PROBABILISTIC GRAMMARS & PROBABILITY DISTRIBUTIONS

1. Probability distributions

Die A : 1, 1, 2, 2, 3, 3 Die B : 1, 2, 2, 3, 3, 3

Roll 1 : Die A

Roll 2 : flip a coin, H → Die A, T → Die B

	A 1	A 2	A 3	B 1	B 2	B 3
A 1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$
A 2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$
A 3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$

$\{1,2,3\} \times \{A,B\} \times \{1,2,3\}$

Random variables : X_1 first roll $\quad X_1 : \Omega \to \{1,2,3\}$

X_2 second roll $\quad X_2 : \Omega \to \{1,2,3\}$

C_2 second die $\quad C_2 : \Omega \to \{A,B\}$

TOTAL $\quad TOTAL : \Omega \to \{2,3,4,5,6\}$

$$Pr(X_1 = 3, C_2 = A, X_2 = 2) = \frac{1}{18}$$

$$Pr(TOTAL = 6) = \frac{1}{18} + \frac{1}{12} = \frac{5}{36}$$

$$Pr(\{x : x \in \Omega \text{ and } TOTAL(x) = 6\}) = \frac{5}{36}$$

	A 1	A 2	A 3	B 1	B 2	B 3
A 1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$
A 2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$
A 3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$

marginal probability

$$\boxed{\Pr(X_1 = 3)} = \frac{1}{18} + \frac{1}{18} + \frac{1}{18} + \frac{1}{36} + \frac{1}{18} + \frac{1}{12} = \frac{1}{3}$$

$$= \sum_{c \in \{A,B\}} \sum_{x \in \{1,2,3\}} \boxed{\Pr(X_1 = 3, C_2 = c, X_2 = x)}$$

joint probability

$$\Pr(\mathrm{TOTAL} \geq 5) = \frac{7}{18}$$

$$\Pr(\mathrm{TOTAL} \geq 5, X_1 = 3) = 3 \times \frac{1}{18} + \frac{1}{12} = \frac{1}{4}$$

$$\frac{\Pr(\mathrm{TOTAL} \geq 5, X_1 = 3)}{\Pr(X_1 = 3)} = \frac{\left(\frac{1}{4}\right)}{\left(\frac{1}{3}\right)} = \frac{3}{4}$$

$$\boxed{\Pr(\mathrm{TOTAL} \geq 5 \mid X_1 = 3)} = \frac{3}{4}$$

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)}$$

conditional probability

conditional probability

Non-independent:

$$Pr(TOTAL \geq 5, X_1 = 3) = \frac{1}{4} \neq Pr(TOTAL \geq 5) \times P(X_1 = 3)$$

$$Pr(TOTAL \geq 5) = \frac{7}{18}$$

$$Pr(TOTAL \geq 5 \mid X_1 = 3) \neq Pr(TOTAL \geq 5)$$

$$Pr(X_1 = 3) = \frac{1}{3}$$

Independent:

$$Pr(X_1 = 3, X_2 = 3) = Pr(X_1 = 3) \times Pr(X_2 = 3)$$

$$Pr(X_1 = 3 \mid X_2 = 3) = Pr(X_1 = 3)$$

Chain rule: $P(A,B) = P(A) \times P(B \mid A)$

$\approx$ SLGS

n-gram

2. Markov models of distributions over Σ^*

probability of "abc" $= Pr(X_1 = a, X_2 = b, X_3 = c, X_4 = STOP)$

$$= Pr(X_1 = a)\, Pr(X_2 = b \mid X_1 = a) \times Pr(X_3 = c \mid X_1 = a, X_2 = b) \times Pr(X_4 = STOP \mid X_1 X_2 X_3 = abc)$$

Assumption ①: Only the last symbol matters.

prob of "abc" $= Pr(X_1 = a) \times Pr(X_2 = b \mid X_1 = a) \times Pr(X_3 = c \mid X_2 = b) \times Pr(X_4 = STOP \mid X_3 = c)$

Assumption ②: Position in the string doesn't matter.

e.g. $Pr(X_2 = b \mid X_1 = a) = Pr(X_{18} = b \mid X_{17} = a)$

So define PSLG:

Σ alphabet

$I : \Sigma \to \text{Prob}$ $\qquad I(x) = Pr(X_1 = x)$

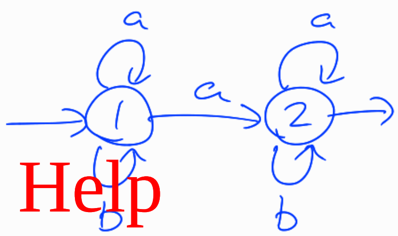
$F : \Sigma \to \text{Prob}$ $\qquad F(x) = Pr(X_{i+1} = \text{STOP} \mid X_i = x)$

$T : \Sigma \times \Sigma \to \text{Prob}$ $\qquad T(x,y) = Pr(X_{i+1} = y \mid X_i = x)$

