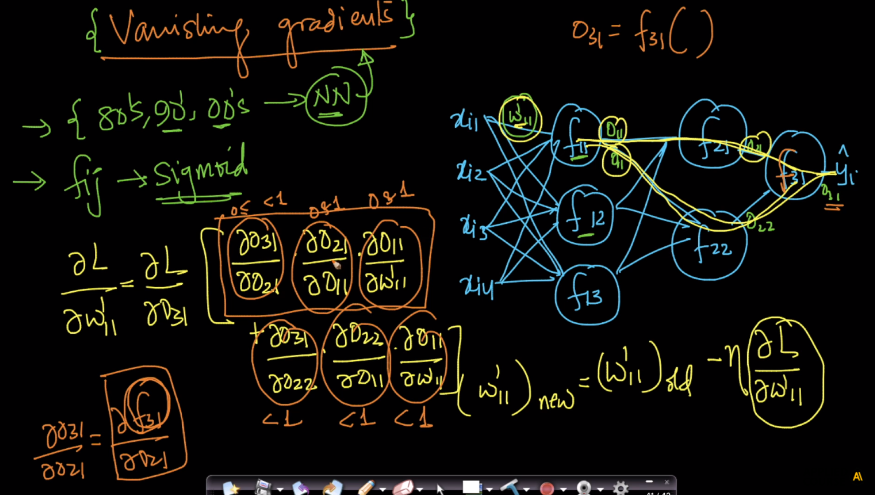
**Vanishing Gradient problem**

Major problems why NN not much used in 80’s, 90’s is because of vanishing gradient.

Vanishing gradient is occur because of activation functions like sigmoid and tanh.

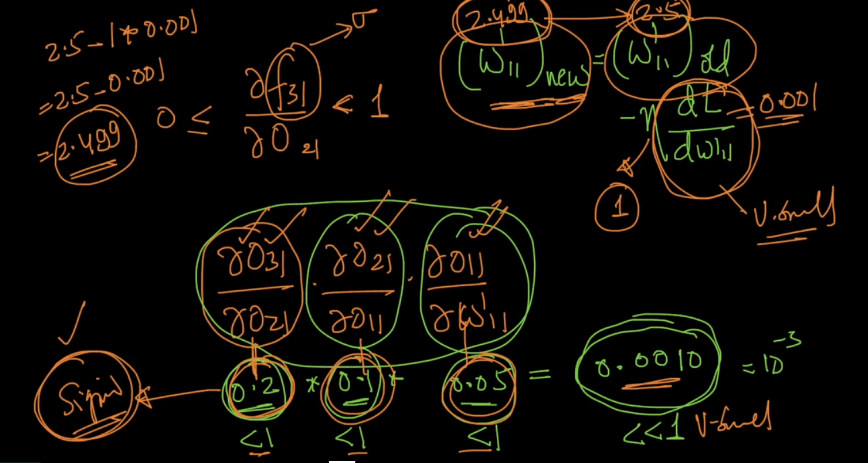
Lets take an example of sigmoid activation function to explain it, as output of this function impacts loss like o/p of sigmoid activation function f31() is O31 which impacts loss, therefore derivative of output O31 is nothing but derivative of f31().

Now as we already know that derivative of sigmoid lies b/w 0 and 1. Now in this all activation functions are sigmoid therefore value of derivative is less than 1 and multiplication of that value gives very low value and sum of very low value doesn’t impacts much.



Now if output of all derivatives is less which gives final derivative ans very very low therefore updating new weights by using old weights and derivative is not much ex : if old w is 2.5 and derivative is 0.001 and eta is 1 then new w is 2.499 which is not much different than old w this is because of low derivative value and this is called as vanishing gradient.

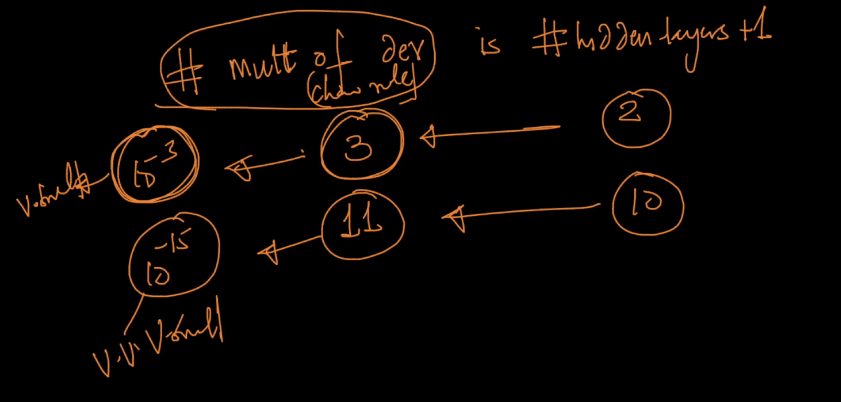
Therefore vanishing gradient means derivative we have with respect to any weights this derivatives becomes very very small because of chain rule and multiplication and because of using sigmoid and tanh because their derivative values is less than 1

