6.034
“Neural” Networks
Peter Szolovits

ai6034.mit.edu

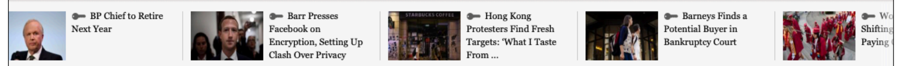
October 7, 2019



THE WALL STREET JOURNAL.

Subscribe Now
50% Off 1 Ye

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine



CIO JOURNAL.

Zillow Develops Neural Network to ‘See’ Like a House Hunter

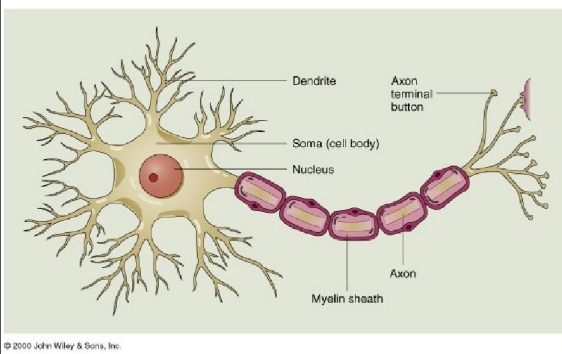
Granite or stainless steel countertops? Zillow’s visual recognition effort can recognize the difference

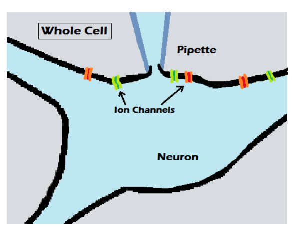
- Data scientists at Zillow Group are developing complex computer programs that detect specific attributes in photographs of homes, which could aid in estimating their value.
- Advances in deep learning, big data and cloud computing have converged to allow the online real estate database firm and others to develop technology that mimics how the human brain processes visual images--a concept still in its early stages and once limited to only the largest technology companies.

https://blogs.wsj.com/cio/2016/11/11/zillow-develops-neural-network-to-see-like-a-home-buyer/

Real Neurons

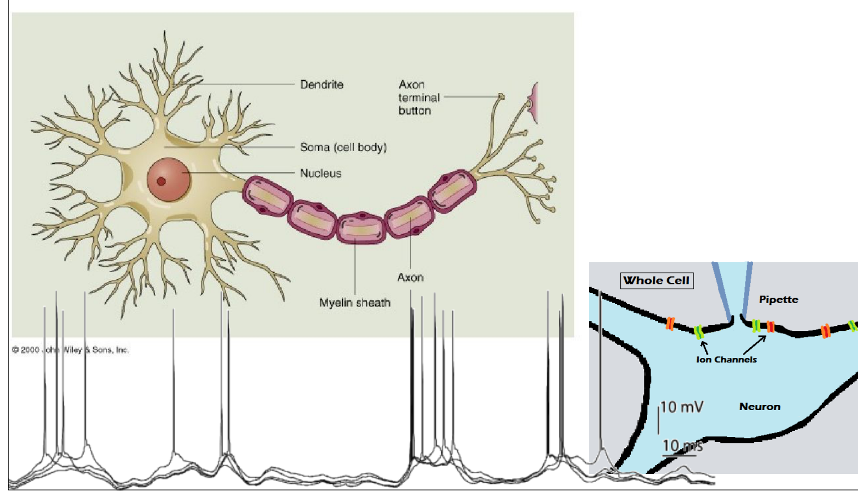
For models of how real neurons work:
https://en.wikipedia.org/wiki/Biological_neuron_model





Real Neurons

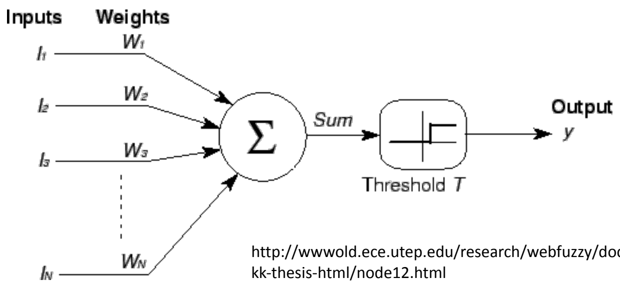
For models of how real neurons work:
https://en.wikipedia.org/wiki/Biological_neuron_model



