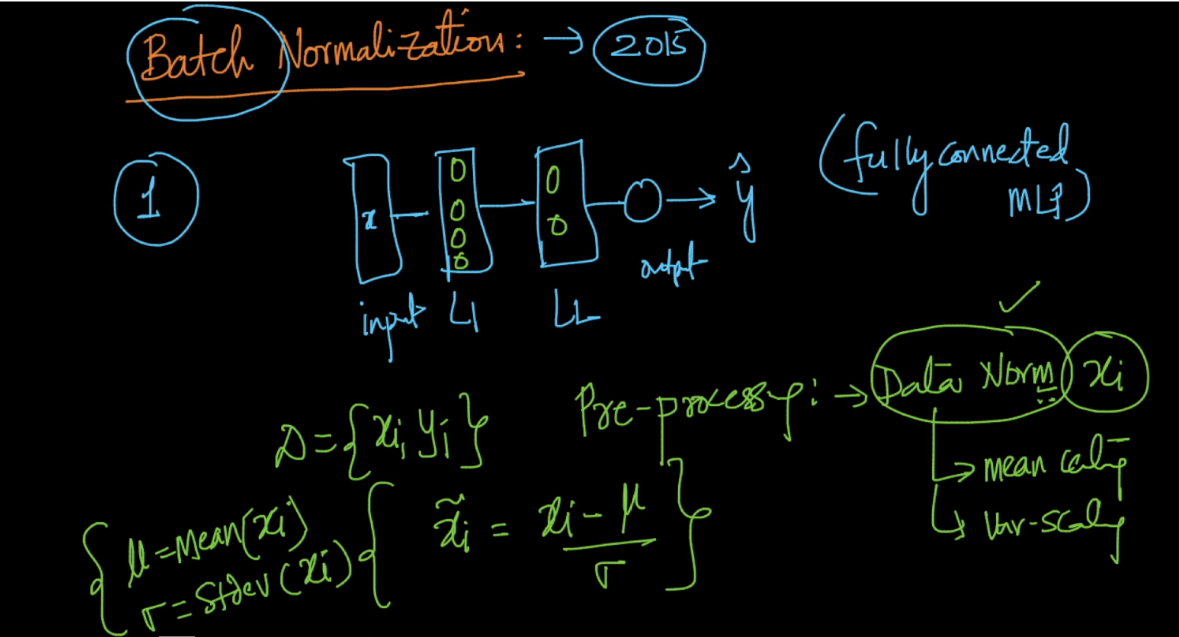
**Batch Normalization**

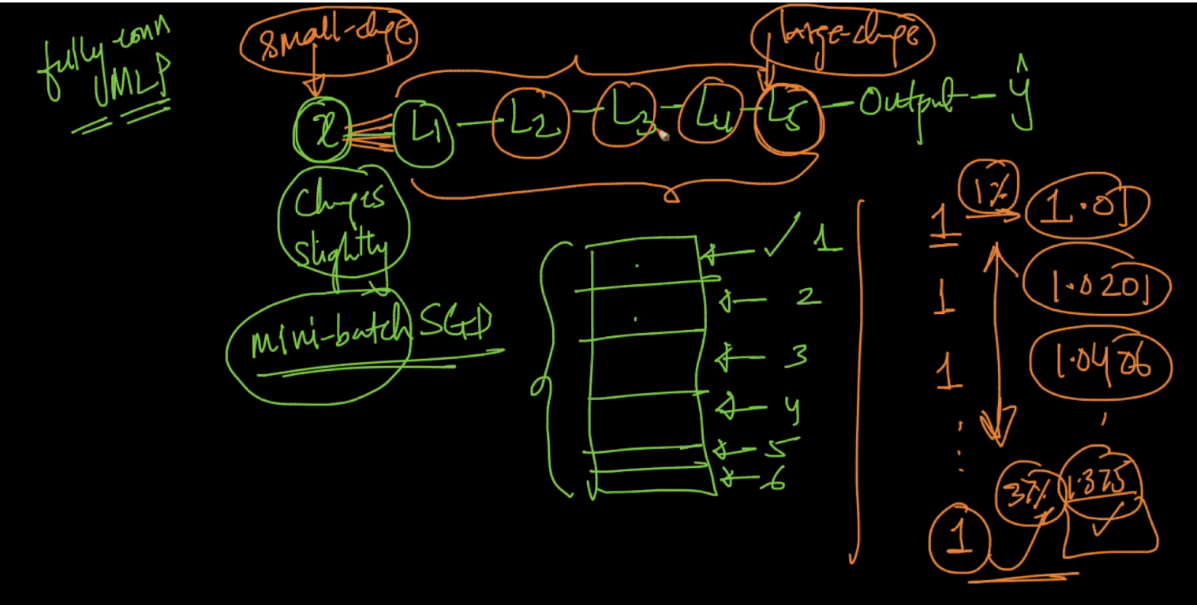
In fully connected MLP network we provide input with pre-processing ie. We do data normalization(mean centering, variance scaling)