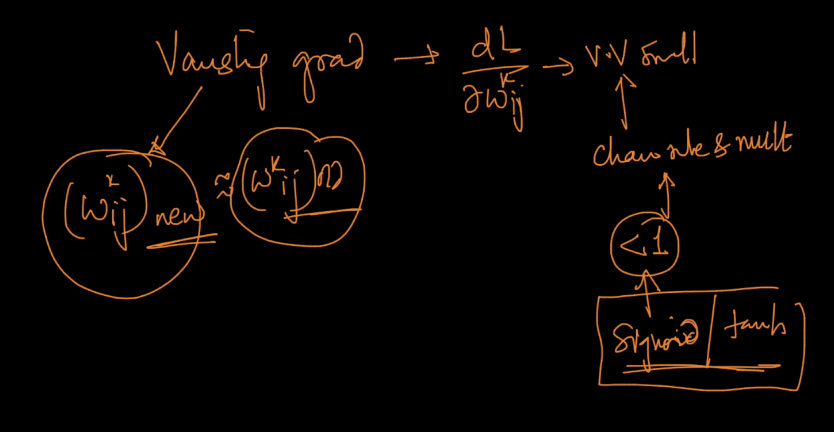


And in multiplication of derivatives in chain rule, no. of multiplication of derivative is equal to no. of hidden layers + 1

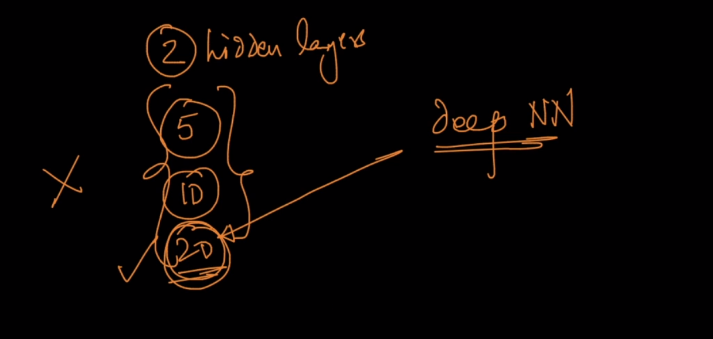
As we saw above examples no. of hidden layers are 2 therefore no. of multiplication of derivative is 3 which gives very small value

If no. of hidden layer is 10 then no. of multiplication of derivative is 11 which gives very very small value

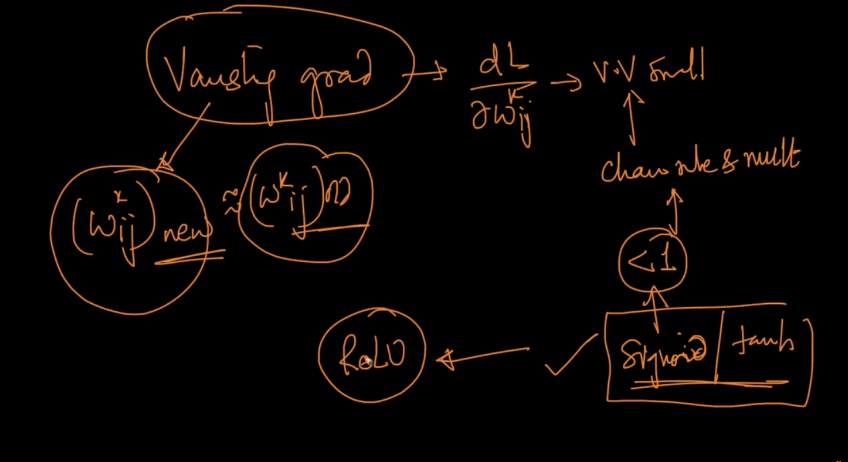




This is the case of only 2 hidden layers suppose the case of 5,10,20 hidden layers then training will become very difficult as no. of hidden layers increases values of derivatives decreases drastically.