McCulloch-Pitts Model

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133. http://doi.org/10.1007/BF02478259

- 1. The activity of the neuron is an "all-or-none" process.
- 2. A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position on the neuron.
- 3. The only significant delay within the nervous system is synaptic delay.
- 4. The activity of any inhibitory synapse absolutely prevents excitation of the neuron at that time.
- 5. The structure of the net does not change with time.



http://wwwold.ece.utep.edu/research/webfuzzy/docs/kk-thesis/kk-thesis-html/node12.html

Publication No.: 9438

MINSKY, Marvin Lee. THEORY OF NEURAL-ANALOG REINFORCEMENT SYSTEMS AND ITS APPLICATION TO THE BRAIN-MODEL PROBLEM.

Princeton University, Ph.D., 1954
Mathematics

Please Note: Find page numbered as 1-8 at the end of film copy.

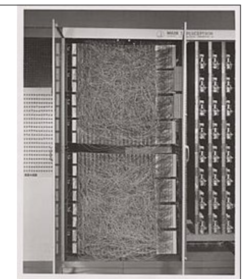
University Microfilms, Inc., Ann Arbor, Michigan

Perceptron

Frank Rosenblatt, 1958, Cornell

- McCulloch-Pitts model of neuron
  - constant input $x_0 = -1$, then $w_0 = T$
  - $y_j(t) = f(\mathbf{w} \cdot \mathbf{x}_j)$
  - f is the threshold function, usually $f(z) = (z > 0)$
- Learning Method:
  - $\{(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_s, d_s)\}$ are the training set, each $\mathbf{x}_j$ an n dimensional vector, d_j the desired (binary) answer
  - $w_i(t+1) = w(t) + r \cdot (d_j - y_j(t))x_{j,i}$ for all features $0 \leq i \leq n$
    - r is the learning rate
- Finds linear separators in n dimensional space

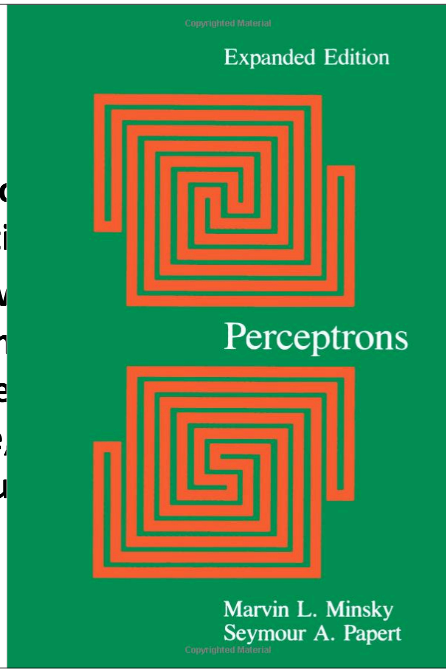
Perceptron

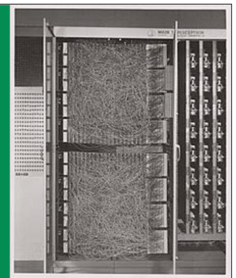


- Meant to be hardware for image recognition, 20x20 photocells
- The New York Times reported the perceptron to be "the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence."

https://en.wikipedia.org/wiki/Perceptron

- Meant t
recogniti
- The New
to be "th
that [the
talk, see,
consciou



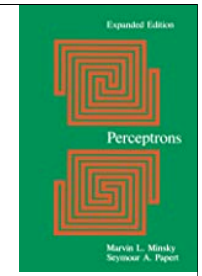


erceptron
computer
to walk,
l be

pedia.org/wiki/Perceptron

Minsky and Papert, 1969

- Single-layer perceptrons cannot model
  - XOR
  - Connectivity
- Blamed for halt to numerical models of intelligence for decades



