Unit 2:

# Supervised Learning

How do we adjust a neural network so that it does what we want it to do?
How do we get the network to *learn*?

By the end of this unit, you will be able to...

- Formulate the problem of supervised learning as an optimization problem.
- List and explain some of the most common loss/cost functions.
- Describe a perceptron, and its limitations.
- Use gradient descent to optimize connection weights.
- Derive and implement the error backpropagation algorithm.
- Use labelled data wisely to train models that are generalizable.
- Explain the problem of vanishing/exploding gradients.
- Employ some methods to improve the convergence of our learning method.

**Neural Learning**

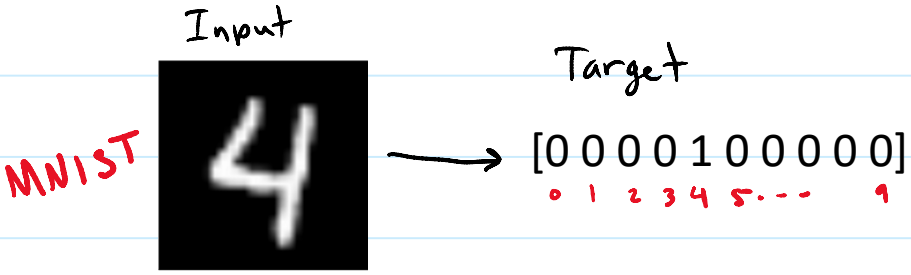
Goal: To formulate the problem of supervised learning as an optimization problem.

Getting a neural network to do what you want usually means finding a set of connection weights that yield the desired behaviour. That is, neural learning is all about adjusting connection weights.

There are three basic categories of learning problems:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

In **supervised learning**, the desired output is known so we can compute the error and use that error to adjust our network.
Example: Given an image of a digit, identify which digit it is.




In **unsupervised learning**, the output is not known (or not supplied), so cannot be used to generate an error signal. Instead, this form of learning is all about finding efficient representations for the statistical structure in the input.
Example: Given spoken English words, transform them into a more efficient representation such as phonemes, and then syllables.

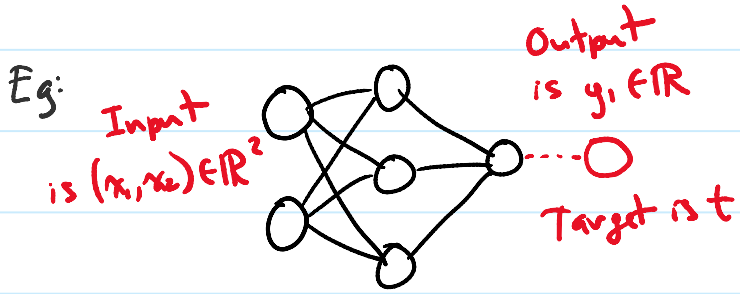
In **reinforcement learning**, feedback is given, but usually less often, and the error signal is usually less specific.
Example: When playing a game of chess, a person knows their play

was good if they win the game. They can try to learn from the moves they made.

In this course, we will mostly focus on supervised learning. But we will look at some examples of unsupervised learning, and possibly some reinforcement learning.

**Supervised Learning**

Our neural network performs some mapping from an input space to an output space.




We are given training data, with many MANY examples of input/target pairs. This data is (presumably) the result of some consistent mapping process. For example, handwritten digits map to numbers. Or,

| A | B | XOR(A,B) |
|---|---|---|
| 1 | 1 | 0 |
| 1 | 0 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 0 |

Input $(A,B) \in \{0,1\}^2$

Output/Target

$y \in \mathbb{R}$ or $y \in [0,1]$ $t \in \{0,1\}$

We are given inputs and their corresponding targets, ~~one pair (or a few pairs) at a time~~. Our task is to alter the connection weights in our network so that our network mimics this mapping.

Our goal is to bring the output as close as possible to the target. But what, exactly, do we mean by "close"? For now, we will use the scalar function $E(y,t)$ as an error function, which returns a smaller value as our outputs are closer to the target.

Two common types of mappings encountered in supervised learning are **regression** and **classification**.

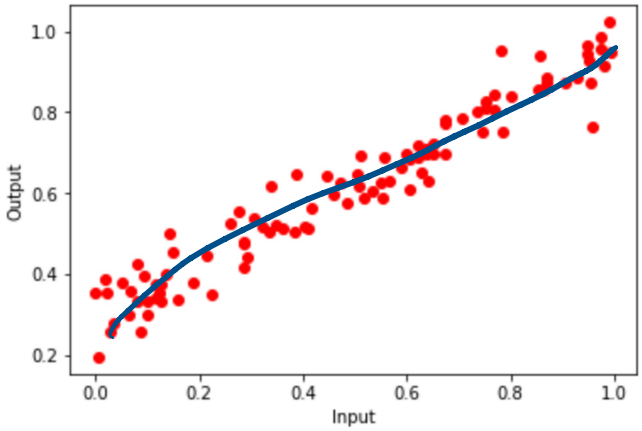
## Regression

Output values are a continuous-valued function of the inputs. The outputs can take on a range of values.

Example:

**Linear Regression**

Outputs fall range of values




## Classification

Outputs fall into a number of distinct categories.

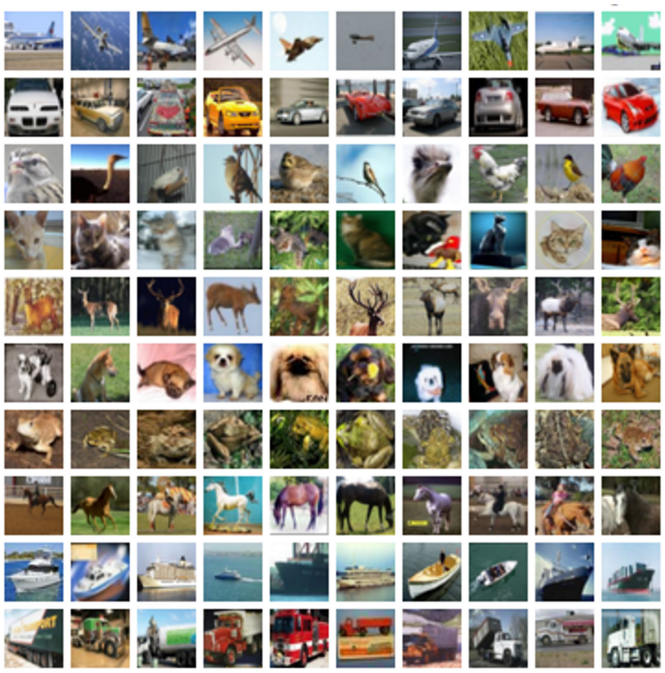
Example:

**MNIST**

| Inputs | Outputs | Inputs | Outputs |
|---|---|---|---|
|  | [0 0 0 0 0 0 0 1 0 0] |  | [0 0 0 0 0 1 0 0 0 0] |
|  | [1 0 0 0 0 0 0 0 0 0] |  | [0 0 0 0 1 0 0 0 0 0] |
|  | [0 0 0 0 0 0 1 0 0 0] |  | [0 0 0 0 0 0 0 0 0 1] |

**CIFAR-10**

Inputs



Outputs

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

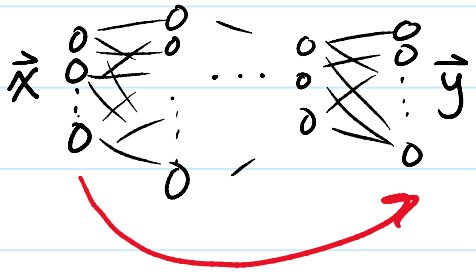
## Optimization

Once we have a cost function, our neural-network learning problem can be formulated as an optimization problem.

Let our network be represented by the mapping $f$, so that

$$\vec{y} = f(\vec{x}; \theta)$$

where $\theta$ represents all the weights and biases.




$$\min_{\theta} \mathbb{E}\left[ E(f(\vec{x};\theta), \vec{t}(\vec{x})) \right]$$

$E(f(\vec{x};\theta))$: output of model; $\vec{t}(\vec{x})$: target; $\vec{x} \in$ data

In other words, find the weights and biases that minimize the expected cost between the outputs and the targets.

## Universal Approximation Theorem

Question: Can we approximate **any** function using a neural network?

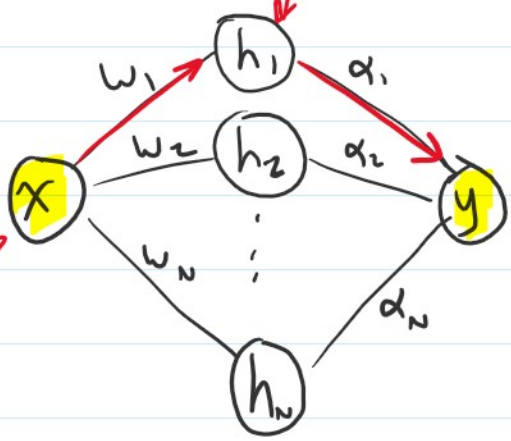
Given a function $f(x)$, can we find the weights $\omega_j, \alpha_j$, and biases $\theta_j, j=1,\dots,N$ such that

$$f(x) \approx \sum_{j=1}^{N} \alpha_j \sigma(\omega_j x + \theta_j)$$

$\alpha_1 h_1$

to arbitrary precision?

$h_1 = \sigma(\omega_1 x + \theta_1)$




Universal Approximation Theorem:

> **Theorem 2.** *Let $\sigma$ be any continuous sigmoidal function. Then finite sums of the form*
>
> continuous functions
>
> $$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(\omega_j x + \theta_j)$$
>
> *are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which*
>
> $$|G(x) - f(x)| < \varepsilon \quad \text{for all} \quad x \in I_n.$$
>
> $I_n = [0,1]^n$

Cybenko G, "Approximation by Superpositions of a Sigmoidal Function", *Math. Control Signals Systems*, 2:303-314, 1989.

