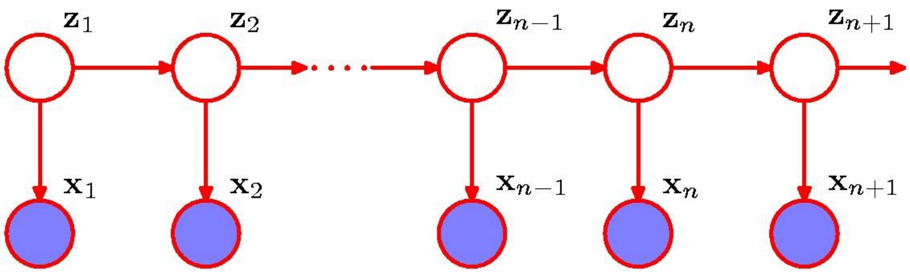
Hidden Markov Models

Training – Selecting model parameters



What we know

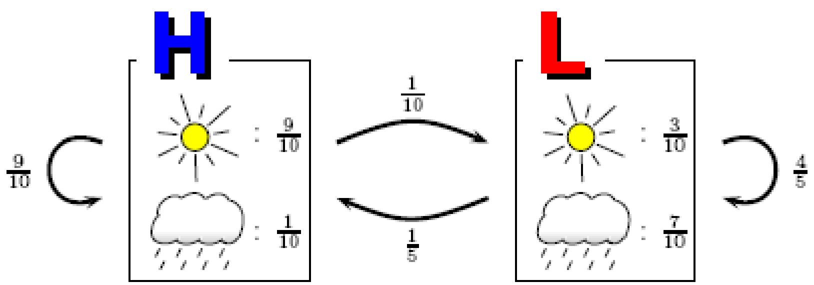
- The terminology and notation of hidden Markov models (**HMMs**)
- The **forward- and backward-algorithms** for determining the likelihood $p(\mathbf{X})$ of a sequence of observations, and computing the **posterior decoding**.
- The **Viterbi-algorithm** for finding the most likely underlying explanation (sequence of latent states) of a sequence of observation
- How to implement the Viterbi-algorithm using log-transform (and the forward- and backward-algorithms using scaling).

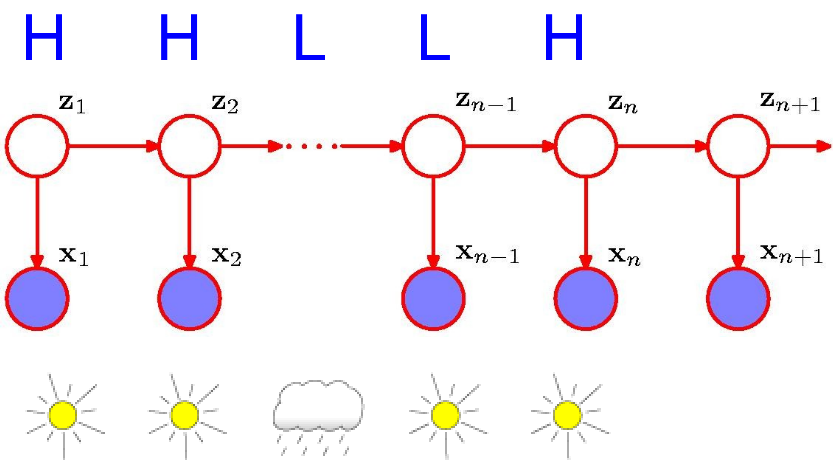
Now

- Training, or how to select model parameters (transition and emission probabilities) to reflect either a set of corresponding $(\mathbf{X},\mathbf{Z})$'s, (or just a set of $\mathbf{X}$'s) ...

Selecting “the right” parameters

Assume that (several) sequences of observations $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ and corresponding latent states $\mathbf{Z}=\{\mathbf{z}_1,\ldots,\mathbf{z}_n\}$ are given ...

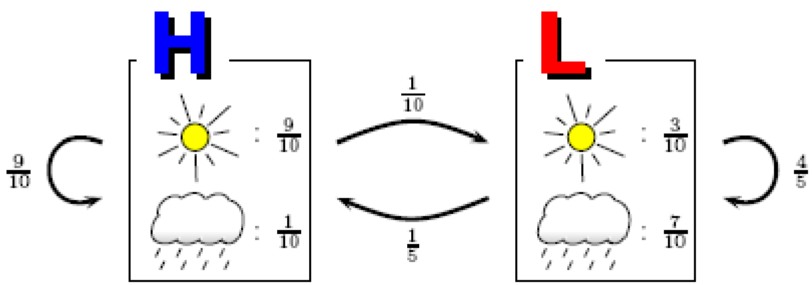


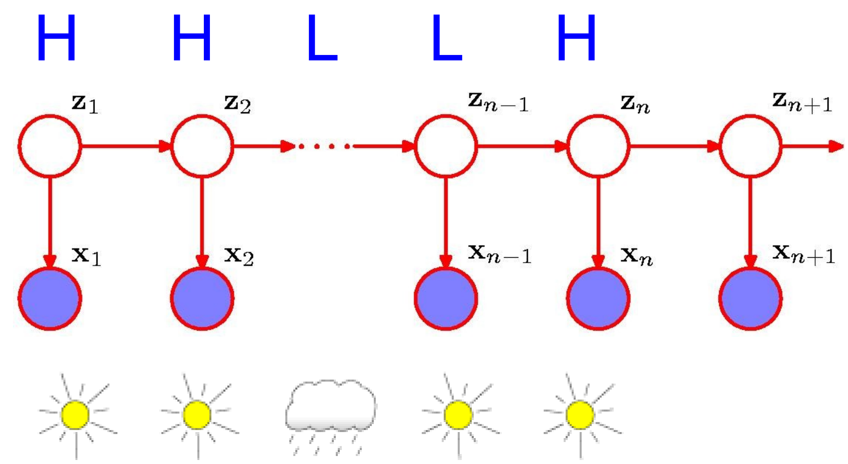


How should we set the model parameters, i.e. transition **A**, **π**, and emission probabilities **Φ**, to make the given (**X**,**Z**)'s most likely?

Selecting "the right" parameters

Assume that (several) sequences of observations **X**={$\mathbf{x}_1,\ldots,\mathbf{x}_n$} and corresponding latent states **Z**={$\mathbf{z}_1,\ldots,\mathbf{z}_n$} are given ...

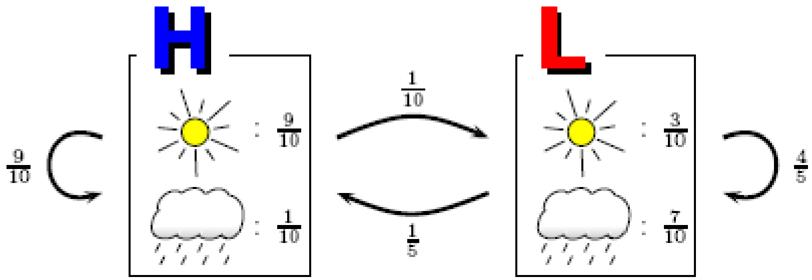


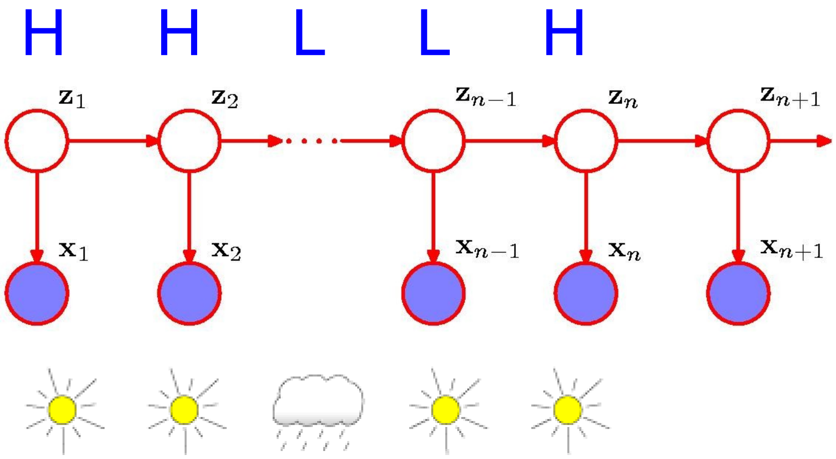


How should we set the model parameters, i.e. transition **A**, **π**, and emission probabilities **Φ**, to make the given (**X**,**Z**)'s most likely?

