Machine Learning I

Lecture 8: Artificial Neural Networks

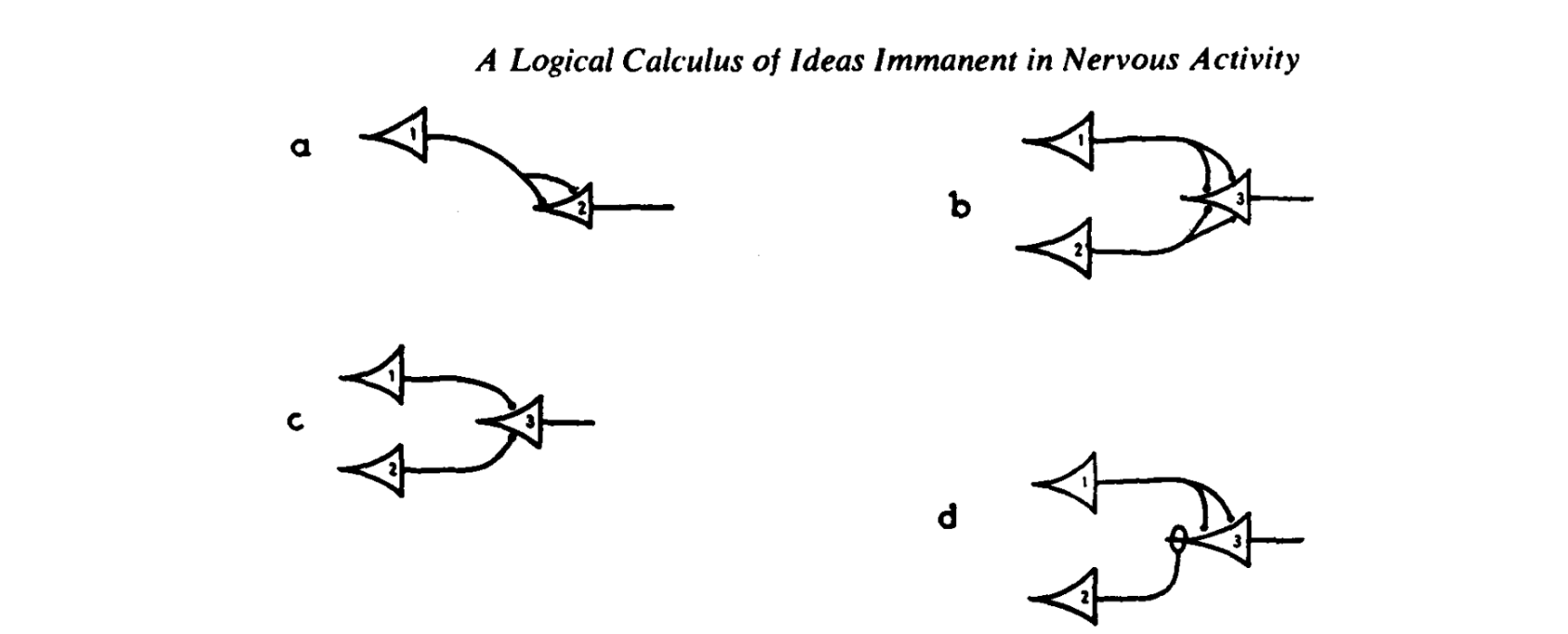
Nathaniel Bade

Northeastern University Department of Mathematics

Table of contents

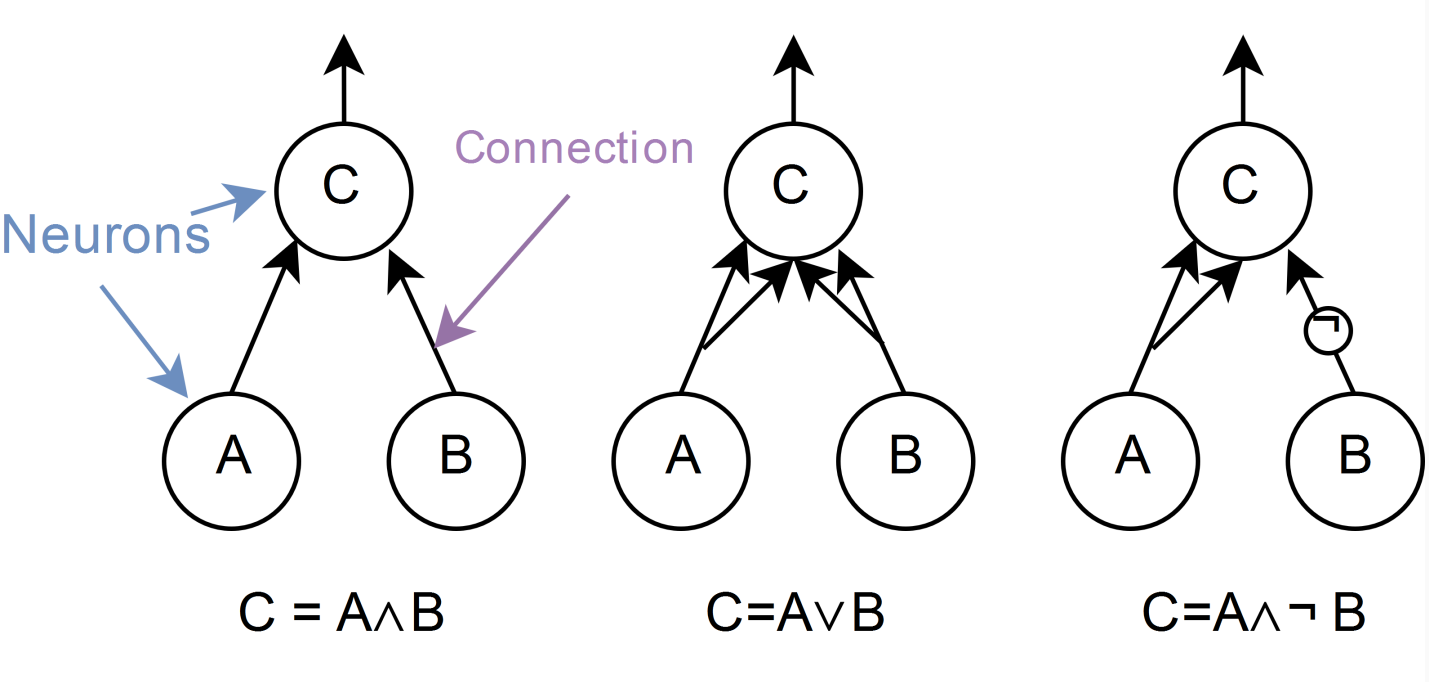
Artificial Neural Networks

Neural Networks



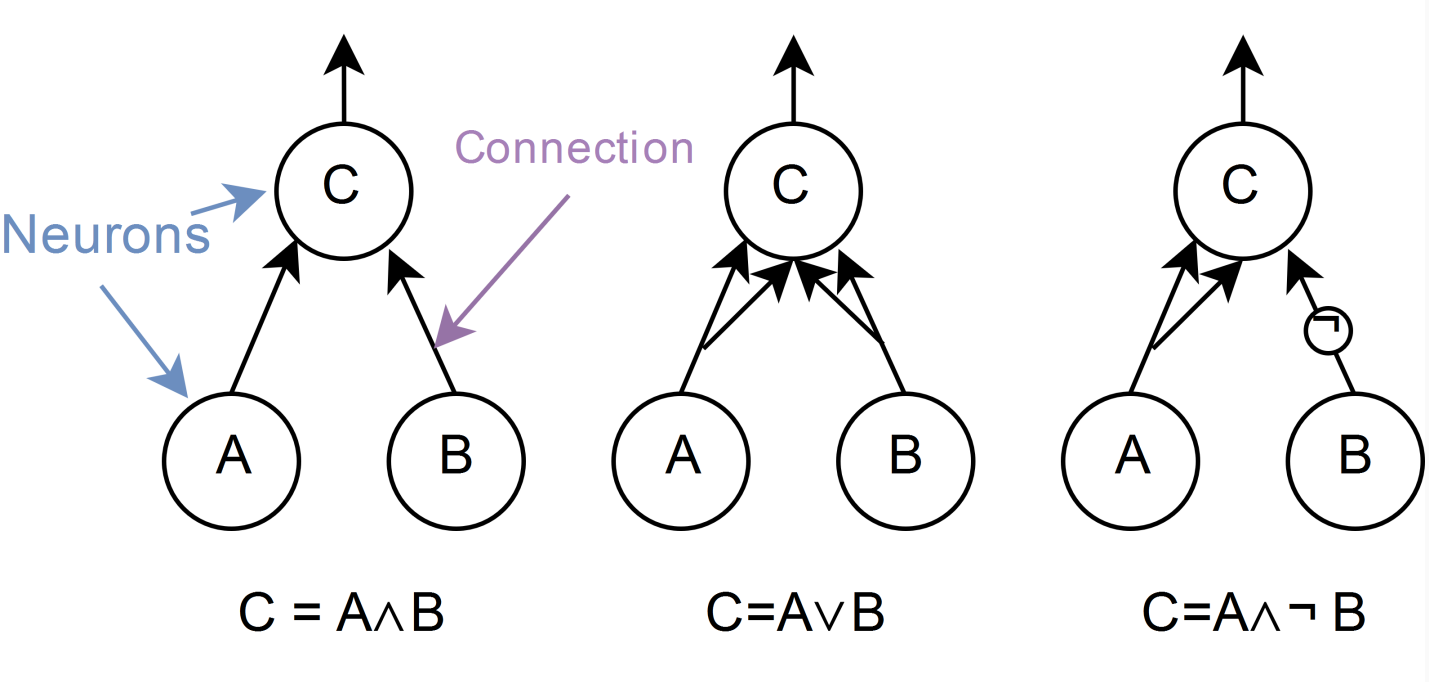
Neural Networks are actually quite old, being introduced in 1943 by neurophysiologist Warren McCulloch and mathematician Walter Pitts to model neurons in the brain using electrical circuits.

Neural Networks as Logic Gates



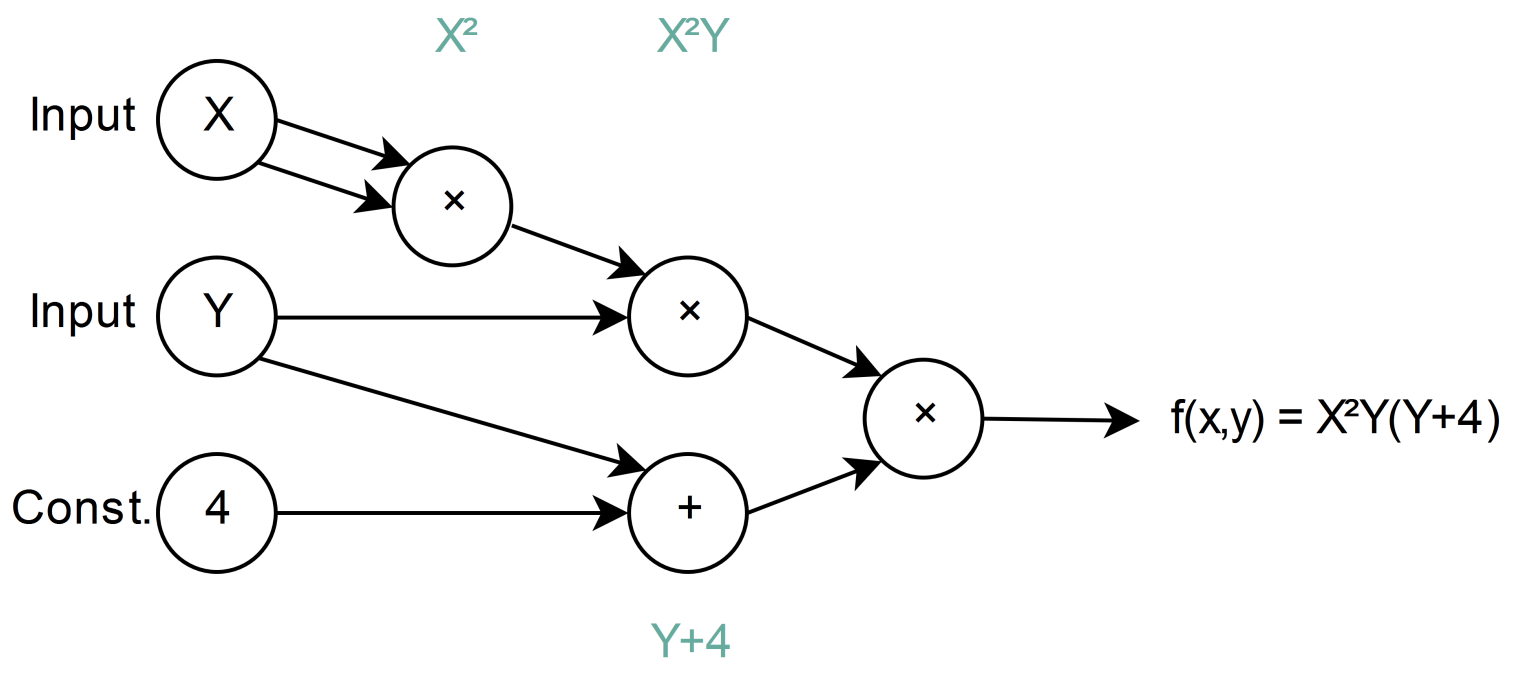
The original neural networks were directed graphs. Each node has a state **on** (firing) or **off** (no firing). Node's also have a threshold t and only fire if more than t imputing nodes are firing.

Neural Networks as Logic Gates



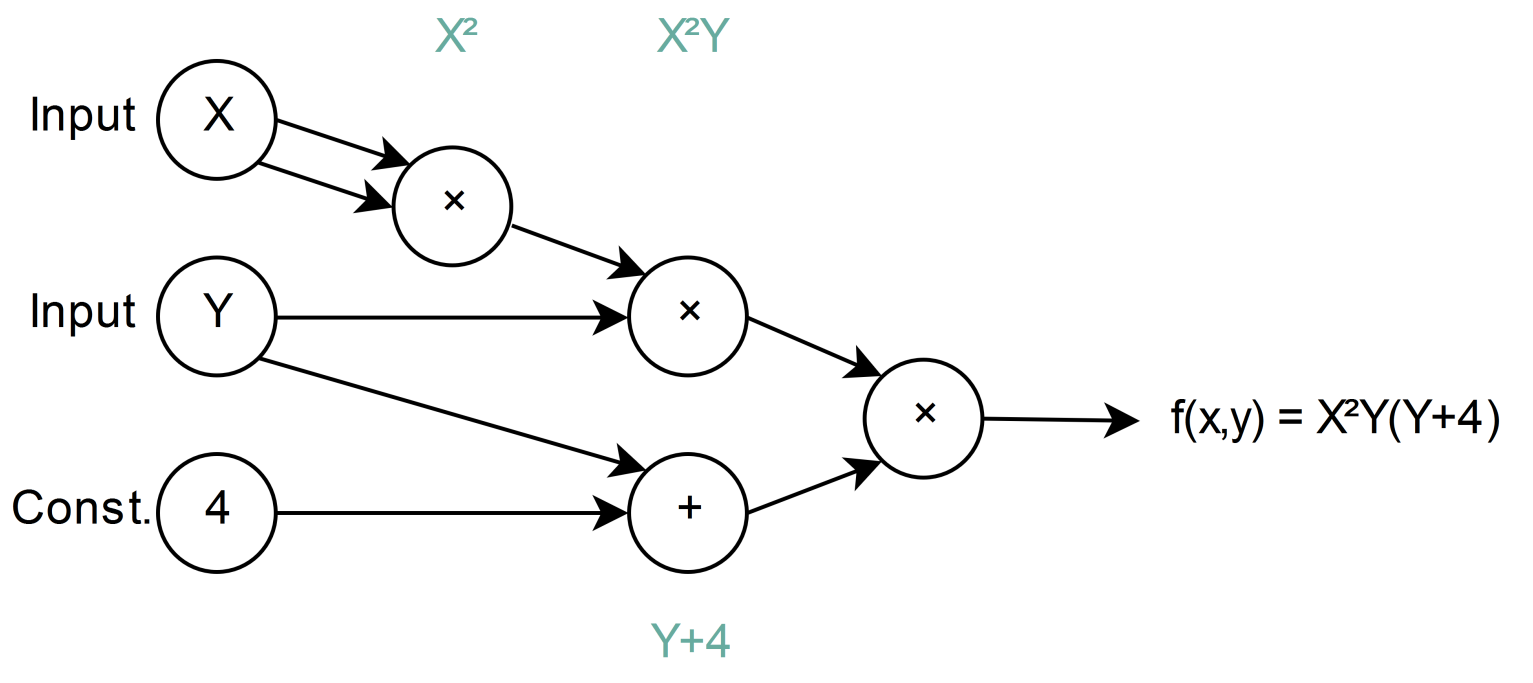
McCulloch and Pitts showed that the standard operations of propositional logic could be built out of strings of such neurons. In the third case, the $\neg$ is a inhibitor connection which stops C from firing. Such connections do appear in biological neurons.

