**Student Name: Guangyu Lin**

**Collaboration Statement:**
 Total hours spent: 2 days
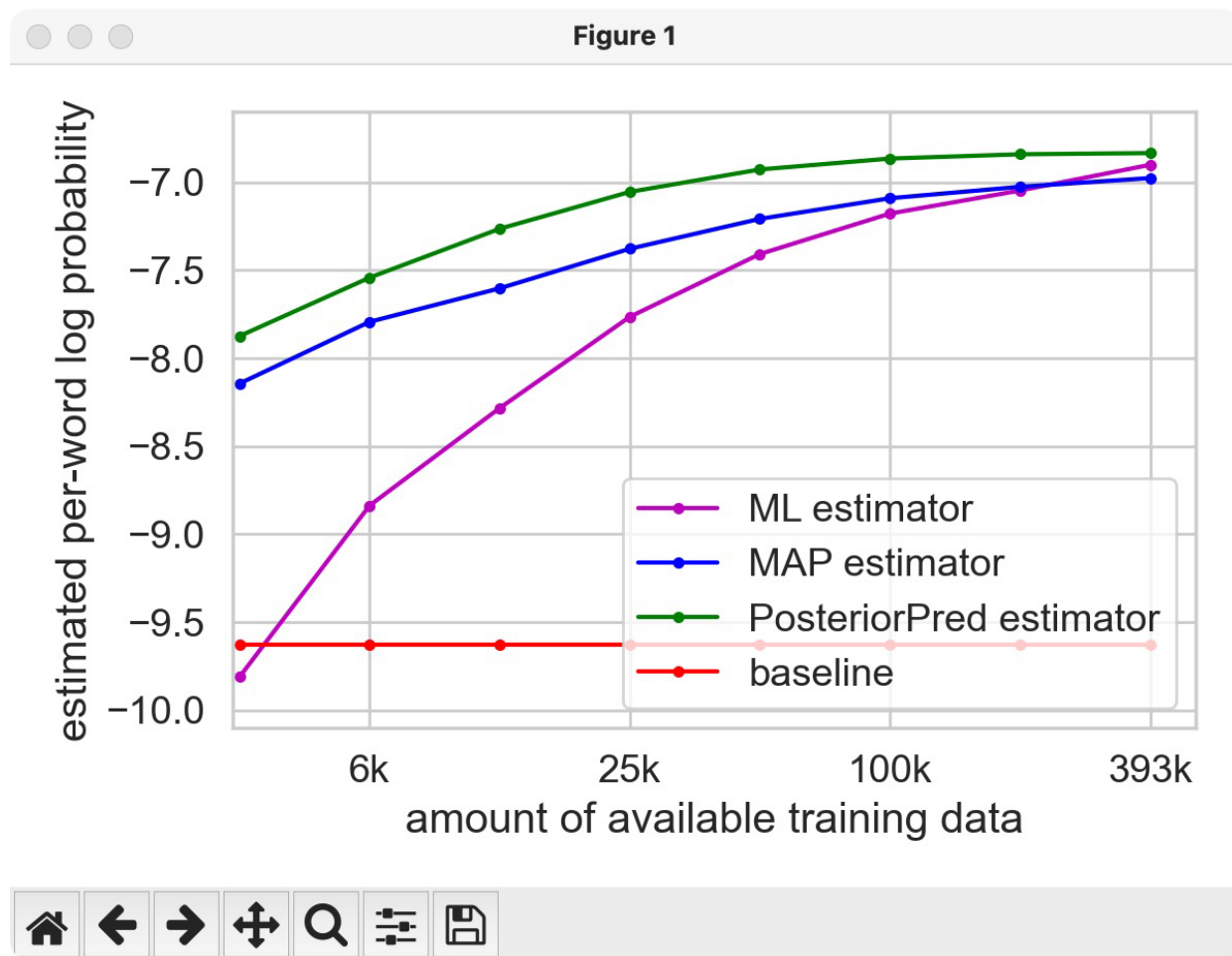 I discussed ideas with these individuals:

• Mingyang Wu

• TODO

• ...

I consulted the following resources:

• office hour with Professor

• TODO

• ...

By submitting this assignment, I affirm this is my own original work that abides by the course collaboration policy.