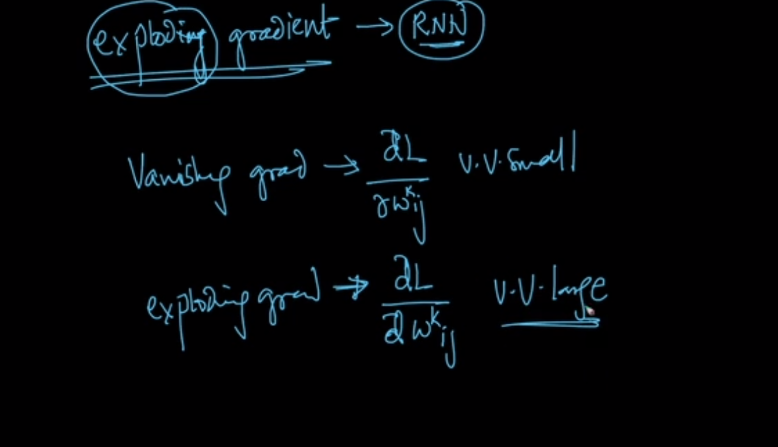
To avoid such problems we use RELU



**Exploding gradient :**

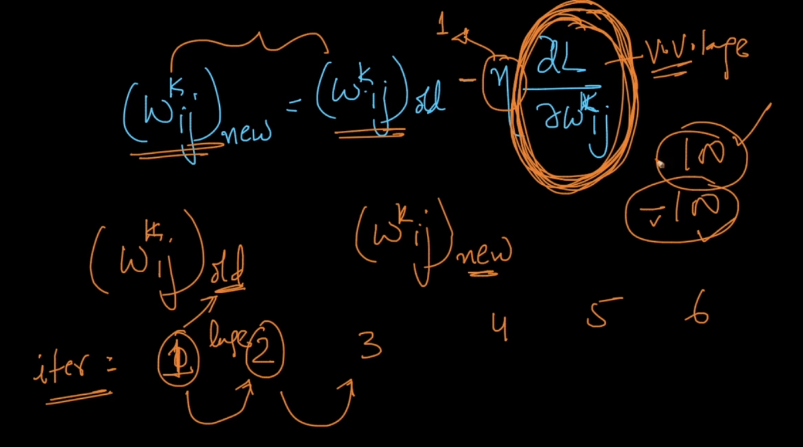
This majorly occurs in recurrent neural network, as vanishing gradient returns very very small value, exploding gradient returns very very large values.

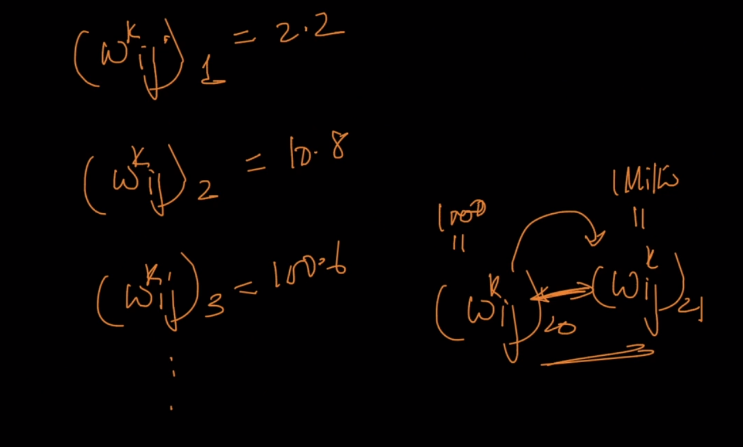
It occurs when depth of network is large.



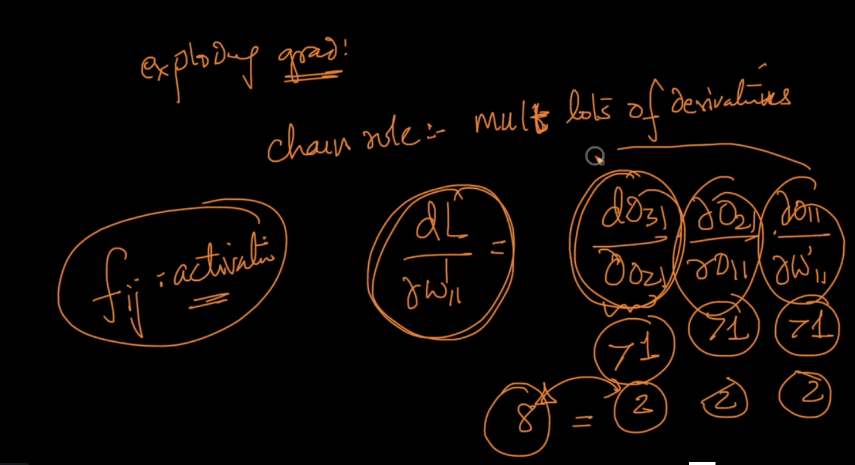
Now lets see what is the effect of exploding gradient.

