$du/dx = 1x^0 = 1$, $dv/dx = \cos(x)$

$$\frac{dy}{dx} = u\frac{dv}{dx} + v\frac{du}{dx}$$

$dy/dx = x\cos(x) + 1\sin(x)$

Quotient Rule

$$y = \frac{u(x)}{v(x)} \qquad \frac{dy}{dx} = \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}$$

Chain Rule

Also known as
function of a function

$$y = u(v(x))$$

$$\frac{dy}{dx} = \frac{du(v)}{dv}\frac{dv}{dx}$$

eg: $y = \sin(x^2)$, $u = \sin(v)$, $v = x^2$

du/dv = cos(v), dv/dx = 2x

$$\frac{dy}{dx} = \frac{du(v)}{dv}\frac{dv}{dx}$$

dy/dx = cos(v) 2x = cos(x^2) 2x

EASY IN PRACTICE (honestly)

The one you will see most

Some examples:

$y_1 = 3x_1$ $\quad \frac{dy_1}{dx_1} = 3$ $\qquad y_1 = w_{11}x_1$ $\quad \frac{\partial y_1}{\partial x_1} = w_{11}$ $\quad \frac{\partial y_1}{\partial w_{11}} = x_1$

partial d's as y a function of 2 variables

$E = u^2 \quad dE/du = 2u \qquad E = u^2 + 2u + 20 \qquad dE/du = 2u + 2$

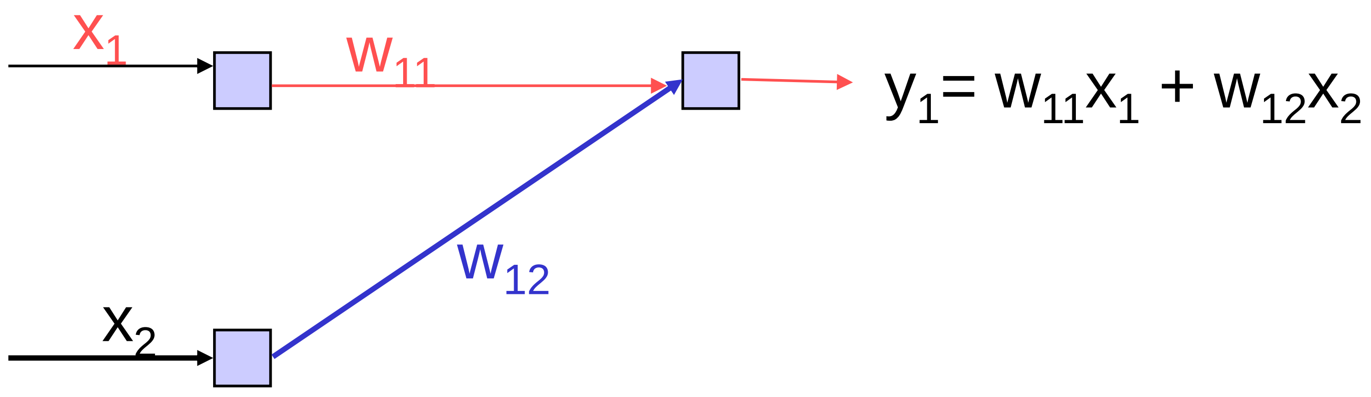
$y_1 = w_{11}x_1 + w_{12}x_2$ $\qquad \frac{\partial y_1}{\partial x_1} = \frac{\partial}{\partial x_1}(w_{11}x_1) + \frac{\partial}{\partial x_1}(w_{12}x_2)$

$\frac{\partial}{\partial x_1}(w_{11}x_1) = w_{11}$ $\qquad \frac{\partial}{\partial x_1}(w_{12}x_2) = 0$ $\qquad$ So $\frac{\partial y_1}{\partial x_1} = w_{11} + 0 = w_{11}$

Similarly, if: $y_1 = w_{11}x_1 + w_{12}x_2 + \ldots + w_{1n}x_n$ $\qquad \frac{\partial y_1}{\partial x_6} = w_{16}$

Sigmoid function: $y = \frac{1}{1 - e^{-a}} = (1 - e^{-a})^{-1}$ Can show $dy/da = y(1-y)$

