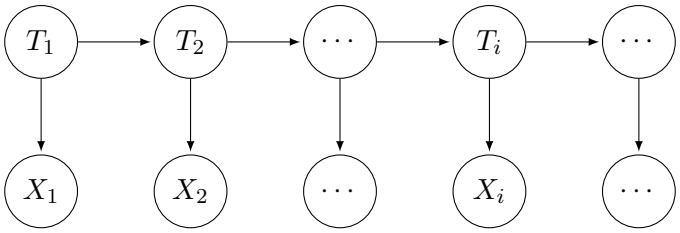
1 20 (15 + 5) points

part a: A representation of the HMM is shown below where the sequence of observable variables $\{X_i\}_{i\geq 1}$ denotes the stream of words and each state T_i denotes the topic of the item from which the corresponding word X_i is generated.



The state space of our model, which we denote $\mathcal{T} = \{1, \cdots, K\}$, is the collection of K possible topics to which each news item can belong. The transition probabilities give a distribution over following states given a current state, and the emission probabilities give a distribution over possible words given a topic. These probabilities can be represented efficiently as

$$P(T_{i+1}|T_i) = \begin{cases} .99 & \text{if } T_{i+1} = T_i \\ \frac{.01}{K-1} & \text{otherwise} \end{cases} \tag{1}$$

$$\mathcal{L}aw(X_i|T_i = k) = \text{Multinomial}(X_i; n = 1, (p_1, \cdots, p_M) = \theta_k). \tag{2}$$

In this formulation, T_i is encocded as an integer in $\mathcal{T}$ and X_i as a one-hot M-vector where M is the size of the collection of words which could appear.

part b: Yes, the variables are stochastically dependent. Intuitively, this can be understood by recognizing that X_i carries information about the state T_i and, due to the dependence of the transition probabilities on the current state, X_i also carries information about the states T_{i+1}, T_{i+2}. Therefore, X_i is informative about the word X_{i+2} since its distribution depends on the state T_{i+2}. More formally, we can write out the conditional probability

$$P(X_{i+2}|X_i) = \sum_{T_{i+2}\in\mathcal{T}} \left[P(X_{i+2}|T_{i+2}) \Big[\sum_{T_{i+1}\in\mathcal{T}} P(T_{i+2}|T_{i+1}) \Big[\sum_{T_i\in\mathcal{T}} P(T_{i+1}|T_i) P(T_i|X_i) \Big] \Big] \right]$$

and notice that it is a function of X_i.

2 30 (15 + 15) points

Part a WLOG we may write:

$$\mu = \begin{bmatrix} \mu_i \\ \mu_{-i} \end{bmatrix} \qquad\qquad \Sigma = \begin{bmatrix} \sigma_i^2 & \Sigma_{-i,i}^T \\ \Sigma_{-i,i} & \Sigma_{-i} \end{bmatrix}$$

where $\mu_i, \sigma_i \in \mathbb{R}^1$, $\mu_{-i}, \Sigma_{-i,i} \in \mathbb{R}^{(d-1)}$, and $\Sigma_{-i} \in \mathbb{R}^{(d-1)\times(d-1)}$. Hence we have

$$x_i \mid x_{-i} \sim \mathcal{N}(\mu_i + \Sigma_{-i,i}^T \Sigma_{-i}^{-1}(x_{-i} - \mu_{-i}), \sigma_i^2 - \Sigma_{-i,i}^T \Sigma_{-i}^{-1} \Sigma_{-i,i}). \tag{3}$$

part b: We write the conditional distribution by the Markov blanket property:

$$P(X_i \mid X_{-i}) = P(X_i \mid X_{i+1}, X_{i-1}) \tag{4}$$

Now, using the simplifying approximation write:

$$\begin{aligned} P(X_i \mid X_{i+1}, X_{i-1}) &\propto P(X_{i+1}, X_{i-1} \mid X_i) P(X_i) \\ &= P(X_{i+1} \mid X_i) P(X_{i-1} \mid X_i) P(X_i) \\ &\propto P(X_{i+1} \mid X_i) P(X_i \mid X_{i-1}) \\ &= \sigma(\theta_{i+1} X_i)^{X_{i+1}} \sigma(-\theta_{i+1} X_i)^{(1-X_{i+1})} \sigma(\theta_i X_{i-1})^{X_i} \sigma(-\theta_i X_{i-1})^{(1-X_i)}. \end{aligned}$$

Note that this will suffice as an answer since knowing the unnormalized form is equivalent to knowing the normalized form. This is due to the fact that we must have $X_i = 0$ or 1. If you can compute $P(X_i = 1|A)$ up to the normalizing constant (say it equals $Cf(1)$ and you can compute $f(1)$) and similarly so for $P(X_i = 0|A) = Cf(0)$, then you're done as

$$\begin{aligned} & P(X_i = 1|A) + P(X_i = 0|A) = C(f(0) + f(1)) = 1 \\ \implies & C = \frac{1}{f(0) + f(1)} \\ \implies & P(X_i = 1|A) = \frac{f(1)}{f(0) + f(1)}. \end{aligned}$$

3 50 (20 + 10 + 20) points

See accompanying jupyter notebook.