Intuition: The parameters should reflect what we have seen ...

Selecting "the right" transition probs





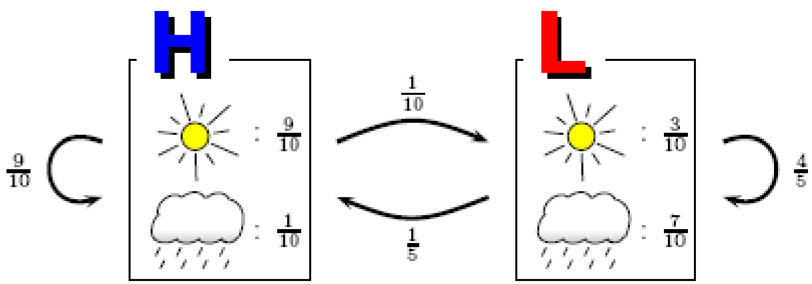
A_{jk} is the probability of a transition from state *j* to state *k*, and π_k is the probability of starting in state *k* ...

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} =$$

How many times is the transition from state *j* to state *k* taken

How many times is a transition from state *j* to any state taken

Selecting "the right" transition probs



H H L L H

A_{jk} is the probability of a transition from state j to state k, and π_k is the probability of starting in state k ...

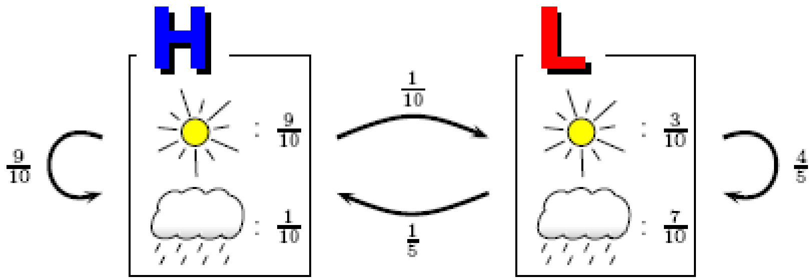
$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} =$$

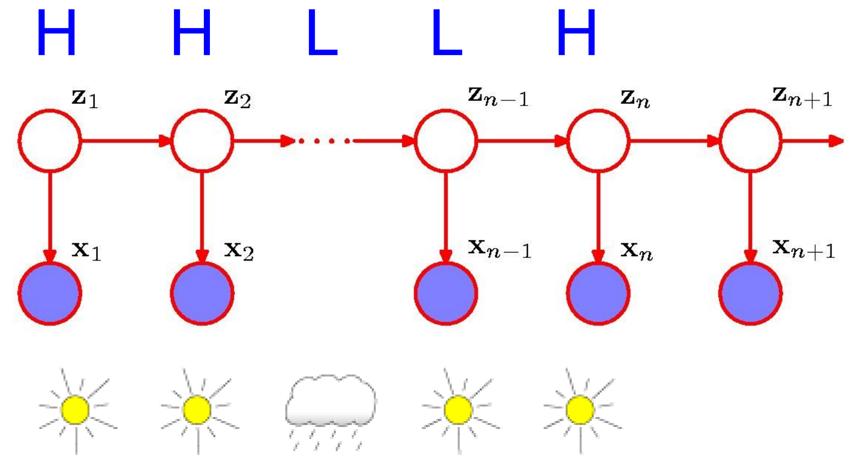
How many times is the transition from state j to state k taken

How many times is a transition from state j to any state taken

$$\pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}}$$

Selecting "the right" emission probs





If we assume discrete observations, then Φ_{ik} is the probability of emitting symbol *i* from state *k* ...

$$\phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}} = \frac{\text{How many times is symbol } i \text{ emitted from state } k}{\text{How many times is a symbol emitted from state } k}$$

Selecting “the right” parameters

Assume that (several) sequences of observations $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ and corresponding latent states $\boldsymbol{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\}$ are given ...

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

We simply count how many times each outcome of the multinomial variables (a transition or emission) is observed ...

Selecting “the right” parameters

Assume that (several) sequences of observations ***X***={***x***$_1$,...,***x***$_n$} and corresponding latent states ***Z***={***z***$_1$,...,***z***$_n$} are given ...

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

We simply count how many times each outcome of the multinomial variables (a transition or emission) is observed ...

This yield a <u>maximum likelihood estimate</u> (MLE) **θ*** of p(**X**,**Z** | **θ**), which is what we mathematically want ...

$$f_{\mathbf{X},\mathbf{Z}}(\Theta) = p(\mathbf{X}, \mathbf{Z} | \Theta) \qquad \Theta^* = \arg\max_{\Theta} f_{\mathbf{X},\mathbf{Z}}(\Theta)$$

Selecting “the right” parameters

Assume that (several) sequences of observations $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ and corresponding latent states $\boldsymbol{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\}$ are given ...

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

We simply count how many times each outcome of the multinomial variables (a transition or emission) is observed ...

This yield a <u>maximum likelihood estimate</u> (MLE) **θ*** of $p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})$, which is what we mathematically want ...

Any problems?

Selecting "the right" parameters

Assume that (several) sequences of observations $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ and corresponding latent states $\boldsymbol{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\}$ are given ...

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

We simply count how many times each outcome of the multinomial variables (a transition or emission) is observed ...

This yield a <u>maximum likelihood estimate</u> (MLE) **θ*** of $p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})$, which is what we mathematically want ...

Any problems? What if e.g. the transition from state *j* to *k* is *not* observed, then probability A_{jk} is set to 0.

Selecting "the right" parameters

Assume that (several) sequences of observations $\boldsymbol{X}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ and corresponding latent states $\boldsymbol{Z}=\{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n\}$ are given ...