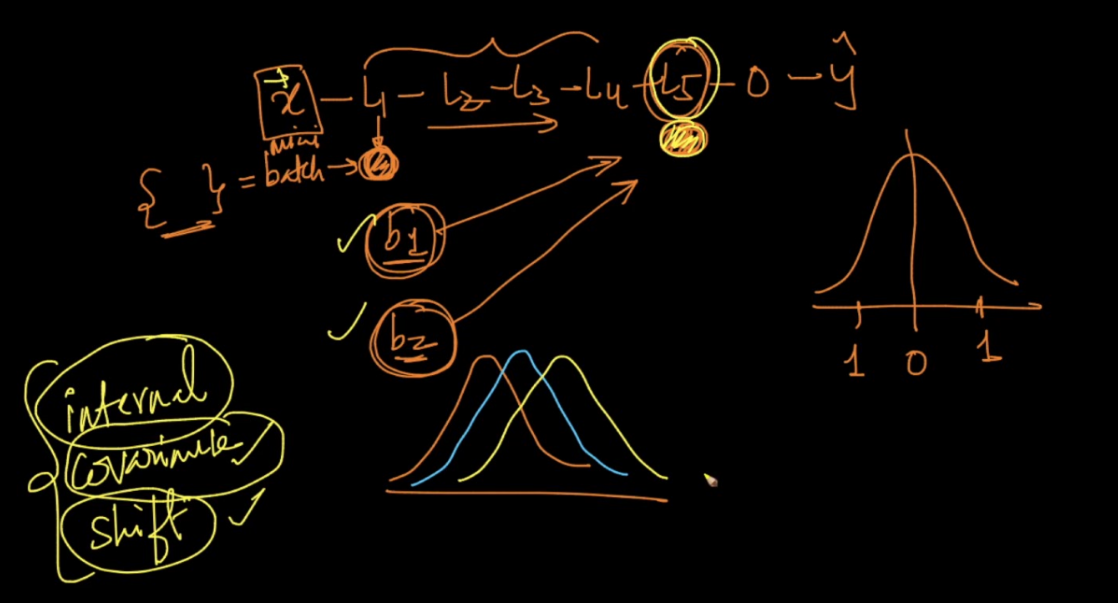
Below image shown multi layer perceptron n we provide mini-batch sgd.

what happen even we provide normalize data by this small change in input can make a large change in layer 5 because it goes through so many computation i.e through layers.

As seen below in example if we provide 1 as an input it remains same till 5th layer but if we provide 1.01 then it changes drastically 37% till 5th layer.