Geoffrey Hinton,
The persistent

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

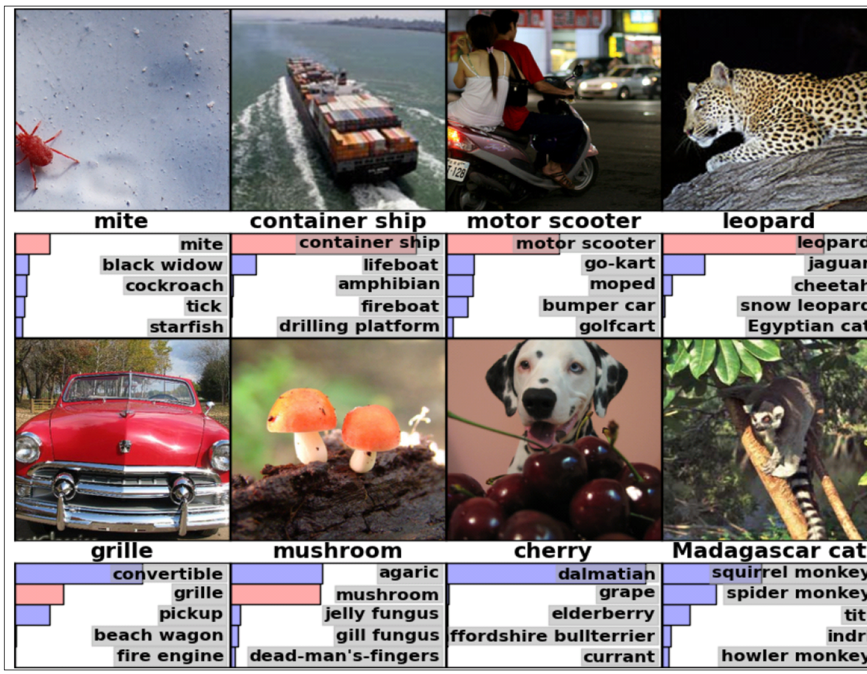
Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

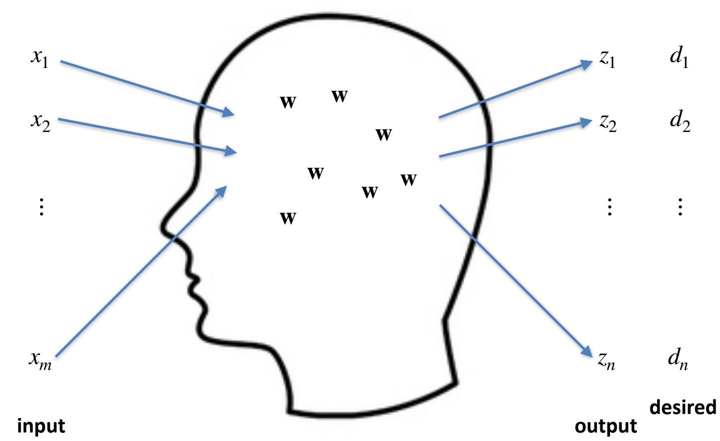
Neural Information Processing Systems, 2012



Are Artificial Neurons = Real Neurons?

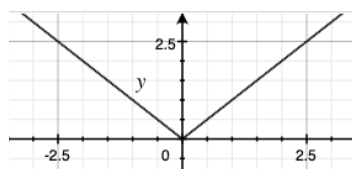
- Refractory period
- Axonal bifurcation — which way does pulse propagate?
- Is information in a spike or a spike train?
  - How is it encoded?
- Do bundles of nerves convey information together?
- Growth and pruning of neural connections

