# CS124: Deep Learning - IIT Ropar


Nihar Ranjan Sahoo
Prime Minister Research Fellow, IIT Bombay


**Instructors:**

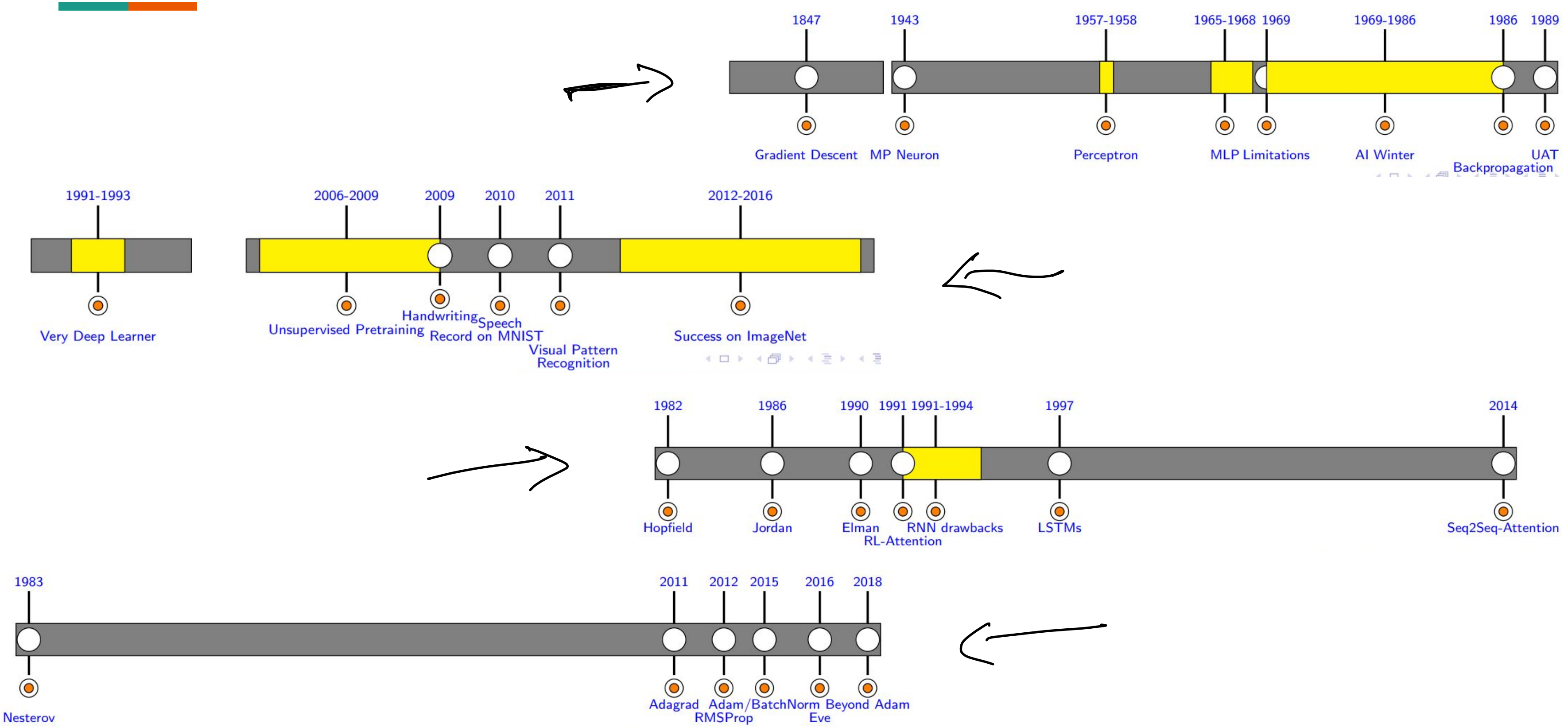
Prof. Mitesh M. Khapra, IIT Madras
Prof. Sudarshan Iyengar, IIT Ropar