A function $\sigma$ is "sigmoidal" if $\sigma(x) = \begin{cases} 1 & \text{as } x \to \infty \\ 0 & \text{as } x \to -\infty \end{cases}$

eg. logistic


The theorem states that

$\exists N$ and $\exists \omega_j, \theta_j, \alpha_j$ for $j=1,\dots,N$ such that

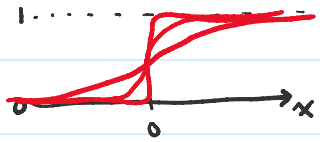
$$|G(x) - f(x)| < \varepsilon \quad \forall x \in I_n$$

**Proof:**

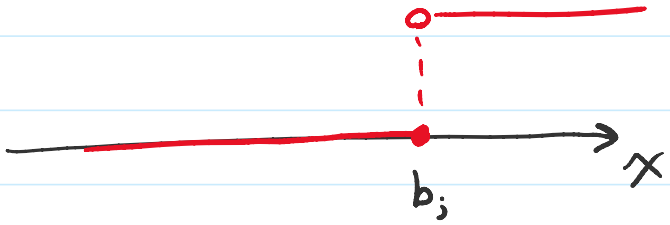
Since we let $\omega_j \to \infty$ for $j = 1, \dots, N$

S'pose we let $w_j \to \infty$ for $j = 1, \ldots, N$

Then $\sigma(w_j x) \xrightarrow{w_j \to \infty} \begin{cases} 0 & \text{for } x \le 0 \\ 1 & \text{for } x > 0 \end{cases}$




Or $\sigma(w_j(x - b_j)) \xrightarrow{w_j \to \infty} \begin{cases} 0 & \text{for } x \le b_j \\ 1 & \text{for } x > b_j \end{cases}$



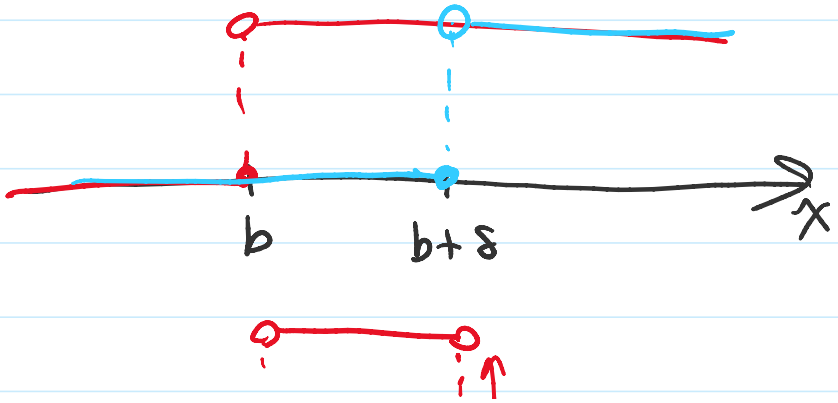

This is the same as the Heaviside step function,

$$H(x) = \lim_{w_j \to \infty} \sigma(w_j x)$$

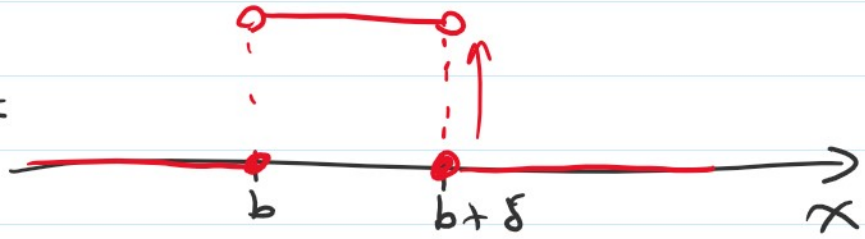
Define: $H(x; b) = \lim_{w \to \infty} \sigma(w(x - b))$

We can use two such functions to create a piece,

$$P(x; b, \delta) = H(x; b) - H(x; b + \delta)$$



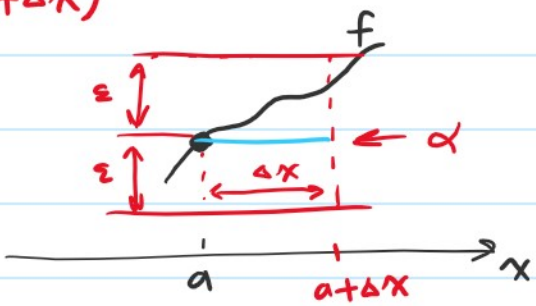
$P(x; b, \delta) =$




Since $f(x)$ is continuous, $\lim_{x \to a} f(x) = f(a)$. $\forall a \in I_n$

$\therefore \exists$ an interval, $(a, a+\Delta x)$ such that

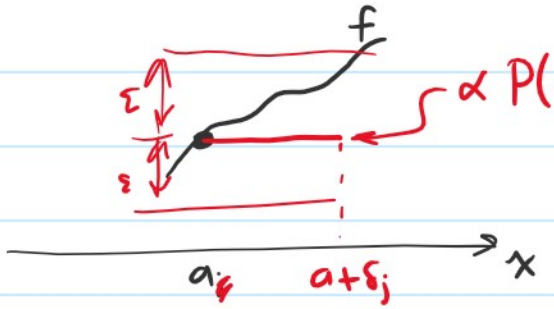
$|f(x) - f(a)| < \varepsilon \quad \forall x \in (a, a+\Delta x)$




Choose $b_j = a$ , $\delta_j = \Delta x$ , and $\alpha_j = f(a)$

$\therefore |f(x) - f(a)| < \varepsilon$ for $a \le x \le a + \delta_j$

$|f(x) - \alpha_j P(x; a, \delta_j)| < \varepsilon$ "




Max error in $[a, a+\delta_j]$ is less than $\varepsilon$.

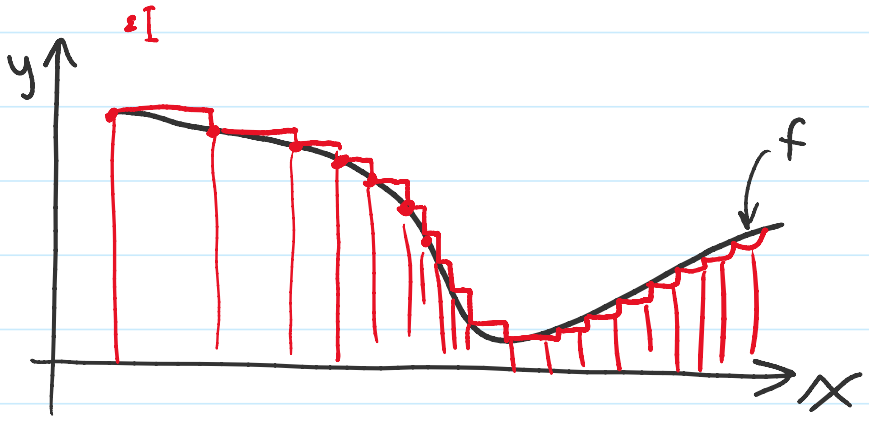
Repeat this process for $x = a = b_j + \delta_j$

Construct

$$G(x) = \sum_{j=1}^{N} \alpha_j P(x; b_j, \delta_j)$$

$\varepsilon I$

$u \uparrow$




So, why would we ever need a neural network with more than one hidden layer?

Answer: The theorem guarantees existance, but makes no claims about N, the # of hidden nodes.

N might grow exponentially as $\varepsilon$ gets smaller.

# Loss Functions

Goal: To become familiar with some of the most common ways to measure error.

We have to choose a way to quantify how close our output is to the target. For this, we use a "cost function", also known as an "objective function". There are many choices, but here are two commonly-used ones.

For input $\vec{x}$, our target is $\vec{t}(\vec{x})$, and the output of our network is $\vec{y}(\vec{x})$.

**Mean Squared Error (MSE)**

$$E(\vec{y}, \vec{t}) = \frac{1}{N} \| \vec{y} - \vec{t} \|_2^2 \qquad N \text{ is the \# of samples in the dataset.}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \| \vec{y}_i - \vec{t}_i \|_2^2$$

The use of MSE as a cost function is often associated with linear activation functions, or ReLU. This is because these activation functions afford a larger output range.

**Cross Entropy (Bernoulli Cross Entropy)**

Consider a function (or network) with a single output that is either 0 or 1. The task of mapping inputs to the correct output (0 or 1) is a classification problem.

$$x \rightarrow \boxed{f(x, \theta)} \rightarrow y \in [0, 1]$$

where the true class is expressed in the target, $t \in \{0, 1\}$.

If we suppose that $y$ is the probability that $x \rightarrow 1$,

$$\therefore P(x \rightarrow 1 \mid \theta) = f(x, \theta)$$

If we suppose that $y$ is the probability that $x \to 1$,

$$y = P(x \to 1 \mid \theta) = f(x, \theta)$$

then we can treat it as a Bernoulli distribution.

i.e. $P(x \to 1 \mid \theta) = y$ ie. $t = 1$

Thus $P(x \to 0 \mid \theta) = 1 - y$ i.e. $t = 0$

The likelihood of our data sample given our model is

$$P(x \to t \mid \theta) = y^t (1-y)^{1-t}$$

$$\log y^t (1-y)^{1-t} = \log y^t + \log (1-y)^{1-t}$$

The task of "learning" would be finding a model ( $\theta$ ) that maximizes this likelihood.

Or, we could equivalently minimize the negative log-likelihood

$$E(y,t) = -\left(t \log y + (1-t)\log(1-y)\right)$$

This log-likelihood formula is the basis of the cross-entropy cost function. Given the dataset,

$$\{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$$

the expected cross entropy over the entire dataset is

$$-\mathbb{E}\left[t \log y + (1-t)\log(1-y)\right]_{\text{data}}$$

$$= -\frac{1}{N} \sum_{i=1}^{N} t_i \log y_i + (1-t_i)\log(1-y_i)$$

Cross entropy assumes that the output values are in the range [0, 1]. Hence, it works nicely with the logistic activation function.

## Categorical Cross-Entropy (Multinoulli Cross-Entropy)

Consider a classification problem that has K classes (K>2). Given an input,

the task of our model is to output the class of the input.
eg. given an image of a digit, determine the digit class

Suppose the probability of getting class $k$ is $p_k$.

i.e. $p(k) = p_k \quad k = 1, \ldots, K, \quad \sum_k p_k = 1$

Suppose we observe a sample of class $k^*$
Thus, the target class vector is

$$\vec{t} = [0_1 \; 0_2 \ldots 1_{k^*} \ldots 0_K]$$

The likelihood of this sample is

$$P(\vec{t} \mid \vec{p}) = p_{k^*}$$

which can also be written

$$P(\vec{t} \mid \vec{p}) = \prod_{k=1}^{K} p_k^{t_k}$$

$$= p_1^0 \cdot p_2^0 \cdots p_{k^*}^1 \cdots p_K^0$$

Taking the logarithm of that gives

$$\log P(\vec{t} \mid \vec{p}) = \sum_{k=1}^{K} t_k \log p_k$$

Suppose our model is given the input $\vec{x}$
The the network's output is

$$\vec{y} = f(\vec{x}, \theta) = P(\vec{t} \mid \theta, \vec{x})$$

We interpret $\vec{y}_k$ as the estimated probability of $\vec{x}$ being from class $k$
The true class is given by the target class vector $\vec{t}$.
Thus, the negative log-likelihood of $\vec{x}$ is

$$E(y, t) = -\sum_{k=1}^{K} t_k \log y_k$$

The expected categorical cross-entropy for a dataset of N samples is

$$\mathbb{E}\left[-\sum_{k=1}^{K} t_k \log y_k\right]_{data} = \frac{-1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} t_{i,k} \log y_{i,k}$$

Since $\sum_k y_k = 1$ , this cost function works well with

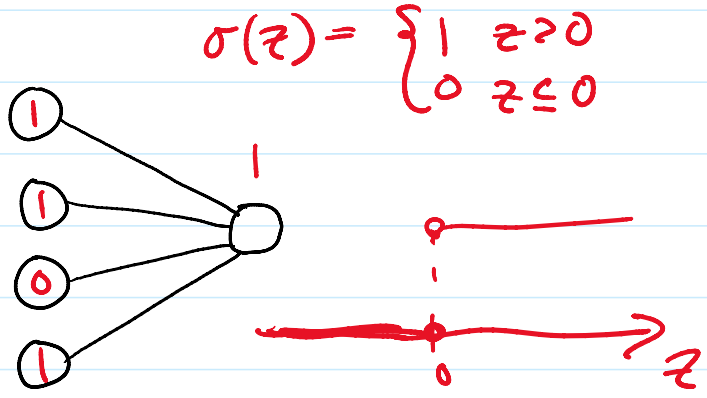
(Look at the new section at the end of Simple Neuron Models.)

