Hidden Markov Models

based on chapters from the book
Durbin, Eddy, Krogh and Mitchison
Biological Sequence Analysis
via Shamir’s lecture notes

music recognition

Delta Pitch: 2 2 0 -2 -2 2 2 -4 -5 5 2 2 6
I2I: 3 1 1 2 1 1 1 1 2 3 1 2
I2I ratio: 5 .5 1 2 1 1 .5 1 .66 5 .5 1
State: e f y η χ α γ φ ϵ θ f y

deal with variations in
- pitch
- timing
- timbre
- …
Stock Market Prediction

- Actual Value versus Forecasted Value for Tata Steel in Rupees over the period 5-9 2009 – 23-9 2011.
- Variations of value over time.

application: gene finding

deal with variations in
- actual sound → actual base (match/substitutions)
- timing → insertions/deletions
**Basic Questions**

Given:
- A sequence of “observations”
- A probabilistic model of our “domain”

Questions:
- Does the given sequence belong to a certain family?
  - Markov chains
  - Hidden Markov Models (HMMs)
- Can we say something about the internal structure of the sequence? (indirect observations)
  - Hidden Markov Models (HMMs)

---

**Introduction Markov Chain Model**

**Characteristics**
- Discrete time
- Discrete space
- No state History
  - Present state only
- States and transitions

**Notations:**
- $P(X)$: probability for event $X$
- $P(X,Y)$: event $X$ and event $Y$
- $P(X|Y)$: event $X$ given event $Y$
A Markov chain model is defined by

- a set of states
  - some states emit a symbol (unique per state)
  - other states (e.g., the begin state) are silent
- a set of transitions with associated probabilities
  - the transitions emanating from a given state define a distribution over the possible next states (i.e., all positive, and sum equals 1)

Markov Model

Markov Model $M = (Q, P, T)$, with

- $Q$ the set of states
- $P$ the set of initial probabilities $p_x$ for each state $x$ in $Q$
- $T = (t_{xy})$ the transition probabilities matrix/graph, with $t_{xy}$ the probability of the transition from state $x$ to state $y$.

This is a first order Markov Model: no history is modeled

An observation $X$ is a sequence of states: $X = x_1 x_2 \ldots x_n$

The probability of an observation $X$ given the model $M$ is equal to:

$$P(X|M) = p_{x_1} t_{x_1 x_2} t_{x_2 x_3} \ldots t_{x_{n-1} x_n} = p_{x_1} \prod_{i=2}^{n} t_{x_{i-1} x_i}$$
A Markov Chain Model Example

- Transition probabilities
  - \( \Pr(x_i = a | x_{i-1} = g) = 0.16 \)
  - \( \Pr(x_i = c | x_{i-1} = g) = 0.34 \)
  - \( \Pr(x_i = g | x_{i-1} = g) = 0.38 \)
  - \( \Pr(x_i = t | x_{i-1} = g) = 0.12 \)

\[ \sum_{x_i} \Pr(x_i | x_{i-1} = g) = 1 \]

over all neighbors \( x_i \)

The Probability of a Sequence for a Markov Chain Model

\[ \Pr(CGCT) = \Pr(C) \Pr(G | C) \Pr(G | G) \Pr(T | G) \]
Markov Chains: Another Example

$M_1$: $Q = \{A, B, C\}$
$P = (1, 0, 0)$
unique starting state A
$T = \begin{pmatrix}
0.7 & 0.3 & 0 \\
0.2 & 0.8 & 0 \\
0.4 & 0.6 & 0.1 \\
\end{pmatrix}$

$M_2$: $T = \begin{pmatrix}
0.6 & 0.4 & 0.3 \\
0.4 & 0.6 & 0.5 \\
0.1 & 0.5 & 0.5 \\
\end{pmatrix}$

$P(\text{AABBCCC}\mid M_1) = 1 \cdot 7 \cdot 3 \cdot 2 \cdot 8 \cdot 6 \cdot 6 \cdot 10^{-6} = 1.2 \cdot 10^{-2}$

$P(\text{AABBCCC}\mid M_2) = 1 \cdot 6 \cdot 4 \cdot 3 \cdot 6 \cdot 5 \cdot 5 \cdot 10^{-6} = 1.1 \cdot 10^{-2}$

Markov Models: Properties

Given some sequence $x$ of length $L$, we can ask:
How probable is the sequence $x$ given our model $M$?