NPTEL Problem Solving Session
Week-1
Date: 30/07/2022

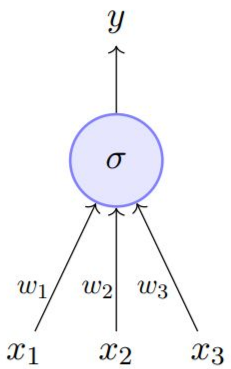
# Outline:

- History of AI
- MP Neuron
- Perceptron
- Problem Solving

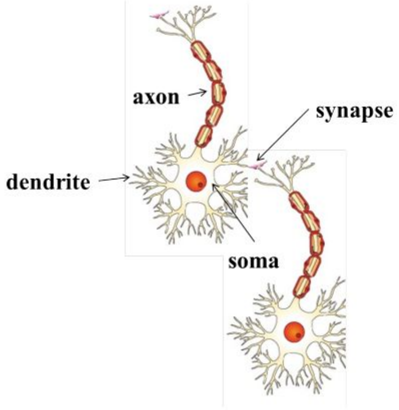
# History


1847
1943
1957-1958
1965-1968
1969
1969-1986
1986
1989
Gradient Descent
MP Neuron
Perceptron
MLP Limitations
AI Winter
UAT
Backpropagation
1991-1993
2006-2009
2009
2010
2011
2012-2016
Very Deep Learner
Unsupervised Pretraining
Handwriting Record on MNIST
Speech
Visual Pattern Recognition
Success on ImageNet
1982
1986
1990
1991
1991-1994
1997
2014
Hopfield
Jordan
Elman
RL-Attention
RNN drawbacks
LSTMs
Seq2Seq-Attention
1983
2011
2012
2015
2016
2018
Nesterov
Adagrad
Adam/BatchNorm
RMSProp
Beyond Adam
Eve

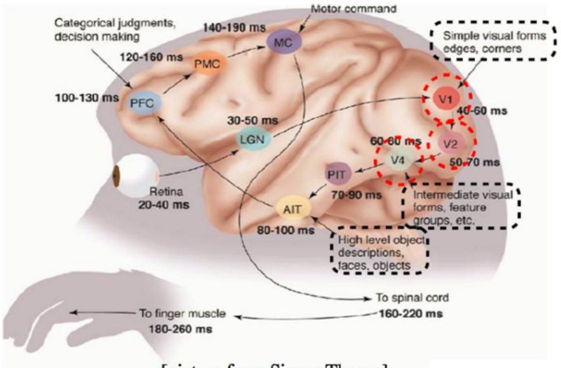
# Neuron



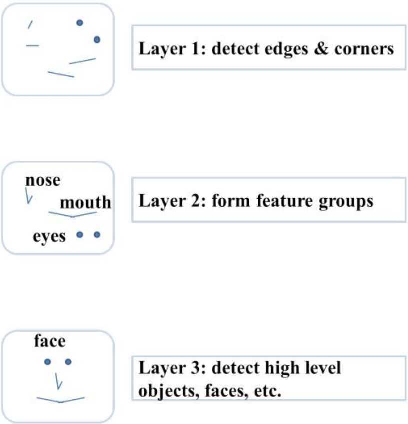

Artificial Neuron




Biological Neurons*

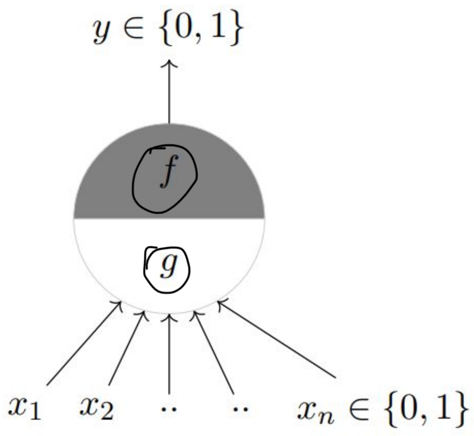



[picture from Simon Thorpe]