Neural Networks as Operation Trees



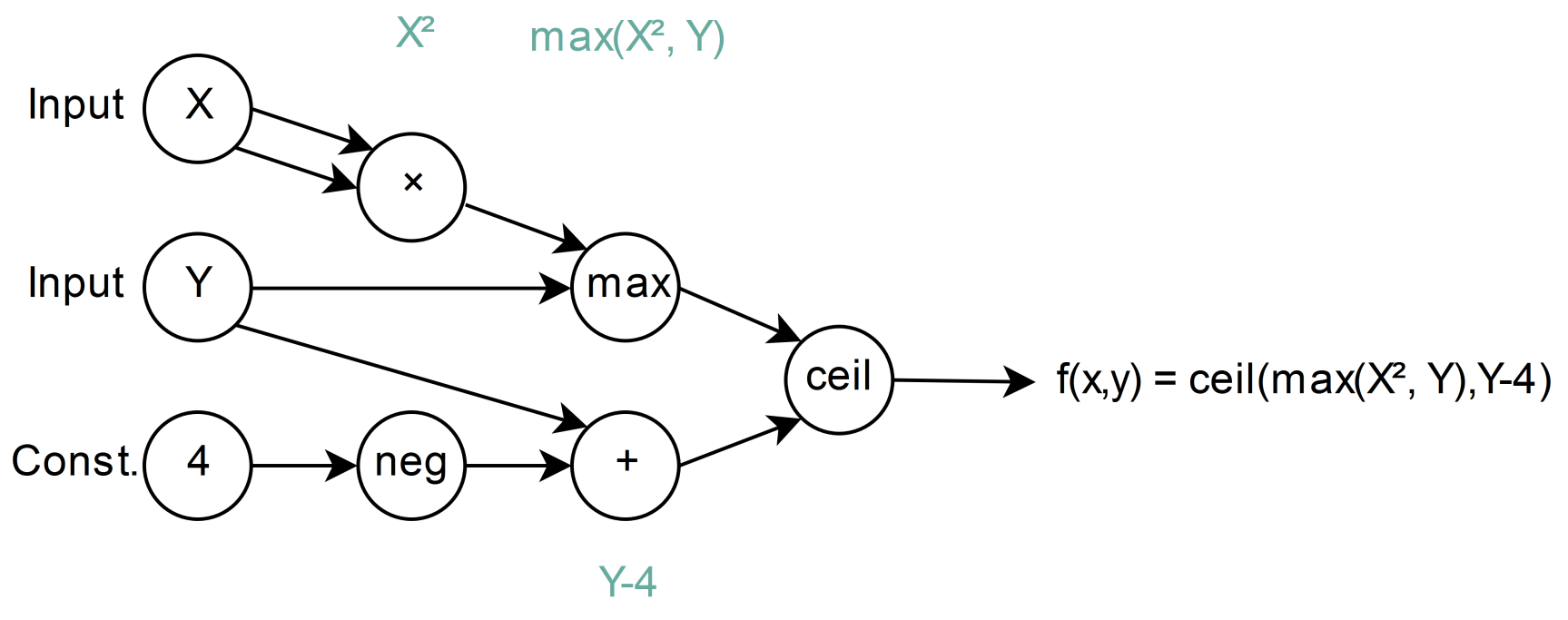
Another view of neural networks (the view most implementations take) is as operation trees. In an operation tree, one constructs a function by giving a sequential list of instructions. Each node represents an operation with a fixed number of inputs and outputs. In an operation tree there is no threshold, and the nodes may output any data type.

Neural Networks as Operation Trees



Viewing neural networks as operation trees give us an intuitive way to build functions out of concrete operational pieces. Furthermore, depending on the allowed nodes it gives a clear representation of nondifferentiable functions.

Neural Networks as Operation Trees



Viewing neural networks as operation trees give us an intuitive way to build functions out of concrete operational pieces. Furthermore, depending on the allowed nodes it gives a clear representation of nondifferentiable functions.

Neural Networks as Operation Trees

Its important to point out that these are two wildly different objects mathematically and philosophically:

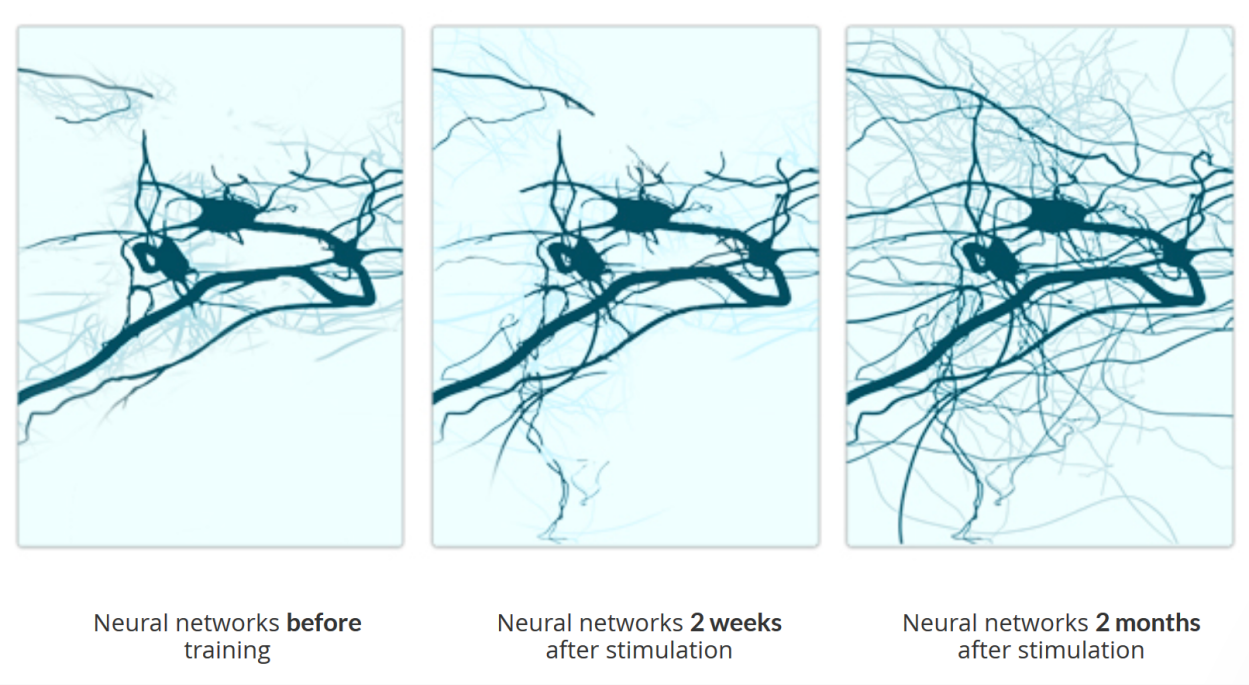
Neural Networks:

- Finite number of types of nodes.
- Binary input and output.
- Interesting behavior comes from topologically.
- Idea: To show how a brain can be constructed from simple pieces.

Operation Trees:

- Nodes are arbitrary functions.
- Arbitrary input and output.
- Topologically trivial as graphs.
- Any function can be constructed so no philosophical interest at full generality.

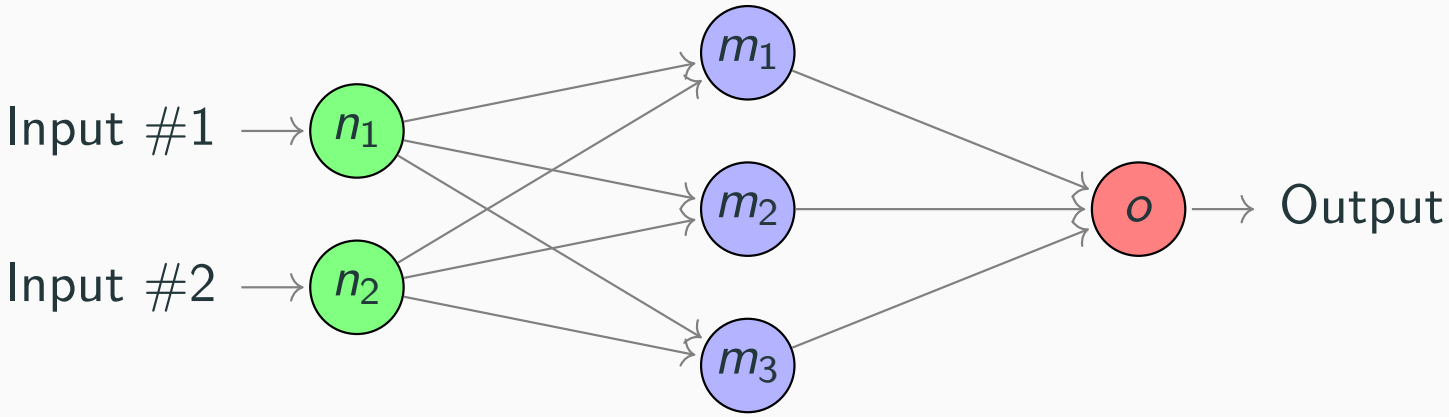
Towards Modern Neurons



In 1949, psychologist Donald Hebb pointed out the neural pathways are strengthened each time they were used, a concept which gave a potential mechanism to human learning.

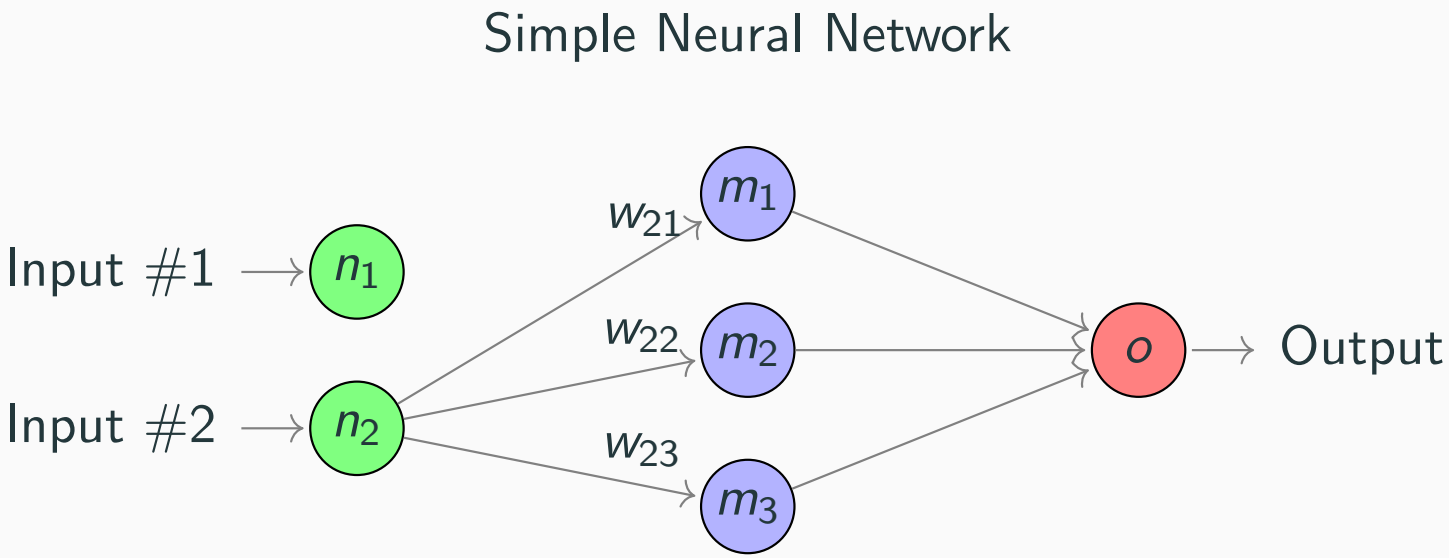
Towards Modern Neurons

Simple Neural Network



In 1949, psychologist Donald Hebb pointed out the neural pathways are strengthened each time they were used, a concept which gave a potential mechanism to human learning. Practically, this means that each of the edges connecting the nodes can carry a multiplicitive weight which appropriately scales the input.

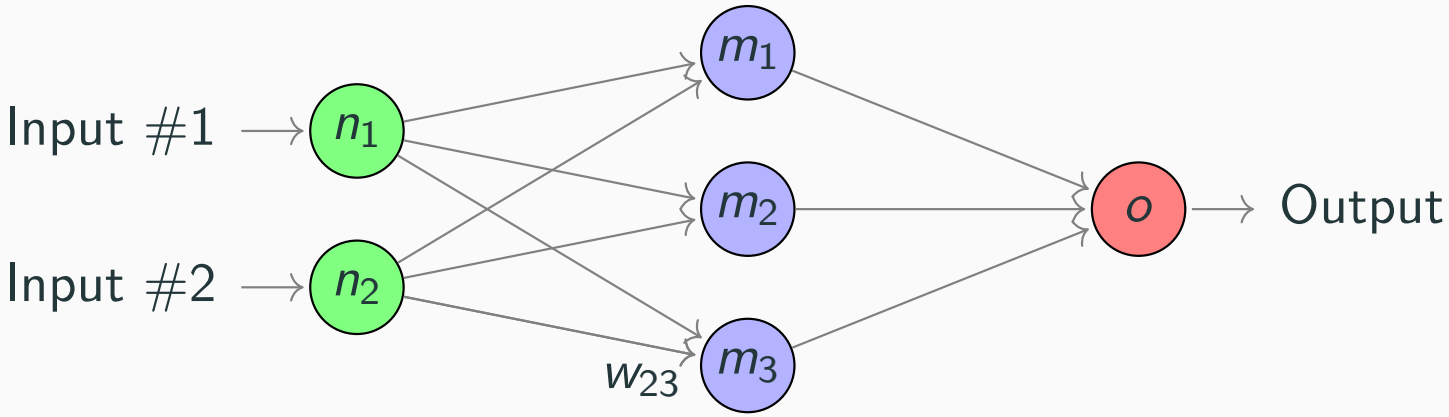
Towards Modern Neurons



While the output was still binary, the weights allowed the connections to be evaluated non-symmetrically. In addition to the weights, these early neural networks incorporated a **Bayesian** firing mechanisms. Both the weights and the firing probability were increased each time the pair of nodes successfully fired.

Towards Modern Neurons

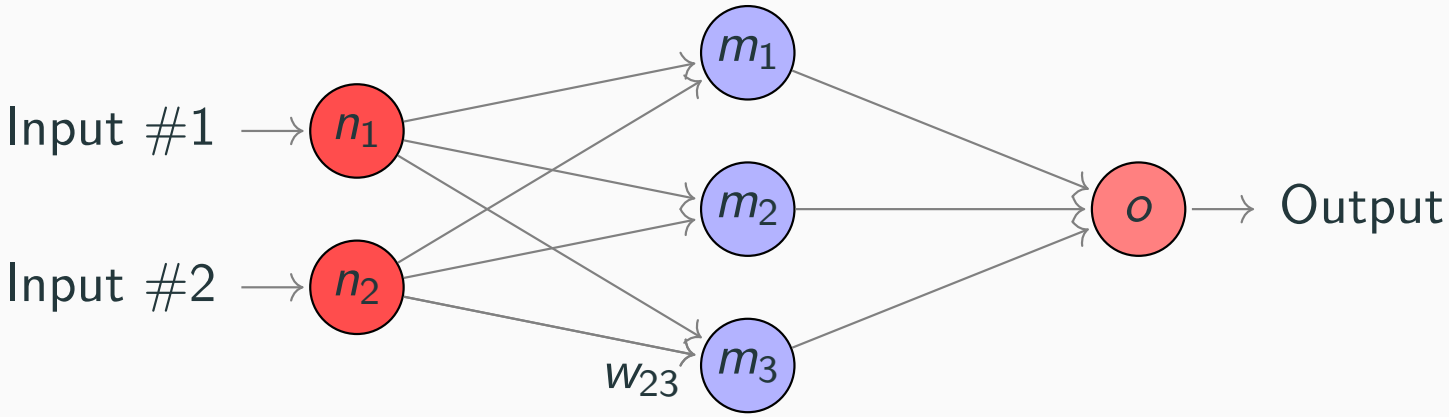
Simple Neural Network



While the output was still binary, the weights allowed the connections to be evaluated non-symmetrically. In addition to the weights, these early neural networks incorporated a **Bayesian** firing mechanisms. Both the weights and the firing probability were increased each time the pair of nodes successfully fired.

