Royal Institute of Technology

MACHINE LEARNING 2 – EM,HMMS (CH 17)

Lecture 4

RELATIONS BETWEEN LOG-LIKELIHOODS AND Q-TERMS

Theorem: for $\boldsymbol{\theta}' = \text{argmax}_{\boldsymbol{\theta}''} Q(\boldsymbol{\theta}'', \boldsymbol{\theta})$

$$\log p(\mathcal{D}|\boldsymbol{\theta}') \geq Q(\boldsymbol{\theta}', \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - R(\boldsymbol{\theta}, \boldsymbol{\theta}) = \log p(\mathcal{D}|\boldsymbol{\theta})$$

So by maximizing Q-term (through ESS) we monotonically increase the likelihood.

The Q-term may not increase in every step!

Expected complete: of all data

$$\begin{aligned}
&\log p(\mathcal{D}|\boldsymbol{\theta}') \\
&= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') \\
&= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\geq \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&= \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n \\
&= \underbrace{\sum_n Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta})}_{Q(\boldsymbol{\theta}'; \boldsymbol{\theta})} - \underbrace{\sum_n R_n(\boldsymbol{\theta}; \boldsymbol{\theta})}_{R_n(\boldsymbol{\theta}; \boldsymbol{\theta})}
\end{aligned}$$

GMM EXPECTED LOG LIKELIHOOD

$$\begin{aligned}
E_{p(z_n|\boldsymbol{x}_n, \boldsymbol{\theta})}\left[l(\theta'; \mathcal{D})\right] &= E_{p(z_n|\boldsymbol{x}_n, \boldsymbol{\theta})}\left[\log \prod_n p(\boldsymbol{x}_n, z_n|\theta')\right] \\
&= \sum_n E\left[\log \prod_c (\pi'_c p(\boldsymbol{x}_n|z_n = c, \theta')^{I(z_n=c)}\right] \\
&= \sum_n \sum_c E\left[I(z_n = c) \log(\pi'_c p(\boldsymbol{x}_n|\theta'_c)\right] \\
&= \sum_n \sum_c p(z_n = c|\boldsymbol{x}_n, \boldsymbol{\theta}) \log(\pi'_c p(\boldsymbol{x}_n|\theta'_c)) \\
&= \sum_n \sum_c r_{nc} \log \pi'_c + \sum_n \sum_c r_{nc} \log p(\boldsymbol{x}_n|\theta'_c)
\end{aligned}$$

where for a one dim Gaussian $\theta_c = (\mu_c, \sigma_c^2)$

DGM WITH CATEGORICAL CPDS - LIKELIHOOD

$$\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$$

- Let $\boldsymbol{\theta}'_{vk\boldsymbol{l}} = p(X_v = k | \boldsymbol{X}_{\mathrm{pa}(v)} = \boldsymbol{l}, \boldsymbol{\theta}')$, then

$$p(X_v = k | \boldsymbol{X}_{\mathrm{pa}(v)} = \boldsymbol{l}, \boldsymbol{\theta}') = \prod_{k,\boldsymbol{l}} (\boldsymbol{\theta}'_{vk\boldsymbol{l}})^{I(X_v = k, \boldsymbol{X}_{\mathrm{pa}(v)} = l)}$$

- The loglikelihood log p(D | Θ´) equals

$$\sum_n \sum_{v,k,\boldsymbol{l}} I(x_{nv} = k, \boldsymbol{x}_{n,\mathrm{pa}(v)} = \boldsymbol{l}) \log \boldsymbol{\theta}'_{v,k,\boldsymbol{l}} = \sum_{v,k,\boldsymbol{l}} N_{v,k,\boldsymbol{l}} \log \boldsymbol{\theta}'_{v,k,\boldsymbol{l}}$$

where $N_{v,k,\boldsymbol{l}} = \sum_n I(x_{nv} = k, \boldsymbol{x}_{n,\mathrm{pa}(v)} = \boldsymbol{l})$