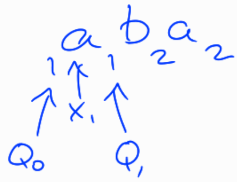
prob of "abc" $= I(a)\, T(a,b)\, T(b,c)\, F(c)$

$$\sum_x I(x) = 1 \qquad \forall x : \sum_y T(x,y) + F(x) = 1$$

3. Probabilistic FSAs



$Pr(Q_0 = 1, X_1 = a, Q_1 = 1, X_2 = a, Q_2 = 2,$
$\quad X_3 = b, Q_3 = 2, Q_4 = \text{STOP})$



PFSA:

Σ alphabet

Q states

$I : Q \to \text{Prob}$ $\qquad I(q) = Pr(Q_0 = q)$

$F : Q \to \text{Prob}$ $\qquad F(q) = Pr(Q_{i+1} = \text{STOP} \mid Q_i = q)$

$\Delta : Q \times \Sigma \times Q \to \text{Prob}$ $\qquad \Delta(q, x, q') = Pr(X_{i+1} = x, Q_{i+1} = q' \mid Q_i = q)$

Assumptions:

1. Only the last state matters.
2. Position in the string/state seq. doesn't matter.

$$= Pr(Q_0 = 1)\, Pr(X_1 = a, Q_1 = 1 \mid Q_0 = 1)$$
$$Pr(X_2 = a, Q_2 = 2 \mid Q_1 = 1)$$
$$Pr(X_3 = b, Q_3 = 2 \mid Q_2 = 2)\; Pr(Q_4 = \text{STOP} \mid Q_3 = 2)$$
$$= I(1)\, \Delta(1, a, 1)\, \Delta(1, a, 2)\, \Delta(2, b, 2)\, F(2)$$

$$\text{prob of "aab"} = \Pr(X_1=a, X_2=a, X_3=b, Q_4=\text{STOP})$$

$$= \sum_{q_0}\sum_{q_1}\sum_{q_2}\sum_{q_3}\Big[I(q_0)\,\Delta(q_0,a,q_1)\,\Delta(q_1,a,q_2)\,\Delta(q_2,b,q_3)\,F(q_3)\Big]$$

3.1 Forward probabilities

prob of $x_1 \cdots x_n$

$$= \Pr(X_1=x_1,\ X_2=x_2,\ \cdots X_n=x_n,\ Q_{n+1}=\text{STOP})$$

$$= \sum_{q_n}\Pr(X_1=x_1,\ X_2=x_2,\ \ldots X_n=x_n,\ Q_n=q_n,\ Q_{n+1}=\text{STOP})$$

$$= \sum_{q_n}\Big[\Pr(X_1 \cdots X_n = x_1 \cdots x_n,\ Q_n=q_n)\ \Pr(Q_{n+1}=\text{STOP} \mid Q_n=q_n)\Big]$$

$$= \sum_{q_n}\Big[\underbrace{\Pr(X_1 \cdots X_n = x_1 \cdots x_n,\ Q_n=q_n)}_{\text{fwd}(x_1 \cdots x_n)(q_n)} \times F(q_n)\Big]$$

$$\boxed{f(w) = \sum_q\Big[\text{fwd}(w)(q) \times F(q)\Big]}$$

$$\Pr(X_1 \cdots X_n = x_1 \cdots x_n,\ Q_n=q)$$

$$= \sum_{q_{n-1}}\Big[\Pr(X_1 \cdots X_{n-1}=x_1 \cdots x_{n-1},\ Q_{n-1}=q_{n-1},\ X_n=x_n,\ Q_n=q)\Big]$$

$$= \sum_{q_{n-1}}\Big[\Pr(X_1 \cdots X_{n-1}=x_1 \cdots x_{n-1},\ Q_{n-1}=q_{n-1}) \times \Pr(X_n=x_n,\ Q_n=q \mid Q_{n-1}=q_{n-1})\Big]$$

$$\boxed{\mathrm{fwd}(x_1 \cdots x_n)(q) = \sum_{q_{n-1}} \left[\mathrm{fwd}(x_1 \cdots x_{n-1})(q_{n-1}) \times \Delta(q_{n-1}, x_n, q) \right]}$$

3.2 Backward probabilities

prob of $x_1 \cdots x_n$

$$= \Pr(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n, Q_{n+1} = \mathrm{STOP})$$

$$= \sum_{q_0} \Pr(Q_0 = q_0, X_1 \cdots X_n = x_1 \cdots x_n, Q_{n+1} = \mathrm{STOP})$$

$$= \sum_{q_0} \left[\Pr(Q_0 = q_0) \Pr(X_1 \cdots X_n = x_1 \cdots x_n, Q_{n+1} = \mathrm{STOP} \mid Q_0 = q_0) \right]$$

$$= \sum_{q_0} \left[I(q_0) \underbrace{\Pr(X_1 \cdots X_n = x_1 \cdots x_n, Q_{n+1} = \mathrm{STOP} \mid Q_0 = q_0)}_{\mathrm{bwd}(x_1 \cdots x_n)(q_0)} \right]$$

$$\mathrm{bwd}(x_1 \cdots x_n)(q)$$

$$= \Pr(X_{i+1} \cdots X_{i+n} = x_1 \cdots x_n, Q_{i+n+1} = \mathrm{STOP} \mid Q_i = q)$$

