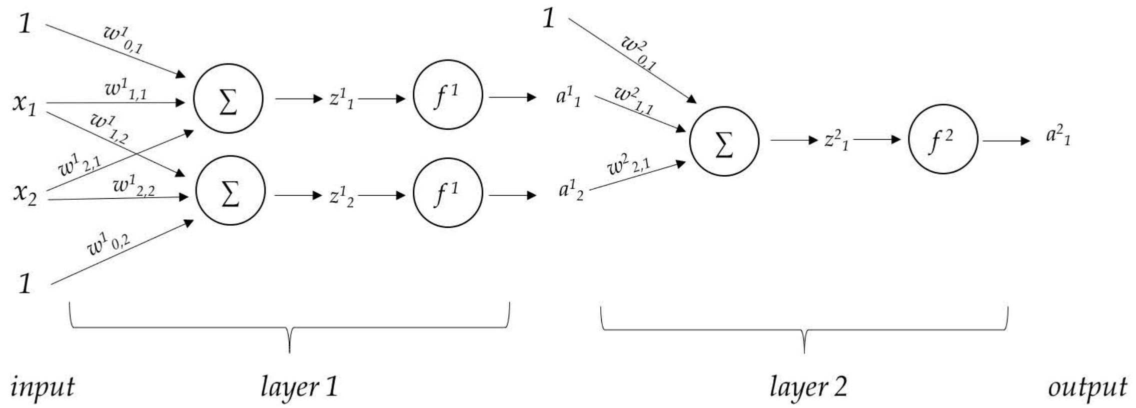
Part B

1. Consider the network below.



Suppose f^1 is the ReLU function and f^2 is the sigmoid activation function.

Assume that the initial weights are

$$w^1_{0,1} = 0, w^1_{1,1} = 1, w^1_{2,1} = -1$$

$$w^1_{0,2} = 0, w^1_{1,2} = -1, w^1_{2,2} = 1$$

$$w^2_{0,1} = 0, w^2_{1,1} = 2, w^2_{2,1} = 2$$

The step size is 0.5. Suppose that the current example is

$$x^{(1)} = [2, 1]^T, y^{(1)} = 0$$

(a) (1 point) What is the output value $y^{(1)} = a^2_1$, given current input $x^{(1)}$ and the current weight values?

(b) (2 points) What is the cross-entropy loss for this example?

(c) (6 points) Write down the numerical values of $\frac{\partial w^i_{j,k}}{\partial L}$ for all the weight parameters in the network, Show your work.

(d) (2 points) What will be the value of the weights after one step of stochastic gradient descent based on $(x^{(1)}, y^{(1)})$?

(e) (2 points) What will be the output value a^2_1 for this same input $x^{(1)}$ with these new weights?

2. (4 points) For a 3-class classification problem, the softmax function takes as input a vector (z_1, z_2, z_3) and returns a vector $(\hat{y}_1, \hat{y}_2, \hat{y}_3)$.

It is computed as

$$S = \sum_j e^{z_j} \qquad y_i = \frac{e^{z_i}}{S}$$

Determine the backpropagation updates for computing $\frac{\partial z_1}{\partial L}$ in terms of $\frac{\partial \hat{y}_1}{\partial L}, \frac{\partial \hat{y}_2}{\partial L}, \frac{\partial \hat{y}_3}{\partial L}$ or in terms of $\frac{\partial \hat{\mathbf{y}}}{\partial L}$