Brain as Transducer

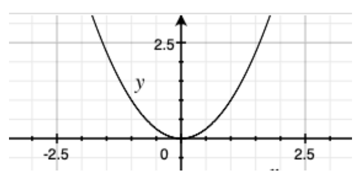


Measuring Performance

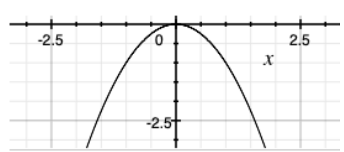
- $P = \| \mathbf{d} - \mathbf{z} \|$



- $P = \| \mathbf{d} - \mathbf{z} \|^2$

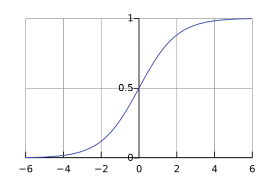


- $P = - \| \mathbf{d} - \mathbf{z} \|^2$

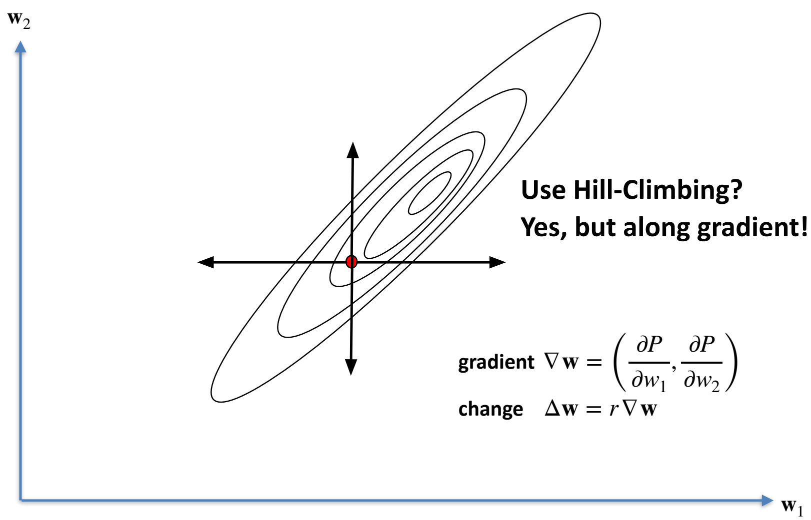


Replace Threshold Function by Sigmoid

- $\sigma(x) = \dfrac{1}{1 + e^{-x}}$



Improving Performance

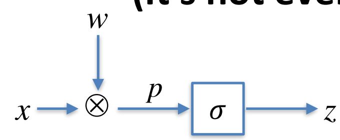


Use Hill-Climbing?
Yes, but along gradient!

gradient $\nabla \mathbf{w} = \left(\frac{\partial P}{\partial w_1}, \frac{\partial P}{\partial w_2}\right)$

change $\Delta \mathbf{w} = r \nabla \mathbf{w}$

Simplest Possible Neural Network (It's not even a network!)



$$P = -\frac{1}{2}(d-z)^2$$

We want $\frac{\partial P}{\partial w}$**; by chain rule,** $= \frac{\partial P}{\partial z}\frac{\partial z}{\partial w}$

$$= \frac{\partial P}{\partial z}\frac{\partial z}{\partial p}\frac{\partial p}{\partial w}$$

Differentiate!

$$\frac{\partial P}{\partial w} = \frac{\partial P}{\partial z}\frac{\partial z}{\partial p}\frac{\partial p}{\partial w} \qquad P = -\frac{1}{2}(d-z)^2$$

$$\frac{\partial P}{\partial z} = d - z$$

$$\frac{\partial p}{\partial w} = x \qquad p = wx$$

$$\frac{\partial z}{\partial p} = \frac{\partial}{\partial p}\sigma(p) = \frac{\partial}{\partial p}\frac{1}{1+e^{-p}} = \cdots = \sigma(p)\cdot(1-\sigma(p)) \qquad z = \sigma(p)$$

$$= z(1-z)$$

So, $\frac{\partial P}{\partial w} = x(d-z)z(1-z)$

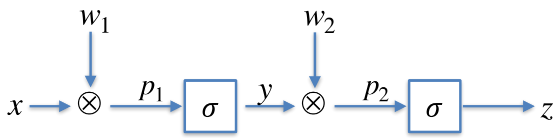
In case you don't believe me…

Here's a detailed derivation:

$$\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] \\
&= \frac{d}{dx}\left(1+e^{-x}\right)^{-1} \\
&= -(1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) \\
&= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) \\
&= \sigma(x) \cdot (1-\sigma(x))
\end{aligned}$$

https://math.stackexchange.com/questions/78575/derivative-of-sigmoid-function-sigma-x-frac11e-x