- **dendrite:** receives signals from other neurons
- **synapse:** point of connection to other neurons
- **soma:** processes the information
- **axon:** transmits the output of this neuron

# MP Neuron



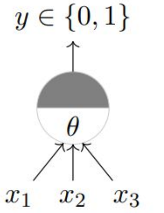

- McCulloch (neuroscientist) and Pitts (logician) proposed a highly simplified computational model of the neuron (1943)
- $g$ aggregates the inputs and the function $f$ takes a decision based on this aggregation
- The inputs can be excitatory or inhibitory
- $y = 0$ if any $x_i$ is inhibitory, else

$$g(x_1, x_2, ..., x_n) = g(\mathbf{x}) = \sum_{i=1}^{n} x_i$$

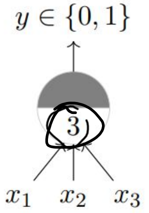
$$\begin{aligned} y = f(g(\mathbf{x})) &= 1 \quad if \quad g(\mathbf{x}) \geq \theta \\ &= 0 \quad if \quad g(\mathbf{x}) < \theta \end{aligned}$$

- $\theta$ is called the thresholding parameter
- This is called Thresholding Logic

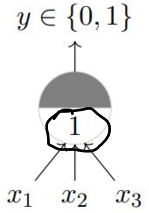
# MP Neuron



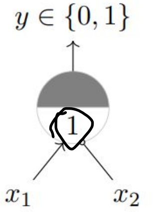

A McCulloch Pitts unit



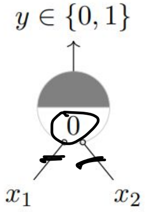

AND function



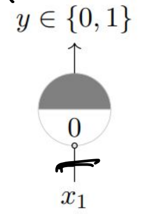

OR function



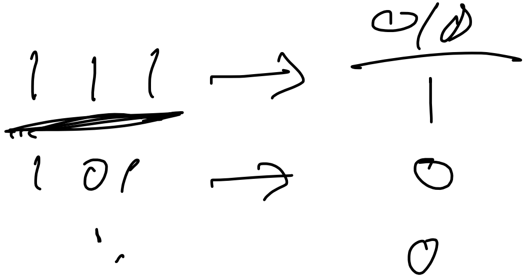

$x_1$ AND $!x_2$*




NOR function



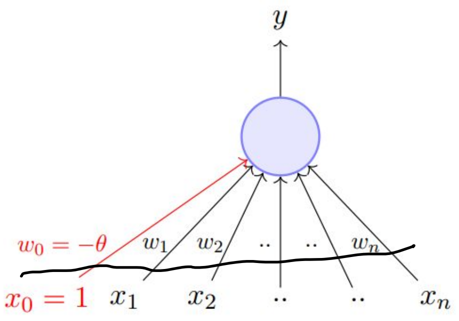

NOT function



The story so far ...

- A single McCulloch Pitts Neuron can be used to represent boolean functions which are linearly separable
- Linear separability (for boolean functions) : There exists a line (plane) such that all inputs which produce a 1 lie on one side of the line (plane) and all inputs which produce a 0 lie on other side of the line (plane)