EXPECTED COMPLETE LOG-LIKELIHOOD

- The expected complete log-likelihood is

$$\sum_n \sum_{v,k,\boldsymbol{l}} E_{p(\boldsymbol{X}|x_n,\theta)}[I(X_v = k, \boldsymbol{X}_{\mathrm{pa}(v)} = \boldsymbol{l})] \log \boldsymbol{\theta}'_{v,k,\boldsymbol{l}} = \sum_{v,k,\boldsymbol{l}} \overline{N}_{v,k,\boldsymbol{l}} \log \boldsymbol{\theta}'_{v,k,\boldsymbol{l}}$$

where

$$\overline{N}_{v,k,\boldsymbol{l}} = \sum_n E_{p(\boldsymbol{X}|\boldsymbol{x}_n,\boldsymbol{\theta})} \left[I(X_v = k, \boldsymbol{X}_{\mathrm{pa}(v)} = \boldsymbol{l}), \right] = \sum_n p(X_v = k, \boldsymbol{X}_{\mathrm{pa}(v)} = \boldsymbol{l} | \boldsymbol{x}_n, \boldsymbol{\theta})$$

Can be computed using GM inference

- We can independently maximize each

$$\sum_k \overline{N}_{v,k,\boldsymbol{l}} \log \boldsymbol{\theta}'_{v,k,\boldsymbol{l}}$$

- Done by setting

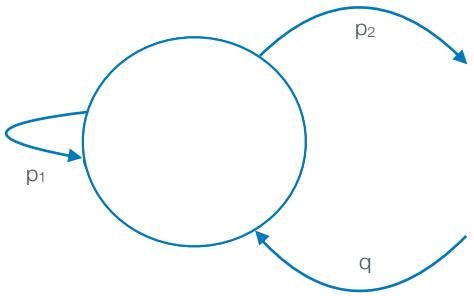
$$\boldsymbol{\theta}'_{v,k,\boldsymbol{l}} = \frac{\overline{N}_{v,k,\boldsymbol{l}}}{\sum_k \overline{N}_{v,k,\boldsymbol{l}}}$$

MAXIMIZED INDEPENDENTLY

MARKOV CHAINS (DISCRETE)

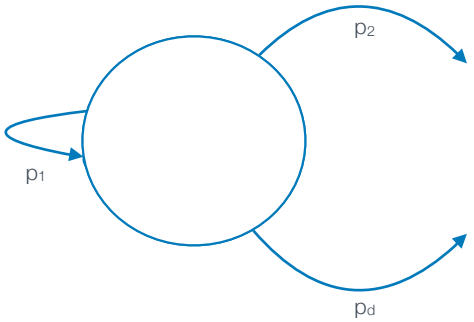
★Directed graph with transition probabilities

★ We observe the sequence of visited vertices



MARKOV CHAINS (DISCRETE)

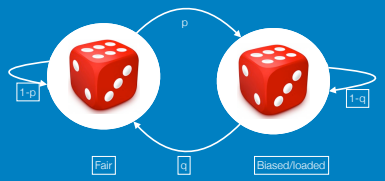
Probabilities on outgoing edges sum to one



$\sum_{i \in [d]} p_i = 1$

WHAT AN HMM DOES

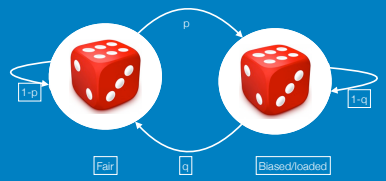
```
Rolls:   6641532161621152346532143566342616552
Die:     LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFF
```



- ★ Starts in the state z_1
- ★ When in state z_t
  - outputs $p(x_t|z_t)$
  - moves to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

WHAT AN HMM DOES

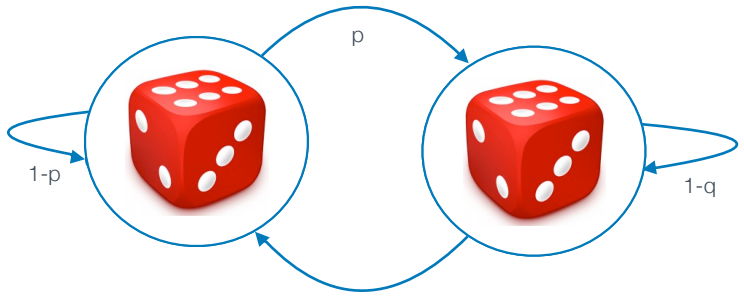
```
Rolls:   6641532161621152346532143566342616552
Die:     LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFF
```



- ★ Starts in the state z_1
- ★ When in state z_t
  - outputs $p(x_t|z_t)$
  - moves to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