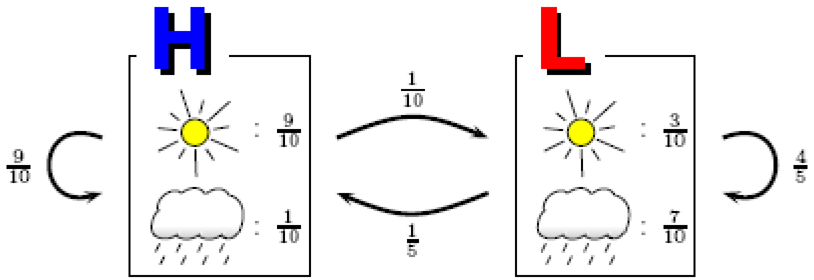
$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

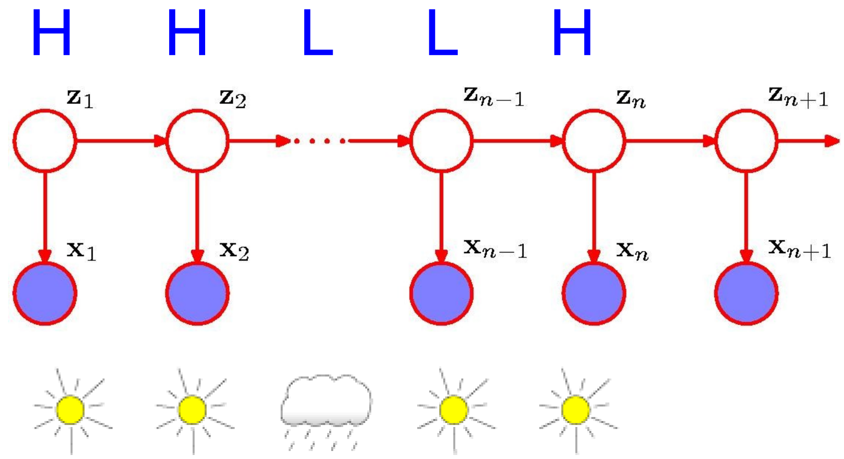
We simply count how many times each outcome of the multinomial variables (a transition or emission) is observed ...

This yield a maximum likelihood estimate (MLE) $\boldsymbol{\theta}^*$ of $p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})$, which is what we mathematically want ...

Any problems? What if e.g. the transition from state *j* to *k* is *not* observed, then probability A_{jk} is set to 0. Practical solution: Assume that every transition and emission is seen once (pseudocount) ...

Example

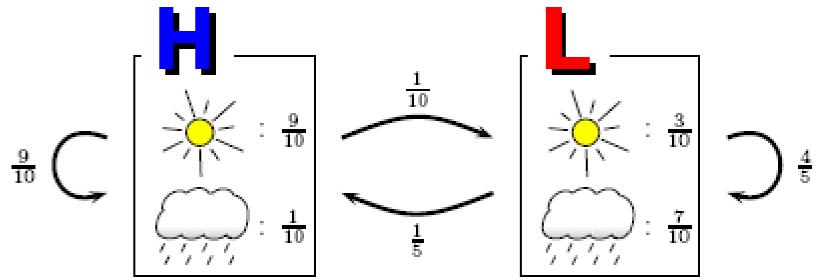


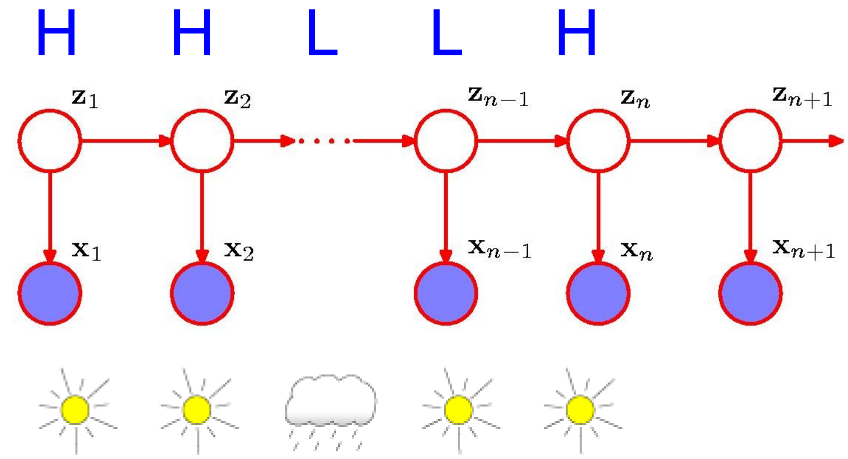


Without pseudocounts:

A_{HH} = 1/2 p(sun|H) = 1
A_{HL} = 1/2 p(rain|H) = 0
A_{LH} = 1/2 p(sun|L) = 1/2
A_{LL} = 1/2 p(rain|L) = 1/2
π_H = 1
π_L = 0

Example





Without pseudocounts:

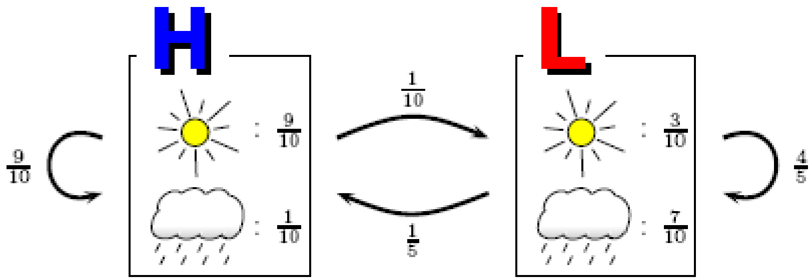
$A_{HH} = 1/2$	p(sun|H) = 1
$A_{HL} = 1/2$	p(rain|H) = 0
$A_{LH} = 1/2$	p(sun|L) = 1/2
$A_{LL} = 1/2$	p(rain|L) = 1/2
$\pi_H = 1$	
$\pi_L = 0$	

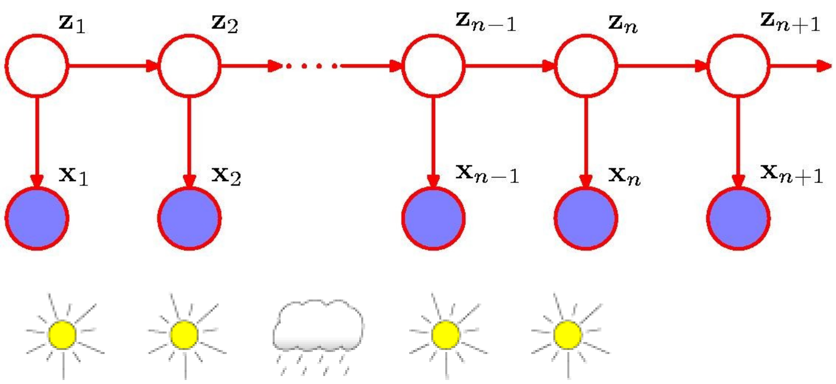
With pseudocounts:

$A_{HH} = 2/4$	p(sun|H) = 4/5
$A_{HL} = 2/4$	p(rain|H) = 1/5
$A_{LH} = 2/4$	p(sun|L) = 2/4
$A_{LL} = 2/4$	p(rain|L) = 2/4
$\pi_H = 2/3$	
$\pi_L = 1/3$	

Selecting "the right" parameters

What if only (several) sequences of observations **X**={$\mathbf{x}_1,\ldots,\mathbf{x}_n$} is given, i.e the corresponding latent states **Z**={$\mathbf{z}_1,\ldots,\mathbf{z}_n$} are unknown?