- For any probabilistic model of sequences, we can write this probability as
  \[ Pr(x) = Pr(x_L, x_{L-1}, \ldots, x_1) \]
  \[ = Pr(x_L \mid x_{L-1}, \ldots, x_1) Pr(x_{L-1} \mid x_{L-2}, \ldots, x_1) \ldots Pr(x_1) \]

- key property of a (1st order) Markov chain: the probability of each $x_i$ depends only on the value of $X_{i-1}$
  \[ Pr(x) = Pr(x_L \mid x_{L-1}) Pr(x_{L-1} \mid x_{L-2}) \ldots Pr(x_2 \mid x_1) Pr(x_1) \]
  \[ = Pr(x_1) \prod_{i=2}^{L} Pr(x_i \mid x_{i-1}) \]
Markov Model: Underflow Problem

- Initial state $x_0$ fixed
- \( \sim \) initial probabilities
- Final state \([not depicted]\)

\[ X = x_1x_2 \ldots x_n \]

\[ P(X|M) = \prod_{i=1}^{n} t_{x_{i-1}x_i} \]

Small values: underflow

\[ t_{0x} = p_x \]

\[ \log P(X|M) = \sum_{i=1}^{n} \log t_{x_{i-1}x_i} \]

Markov Model: Comparing Models

Given:

\[ X = x_1x_2 \ldots x_n \]

\[ P(X|M) = \prod_{i=1}^{n} t_{x_{i-1}x_i} \]

Question: X best explained by which model?

We can calculate:

\[ P(X | M_1) \text{ vs. } P(X | M_2) \]

We want to know:

\[ P(M_1 | X) \text{ vs. } P(M_2 | X) \]

Bayes Rule: \[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]

\[ \frac{P(M_1|X)}{P(M_2|X)} = \frac{P(X|M_1) \cdot P(M_1)}{P(X|M_2) \cdot P(M_2)} \]
There are many cases in which we would like to represent the statistical regularities of some class of sequences:

- genes
- various regulatory sites in DNA (e.g., where RNA polymerase and transcription factors bind)
- proteins in a given family

Markov models are well suited to this type of task.
Markov Chain: An Example Application

- CpG islands
  - CG di-nucleotides are \textit{rarer} in eukaryotic genomes than expected given the marginal probabilities of C and G
  - but the regions upstream of genes (\textit{reading is from 5' to 3'}) are \textit{richer} in CG di-nucleotides than elsewhere – so called CpG islands
  - useful evidence for finding genes
- Application: Predict CpG islands with Markov chains
  - a Markov chain to represent CpG islands
  - a Markov chain to represent the rest of the genome

Markov Chains for Discrimination

- Suppose we want to distinguish CpG islands from other sequence regions
- Given sequences from CpG islands, and sequences from other regions, we can construct
  - a model to represent CpG islands
  - a null model to represent the other regions
- We can then score a test sequence $X$ by:

$$score(X) = \log \frac{Pr(X \mid CpGModel)}{Pr(X \mid nullModel)}$$
Markov Chains for Discrimination

As before we can use the scoring function:

\[ score(X) = \log \frac{\Pr(X \mid CpGModel)}{\Pr(X \mid nullModel)} \]

• Because according to Bayes’ rule we have:

\[
\Pr(CpG \mid X) = \frac{\Pr(X \mid CpG) \Pr(CpG)}{\Pr(X)}
\]

\[
\Pr(null \mid X) = \frac{\Pr(X \mid null) \Pr(null)}{\Pr(X)}
\]

• If we are not taking into account prior probabilities (Pr(CpG) and Pr(null)) of the two classes, then from Bayes’ rule it is clear that we just need to compare Pr(X|CpG) and Pr(X|null) as is done in our scoring function score().

Markov Chain Application: CpG islands

<table>
<thead>
<tr>
<th></th>
<th>island observed frequencies</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>non island</td>
<td>A</td>
<td>C</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>0.180 0.274 0.426 0.120</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.171 0.368 0.274 0.188</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.161 0.339 0.375 0.125</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.079 0.355 0.384 0.182</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In general consecutive CG pairs CG → CG are rare, although ‘islands’ Occur in signal (e.g.) promotor regions.
basic questions

- observation: DNA sequence
- model 1: CpG islands
- model 2: non-islands

• does this sequence belong to a certain family?
  Markov chains
  is this a CpG island (or not)?

