Hidden Markov Model inference with the Viterbi algorithm: a mini-example

In this mini-example, we'll cover the problem of inferring the most-likely state sequence given an HMM and an observation sequence. The problem of parameter estimation is not covered.

Once again, the dynamic program for the HMM trellis on an observation sequence of length n is as follows:

1. Initialize $\delta_0(s) = 1$ for s the start state, and $\delta_0(s) = 0$ for all other states (this is equivalent to having only the start state in the trellis at position zero)
2. For each value $i = 1, \ldots, n$, calculate:
   (a) $\delta_i(s) = \max_{s_{i-1}} P(s_i|s_{i-1})P(w_{i-1}|s_{i-1})\delta_{i-1}(s_{i-1})$
   (b) $\psi_i(s) = \arg\max_{s_{i-1}} P(s_i|s_{i-1})P(w_{i-1}|s_{i-1})\delta_{i-1}(s_{i-1})$
3. Finally, fill out the end state of the trellis (position $n+1$) using the rules in (2) above.

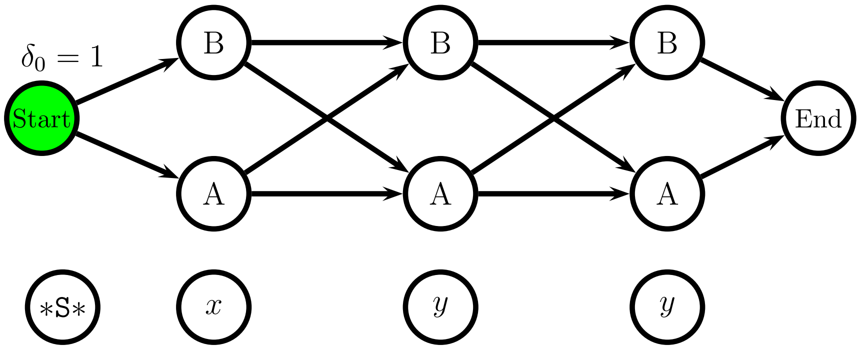
We'll take as our transition probability distribution

	Next		
Current	A	B	End
Start	0.7	0.3	0
A	0.2	0.7	0.1
B	0.7	0.2	0.1

and as our emission probability distribution

	Word		
State	*S*	x	y
Start	1	0	0
A	0	0.4	0.6
B	0	0.3	0.7

Suppose we see the input sequence $x\ y\ y$. We start by constructing the trellis and initializing it with δ_0(*START*) = 1 at the start. The green nodes indicate how much of the sequence can be considered generated after each iteration of trellis-filling:



Next, we calculate the δ values at position 1:

$$\delta_1(A) = \max_{s_0} P(A|s_0)P(*\texttt{S}*|s_0)\delta_0(s_0) \quad (1)$$

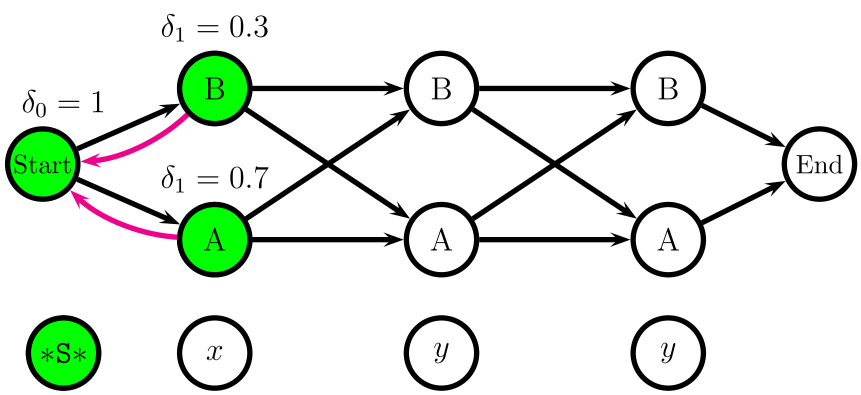
which is simple since there is only one possible value s_0, the start state:

$$\begin{aligned} \delta_1(A) &= 1 \times 1 \times 0.7 && (2)\\ &= 0.7 && (3) \end{aligned}$$

Likewise, we obtain

$$\delta_1(B) = 1 \times 1 \times 0.3 \quad (4)$$

The backtraces are both trivial as well: $\psi_1(A) = \psi_1(B) = *\texttt{S}*_0$