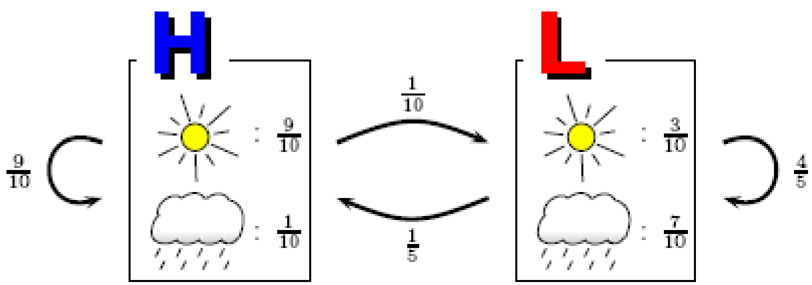


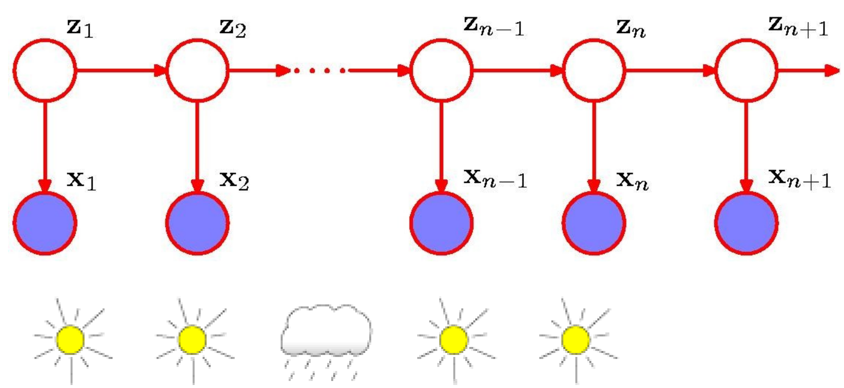


How should we set the model parameters, i.e. transitions **A**, **π**, and emission probabilities **Φ**, to make the given **X**'s most likely?

Selecting “the right” parameters

What if only (several) sequences of observations **X**={$\mathbf{x}_1,\ldots,\mathbf{x}_n$} is given, i.e the corresponding latent states **Z**={$\mathbf{z}_1,\ldots,\mathbf{z}_n$} are unknown?





How should we set the model parameters, i.e. transitions **A**, **π**, and emission probabilities **Φ**, to make the given **X**'s most likely?

Maximize $p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta)$ w.r.t. **θ** ...

Selecting "the right" parameters

What if only (several) sequences of observations **X**={**x**$_1$,...,**x**$_n$} is given, i.e the corresponding latent states **Z**={**z**$_1$,...,**z**$_n$} are unknown?

Direct maximization of the likelihood (or *log*-likelihood) is *hard* ...

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta)$$

How should we set the model parameters, i.e. transitions **A**, **π**, and emission probabilities **Φ**, to make the given **X**'s most likely?

Maximize $p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta)$ w.r.t. **θ** ...

Practical Solution - Viterbi training

A more "practical" thing to do is **Viterbi Training**:

1. Decide on some initial parameter $\boldsymbol{\theta}^0$

2. Find the most likely sequence of states **Z*** explaining **X** using the the Viterbi Algorithm and the current parameters $\boldsymbol{\theta}^i$

3. Update parameters to $\boldsymbol{\theta}^{i+1}$ by "counting" (with pseudo counts) according to (**X**,**Z***).

4. Repeat 2-3 until P(**X**,**Z*** | $\boldsymbol{\theta}^i$) is satisfactory (or the Viterbi sequence of states does not change).

Practical Solution - Viterbi training

A more "practical" thing to do is **Viterbi Training**:

1. Decide on some initial parameter $\boldsymbol{\theta}^0$
2. Find the most likely sequence of states **Z*** explaining **X** using the the Viterbi Algorithm and the current parameters $\boldsymbol{\theta}^i$
3. Update parameters to $\boldsymbol{\theta}^{i+1}$ by "counting" (with pseudo counts) according to (**X**,**Z***).
4. Repeat 2-3 until P(**X**,**Z*** | $\boldsymbol{\theta}^i$) is satisfactory (or the Viterbi sequence of states does not change).

Finds a (local) maximum of:

$$\mathrm{VIT}_{\mathbf{X}}(\Theta) = \max_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \Theta)$$

The identified parameters $\boldsymbol{\theta}^*$ is not a MLE of *p*(**X** | $\boldsymbol{\theta}$), but works "ok"

Summary: Training-by-Counting

Training-by-Counting: We are given a sequence of observations **X**={$\boldsymbol{x_1}$,…,$\boldsymbol{x_n}$} and the corresponding latent states **Z**={$\boldsymbol{z_1}$,…,$\boldsymbol{z_n}$}. We want to find a model:

$$\Theta^*_{\mathrm{TbC}} = \arg\max_{\Theta} p(\mathbf{X}, \mathbf{Z} | \Theta) = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{Z} | \Theta)$$

This can be done analytically by counting the frequency by which each transition and emission occur in the training data (**X, Z**).

If only **X**={$\boldsymbol{x_1}$,…,$\boldsymbol{x_n}$} is given, then we want to find a model:

$$\Theta^*_{\mathbf{X}} = \arg\max_{\Theta} p(\mathbf{X} | \Theta) = \arg\max_{\Theta} \log p(\mathbf{X} | \Theta)$$

Summary: Viterbi Training

Viterbi Training: We are given a sequence of observations **X**={$x_1,\ldots,x_n$}. Pick an initial set of parameters $\boldsymbol{\theta}^0_{\text{Vit}}$ and compute the best explanation of **X** under assumption of these parameters using the Viterbi algorithm:

$$\mathbf{Z}^0_{\text{Vit}} = \arg\max_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \Theta^0_{\text{Vit}}) = \arg\max_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \Theta^0_{\text{Vit}})$$

Compute $\boldsymbol{\theta}^1_{\text{vit}}$ from $\boldsymbol{\theta}^0_{\text{Vit}}$ and $\mathbf{Z}^0_{\text{Vit}}$ using TbC and iterate:

$$\Theta^1_{\text{Vit}} = \arg\max_{\Theta} p(\mathbf{X}, \mathbf{Z}^0_{\text{Vit}} | \Theta) = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{Z}^0_{\text{Vit}} | \Theta)$$

$$\mathbf{Z}^1_{\text{Vit}} = \arg\max_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \Theta^1_{\text{Vit}}) = \arg\max_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \Theta^1_{\text{Vit}})$$

$p(\mathbf{X} | \Theta^i_{\text{Vit}})$ is usually close to $p(\mathbf{X} | \Theta^*_{\mathbf{X}})$, but no guarantees

Expectation Maximization

EM Training: We are given a sequence of observations **X**={$x_1,\ldots,x_n$}. Pick an initial set of parameters $\boldsymbol{\theta}^0_{EM}$ and consider the expectation of log p(**X**, **Z** | **θ**) over **Z** (given **X** and $\boldsymbol{\theta}^0_{EM}$) as a function of **θ**:

$$EM_{\mathbf{X},\Theta^0_{\mathrm{EM}}}(\Theta) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\Theta^0_{\mathrm{EM}}}(\log p(\mathbf{X},\mathbf{Z}|\Theta)) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^0_{\mathrm{EM}}) \log p(\mathbf{X},\mathbf{Z}|\Theta)$$

For HMMs, we can find $\boldsymbol{\theta}^1_{EM}$ analytically, and iterate to get $\boldsymbol{\theta}^i_{EM}$:

$$\Theta^1_{EM} = \arg\max_{\Theta} \mathrm{EM}_{\mathbf{X},\Theta^0_{\mathrm{EM}}}(\Theta)$$

$p(\mathbf{X}|\Theta^i_{\mathrm{EM}})$ converges towards a (local) maximum of $p(\mathbf{X}|\Theta^*_{\mathbf{X}})$

Expectation Maximization

E-Step: Define the Q-function:

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

i.e. the expectation of log p(**X**, **Z** | **θ**) over **Z** (given **X** and **θ**$^{\text{old}}$) as a function of **θ**

M-Step: Maximize Q(**θ**, **θ**$^{\text{old}}$) w.r.t. **θ**

$$\text{EM}_{\mathbf{X},\Theta^{\text{old}}}(\Theta) = E_{\mathbf{Z}|\mathbf{X},\Theta^{\text{old}}}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

$$\Theta^* = \arg\max_{\Theta} \text{EM}_{\mathbf{X},\Theta^{\text{old}}}(\Theta)$$

When iterated, the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$ converges to a (local) maximum

Maximizing the likelihood

Direct maximization of the likelihood (or *log*-likelihood) is *hard* ...

