Assignment 4: Deep Learning Basics
Marissa Como
Zach Corse
Vasu Mahesh

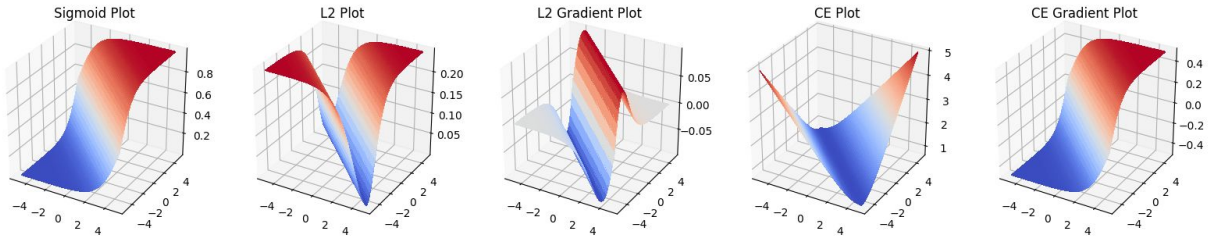
1. Plot Loss and Gradient

1) What's the difference between cross-entropy loss and L2 loss?
The L2 loss has a sharp valley where weight and bias are equal (these offset one another). The CE loss has a similar valley, but does not plateau for weights and biases far from the global loss minimum.

2) What's the difference between the gradients from cross-entropy loss and L2 loss?
The L2 gradient plot has additional inflection points, as well as local minima and maxima. The CE gradient plot, by contrast, is merely an offset sigmoid. Additionally, the L2 loss gradient approaches zero for weights and biases far from equilibrium.

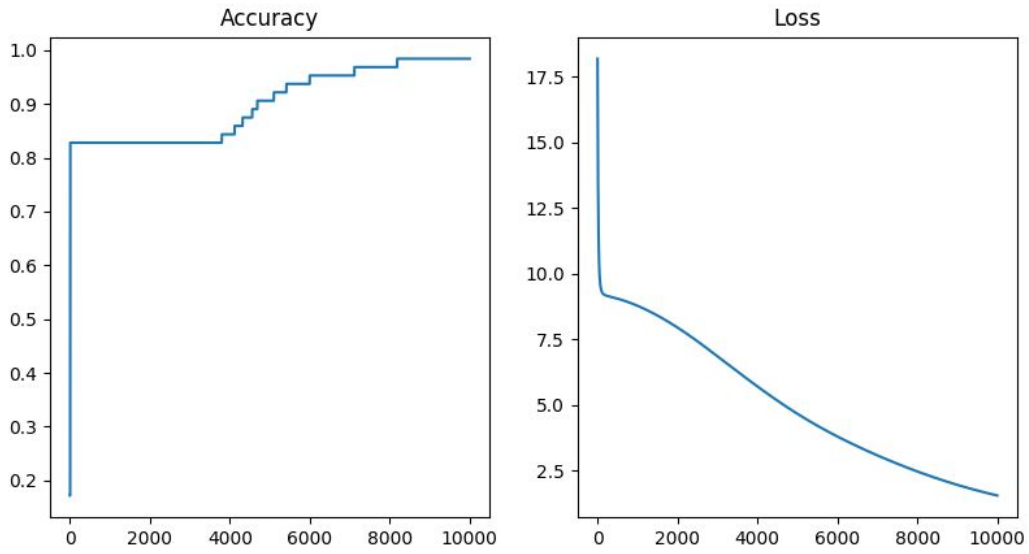
3) Predict how these differences will influence the efficiency of learning.

The gradient of L2 loss disappears after some error. This could lead to the neuron being unable to learn because updates would become too small. CE loss approaches a constant gradient as the total loss moves away from equilibrium, which means that gradient descent should always be possible. Conversely, using L2 error, the neuron runs the risk of incorrect labeling if initial weights for the data set are far from optimal.

