We have $V = \{v_k\}_k$ as the vocabulary and the unigram model $p = \{p_k\}_{k=0}^{|V|-1}$

<u>Prove that:</u> $p_k = \frac{n_k}{\sum_{k=0}^{|V|-1} n_k}$ is optimal, where n_k: number of occurrences of v_k observed.

<u>Proof</u>:

For any model $p = \{p_k\}_{k=0}^{|V|-1}$ used to model the vocabulary V

We know that $\sum_{k=0}^{|V|-1} p_k = 1$ —— ①

must be satisfied.

We also have that the probability of the observations is:

$$p_{obs} = \prod_{k=0}^{|V|-1} (p_k)^{n_k} \quad \text{—— ②}$$

The most optimal model is one that maximizes the probability of observed data

That is, the model $p = \{p_k\}_{k=0}^{|V|-1}$ where the assignment of probabilities p_k satisfies eq ① and maximizes eq ②.

$$\max_{p} \prod_{k=0}^{|V|-1} (p_k)^{n_k} \quad \text{s.t} \quad p = \{p_k\}_{k=0}^{|V|-1} \text{ and } \sum_{k=0}^{|V|-1} p_k = 1$$

This is equivalent to maximizing the log of p_{obs}, since log is a monotonically increasing function

$$\approx \max_{p} \log\left[\prod_{k=0}^{|V|-1} (p_k)^{n_k}\right] \quad \text{s.t} \quad p = \{p_k\}_{k=0}^{|V|-1} \text{ and } \sum_{k=0}^{|V|-1} p_k = 1$$

$$= \max_{p} \sum_{k=0}^{|V|-1} n_K \log p_K \quad \text{s.t} \quad \sum_{k=0}^{|V|-1} p_K = 1$$

Using the Lagrangian method to solve the constrained optimization problem

$$L(p_0, p_1, \ldots, p_{|V|-1}, \lambda) = \sum_{k=0}^{|V|-1} n_K \log p_K - \lambda \left[\left(\sum_{k=0}^{|V|-1} p_K \right) - 1 \right]$$

$$\frac{\partial L(p_0, p_1, \ldots, p_{|V|-1}, \lambda)}{\partial p_K} = 0$$

$$\Rightarrow \frac{n_K}{p_K} - \lambda = 0$$

$$\Rightarrow \boxed{p_K = \frac{n_K}{\lambda}} \quad — (3)$$

From (1)

$$\sum_{k=0}^{|V|-1} p_K = 1$$

$$\Rightarrow \sum_{k=0}^{|V|-1} \frac{n_K}{\lambda} = 1$$

$$\Rightarrow \boxed{\lambda = \sum_{k=0}^{|V|-1} n_K} \quad — (4)$$

Therefore, $p_K = \dfrac{n_K}{\sum_{k=0}^{|V|-1} n_K}$ maximizes p_{obs} while satisfying $\sum_{k=0}^{|V|-1} p_K = 1$