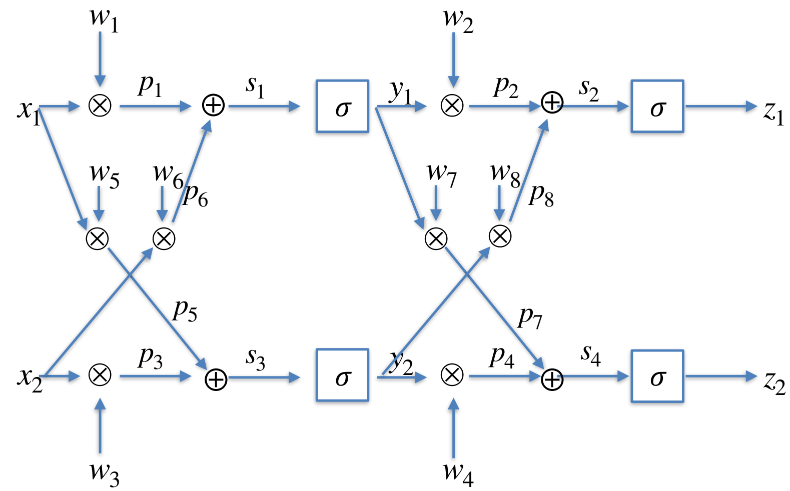
2-layer NN



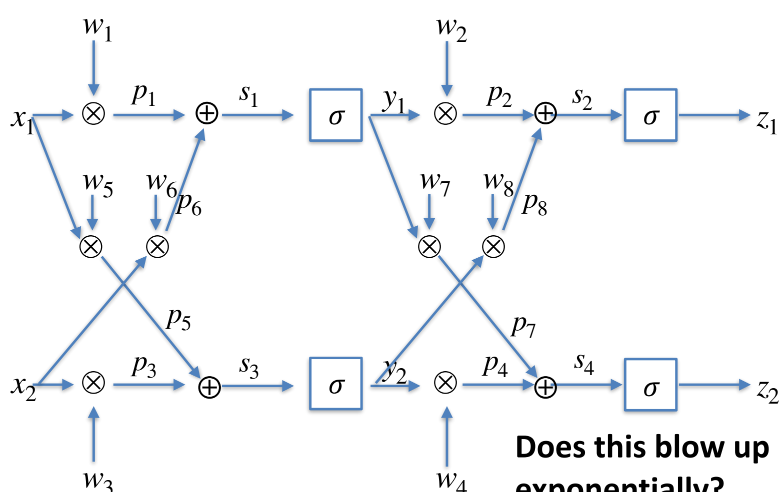
$$\frac{\partial P}{\partial w_2} = \frac{\partial P}{\partial z}\frac{\partial z}{\partial p_2}\frac{\partial p_2}{\partial w_2}$$

$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z}\frac{\partial z}{\partial p_2}\frac{\partial p_2}{\partial y}\frac{\partial y}{\partial p_1}\frac{\partial p_1}{\partial w_2}$$

2-layer 2-wide NN

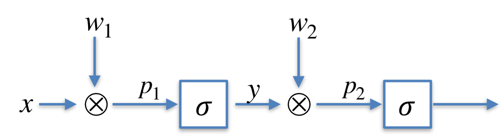


2-layer 2-wide NN



Does this blow up exponentially?

Is this Practical?



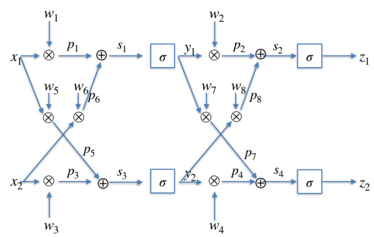
Remember from simple 2-layer NN:

$$\frac{\partial P}{\partial w_2} = \boxed{\frac{\partial P}{\partial z}\frac{\partial z}{\partial p_2}}\frac{\partial p_2}{\partial w_2}$$

$$\frac{\partial P}{\partial w_1} = \boxed{\frac{\partial P}{\partial z}\frac{\partial z}{\partial p_2}}\frac{\partial p_2}{\partial y}\frac{\partial y}{\partial p_1}\frac{\partial p_1}{\partial w_2}$$

- **Re-use**

Lots of Re-Use

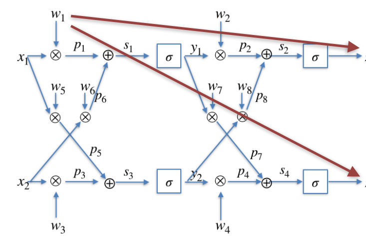


$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1}$$

$$\frac{\partial P}{\partial w_6} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6}$$

- **Complexity is *linear* in depth of network**
  - **and *quadratic* in width**

Lots of Re-Use

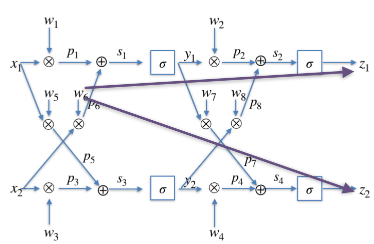


$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1}$$

$$\frac{\partial P}{\partial w_6} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6}$$