EG: Networks

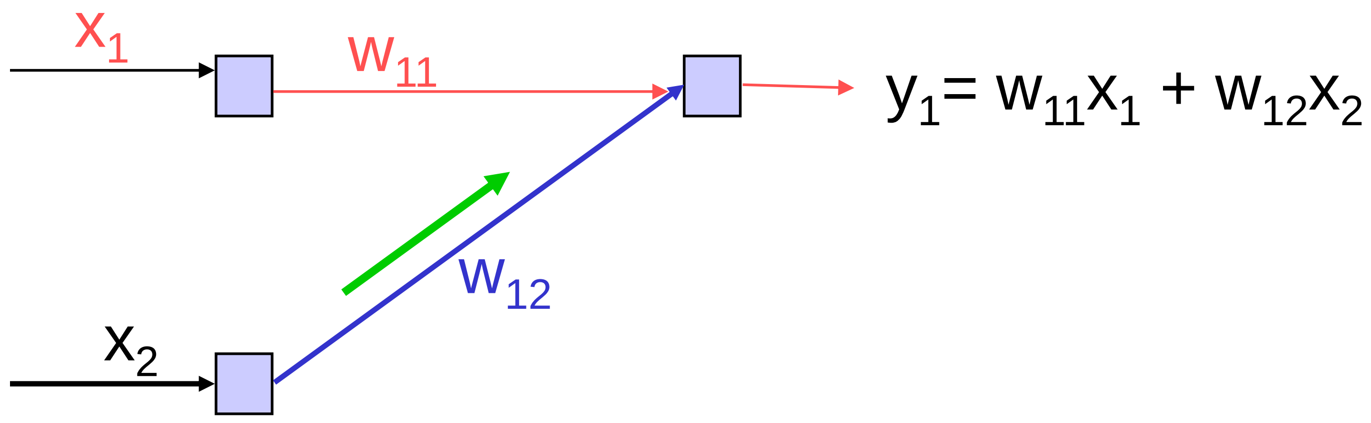


How does a change in x_1 affect the network output? $\frac{\partial y_1}{\partial x_1}$

$$\frac{\partial y_1}{\partial x_1} = \frac{\partial}{\partial x_1}\left(w_{11}x_1\right) + \frac{\partial}{\partial x_1}\left(w_{12}x_2\right)$$

So $\frac{\partial y_1}{\partial x_1} = w_{11}$

Intuitively, the bigger the weight on the connection, the more effect a change in the input along that connection will make.



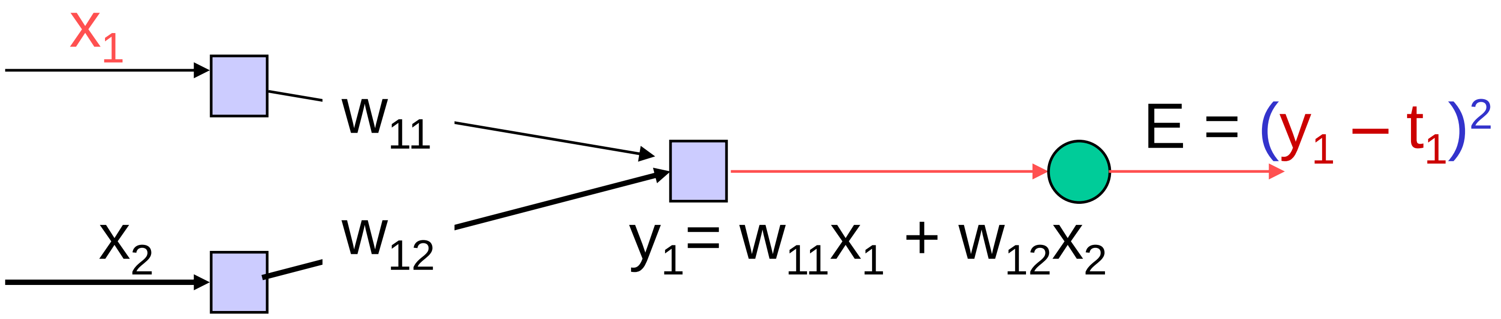
Similarly, w_{12} affects y_1 by:

$$\frac{\partial y_1}{\partial w_{12}} = \frac{\partial}{\partial w_{12}}\left(w_{11}x_1\right) + \frac{\partial}{\partial w_{12}}\left(w_{12}x_2\right) = x_2$$

So intuitively, changing w_{12} has a larger effect the stronger the signal travelling along it

Note also that by following the path of x_2 through the network, you can see which variables affect it

Now suppose we want to train the network so that it outputs a target t_1 when the input is $\underline{x}$. Need an error function E to evaluate how far the output y_1 is from the target t_1



Now how does E change if the output changes? $\dfrac{dE}{dy_1}$

To calculate this we need the chain rule as E is a function of a function of y_1

$$\boxed{\begin{array}{l} E = u(v(y_1)) \\ \dfrac{dE}{dy_1} = \dfrac{du}{dv}\dfrac{dv}{dy_1} \end{array}}$$

