Intuitive Explanation for the weight update eqn:

$$\underline{w}^{(\tau+1)} = \underline{w}^{(\tau)} - \eta \nabla E(\underline{w})$$

[Cauchy, 1847]

$\underline{w}$ → all weights of a neural network

1 layer: perceptron

many layers: MLP

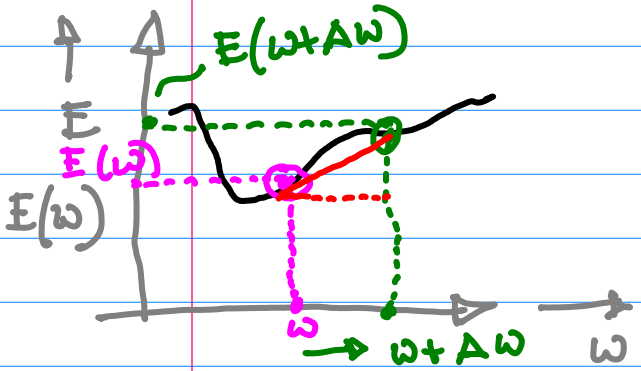
e.g., $w_{ji}^{(1)}$ and $w_{kj}^{(2)}$

∇E → scalar

$\nabla E \rightarrow \begin{bmatrix} \partial E/\partial w \\ \vdots \end{bmatrix}$ vector of all partial derivatives → 'gradient'

$E(\underline{w})$ is an error function of all weights $\underline{w}$

→ This equation is DIMENSIONALLY consistent



$$\text{slope} = \frac{E(w+\Delta w) - E(w)}{(w+\Delta w) - (w)}$$

$$= \frac{E(w+\Delta w) - E(w)}{\Delta w}$$

$$\xrightarrow[\Delta w \to 0]{\lim} \partial E/\partial w$$

Intuition: To get to a local min of the error function, we must go against the gradient/the slope → intuition behind the '−' sign

What is η? ('step size')

Small η → small steps, longer to get to the min

large η → large steps, we may miss the local min

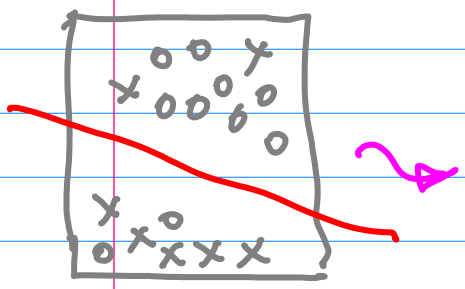
η: step size / learning rate.

A limitation of the Perceptron: seeks a linear decision boundary

$$y(\underline{x}, \underline{w}) = \text{sgn}\left(\sum_{i=1}^{D} w_i x_i + b\right)$$

$$= \text{sgn}(\underline{w}^T \underline{x} + b)$$

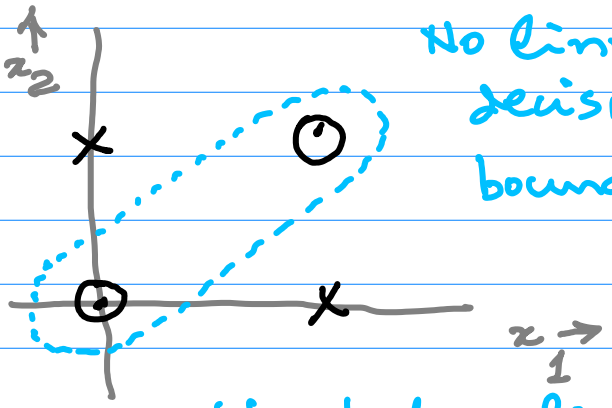
linear eqn



The linear decision boundary is a limitation

'The XOR problem' (Marvin Minsky)