$$\boxed{f(w) = \sum_{q} \left[I(q) \times \mathrm{bwd}(w)(q) \right]}$$

$$\Pr(X_{i+1} \cdots X_{i+n} = x_1 \cdots x_n,\ Q_{i+n+1} = \mathrm{STOP} \mid Q_i = q)$$

$$= \sum_{q'} \Pr(X_{i+1} = x_1,\ Q_{i+1} = q',\ X_{i+2} \cdots X_{i+n} = x_2 \cdots x_n,\ Q_{i+n+1} = \mathrm{STOP} \mid Q_i = q)$$

$$= \sum_{q'} \Big[\Pr(X_{i+1} = x_1,\ Q_{i+1} = q' \mid Q_i = q) \times \underbrace{\Pr(X_{i+2} \cdots X_{i+n} = x_2 \cdots x_n,\ Q_{i+n+1} = \mathrm{STOP} \mid Q_{i+1} = q')}_{\mathrm{bwd}(x_2 \cdots x_n)(q')} \Big]$$

$$\boxed{\mathrm{bwd}(x_1 \cdots x_n)(q) = \sum_{q'} \Big[\Delta(q, x_1, q') \times \mathrm{bwd}(x_2 \cdots x_n)(q') \Big]}$$

Hidden Markov Model (HMM)

Viterbi alg.

Viterbi semiring

$$Q_0 \xrightarrow{X_1} Q_1 \xrightarrow{X_2} Q_2 \xrightarrow{X_3} Q_3$$

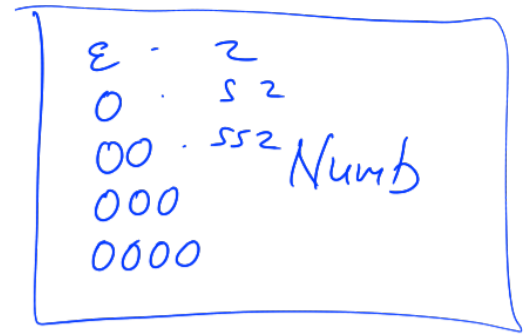
$$\begin{tikzcd}
X_1 & X_2 & X_3 \\
Q_1 \arrow[u] \arrow[r] & Q_2 \arrow[u] \arrow[r] & Q_3 \arrow[u]
\end{tikzcd}$$

emission $\Pr(X_i \mid Q_i)$

transition $\Pr(Q_{i+1} \mid Q_i)$

prob of $x_1 x_2 x_3$ $$= \sum_{q_1} \sum_{q_2} \sum_{q_3} \Big[\Pr(Q_1 = q_1) \Pr(Q_2 = q_2 \mid Q_1 = q_1) \Pr(Q_3 = q_3 \mid Q_2 = q_2) \Pr(X_1 = x_1 \mid Q_1 = q_1) \Pr(X_2 = x_2 \mid Q_2 = q_2) \Pr(X_3 = x_3 \mid Q_3 = q_3) \Big]$$

4 Probabilistic CFGs

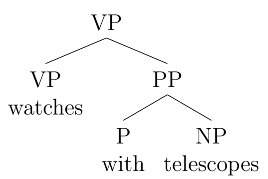


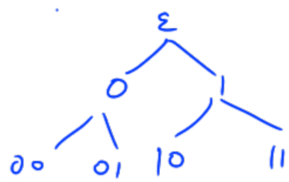
The “hidden structure” here has a more complicated shape than we had to deal with for PFSAs:

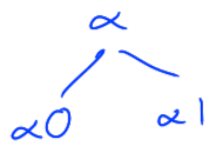
- With a (P)FSA, the hidden additional structure was a collection of *states*, indexed by their position in a *sequence* (i.e. Q_0, Q_1, Q_2, ...).
- Now with a (P)CFG, the hidden additional structure is a collection of *nonterminals*, indexed by their position in a *tree*.

The hidden structure associated with the one tree via which we can generate ‘watches with telescopes’ (as a VP) can be expressed as follows, where the nonterminals are indexed by Gorn addresses:

(1) $\Pr(N_\epsilon = \text{VP}, N_0 = \text{VP}, X_0 = \text{watches}, N_1 = \text{PP}, N_{10} = \text{P}, X_{10} = \text{with}, N_{11} = \text{NP}, X_{11} = \text{telescopes})$







We'll set up a PCFG as the kind of thing that defines joint probabilities like (1).

(2) A probabilistic context-free grammar (PCFG) is a four-tuple $\langle N, \Sigma, I, R\rangle$ where:

- N is a finite set of nonterminal symbols;
- Σ, the alphabet, is a finite set of terminal symbols (disjoint from N);
- $I : N \to [0, 1]$ is a function assigning initial probabilities; and
- $R : N \times ((N \times N) \cup \Sigma) \to [0, 1]$ is a function assigning rule probabilities;

with the additional requirements that

- $\sum_{n \in N} I(n) = 1$;
- for all $n \in N$, $\sum_{\ell \in N} \sum_{r \in N} R(n, \ell, r) + \sum_{x \in \Sigma} R(n, x) = 1$; and
- there are no “dead ends”.