Towards Modern Neurons

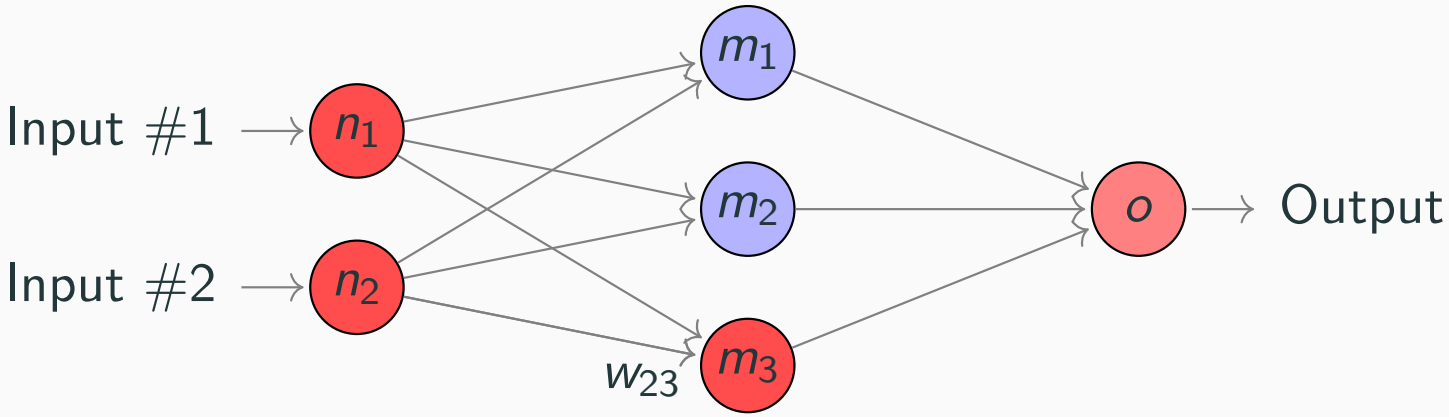
Simple Neural Network



While the output was still binary, the weights allowed the connections to be evaluated non-symmetrically. In addition to the weights, these early neural networks incorporated a **Bayesian** firing mechanisms. Both the weights and the firing probability were increased each time the pair of nodes successfully fired.

Towards Modern Neurons

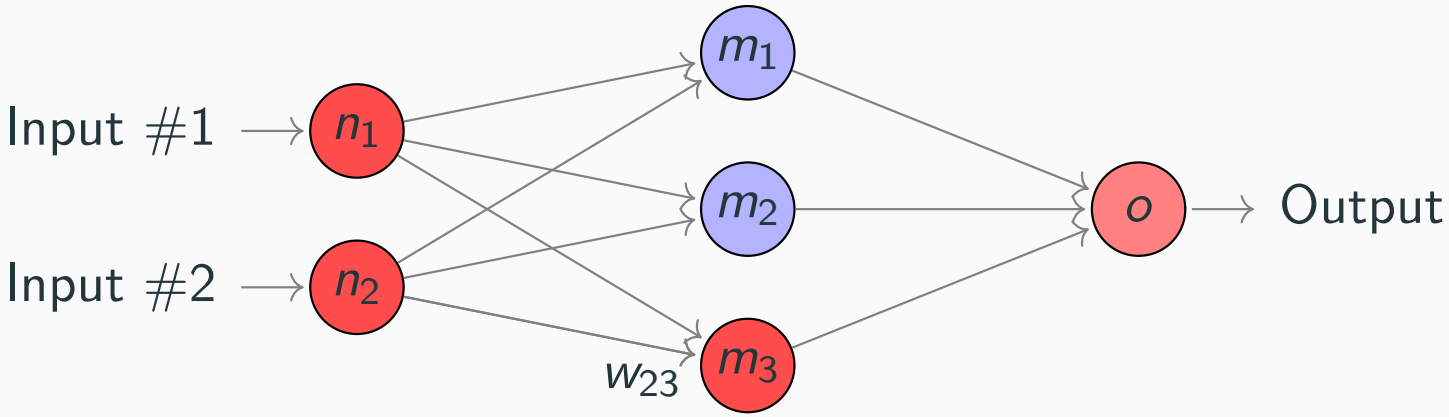
Simple Neural Network



While the output was still binary, the weights allowed the connections to be evaluated non-symmetrically. In addition to the weights, these early neural networks incorporated a **Bayesian** firing mechanisms. Both the weights and the firing probability were increased each time the pair of nodes successfully fired.

Towards Modern Neurons

Simple Neural Network

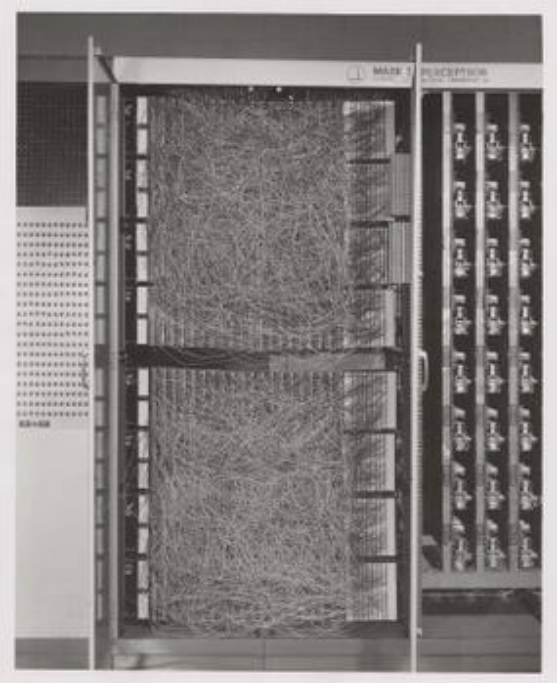


$n_2 \uparrow$, $w_{23} \uparrow$, $m_3 \uparrow$

Here, n_i and m_j represent the firing probabilities. In addition, n_1 and w_{13} would increase in this model.

Linear Classifier, Neural Networks and the Perceptron

Towards Modern Neurons

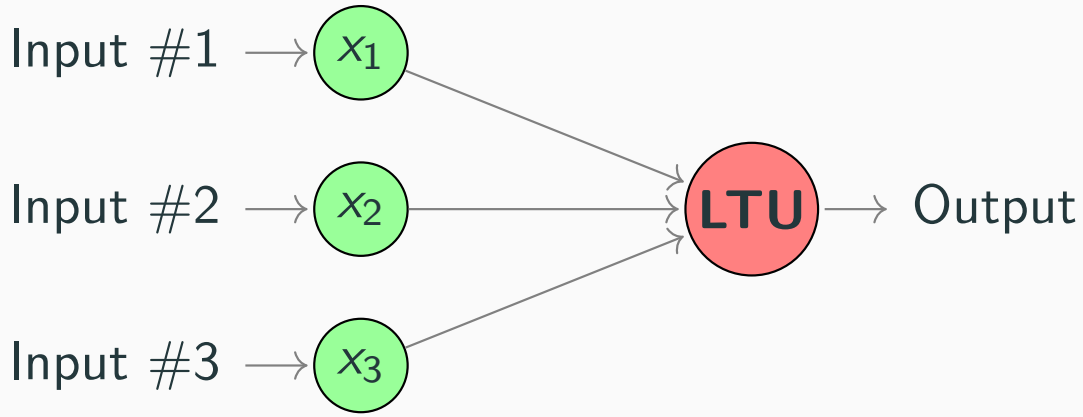


In 1957 Frank Rosenblatt introduced a linear algorithm into the ANN literature called the Perceptron. The perceptron actually entered the world as a hardware device, not as a software algorithms.

In 1957 Frank Rosenblatt introduced a linear algorithm into the ANN literature called the Perceptron. The perceptron is based around a **linear threshold unit** (**LTU**).

Perceptron

Simple Neural Network



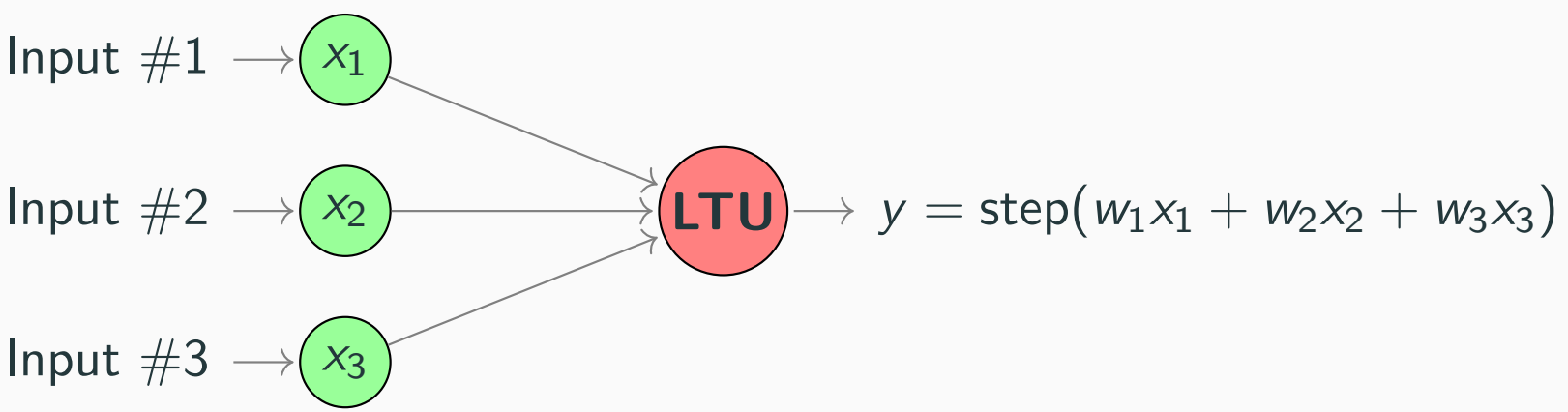
The LTU sums the inputs and then applies a step function:

$$\textbf{LTU} : y = \text{step}(w_1 x_1 + w_2 x_2 + \ldots + w_p x_p)\,,$$

where $\text{step}(x)$ is the Heaviside step function.

$$\text{step}(x) = \begin{cases} 0 & \text{if } x < 0\,, \\ 1 & \text{if } x \geq 0\,. \end{cases}$$

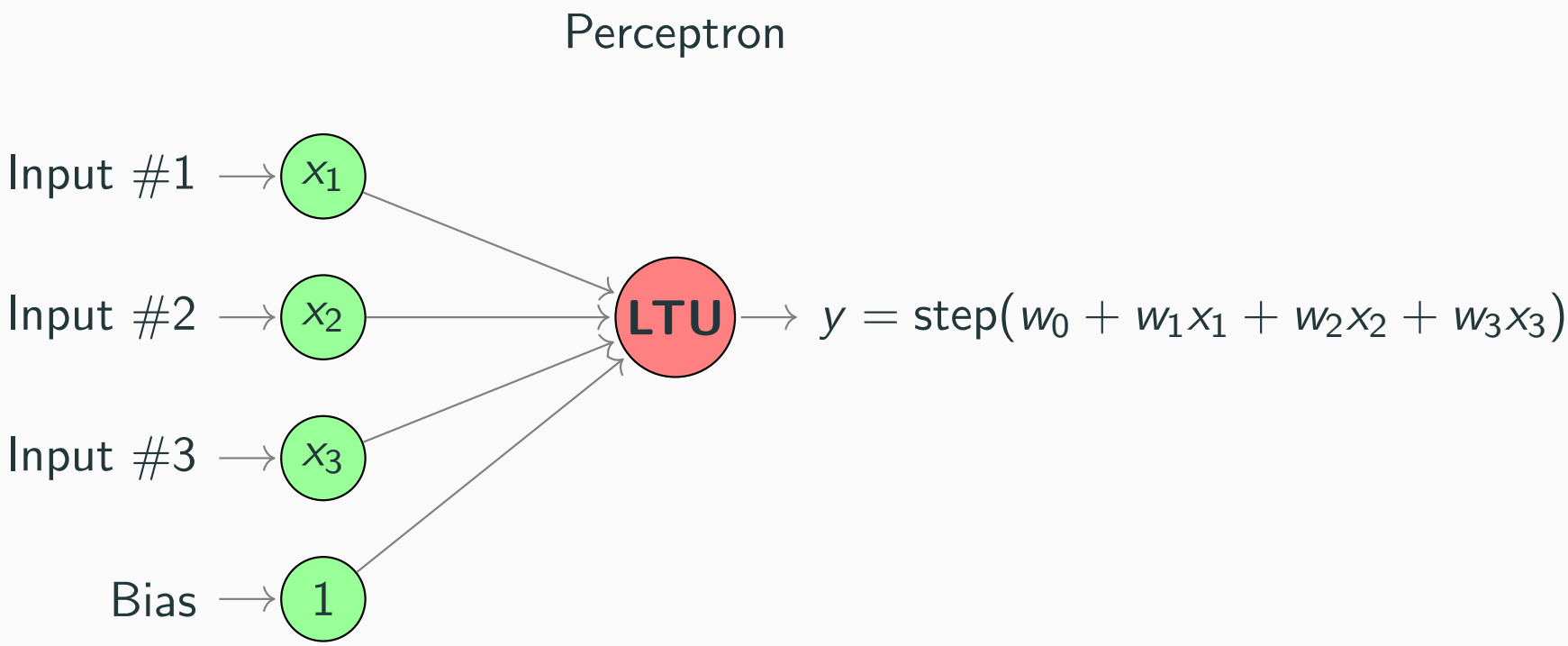
Perceptron



A single LTU gives a binary classifier using the sum of linear functions. Notice that functionally this give the same hypothesis class as (non-affine) linear regression

$$y = \text{step}(x_1 \beta_1 + \ldots + x_p \beta_p)$$

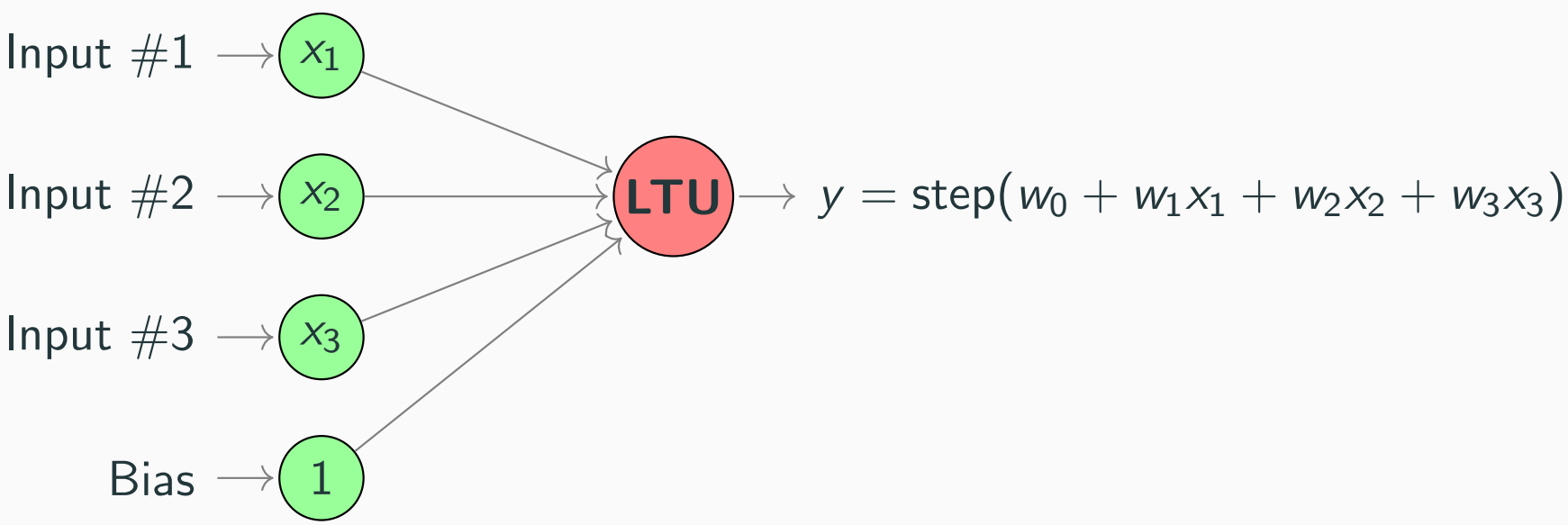
Perceptron



Finally, we can add a constant bias neuron to produce the affine terms w_0. We see that our perceptron can be used to perform categorical linear regression for binary classification.