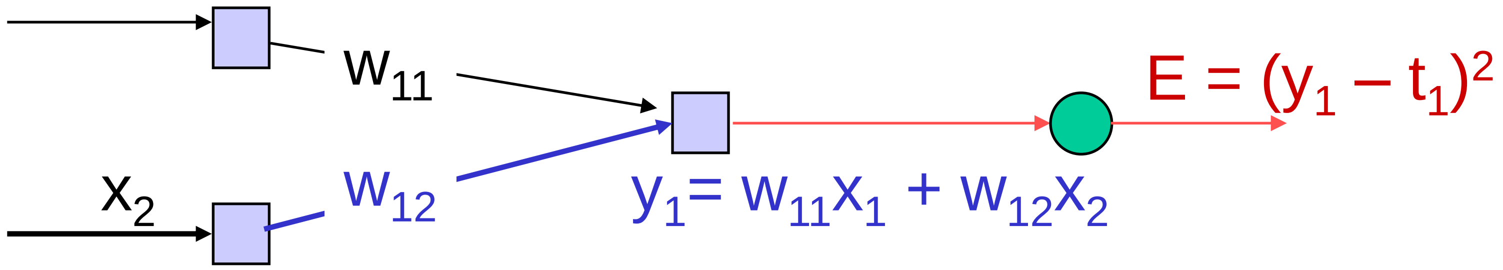
where $v = (y_1 - t_1)$
and $u = v^2$

Now $dv/dy_1 = 1$

and $du/dv = 2v = 2(y_1 - t_1)$

so $\dfrac{dE}{dy_1} = 1 \times 2(y_1 - t_1) = 2(y_1 - t_1)$

We train the network by adjusting the weights, so we need to know how the error varies if eg w_{12} varies ie $\frac{\partial E}{\partial w_{12}}$



To calculate this note that w_{12} affects y_1 by: $\frac{\partial y_1}{\partial w_{12}} = x_2$

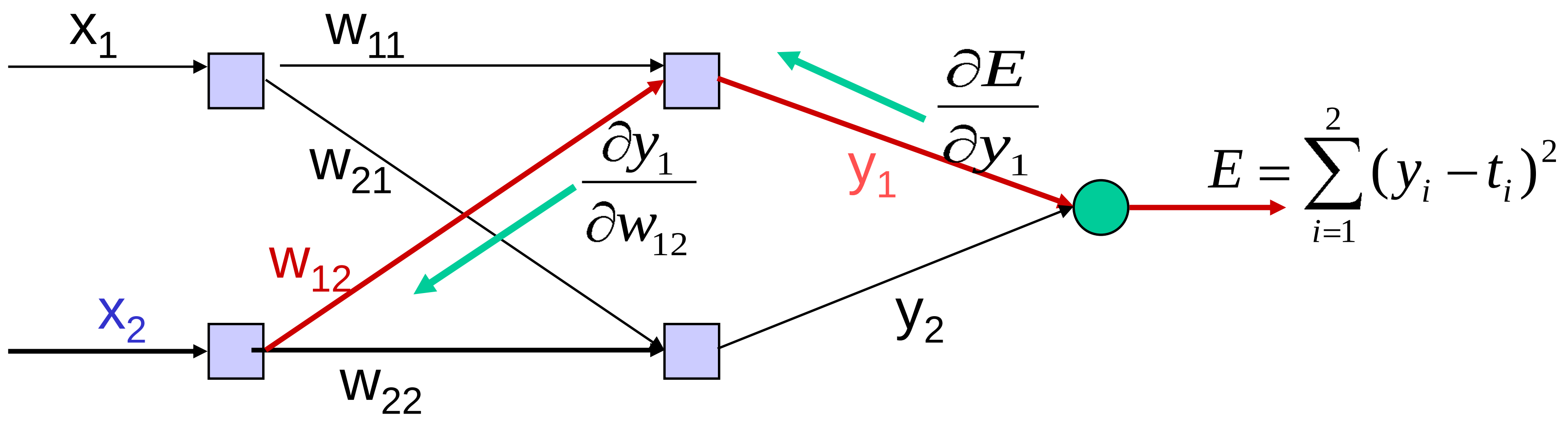
which in turn affects E by: $\frac{\partial E}{\partial y_1} = 2(y_1 - t_1)$

The chain of events: w_{12} affecting y_1 affecting E indicates the chain rule so

$$\frac{\partial E}{\partial w_{12}} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial w_{12}} = x_2\, 2(y_1 - t_1)$$

And in general: $\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial w_{ij}} = 2x_j(y_i - t_i)$ cf backprop Δ_{ij}