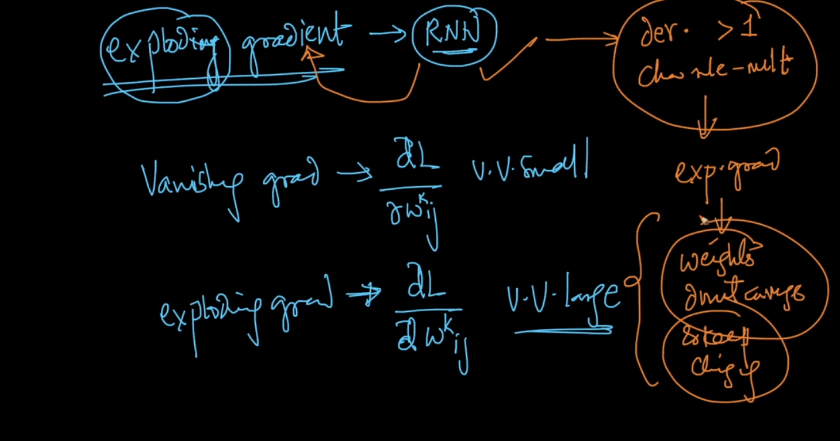
As derivative in exploding gives very very large value therefore new w changes very much and therefore because of it weights doesn’t converge and keeps changing.



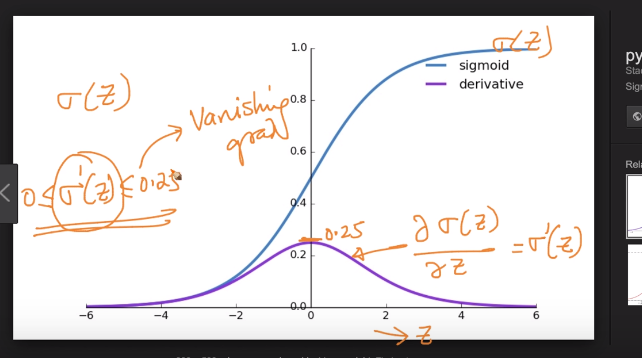


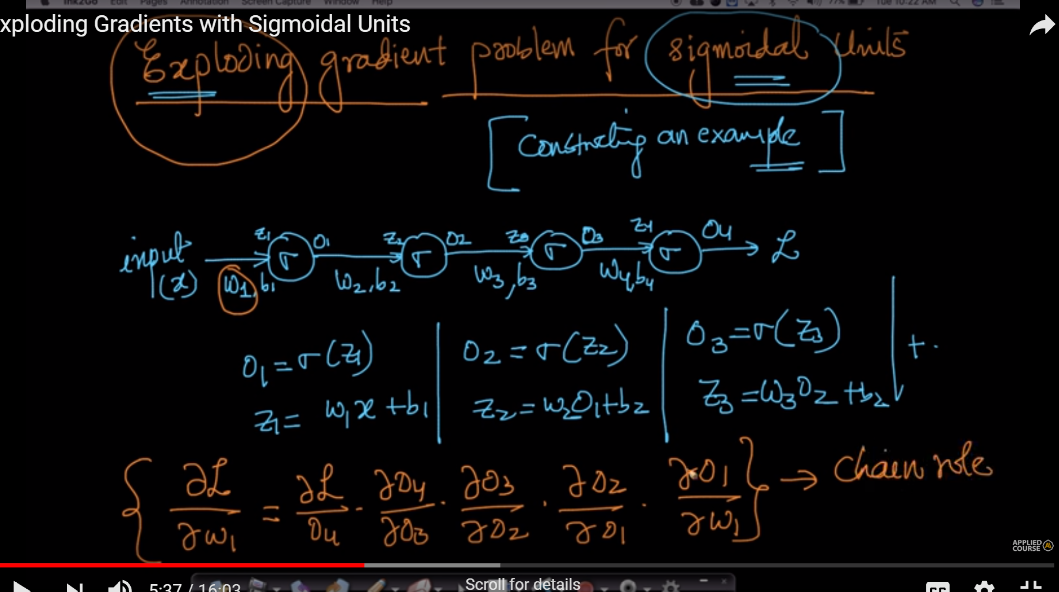
Exploding gradient occurs because of chain rule (multiplication of derivatives) suppose some function returns values more than 1 therefore multiplication of them give large value and therefore while updating new weights it changes very much and therefore weight doesn’t converge and keep changing.