x_2	x_1	$y = x_2 \oplus x_1$
0	0	0 → ⊙
0	1	1 → ×
1	0	1 → ×
1	1	0 → ⊙



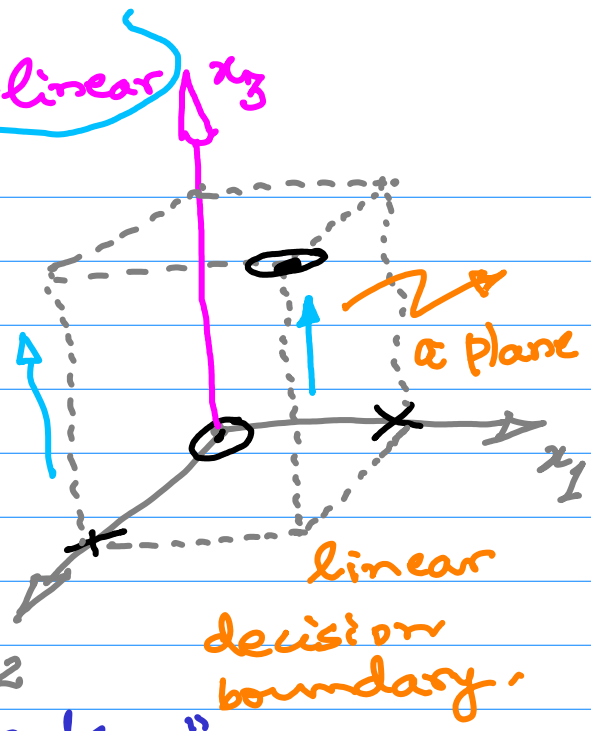
No linear decision boundary

separable, but not linearly separable.

⇒ cannot be solved with a linear classifier such as a basic perceptron.

however, it can be solved by
- Feature transformation + perceptron (by hand!)
- MLP (hidden layer(s)) — this itself has a feature transformation connotation

original features		new feature	typically non-linear
x_2	x_1	$x_3 = x_2 x_1$	$y = x_2 \oplus x_1$
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0



The earlier (1,1) point now "floats up", leading to an infinite number of planes (linear decision boundaries in 3-D) now separating the two classes (much like the concentric circles doughnut 'floating up' over the vada)

The 'Factorisation' in Math/Summation

Where does this appear, and why? Short Answer: Everywhere!

Working Rule:-

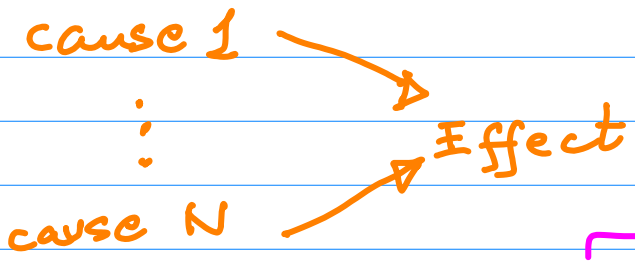
try putting it everywhere, and remove it if it is not required!

→ Probability (conditional/non-conditional)

→ Chain Rule (Calculus)
— total derivate/partial derivative.

PROBABILITY

The general probability factorisation



Bayes Rule

$$p(\text{cause \# i} \mid \text{effect}) = \frac{p(\text{effect} \mid \text{cause \# i})\, p(\text{cause \# i})}{p(\text{effect})}$$

Symmetrical

$$p(A \mid B)\, p(B) = p(B \mid A)\, p(A)$$

$p(A \text{ and } B)$ $\quad$ $p(B \text{ and } A)$

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

a posteriori probability of A — updated probability of A

a priori probability of A

$$x = x + 1$$

$$\Rightarrow x_{new} = x_{old} + 1$$

$$p(A)_{updated} = [\quad] \times p(A)_{initial}$$

$p(A \mid B)$ $\quad$ $p(A)$

→ As such, there is no difference between a conditional probability and an unconditional probability → these are just the updated and initial variants of the same physical quantity.

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

$$p(A|B,C) = \frac{p(B|A,C)\,p(A|C)}{p(B|C)}$$

$p(A|B,C)$: updated prob of A

$p(A|C)$: initial prob of A

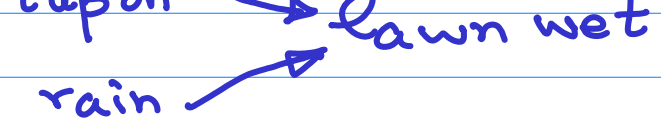
$p(A)$ updated ← Bayes Rule ← $p(A)$ initial

Where do the summations come in

$$p(\text{effect}) = \sum_j p(\text{effect}|\text{cause}\#j)\,p(\text{cause}\#j)$$

$\forall j$ → safely put a summation for all j's

Example: →



Suppose we find the lawn to be wet in the morning. $p(\text{rained last night} \mid \text{lawn wet in the morning})$

[Bayes Rule]

$$p(\text{rain} \mid \text{wet}) = \frac{p(\text{wet} \mid \text{rain})\, p(\text{rain})}{p(\text{wet})}$$

$$= \frac{p(\text{wet} \mid \text{rain})\, p(\text{rain})}{p(\text{wet} \mid \text{rain})\, p(\text{rain}) + p(\text{wet} \mid \text{tap on})\, p(\text{tap on})}$$