# Perceptrons

Goal: To see a simple neural learning algorithm, and understand its limitations.

Let's look at a simple neural network that aims to recognize certain input patterns.

For example, let the output node be a simple threshold neuron, and suppose we want the output node to be 1 when the input is [1, 1, 0, 1], and zero otherwise.




Notice that if we set the weights to [1, 1, 0, 1] (matching the input), then we maximize the input to the output node.

i) $[1\ 1\ 0\ 1] \cdot [1\ 1\ 0\ 1] = 3 \rightarrow \sigma(3) = 1$

(weights · inputs)

But what about other inputs?

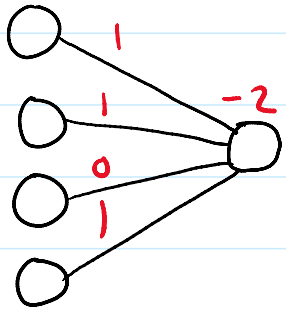
ii) $[1\ 1\ 0\ 1] \cdot [0\ 1\ 1\ 0] = 1 \rightarrow \sigma(1) = 1$

We need the un-matching input to give us a negative value so that the output node returns zero.

Solution: a negative bias

i) $[1\ 1\ 0\ 1] \cdot [1\ 1\ 0\ 1] - 2 \rightarrow \sigma(1) = 1$

ii) $[1\ 1\ 0\ 1] \cdot [0\ 1\ 1\ 0] - 2 \rightarrow \sigma(-1) = 0$




Can we find the weights and bias automatically so that our perceptron produces the correct output for a variety of inputs?

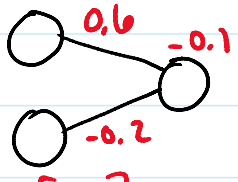
To see an approach, let's look at a 2-D case.

Suppose the 4 different inputs are:

[0, 0], [0, 1], [1, 0], and [1, 1]

And their corresponding outputs are

0, 1, 1, 1 (an "OR" gate)




$w = [0.6\ \ ]$ $b = -0.1$

And their corresponding outputs are
0, 1, 1, 1 (an "OR" gate)

$w = \begin{bmatrix} 0.6 \\ -0.2 \end{bmatrix}$ $b = -0.1$

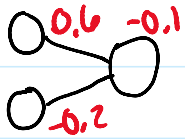
Also, we will use the ~~E~~ error:

$$\Delta E = y - t$$

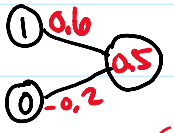
So, if $y=0, t=1 \rightarrow \Delta E = -1$

if $y=1, t=0 \rightarrow \Delta E = 1$

Start with random weights

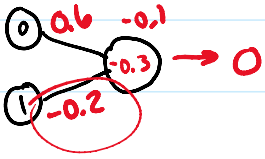



Input [1, 0] target 1




$\Delta E = 0$

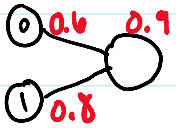
Input [0, 1] target 1 $\quad w \leftarrow w + \Delta w$



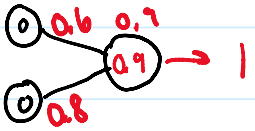

$\Delta E = -1$

$\Delta w = -\Delta E \cdot x$

$\Delta b = -\Delta E$



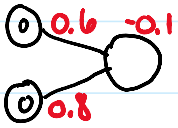

Input [0, 0] target 0



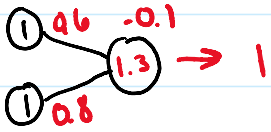

$\Delta E = 1$

$\Delta w = -\Delta E\, x$

$\Delta b = -\Delta E$

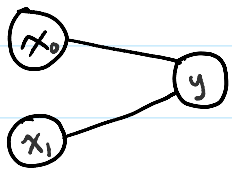



Input [1, 1] target 1




$\Delta E = 0$

Graphical Interpretation

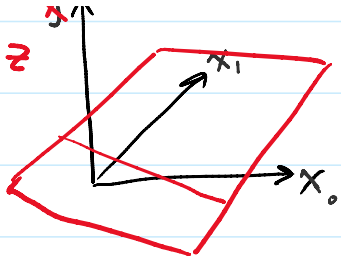



$z = 0.6x_0 + 0.8x_1 - 0.1$

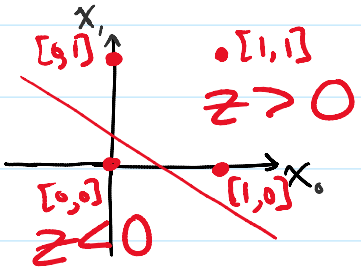
this is a linear equation... the equation of a plane in 3-D.


Looking down on the $x_0$-$x_1$ plane,

$x_1$



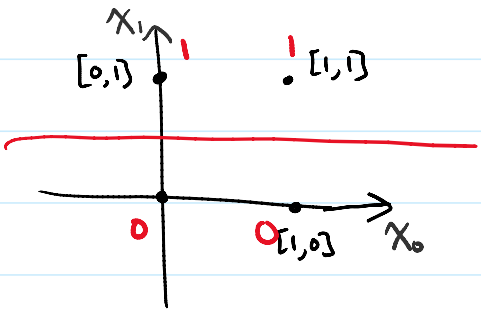

Looking down on the $x_0$-$x_1$ plane,

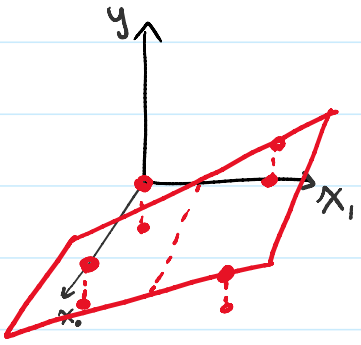



Finding the weights and bias is the same as finding a linear classifier… a linear function that returns a positive value for the inputs that should yield a 1, and a negative value for the inputs that should yield a 0.

Another example:

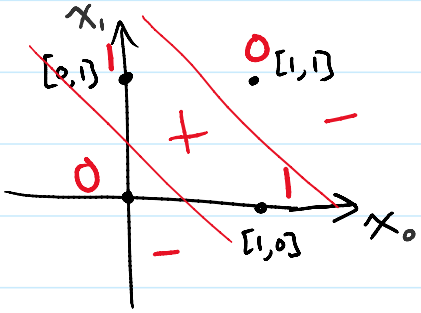
[0, 0] -> 0 [0, 1] -> 1 [1, 0] -> 0 [1, 1] -> 1 "echo $x_1$-value"




A final example:

[0, 0] -> 0 [0, 1] -> 1 [1, 0] -> 1 [1, 1] -> 0 "XOR"




Perceptrons are simple, two-layer neural networks, so only work for linearly separable datasets.

If you want to handle non-linearly-separable data, your neural network is going to need more layers.

But they give us our first glimpse at a learning algorithm.

# Gradient Descent Learning

Goal: To see how we can use a simple optimization method to tune our network weights.

The operation of our network can be written

$$\vec{y} = f(\vec{x}; \theta)$$

So, if our cost function is $E(\vec{y}, \vec{t})$, where $\vec{t}$ is the target, then neural learning becomes the optimization problem

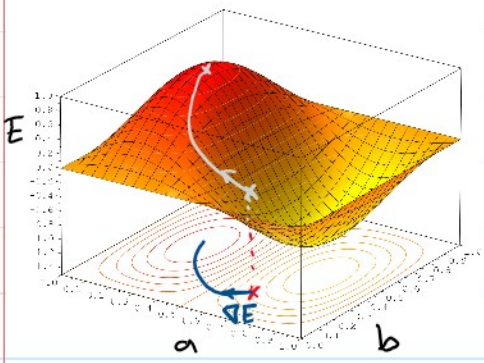
$$E = \min_{\theta} \mathbb{E}\left[ E(f(\vec{x};\theta), \vec{t}(\vec{x})) \right]_{data}$$

We can apply gradient descent to $E$, using the gradient

$$\nabla_\theta E = \left[ \frac{\partial E}{\partial \theta_0} \quad \frac{\partial E}{\partial \theta_1} \quad \cdots \quad \frac{\partial E}{\partial \theta_p} \right]^T$$

## Gradient-Based Optimization

If you want to find a local maximum of a function, you can simply start somewhere, and keep walking uphill.
For example, suppose you have a function with two inputs, $E(a, b)$,
You wish to find $a$ and $b$ to maximize $E$.




https://commons.wikimedia.org/wiki/File:2D_Wavefunction_(2,1)_Surface_Plot.png

We are trying to find the parameters $(\bar{a}, \bar{b})$ that yield the maximum value of $E$.

i.e. $(\bar{a}, \bar{b}) = \underset{(a,b)}{\operatorname{argmax}}\ E(a,b)$

No matter where you are, "uphill" is in the direction of the gradient vector,

$$\nabla E(a,b) = \left[ \frac{\partial E}{\partial a}, \frac{\partial E}{\partial b} \right]^T$$

Gradient ascent is an optimization method where you step in the direction of your gradient vector.
If your current position is $(a_n, b_n)$, then

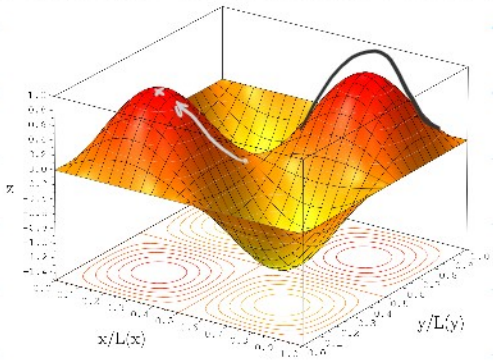
$$(a_{n+1}, b_{n+1}) = (a_n, b_n) + k\, \nabla E(a_n, b_n)$$

where $k$ is your step multiplier.

where $k$ is your step multiplier.

