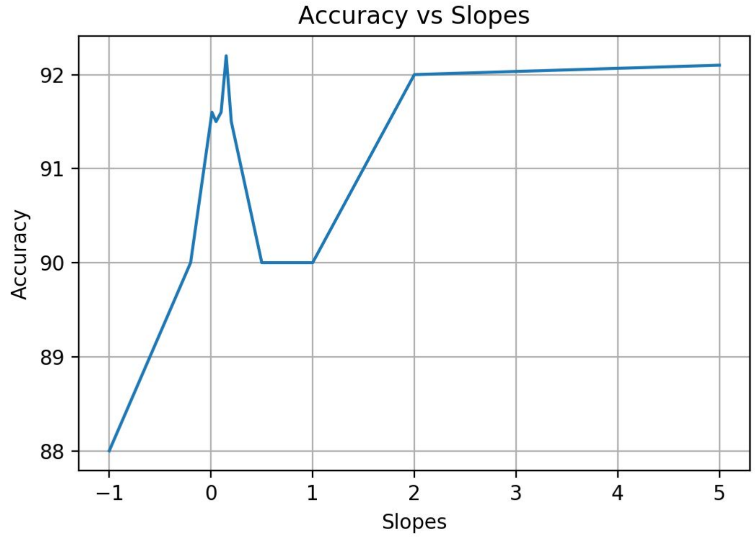
Mehmet Saygın Seyfioglu

1) The following plot shows the results obtained with different slopes in Leaky Relu function. It is obtained by training the default model with the default parameters. (0.01 learning rate, 1000 => 100 => 10 model and sgd, 20 epochs)



It seems the accuracy is decreasing as the slope goes to 1 because the network becomes completely linear as the slope is 1 in the leaky part. The highest accuracy is achieved with a slope of 0.15.

When the slope gets greater than 1, the negative part's derivative gets above to the positive part's derivative. If it stays around 1, the previous gradients becoming indiscriminative as they scaled with similar values, thus slowing down the learning process. When it gets >>1 it largely scales the previous gradient (for this reason we don't want large weights either, thus regularizing the network) thus having exploding gradients that cause overshooting. In other words, as the negative part having a slope that is >>1 we are basically widening the gradient of activation so we are adding more uncertainty, and the network cannot be trained because the uncertainty in the decision boundary becomes too large.

As for the negative case, the activation gradient becomes symmetrical (for the -1 case, it completely becomes symmetrical) Hence, the network tends to cancel out the information thus ending up the slower learning process.

Theoretically, when the slope is 1, the activation becomes linear, therefore not providing any nonlinearity to the network. (x=y for -inf to + inf and constant derivative of 1 for -inf to +inf)

2) If trained with weight decay the slope generally converges to a very small value of ~0.02. If weight decay is not used, after 20 epochs, the network generally provides %92 accuracy and the slope parameter converges to 0.15 which corresponds with what is found in the first question.

3) I tested out several network architectures in various depth and width. The best I get is 5 layer architecture of 400 => 400 => 200 => 100 => 10. This network reaches 98.5% test accuracy with momentum sgd and 0.08 learning rate after 30 epochs (it reaches 97% test accuracy in 4 epochs).