# Perceptron




A more accepted convention,

$$y = 1 \quad if \sum_{i=0}^{n} w_i * x_i \geq 0$$

$$= 0 \quad if \sum_{i=0}^{n} w_i * x_i < 0$$

$$where, \quad x_0 = 1 \quad and \quad w_0 = -\theta$$

$$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i \geq \theta$$

$$= 0 \quad if \sum_{i=1}^{n} w_i * x_i < \theta$$

Rewriting the above,

$$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i - \theta \geq 0$$

$$= 0 \quad if \sum_{i=1}^{n} w_i * x_i - \theta < 0$$

**McCulloch Pitts Neuron**
(assuming no inhibitory inputs)

$$y = 1 \quad if \sum_{i=0}^{n} x_i \geq 0$$

$$= 0 \quad if \sum_{i=0}^{n} x_i < 0$$

**Perceptron**

$$y = 1 \quad if \sum_{i=0}^{n} w_i * x_i \geq 0$$

$$= 0 \quad if \sum_{i=0}^{n} w_i * x_i < 0$$

# Perceptron

**Algorithm:** Perceptron Learning Algorithm

$P \leftarrow inputs \quad with \quad label \quad 1;$ ✓
$N \leftarrow inputs \quad with \quad label \quad 0;$ ✓
Initialize $\mathbf{w}$ randomly;
**while** $!convergence$ **do**
  Pick random $\mathbf{x} \in P \cup N$ ;
  **if** $\mathbf{x} \in P \quad and \quad \mathbf{w}.\mathbf{x} < 0$ **then** ✓
    $\mathbf{w} = \mathbf{w} + \mathbf{x}$ ;
  **end**
  **if** $\mathbf{x} \in N \quad and \quad \mathbf{w}.\mathbf{x} \geq 0$ **then** ✓
    $\mathbf{w} = \mathbf{w} - \mathbf{x}$ ;
  **end**
**end**
//the algorithm converges when all the inputs are classified correctly

$$cos\alpha = \frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||||\mathbf{x}||}$$

- For $\mathbf{x} \in P$ if $\mathbf{w}.\mathbf{x} < 0$ then it means that the angle ($\alpha$) between this $\mathbf{x}$ and the current $\mathbf{w}$ is greater than 90° (but we want $\alpha$ to be less than 90°)
- What happens to the new angle ($\alpha_{new}$) when $\mathbf{w_{new}} = \mathbf{w} + \mathbf{x}$

$$\begin{aligned} cos(\alpha_{new}) &\propto \mathbf{w_{new}}^T\mathbf{x} \\ &\propto (\mathbf{w} + \mathbf{x})^T\mathbf{x} \\ &\propto \mathbf{w}^T\mathbf{x} + \mathbf{x}^T\mathbf{x} \\ &\propto cos\alpha + \mathbf{x}^T\mathbf{x} \\ cos(\alpha_{new}) &> cos\alpha \end{aligned}$$

- Thus $\alpha_{new}$ will be less than $\alpha$ and this is exactly what we want

- For $\mathbf{x} \in N$ if $\mathbf{w}.\mathbf{x} \geq 0$ then it means that the angle ($\alpha$) between this $\mathbf{x}$ and the current $\mathbf{w}$ is less than 90° (but we want $\alpha$ to be greater than 90°)
- What happens to the new angle ($\alpha_{new}$) when $\mathbf{w_{new}} = \mathbf{w} - \mathbf{x}$

$$\begin{aligned} cos(\alpha_{new}) &\propto \mathbf{w_{new}}^T\mathbf{x} \\ &\propto (\mathbf{w} - \mathbf{x})^T\mathbf{x} \\ &\propto \mathbf{w}^T\mathbf{x} - \mathbf{x}^T\mathbf{x} \\ &\propto cos\alpha - \mathbf{x}^T\mathbf{x} \\ cos(\alpha_{new}) &< cos\alpha \end{aligned}$$

- Thus $\alpha_{new}$ will be greater than $\alpha$ and this is exactly what we want

# Perceptron

Theorem

**Definition:** Two sets $P$ and $N$ of points in an $n$-dimensional space are called absolutely linearly separable if $n+1$ real numbers $w_0, w_1, ..., w_n$ exist such that every point $(x_1, x_2, ..., x_n) \in P$ satisfies $\sum_{i=1}^{n} w_i * x_i > w_0$ and every point $(x_1, x_2, ..., x_n) \in N$ satisfies $\sum_{i=1}^{n} w_i * x_i < w_0$.