$$y = \text{step}(\beta_0 + x_1\beta_1 + \ldots + x_p\beta_p)\,.$$

Perceptron



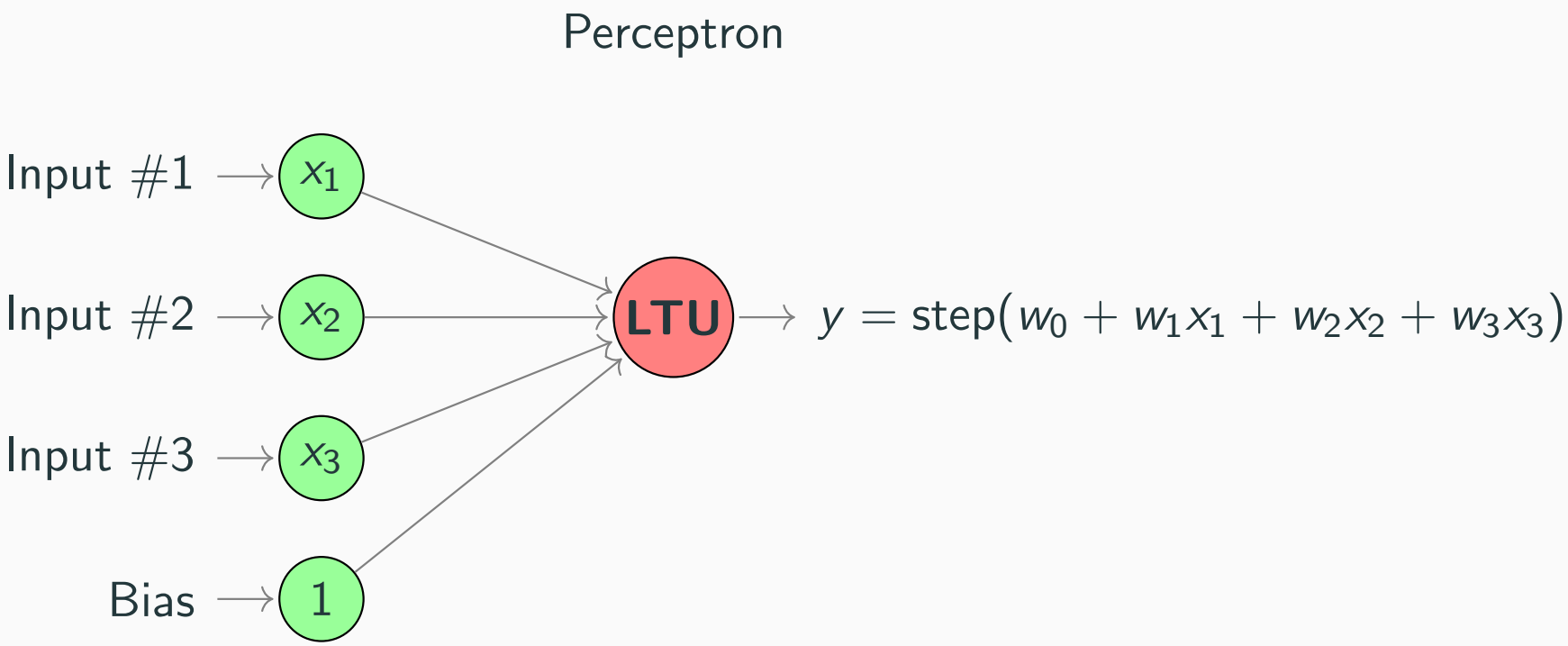
In a modern neural network, we follow the following terminology:

Nodes: Data holding variables.

Input Nodes: A set of p nodes in which we feed the x_i, either for training or prediction. Here, p is the number of features for each input vector.

Bias Node: A constant node.

Perceptron

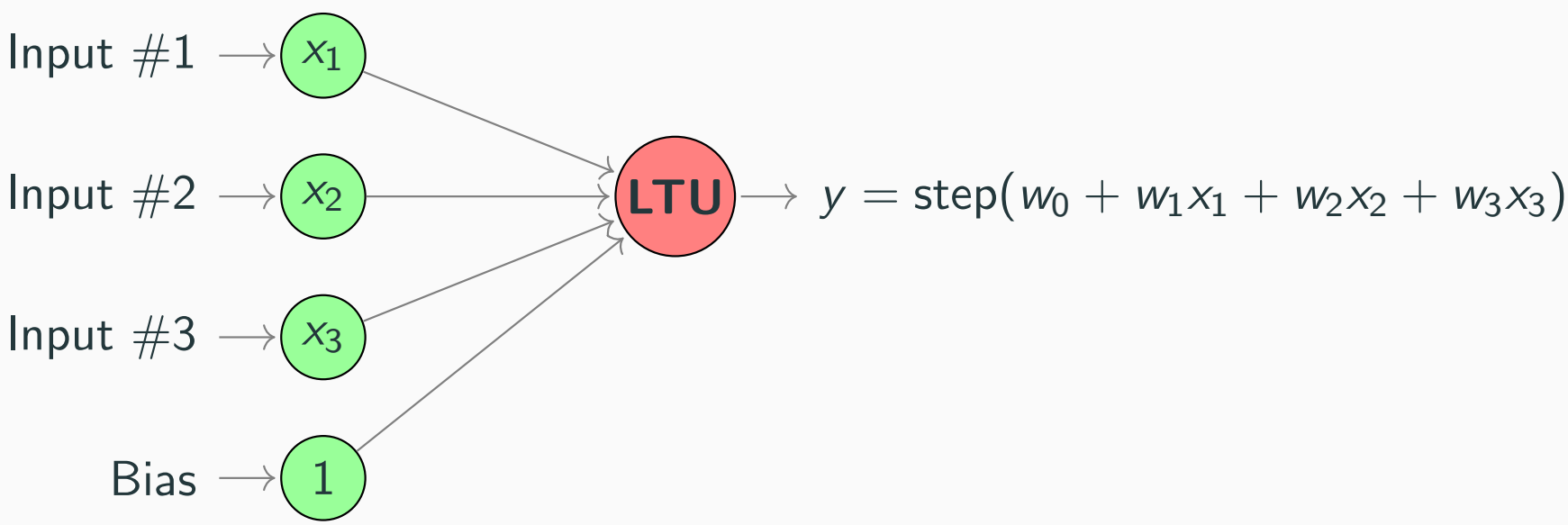


In a modern neural network, we follow the following terminology:

Edges: Each edge connects two nodes and carries a **weight** w. A weight is a trainable constant parameter of the network.

Output Node: Output nodes each hold a single dimension of a K dimensional output Y.

Perceptron



In a modern neural network, we follow the following terminology:

LTU: An LTU node takes and holds the sum over all inputing nodes, weighted by the edge weighting. It then compose that linear combination with an **activation function** to restrict the value range. For the perceptron, the activation function is **step**.

Perception

Training the Perceptron: Write the perceptron function as

$$w_0 + w_1 X_1 + \ldots + w_p X_p = w^T X = \langle w, X \rangle \,.$$

Initially, set all $w^{(0)} = 0$. At each step the, perception randomly samples a point x_i and predicts a label $\hat{y}_i$ based on $w^{(j)}$. We then update to $w^{(j+1)}$ by computing

$$w^{(j+1)} = w^{(j)} + (y_i - \hat{y}_i) x_i \,.$$

If $\hat{y}_i = y_i$, the point is correctly classified and we move on. But if $y_i \neq \hat{y}_i$, then

$$(y_i - \hat{y}_i) = \begin{cases} 1 & \text{if } \hat{y}_i = 0 \,, \\ -1 & \text{if } \hat{y}_i = 1 \,. \end{cases}$$

Perception

We want to see that the new weights

$$w^{(j+1)} = w^{(j)} + (y_i - \hat{y}_i)x_i \,.$$

classify y_i better than the old weights. Notice that

$$\hat{y} = \begin{cases} 0 & \text{if } \langle w, x_i \rangle < 0 \,, \\ 1 & \text{if } \langle w, x_i \rangle > 0 \,. \end{cases}$$

So this means that if $y_i \neq \hat{y}_i$ then

$$(y - \hat{y})\langle w^{(j)}, x_i \rangle < 0 \,.$$

So improving our guess for y_i means increasing $\langle w, x_i \rangle$ if $y_i = 1$, or decreasing it if $y_i = 0$. Either way, we will have been successful if

$$(y_i - \hat{y}_i)\langle w^{(j+1)}, x_i \rangle > (y_i - \hat{y}_i)\langle w^{(j)}, x_i \rangle$$

Perception

It not too hard to see that this will be the case:

$$\begin{aligned}(y_i - \hat{y}_i)\langle w^{(j+1)}, x_i\rangle &= (y - \hat{y})\langle w^{(j)} + (y_i - \hat{y}_i)x_i, x_i\rangle \\ &= (y - \hat{y})w^{(j)} + ||x_i||^2 \\ &> (y - \hat{y})w^{(j)}\,.\end{aligned}$$

The new halfplane "more correctly" classifies x_i. The algorithm stops when all points are classified.

`https://www.youtube.com/watch?v=xpJHhHwR4DQ`

Perception

Theorem: Assume that (x_i, y_i) are separable. Let $R = \max_i ||x_i||$ be roughly the data size and

$$B = \min\{||\beta|| \ : \ y_i \langle x_i, \beta \rangle \geq 1 \, \forall i\}$$

be roughly the gap distance. Then the perceptron algorithm stops after at most $(RB)^2$ iterations.

A couple of points about the perceptron:

- When the data is separable, there are many solutions and which one is found depends on the starting value.
- The finite number of steps can be large, practically, if the gap is small the time to find it is large.
- When the data are not seperable, the algorithm does not converge, and instead falls into a cycle.

Perception

The perceptron is a form of stochastic gradient decent on the loss function

$$\ell(w) = -\sum_{i=1}^{N}(y_i - \hat{y}_i)(x_i^T w)\,.$$

Notice that this loss is non-negative, and only 0 when all labels $\hat{y}$ are correctly fitted.The gradient is

$$\frac{\partial \ell}{\partial w} = -\sum_{i=1}^{N}(y_i - \hat{y}_i)x_i\,.$$

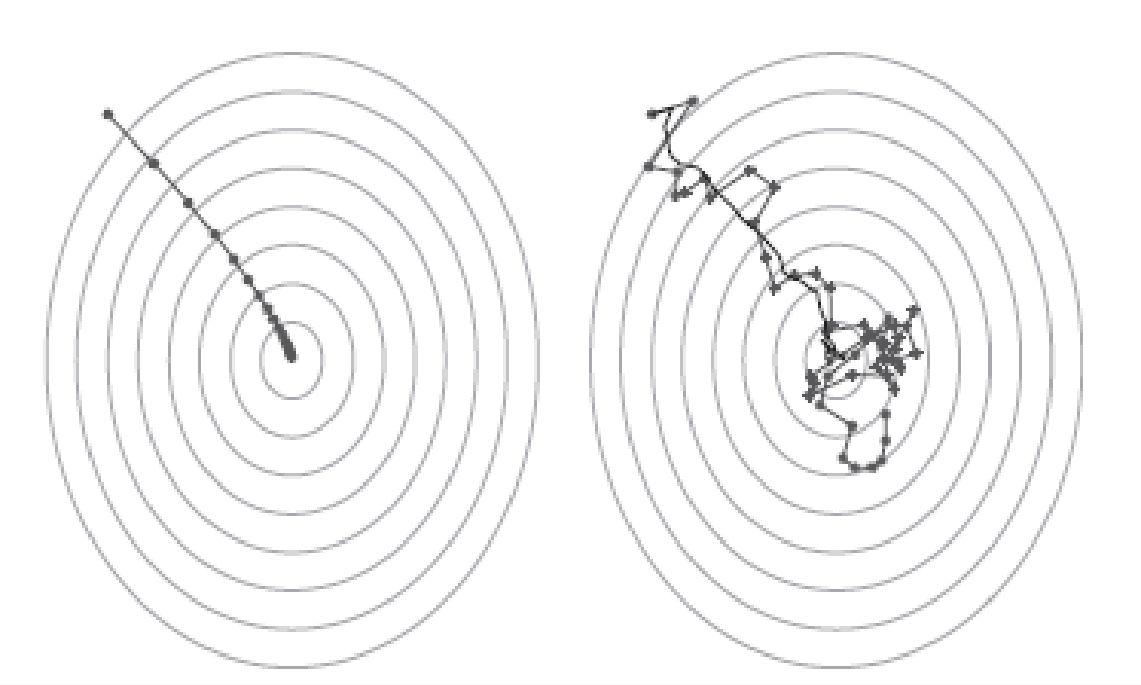
In gradient decent, we would update w and using the full gradient,

$$w = w + \eta\sum_{i=1}^{N}(y_i - \hat{y}_i)x_i\,,$$

but here we are making the SGD update on the single datapoint (x_i, y_i):

$$w = w + (y_i - \hat{y}_i)x_i\,.$$

Perception



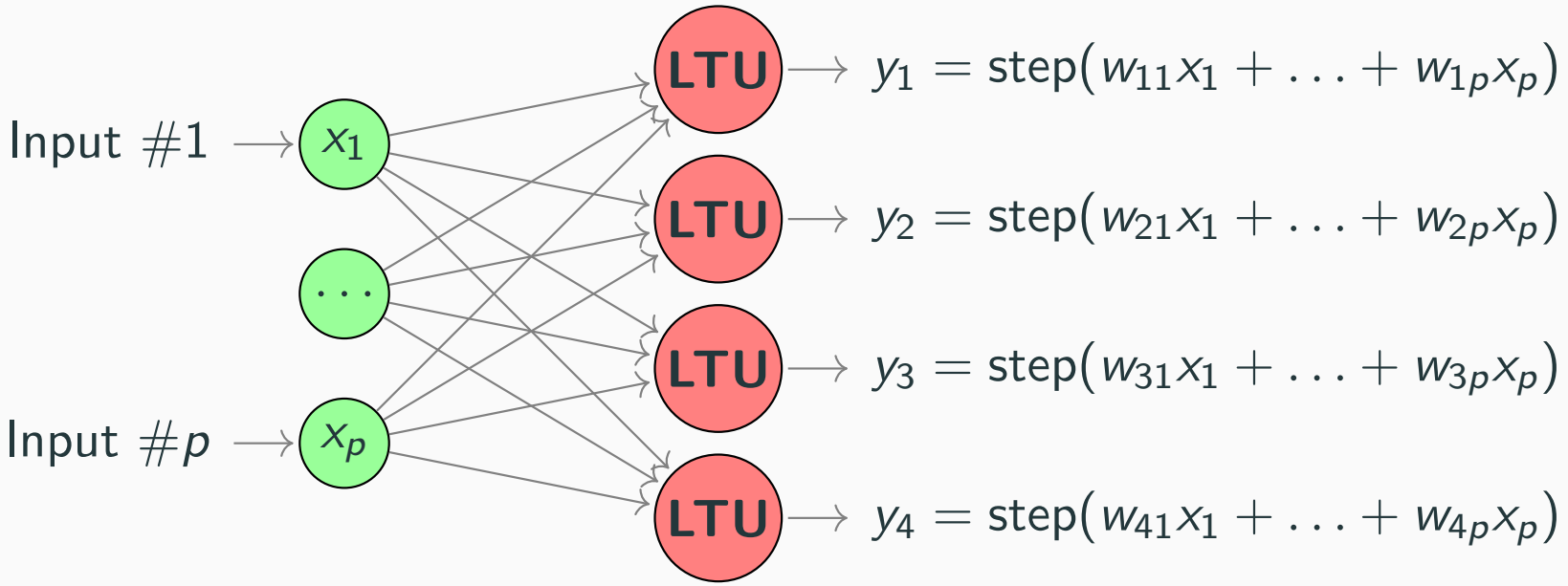
Completing the conversion to Stochasitic gradient decent, we add in a learning parameter η:

$$w = w + (y_i - \hat{y}_i)x_i \, .$$

Multilabel Perceptrons

Perceptron

Multilabel Perceptron



A single LTU gives a binary classifier using the sum of linear functions. For multilable classification we add a LTU nodes for each label.

Notice that we are not post-composing with argmax so it possible that multiple (or none) of the outputs could be 1.

Training the Perceptron

The perceptron training algorithm proposed by Rosenblatt followed Hebb's rule of organization, that is that when one neuron triggers another the connect between then is strengthened.

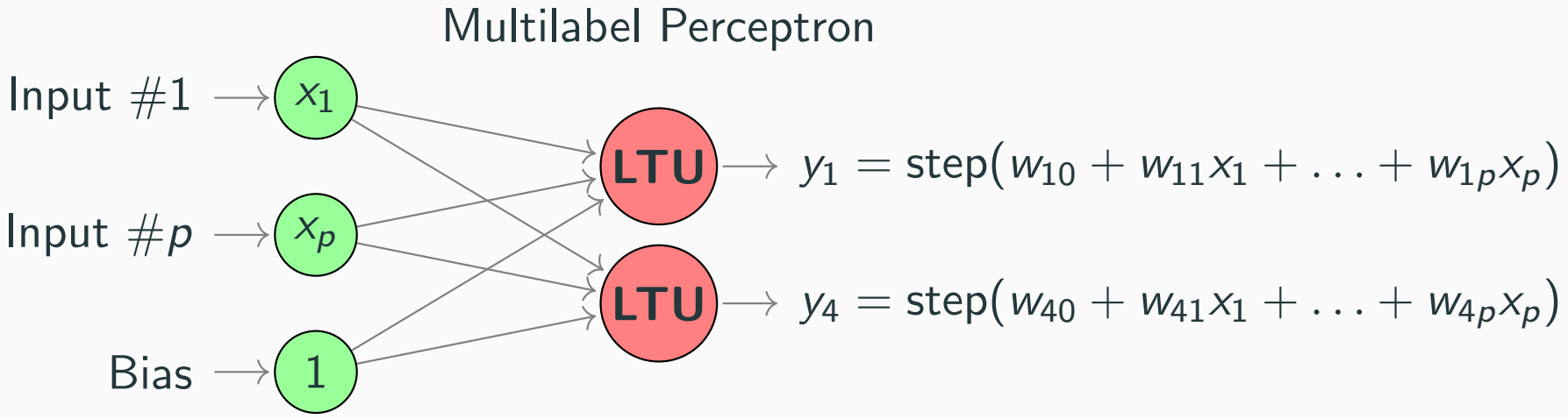
Rosenblatt also wanted to include feedback into the training, only strengthening connections the lead to correct predictions.

