ENM 5310: Data-driven Modeling and Probabilistic Scientific Computing

Lecture #11: Multi-layer perceptrons

Penn
UNIVERSITY of PENNSYLVANIA

Feed-forward neural networks

<u>Pros:</u>

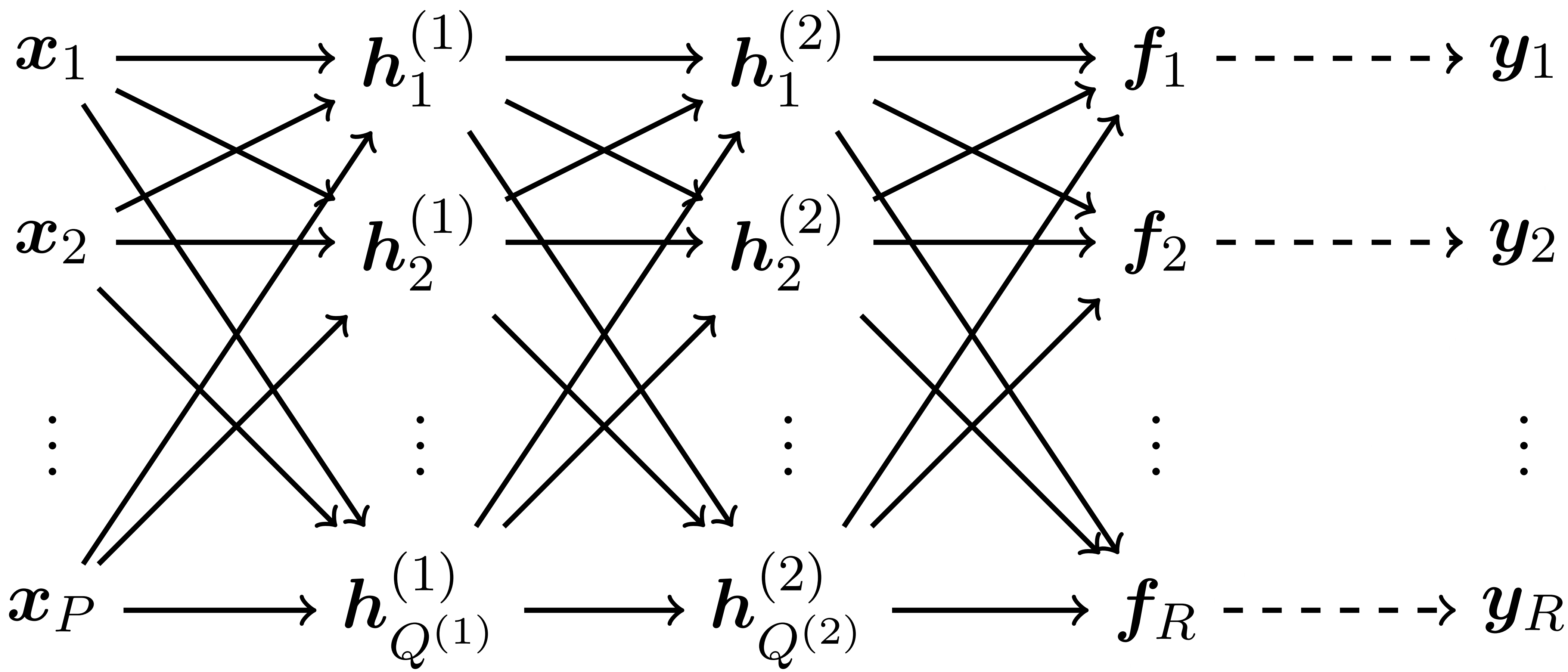
- Adaptive features/basis functions (parametric)
- Flexible non-linear regression models that can approximate any function.
- Scalability to high dimensions.

<u>Cons:</u>

- The likelihood function is no longer a convex function of the model parameters.
- Over-fitting in data-scarce scenarios.
- Results are hard to interpret.

University of Illinois at Urbana-Champaign. Center for Supercomputing Research and Development, & Cybenko, G. (1988)
Continuous valued neural networks with two hidden layers are sufficient.

Feed-forward neural networks



Universal approximation theorem

Theorem 1. *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathrm{T}} x + \theta_j) \tag{2}$$

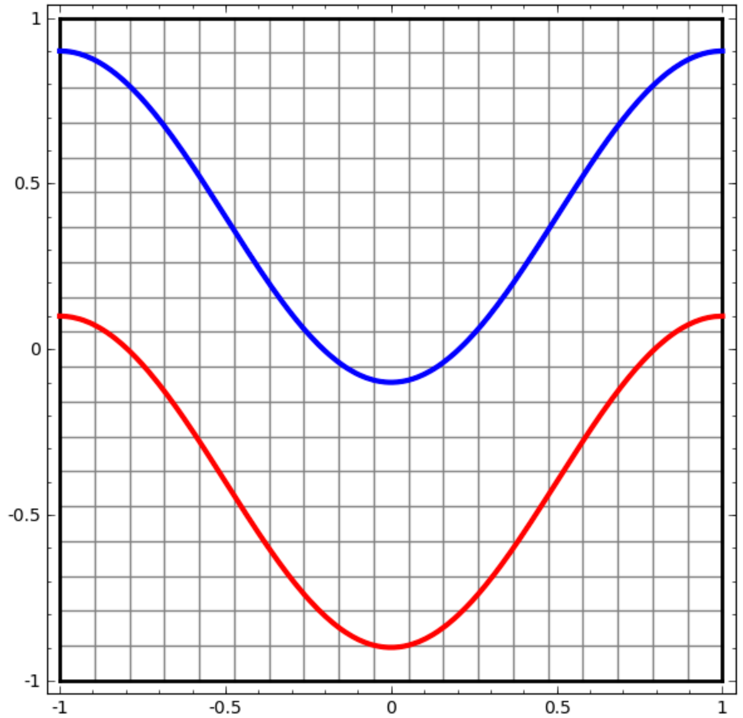
are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \qquad \textit{for all} \quad x \in I_n.$$

