Intelligent Systems

Escuela Técnica Superior de Informática

Universitat Politècnica de València

Block 2 Chapter 6: Forward and Viterbi algorithms

Index

Forward algorithm to compute $P(y|M)$

We define $\alpha(q,t)$ as the probability that a Markov model M generates the sub-string $y_1 \cdots y_t$ and reaches the state q at instant t:

$$\alpha(q,t) = \sum_{\substack{q_1,\ldots,q_t \\ q_t=q}} P(y_1 \cdots y_t,\, q_1,\ldots,q_t)$$

$\alpha(q,t)$ can be computed recursively:

$$\begin{aligned}
\alpha(q,t) &= \sum_{\substack{q_1,\ldots,q_t \\ q_t=q}} P(y_1 \cdots y_t,\, q_1,\ldots,q_t) \\
&= \sum_{\substack{q_1,\ldots,q_{t-1} \\ q'\in Q \\ q_{t-1}=q'}} P(y_1 \cdots y_{t-1},\, q_1,\ldots,q_{t-1}) A_{q',q}\, B_{q,y_t} \\
&= \sum_{q'\in Q} \sum_{\substack{q_1,\ldots,q_{t-1} \\ q_{t-1}=q'}} P(y_1 \cdots y_{t-1},\, q_1,\ldots,q_{t-1}) A_{q',q}\, B_{q,y_t} \\
&= \sum_{q'\in Q} \alpha(q',t-1) A_{q',q}\, B_{q,y_t}
\end{aligned}$$

Forward algorithm (cont.)

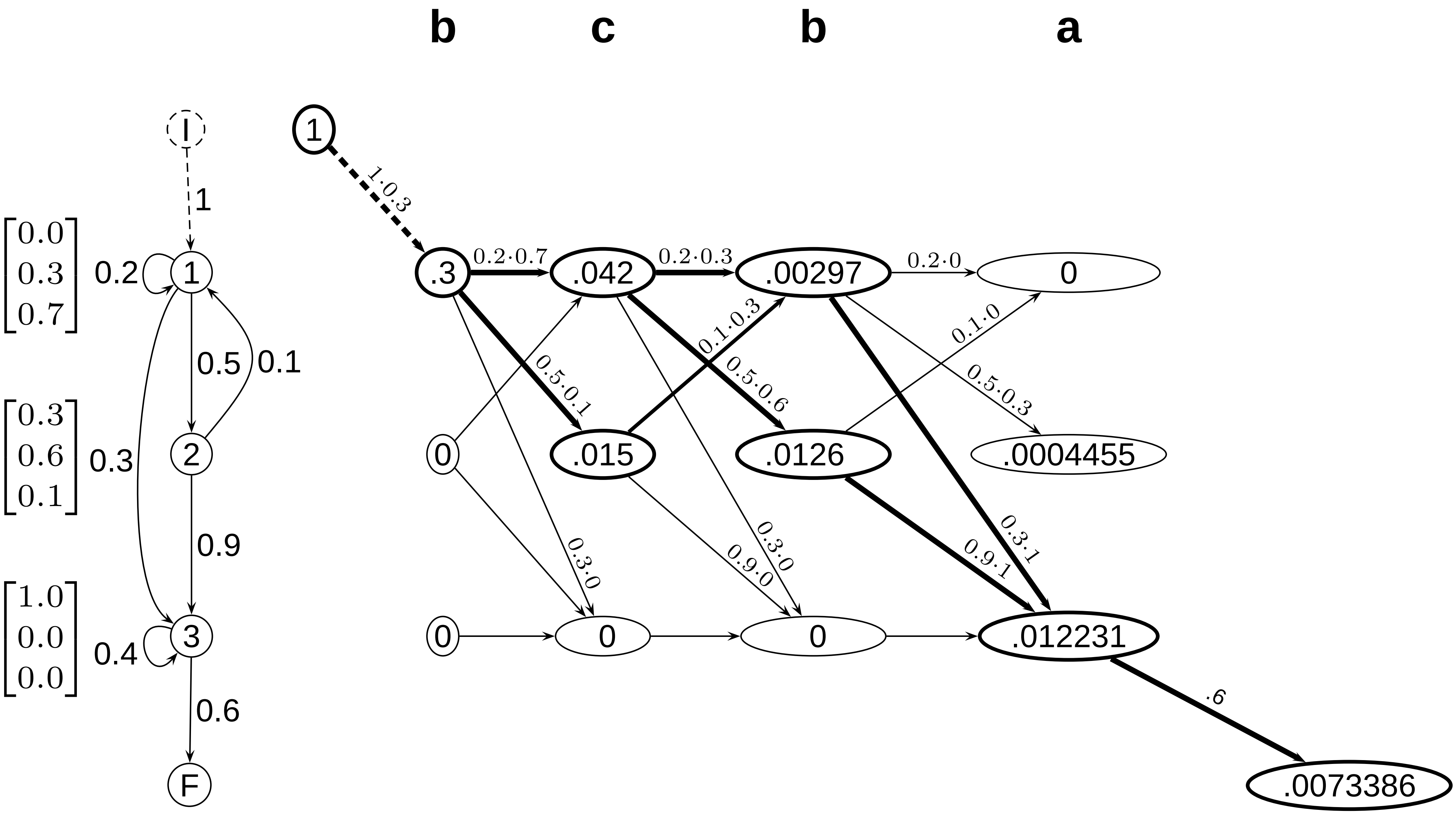
In general:
$$\alpha(q,t) = \begin{cases} \pi_q \, B_{q,y_1} & \textbf{\textit{si}} \;\; t = 1 \\ \sum_{q' \in Q} \alpha(q', t-1) \, A_{q',q} \, B_{q,y_t} & \textbf{\textit{si}} \;\; t > 1 \end{cases}$$