These considerations are what lead to the learning rule

$$w_{ij}^{(n+1)} = w_{ij} + \eta(y_j - \hat{y}_j)x_i \,,$$

with learning rate η, a single training instance (x, y) and the predicted label $\hat{y}_i$. Lets now look at how this works in practice.

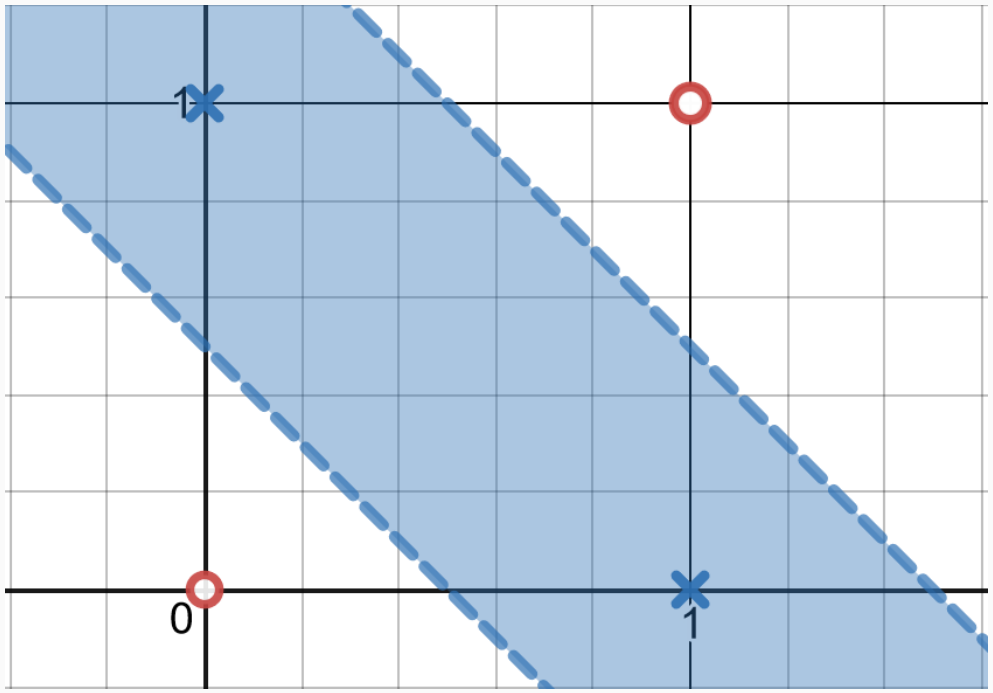
Perceptron



Learning Rule: $w_{ij}^{(n+1)} = w_{ij} + \eta(y_j - \hat{y}_j)x_i\,.$

Pass a single training instance x through the network:
If the prediction on label i is correct $\hat{y}_i = y_i$ and we do not update the weights w_{ij}.
If the prediction on label i is incorrect, $\hat{y}_i \neq y_i$. So if the true labeling is $y_i = 1$, add ηx_i to all weights w_{ij}.If $y_i = 0$, subtract ηx_i from all weights w_{ij}.

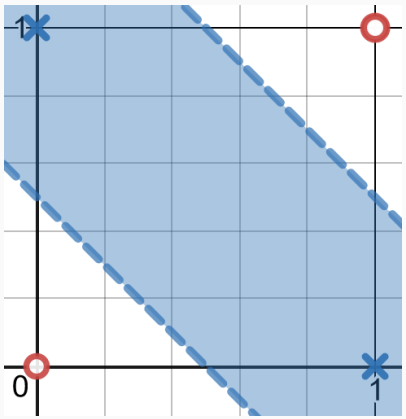
Perceptron



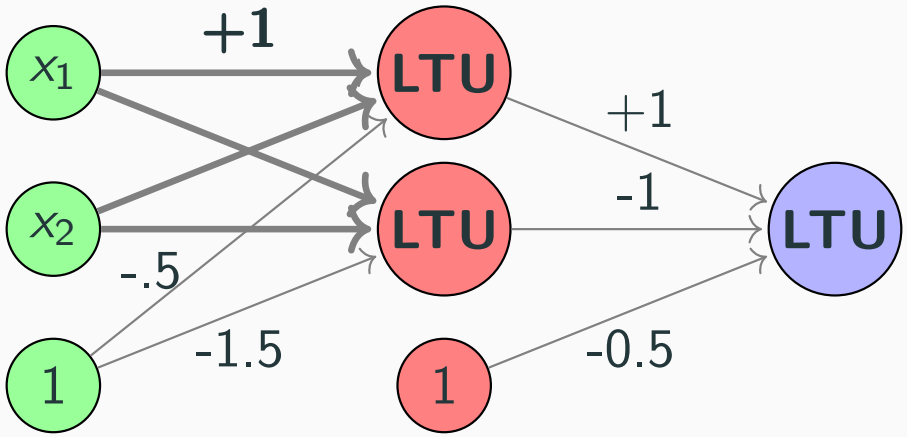
The perceptron itself is one of the first examples of SGD, but it has some serious problems. In particular, it cannot solve some trivial problems like the xor labeling above.

Although now unsurprising (no linear classifier can solve xor) the exceptions for the perceptron were high and when this problem was uncovered in 1969 it lead most researchers to abandon neural networks in favor of functional and logical methods.

Perceptron



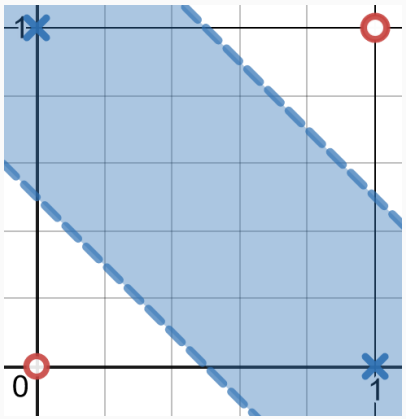
Multilayer Perceptron



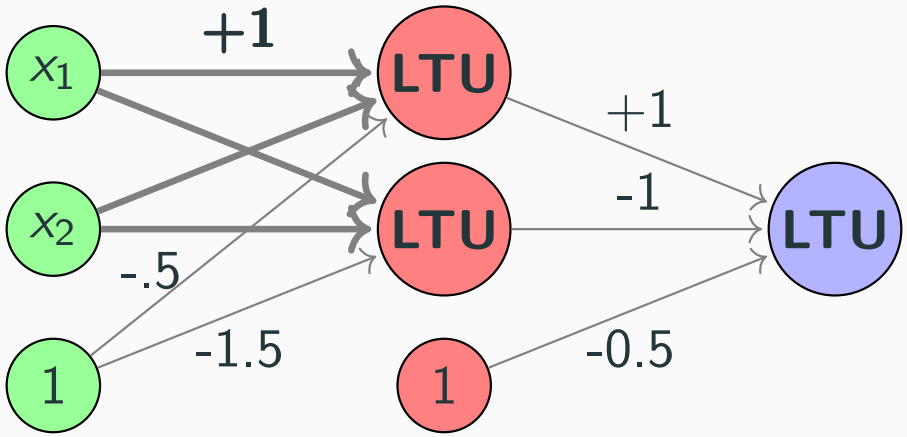
However it was premature to abandon the perceptron: it turns out that if you stack perceptrons you can fit the xor distribution. Consider sending the point $(0, 0)$ through:

$$(0,0) \rightarrow \text{step}(x_1 + x_2 - 0.5, x_1 + x_2 - 1.5) = (0,0)$$
$$\rightarrow \text{step}[(0) - (0) + (-0.5)] = 0$$

Perceptron



Multilayer Perceptron

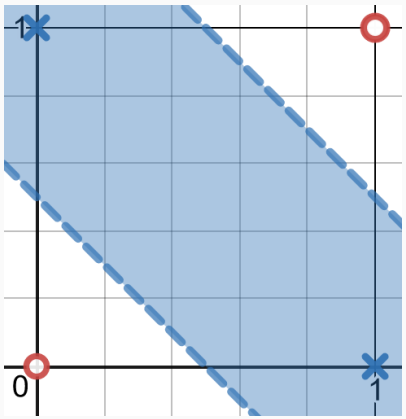


Sending (1,1) through, we get

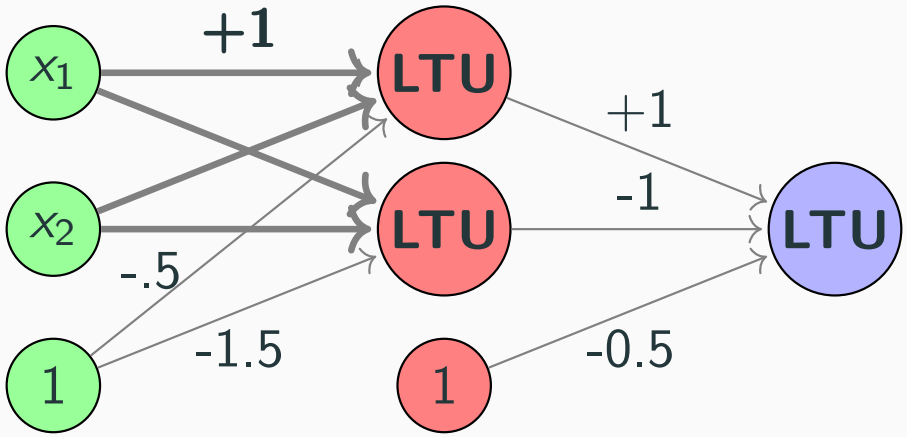
$$(1,1) \rightarrow \text{step}(1+1-0.5, 1+1-1.5) = (1,1)$$
$$\rightarrow \text{step}[(1)-(1)+(-0.5)] = 0$$

(**Exercise:**) Verify that (0,1) and (1,0) are labeled 1.

Perceptron



Multilayer Perceptron



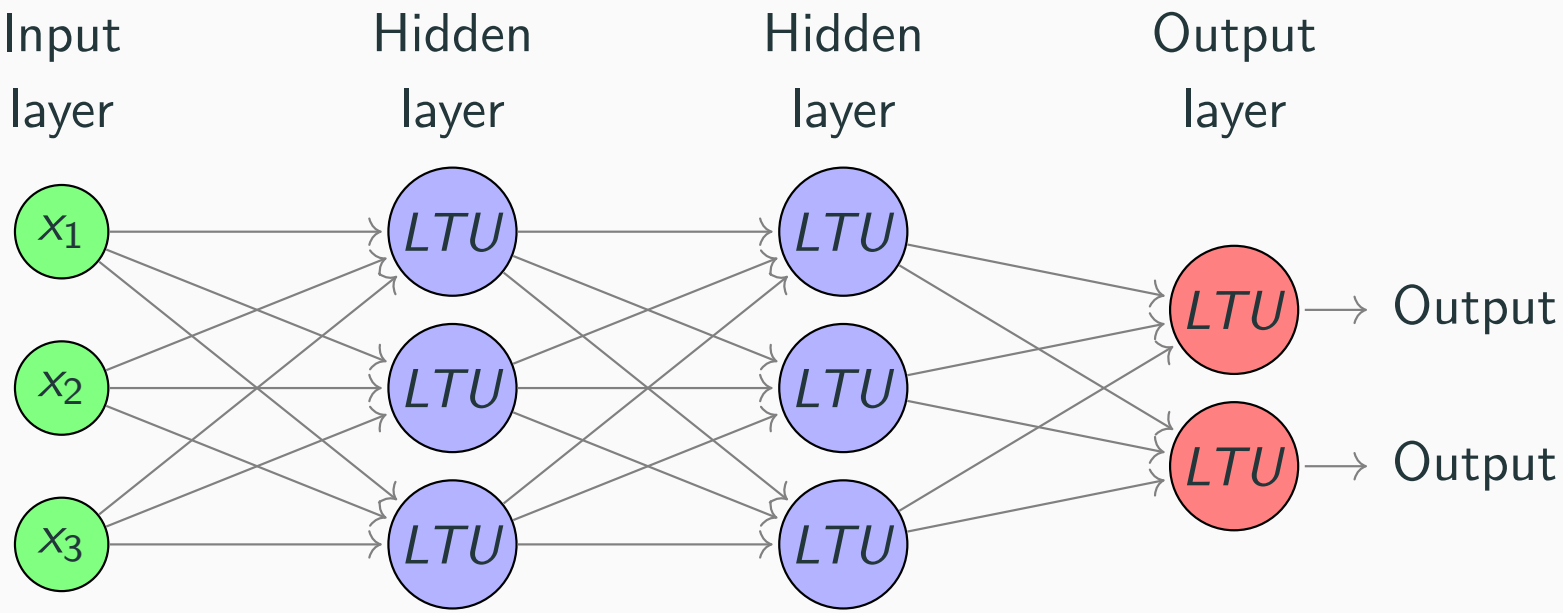
With a little bit of work, one can show that this multilayer perceptron actually is the fitting on the left:

$$\hat{h}(x_1, x_2) = \begin{cases} 1 & 0.5 - x_1 \leq x_2 \leq 1.5 - x_1 \\ 0 & \text{otherwise} \end{cases}$$

Questions: What is the hypothesis class of all two layer two variable perceptrons? How would we construct a perceptron with three linear conditions?

Gradient Decent and Back Propagation

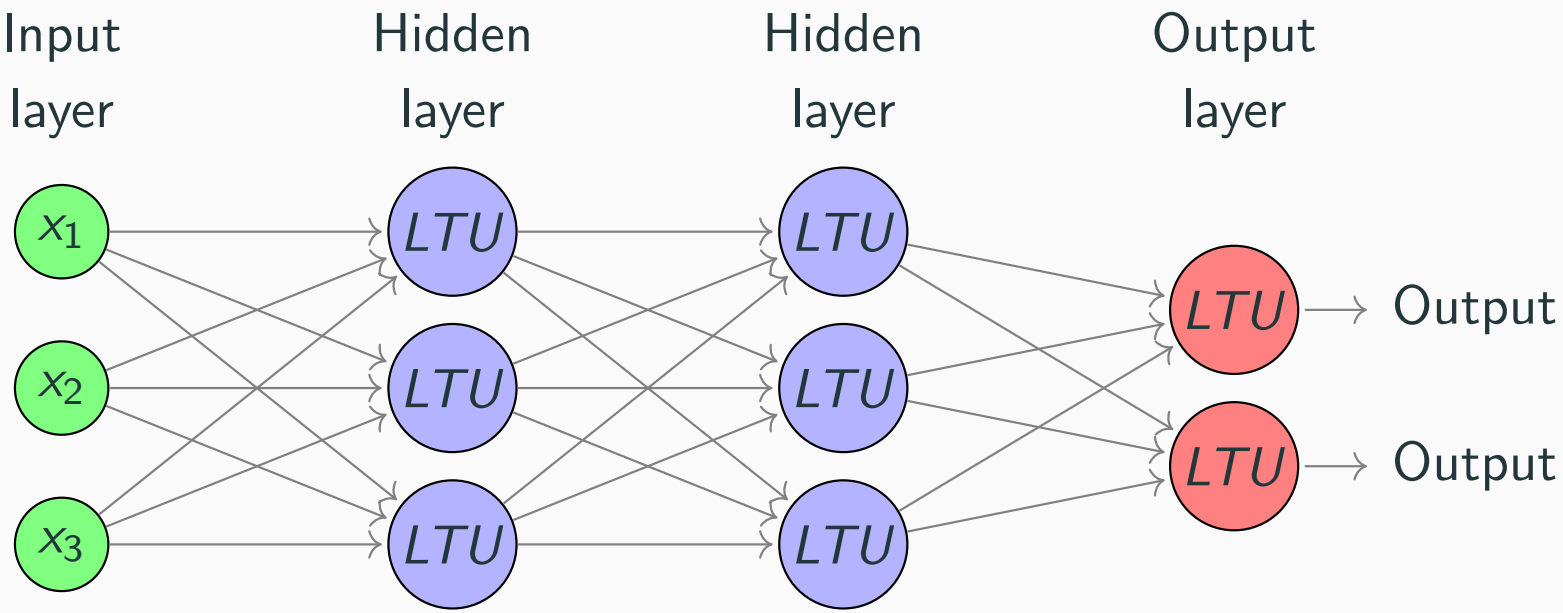
Training the Multilayer Perceptron



A multilayer perceptron (MLP) is composed of one input layer, multiple hidden layers and an output layer. If the network has more than one hidden layer it is called a **deep neural network**.