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \qquad\qquad p(\mathbf{X}|\Theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

Assume that we have valid set of parameters $\boldsymbol{\theta}^{\text{old}}$, and that we want to estimate a set $\boldsymbol{\theta}$ which yields a better likelihood. We can write:

$$\log p(\mathbf{X}|\Theta) \quad = \quad \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}|\Theta)$$

Maximizing the likelihood

Direct maximization of the likelihood (or *log*-likelihood) is *hard* ...

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \qquad\qquad p(\mathbf{X}|\Theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

Assume that we have valid set of parameters θ^{old}, and that we want to estimate a set θ which yields a better likelihood. We can write:

This sums to 1 ...

$$\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}|\Theta)$$

Maximizing the likelihood

Direct maximization of the likelihood (or *log*-likelihood) is *hard* ...

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \qquad\qquad p(\mathbf{X}|\Theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

Assume that we have valid set of parameters $\boldsymbol{\theta}^{\text{old}}$, and that we want to estimate a set $\boldsymbol{\theta}$ which yields a better likelihood. We can write:

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}|\Theta) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) (\log p(\mathbf{X}, \mathbf{Z}|\Theta) - \log p(\mathbf{Z}|\mathbf{X}, \Theta)) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta)
\end{aligned}$$

Maximizing the likelihood

Direct maximization of the likelihood (or *log*-likelihood) is *hard ...*

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \qquad\qquad p(\mathbf{X}|\Theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

Assume that we have valid set of parameters $\boldsymbol{\theta}^{old}$, and that we want to estimate a set $\boldsymbol{\theta}$ which yields a better likelihood. We can write:

$$Q(\Theta, \Theta^{\mathrm{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{X}|\Theta) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) (\log p(\mathbf{X}, \mathbf{Z}|\Theta) - \log p(\mathbf{Z}|\mathbf{X}, \Theta)) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta)
\end{aligned}$$

The expectation (under $\boldsymbol{\theta}^{old}$) of the log-likelihood of the complete data (i.e. observations **X** and underlying states **Z**) as a function of $\boldsymbol{\theta}$

Maximizing the likelihood

Direct maximization of the likelihood (or *log*-likelihood) is *hard* ...

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \qquad\qquad p(\mathbf{X}|\Theta) = \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

Assume that we have valid set of parameters $\boldsymbol{\theta}^{\text{old}}$, and that we want to estimate a set $\boldsymbol{\theta}$ which yields a better likelihood. We can write:

$$\boxed{Q(\Theta, \Theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)}$$

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}|\Theta) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) (\log p(\mathbf{X}, \mathbf{Z}|\Theta) - \log p(\mathbf{Z}|\mathbf{X}, \Theta)) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta) \\
&= Q(\Theta, \Theta^{\text{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta)
\end{aligned}$$

Maximizing the likelihood

Assume that we have valid set of parameters **θ**^{old}, and that we want to estimate a set **θ** which yields a better likelihood. We have:

$$\log p(\mathbf{X}|\Theta) \quad = \quad Q(\Theta, \Theta^{\text{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta)$$

$$\log p(\mathbf{X}|\Theta^{\text{old}}) \quad = \quad Q(\Theta^{\text{old}}, \Theta^{\text{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}})$$

The increase of the log-likelihood can thus be written as:

$$\log p(\mathbf{X}|\Theta) - \log p(\mathbf{X}|\Theta^{\text{old}}) =$$

$$Q(\Theta, \Theta^{\text{old}}) - Q(\Theta^{\text{old}}, \Theta^{\text{old}}) + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}})}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

Maximizing the likelihood

Assume that we have valid set of parameters **θ^{old}**, and that we want to estimate a set **θ** which yields a better likelihood. We have:

$$\log p(\mathbf{X}|\Theta) = Q(\Theta, \Theta^{\mathrm{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta)$$

$$\log p(\mathbf{X}|\Theta^{\mathrm{old}}) = Q(\Theta^{\mathrm{old}}, \Theta^{\mathrm{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}})$$

The increase of the log-likelihood can thus be written as:

$$\log p(\mathbf{X}|\Theta) - \log p(\mathbf{X}|\Theta^{\mathrm{old}}) =$$

$$Q(\Theta, \Theta^{\mathrm{old}}) - Q(\Theta^{\mathrm{old}}, \Theta^{\mathrm{old}}) + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}})}{p(\mathbf{Z}|\mathbf{X}, \Theta)}$$

The relative entropy of p(**Z**|**X**,**θ^{old}**) relative to p(**Z**|**X**,**θ**), i.e. ≥ 0

Maximizing the likelihood

Assume that we have valid set of parameters **θ**^{old}, and that we want to estimate a set **θ** which yields a better likelihood. We have:

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= Q(\Theta, \Theta^{\text{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta) \\
\log p(\mathbf{X}|\Theta^{\text{old}}) &= Q(\Theta^{\text{old}}, \Theta^{\text{old}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}})
\end{aligned}
$$

The increase of the log-likelihood can thus be written as:

$$
\log p(\mathbf{X}|\Theta) - \log p(\mathbf{X}|\Theta^{\text{old}}) \geq Q(\Theta, \Theta^{\text{old}}) - Q(\Theta^{\text{old}}, \Theta^{\text{old}})
$$

By maximizing the expectation Q(**θ**, **θ**^{old}) w.r.t. **θ**, we do not decrease the likelihood, hence name *expectation maximization* ...

EM for HMMs

E-Step: Define the Q-function:

$$Q(\Theta, \Theta^{\mathrm{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\mathrm{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

i.e. the expectation of log p(**X**, **Z** | **θ**) over **Z** (given **X** and **θ**$^{\text{old}}$) as a function of **θ**

M-Step: Maximize Q(**θ**, **θ**$^{\text{old}}$) w.r.t. **θ**

For HMMs Q has a closed form and maximization can be performed explicitly. Iterate until no or little increase in likelihood is observed, or some maximum number of iterations is reached ...

When iterated, the likelihood *p*(**X**|**θ**) converges to a (local) maximum

EM for HMMs

Init: Pick “suitable” parameters (transition and emission probabilities). Observe that if a parameter is initialized to zero, it remains zero ...

E-Step: 1) Run the forward- and backward-algorithms with the current choice of parameters (to get the params of Q-func).

Stop?: 2) Compute the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$, if sufficient (or another stopping criteria is meet) then stop.

M-Step: 3) Compute new parameters using the values stored by the forward- and backward-algorithms. Repeat 1-3.