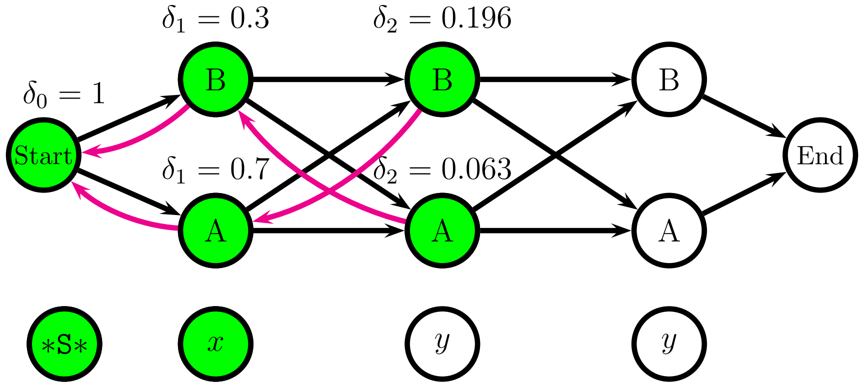
We next calculate the δ values at position 2:

$$\begin{aligned} \delta_2(A) &= \max_{s_1} P(A|s_1)P(*\texttt{S}*|s_1)\delta_1(s_1) && (5)\\ &= \max\{0.2 \times 0.4 \times 0.7, 0.7 \times 0.3 \times 0.3\} && (6)\\ &= \max\{0.056, 0.063\} && (7)\\ &= 0.063 && (8) \end{aligned}$$

This value was higher for $s_1 = B$, hence $\psi_2(A) = B_1$. We also have

$$\delta_2(B) = \max\{\overbrace{0.7 \times 0.4 \times 0.7}^{A}, \overbrace{0.2 \times 0.3 \times 0.3}^{B}\} \tag{9}$$

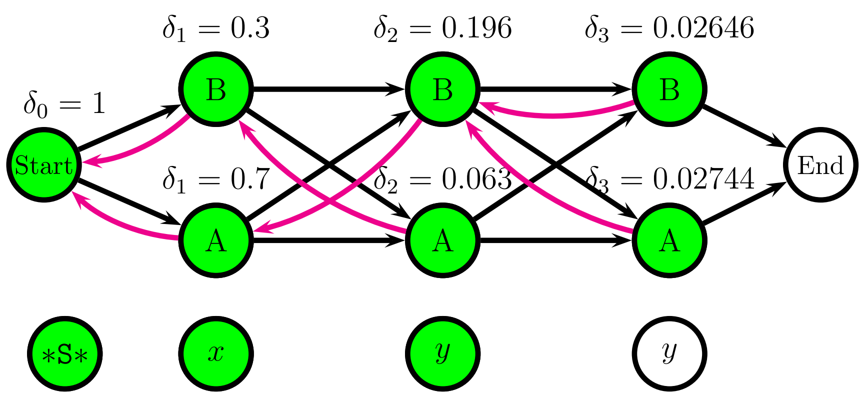
giving us $\psi_2(B) = A_1$:



We recurse one more time for position 3:

$$\delta_3(A) = \max\{\overbrace{0.2 \times 0.6 \times 0.063}^{A}, \overbrace{0.7 * 0.7 * 0.196}^{B}\} \tag{10}$$

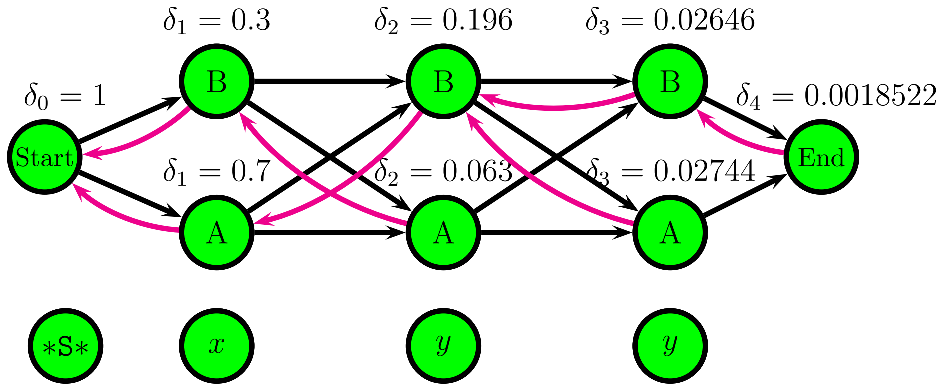
$$\delta_3(B) = \max\{\overbrace{0.7 \times 0.6 \times 0.063}^{A}, \overbrace{0.2 * 0.7 * 0.196}^{B}\} \tag{11}$$



and finally the last time for the end state:

$$\delta_4(End) = \max\{\overbrace{0.1 \times 0.6 \times 0.02744}^{A}, \overbrace{0.1 * 0.7 * 0.02646}^{B}\} \tag{12}$$

giving us



We made it! From the end state we can read off the Viterbi sequence (following the backtraces through to the start state) and its probability:

Viterbi sequence: ABB
$P(ABB, xyy) = 0.00185522$

Note that this is different than the inference than would be made either with a "reverse emission model" where $P(A|x) = 0.4, P(A|y) = 0.3$, which would favor the sequence BBB, or with the transition model alone (which would favor the sequence ABA). The emission and transition models work together to determine the posterior inference.