The parameters

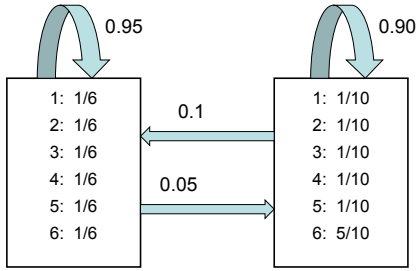
THE OCCASIONALLY DISHONEST CASINO



★ We observe the sequence of dice outcomes of visited vertices

```
Rolls:   6641532161621152346532143566342616552342323151424641566632 46
Die:     LLLLLLLLLLLLLFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFLLLLLLL
```

EMISSION DISTRIBUTIONS



THE JOINT DISTRIBUTION

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})p(\boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1)\left(\prod_{t=2}^{T} p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})\right)\left(\prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{z}_t)\right)$$

Categorial or Gaussian

- ★ Starts in the state z_1
- ★ When in state z_t
  - emits $p(x_t|z_t)$
  - transits to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

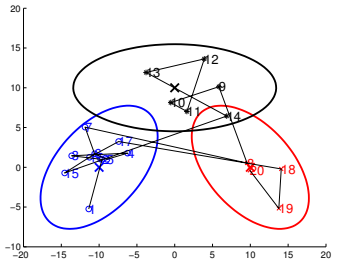
THE JOINT DISTRIBUTION

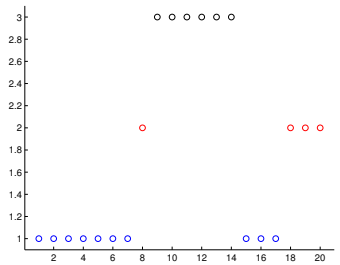
$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T})p(\boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1)\left(\prod_{t=2}^{T} p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})\right)\left(\prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{z}_t)\right)$$

Categorial or Gaussian

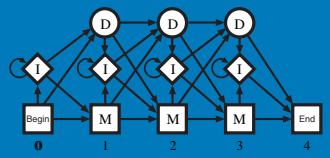
- ★ Starts in the state z_1
- ★ When in state z_t
  - emits $p(x_t|z_t)$
  - transits to $p(z_{t+1}|z_t)$
- ★ Stops after a fixed number of steps or when reaching a stop step

GAUSSIAN EMISSIONS AND HIDDEN STATES



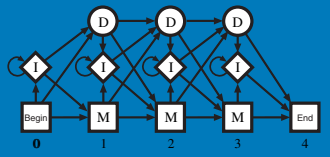


LAYERED OR NOT



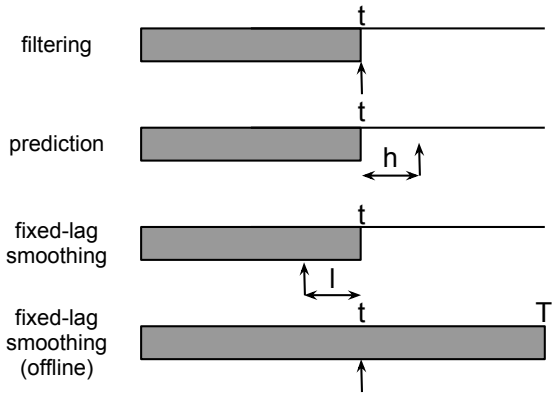
APPLICATIONS OF HMMS

```
      x x . . . x
bat   A G - - - C
rat   A - A G - C
cat   A G - A A -
gnat  - - A A A C
goat  A G - - - C
      1 2 . . . 3
```



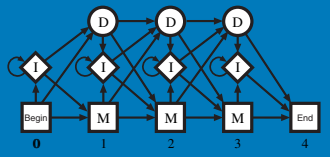
- Automatic speech recognition
- Activity recognition
- Part of speech tagging
- Gene finding
- Protein sequence alignment

TERMINOLOGY X ABOVE Z BELOW



INFERENCE TYPES

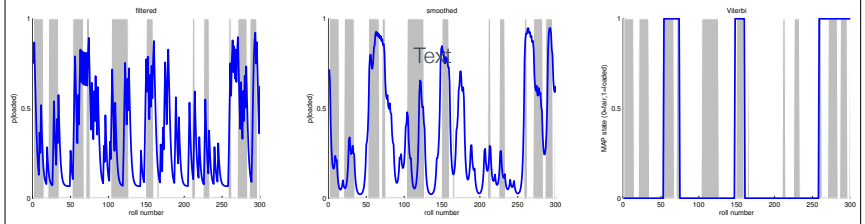
```
      x x . . . x
bat   A G - - - C
rat   A - A G - C
cat   A G - A A -
gnat  - - A A A C
goat  A G - - - C
      1 2 . . . 3
```