The probability of the string $P(y \mid M)$:

$$P(y \mid M) = \sum_{q \in Q} \alpha(q, |y|) \, A_{q,F}$$

- The function $\alpha()$ can be represented as a matrix: $\alpha_{q,t} \equiv \alpha(q,t)$.
- This matrix defines a *multilayer graph* called *trellis*, which allows for an efficient calculation of $\alpha(q, |y|)$ *by Dynamic Programming*.
- Temporal complexity of Forward algorithm: $O(mb)$, where m is the string length and b is the number of state transitions.

Forward algorithm: example (trellis)



Forward algorithm: exercise

Let M be the following Markov model:

$Q = \{1, 2, 3, F\}$

$\Sigma = \{a, b, c\}$

$\pi_1 = \pi_2 = \frac{1}{2}$, $\pi_3 = 0$

A	1	2	3	F
1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
3	0	0	$\frac{1}{2}$	$\frac{1}{2}$

B	a	b	c
1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
3	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

1. Apply the forward algorithm to the string abc.

Exercise: direct resolution

α	a $t=1$	b $t=2$	c $t=3$	
1	$\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$	$\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{3}+$ $\frac{1}{8}\cdot\frac{1}{3}\cdot\frac{1}{3}=\frac{5}{144}$	$\frac{5}{144}\cdot\frac{1}{4}\cdot\frac{1}{6}+$ $\frac{1}{24}\cdot\frac{1}{3}\cdot\frac{1}{6}+$ $\frac{5}{96}\cdot 0\cdot\frac{1}{6}=\frac{13}{3456}$	
2	$\frac{1}{2}\cdot\frac{1}{4}=\frac{1}{8}$	$\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{4}+$ $\frac{1}{8}\cdot\frac{1}{3}\cdot\frac{1}{4}=\frac{1}{24}$	$\frac{5}{144}\cdot\frac{1}{2}\cdot\frac{1}{2}+$ $\frac{1}{24}\cdot\frac{1}{3}\cdot\frac{1}{2}+$ $\frac{5}{96}\cdot 0\cdot\frac{1}{2}=\frac{1}{76}$	
3		$\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{2}+$ $\frac{1}{8}\cdot\frac{1}{3}\cdot\frac{1}{2}=\frac{5}{96}$	$\frac{5}{144}\cdot\frac{1}{4}\cdot\frac{1}{4}+$ $\frac{1}{24}\cdot\frac{1}{3}\cdot\frac{1}{4}+$ $\frac{5}{96}\cdot\frac{1}{2}\cdot\frac{1}{4}=\frac{7}{576}$	
F				$\frac{13}{3456}\cdot 0+$ $\frac{1}{76}\cdot 0+$ $\frac{7}{576}\cdot\frac{1}{2}=\frac{7}{1152}$

Index

Viterbi approximation to $P(y \mid M)$

Given a Markov model $M = (Q, \Sigma, \pi, A, B)$ with final state F, and a string $y = y_1 \cdots y_m \in \Sigma^+$, the probability that M generates y is:

$$P(y \mid M) \;=\; \sum_{z \in Q^+} P(y, z) \;=\; \sum_{q_1, \ldots, q_m \in Q^+} P(y,\, q_1, \ldots, q_m)$$