Gradient **DESCENT** aims to **minimize** your objective function. So, you walk downhill, stepping in the direction **opposite** the gradient vector.

Note that there is no guarantee that you will actually find the global optimum. In general, you will find a local optimum that may or may not be the global optimum.



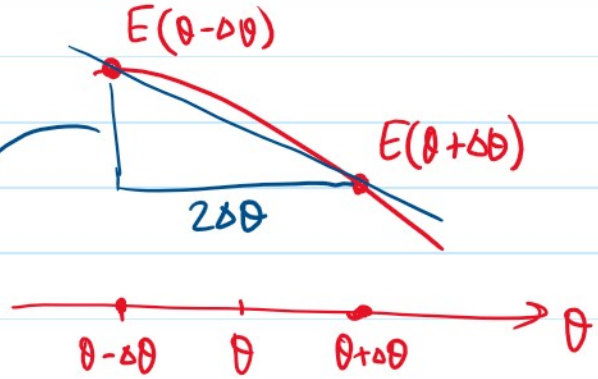

**Approximating the Gradient Numerically**

We can estimate the partial derivatives in the gradient using finite-differencing.

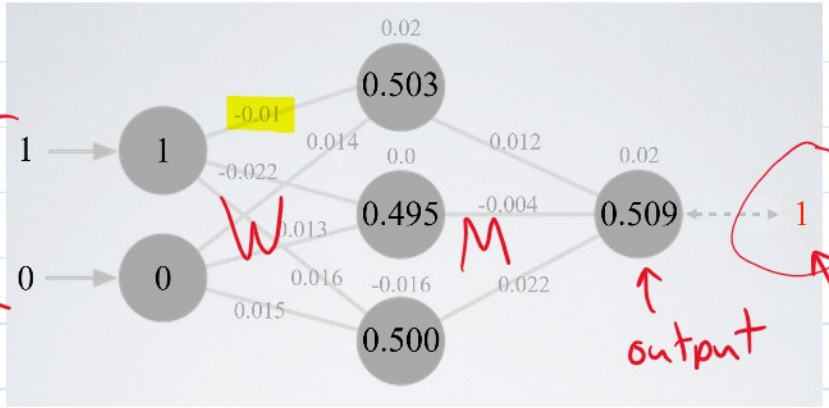
Finite-Difference Approximation

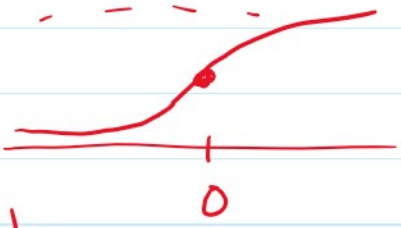
For a function $E(\theta)$, we can approximate $\frac{dE}{d\theta}$ using

$$\frac{dE}{d\theta} \approx \frac{E(\theta+\Delta\theta) - E(\theta-\Delta\theta)}{2\Delta\theta}$$




As an example, consider this network:



It's a neural network, with connection weights and biases shown.

We formulate the problem as

$$\min_{\theta} E\left(f(\vec{x},\theta), \vec{t}(\vec{x})\right)$$

Or, more compactly, $\min_{\theta} \bar{E}(\theta)$ where $\bar{E}(\theta) = E\left(f(\vec{x},\theta), \vec{t}(\vec{x})\right)$

$= (y-t)^2$

Consider $\theta_1$ on its own. With $\theta_1 = -0.01$, our network output is

$y = 0.5089$

This gives

$\bar{E}(0.01) = 0.24113$

What if we perturb $\theta_1$, so that $\theta_1 = -0.01 + \mathbf{0.5} = 0.49$.
Then our output is

$y = 0.5093$

This yields

$\bar{E}(0.49) = 0.240761$

If, instead, we perturb $\theta_1$ so that $\theta_1 = -0.01 - \mathbf{0.5} = -0.51$,
then our output is

$y = 0.5086$

which gives

$\bar{E}(-0.51) = 0.241509$

In summary,

| Parameters | MSE |
|---|---|
| $\theta_1 + \Delta\theta$ | 0.240761 |
| $\theta_1$ | 0.24113 |
| $\theta_1 - \Delta\theta$ | 0.241509 |

We can estimate $\frac{\partial \bar{E}}{\partial \theta_1}$ using

$$\frac{\partial \bar{E}}{\partial \theta_1} \approx \frac{\bar{E}(0.49) - \bar{E}(-0.51)}{2(0.5)}$$

$= -0.0007475$

Obviously, increasing $\theta_1$ seems to be the right thing to do.

$$\theta_1 \leftarrow -0.01 - K(-0.0007475)$$

$$= -0.01 + K(0.0007475)$$

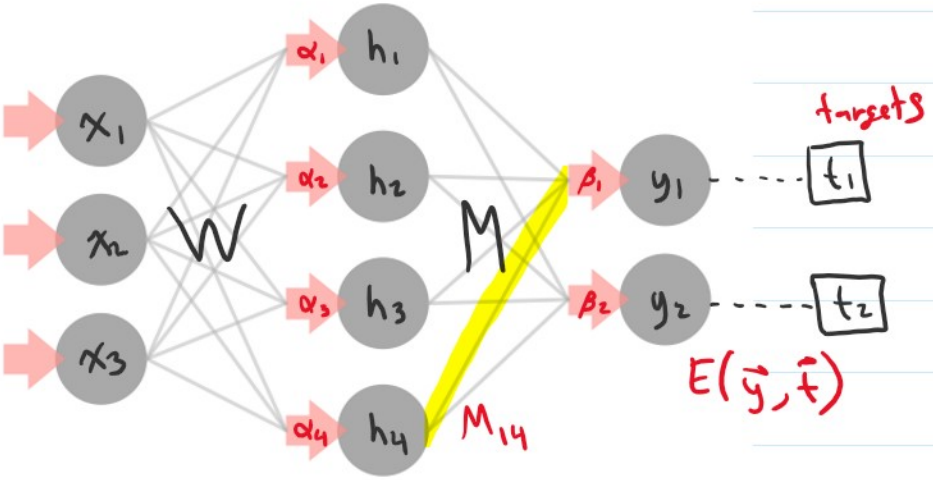
(demo: XOR example)

**Error Backpropagation**

Goal: To find an efficient method to compute the gradients for gradient-descent optimization.

We can apply gradient descent on a multi-layer network, ~~again~~ using chain rule to calculate the gradients of the error with respect to deeper connection weights and biases.

Consider the network




$\alpha_i$ is the input current to hidden node $i$.

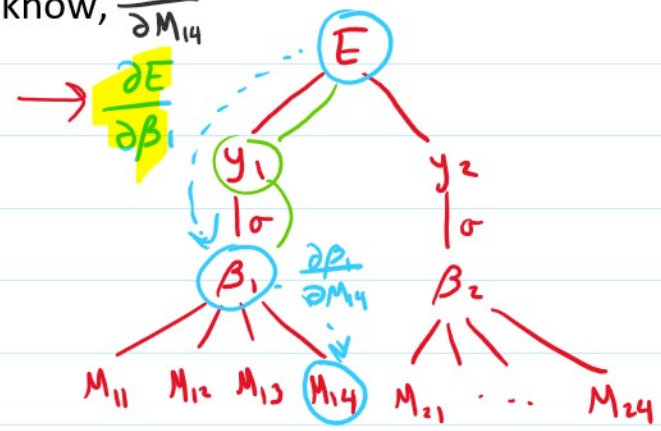
$\beta_j$ is the input current to output node $j$.

For our cost (loss) function, we will use $E(\vec{y}, \vec{t})$

For learning, suppose we want to know, $\frac{\partial E}{\partial M_{14}}$

eg. $M_{14}$

$$\frac{\partial E}{\partial M_{14}} = \frac{\partial E}{\partial \beta_1} \frac{\partial \beta_1}{\partial M_{14}}$$




Recall, $E(\vec{y}, \vec{t})$

$= E(\sigma(M\vec{h} + \vec{b}), \vec{t})$

$\vec{y}$

$\vec{a}$

$= E(\underbrace{\sigma(Mh+b)}_{\vec{\beta}}, T)$

$\therefore \frac{\partial E}{\partial \beta_1} = \frac{\partial E}{\partial y_1}\frac{dy_1}{d\beta_1}$ ← we'll return to this later

Thus, $\frac{\partial E}{\partial M_{14}} = \frac{\partial E}{\partial y_1}\frac{dy_1}{d\beta_1}\frac{\partial \beta_1}{\partial M_{14}}$

Recall $\beta_1 = \sum_{i=1}^{4} M_{1i} h_i + b_1$

$= M_{11}h_1 + M_{12}h_2 + M_{13}h_3 + M_{14}h_4 + b_1$

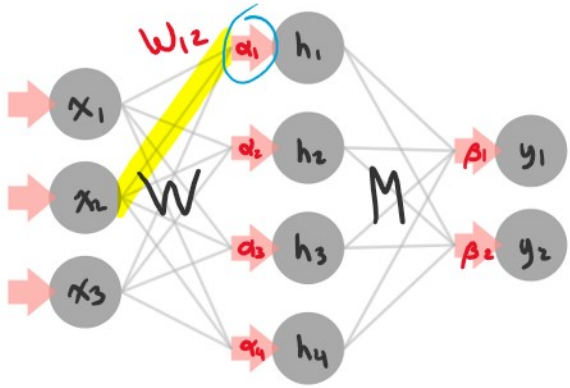
$\frac{\partial \beta_1}{\partial M_{14}} = h_4$

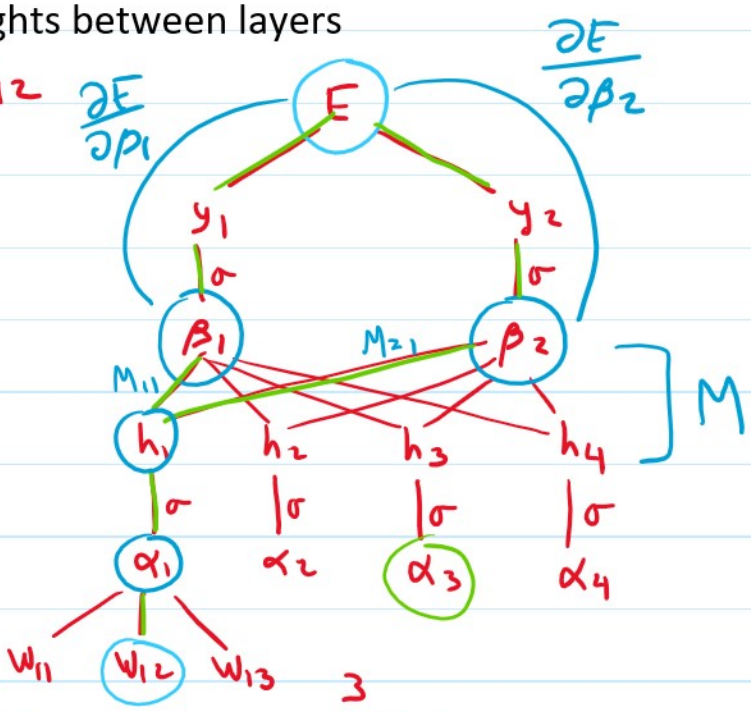
$\frac{\partial E}{\partial \beta_1}$ *

$\therefore \frac{\partial E}{\partial M_{14}} = \frac{\partial E}{\partial y_1}\frac{dy_1}{d\beta_1} h_4$