- Filtering: $p(z_t|x_{1:t})$, online
- Smoothing (MAP): $p(z_t|x_{1:T})$ offline
- Fixed lag smoothing: $p(z_t|x_{1:t+l})$
- Prediction: $p(z_{t+h}|x_{1:t})$
- Viterbi (MAP) argmax $p(z_{1:T}|x_{1:T})$
- Posterior samples: $\sim p(z_{1:T}|x_{1:T})$
- Probability of data: $p(x_{1:T})$
- Parameters: given D & struct.
- Structure and param.: given D

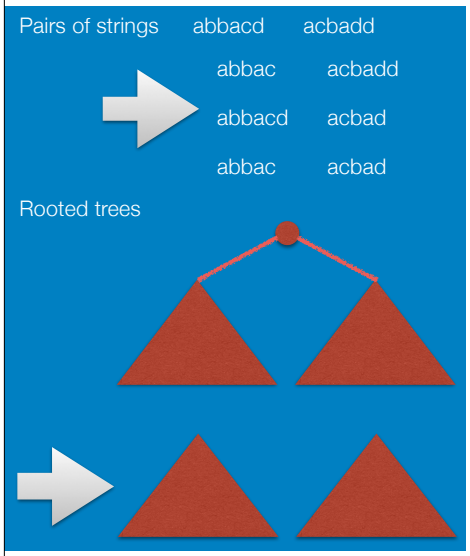
INFERENCE IN THE OCCASIONALLY DISHONEST CASINO

Grey regions are states corresponding to biased die



Text

- Filtering: $p(z_t|x_{1:t})$, online
- Smoothing, MAP state: $p(z_t|x_{1:T})$ offline
- Viterbi, MAP path argmax $p(z_{1:T}|x_{1:T})$



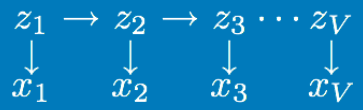
DP

- What is a subproblem?
- What is a subsolution?
- How do we decompose into smaller subproblems?
- How do we combine subsolutions into larger?
- How do we enumerate?
- How many and what time?

DP

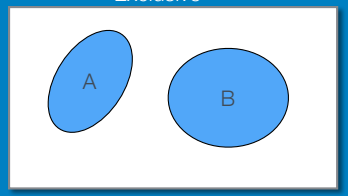
Polynomial many	What is a subproblem?
	What is a subsolution?
Polynomial time	How do we decompose into smaller subproblems?
Polynomial time	How do we combine subsolutions into larger?
Polynomial time	How do we enumerate?
Polynomial time overall	How many and what time?

SPECIAL CASE: HIDDEN MARKOV MODEL (HMM)

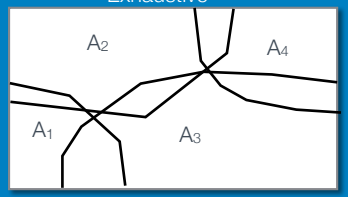


- Z_i hidden
- X_i observable
- Hidden often not observable when training, never when applying