• can we say something about the internal structure?
  Markov Chains: windowing
  where are the CpG islands?

application: CpG islands

<table>
<thead>
<tr>
<th>+</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>-</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
<td>A</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
<tr>
<td>C</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
<td>C</td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
<tr>
<td>G</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
<td>G</td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
<tr>
<td>T</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
<td>T</td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>

score

\[
P(X | \text{ island}) = \frac{\prod_{i=1}^{n} t_{x_{i-1}x_i}^+}{\prod_{i=1}^{n} t_{x_{i-1}x_i}^-}
\]

\[X = \text{ACGT} \quad A \rightarrow C \quad C \rightarrow G \quad G \rightarrow T \]

\[
\frac{0.274 \cdot 0.274 \cdot 0.125}{0.205 \cdot 0.078 \cdot 0.208} = 2.82
\]

Note: A score > 1 is an Indication of a CpG island.
application: CpG islands

LLR = Log-Likelihood Ratio

\[ \log \left( \frac{t^+_x}{t^-_x} \right) \]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.74</td>
<td>0.42</td>
<td>0.58</td>
<td>-0.80</td>
</tr>
<tr>
<td>C</td>
<td>-0.91</td>
<td>0.30</td>
<td>1.81</td>
<td>-0.69</td>
</tr>
<tr>
<td>G</td>
<td>-0.62</td>
<td>0.46</td>
<td>0.33</td>
<td>-0.73</td>
</tr>
<tr>
<td>T</td>
<td>-1.17</td>
<td>0.57</td>
<td>0.39</td>
<td>-0.68</td>
</tr>
</tbody>
</table>

\[ \text{'bits' (log}_2) \]

\[ \log_2 \left( \frac{0.274}{0.078} \right) = 1.81 \]

log-score \ (\log_2)

\[ \log \frac{P(X| \text{island})}{P(X| \text{non})} = \log \prod_{i=1}^n t^+_{x_{i-1}x_i} - \log \prod_{i=1}^n t^-_{x_{i-1}x_i} = \sum_{i=1}^n \log \left( \frac{t^+_{x_{i-1}x_i}}{t^-_{x_{i-1}x_i}} \right) \]

X = ACGT

\[ \log_2 \left( 0.274 \cdot 0.274 \cdot 0.125 / 0.205 \cdot 0.078 \cdot 0.208 \right) = 0.42 + 1.81 - 0.73 = 1.50 \]

CpG Log-Likelihood Ratio

\[ \log \left( \frac{t^+_x}{t^-_x} \right) \]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.74</td>
<td>0.42</td>
<td>0.58</td>
<td>-0.80</td>
</tr>
<tr>
<td>C</td>
<td>-0.91</td>
<td>0.30</td>
<td>1.81</td>
<td>-0.69</td>
</tr>
<tr>
<td>G</td>
<td>-0.62</td>
<td>0.46</td>
<td>0.33</td>
<td>-0.73</td>
</tr>
<tr>
<td>T</td>
<td>-1.17</td>
<td>0.57</td>
<td>0.39</td>
<td>-0.68</td>
</tr>
</tbody>
</table>

LLR(ACGT) = 0.42 + 1.81 - 0.73 = 1.50 \hspace{1cm} (0.37 'bits' per base)

\[ \frac{1.5}{4} = 0.375 \]

- is a (short) sequence a CpG island?
  - compare with observed data \ (normalized for length)
- where (in long sequence) are CpG islands?
  - first approach: sliding window
- ! What would be the length of window?
empirical data

- is a (short) sequence a CpG island?
  compare with observed data (normalized for length)

**Figure 3.2.** The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.

- where (in long sequence) are CpG islands?
  first approach: *sliding window*

![CpGplot](image-url)
Some Notes on: Higher Order Markov Chains

- The Markov property specifies that the probability of a state depends only on the probability of the previous state.
- But we can build more “memory” into our states by using a higher order Markov model.
- In an $n$-th order Markov model

$$\Pr(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_1) = \Pr(x_i \mid x_{i-1}, \ldots, x_{i-n})$$

The probability of the current state depends on the previous $n$ states.
Selecting the Order of a Markov Chain Model

- But the number of parameters we need to estimate for an $n$-th order Markov model grows exponentially with the order
  - for modeling DNA we need $O(4^{n+1})$ parameters (# of state transitions) for an $n$-th order model
- The higher the order, the less reliable we can expect our parameter estimates to be
  - estimating the parameters of a 2$^{nd}$ order Markov chain from the complete genome of E. Coli (5.44 x 10$^6$ bases) , we would see each (length 3) word ~ 85.000 times on average (divide by 4$^3$)
  - estimating the parameters of a 9$^{th}$ order chain, we would see each (length 10) word ~ 5 times on average (divide by 4$^{10}$ ~ 10$^6$)