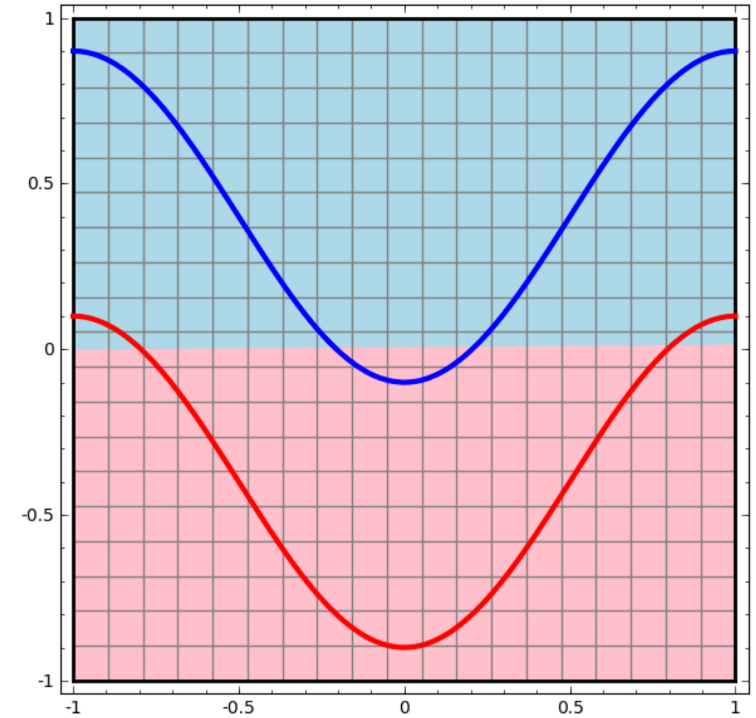
Cybenko, G. (1989) "Approximations by superpositions of sigmoidal functions", Mathematics of Control, Signals, and Systems, 2 (4), 303-314