Trying to find the probability of string y by means of considering all state sequences is impractical

Solution: use ***Viterbi approximation*** to a $P(y \mid M)$ (calculate the most likely/probable sequence of states for generating y)

$$\tilde{P}(y \mid M) = \max_{q_1, \ldots, q_m \in Q^+} P(y,\, q_1, \ldots, q_m)$$

The corresponding *most probable sequence of states* is:

$$\tilde{q} = (\tilde{q}_1, \ldots, \tilde{q}_m) = \operatorname*{argmax}_{q_1, \ldots, q_m \in Q^+} P(y,\, q_1, \ldots, q_m)$$

Viterbi algorithm

We define $V(q,t)$ as the maximum probability that a Markov model reaches state q at instant t and emits the string $y = y_1 \ldots y_t$:

$$V(q,t) = V(q,|y|) = \max_{\substack{q_1,\ldots,q_t \\ q_t=q}} P(y_1 \cdots y_t,\ q_1,\ldots,q_t)$$

Recursive calculation of $V(q,t)$

$t = 1$

$$V(q,t) = V(q,y_1) = P(y_1,q) = P(y_1 \mid q)P(q) = B_{q,y_1}\pi_q$$

$t > 1$

$$V(q,t) =$$

$$\max_{\substack{q_1,\ldots,q_t \\ q_t=q}} P(y_1 \cdots y_t,\ q_1,\ldots,q_t) = \max_{\substack{q_1,\ldots,q_{t-1},q_t \\ q_{t-1}=q' \\ q_t=q}} P(y_1 \cdots y_{t-1},\ q_1,\ldots,q_{t-1}) \cdot A_{q',q}\, B_{q,y_t} =$$

$$\max_{q' \in Q} \max_{\substack{q_1,\ldots,q_{t-1} \\ q_{t-1}=q'}} P(y_1 \cdots y_{t-1},\ q_1,\ldots,q_{t-1}) \cdot A_{q',q}\, B_{q,y_t} =$$

$$\max_{q' \in Q} V(q',t-1) \cdot A_{q',q}\, B_{q,y_t}$$

Viterbi algorithm (cont.)

In general: $$V(q,t) = V(q,|y|) = \begin{cases} \pi_q\, B_{q,y_1} & \textbf{si } t = 1 \\ \max\limits_{q' \in Q} V(q', t-1)\, A_{q',q}\, B_{q,y_t} & \textbf{si } t > 1 \end{cases}$$

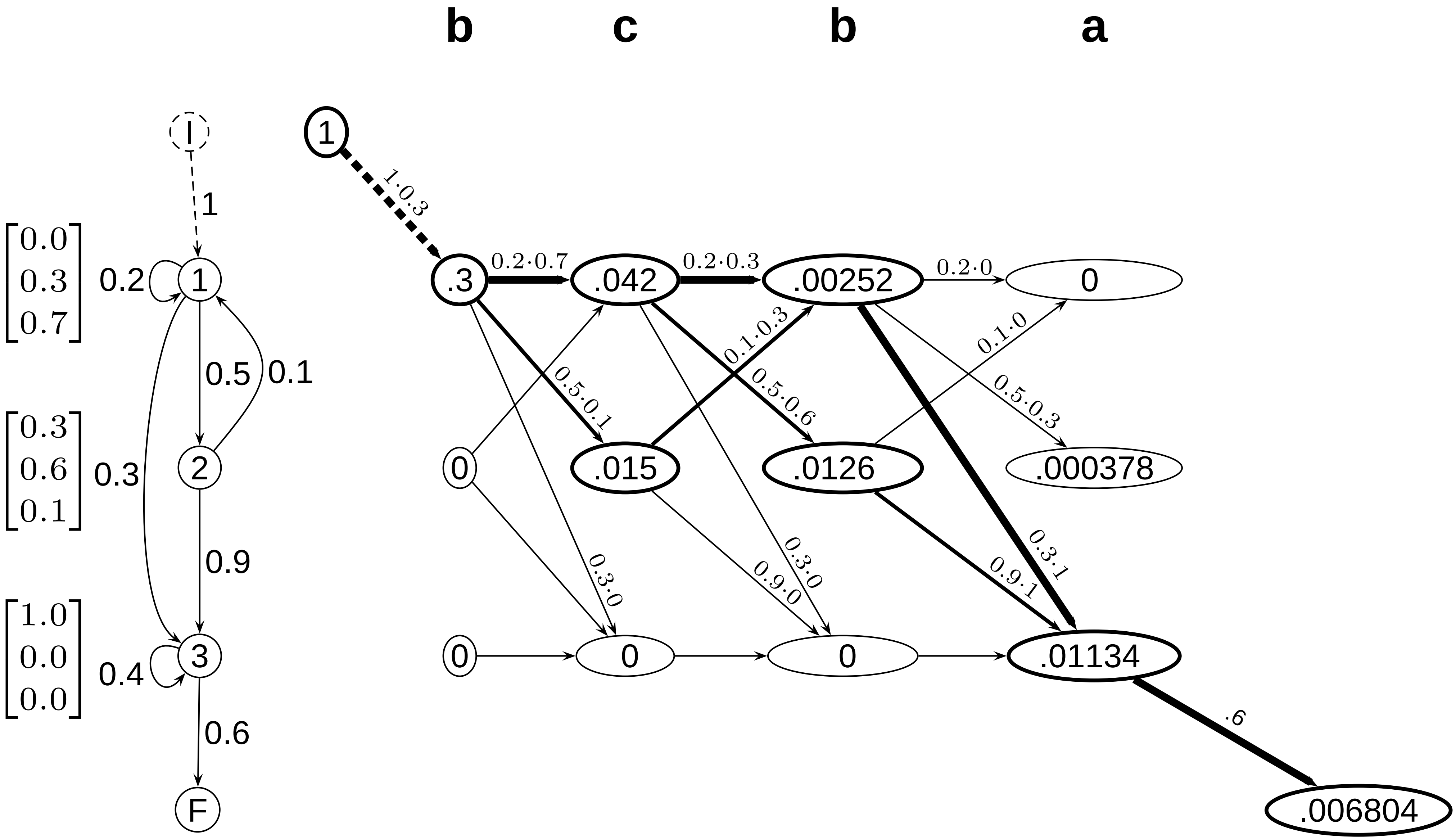
Now, we can replace the calculation of $P(y \mid M)$ by the Viterbi approximation:

$$\tilde{P}(y \mid M) = \max_{q \in Q} V(q, |y|)\, A_{q,F}$$

in other words, rather than finding all the state sequences that generate the string y, when using Viterbi approximation we only consider the most probable state sequence (optimal state sequence), which is the one that maximizes the expression $\max_{q \in Q} V(q, |y|)\, A_{q,F}$.

- Function V can be represented as a matrix: $V_{q,t} \equiv V(q,t)$.
- This function defines a multistage graph called ***trellis*** that allows for the efficient iterative calculation of $V(q, |y|)$ by *Dynamic Programming*.
- The corresponding optimal sequence of states, $\tilde{q}$, is found by tracing the *trellis* backwards.
- The temporal complexity of Viterbi is $O(mb)$ where m is the length of the string and b is the number of state transitions

Viterbi: example (trellis)



Viterbi algorithm: exercise

Let M be a model with:

$Q = \{1, 2, 3, F\}$

$\Sigma = \{a, b, c\}$

$\pi_1 = \pi_2 = \frac{1}{2}$, $\pi_3 = 0$

A	1	2	3	F
1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
3	0	0	$\frac{1}{2}$	$\frac{1}{2}$

B	a	b	c
1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
2	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
3	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

1. Find the *trellis* for the string abc.

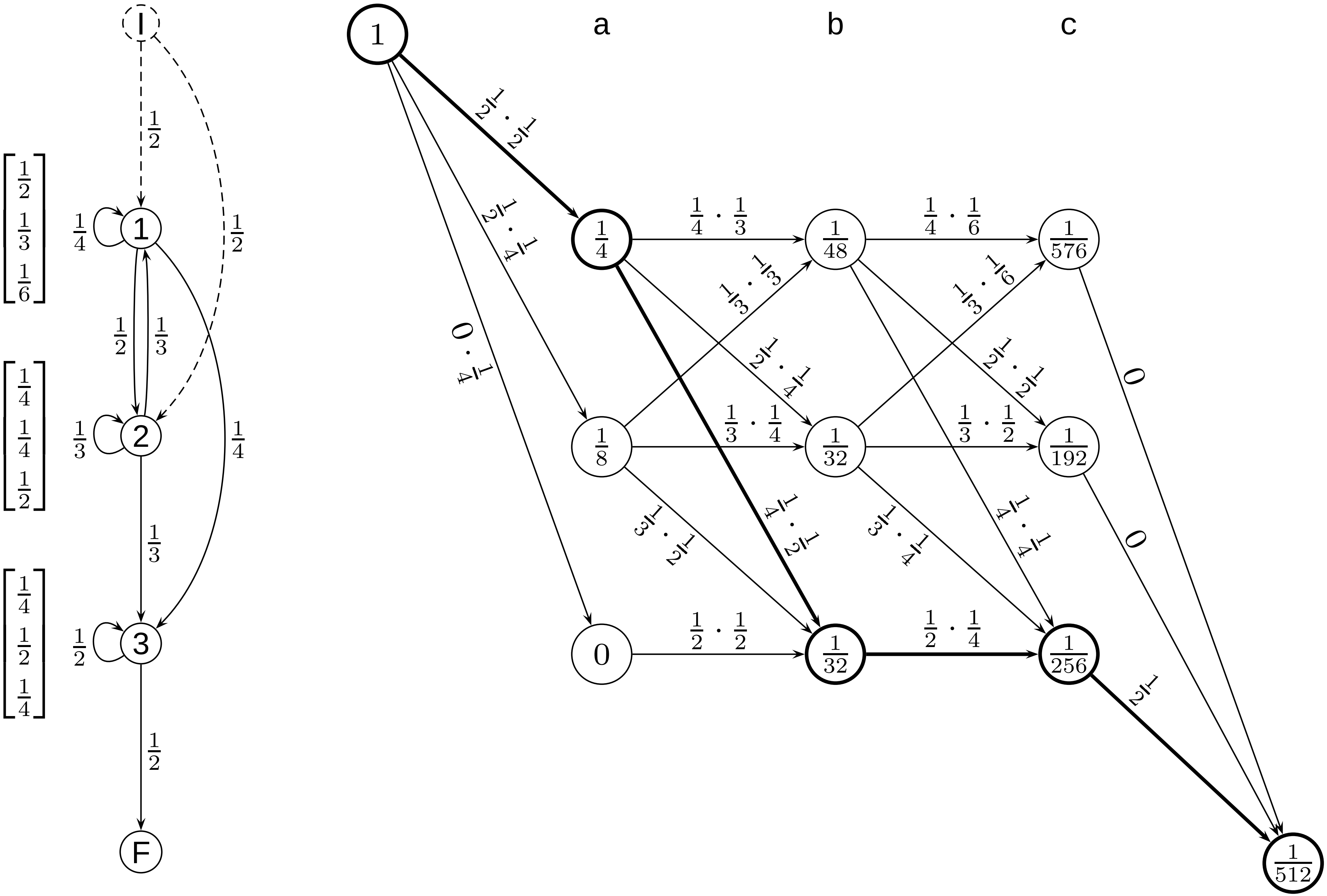
2. Find the optimal state sequence for the string.

Exercise: direct resolution

V	a $t=1$	b $t=2$	c $t=3$	