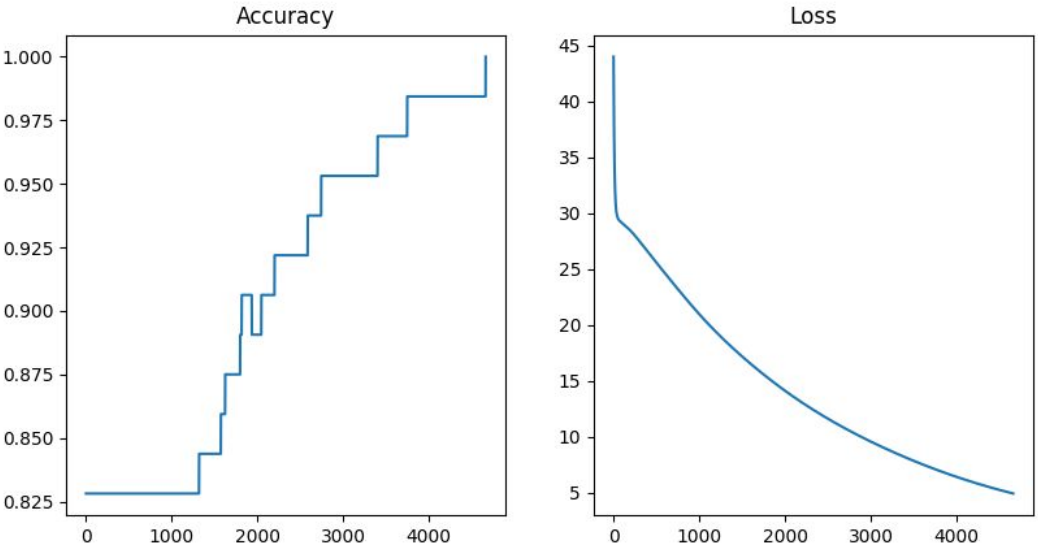


2. Experiment with Fully Connected Network

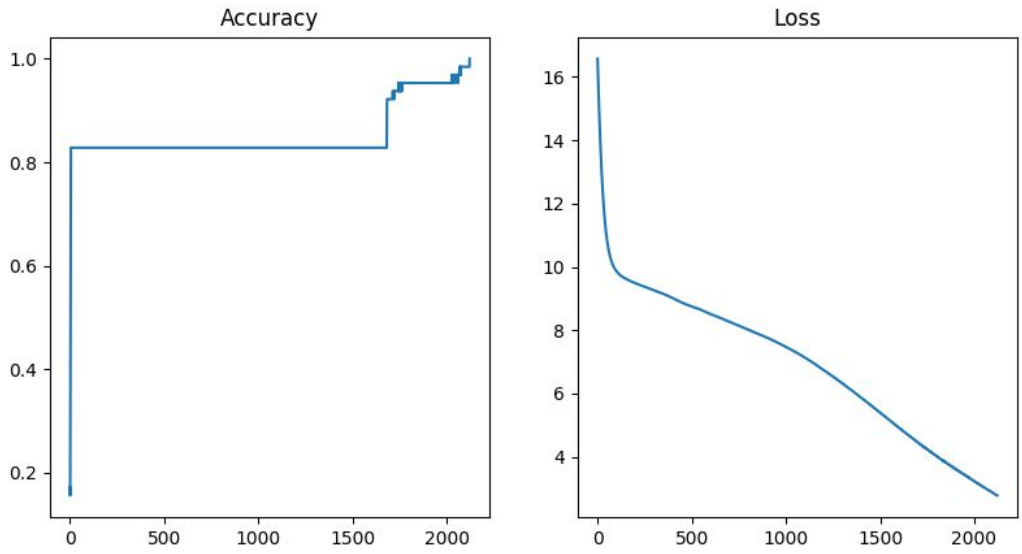
2.1. Sigmoid with L2



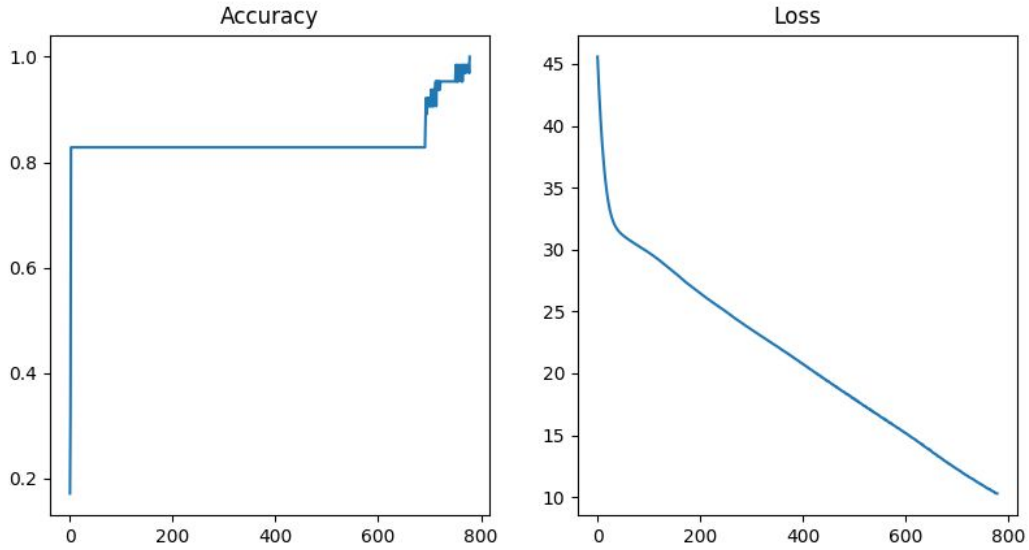
2.2. Sigmoid with CE