Back to 2 (or even N outputs). Consider the route (or more accurately, the variables) by which a change in w_{12} affects E:


x_1
w_{11}
w_{21}
$\frac{\partial y_1}{\partial w_{12}}$
$\frac{\partial E}{\partial y_1}$
y_1
w_{12}
x_2
w_{22}
y_2
$E = \sum_{i=1}^{2} (y_i - t_i)^2$

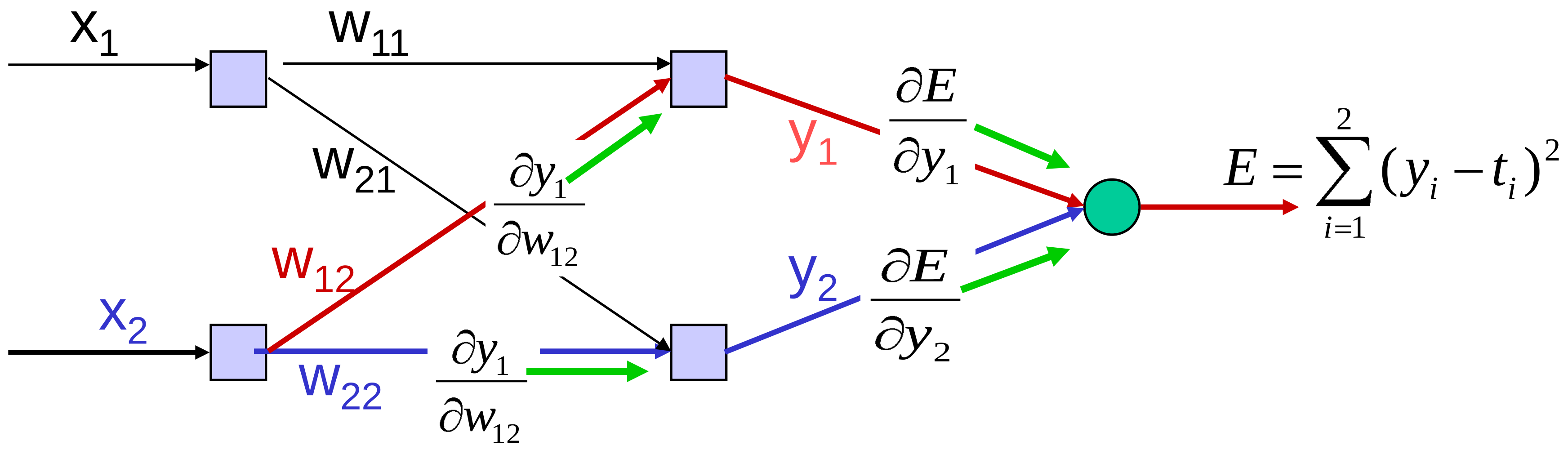
See that w_{12} still affects E via x_2 and y_1 so still have:

$$\frac{\partial E}{\partial w_{12}} = \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial w_{12}} = x_2\, 2(y_1 - t_1)$$

And: $$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial w_{ij}} = 2x_j (y_i - t_i)$$

Visualise the partial derivatives working along the connections

How about how E is changed by x_2: $\frac{\partial E}{\partial x_2}$



x_1
w_{11}
w_{21}
w_{12}
x_2
w_{22}
$\frac{\partial y_1}{\partial w_{12}}$
$\frac{\partial y_1}{\partial w_{12}}$
y_1
y_2
$\frac{\partial E}{\partial y_1}$
$\frac{\partial E}{\partial y_2}$
$E = \sum_{i=1}^{2} (y_i - t_i)^2$

Again look at the route along which x_2 travels:

x_2 changes y_1 along w_{12} which changes E

x_2 also changes y_2 along w_{22} which changes E

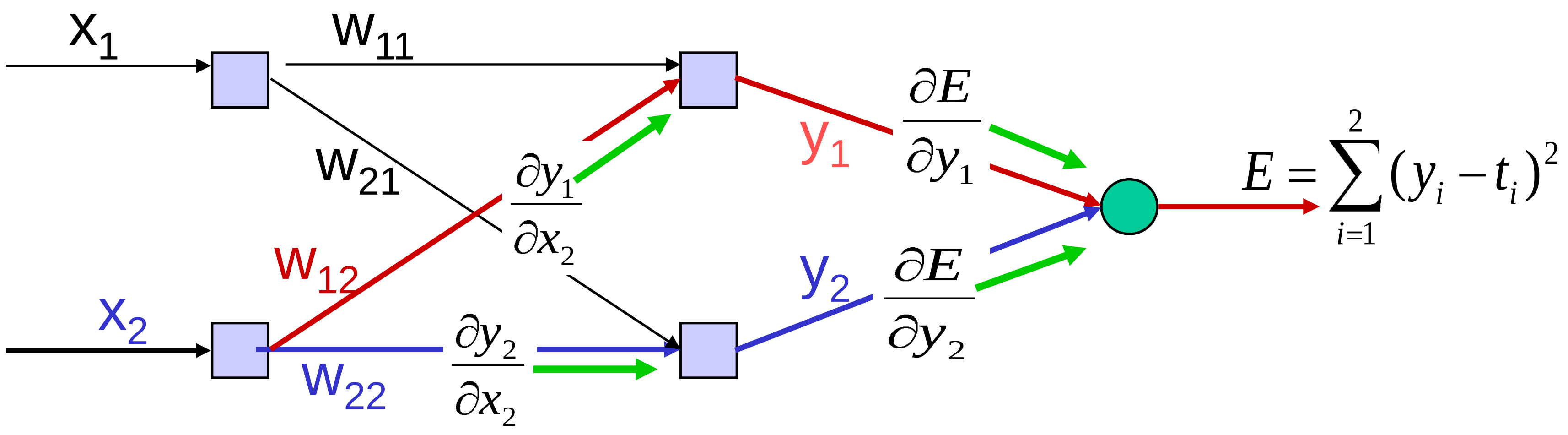
Therefore need to somehow combine the contributions along these 2 routes …

Chain rule for partial differentiation

If y is a function of u and v which are both functions of x then:

$$\boxed{\frac{dy}{dx} = \frac{\partial y}{\partial u}\frac{du}{dx} + \frac{\partial y}{\partial v}\frac{dv}{dx}}$$

Quite intuitive: sum the contributions along each path affecting E

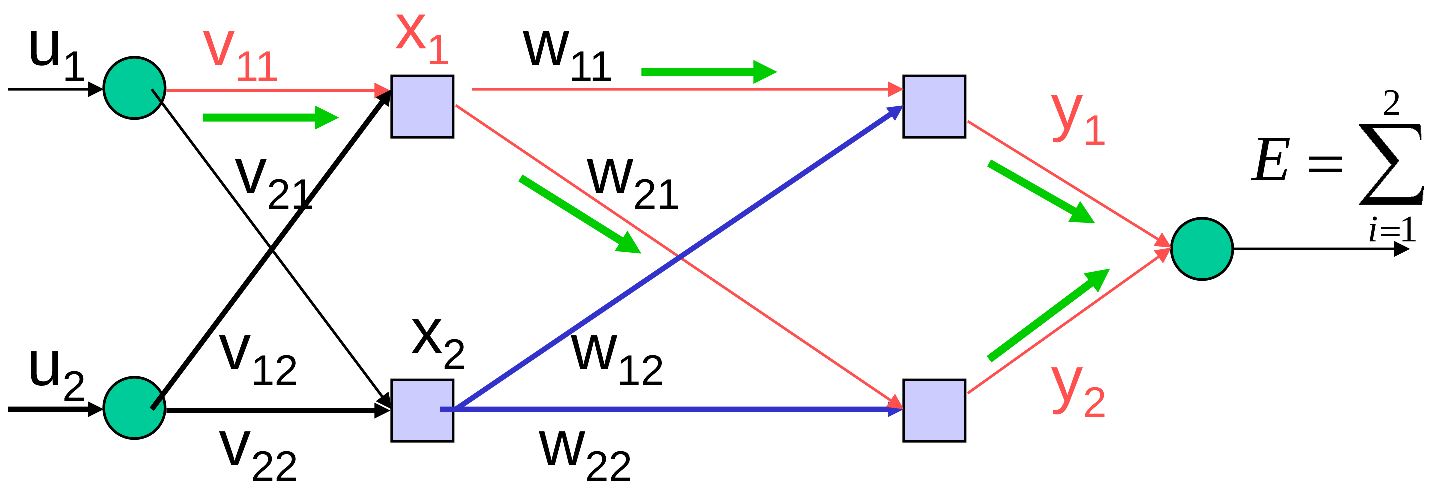


x_1
w_{11}
w_{21}
w_{12}
x_2
w_{22}
$\frac{\partial y_1}{\partial x_2}$
$\frac{\partial y_2}{\partial x_2}$
y_1
y_2
$\frac{\partial E}{\partial y_1}$
$\frac{\partial E}{\partial y_2}$
$E = \sum_{i=1}^{2}(y_i - t_i)^2$

So: $$\frac{\partial E}{\partial x_2} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial E}{\partial y_2}\frac{\partial y_2}{\partial x_2} = 2(y_1 - t_1)\, w_{12} + 2(y_2 - t_2) w_{22}$$

$$= 2 \sum_{\text{Outputs}} (y_i - t_i) w_{i2}$$

What about another layer?



$$E = \sum_{i=1}^{2} (t_i - y_i)^2$$

We need to know how E changes wrt the v's and w's. The calculation for the w's is the same but what about the v's?

v's affect x's which affect y's which affect E.