OK, that works for the connection weights between the top two layers. What about the connection weights between layers deeper in the network?

eg. $W_{12}$




$\frac{\partial E}{\partial W_{12}} = \frac{\partial E}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial W_{12}}$

$\alpha_1 = \sum_{j=1}^{3} W_{1j} x_j + a_j$

$\frac{\partial E}{\partial \alpha_1} = \frac{\partial E}{\partial h_1}\frac{dh_1}{d\alpha_1}$

$= \left(\frac{\partial E}{\partial \beta_1}\frac{\partial \beta_1}{\partial h_1} + \frac{\partial E}{\partial \beta_2}\frac{\partial \beta_2}{\partial h_1}\right)\frac{dh_1}{\cdot}$

$\therefore \frac{\partial \alpha_1}{\partial W_{12}} = x_2$

$$= \left(\frac{\partial E}{\partial \beta_1}\frac{\partial \beta_1}{\partial h_1} + \frac{\partial E}{\partial \beta_2}\frac{\partial \beta_2}{\partial h_1}\right)\frac{dh_1}{d\alpha_1} \qquad \frac{\partial}{\partial w_{12}} - \wedge 2$$

$$h_1 = \sigma(\alpha_1)$$

$$= \left(\frac{\partial E}{\partial \beta_1} M_{11} + \frac{\partial E}{\partial \beta_2} M_{21}\right)\frac{dh_1}{d\alpha_1}$$

we computed these already when we were learning $M$.

$$= \left(M_{11}, M_{21}\right) \cdot \left(\frac{\partial E}{\partial \beta_1}, \frac{\partial E}{\partial \beta_2}\right)\frac{dh_1}{d\alpha_1}$$

More generally, $\vec{x} \in \mathbb{R}^X$, $\vec{h} \in \mathbb{R}^H$, $\vec{y}, \vec{t} \in \mathbb{R}^Y$

$$\frac{\partial E}{\partial \alpha_i} = \left[M_{1i} \cdots M_{Yi}\right] \cdot \left[\frac{\partial E}{\partial \beta_1} \cdots \frac{\partial E}{\partial \beta_Y}\right]\frac{dh_i}{d\alpha_i}$$

This is the $i^{th}$ column of $M$

$$= \left[M_{1i} \cdots M_{Yi}\right] \begin{bmatrix} \frac{\partial E}{\partial \beta_1} \\ \vdots \\ \frac{\partial E}{\partial \beta_Y} \end{bmatrix} \frac{dh_i}{d\alpha_i}$$

For more elements

$$\begin{bmatrix} \frac{\partial E}{\partial \alpha_i} \\ \frac{\partial E}{\partial \alpha_H} \end{bmatrix} =$$

For all elements,

$$\begin{bmatrix} \frac{\partial E}{\partial \alpha_1} \\ \vdots \\ \frac{\partial E}{\partial \alpha_H} \end{bmatrix} = \begin{bmatrix} M_{11} & \cdots & M_{Y1} \\ \vdots & & \\ M_{1H} & \cdots & M_{YH} \end{bmatrix} \begin{bmatrix} \frac{\partial E}{\partial \beta_1} \\ \vdots \\ \frac{\partial E}{\partial \beta_Y} \end{bmatrix} \odot \begin{bmatrix} \frac{dh_1}{d\alpha_1} \\ \vdots \\ \frac{dh_H}{d\alpha_H} \end{bmatrix}$$

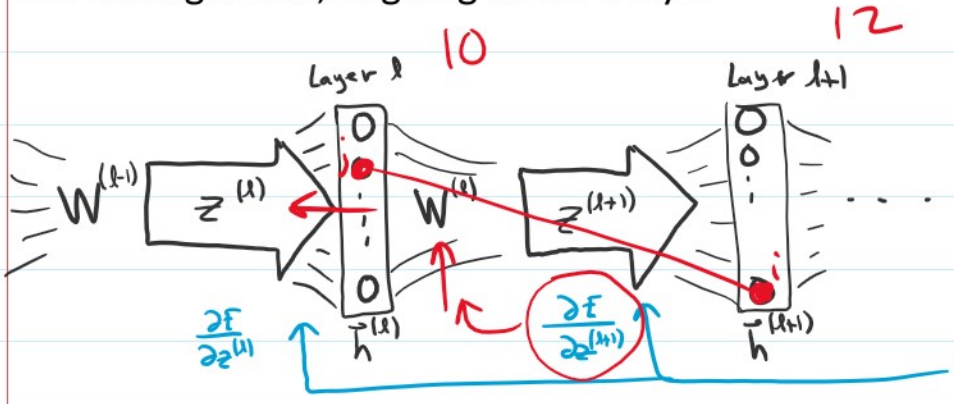
Hadamard Product

$[a\ b] \odot [c\ d] = [ac\ \ bd]$

$$\nabla_\alpha E = \frac{\partial E}{\partial \vec{\alpha}} = \frac{d\vec{h}}{d\vec{\alpha}} \odot M^T \frac{\partial E}{\partial \vec{\beta}} = \frac{d\vec{h}}{d\vec{\alpha}} \odot M^T \nabla_\beta E$$

The most general, in going down a layer

The most general, in going down a layer




Suppose we have $\frac{\partial E}{\partial \vec{z}^{(l+1)}} = \nabla_{l+1} E$

Let $\vec{h}^{(l+1)} = \sigma\left(\vec{z}^{(l+1)}\right) = \sigma\left(W^{(l)} \vec{h}^{(l)} + b^{(l+1)}\right)$

$$\frac{\partial E}{\partial \vec{z}^{(l)}} = \frac{d\vec{h}^{(l)}}{d\vec{z}^{(l)}} \odot \left[W^{(l)}\right]^T \frac{\partial E}{\partial \vec{z}^{(l+1)}}$$

Then, to compute $\frac{\partial E}{\partial W_{ij}^{(l)}}$ ...

$\vec{z}^{(l+1)} = W^{(l)} \vec{h}^{(l)} + \cdots$

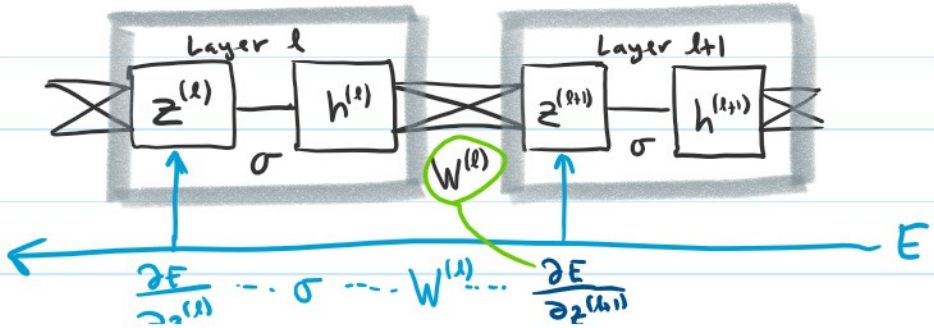
$$\frac{\partial E}{\partial W_{ij}^{(l)}} = \frac{\partial E}{\partial \vec{z}_i^{(l+1)}} \frac{\partial \vec{z}_i^{(l+1)}}{\partial W_{ij}^{(l)}}$$

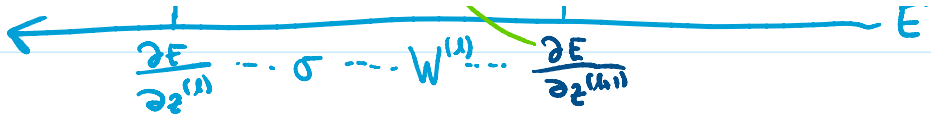
$$= \frac{\partial E}{\partial \vec{z}_i^{(l+1)}} h_j^{(l)}$$

$$\frac{\partial E}{\partial W^{(l)}} = \begin{bmatrix} | \\ \frac{\partial E}{\partial \vec{z}^{(l+1)}} \\ | \end{bmatrix} \left[ - \vec{h}^{(l)} - \right] = \boxed{\text{Same size as } W^{(l)}}$$

outer product

Another graphical summary






$$\frac{\partial E}{\partial z^{(l)}} = \sigma'(z^{(l)}) \odot (W^{(l)})^T \frac{\partial E}{\partial z^{(l+1)}}$$

$$\frac{\partial E}{\partial W^{(l)}} = \frac{\partial E}{\partial z^{(l+1)}} \left[h^{(l)}\right]^T \quad \text{(outer prod)}.$$

END

## Training and Testing

Goal: Develop a process to use our labelled data to generate models that can predict future, unseen samples.

We have seen how to adjust a neural network to get it to learn our training data. Of course, the purpose of training a model is so that we can use it on other samples that aren't in our training set. For this reason, we usually break our data into two pieces:

1. **Training set**: Use most of your labeled data to train your model.
2. **Test set**: Once your model is trained, use the remaining labeled samples to evaluate your model.

Why? Suppose after some trial-and-error, adjusting the hyperparameters, such as:

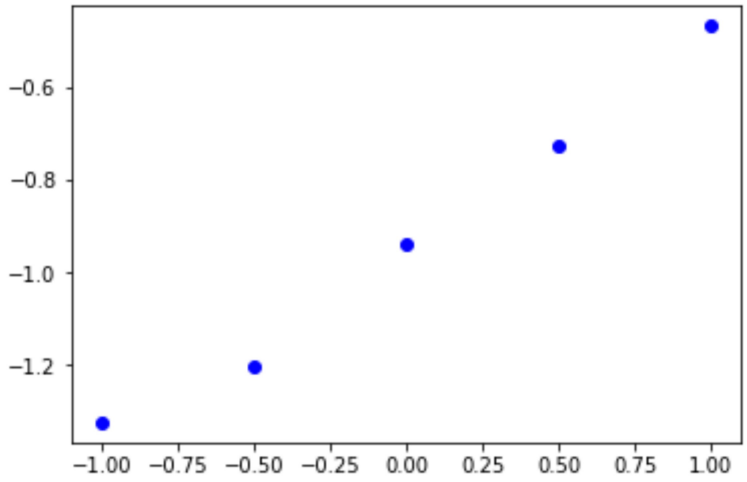
- number of neurons in each layer
- learning rate
- number of epochs
- initial weights

you finally get a low error on the training data. Does that accomplish what you want? Consider an example.