Some intuition

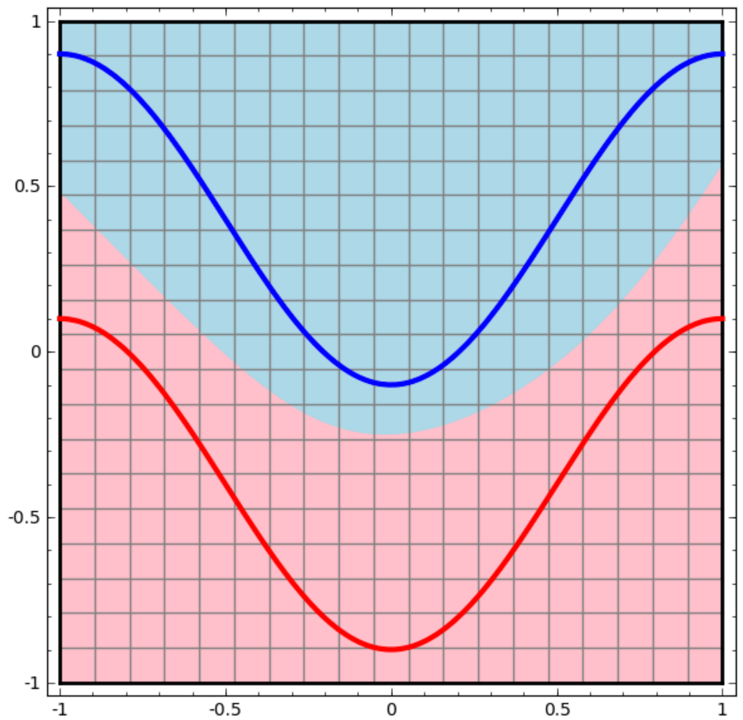
Data to classify



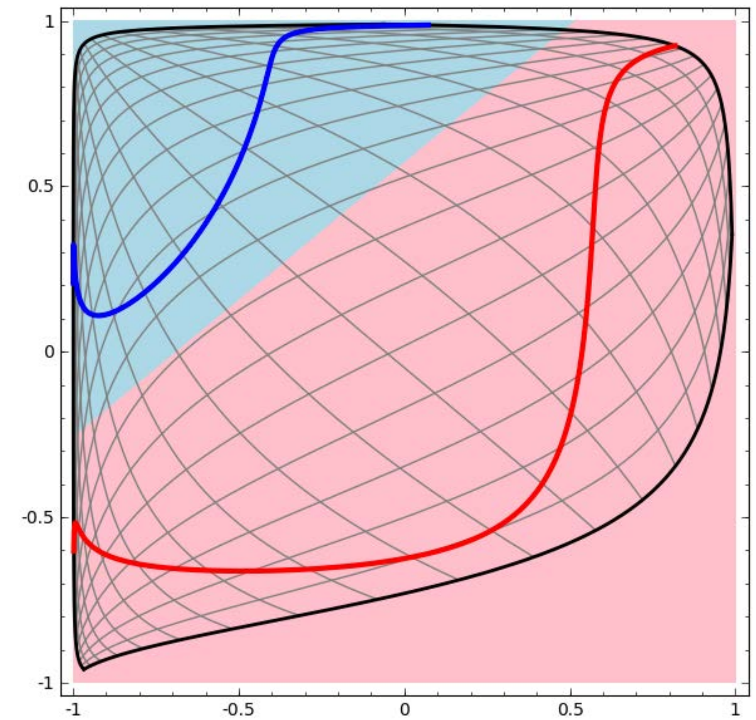
One input, one output layer



One input, one hidden, one output layer

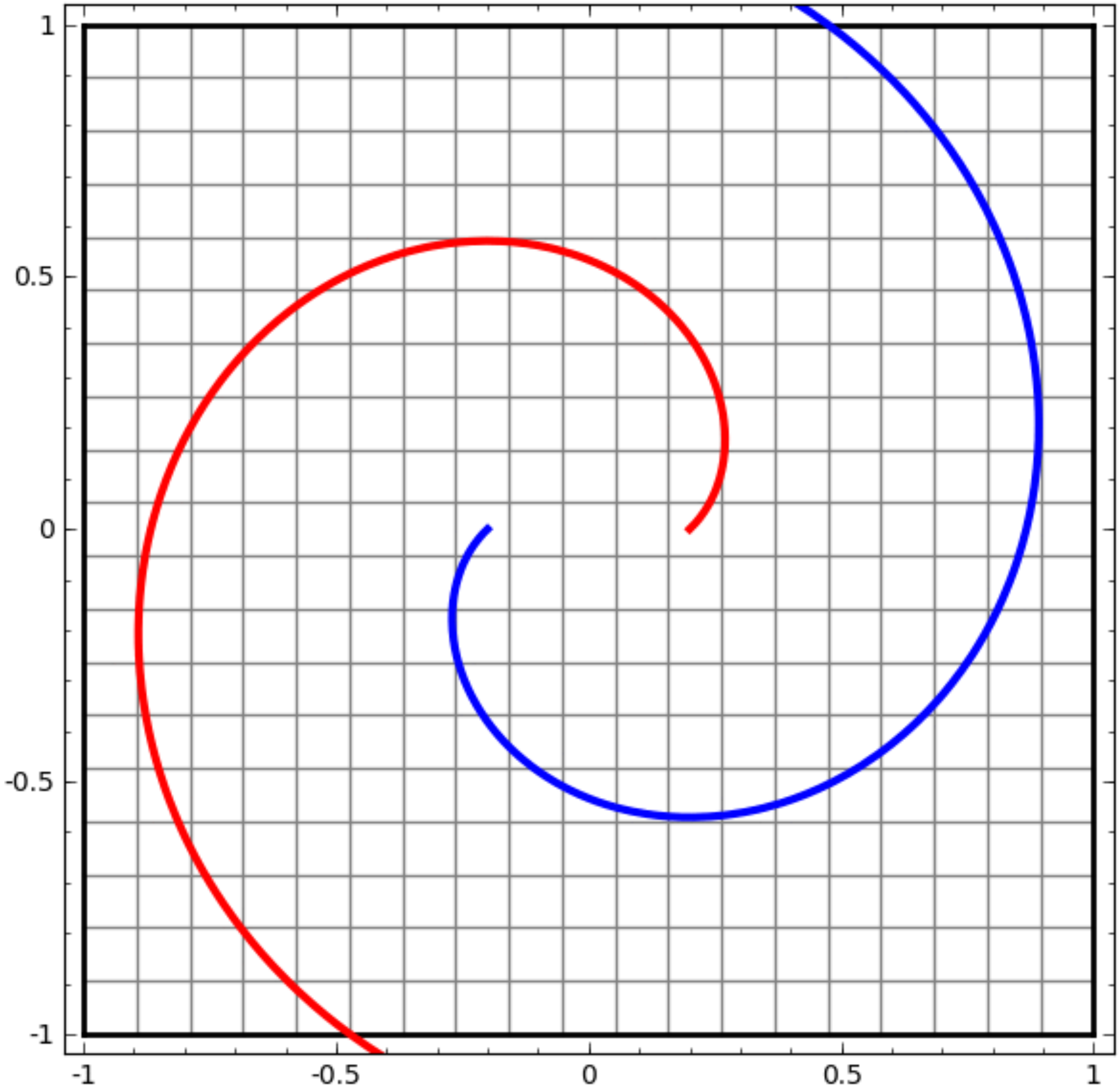


Visualizing the hidden layer



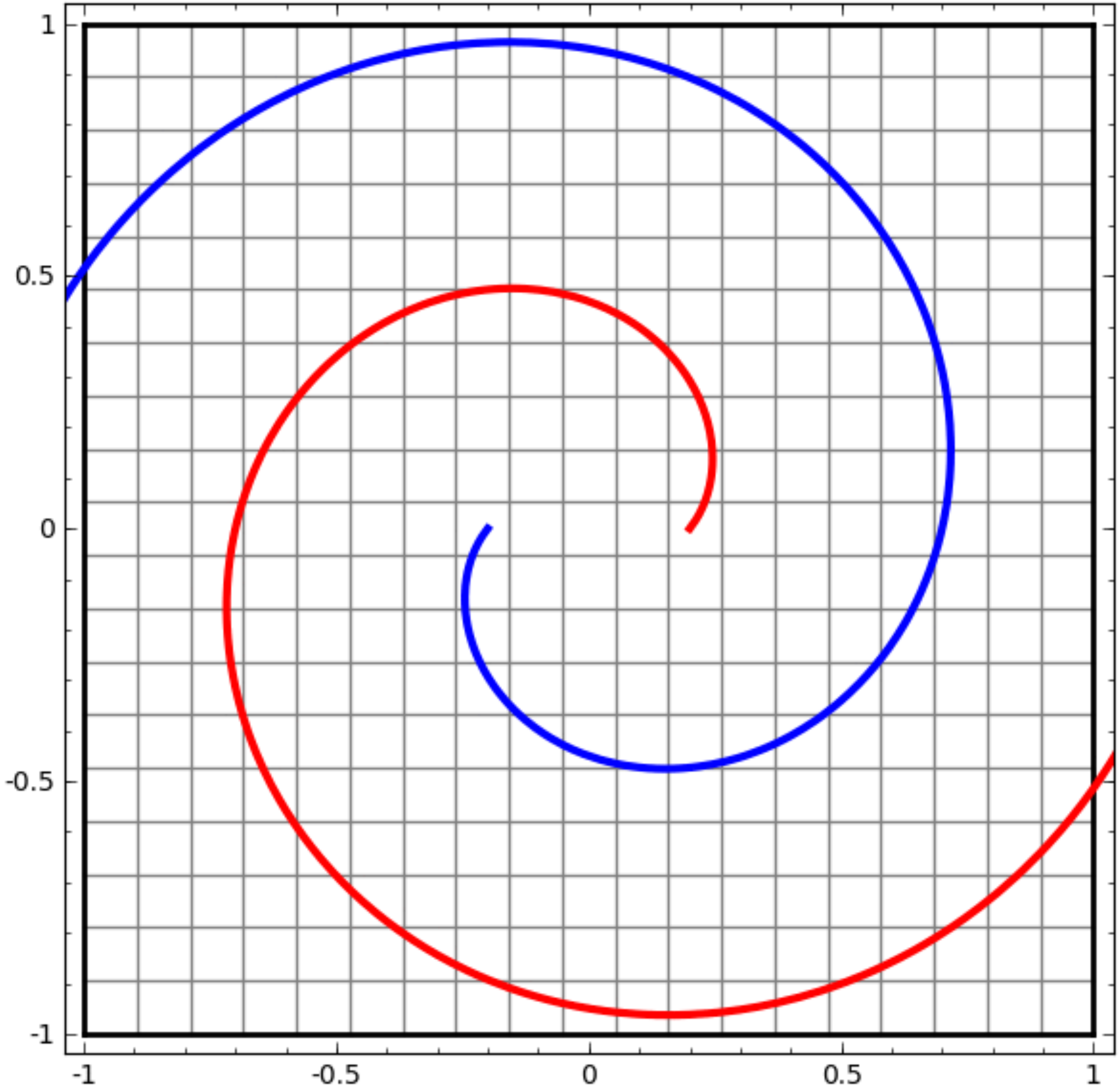
http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Some intuition



