A fully observed (or 'supervised') case of estimation

first suppose a fully observed corpus of obs and state sequences (states in blue, obs in red):

\begin{align*}
\text{one:} & \text{a} & \text{three:} & \text{x} & \text{four:} & \text{z} & \text{five:} & \# \\
\left(= o^0, s^0 \right) \\quad \& \\
\text{one:} & \text{a} & \text{three:} & \text{y} & \text{three:} & \text{x} & \text{five:} & \# \\
\left(= o^1, s^1 \right) \\quad & \\
\text{two:} & \text{b} & \text{four:} & \text{x} & \text{one:} & \text{a} & \text{three:} & \text{y} & \text{five:} & \# \\
\left(= o^2, s^2 \right) \\quad & \\
\end{align*}

The values for $\pi$, $A$ and $B$ derived by relative frequencies in this corpus are:

\begin{align*}
A \\
C(\text{one, three}) &= 3 & C(\text{one}) &= 3 & P(\text{three|one}) &= 1 \\
C(\text{three, four}) &= 1 & C(\text{three}) &= 4 & P(\text{four|three}) &= 0.25 \\
C(\text{three, three}) &= 1 & & P(\text{three|three}) &= 0.25 \\
C(\text{three, five}) &= 2 & & P(\text{five|three}) &= 0.5 \\
C(\text{two, four}) &= 1 & C(\text{two}) &= 1 & P(\text{four|two}) &= 1 \\
C(\text{four, five}) &= 1 & C(\text{four}) &= 2 & P(\text{five|four}) &= 0.5 \\
C(\text{four, one}) &= 1 & & P(\text{one|four}) &= 0.5 \\
B \\
C(\text{one, a}) &= 3 & C(\text{one}) &= 3 & P(\text{a|one}) &= 1 \\
C(\text{three, x}) &= 2 & C(\text{three}) &= 4 & P(\text{x|three}) &= 0.5 \\
C(\text{three, y}) &= 2 & & P(\text{y|three}) &= 0.5 \\
C(\text{two, b}) &= 1 & C(\text{two}) &= 1 & P(\text{b|two}) &= 1 \\
C(\text{five, #}) &= 3 & C(\text{five}) &= 3 & P(\text{#|five}) &= 1 \\
C(\text{four, z}) &= 1 & C(\text{four}) &= 2 & P(\text{z|four}) &= 0.5 \\
C(\text{four, x}) &= 1 & & P(\text{x|four}) &= 0.5 \\
\pi \\
C(s_1 = \text{one}) &= 2 & P(s_1 = \text{one}) &= 0.666 \\
C(s_1 = \text{two}) &= 1 & P(s_1 = \text{two}) &= 0.333
\end{align*}
(brute-force) EM case (unsupervised)

Note that this estimation from a small set of examples has left quite a few options set to have zero probability. This will be handy in the following illustration of the unsupervised case as it keeps the number of alternatives down. In general such zeroes would be a bad idea. Let’s use these probs as initial values in some unsupervised estimation, just using a corpus of observed words.

Let’s suppose this corpus to have much more of the final obs sequence than before:

1× axz#
1× ayx#
10× bxay#

Some worked eg of Parameter Estimation with HMMs: a fully observed case and the 'brute-force' EM unsupervised case

brute-force EM: conditional probs of paths

To apply (brute-force) EM, for each o^i, need for all possible s, p(s|o^i) (ie. the 'responsibility' γ^i(s)). Given the forgoing we get:

\[ o^0: a \times z # \quad s^0: one \ three \ four \ five \ P(s^0|o^0) = 1 \quad (= \gamma^0(s^0)) \]
\[ o^1: a \ y \ x # \quad s^1: one \ three \ four \ five \ P(s^1|o^1) = 0.5 \quad (= \gamma^1(s^1)) \]
\[ o^2: b \times a \ y # \quad s^2: one \ three \ five \ P(s^2|o^2) = 0.5 \quad (= \gamma^2(s^2)) \]
\[ o^{11}: b \times a \ y # \quad s^{11}: two \ four \ one \ three \ five \ P(s^{11}|o^{11}) = 1 \quad (= \gamma^{11}(s^{11})) \]

brute-force EM: considering all the possible paths

On our initial settings what state sequences are possible for these observation sequences?