Higher Order Markov Chains

- An $n$-th order Markov chain over some alphabet $A$ is equivalent to a first order Markov chain over the alphabet of $n$-tuples: $A^n$
- Example: a 2$^{nd}$ order Markov model for DNA can be treated as a 1$^{st}$ order Markov model over alphabet
  
  | AA, AC, AG, AT | Transition probabilities: |
  | CA, CC, CG, CT | P(A|AA), P(A|AC), etc. |
  | GA, GC, GG, GT |
  | TA, TC, TG, TT |
A Fifth Order Markov Chain Equivalent

Pr(GCTAC | A) = Pr(GCTAC)Pr(A | GCTAC)

hidden Markov model

Where (in long sequence) are CpG islands?

- first approach: Markov Chains + windowing
- second approach: hidden Markov model
Hidden Markov Model: A Simple HMM

Given observed sequence AGGCT, which state emits which item?

Another example: Eddy (2004)

An (toy) HMM for 5' splice site recognition.

Figure from: What is a hidden Markov model?
Example: weather

emission probabilities

$$P(H) = 0.1$$

$$P(M) = 0.2$$

$$P(L) = 0.7$$

initial probabilities

$$P_H = 0.4$$

$$P_M = 0.2$$

$$P_L = 0.4$$

transition probabilities

observed weather vs. pressure

$$P(H) = 0.3$$

$$P(M) = 0.4$$

$$P(L) = 0.3$$

Given path

Emissions

P( RCCSS | HHHHH ) = $$1 \cdot 2 \cdot 2 \cdot 7 \cdot 7 = 196 \ (x10^{-5})$$

P( RCCSS | MMMMM ) = $$3 \cdot 4 \cdot 3 \cdot 3 = 432 \ (x10^{-5})$$

P( RCCSS, HHHHH ) = $$4 \cdot 1 \cdot 6 \cdot 2 \cdot 6 \cdot 7 \cdot 6 \cdot 7 = 1016 \ (x10^{-7})$$

P( RCCSS, MMMMM ) = $$2 \cdot 3 \cdot 2 \cdot 4 \cdot 2 \cdot 2 \cdot 3 \cdot 3 = 14 \ (x10^{-7})$$
hidden Markov model

**model** \( M = (\Sigma, Q, T) \)
- **states** \( Q \)
- **transition probabilities** \( t_{pq}, p, q \in Q \)

**observation** \( X = x_1 x_2 \ldots x_n \in \Sigma^* \)
- observe states indirectly ‘hidden’
- **emission probabilities**
  \[ e_{px}, p \in Q, x \in \Sigma \quad e_p(x) \]

**probability**
observation given the model
? there may be many state seq’s

HMM main questions

observation \( X \in \Sigma^* \)

Given HMM \( M \):
- probability of observation \( X \)?
- most probable state sequence?
- how to find the parameters of the model \( M \)? *training*
Given sequence $X$: most probable state vs. optimal path

- most probable state (over all state sequences)
  - posterior decoding
  - using forward & backward probabilities
- most probable path (= single state sequence)
  - Viterbi

**probability of state**

**probability of observation $X$**

*dynamic programming*: probability ending in state $q$ emitting symbol $x_i$

$$f_q(i) = \sum_{p \in Q} f_p(i-1) t_{pq} e_q(x_i)$$
probability of observation $X$

probability observing $x_1, \ldots, x_i$, and ending in state $q$:

$$ f_q(i) = P(x_1 \ldots x_i, \pi_i = q) $$

$$ f_q(i) = \sum_{p \in Q} f_p(i-1) t_{pq} e_q(x_i) $$

‘forward’ probability

$$ P(X) = \sum_{p \in Q} f_p(n) t_{p*} \quad * = \text{end-state} $$

---

Probability of observation: weather

Initial state:
- $p_H = 0.4$
- $p_M = 0.2$
- $p_L = 0.4$

Transitions:
- Remain in H
- Coming from M
- Coming from L

Start: 0 1

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Transition</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>H 0</td>
<td>1:R</td>
<td>4·1</td>
<td>144 (x10^-4)</td>
</tr>
<tr>
<td>M 0</td>
<td>2:C</td>
<td>6·4</td>
<td>576 (x10^-4)</td>
</tr>
<tr>
<td>L 0</td>
<td>4·6</td>
<td>24·4</td>
<td>372 (x10^-4)</td>
</tr>
</tbody>
</table>
posterior decoding

\[ P(\pi_i = q \mid X) \] i-th state equals q

\[ f_q(i) = P(x_1 \ldots x_i, \pi_i = q) \] forward

\[ b_q(i) = P(x_{i+1} \ldots x_n \mid \pi_i = q) \] backward

\[ P(X, \pi_i = q) = f_q(i)b_q(i) \] again:
We cannot try all possibilities Viterbi

\[ P(\pi_i = q \mid X) = \frac{f_q(i)b_q(i)}{P(X)} \]