http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

Some intuition



http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

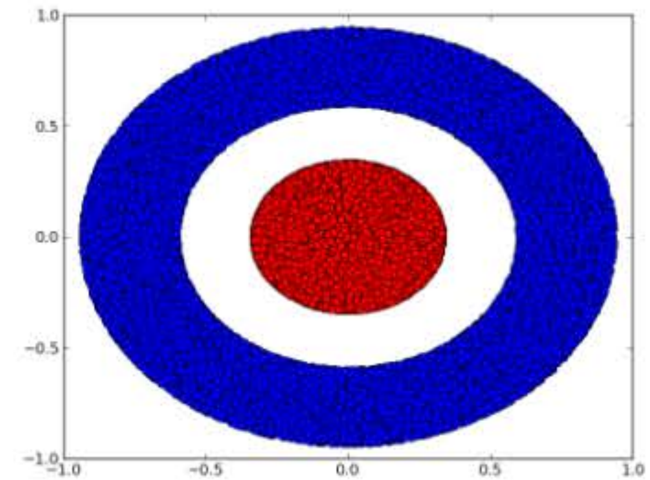
Some intuition

Topology and Classification

Consider a two dimensional dataset with two classes $A, B \subset \mathbb{R}^2$:

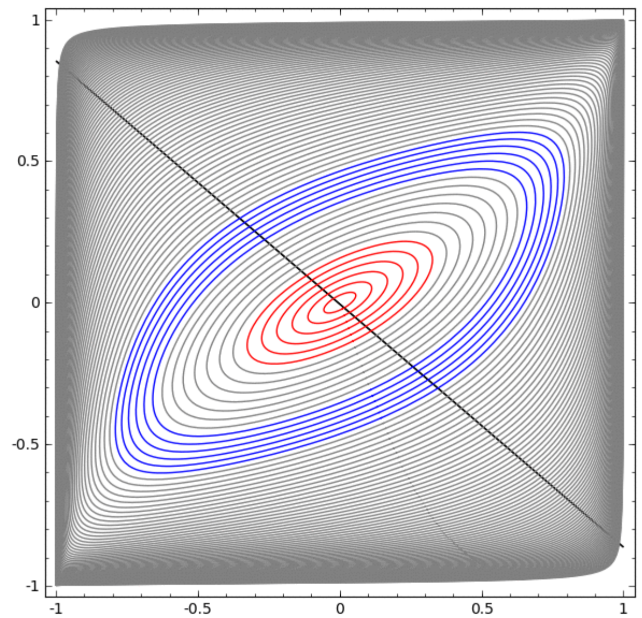
$$A = \{x | d(x, 0) < 1/3\}$$

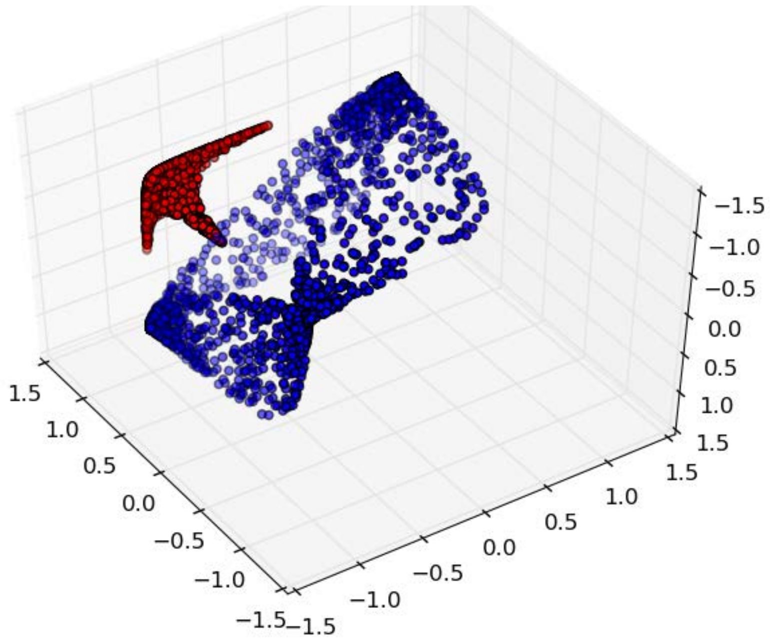
$$B = \{x | 2/3 < d(x, 0) < 1\}$$



A is red, B is blue

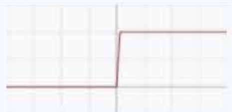
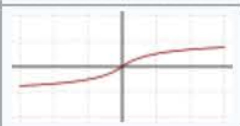
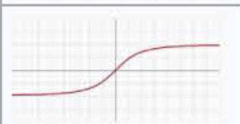
Claim: It is impossible for a neural network to classify this dataset without having a layer that has 3 or more hidden units, regardless of depth.





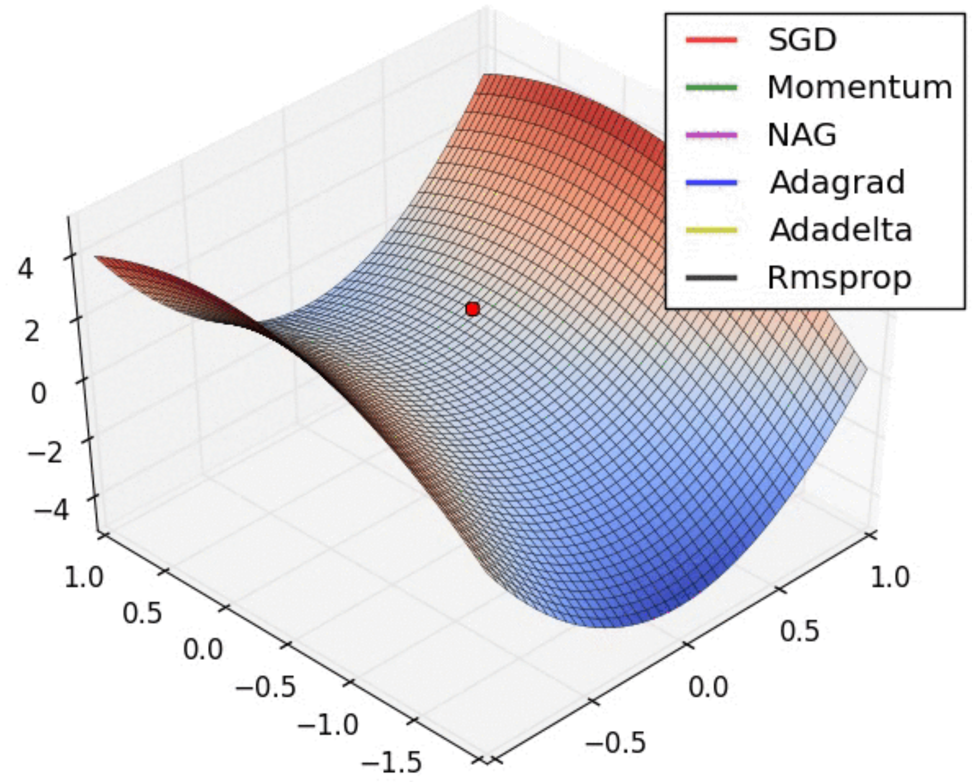
http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/