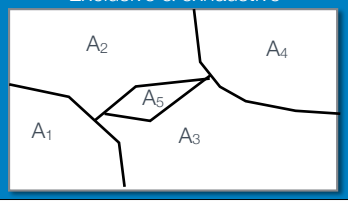
Exclusive



Exhaustive



Exclusive & exhaustive



EXCLUSIVE & EXHAUSTIVE

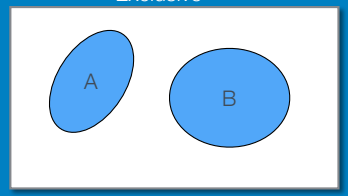
- Exclusive

$$p(A \text{ or } B) = p(A) + P(B)$$

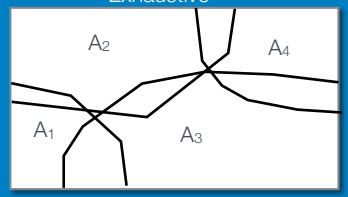
- Exclusive & exhaustive

$$\sum_i p(A_i) = 1$$

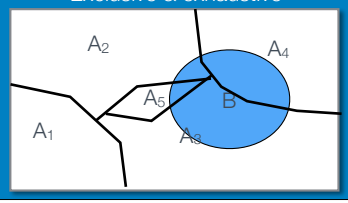
Exclusive



Exhaustive



Exclusive & exhaustive



EXCLUSIVE & EXHAUSTIVE

- Exclusive

$$p(A \text{ or } B) = p(A) + P(B)$$

- Exclusive & exhaustive

$$p(B) = \sum_i p(B, A_i) = \sum_i p(A_i)p(B|A_i)$$

$$p(\boldsymbol{x}_{1:T}) = \sum_{\boldsymbol{z}_{1:T}} p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})$$

$$f_t(k) := p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)$$

FORWARD

- Joint is easy to express
- The sum has exponentially many terms
- The forward variable, f_t, can be computed with DP

$$p(\boldsymbol{x}_{1:T}) = \sum_{\boldsymbol{z}_{1:T}} p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T})$$

$$f_t(k) := p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)$$

$$\begin{aligned}
f_t(k) &= \sum_l p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_{t-1} = l, \boldsymbol{Z}_t = k) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1} | \boldsymbol{Z}_{t-1} = l)}_{\text{emission}} \underbrace{p(\boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l)}_{\text{transition}}
\end{aligned}$$

FORWARD

- Joint is easy to express
- The sum has exponentially many terms
- The forward variable, f_t, can be computed with DP

Zoom in

$$
\begin{aligned}
f_t(k) &= \sum_l p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_{t-1} = l, \boldsymbol{Z}_t = k) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1} | \boldsymbol{Z}_{t-1} = l)}_{\text{emission}} \underbrace{p(\boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l)}_{\text{transition}}
\end{aligned}
$$

Zoom in

K - states

T - length of observation

$$
\begin{aligned}
f_t(k) &= \sum_l p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_{t-1} = l, \boldsymbol{Z}_t = k) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1} | \boldsymbol{Z}_{t-1} = l)}_{\text{emission}} \underbrace{p(\boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l)}_{\text{transition}}
\end{aligned}
$$

T, K

K

Total time $O(TK^2)$

Zoom in

K - states

T - length of observation

$$
\begin{aligned}
f_t(k) &= \sum_l p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_{t-1} = l, \boldsymbol{Z}_t = k) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l p(\boldsymbol{x}_{1:t-2}, \boldsymbol{Z}_{t-1} = l) p(\boldsymbol{x}_{t-1}, \boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l) \\
&= \sum_l \underbrace{f_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{x}_{t-1} | \boldsymbol{Z}_{t-1} = l)}_{\text{emission}} \underbrace{p(\boldsymbol{Z}_t = k | \boldsymbol{Z}_{t-1} = l)}_{\text{transition}}
\end{aligned}
$$

T, K

Constant

Total time $O(TK^2)$

If layered, $O(TK)$

$$p(\boldsymbol{x}_{1:T}) = \sum_k p(\boldsymbol{x}_{1:T}, z_T = k)$$

$$f_t(k) = p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)$$

In general, (e.g. t=T)

$$p(\boldsymbol{x}_{1:t}) = \sum_k f_t(k) p(\boldsymbol{x}_t | \boldsymbol{Z}_t = k)$$

OBSERVATION PROBABILITY

- The final probability is easily obtained

FILTERING

$$p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:t})$$

- Filtering: $p(z_t|x_{1:t})$, online

FILTERING

$$\begin{aligned} p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:t}) &= \frac{p(\boldsymbol{x}_{1:t}, \boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:t})} \\ &= \frac{p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k)p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:t})} \\ &= \frac{f_t(k)p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:t})} \end{aligned}$$

emission

data probability

- Filtering: $p(z_t|x_{1:t})$, online

BACKWARDS

$$b_t(k) := p(\boldsymbol{x}_{t+1:T}|\boldsymbol{Z}_t = k)$$

- DP also for the backward variable b_t

BACKWARDS

$$b_t(k) := p(\boldsymbol{x}_{t+1:T}|\boldsymbol{Z}_t = k)$$

- DP also for the backward variable b_t

$$\begin{aligned} b_t(k) &= \sum_l p(\boldsymbol{x}_{t+1:T}, \boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k) \\ &= \sum_l p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k)p(\boldsymbol{x}_{t+1:T}|\boldsymbol{Z}_{t+1} = l) \\ &= \sum_l p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k)p(\boldsymbol{x}_{t+2:T}|\boldsymbol{Z}_{t+1} = l)p(\boldsymbol{x}_{t+1}|\boldsymbol{Z}_{t+1} = l) \\ &= \sum_l \underbrace{p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k)}_{\text{transition}} \underbrace{b_{t+1}(l)}_{\text{"smaller"}} \underbrace{p(\boldsymbol{x}_{t+1}|\boldsymbol{Z}_{t+1} = l)}_{\text{emission}} \end{aligned}$$

