4 Language Models – Hidden Markov Models

4.1 Computing the parameters of a model

Given the following sequences of pairs `(state,emission)`:

```
(D,the) (N,wine) (V,ages) (A,alone)
(D,the) (N,wine) (N,waits) (V,last) (N,ages)
(D,some) (N,flies) (V,dove) (P,into) (D,the) (N,wine)
(D,the) (N,dove) (V,flies) (P,for) (D,some) (N,flies)
(D,the) (A,last) (N,dove) (V,waits) (A,alone)
```

1. Draw the graph of the resulting *bigram* HMM, and list all non-zero model parameters that we can obtain via MLE from this data.
2. Draw the graph of the resulting *trigram* HMM, and list all non-zero model parameters that we can obtain via MLE from this data.
3. Compute the probability of the following sequence according to each of the two previous models:
   `(D,the) (N,dove) (V,waits) (P,for) (DT,some) (A,last) (N,wine)`

4.2 The Viterbi Algorithm in log-space

In this exercise we will modify the Viterbi algorithm used to compute the most likely state sequence in HMMs. Recall that we can specify an HMM model as $\mu = (A, B, \pi)$, where A is a matrix of transition parameters, B is a matrix of emission parameters, and π is the initial state distribution.

Let's first recall the computations behind basic Viterbi. Given an observation sequence $O = O_1 \dots O_T$ the algorithm computes:

$$\operatorname*{argmax}_{X=X_1 \dots X_T} P_\mu(X \mid O) = \operatorname*{argmax}_{X} \frac{P_\mu(X, O)}{P_\mu(O)} = \operatorname*{argmax}_{X} P_\mu(X, O)$$

The algorithm proceeds by defining quantities $\delta_j(t)$, which keep track of the most likely way of being at state X_j after emmitting $O_1 \dots O_t$. In parallel, the algorithm also computes variable $\psi_j(t)$, which store the transitions that lead to the most likely state sequence. The computations are as follows:

1. Initialization: $\forall j = 1 \dots N: \quad \delta_j(1) = \pi_j b_{j o_1}\ ; \quad \psi_j(1) = 0\ ;$
2. Induction: $\forall t = 1 \dots T,\ \forall j = 1 \dots N:$
   $$\delta_j(t+1) = \max_{1 \le i \le N} \delta_i(t) a_{ij} b_{j o_{t+1}}\ ; \qquad \psi_j(t+1) = \operatorname*{argmax}_{1 \le i \le N} \delta_i(t) a_{ij}\ ;$$
3. Termination: backwards path readout.
   $$\hat{X}_T = \operatorname*{argmax}_{1 \le i \le N} \delta_i(T)\ ; \qquad \forall t = 1..T-1: \ \hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)\ ; \qquad P(\hat{X}) = \max_{1 \le i \le N} \delta_i(T)\ ;$$

Moving to log-space

We are now interested in working in the logarithmic space. That is, we want to redefine the computations so that Viterbi computes $\log(P_\mu(X,O))$ instead of $P_\mu(X,O)$. This extension is important in tasks where probabilities of individual emissions or transitions may be very small, such as in NLP applications where the number of symbols is usually very very large. If individual probabilities are small then products of such probabilities will be very very very small. In such cases, our machines may run out of precision, hence yielding unreliable computations. Working in the log space is a common "trick" that solves the problem. We still want the most likely sequence under μ, but we will compute it as:

$$\begin{aligned}
\underset{X=X_1\ldots X_T}{\operatorname{argmax}} P_\mu(X \mid O) &= \underset{X}{\operatorname{argmax}} \log(P_\mu(X \mid O)) \\
&= \underset{X}{\operatorname{argmax}} \log\left(\frac{P_\mu(X,O)}{P_\mu(O)}\right) \\
&= \underset{x}{\operatorname{argmax}} \log(P_\mu(X,O)) - \log(P_\mu(O)) \\
&= \underset{x}{\operatorname{argmax}} \log(P_\mu(X,O))
\end{aligned}$$

First of all, note that even if we compute log probabilities instead of normal probabilities, the most likely sequence will be the same. This is because the log function preserves the value of the state sequence attaining the maximum probability. Second, as before, we can drop the term $\log(P_\mu(O))$ because O is fixed, and it does not affect the maximum.

Let's assume that the HMM is given in the log-space. That is, $\mu' = (A', B', \pi')$ where

- For any states i and j: $a'_{ij} = \log a_{ij} = \log P(X_{t+1} = s_j \mid X_t = s_i)$
- For any state i and symbol k: $b'_{ik} = \log b_{ik} = \log P(O_t = k \mid X_t = s_i)$
- For any state i: $\pi'_i = \log \pi_i = \log P(X_1 = i)$

Questions:

1. Write $\log(P_\mu(X,O))$ in terms of μ'.
2. Rewrite the recursive expressions for δ and ψ to work with log probabilities.

4.3 Error-augmented Viterbi

In this question we will modify Viterbi to account for a notion of Hamming error of state sequences. In future lectures we will see applications of this algorithm.

Let X and X' be two state sequences of length T. We define an error function that counts the number of different states (also known as Hamming error):

$$\text{error}(X, X') = \sum_{i=1}^{T} I[X_i \neq X'_i]$$

where the function $I[p]$ is an indicator function that returns 1 if predicate p is true and 0 otherwise. For example, $\text{error}(\text{"}abc\text{"}, \text{"}acb\text{"}) = 2$.

For this problem, the input will consist of an observation sequence O together with its *correct* state sequence X^*. The goal is to find the *most-erroneous* sequence under an HMM model specified by μ. That is, we are interested in finding a state sequence that has high probability under our model and also has high error. More formally, we would like to find:

$$\underset{X}{\text{argmax}} \ \log(P(X, O)) + \lambda \cdot \text{error}(X^*, X)$$

where the parameter λ controls the trade-off between the two terms (high values give more importance to the error).

Question:

1. Modify the Viterbi algorithm to solve this problem.

> *HINT:* Note that the error function decomposes in a similar fashion to the computations behind an HMM. Then think of the optimality conditions behind the design of Viterbi that allow us to compute δ and ψ recursively.