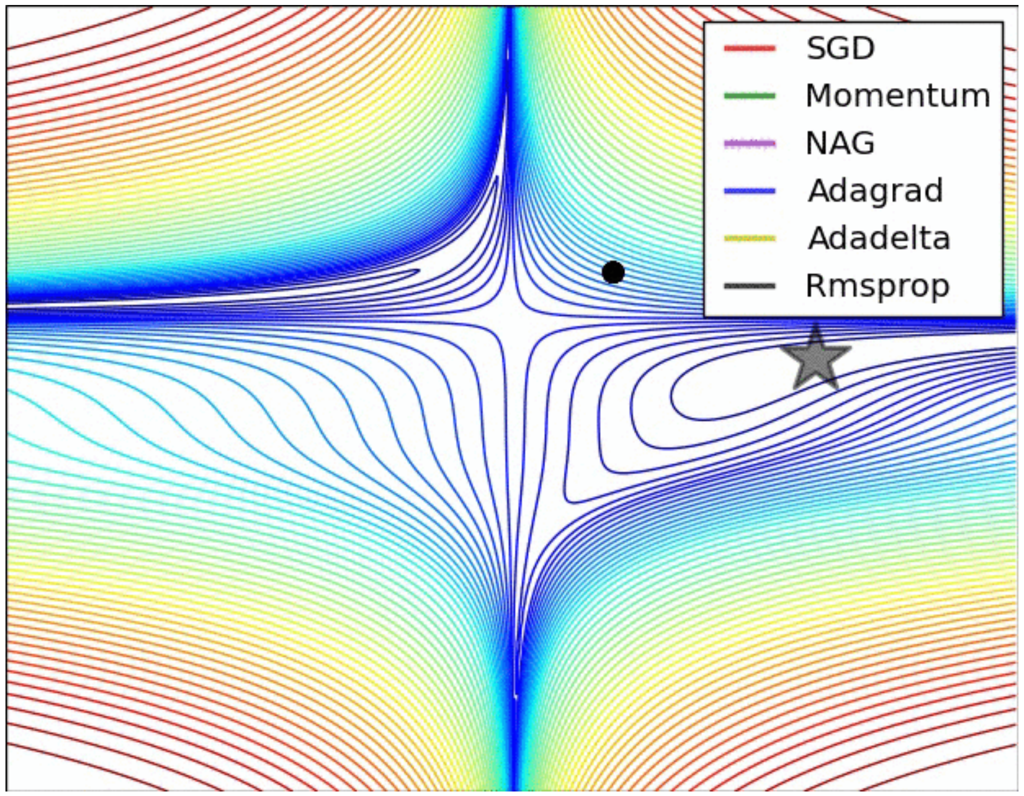
Activation functions

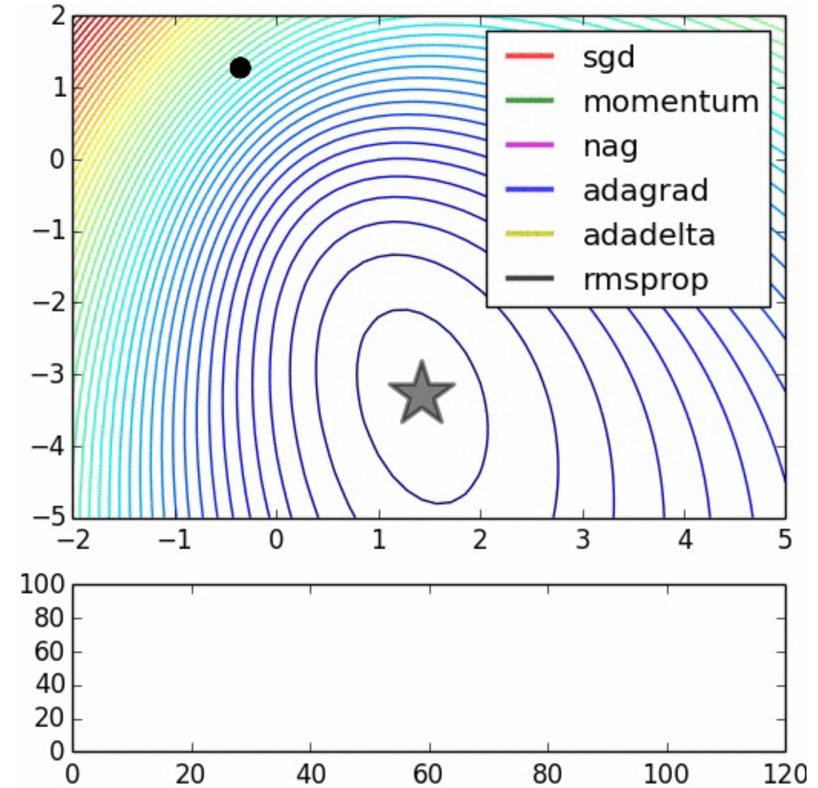
Name	Plot	Equation	Derivative (with respect to *x*)	Range	Order of continuity	Monotonic	Derivative Monotonic	Approximates identity near the origin
Identity		$f(x) = x$	$f'(x) = 1$	$(-\infty, \infty)$	C^∞	Yes	Yes	Yes
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$	$\{0, 1\}$	C^{-1}	Yes	No	No
Logistic (a.k.a. Soft step)		$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$	$(0, 1)$	C^∞	Yes	No	No
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$	$(-1, 1)$	C^∞	Yes	No	Yes
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2+1}$	$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$	C^∞	Yes	No	Yes
Softsign [7][8]		$f(x) = \frac{x}{1+\|x\|}$	$f'(x) = \frac{1}{(1+\|x\|)^2}$	$(-1, 1)$	C^1	Yes	No	Yes
Inverse square root unit (ISRU)[9]		$f(x) = \frac{x}{\sqrt{1+\alpha x^2}}$	$f'(x) = \left(\frac{1}{\sqrt{1+\alpha x^2}}\right)^3$	$\left(-\frac{1}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}}\right)$	C^∞	Yes	No	Yes
Rectified linear unit (ReLU)[10]		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$[0, \infty)$	C^0	Yes	Yes	No
Leaky rectified linear unit (Leaky ReLU)[11]		$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty, \infty)$	C^0	Yes	Yes	No
Parameteric rectified linear unit (PReLU)[12]		$f(\alpha, x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(\alpha, x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$(-\infty, \infty)$	C^0	Yes iff $\alpha \geq 0$	Yes	Yes iff $\alpha = 1$