We take the values assigned by the I and R functions to be giving us “local” conditional probabilities:

(3)
$$\begin{aligned} I(n) &= \Pr(N_\epsilon = n) \\ R(n, \ell, r) &= \Pr(N_{\alpha 0} = \ell, N_{\alpha 1} = r \mid N_\alpha = n) \\ R(n, x) &= \Pr(X_\alpha = x \mid N_\alpha = n) \end{aligned}$$

$\Pr(Q_0 = q)$

And we make the following important *independence assumptions*, from which it will follow that the big joint probability in (1) can be computed as a product of the grammar's local conditional probabilities:

(4)
a. **Limited horizon:**
The choice of how to expand a node N_α is independent of the ancestors of node N_α.

b. **Context-free-ness:**
The choice of how to expand a node N_α is independent of anything to the left or right of node N_α.

c. **Place invariance:**
The choice of how to expand a node N_α is independent of the position α of the node in the tree.

Specifically:

(5) $\Pr(N_\epsilon = \text{VP}, N_0 = \text{VP}, X_0 = \text{watches}, N_1 = \text{PP}, N_{10} = \text{P}, X_{10} = \text{with}, N_{11} = \text{NP}, X_{11} = \text{telescopes})$

$$= \Pr(N_\epsilon = \text{VP}) \Pr(N_0 = \text{VP}, N_1 = \text{PP} \mid N_\epsilon = \text{VP}) \Pr(X_0 = \text{watches} \mid N_0 = \text{VP})$$
$$\Pr(N_{10} = \text{P}, N_{11} = \text{NP} \mid N_1 = \text{PP}) \Pr(X_{10} = \text{with} \mid N_{10} = \text{P}) \Pr(X_{11} = \text{telescopes} \mid N_1 1 = \text{NP})$$
$$= I(\text{VP}) \times R(\text{VP}, \text{VP}, \text{PP}) \times R(\text{VP}, \text{watches}) \times R(\text{PP}, \text{P}, \text{NP}) \times R(\text{P}, \text{with}) \times R(\text{NP}, \text{telescopes})$$

Some notation for dealing with this hidden tree structure:

(6) a. $X_{\widehat{\alpha}}$ is the random variable for the substring *dominated by* the node at position α.
b. $X_{\overleftarrow{\alpha}}$ is the random variable for the substring *to the left of* the node at position α.
c. $X_{\overrightarrow{\alpha}}$ is the random variable for the substring *to the right of* the node at position α.

The probability of a string involves summing over all tree structures:

(7) $$\Pr(X_{\widehat{\epsilon}} = \text{watches with telescopes}) = \sum_t \Pr(X_{\widehat{\epsilon}} = \text{watches with telescopes}, T = t)$$

It's hard to enumerate all those possible choices of t, so we interleave it with the calculation of probabilities for subparts of the tree (just like we did for boolean CFGs).

4.1 Inside probabilities

Let's just add in a specific choice of root nonterminal:

(8) $$\Pr(X_{\widehat{\epsilon}} = x_1 \dots x_n) = \sum_n \Pr(X_{\widehat{\epsilon}} = x_1 \dots x_n, N_\epsilon = n)$$
$$= \sum_n \Pr(N_\epsilon = n) \Pr(X_{\widehat{\epsilon}} = x_1 \dots x_n \mid N_\epsilon = n)$$

A probability of the form $\Pr(X_{\widehat{\alpha}} = w \mid N_\alpha = n)$ is what we previously called $\text{inside}(w)(n)$. So the equation says the same thing as this equation that we've seen before:

(9) $$\boxed{f(w) = \sum_n \left[I(n) \times \text{inside}(w)(n)\right]}$$

We can also break these down recursively:

(10) $\Pr(X_{\widehat{\alpha}} = x_1 \dots x_m \mid N_\alpha = n)$

$$= \sum_i \sum_\ell \sum_r \left[\Pr(N_{\alpha 0} = \ell, N_{\alpha 1} = r, X_{\widehat{\alpha 0}} = x_1 \dots x_i, X_{\widehat{\alpha 1}} = x_{i+1} \dots x_m \mid N_\alpha = n)\right]$$
$$= \sum_i \sum_\ell \sum_r \Big[\Pr(N_{\alpha 0} = \ell, N_{\alpha 1} = r \mid N_\alpha = n) \times \Pr(X_{\alpha 0} = x_1 \dots x_i \mid N_{\alpha 0} = \ell)$$
$$\times \Pr(X_{\alpha 1} = x_{i+1} \dots x_m \mid N_{\alpha 1} = r)\Big]$$

This is the recursive description of inside values that we saw before:

(11) $$\boxed{\text{inside}(x_1 \dots x_m)(n) = \sum_i \sum_\ell \sum_r \left[R(n, \ell, r) \times \text{inside}(x_1 \dots x_i)(\ell) \times \text{inside}(x_{i+1} \dots x_m)(r)\right]}$$

4.2 Outside probabilities

Things work just the same way for outside probabilities.