1	$\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$	$\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{3}=\frac{1}{48}$ $\frac{1}{8}\cdot\frac{1}{3}\cdot\frac{1}{3}=\frac{1}{72}$	$\frac{1}{48}\cdot\frac{1}{4}\cdot\frac{1}{6}=\frac{1}{1152}$ $\frac{1}{32}\cdot\frac{1}{3}\cdot\frac{1}{6}=\frac{1}{576}$ $\frac{1}{32}\cdot 0\cdot\frac{1}{6}=0$	
2	$\frac{1}{2}\cdot\frac{1}{4}=\frac{1}{8}$	$\frac{1}{4}\cdot\frac{1}{2}\cdot\frac{1}{4}=\frac{1}{32}$ $\frac{1}{8}\cdot\frac{1}{3}\cdot\frac{1}{4}=\frac{1}{96}$	$\frac{1}{48}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{192}$ $\frac{1}{32}\cdot\frac{1}{3}\cdot\frac{1}{2}=\frac{1}{192}$ $\frac{1}{32}\cdot 0\cdot\frac{1}{2}=0$	
3		$\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{2}=\frac{1}{32}$ $\frac{1}{8}\cdot\frac{1}{3}\cdot\frac{1}{2}=\frac{1}{48}$	$\frac{1}{48}\cdot\frac{1}{4}\cdot\frac{1}{4}=\frac{1}{768}$ $\frac{1}{32}\cdot\frac{1}{3}\cdot\frac{1}{4}=\frac{1}{384}$ $\frac{1}{32}\cdot\frac{1}{2}\cdot\frac{1}{4}=\frac{1}{256}$	
F				$\frac{1}{576}\cdot 0=0$ $\frac{1}{192}\cdot 0=0$ $\frac{1}{256}\cdot\frac{1}{2}=\frac{1}{512}$

$$\tilde{Q}=(1,3,3,F)$$

Exercise: trellis, graphic resolution



Summary

Evaluation of $P(y \mid M)$

- Probability that the Markov model M generates string y
- Calculation: $P(y \mid M) = \sum_{q_1,\dots,q_m \in Q^+} P(y,\, q_1, \dots, q_m)$.
- $P(y|M)$ can be computed with the Forward algorithm or
- we can use the approximate calculation by Viterbi: $\tilde{P}(y \mid M) = \max_{q \in Q} V(q, |y|)\, A_{q,F}$, which returns the most likely (optimal) sequence of states that generates y

Index

Syntactic-statistical classification

Assume that we have C classes of objects which are represented as strings from Σ^+. That is, one class of strings ($c \in C$) is characterized by a Markov model (M_c) that generates the strings of the class.

The question is: for a new input (string) y, which is the probability that y belongs to class c $(P(c|y))$?. In other words, which is the probability that string y is generated by Markov model M_c ($P(M_c|y)$)? More generally, which is the most probable class c (Markov model M_c) for string y? This is known as the Syntactic-statistical classification

We can use a similar approach as the statistical classification for the feature vector case. That is,

- we are given the prior probability of each class, $P(c)$; i.e. $P(M_1), P(M_2), \ldots, P(M_C)$, and
- we know the conditional probability of each class c; i.e. $P(y|M_c)$ (we calculate this with the Forward algorithm or we approximate the value by Viterbi, $\tilde{P}(y \mid M)$). We have to compute this for every class, i.e. $P(y|M_1), P(y|M_2), \ldots, P(y|M_C)$

then,

- we have to compute the posterior probability of class c, $P(c|y)$ by applying Bayes; i.e. $P(M_1|y), P(M_2|y), \ldots, P(M_c|y)$, and
- we apply the classification rule that returns the most probable class

Syntactic-statistical classification

- ***Prior probability*** of a class c: $P(c), 1 \leq c \leq C$

- ***Conditional probability*** of class c: $P(y \mid M_c)$
  - probability of obtaining string y given that is generated by the Markov model M_c; i.e.: probability that M_c generates string y
  - it is a probability function that models the distribution of strings of c in Σ^* through the Markov model M_c

- ***Posterior probability*** of a class c: $P(c \mid y)$
  - probability that the string y belongs to class c

$$P(c \mid y) = \frac{P(y \mid M_c)P(c)}{P(y)} \quad \text{where} \quad P(y) = \sum_{c'=1}^{C} P(y \mid M_{c'})P(c')$$

- ***Classification rule***: A string $y \in \Sigma^+$ is assigned to a class $\hat{c}(y)$:

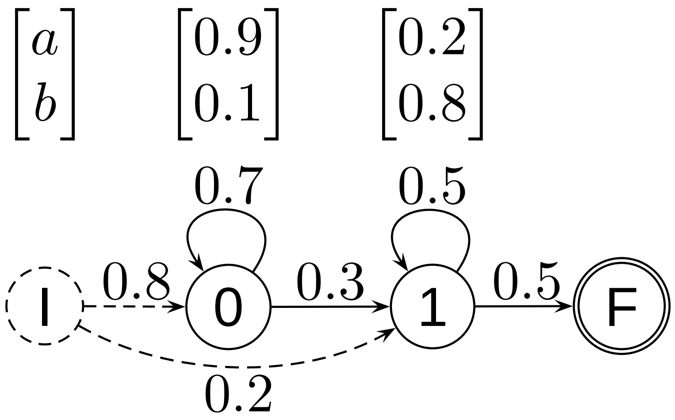
$$\hat{c}(y) = \operatorname*{argmax}_{1 \leq c \leq C} P(c \mid y)$$

Syntactic-statistical classification: exercise

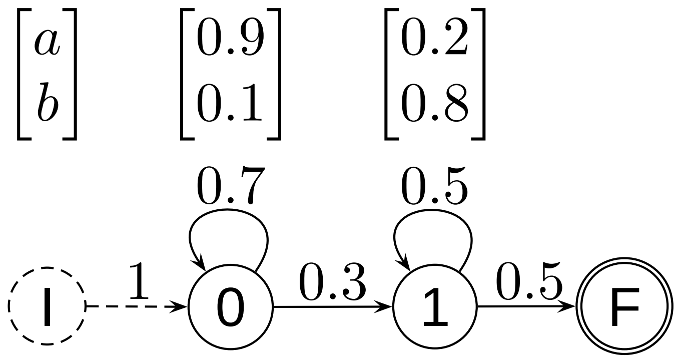
We have a two-class (A and B) classification problem of objects denoted by strings in the alphabet $\Sigma = \{a, b\}$.

