Student Name: Irene Chang

Collaboration Statement:

Total hours spent: 10 hours

I consulted the following resources:

- Piazza
- Pippa
- textbook, lecture notes

Contents

1a: Problem Statement

Prove the following property under a Hidden Markov Model.

$$p(z_{t+1}|x_t, z_t) = p(z_{t+1}|z_t) \tag{1}$$

1a: Solution

By the product rule of probability:

$$p(z_{t+1}, x_t, z_t) = p(z_{t+1}|x_t, z_t)p(x_t, z_t) = p(x_t|z_{t+1}, z_t)p(z_{t+1}, z_t)(1)$$

$$p(x_t, z_t) = p(x_t|z_t)p(z_t)(2)$$

$$p(z_{t+1}, z_t) = p(z_{t+1}|z_t)p(z_t)(3)$$

From the first equality of (1), we can write:

$$p(z_{t+1}|x_t, z_t) = \frac{p(z_{t+1}, x_t, z_t)}{p(x_t, z_t)}$$

And from the second equality of (1), we have:

$$p(z_{t+1}|x_t, z_t) = \frac{p(z_{t+1}, x_t, z_t)}{p(x_t, z_t)} = \frac{p(x_t|z_{t+1}, z_t)p(z_{t+1}, z_t)}{p(x_t, z_t)}$$

From equations (2) and (3):

$$= \frac{p(x_t|z_{t+1}, z_t)p(z_{t+1}|z_t)p(z_t)}{p(x_t|z_t)p(z_t)} = \frac{p(x_t|z_{t+1}, z_t)p(z_{t+1}|z_t)}{p(x_t|z_t)}(*)$$

From HMM assumption B, we know that given the hidden state at time t, the observation at time t is conditionally independent of all other variables in the model, so $p(x_t|z_{t+1}, z_t) = p(x_t|z_t)$

$$(*) = \frac{p(x_t|z_t)p(z_{t+1}|z_t)}{p(x_t|z_t)} = p(z_{t+1}|z_t)$$

This proves that $p(z_{t+1}|x_t, z_t) = p(z_{t+1}|z_t)$

1b: Problem Statement

Prove the following property under a Hidden Markov Model.

$$p(x_{t+1}|x_{1:t}, z_{1:t}) = p(x_{t+1}|z_t) \tag{2}$$

1b: Solution

We have:

$$\begin{aligned}
p(x_{t+1}|x_{1:t}, z_{1:t}) &= \frac{p(x_{t+1}, x_{1:t}, z_{1:t})}{p(x_{1:t}, z_{1:t})} \text{ (Conditional joint probability)} \\
&= \frac{p(z_1)\prod_{t=1}^{t} p(z_{t+1}|z_t)\prod_{t=2}^{t+1} p(x_t|x_{t-1}, z_{1:t})}{p(z_1)\prod_{t=1}^{t} p(z_{t+1}|z_t)\prod_{t=2}^{t} p(x_t|x_{t-1}, z_{1:t})} \text{ (Product rule)} \\
&= p(x_{t+1}|x_t, z_{1:t}) \\
&= \frac{p(x_{t+1}, x_t, z_{1:t})}{p(x_t, z_{1:t})} \text{ (Conditional joint probability)} \\
&= \frac{p(x_{t+1}|z_t)p(z_t|x_t, z_{1:t-1})p(x_t, z_{1:t-1})}{p(z_t|x_t, z_{1:t-1})p(x_t, z_{1:t-1})} \\
&= p(x_{t+1}|z_t)
\end{aligned}$$

Write out an expression for the expected complete log likelihood:

$$\mathbb{E}_{q(z_{1:T}|s)} \left[\log p(z_{1:T}, x_{1:T}|\theta)\right] \tag{3}$$

Use the HMM probabilistic model $p(z_{1:T}, x_{1:T}|\theta)$ and the approximate posterior $q(z_{1:T}|s)$ defined above.

Your answer should be a function of the data x, the local sequence parameters s and $r(s)$, as well as the HMM parameters π, A, ϕ.

2a: Solution

$\mathbb{E}_{q(z_{1:T}|s)} \left[\log p(z_{1:T}, x_{1:T}|\theta)\right] = \mathbb{E}_{q(z_{1:T}|s)} \left[\log \left(p(z_{1:T}|\pi, A) p(x_{1:T}|z_{1:T}, \phi)\right)\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\log p(z_{1:T}|\pi, A) + \log p(x_{1:T}|z_{1:T}, \phi)\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\log p(z_{1:T}|\pi, A)\right] + \mathbb{E}_{q(z_{1:T}|s)} \left[\log p(x_{1:T}|z_{1:T}, \phi)\right]$ (*)

$\mathbb{E}_{q(z_{1:T}|s)} \left[\log p(z_{1:T}|\pi, A)\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\log \left(\prod_{k=1}^K \pi_k^{\delta(z_1,k)} \cdot \prod_{t=2}^T \prod_{j=1}^K \prod_{k=1}^K A_{jk}^{\delta(z_{t-1},j)\delta(z_t,k)}\right)\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\sum_{k=1}^K \log \pi_k^{\delta(z_1,k)} + \sum_{t=2}^T \sum_{j=1}^K \sum_{k=1}^K \log A_{jk}^{\delta(z_{t-1},j)\delta(z_t,k)}\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\sum_{k=1}^K \delta(z_1, k) \log \pi_k\right] +$
$\mathbb{E}_{q(z_{1:T}|s)} \left[\sum_{t=2}^T \sum_{j=1}^K \sum_{k=1}^K \delta(z_{t-1}, j)\delta(z_t, k) \log A_{jk}\right]$
(Since the parameters $\log \pi_k, \log A_{jk}$ are given, assumed to be constant here:)
$= \sum_{k=1}^K \mathbb{E}_{q(z_{1:T}|s)} \left[\delta(z_1, k)\right] \log \pi_k +$
$\sum_{t=2}^T \sum_{j=1}^K \sum_{k=1}^K \mathbb{E}_{q(z_{1:T}|s)} \left[\delta(z_{t-1}, j)\delta(z_t, k)\right] \log A_{jk}$
(It's given that $\mathbb{E}_{q(z_{1:T}|s)}[\delta(z_t, k)] = r_{tk}(s), \mathbb{E}_{q(z_{1:T}|s)}[\delta(z_t, j)\delta(z_{t+1}, k)] = s_{tjk}$)
$= \sum_{k=1}^K r_{1k}(s) \log \pi_k + \sum_{t=2}^T \sum_{j=1}^K \sum_{k=1}^K s_{(t-1)jk} \log A_{jk}$ (1)

$\mathbb{E}_{q(z_{1:T}|s)} \left[\log p(x_{1:T}|z_{1:T}, \phi)\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\log \left(\prod_{t=1}^T \prod_{d=1}^D \prod_{k=1}^K \text{BernPMF}(x_{td}|\phi_{kd})^{\delta(z_t,k)}\right)\right]$
$= \mathbb{E}_{q(z_{1:T}|s)} \left[\sum_{t=1}^T \sum_{d=1}^D \sum_{k=1}^K \log \text{BernPMF}(x_{td}|\phi_{kd})^{\delta(z_t,k)}\right]$
$= \sum_{t=1}^T \sum_{d=1}^D \sum_{k=1}^K \mathbb{E}_{q(z_{1:T}|s)} \left[\delta(z_t, k) \log \text{BernPMF}(x_{td}|\phi_{kd})\right]$
(Since the parameters ... is given)
$= \sum_{t=1}^T \sum_{d=1}^D \sum_{k=1}^K \mathbb{E}_{q(z_{1:T}|s)} \left[\delta(z_t, k)\right] \log \text{BernPMF}(x_{td}|\phi_{kd})$

It's given that $\mathbb{E}_{q(z_{1:T}|s)}[\delta(z_t, k)] = r_{tk}(s)$
$= \sum_{t=1}^{T} \sum_{d=1}^{D} \sum_{k=1}^{K} r_{tk}(s) \log \text{BernPMF}(x_{td}|\phi_{kd})$ (2)
Substitute (1) (2) into (*), we have

$$\mathbb{E}_{q(z_{1:T}|s)} \left[\log p(z_{1:T}, x_{1:T}|\theta)\right] = \sum_{k=1}^{K} r_{1k}(s) \log \pi_k + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} s_{(t-1)jk} \log A_{jk}$$
$$+ \sum_{t=1}^{T} \sum_{d=1}^{D} \sum_{k=1}^{K} r_{tk}(s) \log \text{BernPMF}(x_{td}|\phi_{kd})$$

2b: Problem Statement

Provide a short verbal summary of the update for ϕ_{kd} given below. How should we interpret the numerator? The denominator?

$$\phi_{kd} = \frac{\sum_{t=1}^{T} r_{tk} x_{td}}{\sum_{t=1}^{T} r_{tk}} \tag{4}$$

2b: Solution

In updating ϕ_{kd}, we are trying to maximize the third term in our derived equation of expected complete log likelihood, hence maximizing the overall likelihood. r_{tk} is the probability for the transition to the kth hidden state, for timestep t. Thus, the numerator can be understood as the expected value at dimension d of the hidden state k, since it is the transition probability times the actual value of x_t at dimension d, summing across all t timesteps. The denominator is the sum of all transition probabilities across all timesteps t. Viewing r_{tk}'s as weights, essentially, by dividing the numerator by the denominator, we're finding where the weight is centered among the values of dimension d of x_t's at hidden state k taken across all timesteps. Recall that ϕ_{kd} is the probability that the binary value of dimension of vector x_t will be "on", if generated when time t assigned to hidden state k, so the update means we draw this probability towards a value that's more to the "center", or towards an expected value given the data points.