**Proposition:** If the sets $P$ and $N$ are finite and linearly separable, the perceptron learning algorithm updates the weight vector $\mathbf{w}_t$ a finite number of times. In other words: if the vectors in $P$ and $N$ are tested cyclically one after the other, a weight vector $\mathbf{w}_t$ is found after a finite number of steps $t$ which can separate the two sets.

# Perceptron

**Algorithm:** Perceptron Learning Algorithm

$P \leftarrow inputs \quad with \quad label \quad 1;$
$N \leftarrow inputs \quad with \quad label \quad 0;$
$N^-$ *contains negations of all points in* $N$;
$P' \leftarrow P \cup N^-$;
Initialize **w** randomly;
**while** !*convergence* **do**
  Pick random $\mathbf{p} \in P'$ ;
  $\mathbf{p} \leftarrow \frac{p}{||p||} \quad (so\ now, ||p|| = 1)$ ;
  **if** $\mathbf{w.p} < 0$ **then**
    $\mathbf{w} = \mathbf{w} + \mathbf{p}$ ;
  **end**
**end**
//the algorithm converges when all the inputs are classified correctly
//notice that we do not need the other **if** condition because by construction we want all points in $P'$ to lie in the positive half space $\mathbf{w.p} \geq 0$

**Proof:**

- Now suppose at time step $t$ we inspected the point $p_i$ and found that $w^T \cdot p_i \leq 0$
- We make a correction $w_{t+1} = w_t + p_i$
- Let $\beta$ be the angle between $w^*$ and $w_{t+1}$

$$cos\beta = \frac{w^* \cdot w_{t+1}}{||w_{t+1}||}$$

$$\begin{aligned}
Numerator &= w^* \cdot w_{t+1} = w^* \cdot (w_t + p_i) \\
&= w^* \cdot w_t + w^* \cdot p_i \\
&\geq w^* \cdot w_t + \delta \quad (\delta = min\{w^* \cdot p_i | \forall i \\
&\geq w^* \cdot (w_{t-1} + p_j) + \delta \\
&\geq w^* \cdot w_{t-1} + w^* \cdot p_j + \delta \\
&\geq w^* \cdot w_{t-1} + 2\delta \\
&\geq w^* \cdot w_0 + (k)\delta \quad (By\ induction)
\end{aligned}$$

# Perceptron

**Proof (continued:)**

$$\textit{So far we have,} \quad w^T \cdot p_i \leq 0 \quad (\textit{and hence we made the correction})$$

$$\begin{aligned}
cos\beta &= \frac{w^* \cdot w_{t+1}}{||w_{t+1}||} \quad (\textit{by definition}) \\
Numerator &\geq w^* \cdot w_0 + k\delta \quad (\textit{proved by induction}) \\
Denominator^2 &= ||w_{t+1}||^2 \\
&= (w_t + p_i) \cdot (w_t + p_i) \\
&= ||w_t||^2 + 2w_t \cdot p_i + ||p_i||^2) \\
&\leq ||w_t||^2 + ||p_i||^2 \quad (\because w_t \cdot p_i \leq 0) \\
&\leq ||w_t||^2 + 1 \quad (\because ||p_i||^2 = 1) \\
&\leq (||w_{t-1}||^2 + 1) + 1 \\
&\leq ||w_{t-1}||^2 + 2 \\
&\leq ||w_0||^2 + (k) \quad (\textit{By same observation that we made about } \delta)
\end{aligned}$$

# Perceptron

**Proof (continued:)**

*So far we have,* $w^T \cdot p_i \leq 0$ *(and hence we made the correction)*

$$cos\beta = \frac{w^* \cdot w_{t+1}}{||w_{t+1}||} \quad (by\ definition)$$

$$Numerator \geq w^* \cdot w_0 + k\delta \quad (proved\ by\ induction)$$

$$Denominator^2 \leq ||w_0||^2 + k \quad (By\ same\ observation\ that\ we\ made\ about\ \delta)$$

$$cos\beta \geq \frac{w^* \cdot w_0 + k\delta}{\sqrt{||w_0||^2 + k}}$$

- $cos\beta$ thus grows proportional to $\sqrt{k}$
- As $k$ (number of corrections) increases $cos\beta$ can become arbitrarily large
- But since $cos\beta \leq 1$, $k$ must be bounded by a maximum number
- Thus, there can only be a finite number of corrections ($k$) to $w$ and the algorithm will converge!