HMM main questions

observation \( X \epsilon \Sigma^* \) \Rightarrow most probable state sequence

\[ X: \]

• probability of this observation?
• most probable state sequence?
• how to find the model? training
Viterbi algorithm

most probable state sequence for observation $X$

(1) dynamic programming: $v_q(i)$ probability ending in state $q$ and emitting $x_i$

$$v_q(i) = \max_{p \in Q} v_p(i-1) t_{pq} e_q(x_i)$$

Viterbi algorithm

(2) traceback: most probable state sequence

start with final maximum

$$x_L$$
HMM Example: CpG islands

8 states A⁺ vs A⁻
unique observation each state

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>C</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td>G</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>T</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
</tbody>
</table>

Transition Matrix

HMM for Hidden Coin Tossing

Fig. 2. Three possible Markov models which can account for the results of hidden coin tossing experiments. (a) 1-coin model. (b) 2-coins model. (c) 3-coins model.
### Dishonest Casino Dealer

#### Rolls

<table>
<thead>
<tr>
<th>Rolls</th>
<th>315116246446442453213163164152336251445363165662656666</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>FFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
<tr>
<td>Viterbi</td>
<td>FFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>6511664531326512456366646366631623264552352666666625151631</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
<tr>
<td>Viterbi</td>
<td>LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>22255541666565635643243413151346514353411126414626253356</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>FFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
<tr>
<td>Viterbi</td>
<td>FFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>366163666666621354136616611632525624622552652526526643535336</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
<tr>
<td>Viterbi</td>
<td>LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>23312126534414312331632436365624666626526526666661235254242</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>FFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
<tr>
<td>Viterbi</td>
<td>FFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL</td>
</tr>
</tbody>
</table>
Observation
366163666466232534413661661163252562462255265252266435353336
Viterbi
LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Compare to:
Forward
FFLLLLLLLLLLLLFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL

Posterior (total)
LLLLLLLLLLLL

Sketch: Parameter estimation

training sequences $X^{(i)}$
optimize score $\prod_{i=1}^{n} P(X^{(i)} | \Theta)$ for model $\Theta$.

Markov Chain: state sequences known

- count transitions $pq$ $A_{pq}$
- count emissions $b$ in $p$ $E_p(b)$

divide by
- total transitions in $p$
- emissions in $q$

Laplace correction for dealing with ‘zero’ probabilities.
Adding 1 to each count.
### Baum-Welch

**HMM: state sequences unknown**

#### Baum-Welch training

- Based on given HMM $\Theta$
- Given a training set of sequences $X$
- Determine expected number of transitions and expected number of emissions
- Apply ML and build a new (better) model:
  - ML tries to find a model that gives the training data the highest likelihood
  - Iterate until convergence.

**Note:**
- can get stuck in local maxima
- does not understand the semantics of the states

### Baum-Welch Re-estimation

For the re-estimation we need the expected counts For the transitions and the emissions in the HMM:

- Apply the backward-forward algorithm.

**Probability of state $q$ when emitting $X_i$:**

$$P(\pi_i = q \mid X) = \frac{f_q(i)b_q(i)}{P(x)}$$

**Probability of transition $(p,q)$ after emitting $X_i$:**

$$P(\pi_i = p, \pi_{i+1} = q \mid X, \Theta) = \frac{f_p(i)\cdot t_{pq} \cdot e_q(x_{i+1}) \cdot b_q(i+1)}{P(X)}$$
Baum-Welch

\[ A_{pq} \text{ sum over all training sequences } X \]
\[ \text{sum over all positions } i \]
\[ E_p(b) \text{ sum over all training sequences } X \]
\[ \text{sum over all positions } i \text{ with } x_i = b \]

Estimate parameters by ratio of expected counts.

Baum-Welch training

concerns:
• guaranteed to converge
  target score, not \( \Theta \)
• unstable solutions !
• local maximum

practical
• small values \( \rightarrow \) renormalize

tips:
• repeat for several initial \( \Theta \)
• start with meaningful \( \Theta \)
Viterbi training (sketch):

- determine optimal paths
- re-compute as if paths are known
- score may decrease!

Important Papers on HMM


Furthermore: