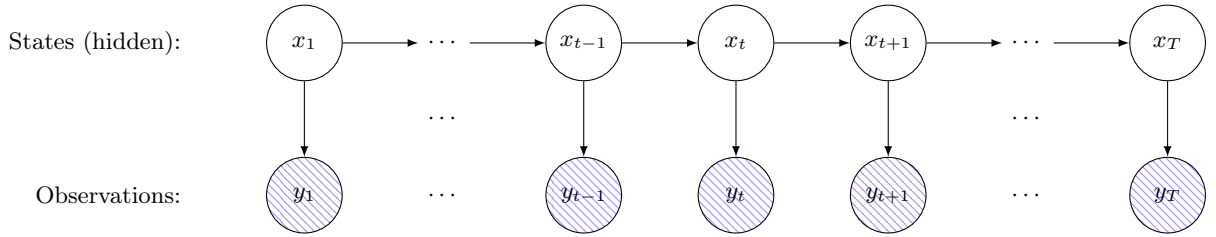


# Chapter 0

## Inference in Hidden Markov Models

A hidden Markov model (HMM) is a graphical model of the form shown below. The top chain is a Markov chain representing the state of some system. Typically the state cannot be observed directly. However, we can observe some (probabilistic) function of the state. For example, the Markov chain can represent the health status of a patient and the observations are symptoms such as temperature, blood pressure, etc. As another example, the Markov chain can represent the part of speech of words in a text, and the observation is the actual word.



The probability distribution for this model factorizes as

$$p(x_1^T, y_1^T; \theta) = p(x_1) \prod_{t=2}^T p(x_t | x_{t-1}) \prod_{t=1}^T p(y_t | x_t).$$

Assuming the Markov chain and the observations are both on discrete spaces, we can complete the model by specifying  $\theta = (\pi, A, B)$ , where:

- The probability distribution  $\pi$  for  $x_1$ ,

$$\pi_i = p(x_1 = i).$$

- The *transition matrix*  $A$  of the Markov chain,

$$A_{ij} = p(x_{t+1} = j | x_t = i).$$

- The *emission matrix*  $B$  describing the probabilities of the observations given the state,

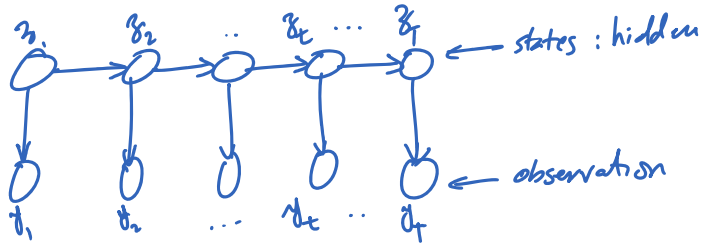
$$B_{ij} = p(y_t = j | x_t = i).$$

Below are three common inference problems associated with HMMs and the methods for solving them. We will not derive the solutions but they can be found in [2].

- Evaluation:  $p(x_t | y_1^T; \theta) \rightarrow$  *forward-backward algorithm* (sum-product).
- Decoding:  $\arg \max_{x_1^T} p(x_1^T | y_1^T; \theta) \rightarrow$  *Viterbi algorithm* (max-product).

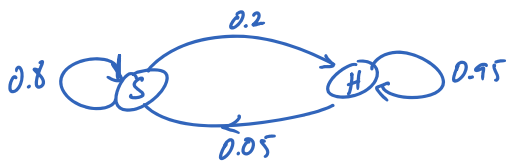
- Learning:  $\arg \max_{\theta} p(y_1^T; \theta) \rightarrow \text{Baum-Welch algorithm (EM)}$ .

*Below are HMM notes from a previous class. Unless I get a chance to go over these in class, they are not part of the course material and are here for self-study. But note that the methods are sum-product, max-product, and EM algorithms, which are part of the course and so reviewing the material below can be helpful in understanding those.*



\* A person can be either sick or healthy : hidden state

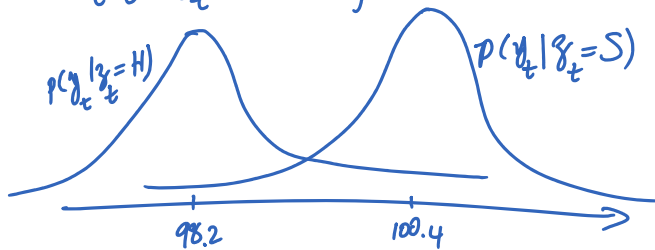
temperature : observation



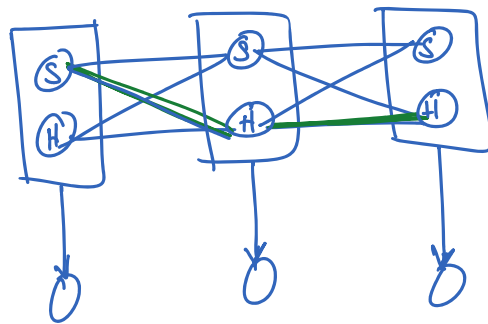
$$p(z_{t+1}=j | z_t=i) = A_{ij} : \text{transition probs.}$$

prob. distr. over initial state  $p(z_1=i) = \pi_i$

$$p(y_t=j | z_t=i) = B_{ij} : \text{emission probs.}$$



Trellis :



paths : configurations  
of hidden states

$$p(z_1^T, y_1^T | \theta) = p(z_1) \left( \prod_{t=2}^T p(z_t | z_{t-1}) \right) \left( \prod_{t=1}^T p(y_t | z_t) \right)$$

$$\theta = (\pi, A, B)$$

$$\pi_i = p(z_1 = i | \theta)$$

$$A_{ij} = p(z_{t+1} = j | z_t = i, \theta)$$

$$B_{ij} = p(y_t = j | z_t = i, \theta)$$

Three HMM problems:

\* Evaluation:  $p(z_t | y^T, \theta)$

- Forward-Backward (Sum-product)

\* Decoding:  $\arg \max_{z_1^T} p(z_1^T | y^T, \theta)$

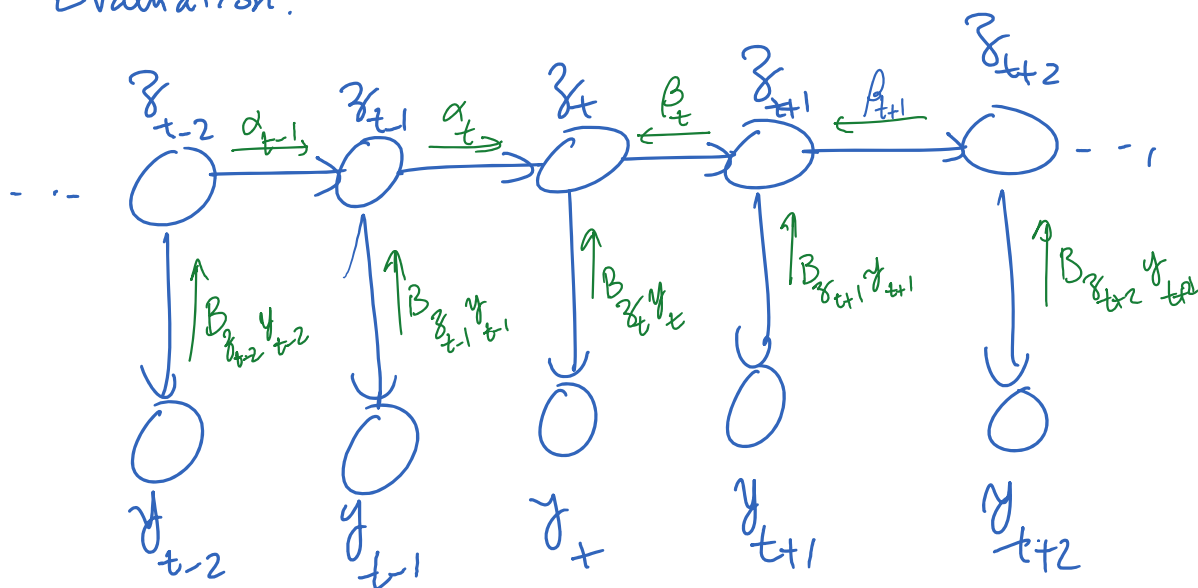
- Viterbi Alg (Max-product)

Copeland & Qualcomm

\* Learning:  $\arg \max_{\theta} p(y^T | \theta)$

- Baum-Welch alg (EM)

Evaluation:



Define

$$\dots \quad \pi_i = p(z_1 = i) \quad t \geq 2 \quad | \quad A_{ij} = p(z_t = j | z_{t-1} = i) \quad t \leq T-1$$

Define

$$\alpha_t(i) = \mu_{z_{t-1} z_t}(z_t=i) \quad t \geq 2 \quad \left| \quad \beta_t(i) = \mu_{z_{t+1} z_t}(z_t=i) \quad t \leq T-1\right.$$

$$\alpha_1(i) = \pi_i \quad \left| \quad \beta_T(i) = 1\right.$$

It can be shown (by induction) that:

$$\alpha_t(i) = p(z_t=i, y_1^{t-1} | \theta), \quad \beta_t(i) = p(y_{t+1}^T | z_t=i, \theta)$$

$$\alpha_t(i) = \sum_j \alpha_{t-1}(j) B_{j y_{t-1}} A_{ji}$$

$$\beta_t(i) = \sum_j \beta_{t+1}(j) B_{j y_{t+1}} A_{ij}$$

Marginals:

$$p(z_t=i | y_1^T, \theta) = \gamma_t(i) \propto \alpha_t(i) \beta_t(i) B_{i y_t}$$

$$p(z_{t-1}=i, z_t=j | y_1^T, \theta) = \zeta_t(i, j) \propto p(y_1^T, z_{t-1}=i, z_t=j | \theta)$$

$$= p(y_1^{t-2}, z_{t-1}=i) \underbrace{p(y_{t-1} | z_{t-1}=i) p(z_t=j | z_{t-1}=i) p(y_t | z_t=j)}_{\downarrow = p(y_{t-1} | y_1^{t-2}, z_{t-1}=i)} p(y_{t+1}^T | z_t=j)$$

$$= \alpha_{t-1}(i) B_{i y_{t-1}} A_{ij} B_{j y_t} \beta_t(j)$$

---

In traditional form of Forward-Backward, forward mgs  
included  $B_{i y_t}$ .

$$\bar{\alpha}_t(i) = p(z_t=i, y_1^t | \theta) = \alpha_t(i) B_{i y_t}$$

$$\bar{\alpha}_t(i) = \sum_j \bar{\alpha}_{t-1}(j) A_{ji} B_{ij} \gamma_t$$

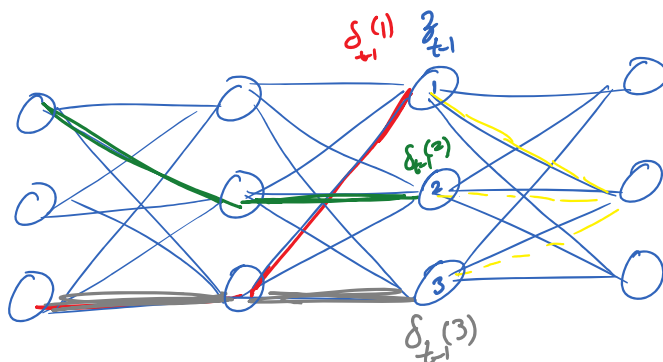
$$\gamma_t(i) \propto \bar{\alpha}_t(i) \beta_t(i)$$

Max-product: choose  $z_T$  as root.