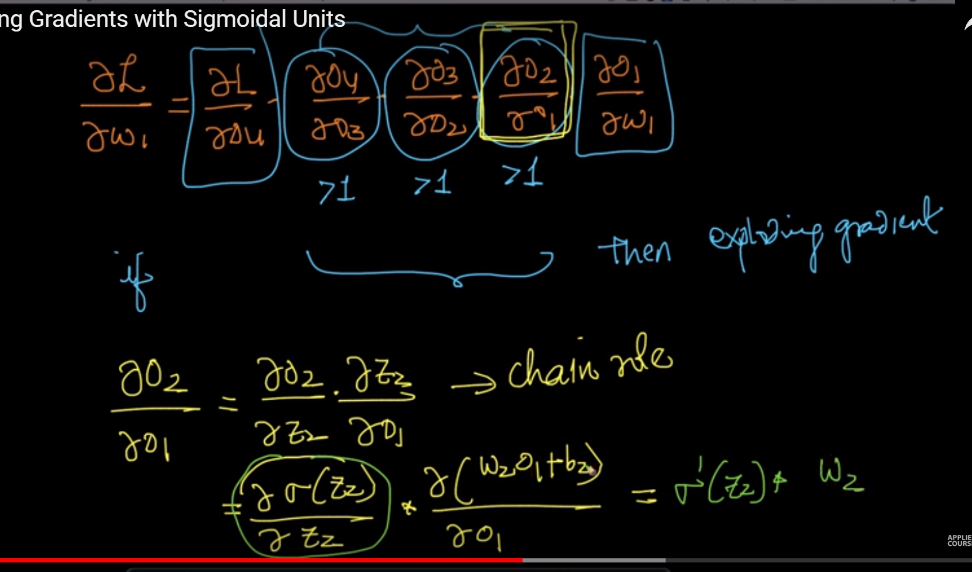


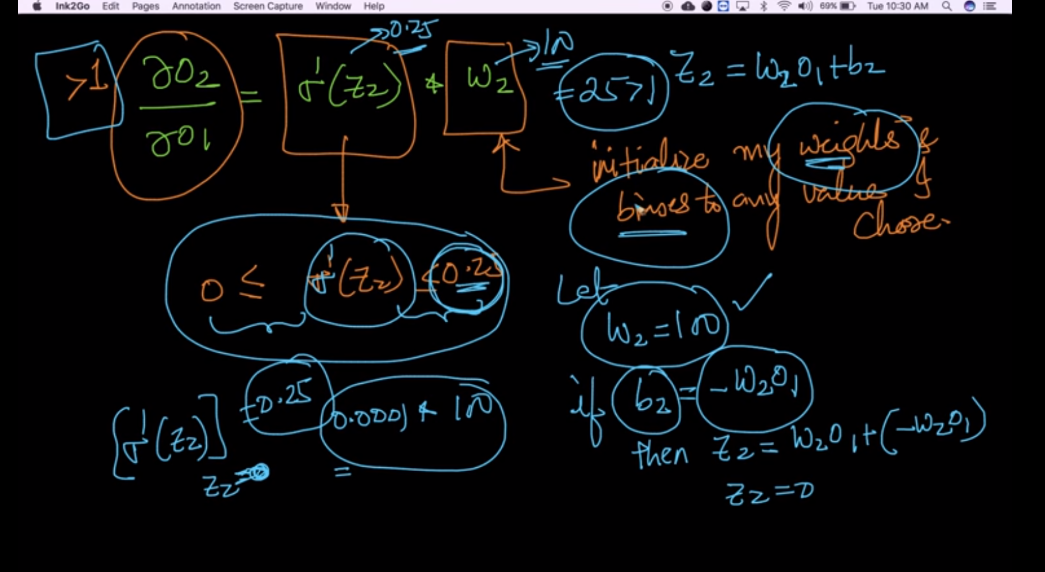


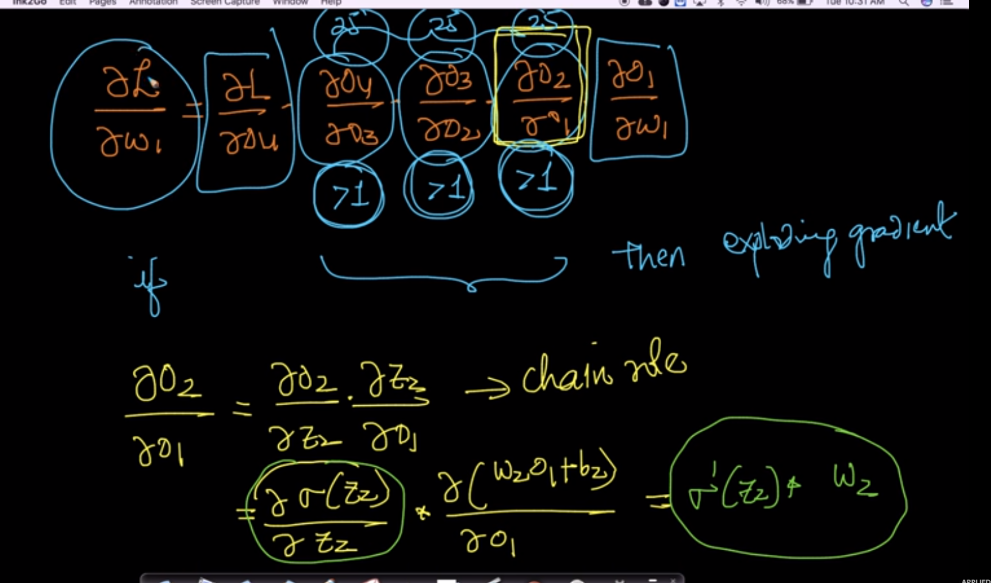
### **How can exploding gradients occur with sigmoid units :**