BACKWARDS

$$b_t(k) := p(\boldsymbol{x}_{t+1:T}|\boldsymbol{Z}_t = k)$$

- DP also for the backward variable b_t

$$\begin{aligned} b_t(k) &= \sum_l p(\boldsymbol{x}_{t+1:T}, \boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k) \\ &= \sum_l p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k) p(\boldsymbol{x}_{t+1:T}|\boldsymbol{Z}_{t+1} = l) \\ &= \sum_l p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k) p(\boldsymbol{x}_{t+2:T}|\boldsymbol{Z}_{t+1} = l) p(\boldsymbol{x}_{t+1}|\boldsymbol{Z}_{t+1} = l) \\ &= \sum_l \underbrace{p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k)}_{\text{transition}} \underbrace{b_{t+1}(l)}_{\text{"smaller"}} \underbrace{p(\boldsymbol{x}_{t+1}|\boldsymbol{Z}_{t+1} = l)}_{\text{emission}} \end{aligned}$$

Running time analysis as before

SMOOTHING

$$p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:T})$$

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

SMOOTHING

$$\begin{aligned} p(\boldsymbol{Z}_t = k|\boldsymbol{x}_{1:T}) &= \frac{p(\boldsymbol{x}_{1:T}, \boldsymbol{Z}_t = k)}{p(\boldsymbol{x}_{1:T})} \\ &\propto p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k) p(\boldsymbol{x}_{t:T}|\boldsymbol{Z}_t = k) \\ &= f_t(k) \underbrace{p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}_{\text{emission}} b_t(k) \end{aligned}$$

- Smoothing, MAP state: $p(z_t|x_{1:T})$, offline

TWO SLICED SMOOTH MARGINALS - MARGINAL OVER PAIRS OF STATES

$$p(\boldsymbol{Z}_t = k, \boldsymbol{Z}_{t+1} = l|\boldsymbol{x}_{1:T})$$

- Can be computed from forward and backward similarly

VITERBI

We want
$\mathbf{argmax}_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$

Not!
$(\mathrm{argmax}_{\boldsymbol{z}_1} p(\boldsymbol{z}_1|\boldsymbol{x}_{t+1:T}), \ldots, \mathrm{argmax}_{\boldsymbol{z}_T} p(\boldsymbol{z}_T|\boldsymbol{x}_{t+1:T}))$

Viterbi variable

$$v_t(k) := \max_{\boldsymbol{z}_{1:t-1}} p(\boldsymbol{z}_{1:t-1}, \boldsymbol{Z}_t = k, \boldsymbol{x}_{1:t})$$

Gives what we want

- MAP path
- Viterbi learning: used, as approximation, to speed up parameter learning
- Again DP now with Viterbi variable
- For the path, use back pointers

VITERBI

We want
$\mathbf{argmax}_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$

Not!
$(\mathrm{argmax}_{\boldsymbol{z}_1} p(\boldsymbol{z}_1|\boldsymbol{x}_{t+1:T}), \ldots, \mathrm{argmax}_{\boldsymbol{z}_T} p(\boldsymbol{z}_T|\boldsymbol{x}_{t+1:T}))$

Viterbi variable

$$v_t(k) := \max_{\boldsymbol{z}_{1:t-1}} p(\boldsymbol{z}_{1:t-1}, \boldsymbol{Z}_t = k, \boldsymbol{x}_{1:t})$$

Gives what we want

$$\begin{aligned}
&\max_{\boldsymbol{z}_{1:T}} p(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}) \\
&= \max_{\boldsymbol{z}_{1:T-1}, k} p(\boldsymbol{z}_{1:T-1}, \boldsymbol{Z}_T = k|\boldsymbol{x}_{1:T}) \\
&= \max_{\boldsymbol{z}_{1:T-1}, k} \frac{p(\boldsymbol{z}_{1:T-1}, \boldsymbol{Z}_T = k, \boldsymbol{x}_{1:T})}{p(\boldsymbol{x}_{1:T})} \\
&= \max_{\boldsymbol{z}_{1:T-1}, k} \frac{v_T(k)}{p(\boldsymbol{x}_{1:T})}
\end{aligned}$$