For a deep network, each layer represents a linear combination of the previous layers, with each output composed by an activation function.

Training the Multilayer Perceptron

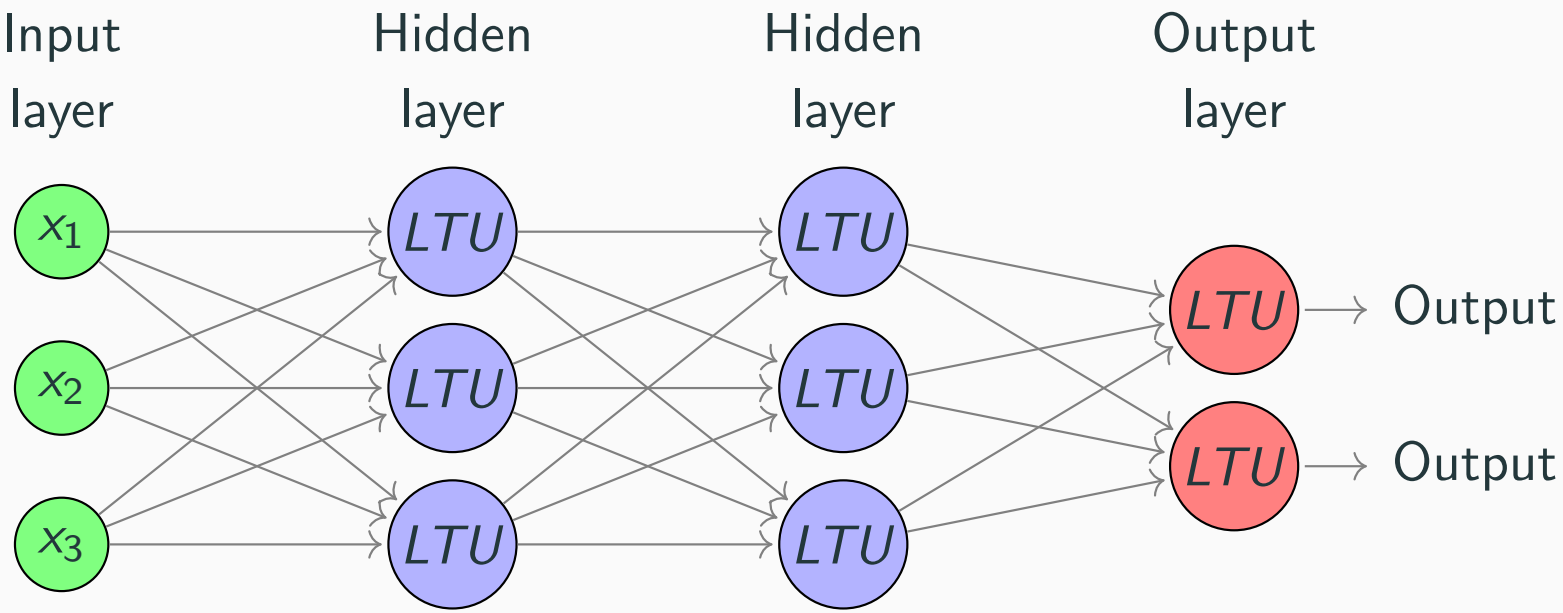


For the j'th hidden layer, let N^j be a column vector of n_j nodes N^j_k. Let σ be an activation function applied component-wise and let $W^j_{\ell k}$ be outgoing weights. The values of the nodes in the $j+1$'st layer will be

$$N^{j+1} = \sigma\left(W^j N^j\right).$$

Here, W^j is a $n_j \times n_{j+1}$ matrix.

Training the Multilayer Perceptron

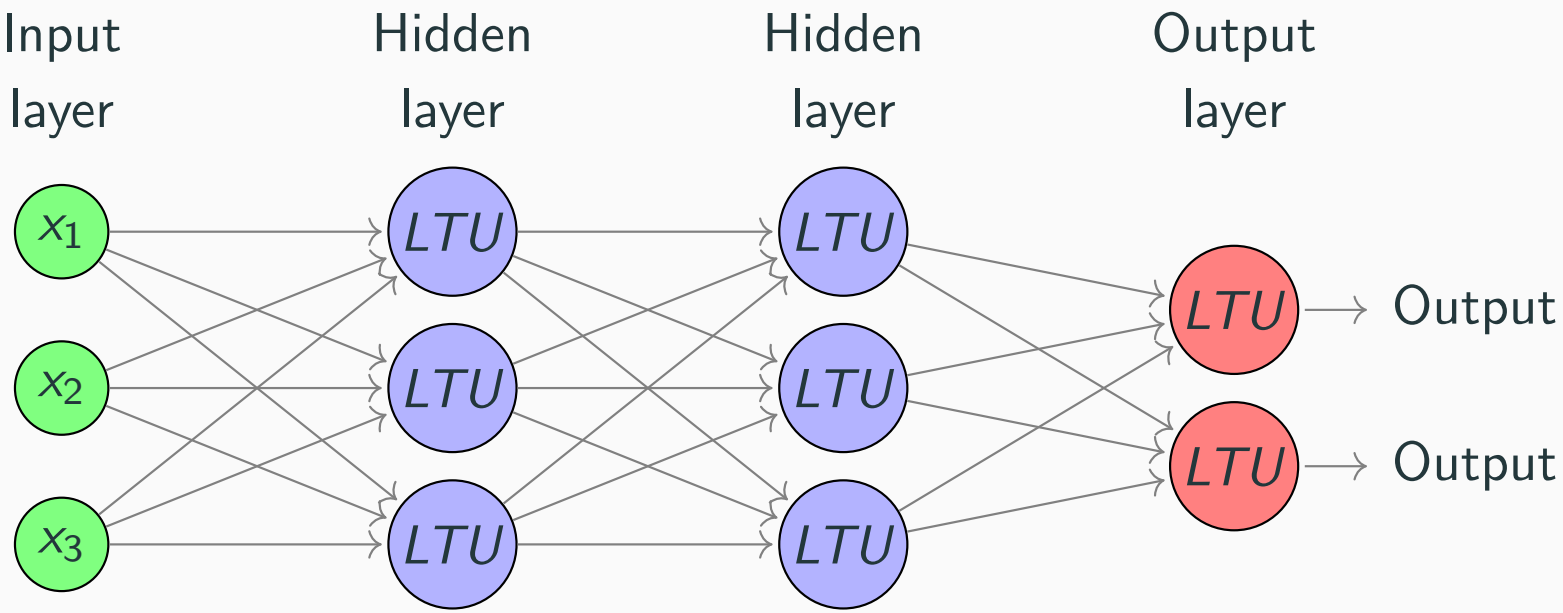


If the input is X and the output is Y, the whole L layer network is a function

$$F(X) = \sigma\left(W^{L-1}\sigma(W^{L-2}\ldots\sigma(W^0X)\ldots)\right).$$

Here, W^j is a $n_k \times n_{k+1}$ matrix.

Training the Multilayer Perceptron

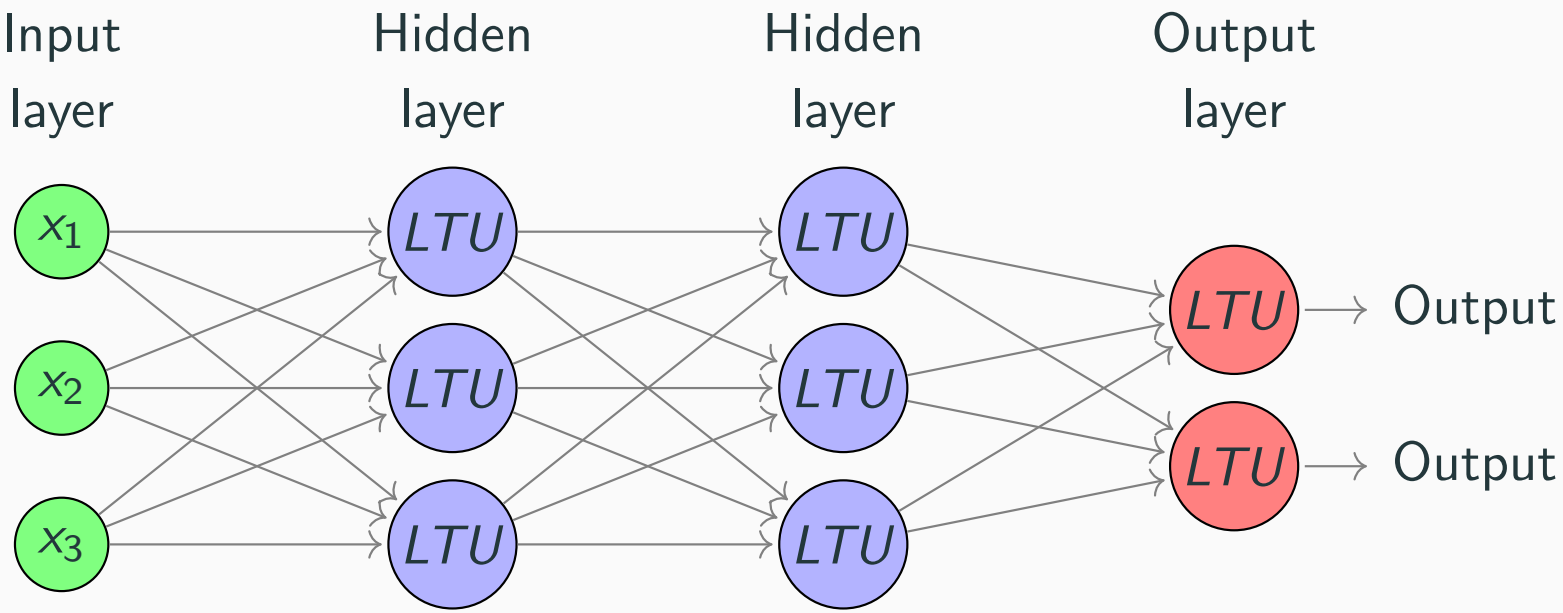


It's clear that the number of weights in a network becomes large *fast*:

$$\#\ \text{Weights:} \qquad |w| = (n_K \times nL - 1^2 \ldots, n_1^2 \times n_0\,.$$

Learning how to effectively train MLP took another 20 years.

Training the Multilayer Perceptron

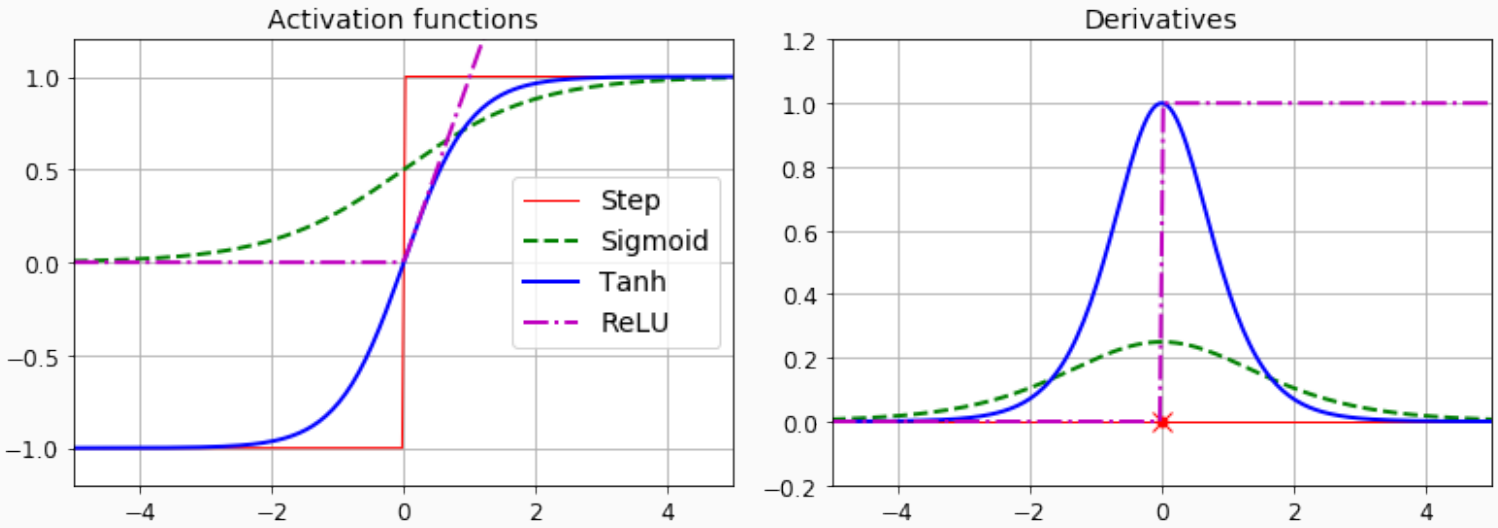


The problem with training wasn't just a large number of constants, but fact that the step function has 0 gradient everywhere, so there is no indication of which direction to move.

To compute the gradient the activation function for the LSU was replaced by the **logistic** or **sigmoid** function

$$\sigma(z) = \frac{1}{1 - e^{-x}}\,.$$

Implementing Gradient Decent



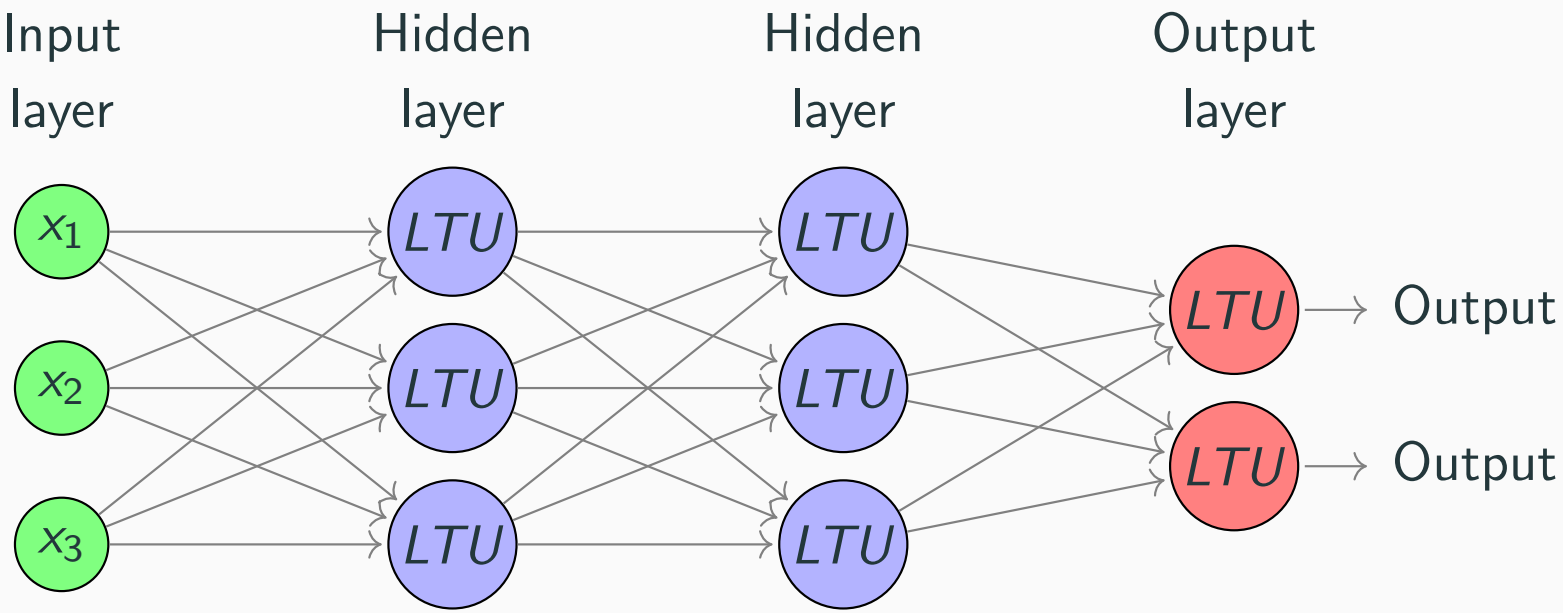
Other popular alternatives to the step function are

$$\tanh(x) = 2\sigma(2x) - 1$$

and

$$\text{ReLU}(x) = \max\{0, x\} = (x)_+$$

Training the Multilayer Perceptron



In 1986, D. E. Rumelhart popularized the idea of **back propagation** to estimate gradients. Although back propagation is the most widely used today, we will first canvas the common methods of gradient estimation.

