Proof that the Unigram Model p is Optimal

Given:

- Vocabulary $V = \{v_k\}_k$
- Unigram model $\mathbf{p} = [p_k]_{k=0}^{|V|-1}$
- n_k is the number of observations of v_k
- $p_k = \frac{n_k}{\sum_k n_k}$ this is the probability of observing v_k is p_k

To prove:

- This unigram model maximizes the probability of the set of observations.

Let N be the total number of observations: $N = \sum_k n_k$

The probability of the set of observations is: $P(\text{observations}) = \prod_k p_k^{n_k}$

We can take natural log for both side: $\log P(\text{observations}) = \sum_k n_k \log(p_k)$

The probabilities must sum to 1 by definition: $\sum_{k=0}^{|V|-1} p_k = 1$

Given the probability constraint:

$\mathcal{L}(p, \lambda) = \sum_{k=0}^{|V|-1} n_k \log p_k + \lambda \left(1 - \sum_{k=0}^{|V|-1} p_k\right)$, where λ is the Lagrange multiplier.

To find the optimal point, take the partial derivative with respect to each p_k and set to zero:

$\frac{\partial \mathcal{L}}{\partial p_k} = \frac{n_k}{p_k} - \lambda = 0$

Thus, $p_k = \frac{n_k}{\lambda}$

Given the constraint: $\sum_k p_k = 1$

$\sum_k \frac{n_k}{\lambda} = 1 \Rightarrow \frac{1}{\lambda} \sum_k n_k = 1 \Rightarrow \lambda = \sum_k n_k$

Substitute λ back, we get:

$p_k = \frac{n_k}{\sum_k n_k}$

This shows that $p_k = \frac{n_k}{\sum_k n_k}$ is optimal for the unigram model.