Answer to the question 2

If all weights and offsets are initialized 0, the output of the ReLU nodes will be 0. So the out put 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>w_1”. In this case, $\sum_{w}^{V^1} p_n(w) = \sum_{w1}^{V} p(w1) = 1$

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 $\sum_{w} p_n(w) = 1$.