$= m\delta$

$c_1 k = O(k) = w^* \cdot w_0 + k\delta$

② $\sqrt{||w_0||^2 + k}$

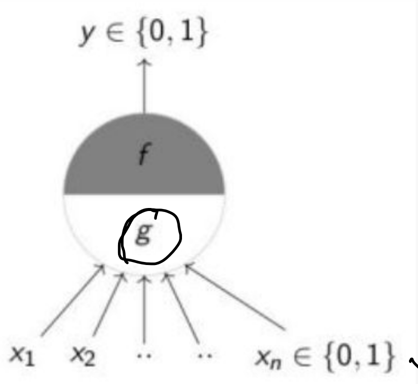
$+1$

$c_3 \sqrt{k} = O(k) \sqrt{c+k}$

$c_2 \cdot \sqrt{k} = O(\sqrt{k})$

# Previous Iteration Problems

1) Given the follwing representation of an MP neuron, What is the role of function ‘g’?



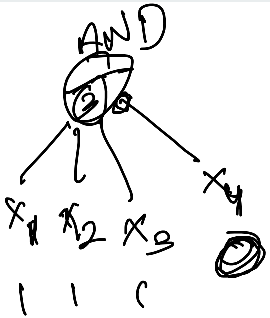

- collect inputs
- aggregate inputs
- takes decision based on inputs
- check validity of inputs

$g = \sum x_i$

$f = g > \theta$

3) What is the thresholding parameter for the MP neuron for the given boolean function,? $(x_1 AND x_2) AND (x_3 AND !x_4)$

- 4
- 3 ✓
- 2
- 1




4) Which of the following Statements is correct?

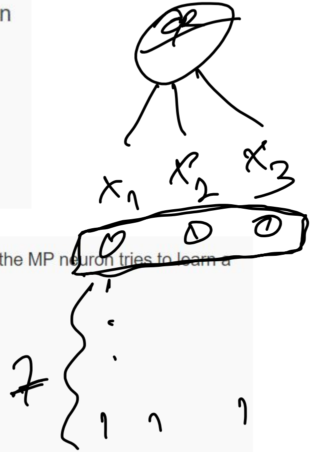
Statement I: Perceptron separates the input space into two sections

Statement II: Only a linearly separable function can be implemented using a single perceptron

- Only I is True
- Only II is True
- Both I and II are True ✓
- Both I and II are False

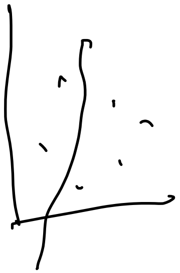
6) Consider an OR function MP neuron with three inputs $x_1, x_2, x_3$. The number of input points can be 8. And the MP neuron tries to learn a decision boundary that is a plane. How many points out of this 8 input points lie below the plane?

- 0
- 1 ✓
- 2
- 3
- 4



7) Pick out the one which best describes the decision boundary that a Mc Culloch Pitt Neuron model learns when the number of inputs is more than three.

- Point
- Line
- Plane
- Hyperplane



9) Which of the following is True for a Perceptron Learning Algorithm?

- weights are initialised with zero values initially
- updating weights takes a finite number of iterations to converge
- weights are updated only by brute force method
- convergence refers to error becoming maximum

min

10) For which of the following gates, the threshold value is 3 for three inputs?

MP neuron

- AND
- OR
- NOR
- XOR