**Question 2**

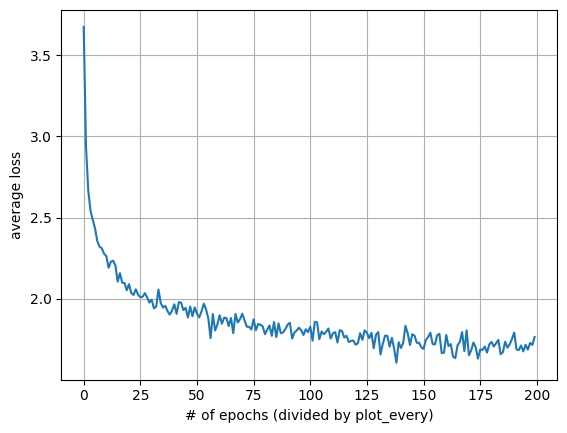
* Advantage of character based language model over word based language model:

Character based language models’ vocabulary is the alphabet or a small set of characters, which means that they are better in handling rare or unseen words. The latter is due to the models ability to “learn” patterns in character sequences rather than just treating unknown words as a “new” word (in comparison to word based language models).

* Advantage of word based language model over character based model:

Since word based language models’ vocabulary is a group of words, they have lower computational overhead. The latter is due to the fact that this group of models treat entire words as a single unit, which leads to them processing fewer tokens when processing text (in comparison to character based language models).

1. Losses plot:



**Question 3**

**Question 5**

1.

1. can be interpreted as a categorical probability distribution due to the following reasons:
   1. since for every , and ’s definition.
   3. Interpretation – Each represents the weight assigned to the corresponding value .
2. From definition, for the categorical distribution to put almost all its weight on a specific , it must be true that: .
3. From the fact that , and the condition from section b we get that and which leads to the fact that:
4. Intuitively, it means that our model gives more attention to the value vector because it finds it more relevant to the query .

2. We define ’ to be the following matrix: .

We define to be . Since the vectors are orthonormal we get that is the projection matrix to the sub space A where lies, thus:

From the fact that and being the projection matrix to the sub space A, we get that:

We finally get:

3. We select where is a large scalar, specifically large enough for the following condition to hold: .

Similarly, we can get that .

For every we get:

Now we calculate :

3.a. Since the covariance matrices are for vanishingly small we get that for all . .

We select where is a large scalar, specifically large enough for the following condition to hold: .

Similarly, we can get that .

For every we get:

Now we calculate :

1. Since we use from section a, for all we get that still . This leads to the attention output being primarily influenced by and , . What’s changed from section is that now the covariance for item has changed, the variability in ’s norm causes fluctuations in .  
   When ’s norm is larger, becomes more dominant, increasing ’s contribution to (in comparison to ). Conversely, when ’s norm is smaller, becomes less dominant, reducing ’s contribution to (in comparison to ).   
   The variability in ​'s norm introduces higher variance in compared to section a, where all keys had consistent norms (since ’s covariances were vanishingly small). So, in comparison to section a, the qualitative behaviour of depends strongly on 's sampled magnitude.

4.a. We select and where is a large scalar, specifically large enough for the following condition to hold: .

We will first focus on :

For every we get:

From here we get that:

Similarly, for we get , which leads to .

Finally, we get:

b. TODO

The query vectors , focus on and , respectively. Since has larger

Since we use and from section a, and . Which leads to being primarily influenced by and being primarily influenced by . What’s changed from section is that now the covariance for item has changed, the variability in ’s norm causes fluctuations in .  
When ’s norm is larger, becomes more dominant, increasing ’s contribution to (in comparison to ). Conversely, when ’s norm is smaller, becomes less dominant, reducing ’s contribution to (in comparison to ).   
The variability in ​'s norm introduces higher variance in compared to section a, where all keys had consistent norms (since ’s covariances were vanishingly small). So, in comparison to section a, the qualitative behaviour of depends strongly on 's sampled magnitude.