Implementing Gradient Decent

There are several popular ways of implementing gradient decent. For a loss function like MSE, gradient decent updates the parameters at each step by

$$\beta^{(n+1)} = \beta^{(n)} - \eta \nabla_\beta \text{MSE}(\beta^{(n)}).$$

If there is a closed form formula for MSE, the update step can be computed explicitly. This is known as **manual differentiation** or **pen and paper differentiation**.

If the nodes are fairly out of the box, we can try to use automatic **symbolic differentiation** to pass through the graph and create a gradient. There are some decent solvers out there and this is an active area of development, but unfortunately it often yields huge redundant computations that allow errors to propagate. It is also must be implemented manually on arbitrary code.

Numerical Differential

Numerical differentiation is the simplest general solver. For a network $Y = h_\beta(X)$ with inputs X and outputs Y we implement partial differentiation by passing a data point (or points) through the network and seeing how a small change in β effects it

$$\frac{\partial h_\beta}{\partial \beta} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{h_{\beta+\epsilon}(X_i) - h_\beta(X_i)}{\epsilon}, \qquad \epsilon \ll 1 .$$

For a large network, this takes many calls to the function while also being numerically suspect. It is however a good first pass test and it is easy to implement.

Autodiff

Forward mode autodiff is a mixture of numerical and symbolic differentiation. It relies on **dual numbers** (**infinitesimals** in the math context), that is we write

$$a + b\epsilon$$

where ϵ is "small," ie $\epsilon^2 = 0$. In memory, dual numbers are stored as a pair of floats (a, b).

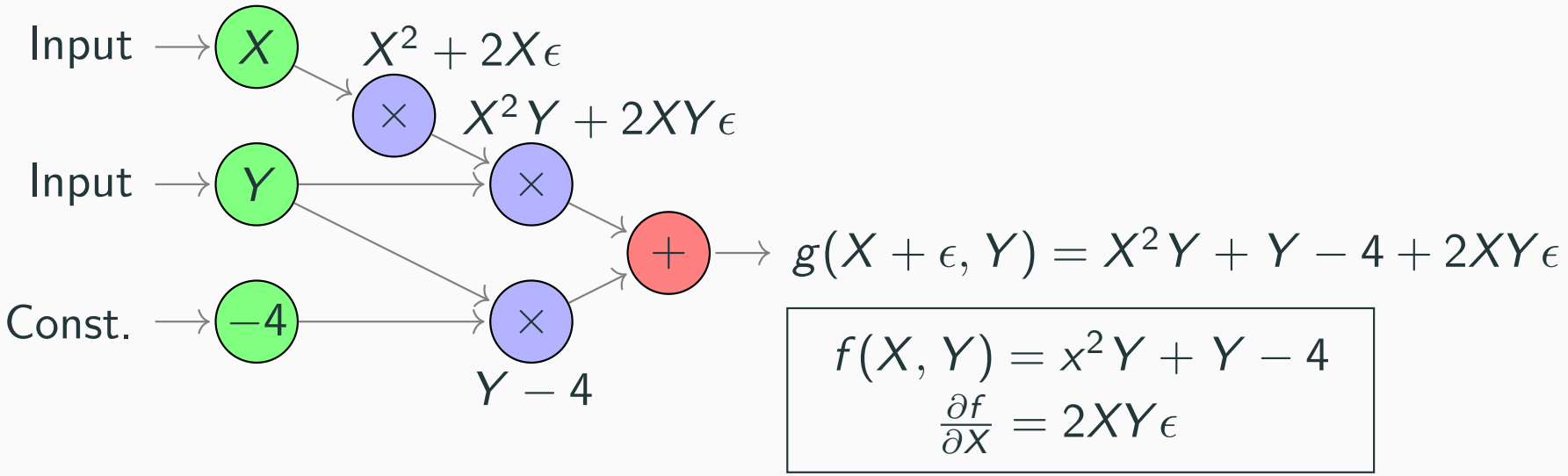
Dual numbers have slightly nonstandard multiplication and exponentiation that encode the rules of differentiation implicitly:

$$(a + b\epsilon)(c + d\epsilon) = ac + (bc + ad)\epsilon\,, \qquad e^{a+b\epsilon} = e^a(1 + b\epsilon)\,.$$

Both equations are derive by expanding $e^{b\epsilon}$ and using ϵ^2 and higher order terms to 0. In fact, if $f(x)$ is smooth than by Taylors theorem

$$f(a + b\epsilon) = f(a) + b\epsilon f'(a)\,.$$

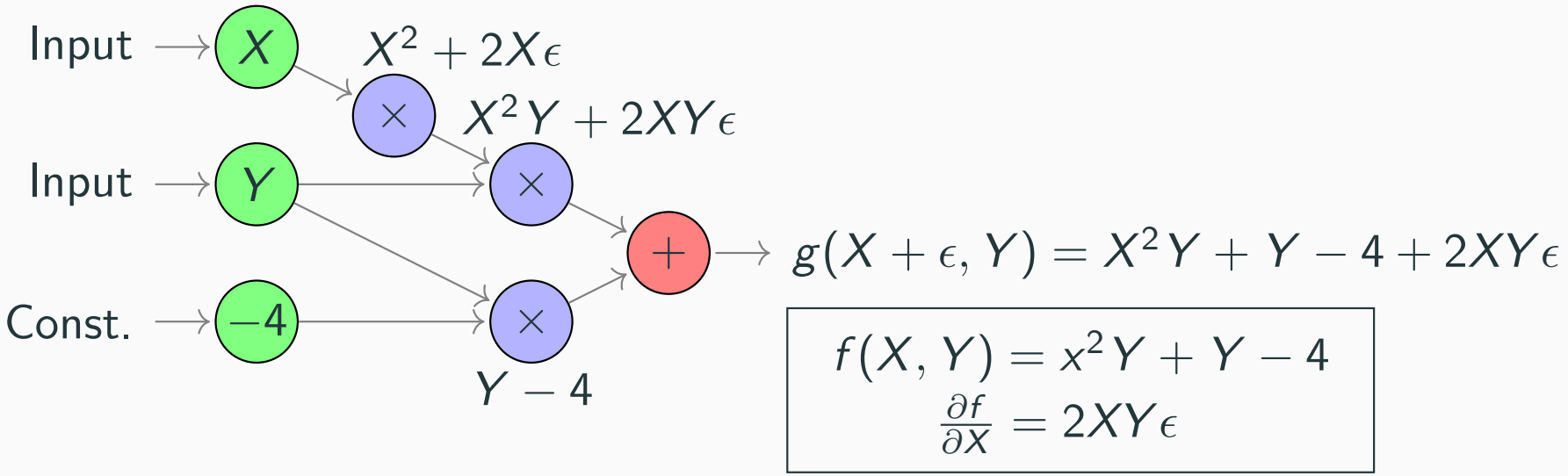
Forward Autodiff



Since $f(a + b\epsilon) = f(a) + b\epsilon f'(a)$ gives you both $f(a)$ and $f'(a)$ in one go, implementing forward autodiff comes down to creating a new graph that computes the result of passing the dual number through the graph.

Of course, we intend to differentiate with respect to the weights, but that is equivalent to treating the weights as dual numbers.

Reverse Autodiff



Forward autodiff is exact (unlike numerical approximation) but it still involves passing through the graph once for each weight β. For a graph with many connection weights this cause take a long time. **Reverse autodiff**, or **back propagation** on the other hand passes through the graph exactly twice.

Back Propagation

Reverse Autodiff

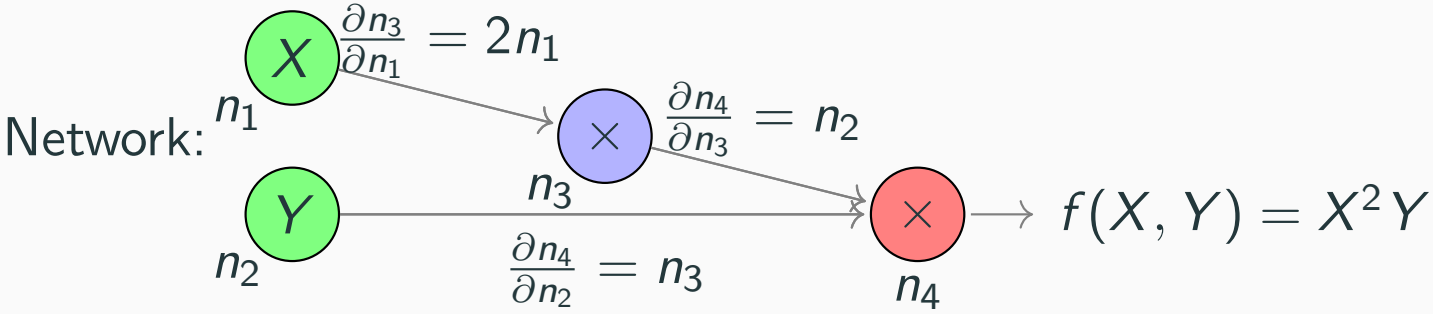
Reverse autodiff constructs a new graph containing the derivative of each input node with respect to each output node. It then uses the chain rule

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial n}\frac{\partial n}{\partial x}$$

to compute the derivative by filling out a dual graph one step at a time.

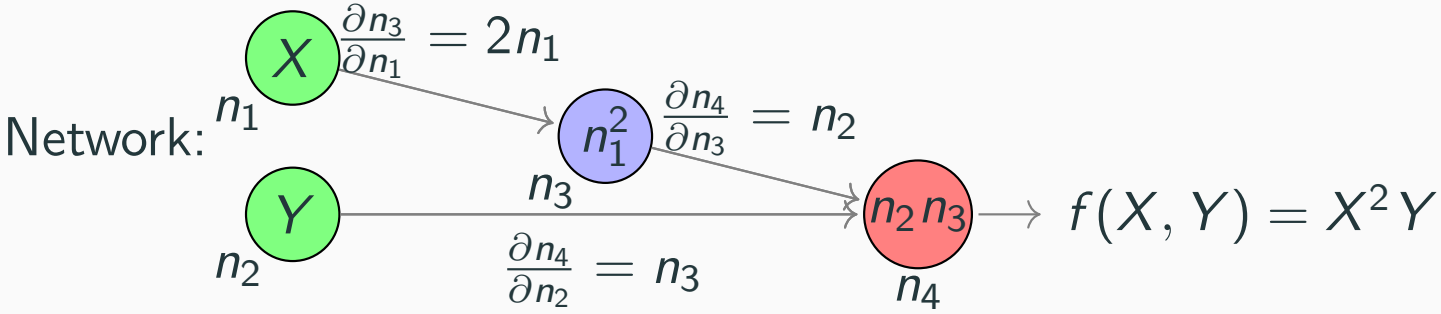
Back propagation uses reverse auto differentiation to compute the gradient for a fixed input-output pair (x_i, y_i). The speed increase comes from the fact that computing $\frac{\partial F}{\partial w}$ for each weight requires many redundant calculations. By keeping track of these, and computing the chain rule in a reverse layer by layer fashion, we are able to compute the gradient via an adjoint network graph.

Reverse Autodiff



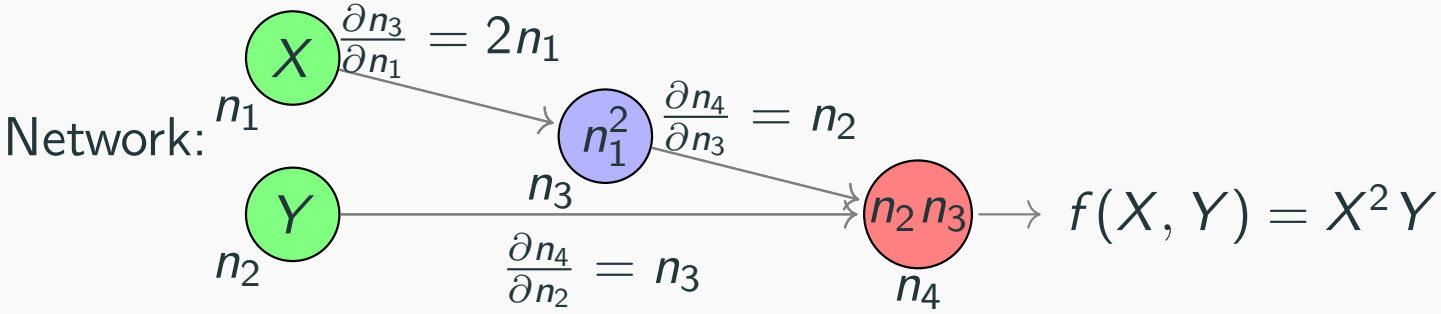
First, denote the value of each node by n_i.

Reverse Autodiff



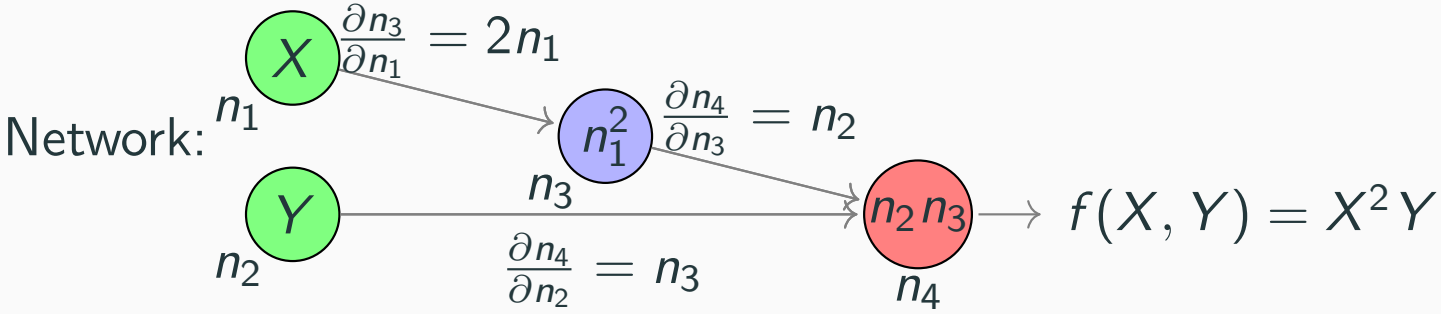
First, denote the value of each node by n_i. Second, propagating the node values through, compute the partial derivative at each connection. This can be done locally at each node upon constructing the graph.

Reverse Autodiff



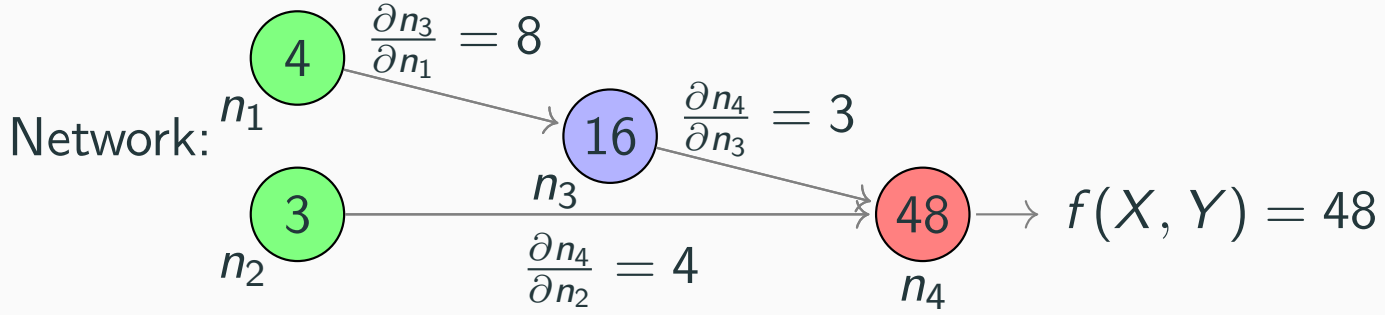
First, denote the value of each node by n_i. Second, propagating the node values through, compute the partial derivative at each connection. This can be done locally at each node upon constructing the graph. This constructs the gradient graph.