EM for HMMs

We want a closed form for $Q(\Theta, \Theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) &= p(\mathbf{z}_1|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \phi) \\
&= \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{n=2}^{N} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}} \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}
\end{aligned}
$$

EM for HMMs

We want a closed form for $Q(\Theta, \Theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) &= p(\mathbf{z}_1|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \phi) \\ &= \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{n=2}^{N} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}} \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}} \end{aligned}$$

$$p(\mathbf{z}_1|\pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \qquad p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}}$$

$$p(\mathbf{x}_n|\mathbf{z}_n, \phi) = \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}$$

EM for HMMs

We want a closed form for $Q(\Theta, \Theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) &= p(\mathbf{z}_1|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \phi) \\ &= \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{n=2}^{N} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}} \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}} \end{aligned}$$

Taking the log yields:

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} z_{1k} \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} z_{n-1,j} z_{nk} \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log p(\mathbf{x}_n|\phi_k)$$

EM for HMMs

We want a closed form for $Q(\Theta, \Theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) &= p(\mathbf{z}_1|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n, \phi) \\ &= \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{n=2}^{N} \prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}} \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}} \end{aligned}$$

Taking the log yields:

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} z_{1k} \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} z_{n-1,j} z_{nk} \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log p(\mathbf{x}_n|\phi_k)$$

Taking the expectation (under $\boldsymbol{\theta}^{\text{old}}$ and **X**) over **Z** yields Q($\boldsymbol{\theta}$, $\boldsymbol{\theta}^{\text{old}}$), i.e:

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} E(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} E(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} E(z_{nk}) \log p(\mathbf{x}_n|\phi_k)$$

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} E(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} E(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} E(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

E-Step: To calculate *Q*, we must compute the expectations $E(z_{1k})$, $E(z_{nk})$, and $E(z_{n-1,j} z_{nk})$. Consider the probabilities:

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{X}, \Theta^{\text{old}})$$

A *K*-vector where entry *k is* the prob $\gamma(z_{nk})$ of being in state k in the n'th step ...

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}, \Theta^{\text{old}})$$

A *KxK*-table where entry (*j,k*) *is* the prob $\xi(z_{n-1,j}\ z_{nk})$ of being in state *j* and *k* in the (n-1)'th and n'th step ...

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} E(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} E(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} E(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

E-Step: To calculate Q, we must compute the expectations $E(z_{1k})$, $E(z_{nk})$, and $E(z_{n-1,j} z_{nk})$. Consider the probabilities:

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{X}, \Theta^{\text{old}})$$

A *K*-vector where entry *k is* the prob $\gamma(z_{nk})$ of being in state k in the n'th step ...

binary variables

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}, \Theta^{\text{old}})$$

A *KxK*-table whe[illegible] *j*,*k*) *is* the prob $\xi(z_{n-1,j}\ z_{nk})$ of being in state *j* and *k* in the (n-1)'th and n'th step ...

Fact: The expectation of a binary variable *z* is just $p(z=1)$ ...

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} E(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} E(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} E(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

E-Step: To calculate Q, w
$E(z_{nk})$, and $E(z_{n-1,j}, z_{nk})$. Co

$$\mathrm{E}(z_{nk}) = \gamma(z_{nk})$$

$$\mathrm{E}(z_{n-1,j} z_{nk}) = \xi(z_{n-1,j} z_{nk})$$

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{X}, \Theta^{\text{old}})$$

prob $\gamma(z_{nk})$ of being in state k in the n'th step ...

binary variables

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}, \Theta^{\text{old}})$$

A *KxK*-table where
is the prob $\xi(z_{n-1,j}\ z_{nk})$ of being in state *j* and *k* in the (n-1)'th and n'th step ...

Fact: The expectation of a binary variable *z* is just *p*(*z*=1) ...

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

M-Step: If we assume discrete observables x_i, then maximizing the above w.r.t. **θ**, i.e. *A, π*, and *Φ,* yields:

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

M-Step: If we assume discrete observables x_i, then maximizing the above w.r.t. **θ**, i.e. *A, π*, and *Φ,* yields:

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j} z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j} z_{nl})} = \frac{\text{Expected number of transitions from state } j \text{ to state } k}{\text{Expected number of transitions from state } j \text{ to any state}}$$

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}$$

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

M-Step: If we assume discrete observables x_i, then maximizing the above w.r.t. **θ**, i.e. *A, π*, and *Φ,* yields:

$$\phi_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\text{Expected number of times symbol } i \text{ is emitted from state } k}{\text{Expected number of times a symbol is emitted from state } k}$$

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

M-Step: If we assume discrete observables x_i, then maximizing the above w.r.t. **θ**, i.e. A, π, and Φ, yields:

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \qquad \pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

EM for HMMs

$$Q(\Theta, \Theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(\mathbf{x}_n | \phi_k)$$

M-Step: If we assume discrete observables x_i, then maximizing the above w.r.t. **θ**, i.e. *A, π*, and *Φ,* yields:

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \qquad \pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

Compare this to the formulas when **X** and **Z** where given:

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

Computing γ and ξ

$$\begin{aligned}
\gamma(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}
\end{aligned}$$

$$\begin{aligned}
\xi(\mathbf{z}_{n-1},\mathbf{z}_n) &= p(\mathbf{z}_{n-1},\mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1},\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_{n-1})\beta(\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)}{p(\mathbf{X})}
\end{aligned}$$

Can be computed efficiently using the forward- and backward-algorithm

Computing the new parameters

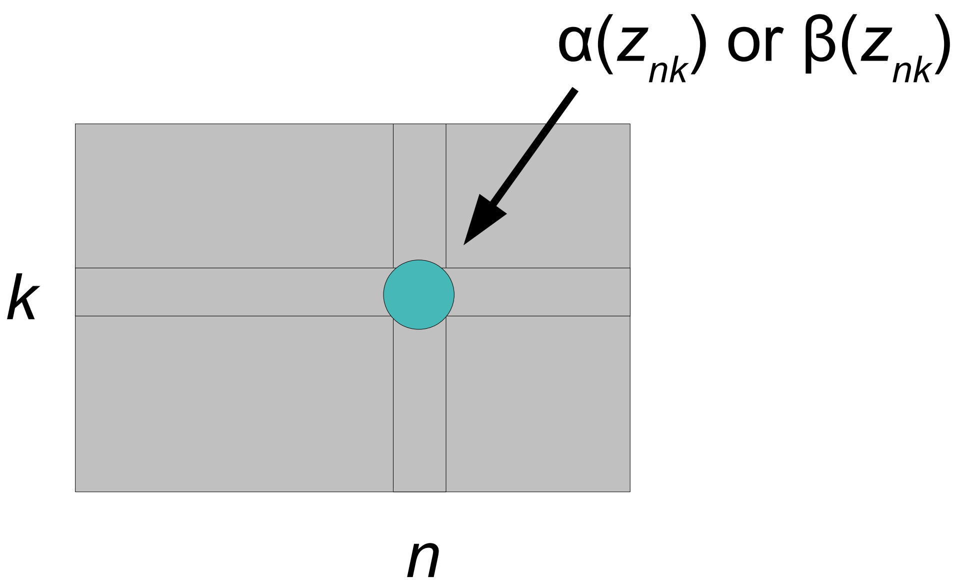
$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} = \frac{\alpha(z_{1k})\beta(z_{1k})/p(\mathbf{X})}{\sum_{j=1}^{K} \alpha(z_{1j})\beta(z_{1j})/p(\mathbf{X})} = \frac{\alpha(z_{1k})\beta(z_{1k})}{\sum_{j=1}^{K} \alpha(z_{1j})\beta(z_{1j})}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K}\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} = \frac{\sum_{n=2}^{N} \alpha(z_{n-1,j})\beta(z_{nk})p(\mathbf{x}_n|\phi_k)A_{jk}}{\sum_{l=1}^{K}\sum_{n=2}^{N} \alpha(z_{n-1,j})\beta(z_{nl})p(\mathbf{x}_n|\phi_l)A_{jl}}$$

$$\phi_{ik} = \frac{\sum_{n=1}^{N} \alpha(z_{nk})\beta(z_{nk})x_{ni}}{\sum_{n=1}^{N} \alpha(z_{nk})\beta(z_{nk})}$$

$$\gamma(\mathbf{z}_n) = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})}$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \frac{\alpha(\mathbf{z}_{n-1})\beta(\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)}{p(\mathbf{X})}$$