2.3. ReLu with L2



2.4. ReLu with CE



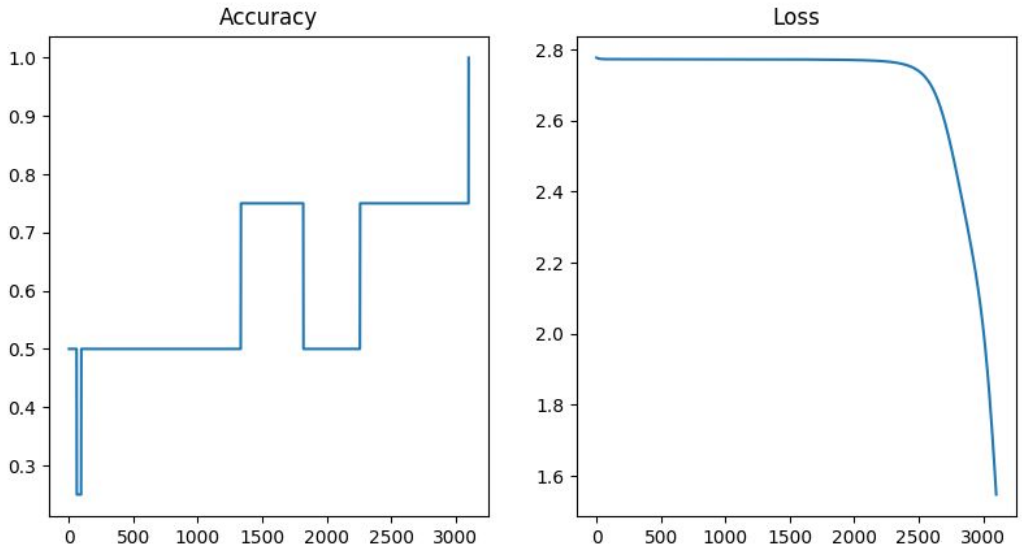
2.5. Sort the settings according to the training iterations to reach 100% accuracy, and explain the reasons of different convergence rates.