# 1a



Figure 1

# 1b

When N get larger and becomes infinite, there will be no situation of unseen count since every word will be in the vocabulary. Thus, there are also no problem of word appear and we never seen it in our training set since our train set size is large enought. So it turns out the probability will be

$$\frac{n_v}{N}$$

For the MAP estimator, when N becomes infinite, the original expression whcih is

$$\frac{n_v + \alpha - 1}{N + V(\alpha - 1)}$$

since alpha and V are fixed, with the N becomes infinitem, it will also becomes

$$\frac{n_v}{N}$$

For the Posterior Predictive estimator, since the expression is also contains alpha and V, the original expression which is

$$\frac{n_v + \alpha}{N + V\alpha}$$
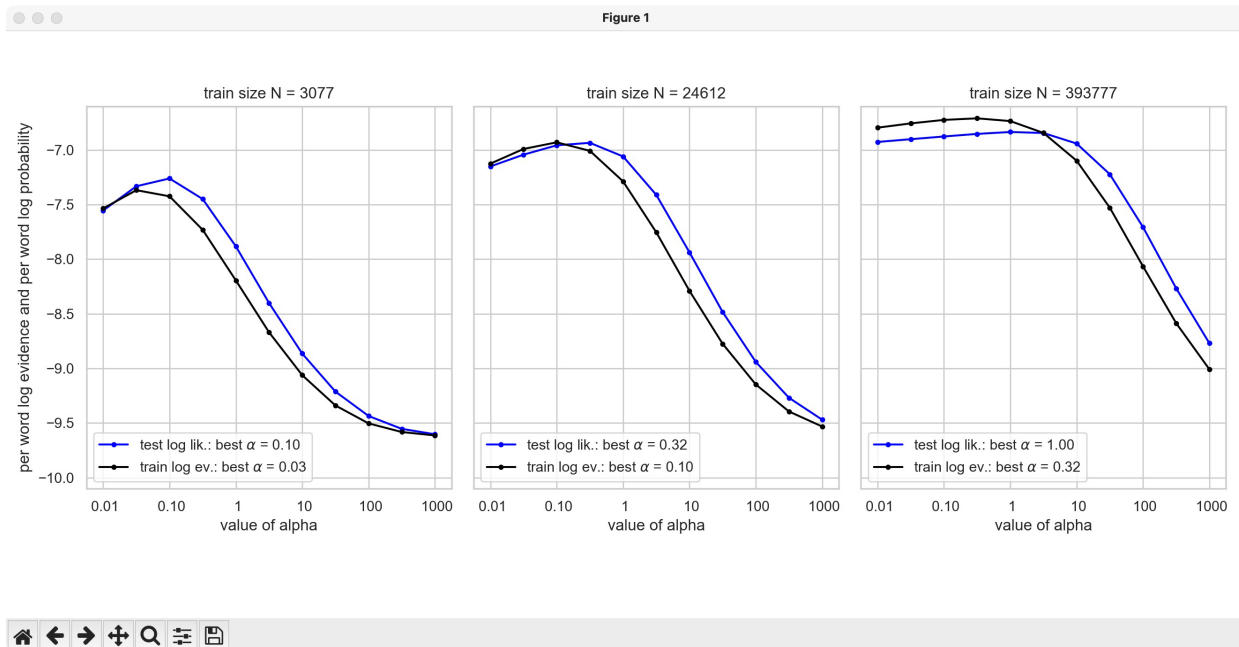
can be written as

$$\frac{n_v}{N}$$

Therefore, all three estimators' expression will become same when N becomes infinite. ML estimator will not be superior to the others when we have enough training data.

## 1c

In figure 1a, I graphed the baseline as a red line, which indicates that for relatively small datasets, the baseline may outperform the ML estimator. However, as the amount of training data increases, the ML estimator is expected to outperform the baseline.

## 2a

train size N = 3077     train size N = 24612     train size N = 393777

test log lik.: best $\alpha$ = 0.10
train log ev.: best $\alpha$ = 0.03

test log lik.: best $\alpha$ = 0.32
train log ev.: best $\alpha$ = 0.10

test log lik.: best $\alpha$ = 1.00
train log ev.: best $\alpha$ = 0.32

# 2b

The utilization of the evidence-on-train approach in 2a confers the advantage of curbing overfitting by taking into account the model's complexity during the hyperparameter tuning process, thereby ensuring that the model can generalize well to new data beyond the training dataset.

In contrast, the conditional-likelihood-on-validation approach enables efficient identification of the optimal model by assessing its performance on both the training and validation datasets, resulting in a closer approximation to real-world performance in the test dataset. However, the efficacy of this approach is dependent on the presence of a relevant validation dataset that accurately represents the novel data to be encountered.