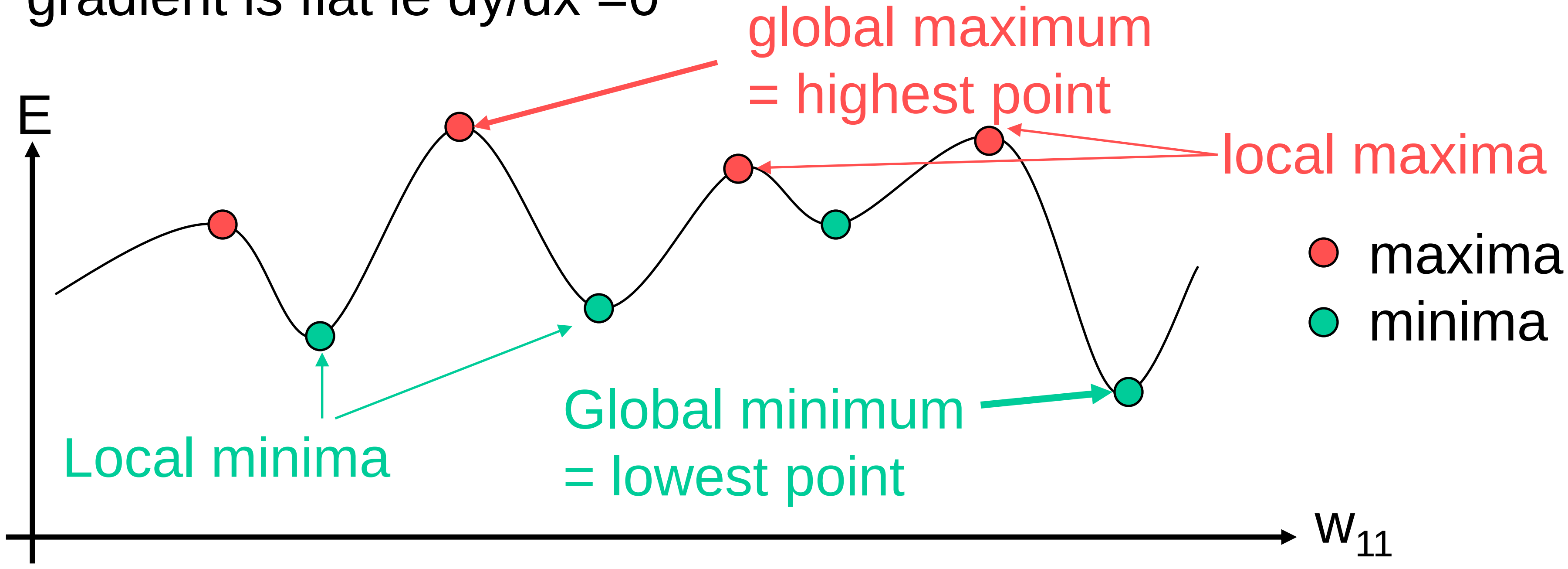
So we need: $\dfrac{\partial E}{\partial v_{ji}}$ which involves finding $\dfrac{\partial E}{\partial x_j}$ and $\dfrac{\partial x_j}{\partial v_{ji}}$

Thus: $$\frac{\partial E}{\partial v_{ji}} = \frac{\partial E}{\partial x_j}\frac{\partial x_j}{\partial v_{ji}} = \left(\sum_{k=1}^{2} \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial x_j} \right) u_i$$

Seems complicated but can do via matrix multiplication

Finding Maxima/Minima (Optima)

At a maximum or minimum (optimum, stationary point etc) the gradient is flat ie dy/dx =0



To find max/min calculate dy/dx and solve dy/dx =0

Eg: $y = x^2$, dy/dx = 2x so at optima: dy/dx = 0 => 2x =0 ie x=0

Similarly for a function of n variables $y=f(x_1,x_2, ...,x_n)$, at optima $\frac{\partial y}{\partial x_i} = 0$ for i = 1, …, n

Why do we want to find optima? Error minimisation

In neural network training have the error function $E = \Sigma(\underline{y}_i - \underline{t}_i)^2$
At the global minimum of E, outputs are close as possible to targets, so network is 'trained'

Therefore to train a network 'simply' set dE/dw=0 and solve

However, In many cases (especially neural network training) cannot solve dy/dx=0.