- Sigmoid with L2 (10,000 iterations)
- Sigmoid with CE (~5,000 iterations)
- ReLu with L2 (~2,500 iterations)
- ReLu with CE (~800 iterations)

As explained in part 1, cross entropy is more robust than L2 as a loss function for learning. We know that ReLu is more computationally efficient, as it does not

need to calculate an expensive exponential. Additionally, ReLu does not have a vanishing gradient, making it a more robust choice of loss function.

3. Solving XOR with a 2-layer Perceptron

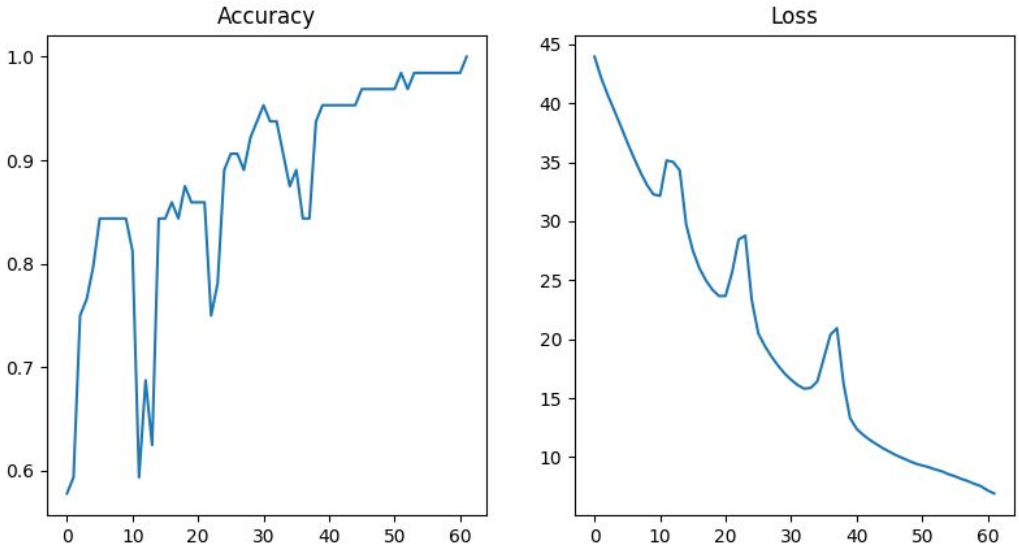


1) What will happen if we don't use an activation function in the hidden layer? Is the network be able to learn the XOR function?

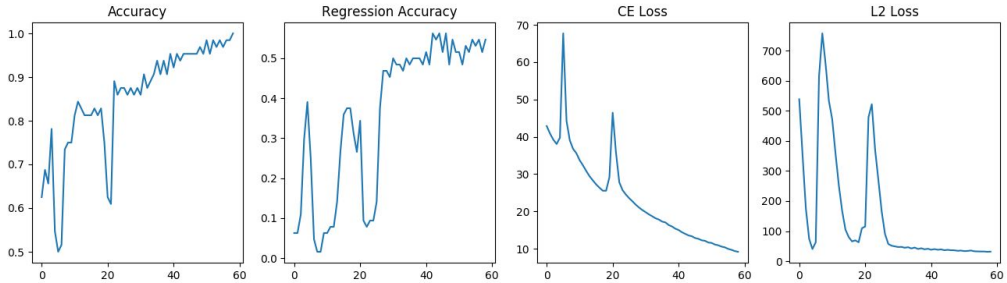
The nonlinear mapping we use in the activation layer allows us to fit a single hyperplane perceptron that properly labels samples on other side of it. The XOR function without this mapping is not linearly separable, so the network will not be able to learn it.

4. Experiment with Convolutional Network

4.1. CE loss + Accuracy vs iterations



4.2. L2 regression with CE loss vs iterations



4.3. Including the L2 regression in our neural net helps it converge faster, and labeling accuracy improves and stabilizes at higher iteration counts. Conversely, we see that without this branch in the neural net, loss increases three times (three blips) as opposed to two.