Reverse Autodiff



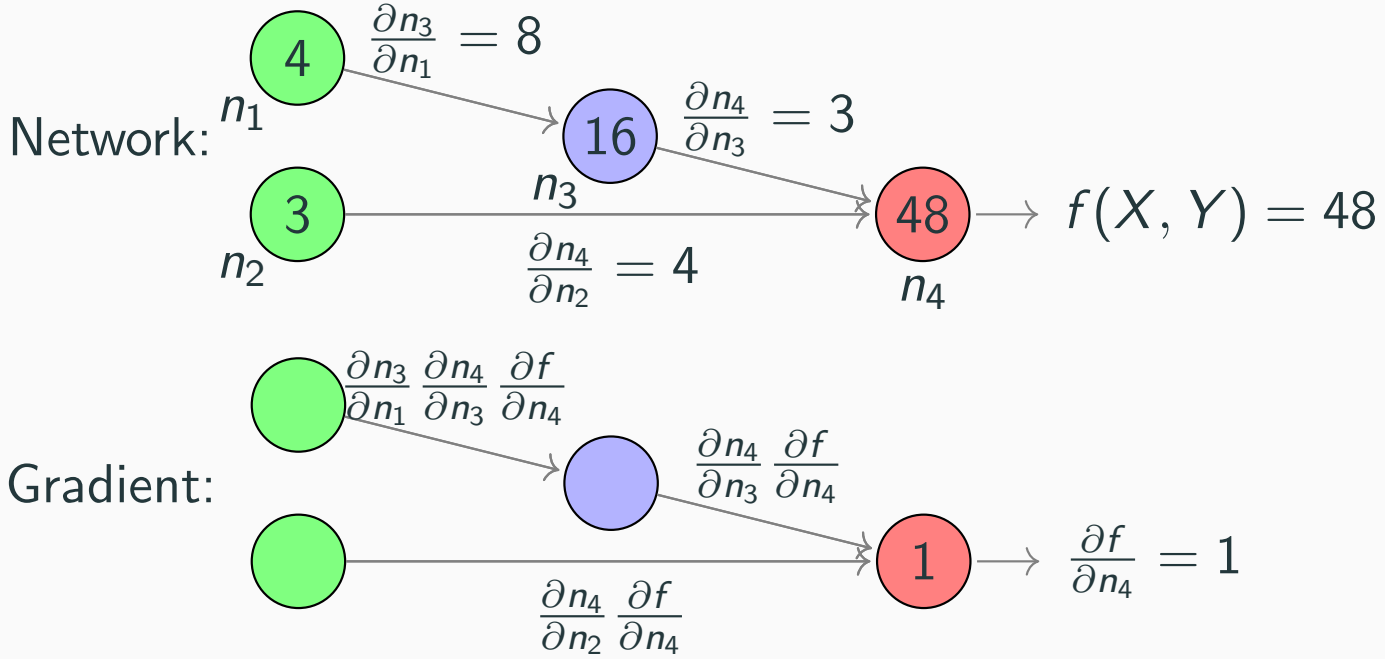
At runtime, for fixed (X, Y), say (4,3), we compute the gradient as follows. Pass (4,3) through the graph, filling all nodes.

Reverse Autodiff



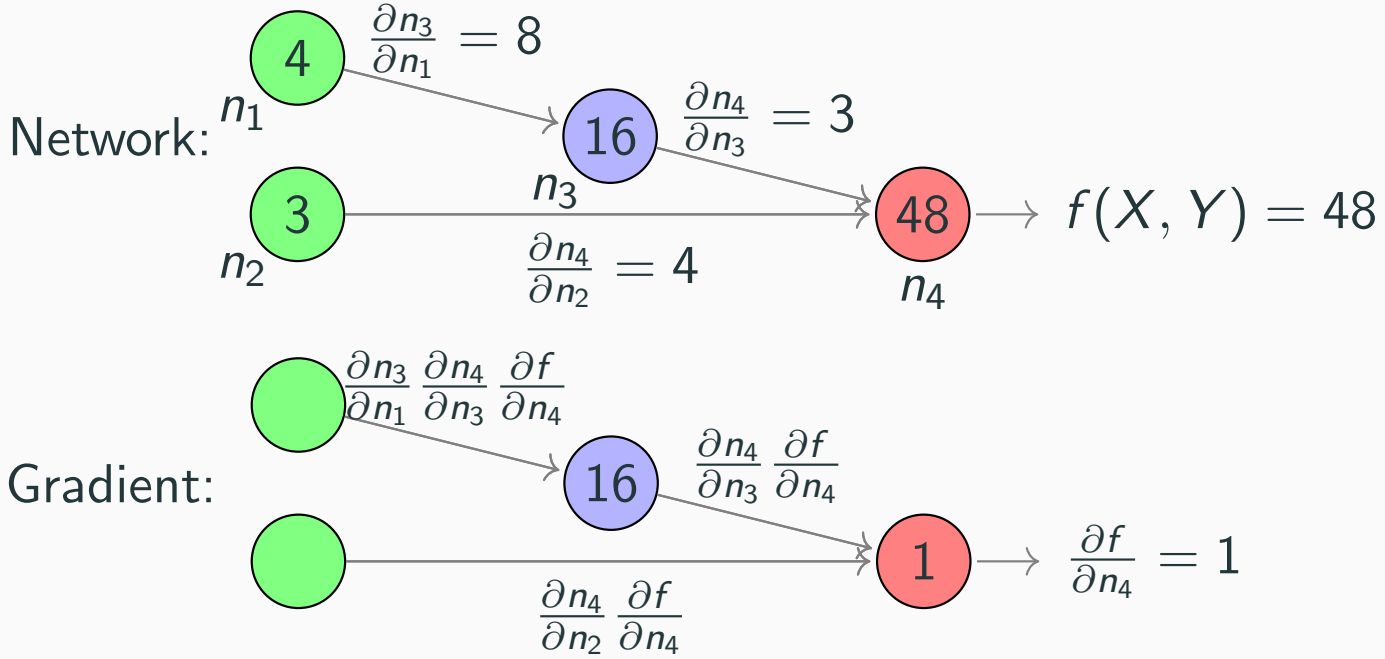
At runtime, for fixed (X, Y), say (4,3), we compute the gradient as follows. Pass (4,3) through the graph, filling all nodes.

Reverse Autodiff



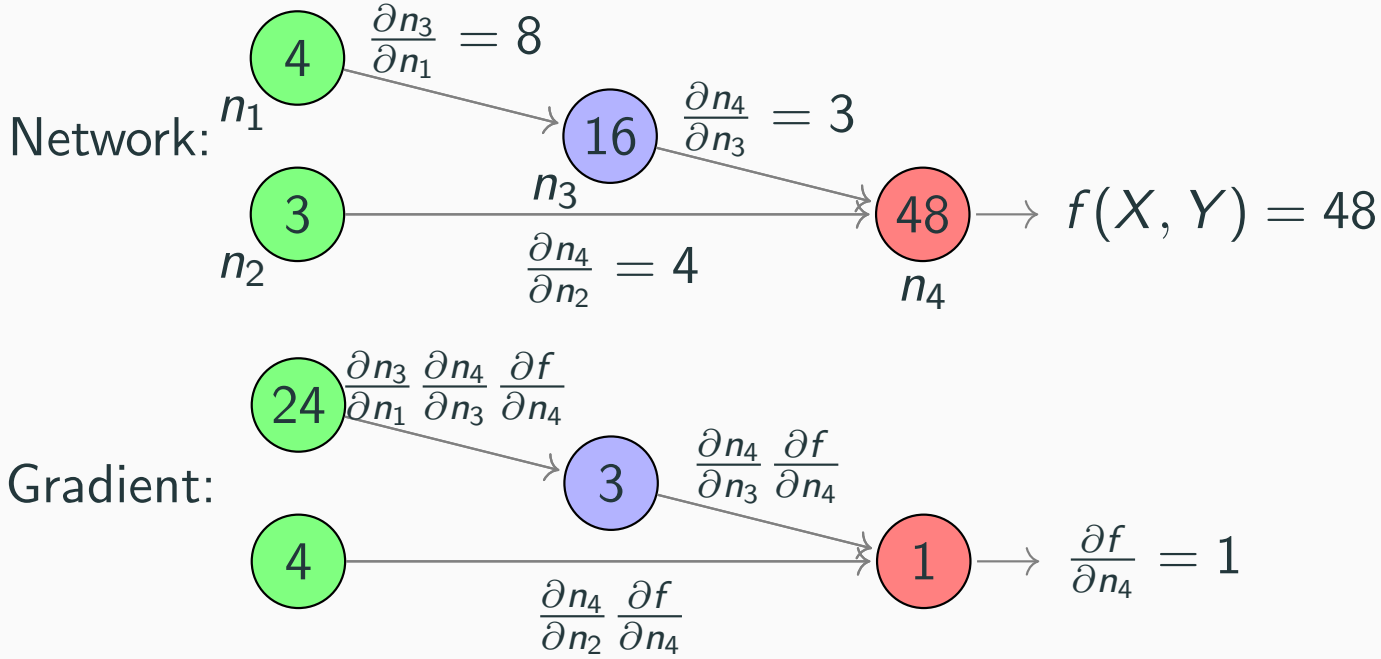
At runtime, for fixed (X, Y), say (4,3), we compute the gradient as follows.
Pass (4,3) through the graph, filling all nodes.
Second, introduce a new graph and back fill it using the values computed in the forward graph. Starting with $\frac{\partial f}{\partial n_4} = 1$, multiply backwards until the graph is filled.

Reverse Autodiff



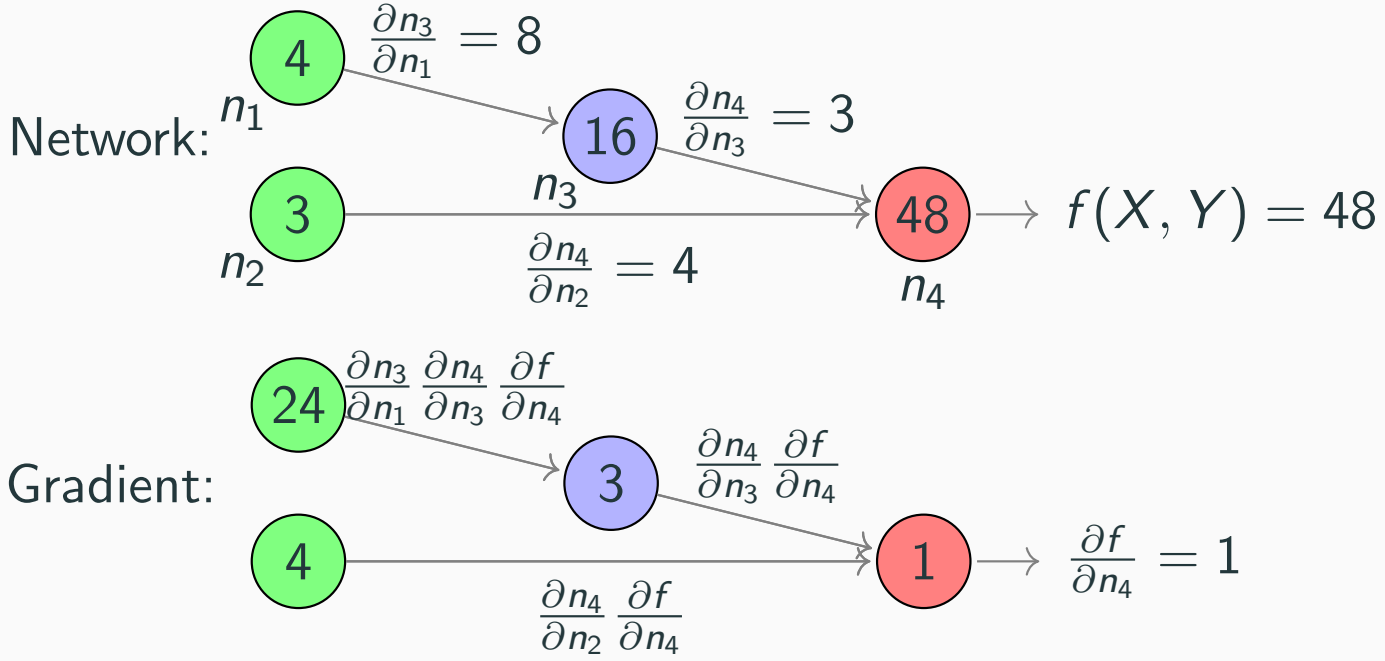
At runtime, for fixed (X, Y), say (4,3), we compute the gradient as follows.
Pass (4,3) through the graph, filling all nodes.
Second, introduce a new graph and back fill it using the values computed in the forward graph. Starting with $\frac{\partial f}{\partial n_4} = 1$, multiply backwards until the graph is filled.

Reverse Autodiff



At runtime, for fixed (X, Y), say (4,3), we compute the gradient as follows.
Pass (4,3) through the graph, filling all nodes.
Second, introduce a new graph and back fill it using the values computed in the forward graph. Starting with $\frac{\partial f}{\partial n_4} = 1$, multiply backwards until the graph is filled.

Reverse Autodiff

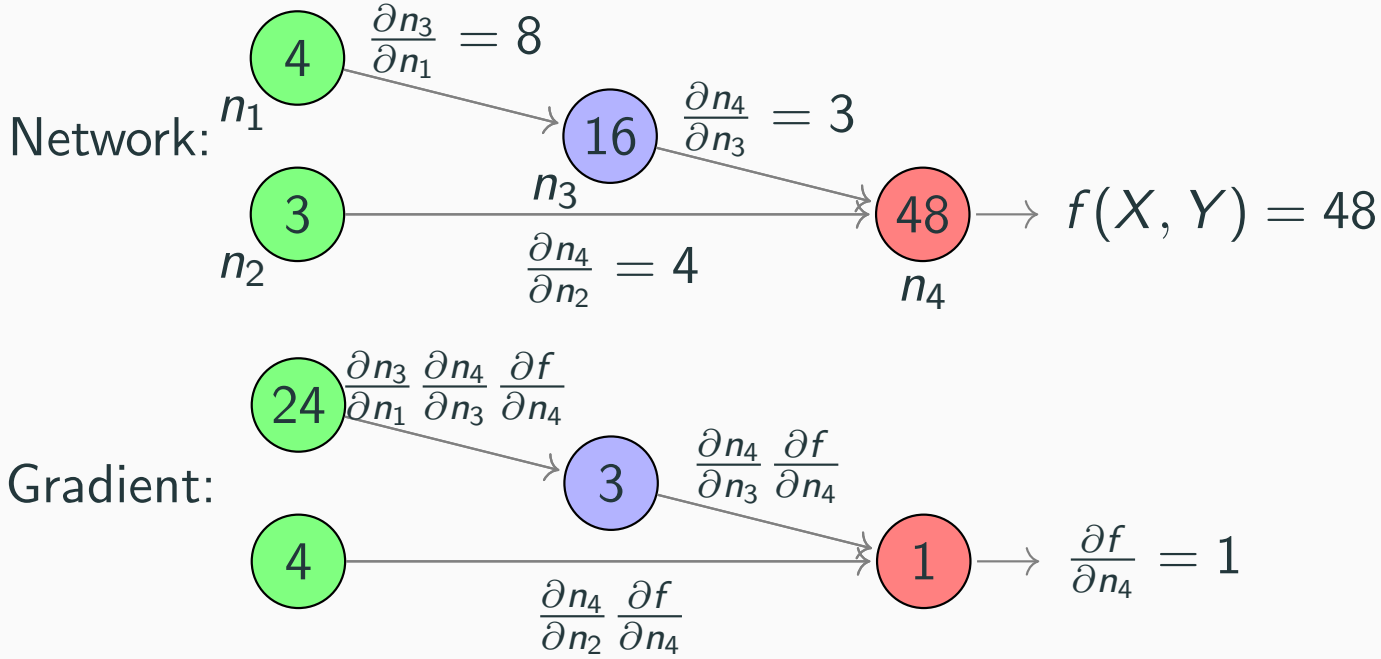


We are just computing the chain rule

$$\frac{\partial f}{\partial X}(4,3) = \frac{\partial n_1}{\partial x}\frac{\partial n_3}{\partial n_1}\frac{\partial n_4}{\partial n_3}\frac{\partial f}{\partial n_4}(4,3) = 24$$

By precomputing the derivatives at each node, reverse autodiff allows us to very quickly compute the whole gradient in two passes though the graph.

Reverse Autodiff



There is of course one caveat: If we have many output labels f_i or a node that is not differentiable we may not be able to use reverse autodiff.

The first problem is simply an limitation of the algorithm, but the second we can often smooth away by replacing a node with a smooth function, or by using a subgradient.

Training Neural Networks

Table 9-2. Main solutions to compute gradients automatically

Technique	Nb of graph traversals to compute all gradients	Accuracy	Supports arbitrary code	Comment
Numerical differentiation	$n_{\text{inputs}} + 1$	Low	Yes	Trivial to implement
Symbolic differentiation	N/A	High	No	Builds a very different graph
Forward-mode autodiff	n_{inputs}	High	Yes	Uses *dual numbers*
Reverse-mode autodiff	$n_{\text{outputs}} + 1$	High	Yes	Implemented by TensorFlow

The four nontrivial methods of computing gradients for gradient decent are summarized in this tabel taken from Chapter 9 or Geron. Tensorflow has automatically implemented reverse autodiff and as long as you use it's constructors it can compute gradients for you.

If you intend to implement a new kind of neuron you will have to code in compatibility with autodiff by hand, but after you do it once tensorflow will take it forward.

Reference

McCulloch and Pitts "A logical calculus of ideas immanent is nervous activity."
`http://www.cse.chalmers.se/~coquand/AUTOMATA/mcp.pdf`

Neuroplacsicity Picture:
`https://www.cognifit.com/brain-plasticity-and-cognition`

This lecture is based on Chapters 9 - 10 of Geron "Hands on Machine Learning with Scikit-Learn and Tensorflow."