In such cases can use gradient descent to change the weights in such a way that the error decreases

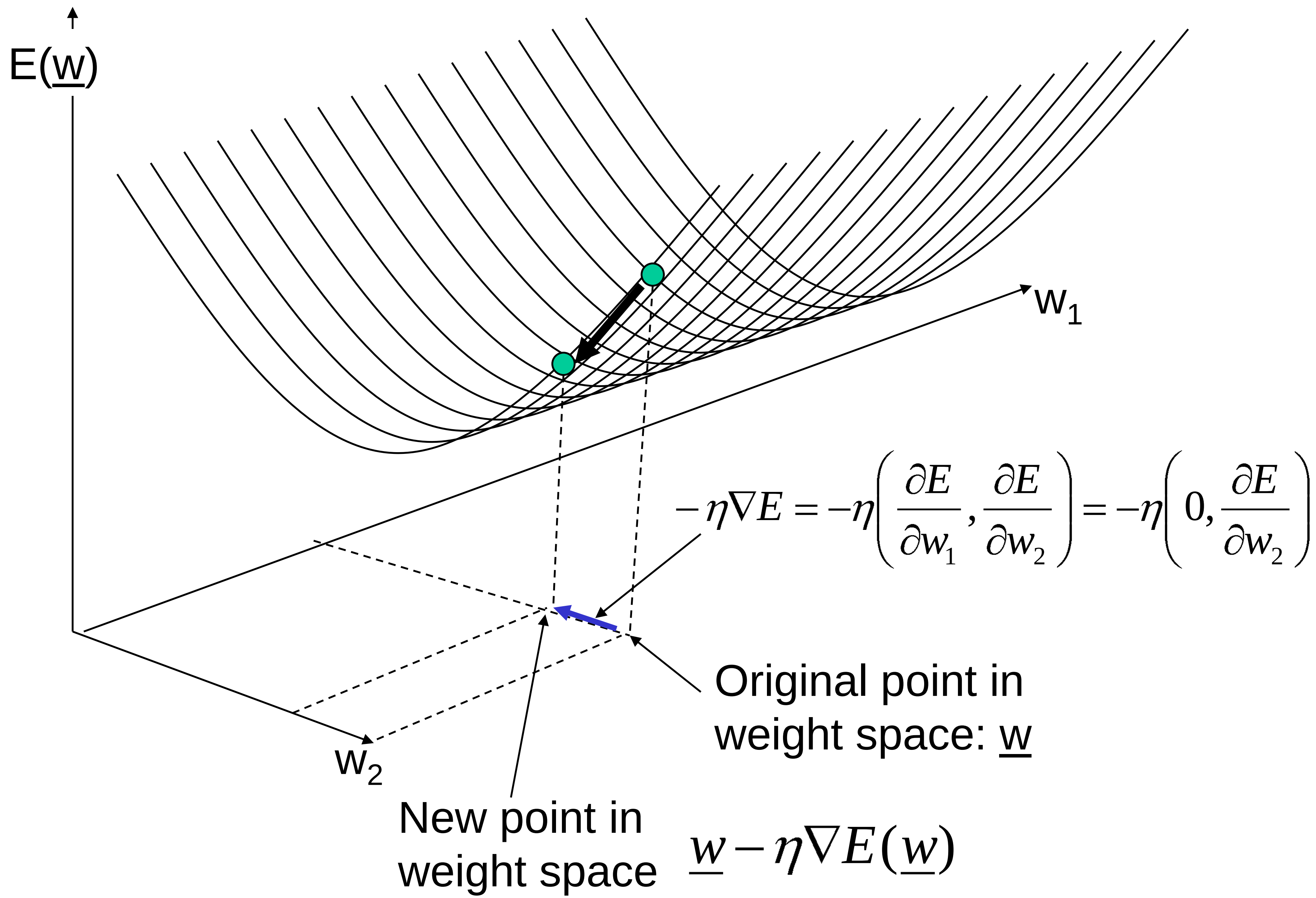
Gradient Descent

Analogy is being on a foggy hillside where we can only see the area around us.
To get to the valley floor a good plan is to take a step downhill.
To get there quickly, take a step in the steepest direction downhill ie in the direction of –grad.
As we can't see very far and only know the gradient locally only take a small step