As an experiment, consider noisy samples coming from the ideal mapping,

$$y = 0.4x - 0.9$$

We can only get noisy samples from that mapping.




Our training dataset has 5 samples. And since this is a regression problem, we will use a identity activation function on the output, and MSE as a loss function.

Training usually entails going through the training data repeatedly, updating the network weights as we go. Each pass through the data is called an epoch

Let's create a neural network to learn this mapping.

```
net = Network([1, 1000, 1])
```

# inputs # hidden # outputs.

Before training...

```
Training MSE = 0.955758517256
```

Call the learning function for many epochs.

```
net.Learn([training_input, training_output], epochs=40000, lrate=0.01)
```

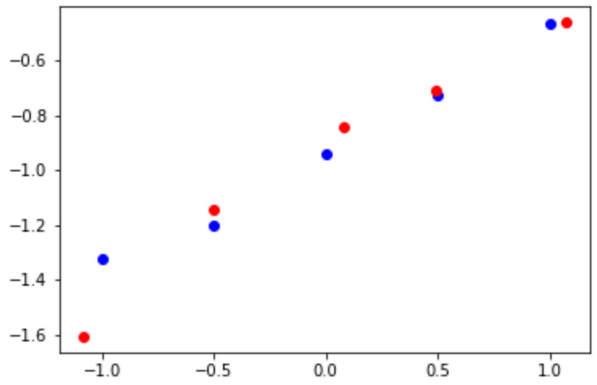
After going through the dataset MANY times, our average loss is

```
Training MSE = 0.000692658082161
```

SUCCESS!

Let's bask in the glory of our brilliance, and demonstrate how

great we are on another sampling of data.




Red samples are 5 new observations

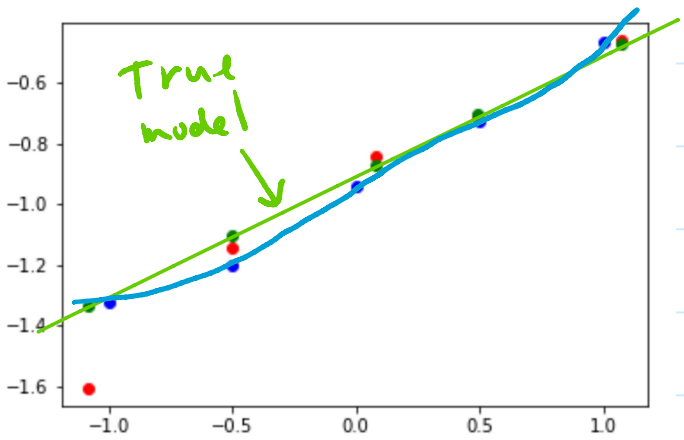
Our average loss on this new set of samples is

```
Test MSE = 0.0156456136501
```

Uh-oh!

This is not as good as our training error!

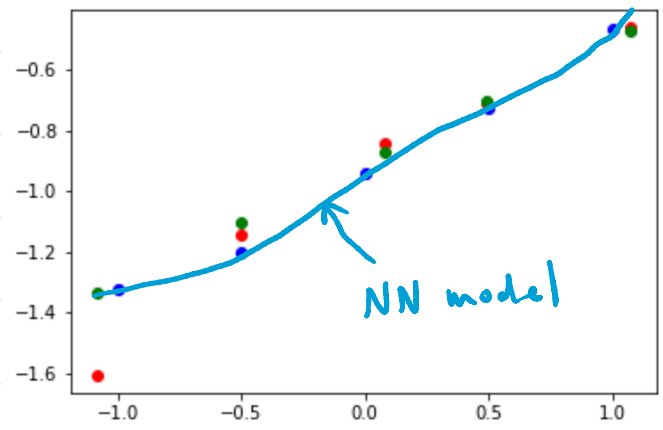
In fact, what happens when we give it a perfect dataset, without noise?




```
Perfect MSE = 0.00161413205859
```

Even perfect data has a higher error than the test data. ???

The false sense of success we get from the results on our training dataset is known as overfitting.



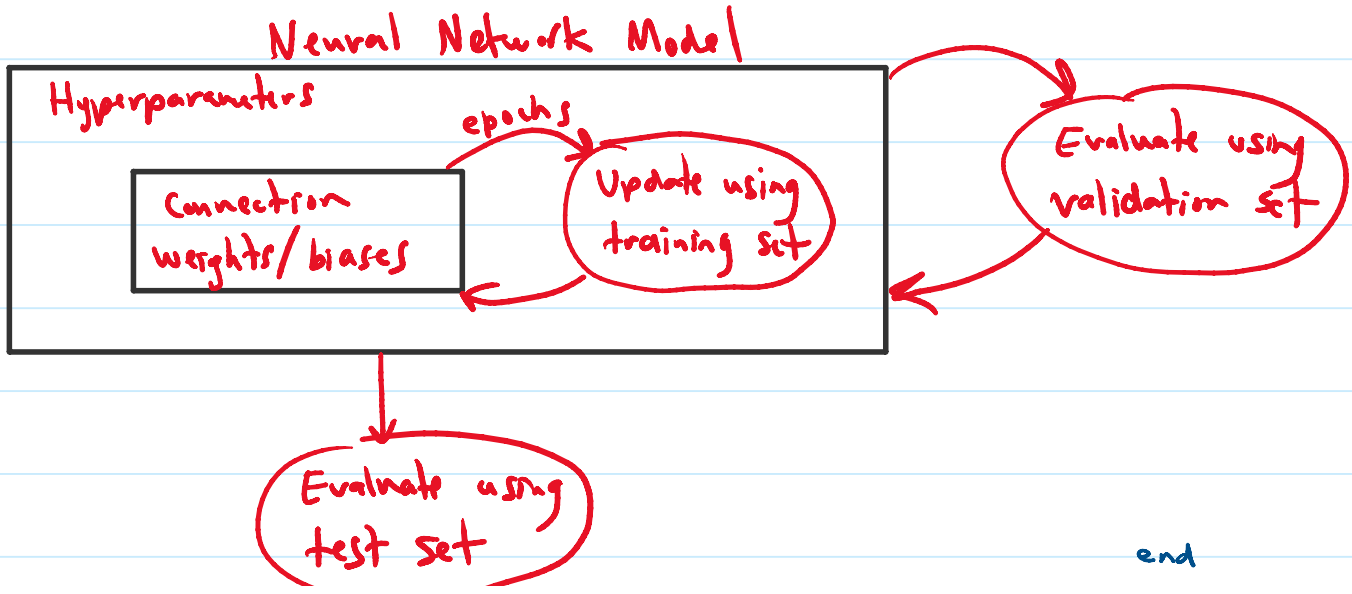

The model starts to fit the noise specific to the training set, rather than just to the underlying model.

Recall that our sole purpose was to create a model to predict the output for samples it hasn't seen. How can we avoid overfitting?

## Validation

If we want to estimate how well our model will generalize to samples it hasn't trained on, we can withhold part of the training set and try our model on that "validation set". Once our model does reasonably well on the validation set, then we have more confidence that it will perform reasonably well on the test set.

It's common to use a random subset of the training set as a validation set.



test set

end

# Overfitting

Goal: See some tricks for how to mitigate ~~against~~ overtraining.

We saw that if a model has enough degrees of freedom, it can become hyper-adapted to the training set, and start to fit the noise in the dataset.

Training error is very small

This is a problem because the model does not generalize well to new samples.

Test error is much bigger than training error

There are some strategies to try to stop our network from trying to fit the noise.

# Regularization

**Weight Decay**

We can limit overfitting by creating a preference for solutions with smaller weights, achieved by adding a term to the loss function that penalizes for the magnitude of the weights.

$$\tilde{E}(y,t;\theta) = E(y,t;\theta) + \lambda \|\theta\|_F^2$$

$$\|\theta\|_F = \sqrt{\sum_j \theta_j^2} \quad \text{"Frobenius Norm"}$$

$$\|\theta\|_F^2 = \theta_1^2 + \theta_2^2 + \cdots + \theta_N^2$$

How does this change our gradients, and thus our update rule?

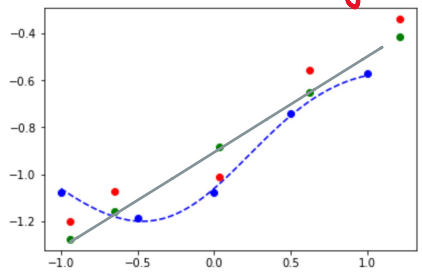
$$\frac{\partial \tilde{E}}{\partial \theta_i} = \frac{\partial E}{\partial \theta_i} + 2\lambda\theta_i$$

$$\theta_{new} \leftarrow \theta_{old} - k\frac{\partial \tilde{E}}{\partial \theta_{old}}$$

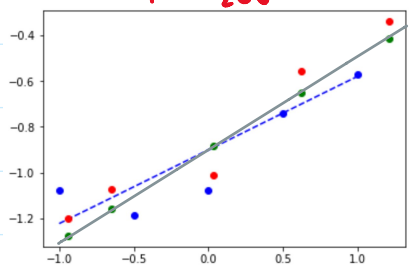
decay

$$\theta_{old} - k\frac{\partial E}{\partial \theta_{old}} - 2k\lambda\theta_{old}$$

$\lambda = 0$ (no regn)




$\lambda = \frac{1}{200}$



```
Training MSE = 0.00596061335994
    Test MSE = 0.00574788521961
 Perfect MSE = 0.00178339752734
```

$\lambda$ controls the weight of the regularization term.

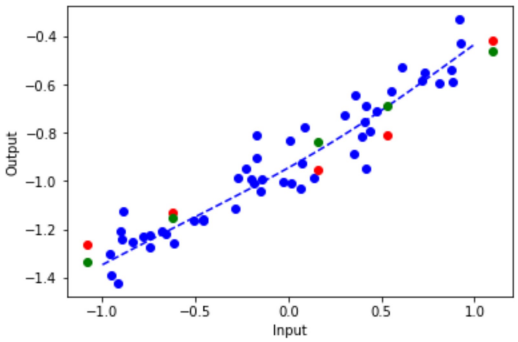
One can also use different norms. For example, it is common to use the L1 norm,

$$L_1(\theta) = \sum_i |\theta_i| \qquad \frac{\partial L_1}{\partial \theta} = \text{sign}(\theta)$$

The L1 norm tends to favour sparcity (most weights are close to zero, with only a small number of non-zero weights).

**Data Augmentation**

Another approach is to include a wider variety of samples in your training set, so that the model is less likely to focus its efforts on the noise of a few.




```
Training MSE = 0.00694267279424
    Test MSE = 0.00685380408996
 Perfect MSE = 0.00627912229049
```

Where does this extra data come from?

