

*The first gradient part is wrong. Right notes:*

*If you have a n-examples batch, each example can have one set of gradient, after calculating each gradient for each example, simply sum them up and take the average value of the sum, which is the final gradient of current batch.*

*A single neuron can be used to implement a binary classifier (e.g. binary Softmax or binary SVM classifiers)* *that is, activation function of each node can use the model of loss function.*

*Usual procedure:*

1. *Forward propagation, get overall scores*
2. *Use scores to derive overall loss*
3. *Use loss to derive loss function and implement back propagation*
4. *Derive gradient of each parameter W from back propagation*
5. *Implement gradient descent*