Algorithm

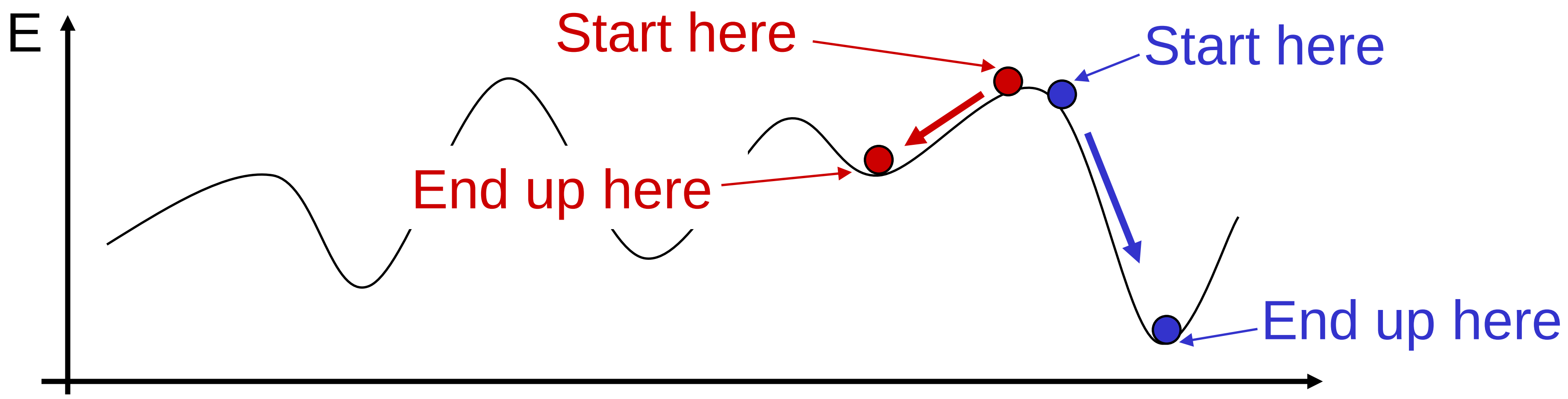
- Start at a random position in space, $\underline{w}$
- Calculate $\nabla E(\underline{w})$
- Update $\underline{w}$ via: $\underline{w}_{new} = \underline{w}_{old} - \eta \nabla E(\underline{w})$
- Repeat from 2

η is the learning rate and governs the size of the step taken. It is set to a low value and often changed adaptively eg if landscape seems 'easy' (not jagged or steep) increase step-size etc



E(w)
w1
w2
$-\eta\nabla E = -\eta\left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}\right) = -\eta\left(0, \frac{\partial E}{\partial w_2}\right)$
Original point in weight space: w
New point in weight space
$\underline{w} - \eta\nabla E(\underline{w})$

Problems of local minima. Gradient descent goes to the closest local minimum and stays there:



Other problems: length of time to get a solution especially if there are flat regions in the search space; Can be difficult to calculate grad

Many variations which help with these problems eg:

- Introduce more stochasticity to ‘escape’ local minima
- use higher order gradients to improve convergence
- Approximate grad or use a proxy eg take a step, see if lower down or not and if so, stay there eg hill-climbing etc

Higher order derivatives

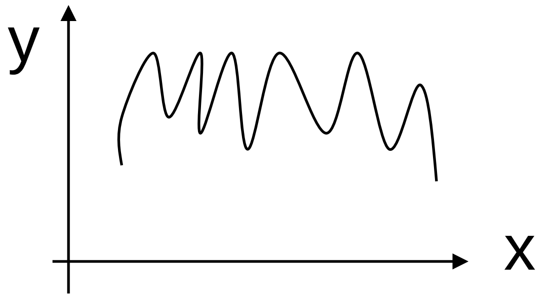
Rates of rates of change. In 1 dimension $\frac{d^3y}{dx^3}$

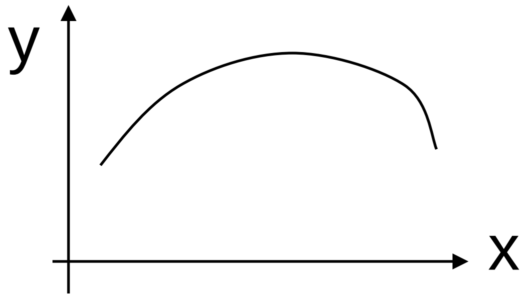
In many dimensions a vector $\nabla^n y$

Eg velocity dy/dx is 1st order, acceleration: d^2y/dx^2 is 2nd order

The higher the order, the more information known about y and so used eg improve gradient descent search

Also used to tell whether a function (eg mapping produced by a network) is smooth as it calculates the curvature at a point





$\sum_x \left|\nabla^2 y\right|$ is greater for the first mapping than the 2nd