- MAP path
- Viterbi learning: used, as approximation, to speed up parameter learning
- Again DP now with viterbi variable
- For the path, use back pointers

DP FOR VITERBI

$$\begin{aligned}
v_t(k) &= \max_{\boldsymbol{z}_{1:t-1}} p(\boldsymbol{z}_{1:t-1}, \boldsymbol{Z}_t = k, \boldsymbol{x}_{1:t}) \\
&= \max_l \max_{\boldsymbol{z}_{1:t-2}} p(\boldsymbol{z}_{1:t-2}, \boldsymbol{Z}_{t-1} = l, \boldsymbol{Z}_t = k, \boldsymbol{x}_{1:t}) \\
&= \max_l \max_{\boldsymbol{z}_{1:t-2}} p(\boldsymbol{z}_{1:t-2}, \boldsymbol{Z}_{t-1} = l, \boldsymbol{x}_{1:t-1}) p(\boldsymbol{Z}_t = k, \boldsymbol{x}_t|\boldsymbol{Z}_{t-1} = l) \\
&= \max_l \max_{\boldsymbol{z}_{1:t-2}} \underbrace{v_{t-1}(l)}_{\text{smaller}} \underbrace{p(\boldsymbol{Z}_t = k|\boldsymbol{Z}_{t-1} = l)}_{\text{transition}} \underbrace{p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}_{\text{emission}}
\end{aligned}$$

- This gives the probability of the MAP path
- Using backpointers the path can be obtained

FORWARD FILTERING, BACKWARDS SAMPLING

$$\boldsymbol{z}^s_{1:T} \sim p(\boldsymbol{Z}_{1:T}|\boldsymbol{x}_{1:T})$$

$$\begin{aligned}
p(\boldsymbol{Z}_t = k|\boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_{1:t}) &= \frac{p(\boldsymbol{Z}_t = k, \boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_{1:t})}{p(\boldsymbol{x}_{1:t}, \boldsymbol{Z}_{t+1} = l)} \\
&= \frac{p(\boldsymbol{x}_{1:t-1}, \boldsymbol{Z}_t = k) p(\boldsymbol{Z}_{t+1} = l, \boldsymbol{x}_t|\boldsymbol{Z}_t = k)}{f_{t+1}(l)} \\
&= \frac{f_t(k) p(\boldsymbol{Z}_{t+1} = l|\boldsymbol{Z}_t = k) p(\boldsymbol{x}_t|\boldsymbol{Z}_t = k)}{f_{t+1}(l)}
\end{aligned}$$

- Sample from posterior
- Sample in order $z_T, \ldots, z_1$
- Start somewhat differently

LEARNING TRANSITION AND EMISSION PARAMETERS - FULLY OBSERVED DATA

- Parameters
  - transition $A_{kl} = p(Z_t = l | Z_{t-1} = k)$
  - z_0 always q* and A_{q^*k} is start probability
  - emission
    $$B_{ks} = p(X_t = s | Z_t = k)$$
- Likelihood

$$L(\boldsymbol{\theta}; \mathcal{D}) = \prod_{n=1}^{N} \prod_{t=1}^{T} \left[\prod_{k,s} B_{ks}^{I(x_{n,t}=s, z_{n,t}=k)} \prod_{k,l} A_{kl}^{I(z_{n,t-1}=k, z_{n,t}=l)} \right]$$

LEARNING TRANSITION AND EMISSION PARAMETERS - FULLY OBSERVED DATA

- Parameters
  - transition $A_{kl} = p(Z_t = l | Z_{t-1} = k)$
  - z_0 always q* and A_{q^*k} is start probability
  - emission
    $$B_{ks} = p(X_t = s | Z_t = k)$$
- Loglikelihood

$$l(\boldsymbol{\theta}; \mathcal{D}) = \sum_{n=1}^{N} \sum_{t=1}^{T} [\sum_{k,s} I(x_{n,t} = s, z_{n,t} = k) \log B_{ks} + \sum_{k,l} I(z_{n,t-1} = k, z_{n,t} = l) \log A_{kl}]$$