- **Complexity is *linear* in depth of network**
  - **and *quadratic* in width**

Lots of Re-Use

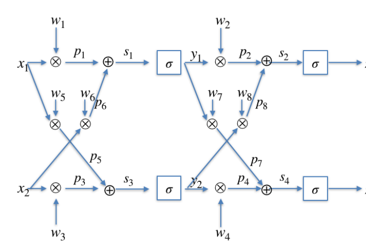


$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1}$$

$$\frac{\partial P}{\partial w_6} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6}$$

- **Complexity is *linear* in depth of network**
  - **and *quadratic* in width**

Lots of Re-Use

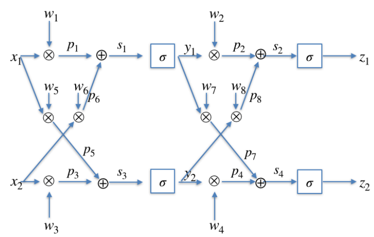


$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1}$$

$$\frac{\partial P}{\partial w_6} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6} + \frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6}$$

- **Complexity is *linear* in depth of network**
  - **and *quadratic* in width**

Lots of Re-Use



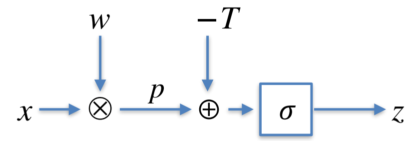
$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1} + \boxed{\frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_1}\frac{\partial p_1}{\partial w_1}$$

$$\frac{\partial P}{\partial w_6} = \frac{\partial P}{\partial z_1}\frac{\partial z_1}{\partial s_2}\frac{\partial s_2}{\partial p_2}\frac{\partial p_2}{\partial y_1}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6} + \boxed{\frac{\partial P}{\partial z_2}\frac{\partial z_2}{\partial s_4}\frac{\partial s_4}{\partial p_7}\frac{\partial p_7}{\partial y_1}}\frac{\partial y_1}{\partial s_1}\frac{\partial s_1}{\partial p_6}\frac{\partial p_6}{\partial w_6}$$

- **Complexity is *linear* in depth of network**
  - **and *quadratic* in width**

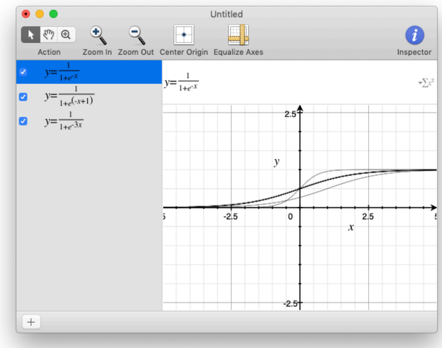
Revisit the Sigmoid to Fit Probabilities (Logistic Regression)

$$z = \frac{1}{1 + e^{-wx+T}}$$

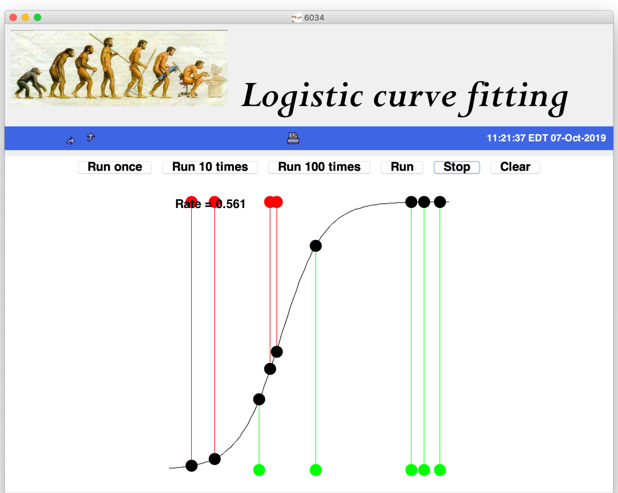


Want to adjust w, T so the z's match probability of the data we see.

Same partial derivative tricks as before.



Fitting a Single Neuron to Data



SoftMax

- **Interpret (continuous) outputs of many final-layer neurons as probabilities**

$$p(\textbf{result } i) = \frac{z_i}{\sum_j z_j}$$