$$\text{Define } \delta_t(i) = \mu_{z_{t-1} z_t} (z_t = i), \quad t \geq 2 \quad \delta_1(i) = \pi_i$$

Can be shown that

$$\delta_t(i) = \max_{z_1^{t-1}} p(z_1^{t-1}, z_t = i, y_1^{t-1} | \theta)$$



$$\delta_t(i) = \max_j \delta_{t-1}(j) B_{ji} A_{ji}$$

$$\text{prob of the max-prob path} = \max_j \delta_T(j) B_{jT}$$

\* Learning: EM / Baum-Welch

Assume complete data:  $z_1^T, y_1^T$

Estimate  $\theta = (\pi, A, B)$

$$\hat{\pi}_i = \begin{cases} 1 & z_1 = i \\ 0 & \text{else} \end{cases}$$

$$\hat{A}_{ij} = \frac{\sum_{t=1}^{T-1} I(z_t = i, z_{t+1} = j)}{\sum_{t=1}^{T-1} I(z_t = i)}$$

$$\hat{A}_{ij} = \frac{\sum_{t=1}^T I(z_t = i, y_t = j)}{\sum_{t=1}^T I(z_t = i)}$$

$$\hat{B}_{ij} = \frac{\sum_{t=1}^T I(z_t = i, y_t = j)}{\sum_{t=1}^T I(z_t = i)}$$

log likelihood of the complete data

$$\ln p(z_1^T, y_1^T | \theta) = \ln \pi_{z_1} + \sum_{t=2}^T \ln A_{z_{t-1} z_t} + \sum_{t=1}^T \ln B_{z_t y_t}$$

E-step

$$Q(\theta | \theta') = E[\ln p(z_1^T, y_1^T | \theta) | y_1^T, \theta']$$

$$E[\ln \pi_{z_1} | y_1^T, \theta'] = \sum_i (\ln \pi_i) \underbrace{p(z_1 = i | y_1^T, \theta')}_{\gamma'_1(i)} = \sum_i \gamma'_1(i) \ln \pi_i$$

$$E\left[\sum_{t=2}^T \ln A_{z_{t-1} z_t} | y_1^T, \theta'\right] = \sum_{t=2}^T \sum_i \sum_j (\ln A_{ij}) \underbrace{p(z_t = j, z_{t-1} = i | y_1^T, \theta')}_{\sum_t \gamma'_t(i, j)}$$

$$= \sum_i \sum_j \left( \sum_{t=2}^T \gamma'_t(i, j) \right) \ln A_{ij}$$

$$E\left[\sum_{t=1}^T \ln B_{z_t y_t} | y_1^T, \theta'\right] = \sum_i \sum_{t=1}^T \underbrace{p(z_t = i | y_1^T, \theta')}_{\gamma'_t(i)} \ln B_{i y_t}$$

ML for  $p$  if the LL =  $\sum_j n_j \ln p_j \Rightarrow p_j \propto n_j$

$$\pi_i \propto \gamma'_1(i) \quad A_{ij} \propto \sum_{t=2}^T \gamma'_t(i, j)$$

$$\Rightarrow \sum_i \sum_j \left( \sum_{t=1}^T \gamma'_t(i) I(y_t = j) \right) \ln B_{ij} \Rightarrow B_{ij} \propto \sum_{t=1}^T \gamma'_t(i) I(y_t = j)$$

Helpful references: [2, 1]

## References

- [1] Christopher M. Bishop. *Pattern Recognition And Machine Learning*. New York: Springer, 2006. URL: <http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%5C%20-%5C%20Pattern%5C%20Recognition%5C%20And%5C%20Machine%5C%20Learning%5C%20-%5C%20Springer%5C%20%5C%202006.pdf> (visited on 02/14/2017).
- [2] Bruce Hajek. *Random Processes for Engineers*. Illinois, 2014. URL: <http://hajek.ece.illinois.edu/Papers/randomprocJuly14.pdf> (visited on 01/30/2017).