1. Let $X_1, \ldots, X_n$ be a sequence of independently and identically distributed random variables taking values in a discrete and finite set $\mathcal{X}$, with common probability $P : \mathcal{X} \to [0, 1]$. Argue why the probability of observing a particular realization $X_1 = x_1, \ldots, X_n = x_n$ is given by

$$\log P(x_1, \ldots, x_n) = \log \prod_{t=1}^{n} P(x_t) = \sum_{x \in \mathcal{X}} N(x) \log P(x),$$

where $N(x)$ is the count of the symbol $x$ in the sequence $x_1, \ldots, x_n$.

Relate the formula derived above to Equation (16) in Chapter 4, page 68, of the textbook.

2. We discussed linear interpolation for smoothing a bigram language model in class, namely

$$P(w|v) = \lambda f(w|v) + (1 - \lambda) f(w),$$

where $f(\cdot|\cdot)$ and $f(\cdot)$ denote the appropriate relative frequency estimates, and $\lambda$ is chosen so as to maximize the probability of some held-out data.

This problem considers a few alternative strategies for smoothing a bigram language model by directly modifying the counts observed in the training data. In particular, let $C(v, w)$ denote the count of the bigram $\langle v, w \rangle$ in the training text, and let $C^*(v, w)$ be the modified count. For some constant $\theta > 0$, consider the three cases

(i) $C^*(v, w) = C(v, w) + \theta$,
(ii) $C^*(v, w) = C(v, w) + \theta C(v)$, and
(iii) $C^*(v, w) = C(v, w) + \theta C(v)f(w)$.

In each case, the smoothed bigram probability is calculated as

$$P^*(w|v) = \frac{C^*(v, w)}{\sum_{w' \in \mathcal{V}} C^*(v, w')} = \frac{C^*(v, w)}{C^*(v)}.$$  

(a) Let $N(v, w)$ denote the count of a bigram $\langle v, w \rangle$ in the held-out text. Compute, for (i)-(iii), an expression for the $\theta$ that maximizes the probability of the held-out text.
(b) Show that if \( N(v, w) = C(v, w) \) for all bigrams, then the optimal value is \( \theta = 0 \) in each of (i)-(iii). Why is this satisfactory?

(c) Show that for each of (i)-(iii), \( P^\star \) may be written as a linear interpolation. Identify the interpolation weight \( \lambda \), and comment on the merits or drawbacks of the interpolation.

3. Consider the back-off bigram language model

\[
P_{BO}(w|v) = \begin{cases} 
\frac{C(v, w) - \delta_1}{C(v)} & \text{if } C(v, w) > 0, \text{ and} \\
\alpha_1(v) f(w) & \text{otherwise}
\end{cases}
\]

where \( \delta_1 \in (0, 1) \) is sometimes called a constant discount coefficient, \( f(\cdot) \) denotes unigram relative frequencies and \( \alpha_1(v) \), called the back-off weight, is chosen to make \( P_{BO}(\cdot|v) \) a bona fide probability.

(a) Develop an expression for \( \alpha_1(v) \) in terms of the discount coefficient \( \delta_1 \), unigram probabilities \( f(\cdot) \), and bigram counts \( C(\cdot, \cdot) \).

(b) Write a corresponding formula for replacing \( f(w) \) with a back-off unigram model \( P_{BO}(w) \) that uses a discount coefficient \( \delta_0 \) and backs off to a uniform distribution on the entire vocabulary. Develop an expression for the back-off weight \( \alpha_0 \).

(c) Does replacing \( f(w) \) with \( P_{BO}(w) \) necessitate recomputation of \( \alpha_1(v) \)? What does this say about the sequence in which back-off weights in an back-off \( N \)-gram model should be computed?