If we add in just a single chosen leaf node at address α:

(12) $$\begin{aligned}
&\Pr(X_{\widehat{\epsilon}} = x_1 \dots x_m) \\
&\quad = \sum_{n \in N} \Big[\Pr(N_\alpha = n, X_{\overleftarrow{\alpha}} = x_1 \dots x_{i-1}, X_{\widehat{\alpha}} = x_i, X_{\overrightarrow{\alpha}} = x_{i+1} \dots x_m) \Big] \\
&\quad = \sum_{n \in N} \Big[\Pr(N_\alpha = n, X_{\overleftarrow{\alpha}} = x_1 \dots x_{i-1}, X_{\overrightarrow{\alpha}} = x_{i+1} \dots x_m) \times \Pr(X_{\widehat{\alpha}} = x_i \mid N_\alpha = n) \Big]
\end{aligned}$$

which we saw before as:

(13) $$f(x_1 \dots x_m) = \sum_n \big[\text{outside}(x_1 \dots x_{i-1}, x_{i+1} \dots x_m)(n) \times R(n)(x_i) \big]$$

And the recursive description:

(14) $$\begin{aligned}
&\Pr(X_{\overleftarrow{\alpha}} = y_1 \dots y_n, X_{\overrightarrow{\alpha}} = z_1 \dots z_m, N_\alpha = x) \\
&\quad = \Pr(X_{\overleftarrow{\beta 0}} = y_1 \dots y_n, X_{\overrightarrow{\beta 0}} = z_1 \dots z_m, N_{\beta 0} = x) \\
&\qquad + \Pr(X_{\overleftarrow{\beta 1}} = y_1 \dots y_n, X_{\overrightarrow{\beta 1}} = z_1 \dots z_m, N_{\beta 1} = x) \\
&\quad = \sum_i \sum_p \sum_r \Big[\Pr(N_{\beta 0} = x, N_{\beta 1} = r \mid N_\beta = p) \times \Pr(X_{\widehat{\beta 1}} = z_1 \dots z_i \mid N_{\beta 1} = r) \\
&\qquad\qquad \times \Pr(X_{\overleftarrow{\beta}} = y_1 \dots y_n, X_{\overrightarrow{\beta}} = z_{i+1} \dots z_m, N_\beta = p) \Big] \\
&\qquad + \sum_i \sum_p \sum_\ell \Big[\Pr(N_{\beta 0} = \ell, N_{\beta 1} = x \mid N_\beta = p) \times \Pr(X_{\widehat{\beta 0}} = y_{i+1} \dots y_n \mid N_{\beta 0} = \ell) \\
&\qquad\qquad \times \Pr(X_{\overleftarrow{\beta}} = y_1 \dots y_i, X_{\overrightarrow{\beta}} = z_1 \dots z_m, N_\beta = p) \Big]
\end{aligned}$$

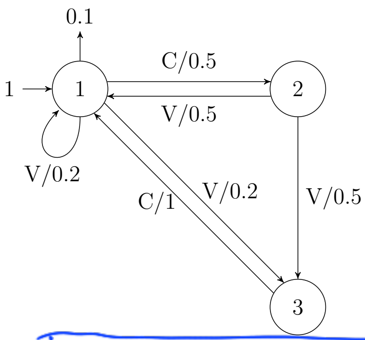
which we saw before as:

(15) $$\begin{aligned}
&\text{outside}_G(y_1 \dots y_m, z_1 \dots z_q)(n) \\
&\quad = \sum_i \sum_p \sum_r \Big[R(p, n, r) \times \text{inside}_G(z_1 \dots z_i)(r) \times \text{outside}_G(y_1 \dots y_m, z_{i+1} \dots z_q)(p) \Big] \\
&\qquad + \sum_i \sum_p \sum_\ell \Big[R(p, \ell, n) \times \text{inside}_G(y_{i+1} \dots y_m)(\ell) \times \text{outside}_G(y_1 \dots y_i, z_1 \dots z_q)(p) \Big]
\end{aligned}$$

5 Probabilities of structure given a string

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

This PFSA assigns joint probabilities, each of which we can think of as the probability of a certain string along with a certain syllable structure.



V C V
1 1 2 1
1 3 1 1

$\Pr(X_1 \cdots X_3 = VCV, Q_0 \cdots Q_4 = 1121) = 0.005$

$\Pr(X_1 \cdots X_3 = VCV, Q_0 \cdots Q_4 = 1311) = 0.004$

$\Pr(X_1 \cdots X_3 = VCV) = 0.005 + 0.004 = 0.009$

$$\frac{0.004}{0.009} = \frac{\Pr(X_1 \cdots X_3 = VCV, Q_0 \cdots Q_4 = 1311)}{\Pr(X_1 \cdots X_3 = VCV)}$$

$\Pr(Q_0 \cdots Q_4 = 1311 \mid X_1 \cdots X_3 = VCV)$

x_1	x_2	x_3	x_4	x_5
O	O	B	I	

$\Pr(Q_3 = 1 \mid X_1 \cdots X_{10} = CVCVCVCVCV)$

$$= \frac{\Pr(Q_3 = 1, X_1 \cdots X_{10} = CVCVCVCVCV)}{\Pr(X_1 \cdots X_{10} = CVCVCVCVCV)}$$

$$= \frac{\Pr(X_1 \cdots X_3 = CVC, Q_3 = 1, X_4 \cdots X_{10} = VCVCVCV)}{\Pr(\_\_\_)}$$

$Q_{11} = STOP$

$= \Pr(X_1 \cdots X_3 = CVC, Q_3 = 1) \Pr(X_4 \cdots X_{10} = VCVCVCV \mid Q_3 = 1) / \Pr(\_\_\_)$

$= fwd(CVC)(1) \times bwd(VCVCVCV)(1) / \Pr(\_\_\_)$

6 Fitting a probabilistic model to data

6.1 Three-sided coin example

Suppose we have a "three-sided coin", with faces A, B and C.