<table>
<thead>
<tr>
<th>observations</th>
<th>possible completions</th>
</tr>
</thead>
<tbody>
<tr>
<td>o^0: a x z #</td>
<td>one:a three:x four:x five:# (= o^0, s^0)</td>
</tr>
<tr>
<td>o^1: a y x #</td>
<td>one:a three:y four:x five:# (= o^1, s^1)</td>
</tr>
<tr>
<td>o^2: b a y #</td>
<td>two:b four:x one:a three:y five:# (= o^2, s^3)</td>
</tr>
</tbody>
</table>

\[ o^0 \text{ and the 10 copies of } o^2 \text{ get actually get only } 1 \text{ possible completion each}^{1} \]
\[ o^1 \text{ though gets two possible completions, equally likely ones} \]

this is because when you look to the terms that differ between \( P(o^1, s^1) \) and \( P(o^1, s^3) \), numerically there’s no difference

\[
P(\text{four|three}) = 0.25 \quad \text{P(x|four}) = 0.5 \quad \text{P(five|four}) = 0.5 \]
\[
P(\text{three|three}) = 0.25 \quad \text{P(x|three}) = 0.5 \quad \text{P(five|five}) = 0.5 \]

\[^{1}\text{the 'hidden' states are not really very hidden!}\]

A and B: expected counts and ratios

will compute expected counts of events relevant for A and B based on the virtual corpus of completions:

<table>
<thead>
<tr>
<th>observations</th>
<th>'count' of completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>o^0: a x z #</td>
<td>1× one:a three:x four:x five:#</td>
</tr>
<tr>
<td>o^1: a y x #</td>
<td>0.5× one:a three:y four:x five:#</td>
</tr>
<tr>
<td>o^2: b a y #</td>
<td>10× two:b four:x one:a three:y five:#</td>
</tr>
</tbody>
</table>

Note although x can be generated from four or three, in o^2 only four is possible, so four:x is more frequent in this virtual corpus than it was in the fully observed corpus we started with.

We derive expected counts for 'state-then-state' and 'obs-with-state', then derive new A and B, eg.

\[ A^1 \quad E(\text{three, four}) = 1.5 \quad E(\text{three}) = 12.5 \quad P(\text{four|three}) = 1.5/12.5 \]
\[ E(\text{three, three}) = 0.5 \quad P(\text{three|three}) = 0.5/12.5 \]
\[ E(\text{four, five}) = 10.5 \quad E(\text{four}) = 11.5 \quad P(\text{five|four}) = 1.5/11.5 \]
\[ B^1 \quad E(\text{three, x}) = 1.5 \quad E(\text{three}) = 12.5 \quad P(\text{x|three}) = 1.5/12.5 \]
\[ E(\text{three, y}) = 11 \quad E(\text{four, x}) = 10.5 \quad P(\text{x|four}) = 10.5/11.5 \]

Some worked eg of Parameter Estimation with HMMs: a fully observed case and the 'brute-force' EM unsupervised case
the virtual corpus of completions:

<table>
<thead>
<tr>
<th>observations</th>
<th>'count' of completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>o0 a x z #</td>
<td>1 x one:a three:x four:z five: #</td>
</tr>
<tr>
<td>o1 a y x #</td>
<td>0.5 x one:a three:y four:x five: #</td>
</tr>
<tr>
<td>o2 b x a y #</td>
<td>10 x two:b four:x one:a three:y five: #</td>
</tr>
</tbody>
</table>

derive expected counts of 'state at start', then new $\pi$:

$$
\pi^1 \quad E(s_1 = \text{one}) = 2 \quad P(s_1 = \text{one}) = 2/12 \\
E(s_1 = \text{two}) = 10 \quad P(s_1 = \text{two}) = 10/12 
$$

after one iteration

so after one iteration of brute force EM the new values for $\pi$ and $A$ and $B$ are somewhat different from the starting values, here's some:

$$
\begin{align*}
\pi^1 & \quad P(\text{start} = \text{one}) = 2/12 & P(\text{four|three}) = 1.5/12.5 & P(\text{x|three}) = 1.5/12.5 \\
& \quad P(\text{start} = \text{two}) = 10/12 & P(\text{three|three}) = 0.5/12.5 & P(\text{x|four}) = 10.5/11.5 \\
& & P(\text{five|three}) = 10.5/12.5 & P(\text{five|four}) = 1.5/11.5 \\
\end{align*}
$$

For example, concerning $a \ y \ x \ #$, whereas previously there two state sequences that could pair equally well with this, now there is an asymmetry.

$$
\begin{align*}
P(\text{one: a three: y four: x five: #}) & = 1.5/12.5 \times 10.5/11.5 \times 1.5/11.5 = 0.0143 \\
P(\text{one: a three: y three: x five: #}) & = 0.5/12.5 \times 1.5/12.5 \times 10.5/12.5 = 0.0040 
\end{align*}
$$

So one round of EM has definitely changed the probabilities; really all this should be every expectable given previous study of EM applied in (i) the coin tossing scenario – hidden A vs B choice (ii) the IBM Model 1 scenario – hidden alignment. We 'just' have a different hidden variable here – the hidden state sequence. As the main slides expound upon, the bigger point is that this brute-force approach is not computationally feasible.