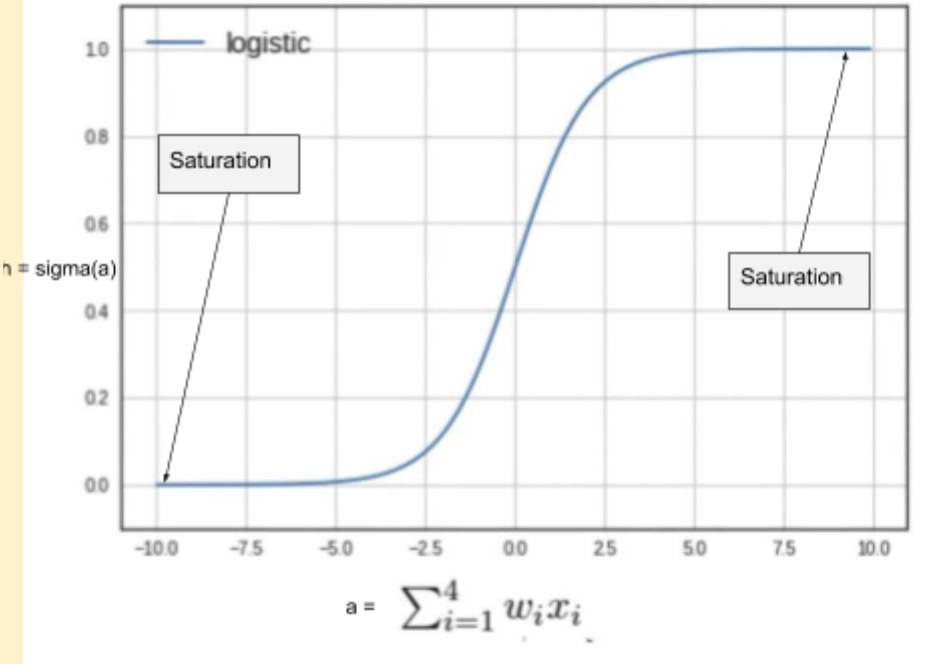
PadhAI: Activation Functions & Initialization Methods

One Fourth Labs

Zero centered functions

1. Why do logistic neurons saturate?
   a. Consider a pre activation function: $a = \Sigma^{n}_{i=1} w_n x_n$
   b. And the activation function $h = \sigma(a)$



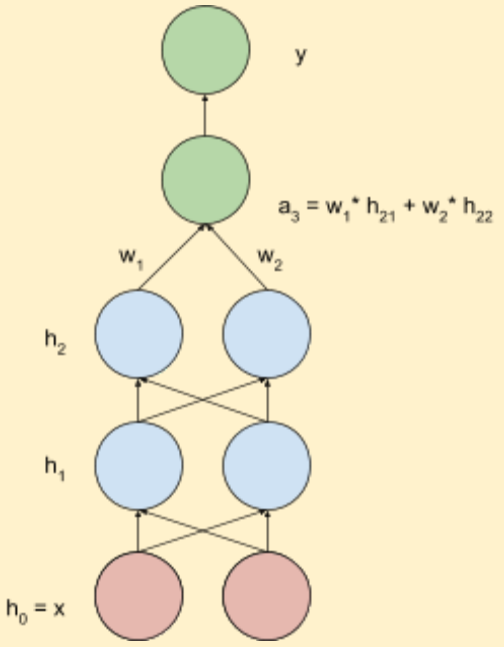
   c. In cases where the weights are initialised to very high or very low values, the weighted summation term (a) will become very large or very small (very negative).
   d. This could result in the neuron attaining saturation
   e. Remember to initialise the weights to small values
2. Another shortcoming with the logistic function is that it is not zero centered
   a. Zero centered: The function is spread out equidistant around the 0 point, i.e. it takes an equal number of positive and negative values.
   b. The logistic function ranges from 0 to 1
   c. The tanh function is a zero centered sigmoid function

PadhAI: Activation Functions & Initialization Methods

One Fourth Labs

---

3. Consider the simple neural network with logistic sigmoid neurons



   a. Consider the following gradients

   b. $\nabla w_1 = (\frac{\partial L(w)}{\partial y} \frac{\partial y}{\partial h_3} \frac{\partial h_3}{\partial a_3}) * \frac{\partial a_3}{\partial w_1} = (\frac{\partial L(w)}{\partial y} \frac{\partial y}{\partial h_3} \frac{\partial h_3}{\partial a_3}) * h_{21}$

   c. $\nabla w_2 = (\frac{\partial L(w)}{\partial y} \frac{\partial y}{\partial h_3} \frac{\partial h_3}{\partial a_3}) * \frac{\partial a_3}{\partial w_2} = (\frac{\partial L(w)}{\partial y} \frac{\partial y}{\partial h_3} \frac{\partial h_3}{\partial a_3}) * h_{22}$

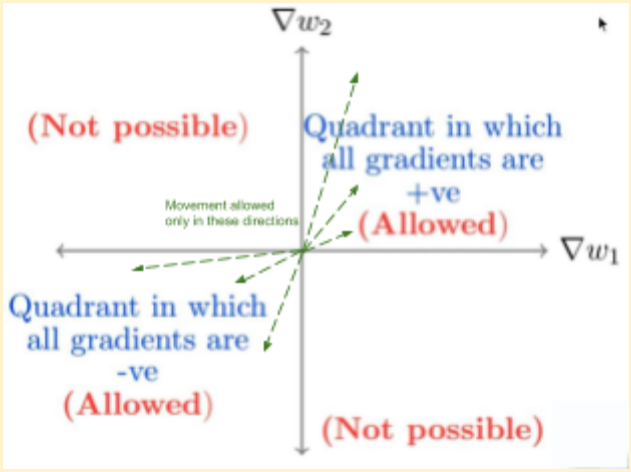
   d. The bracketed terms are common

   e. Both h_{21} and h_{22} are outputs of the logistic function, so they are always positive (i.e. ranging from 0 to 1)

   f. Due to this, at all times, both ∇w_1 and ∇w_2 will always be of the same sign, either positive or negative. They cannot be different from each other since the bracketed part is common between them and the logistic function output is always positive

   g. The gradients w.r.t all the weights connected to the same neuron are either all +ve or all -ve

   h. Thus, this limits the directions in which the weights can be updated



4. Thus, we cannot arrive at the local minima as fast as possible by moving in all directions.
5. Also, logistic function is computationally expensive because of e^x