⋮

Modern SGD variants







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

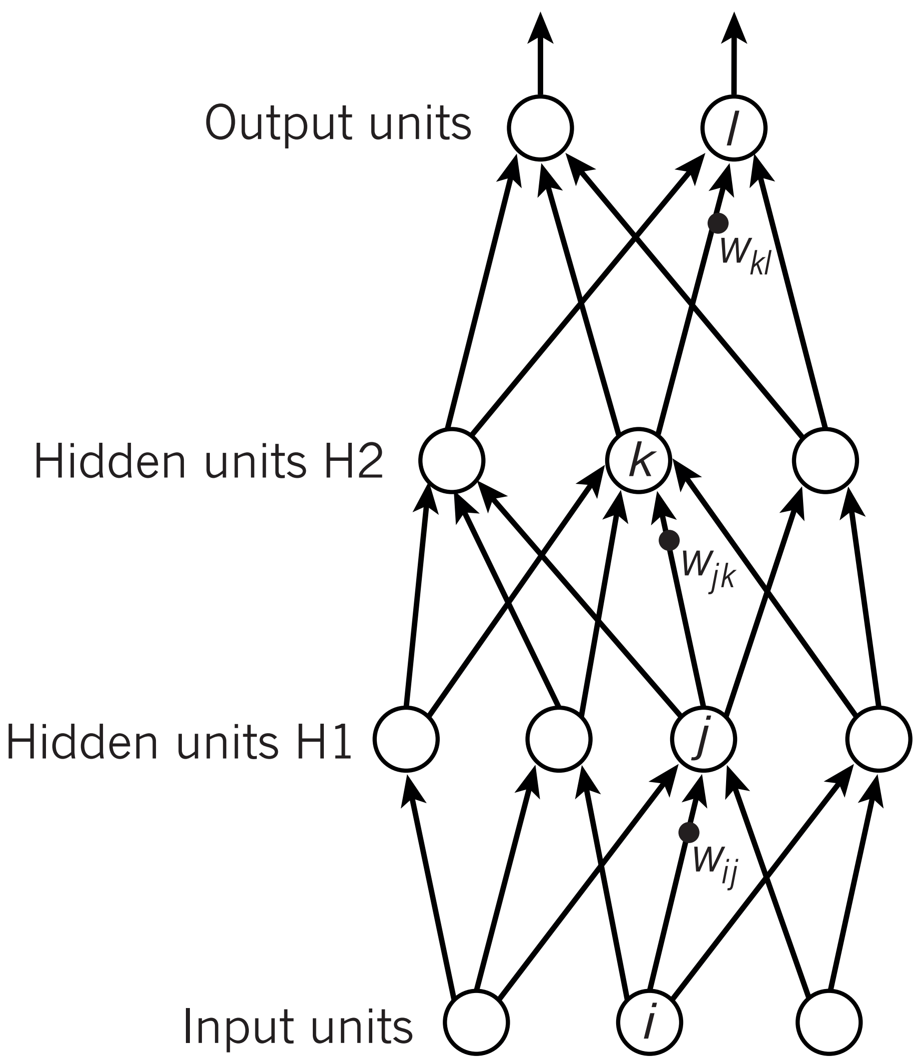
https://distill.pub/2017/momentum/

http://louistiao.me/notes/visualizing-and-animating-optimization-algorithms-with-matplotlib/

*animation credit: Alec Redford

Back-propagation

Forward pass



$$y_l = f(z_l)$$

$$z_l = \sum_{k \,\varepsilon\, \mathrm{H2}} w_{kl}\, y_k$$

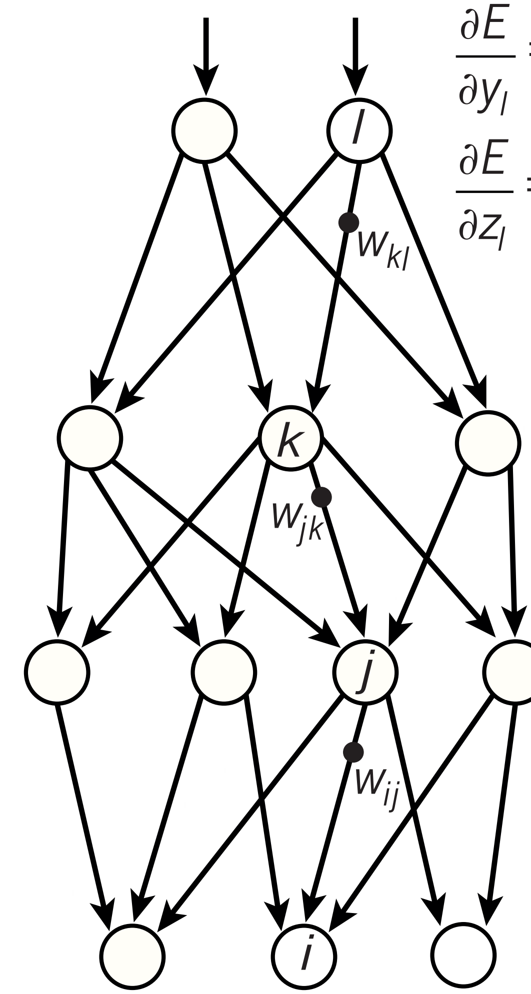
$$y_k = f(z_k)$$

$$z_k = \sum_{j \,\varepsilon\, \mathrm{H1}} w_{jk}\, y_j$$

$$y_j = f(z_j)$$

$$z_j = \sum_{i \,\varepsilon\, \mathrm{Input}} w_{ij}\, x_i$$

Backward pass



$$\frac{\partial E}{\partial y_l} = y_l - t_l$$

$$\frac{\partial E}{\partial z_l} = \frac{\partial E}{\partial y_l}\,\frac{\partial y_l}{\partial z_l}$$

$$\frac{\partial E}{\partial y_k} = \sum_{l \,\varepsilon\, \mathrm{out}} w_{kl}\,\frac{\partial E}{\partial z_l}$$