MAXIMIZING LOG-LIKELIHOOD - COMPLETE DATA

$$l(\boldsymbol{\theta}; \mathcal{D}) = \sum_{n=1}^{N} \sum_{t=1}^{T} \left[\sum_{k,s} I(x_{n,t} = s, z_{n,t} = k) \log B_{ks} + \sum_{k,l} I(z_{n,t-1} = k, z_{n,t} = l) \log A_{kl} \right]$$

$$= \sum_{k,s} \left[\underbrace{\sum_{n=1}^{N} \sum_{t=1}^{T} I(x_{n,t} = s, z_{n,t} = k)}_{M_{k,s}} \right] \log B_{ks} + \sum_{k,l} \left[\underbrace{\sum_{n=1}^{N} \sum_{t=1}^{T} I(z_{n,t-1} = k, z_{n,t} = l)}_{N_{k,l}} \right] \log A_{kl}$$

- Maximized by

$$B_{ks} = M_{k,s} / \sum_{s} M_{k,s} = M_{k,s} / N_k \quad \text{and} \quad A_{kl} = N_{k,l} / \sum_{l} N_{k,l} = N_{k,l} / N_k$$

HIDDEN VARIABLES - ONE EM-STEP MAXIMIZING Q

$$\sum_{n=1}^{N} E_{p(\boldsymbol{Z}|\boldsymbol{x}_n, \boldsymbol{\theta})} [l(\theta'; \boldsymbol{Z}, \boldsymbol{x}_n)]$$

$$= \sum_{n=1}^{N} E_{p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n, \boldsymbol{\theta})} [l(\theta'; \boldsymbol{Z}, \boldsymbol{X})]$$

$$= \sum_{n=1}^{N} E_{p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{X}=\boldsymbol{x}_n, \boldsymbol{\theta})} \left[\sum_{t=1}^{T} \left[\sum_{k,s} I(\boldsymbol{X}_t = s, \boldsymbol{Z}_t = k) \log B'_{ks} + \sum_{k,l} I(\boldsymbol{Z}_{t-1} = k, \boldsymbol{Z}_t = l) \log A'_{kl} \right] \right]$$

$$= \sum_{k,s} \left[\underbrace{\sum_{n=1}^{N} \sum_{t=1}^{T} p(\boldsymbol{X}_t = s, \boldsymbol{Z}_t = k) | \boldsymbol{X} = \boldsymbol{x}_n, \boldsymbol{\theta})}_{\overline{M}_{k,s}} \right] \log B'_{ks}$$

$$+ \sum_{k,l} \left[\underbrace{\sum_{n=1}^{N} \sum_{t=1}^{T} p(\boldsymbol{Z}_{t-1} = k, \boldsymbol{Z}_t = l | \boldsymbol{X} = \boldsymbol{x}_n, \boldsymbol{\theta})}_{\overline{N}_{k,l}} \right] \log A'_{kl}$$

LEARNING TRANSITION AND EMISSION PARAMETERS - HIDDEN VARIABLE

$$\sum_{n=1}^{N} E_{p(\boldsymbol{Z}|\boldsymbol{x}_n,\boldsymbol{\theta})}[l(\boldsymbol{\theta}';\boldsymbol{Z},\boldsymbol{x}_n)]$$

$$= \sum_{k,s} \left[\underbrace{\sum_{n=1}^{N}\sum_{t=1}^{T} p(\boldsymbol{X}_t = s, \boldsymbol{Z}_t = k | \boldsymbol{X} = \boldsymbol{x}_n, \boldsymbol{\theta})}_{\overline{M}_{k,s}} \right] \log B'_{ks}$$

$$+ \sum_{k,l} \left[\underbrace{\sum_{n=1}^{N}\sum_{t=1}^{T} p(\boldsymbol{Z}_{t-1} = k, \boldsymbol{Z}_t = l | \boldsymbol{X} = \boldsymbol{x}_n, \boldsymbol{\theta})}_{\overline{N}_{k,l}} \right] \log A'_{kl}$$

jl

- Maximized by

$$B'_{ks} = M'_{k,s} / \sum_s M'_{k,s} = M'_{k,s}/N'_k \quad \text{and} \quad A'_{kl} = N'_{k,l} / \sum_l N'_{k,l} = N'_{k,l}/N'_k$$

THE END