The prior probabilities of the classes are $P(A) = 0.6$ y $P(B) = 0.4$. The conditional probabilities of the classes are characterized by the following Markov models:

Model M_A: $P(y \mid A) = P(y \mid M_A)$



Model M_B: $P(y \mid B) = P(y \mid M_B)$



Let $y =$ aab. Calculate $P(y \mid A)$ and $P(y \mid B)$, and then $P(A \mid y)$ and $P(B \mid y)$, and classify y by minimum classification error.

Exercise: solution

$$P(y \mid M_A)$$
$$\begin{aligned}
&= P(aab, q_1q_2q_3 = 001 \mid A)\\
&+ P(aab, q_1q_2q_3 = 011 \mid A)\\
&+ P(aab, q_1q_2q_3 = 111 \mid A)\\
&= (0.8 \cdot 0.9)\ (0.7 \cdot 0.9)\ (0.3 \cdot 0.8)\ 0.5\\
&+ (0.8 \cdot 0.9)\ (0.3 \cdot 0.2)\ (0.5 \cdot 0.8)\ 0.5\\
&+ (0.2 \cdot 0.3)\ (0.5 \cdot 0.2)\ (0.5 \cdot 0.8)\ 0.5\\
&= 0.0544 + 0.0086 + 0.0008 = \ 0.0638
\end{aligned}$$

$$P(y \mid M_B)$$
$$\begin{aligned}
&= P(aab, q_1q_2q_3 = 001 \mid B)\\
&+ P(aab, q_1q_2q_3 = 011 \mid B)\\
&= (1 \cdot 0.9)\ (0.7 \cdot 0.9)\ (0.3 \cdot 0.8)\ 0.5\\
&+ (1 \cdot 0.9)\ (0.3 \cdot 0.2)\ (0.5 \cdot 0.8)\ 0.5\\
&= 0.0680 + 0.0108\\
&= 0.0788
\end{aligned}$$

$$P(A \mid y) = \frac{P(y \mid M_A)\, P(A)}{\sum_{c'} P(y \mid M_{c'})\, P(c')} = \frac{0.0638 \cdot 0.6}{0.0638 \cdot 0.6 + 0.0788 \cdot 0.4} = 0.5484$$

$$P(B \mid y) = 1 - P(A \mid y) = 0.4516$$

$$\hat{c}(y) = \operatorname*{argmax}_{c=A,B} P(c \mid y) = A$$

Summary

Classification: $P(c \mid y)$

Probability that the string y belongs to class c

$$P(c \mid y) = \frac{P(y \mid M_c)P(c)}{P(y)} \quad \text{where} \quad P(y) = \sum_{c'=1}^{C} P(y \mid M_{c'})P(c')$$

and $P(y \mid M_c)$ and $P(y \mid M_{c'})$ can be approximated by Viterbi

Exercise: syntactic-statistical classification by using Viterbi

In practice, the conditional probabilities of the classes are typically approximated by Viterbi. Let's get back to the exercise in page 21:

$$\begin{aligned}\tilde{P}(y \mid M_A) \\ &= \max(P(aab, q_1q_2q_3 = 001 \mid A), \\ &\qquad P(aab, q_1q_2q_3 = 011 \mid A), \\ &\qquad P(aab, q_1q_2q_3 = 111 \mid A)) \\ &= \max(0.0544, 0.0086, 0.0008) \\ &= 0.0544\end{aligned}$$

$$\begin{aligned}\tilde{P}(y \mid M_B) \\ &= \max(P(aab, q_1q_2q_3 = 001 \mid B), \\ &\qquad P(aab, q_1q_2q_3 = 011 \mid B)) \\ &= \max(0.0680, 0.0108) \\ &= 0.0680\end{aligned}$$

$$\tilde{P}(A \mid y) = \frac{\tilde{P}(y \mid M_A)\, P(A)}{\sum_{c'} \tilde{P}(y \mid c')\, P(c')} = \frac{0.0544 \cdot 0.6}{0.0544 \cdot 0.6 + 0.0680 \cdot 0.4} = 0.5455$$