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k}\,\frac{\partial y_k}{\partial z_k}$$

$$\frac{\partial E}{\partial y_j} = \sum_{k \,\varepsilon\, \mathrm{H2}} w_{jk}\,\frac{\partial E}{\partial z_k}$$

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial y_j}\,\frac{\partial y_j}{\partial z_j}$$

The forward pass

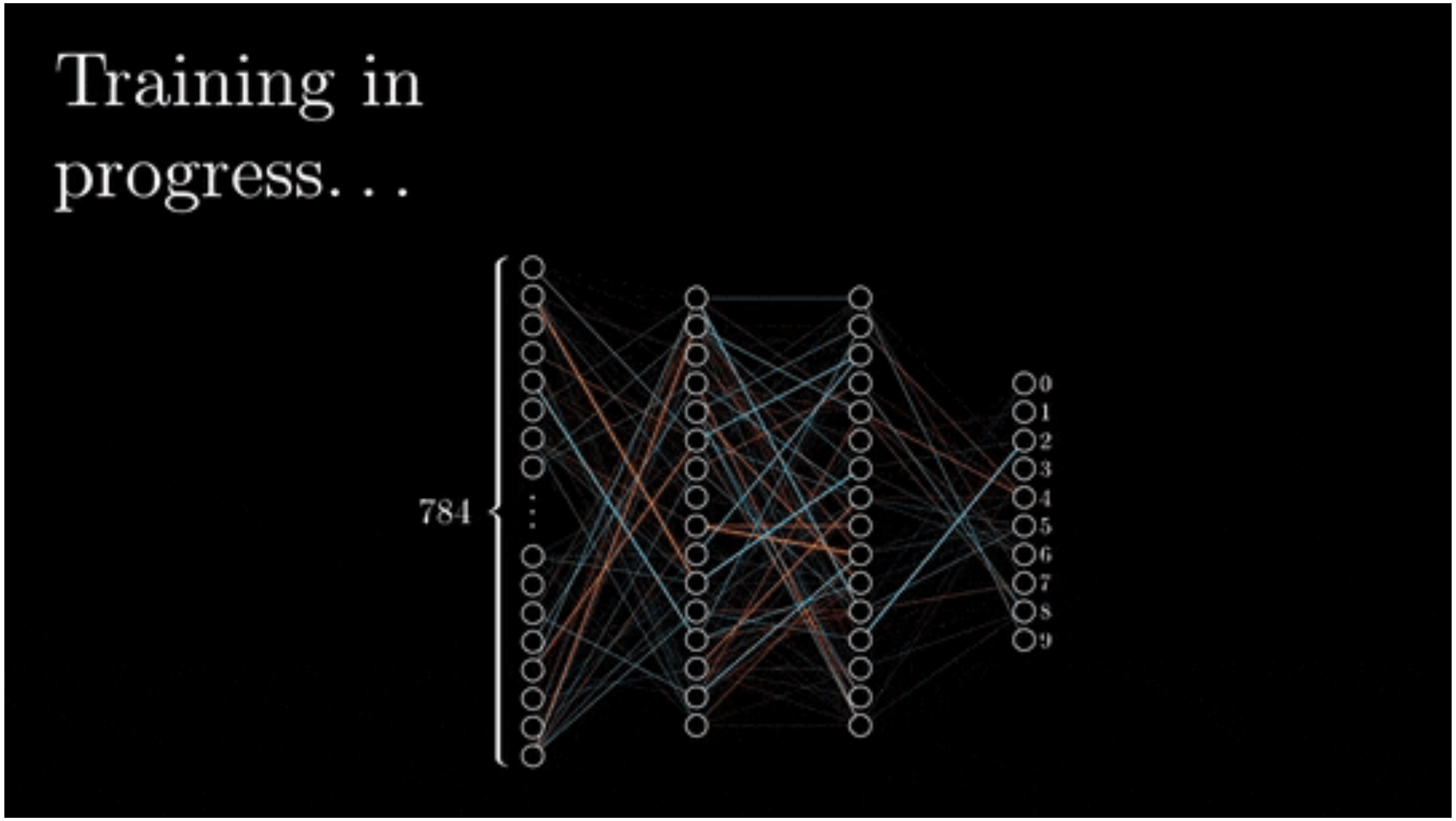
$$
\begin{array}{ccccc}
F = A_\ell W_\ell + b_\ell, & F \in \mathbb{R}^{n \times p_{\ell+1}}, & A_\ell \in \mathbb{R}^{n \times p_\ell}, & W_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}, & b_\ell \in \mathbb{R}^{1 \times p_{\ell+1}}, \\
A_\ell = \tanh(H_\ell), & A_\ell \in \mathbb{R}^{n \times p_\ell}, & H_\ell \in \mathbb{R}^{n \times p_\ell}, & & \\
H_\ell = A_{\ell-1} W_{\ell-1} + b_{\ell-1}, & H_\ell \in \mathbb{R}^{n \times p_\ell}, & A_{\ell-1} \in \mathbb{R}^{n \times p_{\ell-1}}, & W_{\ell-1} \in \mathbb{R}^{p_{\ell-1} \times p_\ell}, & b_{\ell-1} \in \mathbb{R}^{1 \times p_\ell}, \\
A_{\ell-1} = \tanh(H_{\ell-1}), & A_{\ell-1} \in \mathbb{R}^{n \times p_{\ell-1}}, & H_{\ell-1} \in \mathbb{R}^{n \times p_{\ell-1}}, & & \\
H_{\ell-1} = A_{\ell-2} W_{\ell-2} + b_{\ell-2}, & H_{\ell-1} \in \mathbb{R}^{n \times p_{\ell-1}}, & A_{\ell-2} \in \mathbb{R}^{n \times p_{\ell-2}}, & W_{\ell-2} \in \mathbb{R}^{p_{\ell-2} \times p_{\ell-1}}, & b_{\ell-2} \in \mathbb{R}^{1 \times p_{\ell-1}}, \\
A_{\ell-2} = \tanh(H_{\ell-2}), & A_{\ell-2} \in \mathbb{R}^{n \times p_{\ell-2}}, & H_{\ell-2} \in \mathbb{R}^{n \times p_{\ell-2}}, & & \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
H_2 = A_1 W_1 + b_1, & H_2 \in \mathbb{R}^{n \times p_2}, & A_1 \in \mathbb{R}^{n \times p_1}, & W_1 \in \mathbb{R}^{p_1 \times p_2}, & b_1 \in \mathbb{R}^{1 \times p_2}, \\
A_1 = \tanh(H_1), & A_1 \in \mathbb{R}^{n \times p_1}, & H_1 \in \mathbb{R}^{n \times p_1}, & & \\
H_1 = X W_0 + b_0, & H_1 \in \mathbb{R}^{n \times p_1}, & X \in \mathbb{R}^{n \times p_0}, & W_0 \in \mathbb{R}^{p_0 \times p_1}, & b_0 \in \mathbb{R}^{1 \times p_1},
\end{array}
$$