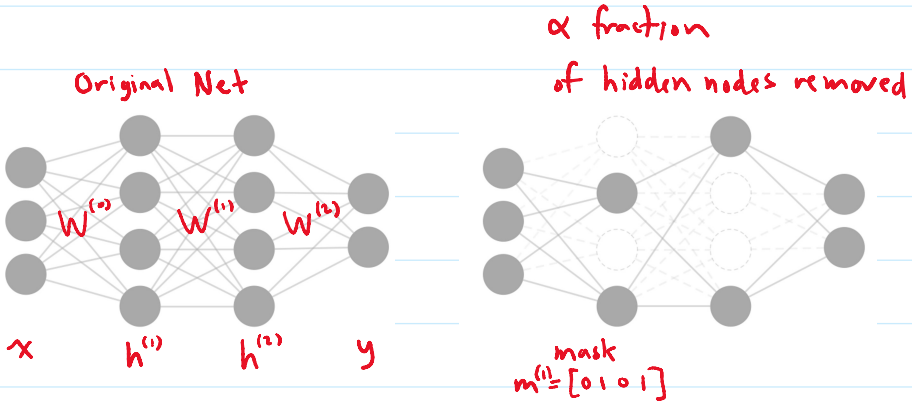
For image-recognition datasets, one can generate more samples by shifting or rotating the images. Those transformations presumably do not change the labelling.

**Dropout**

The last method we will talk about is the most bizarre.

While training using the dropout method, you systematically ignore a large fraction (typically half) of the hidden nodes for each sample. That is, given a dropout probability, $\alpha$, each hidden node will be dropped with probability $\alpha$.

A dropped node is temporarily taken off-line and set to zero.




Do both a feedforward and backprop pass with this diminished network.

$$z^{(1)} = W^{(0)}x + b^{(1)} \Rightarrow \hat{h}^{(1)} = \sigma(z^{(1)}) \in \mathbb{R}^{N^{(1)}}$$

Then apply mask $m^{(1)} = \{0,1\}^{N^{(1)}}$

$$h^{(1)} = \hat{h}^{(1)} \odot m^{(1)}$$

Important caveat:

Removing a bunch of the hidden nodes reduces the input current to the next layer. The weights learned in this context will not work properly in the full network.
We have to scale the activities from a diminished layer in order to give reasonable inputs to the next layer.

We scale up the remaining activities of the diminished layer.
Suppose only $R$ of the $N^{(1)}$ nodes remained.
Let $h^{(1)}$ be the scaled activities, where $h^{(1)} = \frac{N^{(1)}}{R} h^{(1)}$
This give us

$$\mathbb{E}\left[\sum_i h_i^{(1)}\right] = \mathbb{E}\left[\sum_i \hat{h}_i^{(1)}\right]$$

on average $\frac{N^{(1)}}{R} = \frac{1}{1-\alpha}$

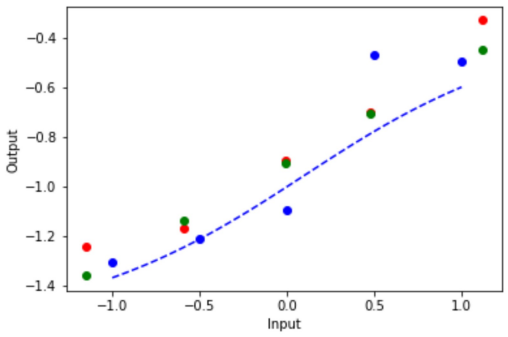
We store these scaled activities in our hidden nodes, but also record the scale factor, $\frac{N^{(1)}}{R}$.

During backprop, suppose we have to make sure we undo this scaling as we pass the gradient back. We want to apply weight updates as though we were using the full network (no dropouts).
For example,

$$\frac{\partial E}{\partial W^{(1)}} = \frac{\partial E}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial W^{(1)}} = \frac{\partial E}{\partial z^{(2)}} h^{(1)} = \frac{\partial E}{\partial z^{(2)}} \left(\frac{R}{N^{(1)}} \hat{h}^{(1)}\right)$$

We have to un-scale the hidden node activities during backprop

We have to un-scale the hidden node activities during backprop.




```
Training MSE = 0.0234801932225
    Test MSE = 0.0212572143563
 Perfect MSE = 0.00874342134987
```

Why does dropout work?

- It's akin to training a bunch of different networks and combining their answers. Each diminished network is like a contributor to this consensus strategy.
- Dropout disallows sensitivity to particular combinations of nodes. Instead, the network has to seek a solution that is robust to loss of nodes.

End of L12

# Enhancing Optimization

Goal: To learn some methods that help learning go faster.

Suppose our training set is

$$\{(x_1,t_1),\cdots,(x_P,t_P)\} \quad \text{where } \vec{x}\in\mathbb{R}^X,\ \vec{t}\in\mathbb{R}^Y$$

Notice that we can put all of our inputs into a single matrix

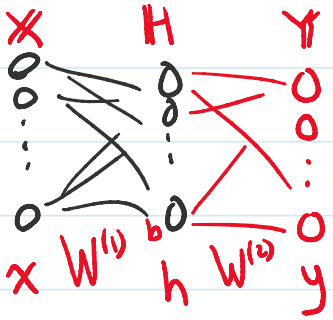
$$X=\left[x_1\middle|x_2\middle|\cdots\middle|x_P\right]$$

As well as our targets

$$T=\left[t_1\middle|\cdots\middle|t_P\right]$$

Does this help us? Yes!

Consider the 1st hidden layer...




$$\vec{z}^{(1)}=W^{(1)}\vec{x}+\vec{b} \qquad (H\times X)$$

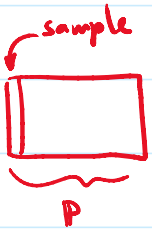
Then $\vec{h}^{(1)}=\sigma(\vec{z}^{(1)})$

But we can process all $P$ inputs at once!

$$Z^{(1)}=W^{(1)}X+b\underbrace{[1\cdots 1]}_{P} \qquad (H\times X)(X\times P)$$

$$[\,\cdots\,]=\left[b\middle|b\middle|\cdots\middle|b\right]$$

Then, $H^{(1)}=\sigma(Z^{(1)})$

At the top layer, we get $Y\in\mathbb{R}^{Y\times P}=$ 


$$E(Y,T)=\frac{1}{P}\sum_{p=1}^{P}E(y_p,t_p)$$

Now, working our way back down,

$$\frac{\partial E}{\partial}\quad(Y-T)\perp\quad\cdots P$$

Now, working our way back down,

$$\frac{\partial E}{\partial Z^{(n+1)}} = (Y - T)\frac{1}{P} \qquad Y \times P$$

And going down one layer

$$\frac{\partial E}{\partial Z^{(1)}} = \frac{dH}{dZ^{(1)}} \odot \left(W^{(1)}\right)^T \frac{\partial E}{\partial Z^{(n+1)}}$$

$H \times P \quad H \times P \quad H \times Y \quad Y \times P$

Then,

$$\frac{\partial E}{\partial W^{(1)}} = \frac{\partial E}{\partial Z^{(n+1)}} \left(H^{(1)}\right)^T$$

$Y \times H \quad Y \times P \quad P \times H$

$H = 20 \quad Y = 10$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} 4 & 5 \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 8 & 10 \\ 12 & 15 \end{bmatrix}$$

$$= \sum_{10} \begin{bmatrix} \frac{\partial E}{\partial Z_1^{(n+1)}} \end{bmatrix} \begin{bmatrix} - h_1^{(1)} - \end{bmatrix} + \cdots + \begin{bmatrix} \frac{\partial E}{\partial Z_P^{(n+1)}} \end{bmatrix} \begin{bmatrix} - h_P^{(1)} - \end{bmatrix}$$

$\leftarrow 20 \rightarrow$

Hence, we can use the same formulas to process a whole batch of samples.

One problem: processing the entire dataset for a single update to the weights can be slow. Instead, there is an intermediate approach.

**Stochastic Gradient Descent**

Computing the gradient of the cost function can be very expensive and time-consuming, especially if you have a huge training set.

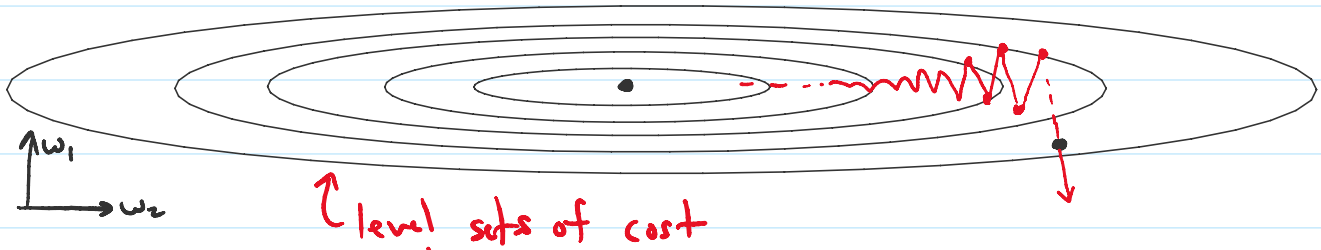
Rather than compute the full gradient, we can try to get a cheaper estimate by computing the gradient from a random sampling.

We use the estimate from this batch to update our weights, and then choose subsequent batches from the remaining samples.

This method is called Stochastic Gradient Descent.

## Momentum

Consider gradient descent optimization in this situation...




level sets of cost function in space of connection weights.

The shape of the cost function causes oscillation, and this back-and-forth action can be inefficient.

Instead, we can smooth out our trajectory using

Recall from physics,

$$\frac{dD}{dt} = V \quad \text{(velocity)} \qquad \frac{dV}{dt} = A - \hat{r}V \quad \text{(acceleration)}$$

So, solving numerically using Euler's method...

$$D_{n+1} = D_n + \Delta t\, V_n \qquad (1)$$

$$V_{n+1} = (1-r)V_n + \Delta t\, A_n$$

$r$ is resistance from friction

When driving, this is what you control. eg. gas pedal, brakes, steering.

In our previous optimization method, $D$ represented our weights, and $V$ was like our error gradients. Then we updated our weights using (1)

Distance $D_{n+1} = D_n + \Delta t\, V_n$

Weights $W_{n+1} = W_n - K\frac{\partial E}{\partial w_n}$

But we can instead treat our error gradients as $A$, and integrate to get an accumulated weight update, akin to $V$

But we can instead treat our error gradients as $A$, and integrate to get an accumulated weight update, akin to $V$.
In fact, let's call it $V$.
It's like our weights are dictated by our location in parameter space,


and we move around weight space, accelerated by the error gradients. We build speed if we get a lot of acceleration in the same direction.
We gain momentum.