We know that there's some number w (and $0 \leq w \leq 1$) such that, on any particular flip of the coin:

- the probability of A is w^2,
- the probability of B is $w(1-w)$, and
- the probability of C is $1-w$.

So, for example:

(16)

	$w=0$	$w=0.25$	$w=0.5$	$w=0.75$	$w=1$
probability of A	0	0.0625	0.25	0.5625	1
probability of B	0	0.1875	0.25	0.1875	0
probability of C	1	0.75	0.5	0.25	0

We flip the coin ten times and get three As, five Bs and two Cs. In general, the probability or *likelihood* of this outcome is:

(17) $$(\text{probability of A})^3 \times (\text{probability of B})^5 \times (\text{probability of C})^2$$

This value will depend on the value of w:

(18) a. If w were 0.25, then the likelihood would be $(0.0625)^3 \times (0.1875)^5 \times (0.75)^2 \approx 3.18 \times 10^{-8}$.
b. If w were 0.5, then the likelihood would be $(0.25)^3 \times (0.25)^5 \times (0.5)^2 \approx 3.81 \times 10^{-6}$.
c. If w were 0.75, then the likelihood would be $(0.5625)^3 \times (0.1875)^5 \times (0.25)^2 \approx 2.58 \times 10^{-6}$.

So of these three options, $w = 0.5$ makes the likelihood the greatest. But which is the best option *overall*?

(19) $$\begin{aligned} \text{likelihood } L(w) &= (\text{probability of A})^3 \times (\text{probability of B})^5 \times (\text{probability of C})^2 \\ &= (w^2)^3 \times (w(1-w))^5 \times (1-w)^2 \\ &= w^{11}(1-w)^7 \end{aligned}$$

(20) $$\begin{aligned} \frac{dL(w)}{dw} &= 11w^{10} \times (1-w)^7 + w^{11} \times \big(-7(1-w)^6\big) \qquad \text{(product rule for differentiation)} \\ &= w^{10}(1-w)^6\big(11(1-w) - 7w\big) \end{aligned}$$

This derivative is zero when:

(21) $w = 0$ or $1 - w = 0$, $w = 1$ or
$$\begin{aligned} 11(1-w) - 7w &= 0 \\ 11 - 11w - 7w &= 0 \\ 11 &= 18w \\ w &= \tfrac{11}{18} \approx 0.611 \end{aligned}$$

Of these, the only local maximum is $w \approx 0.611$. So this is the value of w which assigns the highest likelihood to the observed flips. (And with this value of w, the probabilities of A, B and C are about 0.373, 0.238 and 0.389, respectively.)

100 flips: 63 H 37 T

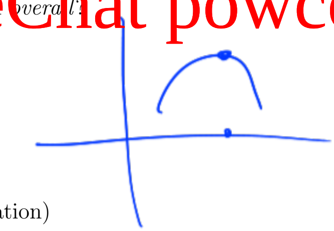
Three guesses: 0.5 0.6 0.7

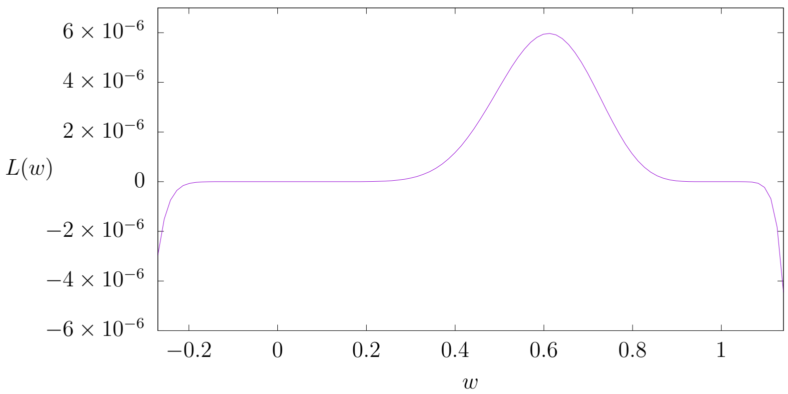
$0.5 \rightsquigarrow 0.5^{63}\, 0.5^{37} = 7.9 \times 10^{-31}$

$0.6 \rightsquigarrow 0.6^{63}\, 0.4^{37} = 2.0 \times 10^{-29}$

$0.7 \rightsquigarrow 0.7^{63}\, 0.3^{37} = 7.8 \times 10^{-30}$

$0.63 \rightsquigarrow 0.63^{63}\, 0.37^{37} = \underbrace{2.4 \times 10^{-29}}_{\text{likelihood}}$





$\lambda w.\ -w^{11} * (1-w)^7$

```
$ python3 -q
>>> from scipy.optimize import minimize_scalar
>>> minimize_scalar(lambda w: - w**11 * (1-w)**7)
fun: -5.9717377174644157e-06
nfev: 13
nit: 12
success: True
x: 0.61111111104523153
>>>
```


6.2 Two-sided coin example

Suppose now we have a more traditional two-sided coin, with faces H and T.