The backward pass

$$\mathcal{L} := \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p_{\ell+1}} (F_{i,j} - Y_{i,j})^2 \longrightarrow \nabla_\theta \mathcal{L}(\theta)$$

$$\begin{aligned}
G_\ell &= F - Y \in \mathbb{R}^{n \times p_{\ell+1}}, & \tfrac{\partial \mathcal{L}}{\partial W_\ell} &= A_\ell^T G_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}, & \tfrac{\partial \mathcal{L}}{\partial b_\ell} &= \mathbb{1}^T G_\ell \in \mathbb{R}^{1 \times p_{\ell+1}}, \\
G_{\ell-1} &= (1 - A_\ell \odot A_\ell) \odot (G_\ell W_\ell^T) \in \mathbb{R}^{n \times p_\ell}, & \tfrac{\partial \mathcal{L}}{\partial W_{\ell-1}} &= A_{\ell-1}^T G_{\ell-1} \in \mathbb{R}^{p_{\ell-1} \times p_\ell}, & \tfrac{\partial \mathcal{L}}{\partial b_{\ell-1}} &= \mathbb{1}^T G_{\ell-1} \in \mathbb{R}^{1 \times p_\ell}, \\
G_{\ell-2} &= (1 - A_{\ell-1} \odot A_{\ell-1}) \odot (G_{\ell-1} W_{\ell-1}^T) \in \mathbb{R}^{n \times p_{\ell-1}}, & \tfrac{\partial \mathcal{L}}{\partial W_{\ell-2}} &= A_{\ell-2}^T G_{\ell-2} \in \mathbb{R}^{p_{\ell-2} \times p_{\ell-1}}, & \tfrac{\partial \mathcal{L}}{\partial b_{\ell-2}} &= \mathbb{1}^T G_{\ell-2} \in \mathbb{R}^{1 \times p_{\ell-1}}, \\
&\vdots & &\vdots & &\vdots \\
G_1 &= (1 - A_2 \odot A_2) \odot (G_2 W_2^T) \in \mathbb{R}^{n \times p_2}, & \tfrac{\partial \mathcal{L}}{\partial W_1} &= A_1^T G_1 \in \mathbb{R}^{p_1 \times p_2}, & \tfrac{\partial \mathcal{L}}{\partial b_1} &= \mathbb{1}^T G_1 \in \mathbb{R}^{1 \times p_2}, \\
G_0 &= (1 - A_1 \odot A_1) \odot (G_1 W_1^T) \in \mathbb{R}^{n \times p_1}, & \tfrac{\partial \mathcal{L}}{\partial W_0} &= X^T G_0 \in \mathbb{R}^{p_0 \times p_1}, & \tfrac{\partial \mathcal{L}}{\partial b_0} &= \mathbb{1}^T G_0 \in \mathbb{R}^{1 \times p_1}.
\end{aligned}$$

Backpropagation



$$
\begin{aligned}
G_\ell &= F - Y \in \mathbb{R}^{n \times p_{\ell+1}}, & \frac{\partial \mathcal{L}}{\partial W_\ell} &= A_\ell^T G_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}, & \frac{\partial \mathcal{L}}{\partial b_\ell} &= \mathbb{1}^T G_\ell \in \mathbb{R}^{1 \times p_{\ell+1}}, \\
G_{\ell-1} &= (1 - A_\ell \odot A_\ell) \odot (G_\ell W_\ell^T) \in \mathbb{R}^{n \times p_\ell}, & \frac{\partial \mathcal{L}}{\partial W_{\ell-1}} &= A_{\ell-1}^T G_{\ell-1} \in \mathbb{R}^{p_{\ell-1} \times p_\ell}, & \frac{\partial \mathcal{L}}{\partial b_{\ell-1}} &= \mathbb{1}^T G_{\ell-1} \in \mathbb{R}^{1 \times p_\ell}, \\
G_{\ell-2} &= (1 - A_{\ell-1} \odot A_{\ell-1}) \odot (G_{\ell-1} W_{\ell-1}^T) \in \mathbb{R}^{n \times p_{\ell-1}}, & \frac{\partial \mathcal{L}}{\partial W_{\ell-2}} &= A_{\ell-2}^T G_{\ell-2} \in \mathbb{R}^{p_{\ell-2} \times p_{\ell-1}}, & \frac{\partial \mathcal{L}}{\partial b_{\ell-2}} &= \mathbb{1}^T G_{\ell-2} \in \mathbb{R}^{1 \times p_{\ell-1}}, \\
&\vdots & &\vdots & &\vdots \\
G_1 &= (1 - A_2 \odot A_2) \odot (G_2 W_2^T) \in \mathbb{R}^{n \times p_2}, & \frac{\partial \mathcal{L}}{\partial W_1} &= A_1^T G_1 \in \mathbb{R}^{p_1 \times p_2}, & \frac{\partial \mathcal{L}}{\partial b_1} &= \mathbb{1}^T G_1 \in \mathbb{R}^{1 \times p_2}, \\
G_0 &= (1 - A_1 \odot A_1) \odot (G_1 W_1^T) \in \mathbb{R}^{n \times p_1}, & \frac{\partial \mathcal{L}}{\partial W_0} &= X^T G_0 \in \mathbb{R}^{p_0 \times p_1}, & \frac{\partial \mathcal{L}}{\partial b_0} &= \mathbb{1}^T G_0 \in \mathbb{R}^{1 \times p_1}.
\end{aligned}
$$

https://maziarraissi.github.io/teaching/3_backpropagation/

Modular implementation