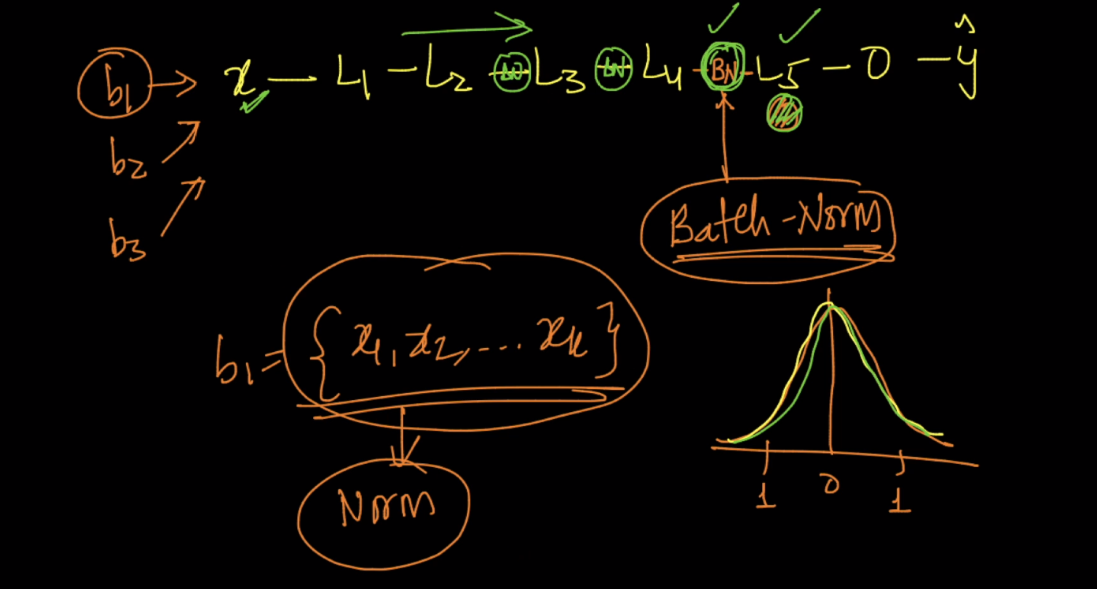


Now what happen by this drastic changes is the distribution of data is changes means at layer 5 b1 have some dist. then b2 have some other distrib. And so on and this problem is known as internal covariance shift (it internal because within the network we are seeing this problem, covariance means generalization of variance to a multiple vector, covariance shift because here variance is shifting and also mean is shifting)



Now the solution for this internal covariance shift is Batch normalization what it means is we can think of it as a layer we provide batch normalization b/w layers and it normalize data and thus provide same distribution data of every batch to next layer.

We can provide it before last layer, second last layer ,anywhere but it works mainly in deep neural network.



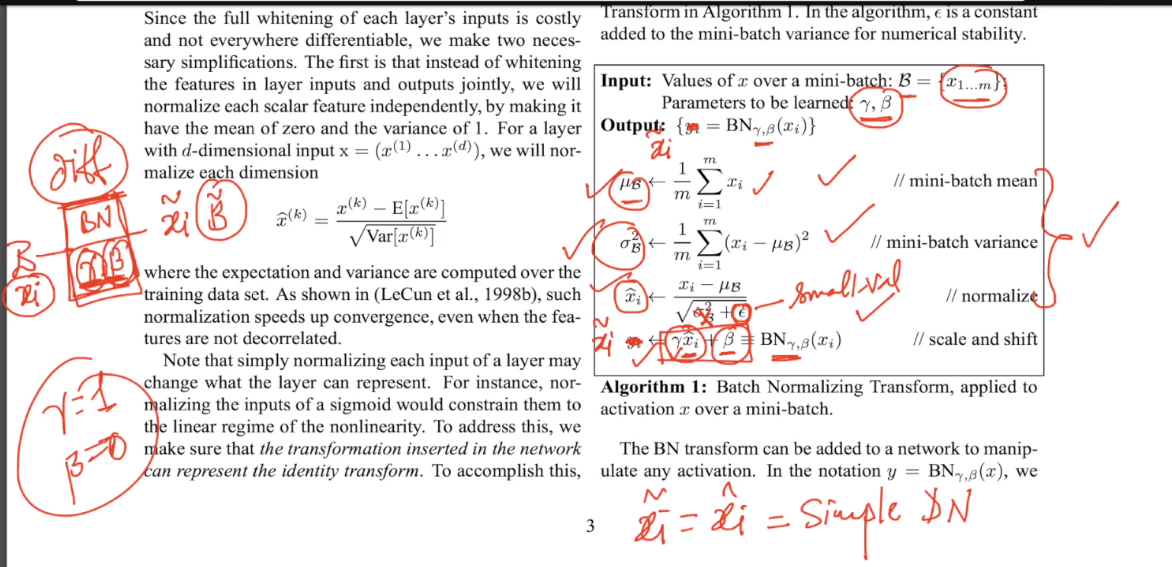
Batch Normalization algorithm:

It uses two parameters gamma and beta.

Algorithm is shown below, at normalize step it uses epsilon it is small value and it is used because if variance is 0 then it shouldn’t be like /0

And final output x\_tilde is given by gamma\*x\_hat (this means scaling because we are multiplying ) + beta (shifting because we are adding beta)

And batch normalization learned parameter gamma and beta by back propogation, and we can learn it by back propogation because here everything is differentiable(mean, variance, normalize, scale and shift)

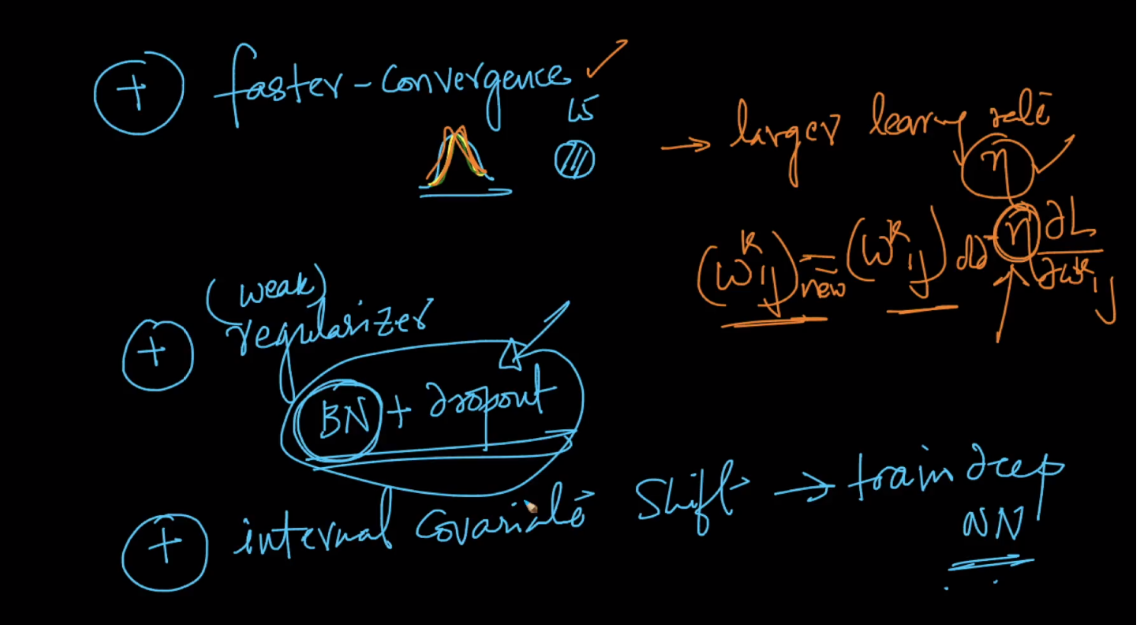


Advantages of batch normalization :

1. Faster convergence :

This is because if we provide same distribution data batch to layers then we can take larger learning rate eta and by this it updates fastly.

1. It’s a weak regularizer we can use BN+dropout for good regularizer
2. It remove internal covariance shift therefore we can train deep neural network.



Comments :