We know that there's some number w (and $0 \le w \le 1$) such that, on any particular flip of the coin:

- the probability of H is w, and
- the probability of T is $1 - w$.

prob = relative freq.

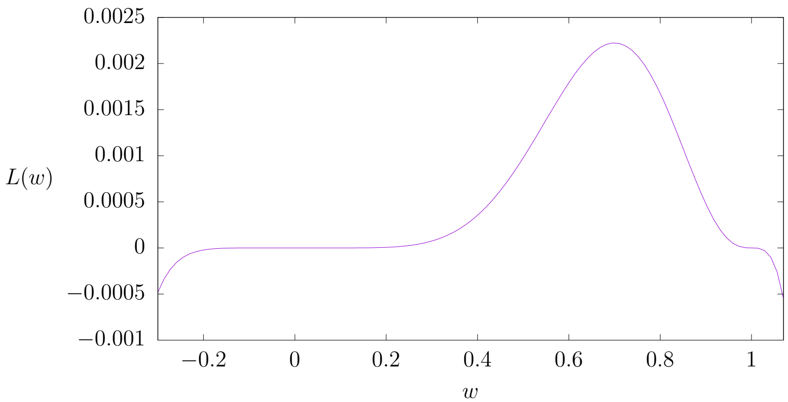
We flip the coin ten times and get seven Hs and three Ts.

(22)
$$\begin{aligned} \text{likelihood } L(w) &= (\text{probability of H})^7 \times (\text{probability of T})^3 \\ &= w^7(1-w)^3 \end{aligned}$$

(23)
$$\begin{aligned} \frac{dL(w)}{dw} &= 7w^6 \times (1-w)^3 + w^7 \times \big(-3(1-w)^2\big) && \text{(product rule for differentiation)} \\ &= w^6(1-w)^2\big(7(1-w) - 3w\big) \end{aligned}$$

This derivative is zero when:

(24)
$$\begin{array}{ccccc} w = 0 & \text{or} & 1 - w = 0 & \text{or} & 7(1-w) - 3w = 0 \\ & & w = 1 & & 7 - 7w - 3w = 0 \\ & & & & 7 = 10w \\ & & & & w = \frac{7}{10} \end{array}$$




```
$ python3 -q
>>> from scipy.optimize import minimize_scalar
>>> minimize_scalar(lambda w: - w**7 * (1-w)**3)
fun: -0.0022235660999999998
nfev: 13
nit: 12
success: True
x: 0.7000000001004768
>>>
```

$\lambda w.\, w^7(1-w)^3$

6.3 Two-parameter three-sided coin example

Suppose now we have another three-sided coin, with faces A, B and C.

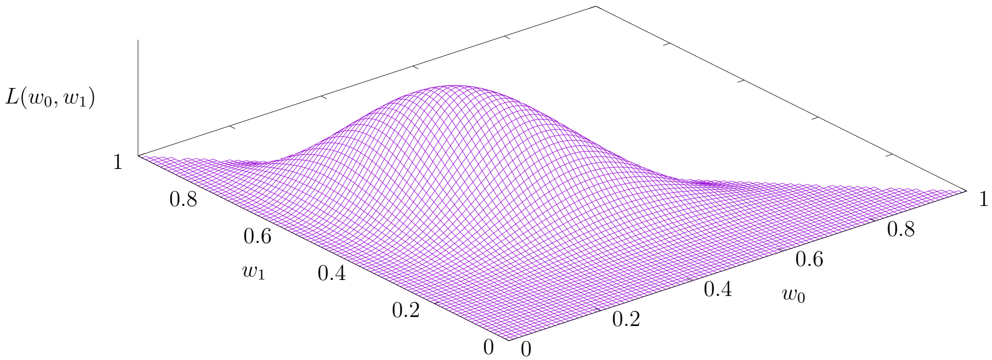
We know that there are two numbers w_0 and w_1 such that, on any particular flip of the coin:

- the probability of A is w_0,
- the probability of B is w_1, and
- the probability of C is $(1 - w_0 - w_1)$.

We flip the coin ten times and get three As, five Bs and two Cs.

(25)
$$\begin{aligned}\text{likelihood } L(w_0, w_1) &= (\text{probability of A})^3 \times (\text{probability of B})^5 \times (\text{probability of C})^2 \\ &= {w_0}^3 \times {w_1}^5 \times (1 - w_0 - w_1)^2\end{aligned}$$

High-school calculus is less help here, but the principles are the same.




```
$ python3 -q
>>> from scipy.optimize import minimize
>>> l = lambda w: - (w[0]**3 * w[1]**5 * (1 - w[0] - w[1])**2)
>>> c = #{'type': 'ineq', 'fun': (lambda w: 1 - w[0] - w[1])#}
>>> b = [(0,1),(0,1)]
>>> minimize(l, (0.2,0.2), bounds=b, constraints=c, tol=1e-10)
fun: -3.3749999736171599e-05
jac: array([ -3.45876288e-08,  -9.02400643e-09,   0.00000000e+00])
message: 'Optimization terminated successfully.'
nfev: 70
nit: 17
njev: 17
status: 0
success: True
x: array([ 0.29998248,  0.50000868])
>>>
```