Computing the new parameters

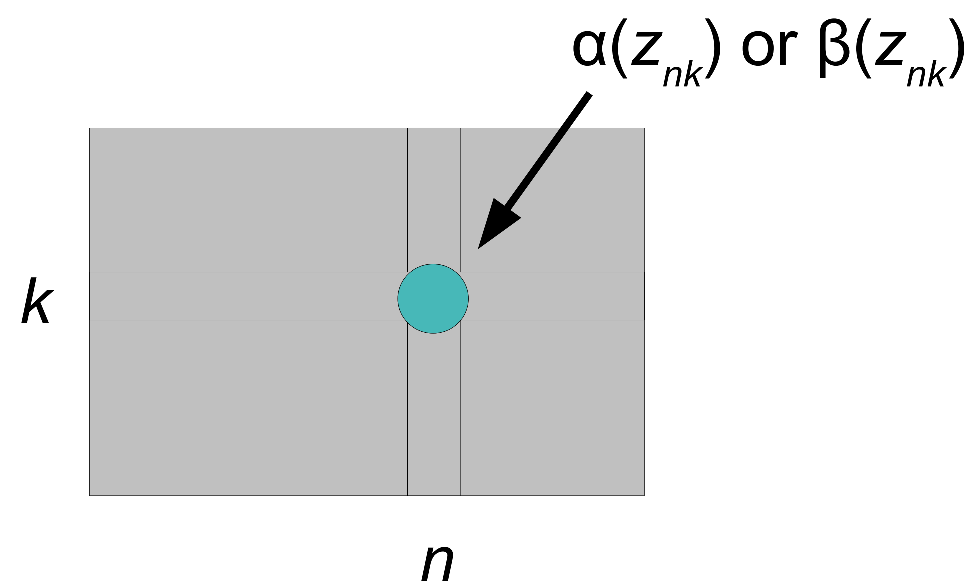
$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} = \frac{\alpha(z_{1k})\beta(z_{1k})/p(\mathbf{X})}{\sum_{j=1}^{K} \alpha(z_{1j})\beta(z_{1j})/p(\mathbf{X})} = \frac{\alpha(z_{1k})\beta(z_{1k})}{\sum_{j=1}^{K} \alpha(z_{1j})\beta(z_{1j})}$$

The old parameters

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K}\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} = \frac{\sum_{n=2}^{N} \alpha(z_{n-1,j})\beta(z_{nk})p(\mathbf{x}_n|\phi_k)A_{jk}}{\sum_{l=1}^{K}\sum_{n=2}^{N} \alpha(z_{n-1,j})\beta(z_{nl})p(\mathbf{x}_n|\phi_l)A_{jl}}$$

$$\phi_{ik} = \frac{\sum_{n=1}^{N} \alpha(z_{nk})\beta(z_{nk})x_{ni}}{\sum_{n=1}^{N} \alpha(z_{nk})\beta(z_{nk})}$$

The new parameters



EM for HMMs - Summary

Init: Pick “suitable” parameters (transition and emission probabilities). Observe that if a parameter is initialized to zero, it remains zero ...

E-Step: 1) Run the forward- and backward-algorithms with the current choice of parameters (to get t.he params of Q-func).

Stop?: 2) Compute the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$, if sufficient (or another stopping criteria is meet) then stop.

M-Step: 3) Compute new parameters using the values stored by the forward- and backward-algorithms. Repeat 1-3.

Running time per iteration:

$O(K^2N + KK + K^2NK + KDN)$, where D is number of observable symbols

By using memorization in 3), we can improve it to $O(K^2N + KDN)$

Using the scaled values in EM

$$
\begin{aligned}
\gamma(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \hat{\alpha}(\mathbf{z}_n)\hat{\beta}(\mathbf{z}_n)
\end{aligned}
$$

$$
\begin{aligned}
\xi(\mathbf{z}_{n-1},\mathbf{z}_n) &= p(\mathbf{z}_{n-1},\mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1},\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_{n-1})\beta(\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \hat{\alpha}(\mathbf{z}_{n-1})\hat{\beta}(\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)/c_n
\end{aligned}
$$

Can be computed using the modified forward- and backward-algorithm

Using the scaled val

$$p(\mathbf{X}) = \prod_{n=1}^{N} c_n$$

$$\alpha(\mathbf{z}_n) = \left(\prod_{m=1}^{n} c_m\right)\hat{\alpha}(\mathbf{z}_n)$$

$$\beta(\mathbf{z}_n) = \left(\prod_{m=n+1}^{N} c_m\right)\hat{\beta}(\mathbf{z}_n)$$

$$\begin{aligned}
\gamma(\mathbf{z}_n) &= p(\mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \hat{\alpha}(\mathbf{z}_n)\hat{\beta}(\mathbf{z}_n)
\end{aligned}$$

$$\begin{aligned}
\xi(\mathbf{z}_{n-1},\mathbf{z}_n) &= p(\mathbf{z}_{n-1},\mathbf{z}_n|\mathbf{X}) \\
&= \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1},\mathbf{z}_{n-1})p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_{n-1})\beta(\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)}{p(\mathbf{X})} \\
&= \hat{\alpha}(\mathbf{z}_{n-1})\hat{\beta}(\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)/c_n
\end{aligned}$$

Error in book

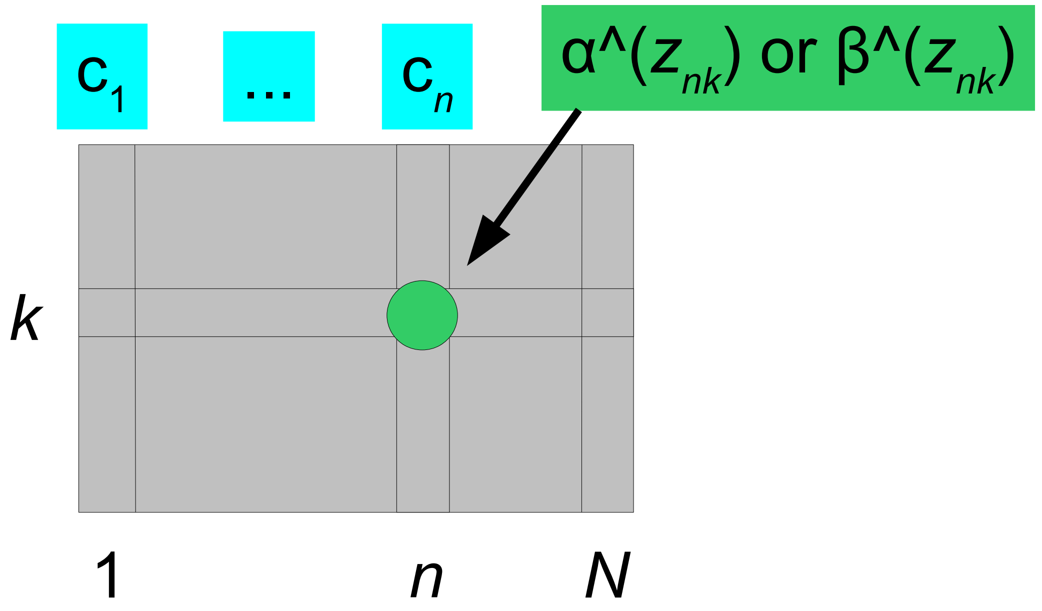
Can be computed using the modified forward- and backward-algorithm

Computing the new parameters

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} = \frac{\hat{\alpha}(z_{1k})\hat{\beta}(z_{1k})}{\sum_{j-1}^{K} \hat{\alpha}(z_{1k})\hat{\beta}(z_{1k})}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} = \frac{\sum_{n=2}^{N} \hat{\alpha}(z_{n-1,j})\hat{\beta}(z_{nk})p(\mathbf{x}_n|\phi_k)A_{jk}/c_n}{\sum_{l=1}^{K} \sum_{n=2}^{N} \hat{\alpha}(z_{n-1,j})\hat{\beta}(z_{nl})p(\mathbf{x}_n|\phi_l)A_{jl}/c_n}$$

$$\phi_{ik} = \frac{\sum_{n=1}^{N} \hat{\alpha}(z_{nk})\hat{\beta}(z_{nk})x_{ni}}{\sum_{n=1}^{N} \hat{\alpha}(z_{nk})\hat{\beta}(z_{nk})}$$



Summary

- Selecting parameters by **counting** to reflect a set of (**X**,**Z**)'s, i.e. if full information about observables and corresponding latent values is given.