For each weight $W_{ij}$, we also calculate $V_{ij}$.
Or, in matrix form, for each $W^{(l)}$, we have $V^{(l)}$

$$V^{(l)} \leftarrow (1-r)V^{(l)} + \frac{\partial E}{\partial W^{(l)}}$$

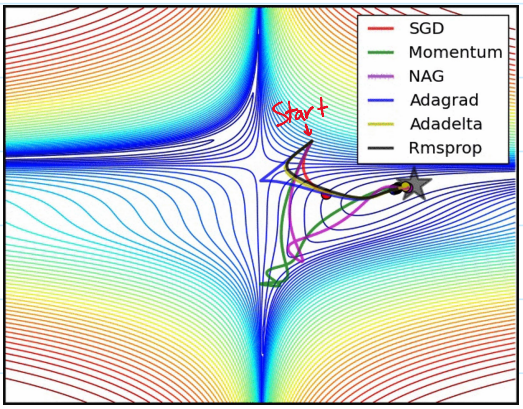
Or, as is commonly used,

$$V^{(l)} \leftarrow \beta V^{(l)} + (1-\beta)\frac{\partial E}{\partial W^{(l)}}$$

Then, update our weights using

$$W^{(l)} \leftarrow W^{(l)} - \kappa V^{(l)}$$

Not only does this smooth out oscillations, but can also help to avoid getting stuck in local minima.




http://ruder.io/optimizing-gradient-descent/

End of 14th lecture

# Deep Neural Networks

Goal: To see the advantages and disadvantages of deep neural networks: representational power vs. vanishing or exploding gradients.

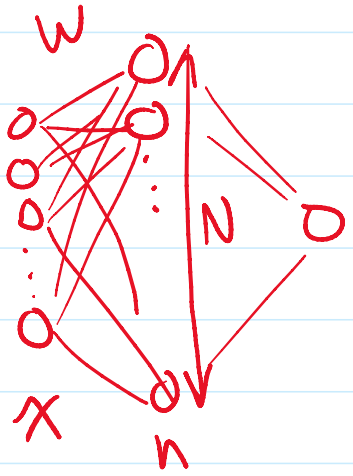
How many layers should our neural network have?

Recall the Universal Approximation Theorem:

> **Theorem 2.** *Let $\sigma$ be any continuous sigmoidal function. Then finite sums of the form*
>
> $$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j x + \theta_j)$$
>
> *are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which*
>
> $$|G(x) - f(x)| < \varepsilon \qquad \text{for all} \quad x \in I_n.$$

Cybenko G, "Approximation by Superpositions of a Sigmoidal Function", *Math. Control Signals Systems*, 2:303-314, 1989.

Thus, we really only ever need one hidden layer. But is that the best approach, either in number of nodes, or learning efficiency? No, it can be shown that such a shallow network could require an exponentially large number of nodes (ie. A really big $N$ ) to work.

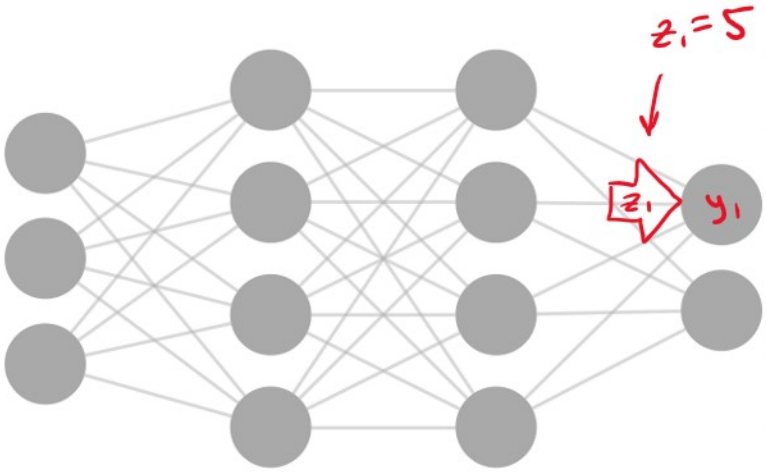



So, a deeper network is preferred in many cases. (see Python example)

So, why don't we always use really deep networks?

# Vanishing Gradients

Suppose the initial weights and biases were large enough that the input current to many of the nodes was not too close to zero. As an example, consider one of the output nodes.



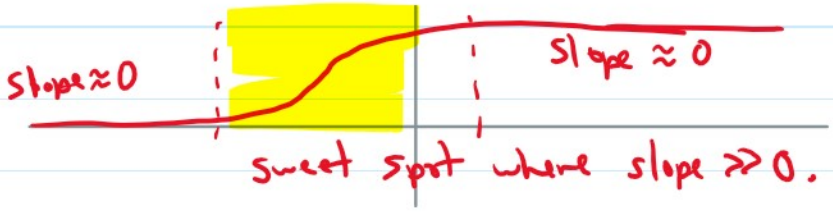

$$y_1 = \sigma(z_1) = \frac{1}{1+e^{-5}} = 0.9933\ldots$$

$$\frac{dy_1}{dz_1} = y_1(1-y_1) = 0.0066$$

Compare that to if the input current was 0.1.

$y_1 = \sigma(0.1) = 0.525$

$\frac{dy_1}{dz_1} = 0.249$ Almost 40x larger than




Hence, the updates to the weights will be smaller when the input currents are large in magnitude.

What about the next layer down?

Suppose $\frac{\partial E}{\partial z^{(4)}} \approx 0.1$

What if the inputs to the penultimate layer were around 4 in magnitude?
Then the corresponding slopes of their sigmoid functions will also be small.

$\sigma(4) = 0.982$

$\sigma'(4) = 0.0177$

Recall that

$$\frac{\partial E}{\partial z^{(3)}} = \frac{dh^{(3)}}{dz^{(3)}} \odot \left(W^{(3)}\right)^T \frac{\partial E}{\partial z^{(4)}}$$

Recall that

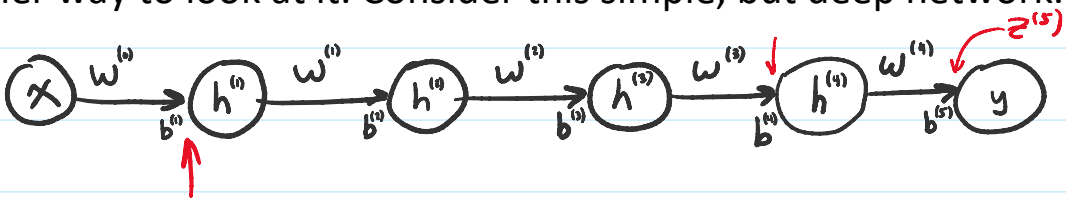
$$\frac{\partial E}{\partial z^{(3)}} = \frac{dh^{(3)}}{dz^{(3)}} \odot \left(W^{(3)}\right)^T \frac{\partial E}{\partial z^{(4)}}$$

$$= (\text{approx } 0.0177) \odot \left(W^{(3)}\right)^T (\text{approx } 0.1)$$

$$= (\text{approx. } 0.00177) \left(W^{(3)}\right)^T$$

And it gets smaller and smaller as you go deeper.
When this happens, learning comes to a halt, especially in the deep layers. This is often called the
vanishing gradients problem.

Another way to look at it. Consider this simple, but deep network.




Start with the loss on the output side: $E(y,t)$
The gradient w.r.t. the input current of the output node is

$$\frac{\partial E}{\partial z^{(5)}} = y - t$$

Then, using backprop, we can compute a single formula for

$$\frac{\partial E}{\partial z^{(4)}} = (y-t)\, w^{(4)} \sigma'(z^{(4)})$$

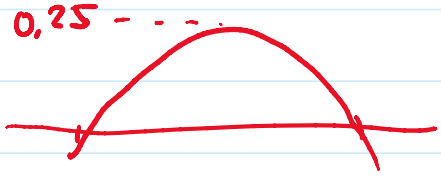
Going deeper...

$$\frac{\partial E}{\partial z^{(1)}} = (y-t)\, w^{(4)} \sigma'(z^{(4)})\, w^{(3)} \sigma'(z^{(3)})\, w^{(2)} \sigma'(z^{(2)})\, w^{(1)} \sigma'(z^{(1)})$$

What is the steepest slope that $\sigma(z)$ attains?

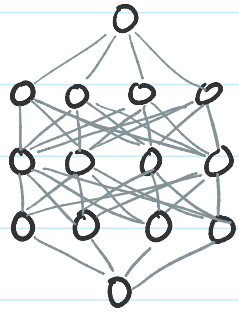
$$\sigma(z) = \sigma(z)(1-\sigma(z))$$

for $0 \le \sigma(z) \le 1$



All else being equal, the gradient goes down by a factor of at least 4 each layer.
We can see this if we look at the norm of the gradients at each layer.

$$\text{i.e. } \left\|\frac{\partial E}{\partial z^{(i)}}\right\|^2 = \sum_j \left(\frac{\partial E}{\partial z_j^{(i)}}\right)^2$$



5 layers
layer 4 norm=0.7598 [ 0.76]

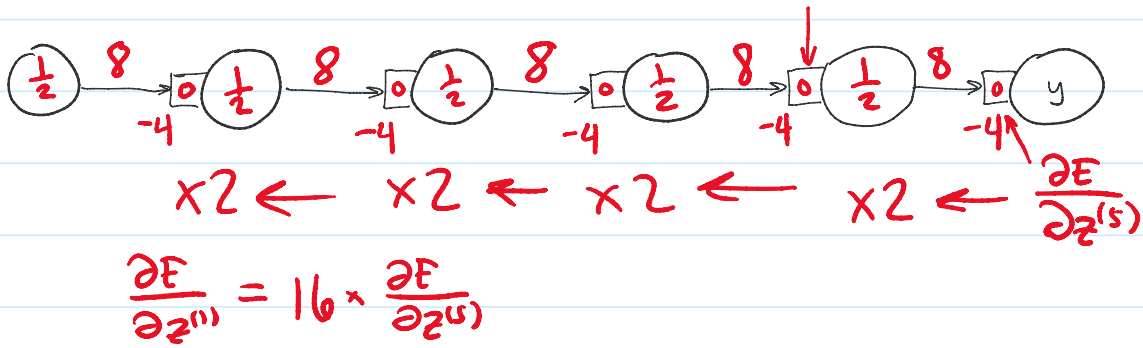
layer 3 norm=0.2266 [-0.038 -0.217 -0.039 -0.039]

layer 2 norm=0.0316 [ 0.018 -0.018 -0.007  0.018]

layer 1 norm=0.0094 [-0.004  0.002  0.001  0.008]

## Exploding Gradients

A similar, though less frequent phenomenon can result in very large gradients.




This situation is more rare since it only occurs when the weights are high and the biases compensate so that the input current lands in the sweet spot of the logistic curve.