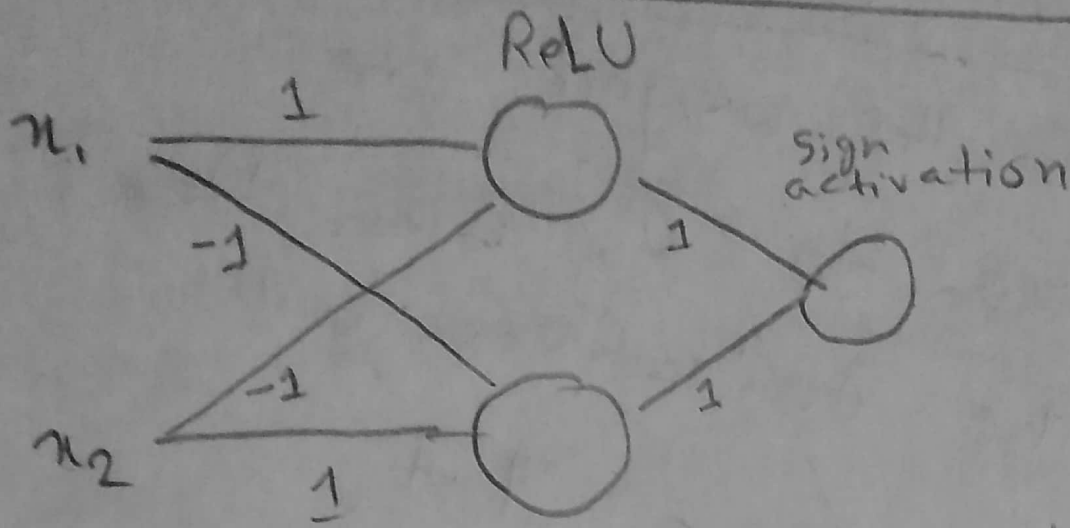


Answer to the Question 1



(input) ReLU (Hidden layer) (output layer)

All weights are shown in the figure. ReLU is used in the hidden layer and Sign activation function is used as the output layer.

Answer to the question 2

If all weights and offsets are initialized 0, the output of the ReLU nodes will be 0. So the output of them will be also 0. It means on the output node, both hidden units will be same influence and ultimately same gradients. So, the network will not be able to learn anything. Consequently, gradient descent will not learn the desired function from this initialization.

Answer to the question 3

Base case:

Let's we have a corpus where the sentences length is 1. So, the sentences look like "<s>w₁". In this case, $\sum_w^{V^1} p_n(w) = \sum_{w_1}^V p(w_1) = 1$

Inductive step:

Assume that $\sum_w^{V^{m-1}} p_n(w) = 1$ when the maximum sentence length is m-1. We want to show that is also holds for $\sum_w^{V^m} p(w) = 1$. We can think this summation as the product of two summations: 1) summation of probabilities of all sentences considering max length m and 2) summation of all n-gram probabilities for all possible sentences given the first. Part 1) equals 1 by the n-gram probability assumption from the question. Part 2) equals 1 by the inductive hypothesis. So $\sum_w p_n(w) = 1$.

Answer to the Question 9

From problem 3, we can say RNN Language models are valid which means

$$\sum_{\omega} p_n(\omega) = 1$$

Ensemble of RNN is also valid.

Because for ensemble of RNN, $p(\omega) = \frac{1}{R} \sum_{n=1}^R p_n(\omega)$

$$\begin{aligned}\sum_{\omega} p(\omega) &= \sum_{\omega} \frac{1}{R} \sum_{n=1}^R p_n(\omega) \\ &= \frac{1}{R} \sum_{n=1}^R \sum_{\omega} p_n(\omega) \\ &= \frac{1}{R} \sum_{n=1}^R 1 \\ &= 1\end{aligned}$$