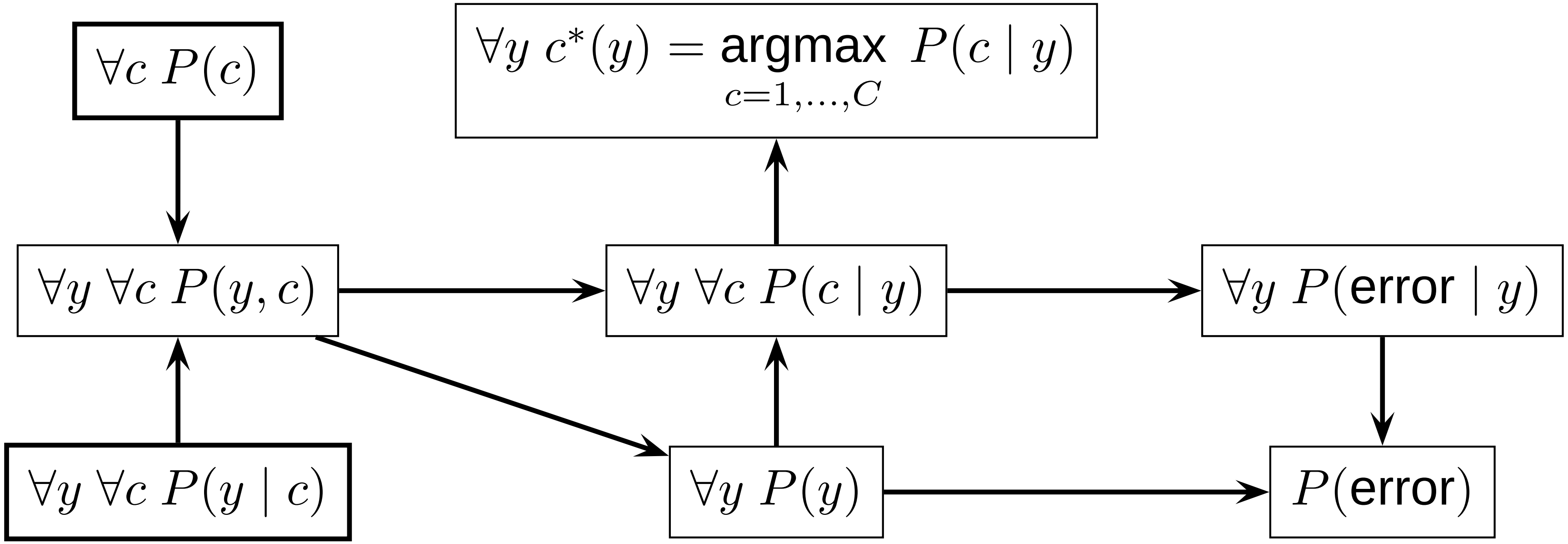
$$\tilde{P}(B \mid y) = 1 - \tilde{P}(A \mid y) = 0.4545$$

$$\tilde{c}(y) = \underset{c=A,B}{\text{argmax}}\, \tilde{P}(c \mid y) = A$$ identical result to the one in page 21

Index

Annex: calculations in statistical classification (summary)

The statistical approach for classifying objects represented as feature vectors is also valid for objects represented as strings of symbols in a given alphabet ($y \in \Sigma^+$):



Exercise:

- Give name and formula to the nodes in the chart
- Calculate $P(c)$ from $P(y, c) \; \forall y$.
- Calculate $P(y \mid c)$ from $P(y, c)$ and $P(c)$.
- Calculate $P(y, c)$ from $P(c \mid y)$ and $P(y)$.

Annex: calculations in statistical classification (summary)

$P(c)$	***Prior probability*** of class c
$P(y \mid c)$	***Conditional probability*** of class c
$P(y, c) = P(c)\, P(y \mid c)$	***Joint probability*** of a class c and string y
$P(y) = \sum_{c=1,\ldots,C} P(y, c)$	***Unconditional probability*** of a string y
$P(c \mid y) = \dfrac{P(c)\, P(y \mid c)}{P(y)}$	***Posterior probability*** of class c (for string y)
$c^*(y) = \text{argmax}_{c=1,\ldots,C}\ P(c \mid y)$	***Bayes decision rule*** for min. classification error
$P(\text{error} \mid y) = 1 - \max_{c=1,\ldots,C} P(c \mid y)$	***Local Bayes error*** (minimum probability of error)
$P(\text{error}) = \sum_{y \in \Sigma^+} P(y)\, P(\text{error} \mid y)$	***Global Bayes error*** (min. average prob. of error)

$$P(c) = \sum_{y \in \Sigma^+} P(y, c) \qquad P(y \mid c) = \frac{P(y, c)}{P(c)} \qquad P(y, c) = P(c)\, P(y \mid c)$$