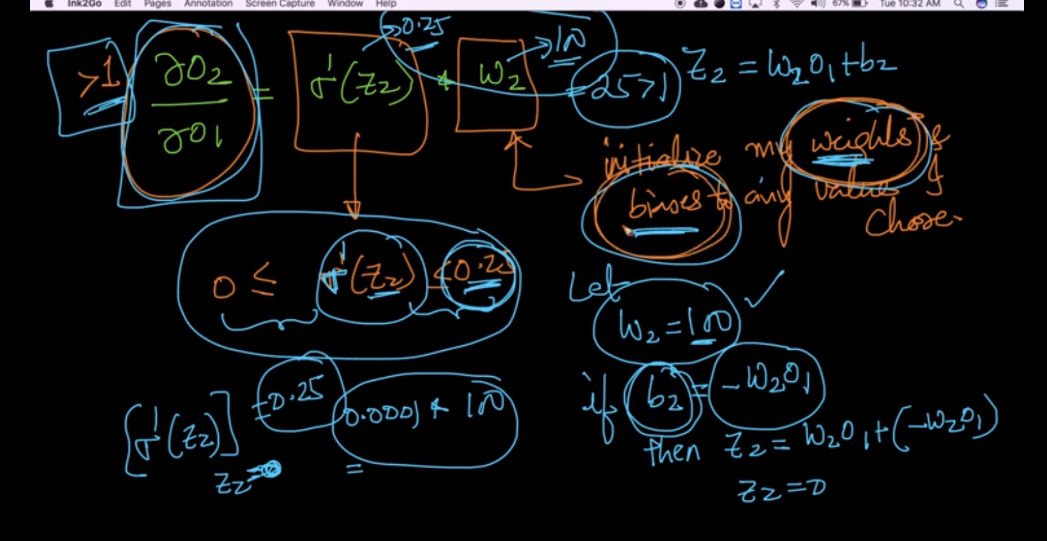




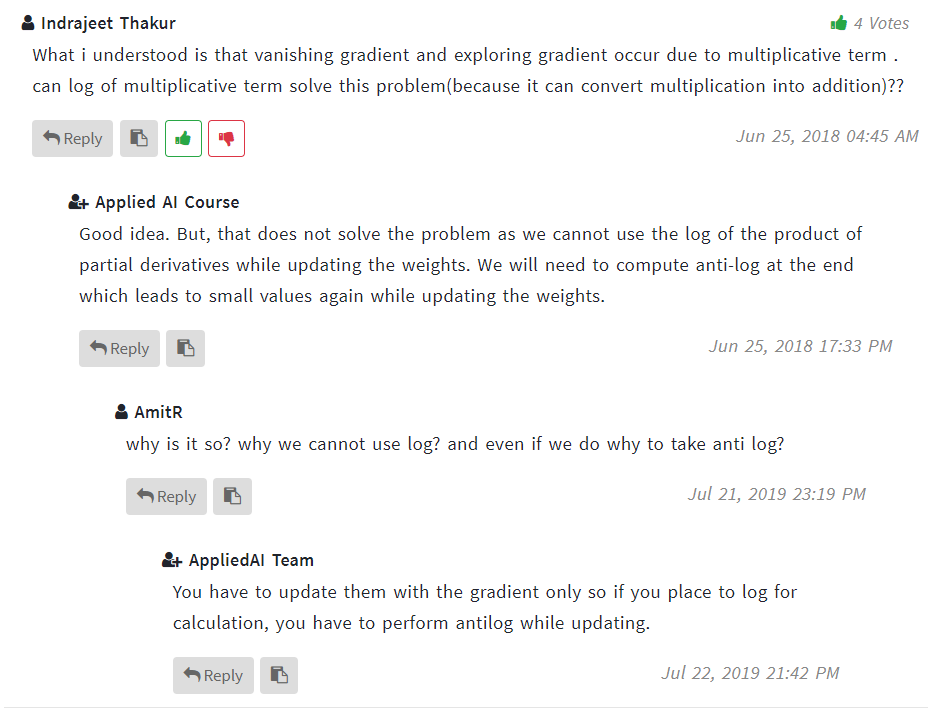


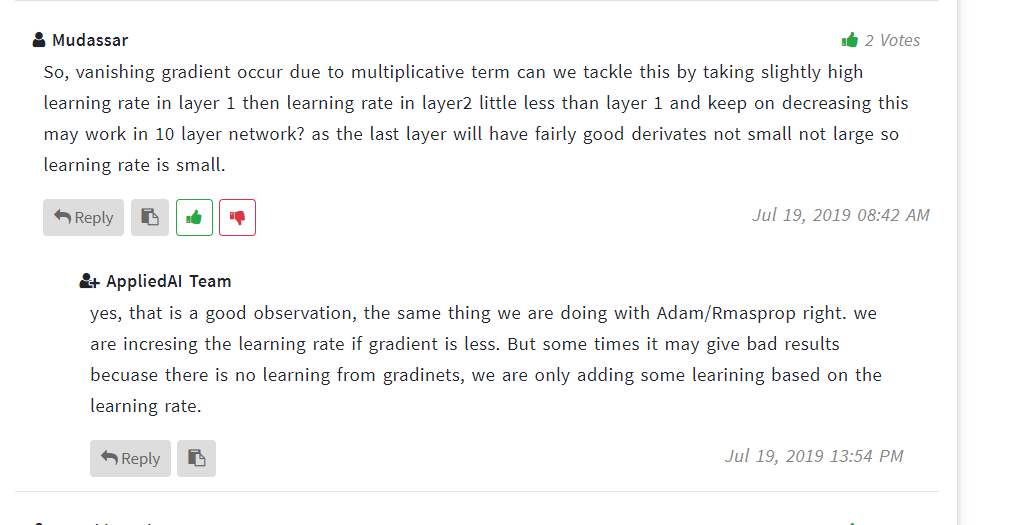


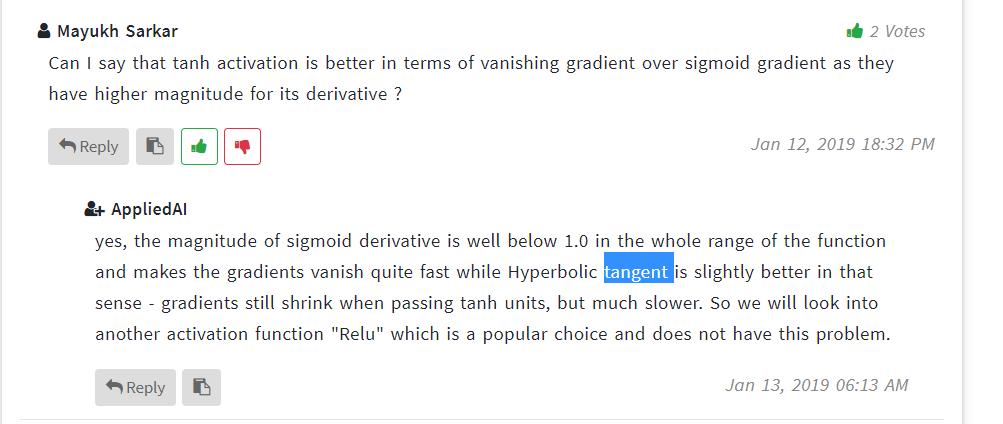


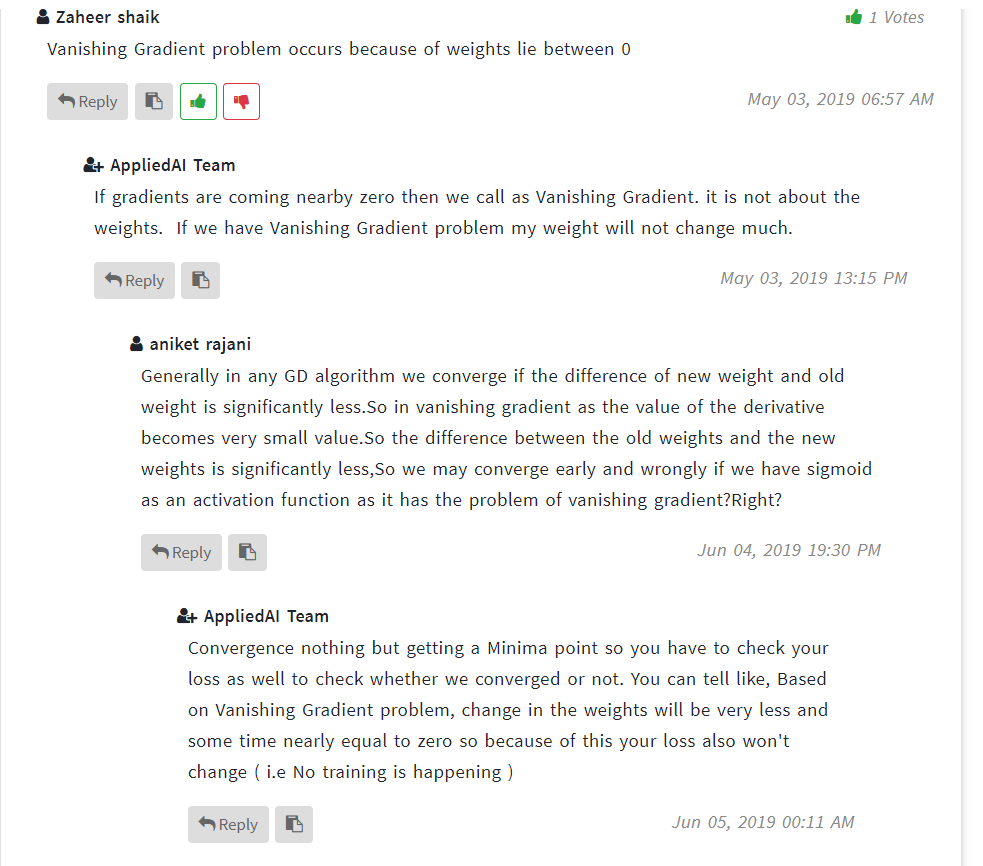


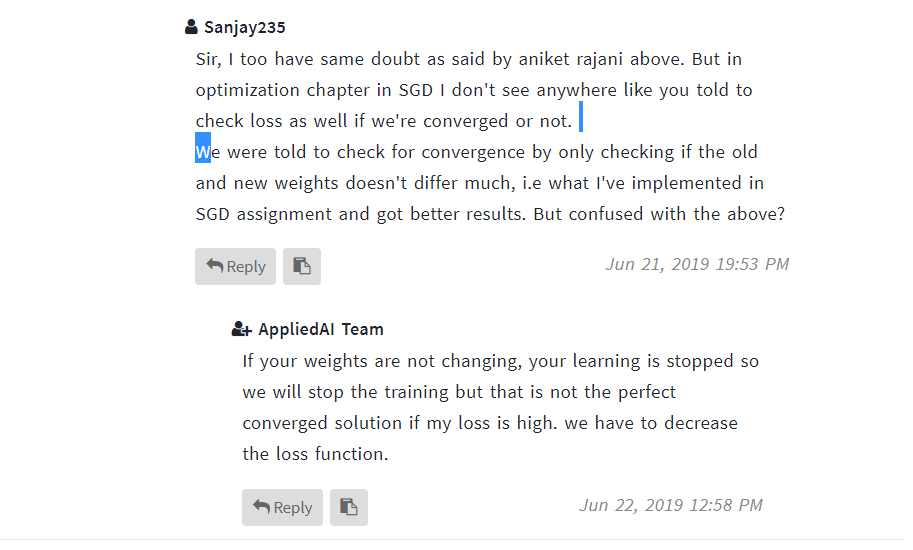