- Selecting parameters by **Viterbi Training** or **Expectation Maximization** to reflect a set of **X**'s, i.e. if only information about observables is given.

Summary

- Selecting parameters by **counting** to reflect a set of (**X**,**Z**)'s, i.e. if full information about observables and corresponding latent values is given.

- Selecting parameters by **Viterbi Training** or **Expectation Maximization** to reflect a set of **X**'s, i.e. if only information about observables is given.

How to deal with multiple “training sequences”?

When multiple (*X*, Z)'s are given ...

Assume that (several) sequences of observations **X**={**x**$_1$,...,**x**$_n$} and corresponding latent states **Z**={**z**$_1$,...,**z**$_n$} are given ...

$$A_{jk} = \frac{\sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{z_{1k}}{\sum_{j=1}^{K} z_{1j}} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{n=1}^{N} z_{nk}}$$

... just sum each nominator and denominator over all (**X**,**Z**)'s, i.e. we divide total counts ...

$$A_{jk} = \frac{\sum_{(\mathbf{X},\mathbf{Z})} \sum_{n=2}^{N} z_{n-1,j} z_{nk}}{\sum_{(\mathbf{X},\mathbf{Z})} \sum_{l=1}^{K} \sum_{n=2}^{N} z_{n-1,j} z_{nl}} \qquad \pi_k = \frac{\sum_{(\mathbf{X},\mathbf{Z})} z_{1k}}{\sum_{(\mathbf{X},\mathbf{Z})} \sum_{j=1}^{K} z_{1j}}$$

$$\phi_{ik} = \frac{\sum_{(\mathbf{X},\mathbf{Z})} \sum_{n=1}^{N} z_{nk} x_{ni}}{\sum_{(\mathbf{X},\mathbf{Z})} \sum_{n=1}^{N} z_{nk}}$$

When multiple *X*'s are given ...

Assume that a set sequences of observations **X**={**x**$_1$,...,**x**$_n$} is given

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \qquad \pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

... just sum each nominator and denominator over all **X**'s, i.e. we divide total expectation, and we must run the forward- and backward algorithms for each training sequence **X** ...

$$A_{jk} = \frac{\sum_{\mathbf{X}} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{\mathbf{X}} \sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \qquad \pi_k = \frac{\sum_{\mathbf{X}} \gamma(z_{1k})}{\sum_{\mathbf{X}} \sum_{j=1}^{K} \gamma(z_{1j})}$$

$$\phi_{ik} = \frac{\sum_{\mathbf{X}} \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{\mathbf{X}} \sum_{n=1}^{N} \gamma(z_{nk})}$$

Summary: Training-by-Counting

Training-by-Counting: We are given a sequence of observations **X**={$x_1,\ldots,x_n$} and the corresponding latent states **Z**={$z_1,\ldots,z_n$}. We want to find a model:

$$\Theta^*_{\mathrm{TbC}} = \arg\max_{\Theta} p(\mathbf{X}, \mathbf{Z}|\Theta) = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

This can be done analytically by counting the frequency by which each transition and emission occur in the training data (**X, Z**).

Summary: Training-by-Counting

Training-by-Counting: We are given a sequence of observations $\mathbf{X}=\{x_1,\ldots,x_n\}$ and the corresponding latent states $\mathbf{Z}=\{z_1,\ldots,z_n\}$. We want to find a model:

$$\Theta^*_{\mathrm{TbC}} = \arg\max_{\Theta} p(\mathbf{X}, \mathbf{Z}|\Theta) = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

This can be done analytically by counting the frequency by which each transition and emission occur in the training data (**X, Z**).

If only $\mathbf{X}=\{x_1,\ldots,x_n\}$ is given, then we want to find a model:

$$\Theta^*_{\mathbf{X}} = \arg\max_{\Theta} p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta)$$

Finding $\boldsymbol{\theta}^*_{\mathbf{X}}$ is hard. We have seen two approaches.

Summary: Viterbi Training

Viterbi Training: We are given a sequence of observations **X**={$x_1,\ldots,x_n$}. Pick an initial set of parameters θ^0_{Vit} and compute the best explanation of **X** under assumption of these parameters using the Viterbi algorithm:

$$\mathbf{Z}^0_{\text{Vit}} = \arg\max_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \Theta^0_{\text{Vit}}) = \arg\max_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \Theta^0_{\text{Vit}})$$

Compute θ^1_{vit} from θ^0_{Vit} and $\mathbf{Z}^0_{Vit}$ using TbC and iterate:

$$\Theta^1_{\text{Vit}} = \arg\max_{\Theta} p(\mathbf{X}, \mathbf{Z}^0_{\text{Vit}} | \Theta) = \arg\max_{\Theta} \log p(\mathbf{X}, \mathbf{Z}^0_{\text{Vit}} | \Theta)$$

$$\mathbf{Z}^1_{\text{Vit}} = \arg\max_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \Theta^1_{\text{Vit}}) = \arg\max_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \Theta^1_{\text{Vit}})$$

$p(\mathbf{X} | \Theta^i_{\text{Vit}})$ is usually close to $p(\mathbf{X} | \Theta^*_{\mathbf{X}})$, but no guarantees

Summary: Expectation Maximization

EM Training: We are given a sequence of observations $\mathbf{X}=\{x_1,\ldots,x_n\}$. Pick an initial set of parameters $\boldsymbol{\theta}^0_{EM}$ and consider the expectation of log p($\mathbf{X}$, $\mathbf{Z}$ | $\boldsymbol{\theta}$) over $\mathbf{Z}$ (given $\mathbf{X}$ and $\boldsymbol{\theta}^0_{EM}$) as a function of $\boldsymbol{\theta}$:

$$EM_{\mathbf{X},\Theta^0_{\mathrm{EM}}}(\Theta) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\Theta^0_{\mathrm{EM}}}(\log p(\mathbf{X},\mathbf{Z}|\Theta)) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^0_{\mathrm{EM}}) \log p(\mathbf{X},\mathbf{Z}|\Theta)$$

For HMMs, we can find $\boldsymbol{\theta}^1_{EM}$ analytically, and iterate to get $\boldsymbol{\theta}^i_{EM}$:

$$\Theta^1_{EM} = \arg\max_{\Theta} \mathrm{EM}_{\mathbf{X},\Theta^0_{\mathrm{EM}}}(\Theta)$$

$p(\mathbf{X}|\Theta^i_{\mathrm{EM}})$ converges towards a (local) maximum of $p(\mathbf{X}|\Theta^*_{\mathbf{X}})$