PSLG

$$\Pr(X_i = a \mid X_{i-1} = b) = \frac{\text{count}(ba)}{\text{count}(b)}$$

PFSA

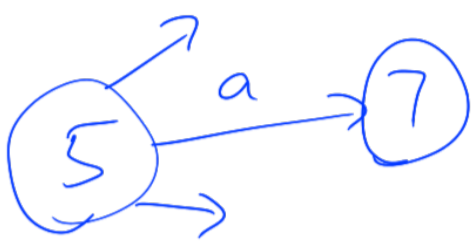
$$\Pr(X_i = a, Q_i = 7 \mid Q_{i-1} = 5) = \ ?$$

labeled data $\rightsquigarrow$ rel. freq.

unlabeled data $\rightsquigarrow$ expectation maximization (EM)

6.4 Relative frequency and likelihood

So in general: when our hypothesis space is set up with parameters corresponding to each possible outcome of a multinomial distribution, taking probabilities to be relative frequencies maximizes likelihood.



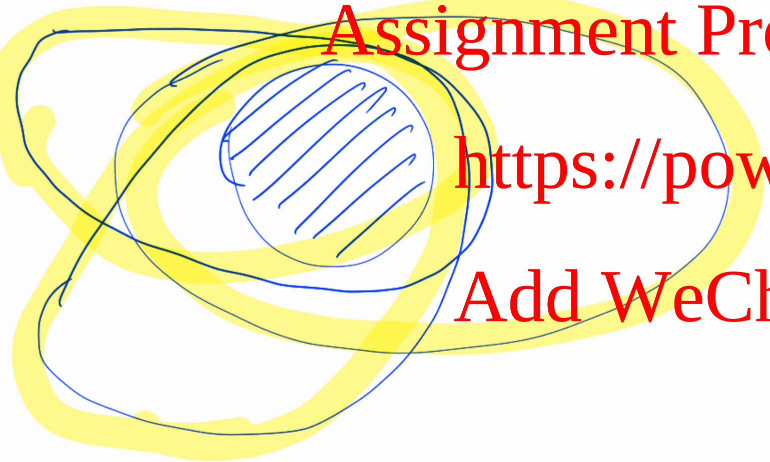
WFSAs

coda → 0.9

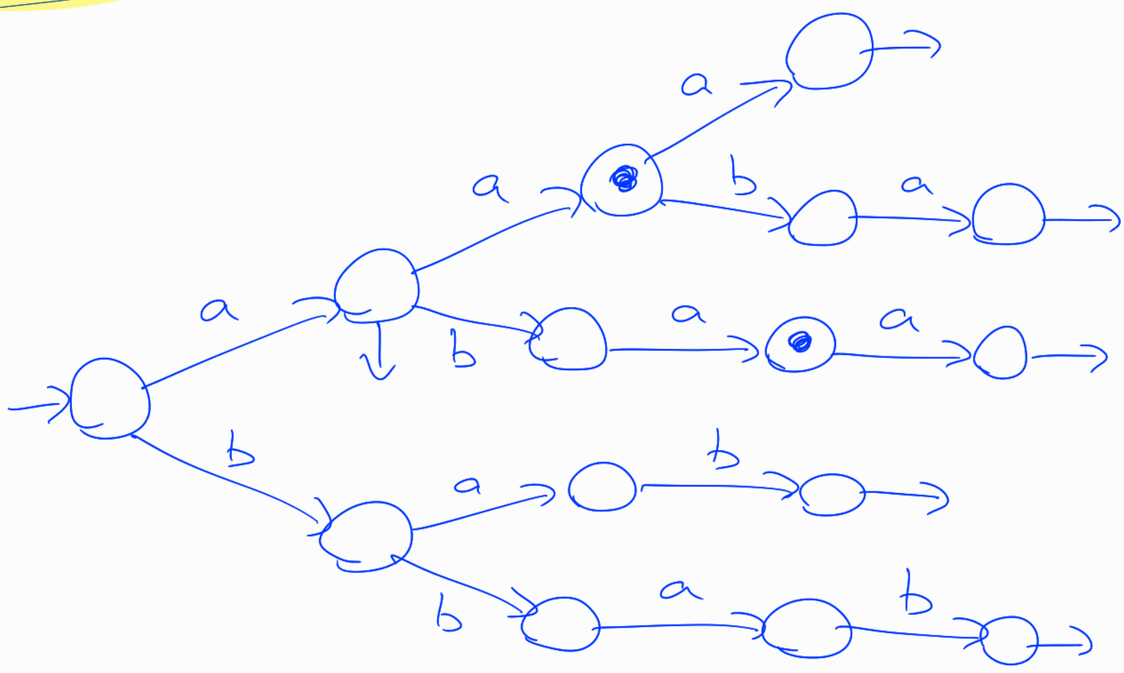
onset → 1.0

"well formedness"

	w/a	prob
The dog barked	0.99	0.001
The cat barked	0.99	0.00000001
What do you wonder who bought?	0.6	0.0001
What do you think John bought?	0.9	
Which aardvark do you think the swimming pool met?	0.9	0.0000000001



a
aaa
aaba
abaa
bab
bbab



first = last
odd as