**Comments :**

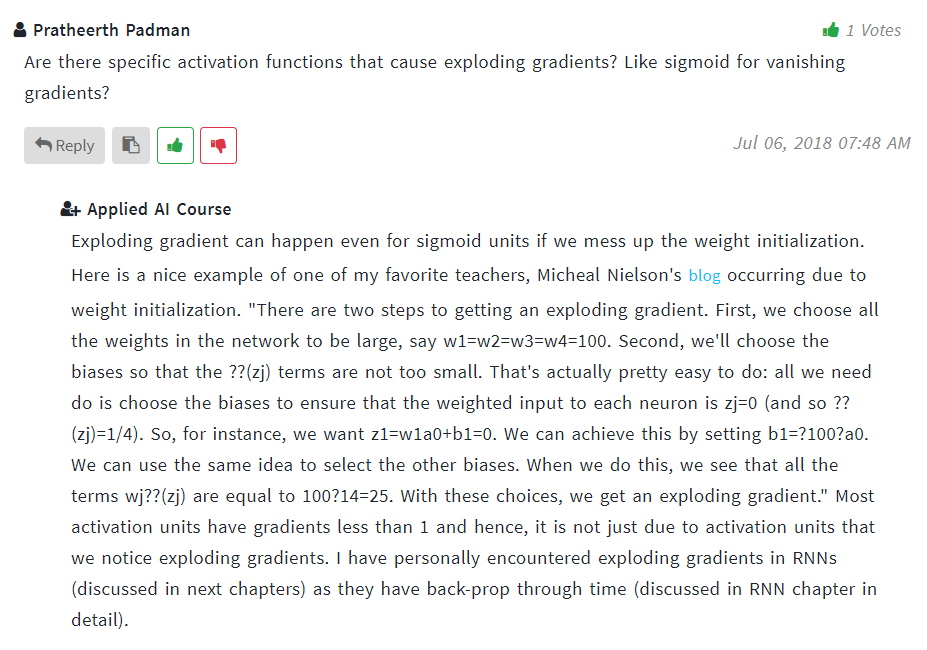












**Link :** <http://neuralnetworksanddeeplearning.com/chap5.html>